Markov chain Monte Carlo with mixtures of singular distributions

Raphael Gottardo and Adrian E. Raftery

Technical Report no. 470
Department of Statistics
University of Washington

October 7, 2004

Abstract

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 paper, we introduce a simple framework for using MCMC algorithms in Bayesian computation with mixtures of singular distributions. The idea is to find a common dominating measure that allows the use of traditional Metropolis-Hastings algorithms. We show how our formulation can be used in Bayesian model selection. Using our formulation, when the full conditionals are available, the Gibbs sampler can be used. We compare our formulation with the reversible jump approach, and show that the two are closely related. We give results for four examples, involving testing a normal mean, variable selection in regression, and hypothesis testing for differential expression under multiple conditions in gene expression data. This allows us to compare the three methods considered: Metropolis-Hastings with singular measures, Gibbs sampler with singular measures, and reversible jump. In our examples, we found the Gibbs sampler with singular measures to be more precise and to need considerably less computer time than the other methods.

Contents

List of Tables

List of Figures

1 Introduction

Mixtures of singular measures 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. On the other hand one could 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 of 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-Nykodim derivative) is not as intuitive. However the difficulty is not great and the goal of this paper 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 is general and can be used with general dominating measures such as sums of singular measures, they are rarely used in practice. Simple mixtures of singular distributions such as a point mass and a continuous distribution (with respect to 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), perhaps because of the nonstandard formulation. Such mixtures also arise with Dirichlet processes in Bayesian density estimation (Escobar and West 1995; Neal 2000). Neal (2000) devised several MCMC algorithms for Dirichlet process mixture models. There is a need for a general formulation for MCMC computations with mixtures of singular measures. Note that the term "mixture of 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 singular measures 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 sub-manifold of

$\mathbb{R}^n$. Recently, there has been a great deal of work on MCMC algorithms for Bayesian model selection. Madigan and York (1995) and Raftery, Madigan, and Hoeting (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 singular measures, 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 show that it can be used as an alternative to of reversible jump. Their approach is limited and computationally intensive. In this paper, we show that reversible jump can be viewed in terms of a mixture of singular distributions. Using examples, we show the direct relationship between the two formulations.

The paper is organized as follows. In Section 2, we introduce some notation and show how one can derive Radon-Nikodym derivatives (densities) for mixtures of 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 singular distributions as 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 Mixtures of singular measures and Radon-Nikodym derivatives

We denote the n-dimensional Lebesgue measure by λ_n, the Dirac measure concentrated at $\boldsymbol{x}$ by $\delta_{\boldsymbol{x}}$, and the n-dimensional Borel σ-algebra of subsets of $\mathbb{R}^n$ by $\mathcal{B}_n$. Our basic set-up is the usual n-dimensional Euclidean space $\mathbb{R}^n$ equipped with its Borel subsets $\mathcal{B}_n$. The following theorem gives a way of explicitly writing down Radon-Nikodym derivatives for mixtures of singular measures.

Theorem 1 *Let π and $\nu_i, i \in I$, be σ-finite Borel measures, where I is a countable set. Assume that the ν_i's are mutually singular, i.e. $\nu_i \perp \nu_j$ where $i, j \in I$ and $i \neq j$. Suppose*

that there exist Borel probability measures π_i *such that* $\pi = \sum_{i\in I} w_i\pi_i$ *where* $w_i \in [0,1]$, $\sum_{i\in I} w_i = 1$ *and* π_i *is dominated by* ν_i. *Then the Radon-Nikodym derivative of* π *with respect to* $\sum_{i\in I}\nu_i$ *is given by*

$$\frac{d\pi}{d\sum_{i\in I}\nu_i}(\boldsymbol{x}) = \sum_{i\in I} w_i \frac{d\pi_i}{d\nu_i}(\boldsymbol{x})\mathbf{1}_{S_i}(\boldsymbol{x}),$$

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

The proof is given in the Appendix. The theorem says that the Radon-Nikodym derivative of π with respect to the general dominating measure can be expressed as a mixture of the individual Radon-Nikodym derivatives. In general the individual Radon-Nikodym derivative of π_i with respect to ν_i is known or easy to derive. For example if ν_i is the Lebesgue measure, the Radon-Nikodym derivative is simply the usual Lebesgue density. Note, however, that the sets S_i are crucial to get a proper density. The sets S_i are there to make sure that a set does not contribute to more than one component, which would not be legitimate since the dominating measures are singular. If a set is counted more than once the density would not even integrate to one!

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, we can give a general guideline on how to derive them. In order to do so, we need to introduce a notion of dimension. The definition of dimension we need is the Hausdorff dimension which is related to the Hausdorff measure. The Hausdorff measure can be seen as an extension of the Lebesgue measure to measure sub-manifolds, and in general small sets of $\mathbb{R}^n$. In $\mathbb{R}^n$, the n-dimensional Hausdorff measure and the n-dimensional Lebesgue measure agree. The Hausdorff dimension of the set $A \subset \mathbb{R}^n$ is defined to be

$$\dim_H(A) \equiv \inf\{0 \leq s \leq \infty | \mathcal{H}_s(A) = 0\},$$

where $\mathcal{H}_s$ is the s-dimensional Hausdorff measure; for the full definition see Falconer (2003). An important consequence of the definition is that a smooth k-dimensional sub-manifold of $\mathbb{R}^n$ has dimension k. Here, we restrict ourselves to integer-valued dimensions, as we are not interested in Cantor-like or fractal-like sets with fractional Hausdorff dimensions. In most applications, the supports of the measures ν_i will be submanifolds or portions of submanifolds. The only relevant properties of Hausdorff dimension for this paper are that a countable set has dimension 0, lines and segments (and in general smooth curves) have dimension 1, planes and smooth surfaces have dimension 2, and so on.

Having an exact definition of dimension, we can define each set S_i in terms of the supports of the measures ν_i. If we denote by C_i the support of ν_i, $1 \leq i \leq n$, the set S_i can be written as

$$S_i = C_i \setminus \{\cup_{j \in I_i} C_j\}, \tag{1}$$

where $I_i = \{j \neq i : \dim_H(C_j) \leq \dim_H(C_i)\}$. Note that the exact value of the dimension is not too important, as long as we can order the supports by dimension. When the supports are nice sets such as lines, planes, etc, ordering the sets by dimension is easy and intuitive.

Example 1: Mixture of discrete and continuous random variables. Let X_1 be a discrete random variable with countable support C_1 and induced probability measure $\pi_1 = \sum_{x \in C_1} \pi_1(x)\delta_x$ dominated by the discrete measure $\nu = \sum_{x \in C_1} \delta_x$. Let X_2 be a continuous random variable with probability measure π_2 dominated by the Lebesgue measure λ. Define a third random variable Y equal to X_1 with probability w and to X_2 with probability $(1-w)$. The support of ν, C_1, is countable and has Hausdorff dimension 0, and $C_2 \equiv \mathbb{R}$ has Hausdorff dimension 1. Therefore, using (1), we define the sets $S_1 = C_1$ and $S_2 = C_2 \setminus C_1$. Applying Theorem 1, we obtain the Radon-Nikodym derivative of π with respect to $(\nu + \lambda)$, namely

$$\begin{aligned}
\frac{d\pi}{d(\nu+\lambda)}(x) &= w\frac{d\pi_1}{d\nu}(x)\mathbf{1}_{S_1}(x) + (1-w)\frac{d\pi_2}{d\lambda}(x)\mathbf{1}_{S_2}(x) \;\; a.e.\;\; \nu+\lambda, \\
\frac{d\pi}{d(\nu+\lambda)}(x) &= w\sum_{y \in C_1} \pi_1(y)\delta_y(x) + (1-w)f(x)\mathbf{1}_{C_2 \setminus C_1}(x) \;\; a.e.\;\; \nu+\lambda,
\end{aligned}$$

where $f(x)$ is the Lebesgue density (Radon-Nikodym) derivative of X_2, and "a.e. $\nu + \lambda$" means almost everywhere with respect to the measure $\nu + \lambda$, i.e. everywhere except on a set that has $(\nu + \lambda)$-measure equal to zero. Note that it is crucial to remove C_1 from C_2 for the density to be a valid density with respect to $(\nu + \lambda)$.

3 Application to MCMC

In this section, we consider Markov chain Monte Carlo algorithms for mixtures of singular distributions. We show how it is possible to use the regular Metropolis-Hastings algorithm, where Theorem 1 is used to derive the required densities.

Suppose that we wish to sample from a distribution for a variable $\boldsymbol{x}$ with distribution π and associated measure $\pi(\boldsymbol{dx})$ dominated by a σ-finite measure ν. In this paper, we shall be concerned with a posterior distribution $\pi(\boldsymbol{x}|\boldsymbol{y})$ even though the result presented could be applied in a more general setting. Abusing notation, we will use $\pi(\boldsymbol{x})$ to denote the target

distribution, in this case $\pi(\boldsymbol{x}|\boldsymbol{y})$. One of the most widely used algorithms for generating such chains is the Metropolis-Hastings procedure. 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 Lebesgue measure or to counting measure.

In Markov chain Monte Carlo, one constructs a Markov Chain with invariant distribution π. Let Q be a Markov transition kernel of the form

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

and define

$$\alpha(\boldsymbol{x}, \boldsymbol{x'}) = \min\left\{\frac{\pi(\boldsymbol{x'})q(\boldsymbol{x'}, \boldsymbol{x})}{\pi(\boldsymbol{x})q(\boldsymbol{x}, \boldsymbol{x'})}, 1\right\}. \tag{2}$$

If the chain is currently at a point $\boldsymbol{X_n} = \boldsymbol{x}$, then a new candidate is generated according to the distribution $Q(\boldsymbol{x}, \cdot)$ and the new point is accepted with probability $\alpha(\boldsymbol{x}, \boldsymbol{y})$. This defines a Metropolis-Hastings kernel K with density k,

$$k(\boldsymbol{x}, \boldsymbol{x'}) = \alpha(\boldsymbol{x}, \boldsymbol{x'})q(\boldsymbol{x}, \boldsymbol{x'})\left(1 - \delta_{\boldsymbol{x}}(\boldsymbol{x'})\right) + \left(1 - \int \alpha(\boldsymbol{x}, \boldsymbol{x'})q(\boldsymbol{x}, \boldsymbol{x'})\boldsymbol{dx'}\right)\delta_{\boldsymbol{x}}(\boldsymbol{x'}). \tag{3}$$

If K is π-irreducible, Harris-recurrent and aperiodic, the Markov chain formed will converge to the unique stationary distribution 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 as 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.

The general result extends to mixtures of (σ-finite) singular measures. The main difference 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 below. In the next section, we will show the relationship with the reversible jump methodology of Green (1995).

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 that the ν_i's are mutually singular.

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, it will be convenient to write the kernel Q as

$$Q(\boldsymbol{x}, \boldsymbol{dx'}) = \sum_{i \in I} p_i(\boldsymbol{x}) Q_i(\boldsymbol{x}, \boldsymbol{dx'}), \tag{4}$$

where the Q_i's are singular transition kernels in the sense that $Q_i(\boldsymbol{x}, \boldsymbol{dx'}) = q_i(\boldsymbol{x}, \boldsymbol{x'})\nu_i(\boldsymbol{dx'})$, and the $p_i(\boldsymbol{x})$'s are $\mathcal{B}_n$-measurable functions satisfying $\sum_{i \in I} p_i(\boldsymbol{x}) = 1$. Clearly, Q is a transition kernel since for each fixed $\boldsymbol{x}$, $Q(\boldsymbol{x}, \cdot)$ is a measure on $\mathcal{B}_n$ and

$$Q(\boldsymbol{x}, \mathbb{R}^n) = \sum_{i \in I} p_i(\boldsymbol{x}) Q_i(\boldsymbol{x}, \mathbb{R}^n) = \sum_{i \in I} p_i(\boldsymbol{x}) = 1.$$

In most cases it will be convenient to take

$$p_i(\boldsymbol{x}) \equiv \sum_{j \in I} p_{ji} \mathbf{1}_{S_j}(\boldsymbol{x}), \tag{5}$$

where the sets S_j are defined by (1). For Q to be π-irreducible it is necessary for the discrete transition kernel $P \equiv (p_{ij})$ to be irreducible on the set I. The π-irreducibility of Q and the associated Metropolis-Hastings kernel will depend on the Q_i's and will need to be verified for each case. Irreducibility is usually easily checked. In some of the examples explored here, we chose the proposal

$$Q(\boldsymbol{x}, \boldsymbol{dx'}) = \sum_{i \in I} p_i(\boldsymbol{x}) Q_i(\boldsymbol{dx'}), \tag{6}$$

where $p_i(\boldsymbol{x})$ is defined by (5) and Q_i is an independent kernel satisfying $Q_i(A) > 0$ for any $A \in \mathcal{B}_n$ such that $\pi_i(A) > 0$.

For each $\boldsymbol{x} \in \mathbb{R}^n$, $Q(\boldsymbol{x}, \cdot)$ as given by (4) is a mixture of singular proposal measures and therefore the density can easily be obtained as a mixture of the densities q_i using Theorem 1. 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). Tipically, for the Metropolis-Hastings algorithm, irreducibility implies Harris recurrence (Tierney 1994). Additionaly if $K(\boldsymbol{x}, \cdot)$ is absolutely continuous with respect to π, Harris recurrence is guaranteed for the Gibbs sampler (Tierney 1994). This remains true for hybrid samplers such as the variable-at-a-time Metropolis-Hastings algorithm, though the proof is not trivial (Chan

and Geyer 1994). We refer the reader to Tierney (1994) and Roberts and Tweedie (1996) for more details on convergence properties of Metropolis-Hastings algorithms.

4 Applications

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

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

$$\begin{aligned} y_j &= \gamma + \epsilon_j, \\ (\epsilon_j|\psi) &\sim \mathrm{N}(0, \psi^{-1}). \end{aligned} \tag{7}$$

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

$$\gamma \sim w\delta_0 + (1-w)\mathrm{N}(0, \psi_\gamma^{-1}), \tag{8}$$

which is a mixture of a point mass at 0 and a Gaussian distribution. Using Example 1, its density is $w\delta_0(\gamma) + (1-w)\mathrm{N}(\gamma; 0, \psi_\gamma^{-1})(1-\delta_0(\gamma))$.

We let the prior for the precision ψ be Gamma(ξ_1, ξ_2), which has mean ξ_1/ξ_2 and variance ξ_1/ξ_2^2. Here the target distribution is $\pi(\gamma, \psi|\boldsymbol{y})$. As in traditional MCMC, we can form two kernels (one for each parameter) and combine them to form an ergodic Markov chain with stationary distribution π. The update for ψ presents nothing complicated as the dominating measure is the Lebesgue measure. Therefore we shall be concerned only with the update of γ conditional on ψ.

We can use the usual Metropolis-Hastings algorithm to update γ. We need to define an irreducible Markov transition kernel (proposal), $Q(\gamma, d\gamma') = q(\gamma, \gamma')\nu(d\gamma)$, that is absolutely continuous with respect to $\nu = \delta_0 + \lambda$, where δ_0 is the Dirac measure concentrated at 0, and λ is Lebesgue measure. In other words we have to make sure we propose a move to zero as well as to the remainder of the real line. This can be done using (4) which we rewrite as

$$q(\gamma, \gamma') = p(\gamma)\delta_0(\gamma') + (1-p(\gamma))q^*(\gamma, \gamma'), \tag{9}$$

where $p(\gamma)$ is a number between 0 and 1, possibly depending on γ. In other words, with probability $p(\gamma)$ we propose a move to zero (independently of the current values) and with probability $(1-p(\gamma))$ we propose a value according to q^*, a kernel that is absolutely continuous with respect to Lebesgue measure. For the Metropolis-Hastings algorithm, the acceptance probability is given by

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

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 (9) is clearly not symmetric as $q(0,\gamma') \neq q(\gamma',0)$ for $\gamma' \neq 0$. Any other proposal will have the same problem because of the singularity between the two measures δ_0 and λ.

The Gibbs sampler can also be used, as the full conditional is available and is given by

$$\pi(\gamma|\psi,\boldsymbol{y}) = w^*\delta_0 + (1-w^*)\mathrm{N}\left(\psi\sum_i y_i/(n\psi+\psi_\gamma),(n\psi+\psi_\gamma)^{-1}\right),$$

where

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

We now compare the performance of the Gibbs sampler to the Metropolis-Hastings algorithm with two different proposals. The first proposal is a generalization of the random walk Metropolis proposal whose density is given by

$$q_1(\gamma,\gamma') = 0.5\delta_0(\gamma') + 0.5\mathrm{N}(\gamma';\gamma,\sigma_1^2)(1-\delta_0(\gamma')),$$

where σ_1^2 is a fixed number. From γ, this proposes 0 with probability 0.5, and proposes a random walk step with probability 0.5. The second proposal q_2, is the concatanation of two proposals,

$$q_{21}(\gamma,\tilde{\gamma}) = (1-\delta_0(\gamma))\delta_0(\tilde{\gamma}) + \delta_0(\gamma)\mathrm{N}(\tilde{\gamma};\hat{\gamma},\sigma_2^2)(1-\delta_0(\tilde{\gamma})),$$

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

$$q_{22}(\tilde{\gamma},\gamma') = \delta_0(\tilde{\gamma})\delta_{\gamma'}(\tilde{\gamma}) + (1-\delta_0(\tilde{\gamma}))q^+(\gamma'),$$

where q^+ is the full conditional of γ for the model with a non zero component. From γ, this proposes 0 if $\gamma \neq 0$ and a non-zero γ if $\gamma = 0$. In the latter case, this non zero γ' can be simulated in two steps. The first step based on q_{21} proposes $\tilde{\gamma}$ from a $\mathrm{N}(\hat{\gamma},\sigma_2^2)$ density where $\hat{\gamma}$ is the sample mean and σ_2^2 is a fixed number. If this is accepted, the second step is a "partial" Gibbs step based on the full conditional for the model with the non-zero

component. This last step, known as within model move in the reversible jump literature, is not necessary for the Markov Chain to be ergodic but greatly improve 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 (7) and (8) using $w = 0.5$, *i.e.* each component is equally likely *a priori*. We fixed $\psi_\gamma = 0.01$, $\xi_1 = 1$ 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.

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). All three methods considered did well in estimating this posterior probability, with accurate estimates based on 10,000 iterations, and low variability of the estimates, as assessed from ten million iterations divided into batches. The Metropolis-Hastings algorithm based on the proposal q_2 performed better than the one based on q_1. It moved between components almost twice as often, not surprisingly since q_2 forces moves between components. The Gibbs sampler did essentially as well as Metropolis-Hastings with proposal q_2 in estimating the posterior probability, and it used much less computer time. Overall, therefore, the Gibbs sampler performed best among these three methods for this example.

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 10,000,000 iterations into 1,000 batches of 10,000 iterations each. The "truth" was obtained from ten million iterations; the estimates from the three algorithms agreed to within three digits. PMBC is the percentage of moves that were between mixture components.

	Truth	MH (q_1)		MH (q_2)		Gibbs	
		Estimate	sd	Estimate	sd	Estimate	sd
$\pi(\gamma = 0\|\boldsymbol{y})$	0.867	0.858	0.0054	0.870	0.0031	0.865	0.0036
PMBC(%)		13		25		24	
CPU (μs/iter)		23		24		5	

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 predicts an outcome variable. Given a dependent variable $\boldsymbol{Y}$ and a set of potential regressors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$, we wish to compare models of the form $\boldsymbol{Y} = \beta_0 + \boldsymbol{X}_{i_1}\beta_{i_1} + \cdots + \boldsymbol{X}_{i_q}\beta_{i_q}$, where $\boldsymbol{X}_{i_1}, \ldots, \boldsymbol{X}_{i_q}$ is a selected subset of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$. For this problem we take an approach similar to that of George and McCulloch (1997). However, our model is more robust with t-distributed errors. We assume a standard linear model to describe the relationship between the observed 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's are independent of the ϵ's, we have $\frac{\epsilon_i}{\sqrt{\varpi_i}} \sim t_{(\nu,0,\psi^{-1})}$, i.e. the errors have a t distribution with ν degrees of freedom and scale parameter ψ^{-1}. The advantage of writing the model this way is that, conditioning on the ϖ_i, the sampling errors are again normal, but with different precisions.

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

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

where w is the prior probability for each variable of being in the model, $S_{X_k}^2$ is the empirical variance of the k-th predictor, S_Y^2 is the empirical variance of the observed variables and σ_β^2 is a common variance parameter. The prior for the variance parameter σ_β^2 is taken uniform on the interval $[0, 1]$. 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 normal with a large variance centered at the least square estimate $\hat{\beta}_0$, 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 $\{1, 2, \ldots, 10, 20, \ldots, 100\}$. For each β_k, we use $w = 0.5$ which makes every model equally likely *a priori*.

As in Example 2, both the usual Metropolis-Hastings and the Gibbs sampler can be used.

The full conditionals for the β_k's are given by

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

where

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

the residual r_{ik} is defined by

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

and

$$\psi_{\beta_k} = \left(\frac{S_Y^2}{S_{X_k}^2}\sigma_\beta^2\right)^{-1}.$$

Again we wish to compare the Metropolis-Hastings algorithm with the proposals q_1 and q_2 given 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, Raftery, and Madigan (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" which is the percent of unconverted ammonia that escapes from the plant. There are three independent variables, $\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3$. The airflow $\boldsymbol{X}_1$ measures the rate of operation of the plant. The nitric oxides produced are absorbed in a counter-current absorption tower: $\boldsymbol{X}_2$ is the inlet temperature of cooling water circulating through coils in this tower and $\boldsymbol{X}_3$ 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 $\boldsymbol{X}_3$ (acid concentration) should be dropped from the model and that observations 1,3,4 and 21 are outliers.

The comparison of the different algorithms is given in Table 2. This time, the Gibbs sampler outperforms the other two algorithms. This is not surprising. In this example, during the posterior exploration, variables will be in and out of the model changing the relative estimates of the β's. As a consequence it is hard to construct an efficient proposal.

Table 2: Comparison of the estimated model posterior 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 10,000,000 iterations into 1,000 batches of 10,000 iterations each. The "truth" was obtained from 10 billion iterations on the basis of which the estimates from the three algorithms agreed to within three digits. PMBC is the percentage of the moves that were between mixture components. The Gibbs sampler performed better than the Metropolis-Hastings samplers.

	Truth	MH (q_1)		MH (q_2)		Gibbs	
		Estimate	sd	Estimate	sd	Estimate	sd
$\pi(\beta_1 \neq 0 \mid \boldsymbol{y})$	1.000	1.000	0.009	1.000	0.000	1.000	0.000
$\pi(\beta_2 \neq 0 \mid \boldsymbol{y})$	0.883	1.000	0.282	1.000	0.140	0.808	0.125
$\pi(\beta_3 \neq 0 \mid \boldsymbol{y})$	0.112	0.462	0.233	0.045	0.110	0.131	0.093
PMBC % (β_1)		0.0		0.0		0.0	
PMBC % (β_2)		0.2		0.2		0.5	
PMBC % (β_3)		0.7		1.2		1.6	
CPU (μs/iter)		105		110		75	

The Gibbs sampler is adaptive and automatic in the sense that the probability of switching between components is calculated at each iteration and depends on the current value of the other coefficients currently in the model. The average number of moves between components is the highest and the estimated variances for the estimate of the posterior probabilities are smaller. Finally the computing time is greatly reduced.

In this example, we are usually interested in the posterior probabilities of all $2^3 = 8$ possible models. This can be easily computed from the β_k's. The model posterior probabilities are summarized in Table 3. The posterior probabilities are consistent with the general consensus about the data.

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.	0.12	0.78	0.01	0.09

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

known to be outliers, are downweighted by our model.

Table 4: Posterior weights, i.e. posterior means of the ϖ's, associated with each observation of the *Stack Loss* Data. Observations with small weights are downweighted during the estimation. Observations 1, 3, 4 and 21 have small weights suggesting that they might be outliers.

Obs.	1	2	3	4	5	6	7	8	9	10	11
Weight	0.41	0.91	0.38	0.28	1.17	0.95	1.15	1.19	1.02	1.27	1.26
Obs.	12	13	14	15	16	17	18	19	20	21	
Weight	1.21	0.62	0.71	1.19	1.31	1.37	1.38	1.16	0.73	0.23	

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 are part of a new class of biotechnologies that allows the monitoring of thousands of genes simultaneously under different biological or experimental conditions. One of the main tasks with microarrays is the identification of differentially expressed genes under the different conditions. Hedenfalk et al. (2001) conducted 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. Hedenfalk et al. (2001) examined 22 breast cancer tumor samples: 7 tumors with BRCA1, 8 tumors with BRCA2 and 7 sporadic tumors, i.e. with neither mutation. The goal of the experiment was to study the expression patterns of 3226 genes under the three conditions and detect genes whose expression changed in at least one of the conditions. For illustrative purposes, we show the results for one gene. 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:

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

where γ_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 $\gamma_1 = \gamma_2 = \gamma_3$. In this example the alternative hypothesis is more complex due to the number of patterns possible. The prior distribution needs to include all such possible patterns. We therefore

Table 5: Log transformed measurements of one gene of the BRCA dataset. Each column corresponds to a different condition.

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	

consider the following prior, whose density is

$$\begin{aligned}(\boldsymbol{\gamma}|\boldsymbol{\psi}_{\boldsymbol{\gamma}}, \boldsymbol{w}) \;\sim\;& w_1 \mathrm{N}(\gamma_1; 0, \psi_{\gamma_{123}}^{-1})\mathbf{1}_{[\gamma_1=\gamma_2=\gamma_3]} \\ +\;& w_2 \mathrm{N}(\gamma_1; 0, \psi_{\gamma_1}^{-1})\mathrm{N}(\gamma_2; 0, \psi_{\gamma_{23}}^{-1})\mathbf{1}_{[\gamma_1\neq\gamma_2=\gamma_3]} \\ +\;& w_3 \mathrm{N}(\gamma_2; 0, \psi_{\gamma_2}^{-1})\mathrm{N}(\gamma_1; 0, \psi_{\gamma_{13}}^{-1})\mathbf{1}_{[\gamma_1=\gamma_3\neq\gamma_2]} \\ +\;& w_4 \mathrm{N}(\gamma_3; 0, \psi_{\gamma_3}^{-1})\mathrm{N}(\gamma_1; 0, \psi_{\gamma_{12}}^{-1})\mathbf{1}_{[\gamma_1=\gamma_2\neq\gamma_3]} \\ +\;& w_5 \mathrm{N}(\gamma_1; 0, \psi_{\gamma_1}^{-1})\mathrm{N}(\gamma_2; 0, \psi_{\gamma_2}^{-1})\mathrm{N}(\gamma_3; 0, \psi_{\gamma_3}^{-1})\mathbf{1}_{[\gamma_1\neq\gamma_2\neq\gamma_3]}, \end{aligned} \tag{10}$$

where $\boldsymbol{\psi}_{\boldsymbol{\gamma}} = (\psi_{\gamma_1}, \psi_{\gamma_2}, \psi_{\gamma_3}, \psi_{\gamma_{12}}, \psi_{\gamma_{13}}, \psi_{\gamma_{23}}, \psi_{\gamma_{123}})$ is the vector of precisions and $\boldsymbol{w}$ 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 dominating measure on $(\mathbb{R}^3, \mathcal{B}_3)$, namely

$$\nu(\cdot) = \lambda_1(\Delta \cap \cdot) + \lambda_2(P_{\gamma_1} \cap \cdot) + \lambda_2(P_{\gamma_2} \cap \cdot) + \lambda_2(P_{\gamma_3} \cap \cdot) + \lambda_3(\cdot),$$

where Δ is the line $\gamma_1 = \gamma_2 = \gamma_3$, P_{γ_1} is the plane $\gamma_2 = \gamma_3$, P_{γ_2} is the plane $\gamma_1 = \gamma_3$, P_{γ_3} is the plane $\gamma_1 = \gamma_2$ and λ_k denotes the k dimensional Lebesgue measure. In this case, we directly defined the distribution in terms of the density given by (10), in order to minimize the space. The distribution and density are well defined, and so we can use the usual Metropolis-Hastings algorithm. The target distribution is $\pi(\boldsymbol{\gamma}, \psi|\boldsymbol{y})$, but we are concerned only with $\boldsymbol{\gamma}$.

It is harder to construct an efficient proposal for the Metropolis-Hastings algorithm than in the last example because of the greater number of 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

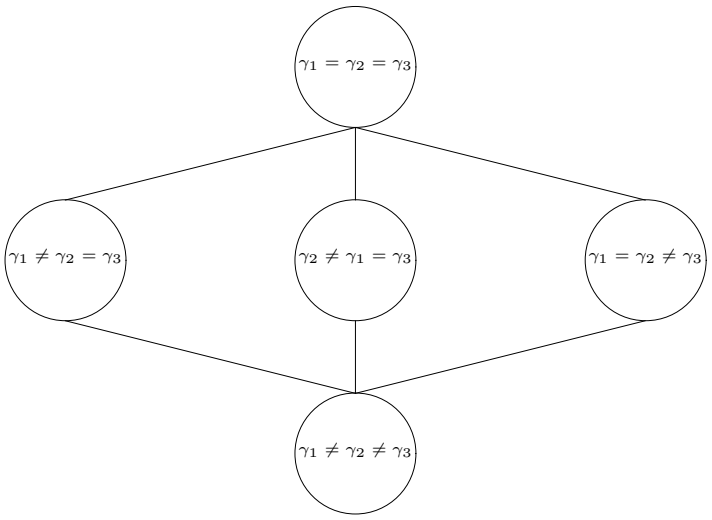

Figure 1: Local move graph for the three-way comparison proposal. Δ is the line $\gamma_1 = \gamma_2 = \gamma_3$.

form given by equations (4) and (5), with $p_{ij} > 0$ if there is an edge between i and j (Figure 1). We consider only one kernel, which is similar to q_2 of Example 2-3 adapted to the local move structure.

From the current value of $\boldsymbol{\gamma}$, 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 $\boldsymbol{\gamma}$ from a Gaussian proposal centered at the least squares estimates for the corresponding component. The widths of the proposals are taken to be the same for all the Gaussian proposals. For example, given that the current $\boldsymbol{\gamma}$ is from the null component, i.e. $\gamma_1 = \gamma_2 = \gamma_3$, we first select one of the three accessible components with probability 1/3 each. Then we generate new values for $\boldsymbol{\gamma}$, independent of the current values, from a Gaussian proposal centered at the constrained least squares estimates. If the second component is chosen then the constraint is $\gamma_2 = \gamma_3$. Similarly to q_2, within a given component of the mixture each parameter is updated using a partial Gibbs step based on the full conditional for the model with only that particular component. This last step is known as within model move in the reversible jump literature. Once again, the full conditional for $\boldsymbol{\gamma}$ is available and the Gibbs sampler can be used. The full conditional is given in Appendix A.2.

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 aver-

Table 6: Comparison of the model posterior probabilities for the five models computed with each algorithm in Example 4. Estimates were computed from 10,000 iterations with 1,000 burn-in iterations. The standard deviations were computed by dividing a chain of 10,000,000 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 moves that are between mixture components. The estimates of the posterior probabilities agree well. The standard deviations are smaller for the Gibbs sampler.

	Truth	MH		Gibbs		RJ	
		Estimate	sd	Estimate	sd	Estimate	sd
$\pi(\gamma_1 = \gamma_2 = \gamma_3 \mid \boldsymbol{y})$	0.085	0.081	0.012	0.084	0.004	0.081	0.009
$\pi(\gamma_1 \neq \gamma_2 = \gamma_3 \mid \boldsymbol{y})$	0.023	0.022	0.006	0.025	0.001	0.026	0.004
$\pi(\gamma_2 \neq \gamma_2 = \gamma_3 \mid \boldsymbol{y})$	0.850	0.851	0.017	0.848	0.004	0.824	0.013
$\pi(\gamma_1 = \gamma_2 \neq \gamma_3 \mid \boldsymbol{y})$	0.007	0.008	0.002	0.007	0.001	0.009	0.001
$\pi(\gamma_1 \neq \gamma_2 \neq \gamma_3 \mid \boldsymbol{y})$	0.035	0.038	0.004	0.036	0.002	0.061	0.003
PMBC (%)		5		24		8	
CPU (μs/iter)		17		5		17	

age number of moves between components is much greater for the Gibbs sampler than for Metropolis-Hastings, indicating better mixing, and the computing time was much shorter.

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. 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 $\pi(k, \boldsymbol{\theta}_k|\boldsymbol{y})$, which can be decomposed as

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

Using this formulation, the sample space can be represented by $S = \cup_{k\in\mathcal{K}}\{k\} \times \mathbb{R}^{n_k}$. Let $\boldsymbol{x} = (k, \boldsymbol{\theta}_k)$, and let $\pi(\boldsymbol{x}) \equiv \pi(\boldsymbol{x}|\boldsymbol{y})$ denote the the target distribution. Even though the dimension of $\boldsymbol{x}$ is allowed to change, Green (1995) showed that it is still possible to use the standard 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 $\boldsymbol{x}$, we generate r random numbers $\boldsymbol{u}$ 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: $\boldsymbol{x}' = h(\boldsymbol{x}, \boldsymbol{u})$. The reverse transformation will be made with the aid of random numbers $\boldsymbol{u}' \sim g'$, giving $\boldsymbol{x} = h'(\boldsymbol{x}', \boldsymbol{u}')$. Assuming that the transformation from $(\boldsymbol{x}, \boldsymbol{u})$ to $(\boldsymbol{x}', \boldsymbol{u}')$ is a diffeomorphism, Green (1995) showed that a valid choice for the acceptance probability in the usual Metropolis-Hastings algorithm is given by

$$\alpha(\boldsymbol{x}, \boldsymbol{x}') = \min\left\{1, \frac{\pi(\boldsymbol{x}')g'(\boldsymbol{u}')}{\pi(\boldsymbol{x})g(\boldsymbol{u})}\left|\frac{\partial(\boldsymbol{x}', \boldsymbol{u}')}{\partial(\boldsymbol{x}, \boldsymbol{u})}\right|\right\}, \tag{11}$$

where

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

The reversible jump formulation was introduced to handle cases where the dimension of the parameter vector $\boldsymbol{\theta}$ can change from model to model. The notion of dimension here is different from the Hausdorff dimension introduced earlier. However one could fix the number of parameters to be the same and allow some of the parameters to vanish or to lie in a hyperplane, reducing the Hausdorff dimension of the support. One example of this is provided by nested models with linear constraints on the parameters, and MCMC with singular measures could be applied to that case. When the Jacobian term is equal to one, the two formulations are equivalent in the sense that one could write down two algorithms with the same acceptance probability. This is the case in Example 2, as we will show below. The Jacobian term present in (11) 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 (11) for the detailed balance condition to be satisfied (Green 1995). Such a change of parameter is not necessary.

Thinking about the problem in terms of singular measures 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, $\mathcal{M}_0 : \gamma = 0$ and $\mathcal{M}_1 : \gamma \neq 0$. For the first model $\theta_0 = (\psi)$, and for the second model

we have one more parameter, $\boldsymbol{\theta}_1 = (\gamma, \psi)$. The sample space is $S = \{0\} \times \theta_0 \cup \{1\} \times \boldsymbol{\theta}_1$. As ψ is common to the two models, we shall be concerned only with the update of θ conditional on ψ. At the current state $\boldsymbol{x} = (k, \boldsymbol{\theta}_k)$, we can generate a new value $\boldsymbol{x'}$ according to

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

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

$$\alpha(\boldsymbol{x}, \boldsymbol{x'}) = \begin{cases} 1 \wedge r(\boldsymbol{x}, \boldsymbol{x'}) & \text{if} \quad k = 1 \\ 1 \wedge r(\boldsymbol{x'}, \boldsymbol{x})^{-1} & \text{if} \quad k = 0, \end{cases} \tag{12}$$

where

$$r(\boldsymbol{x}, \boldsymbol{x'}) = \frac{f(\boldsymbol{y}|\psi)(1-w)g(\gamma)}{f(\boldsymbol{y}|\gamma, \psi)\mathrm{N}(\gamma; 0, \psi_\gamma^{-1})w}. \tag{13}$$

On the other hand we could use the prior mixture distribution given by (8), whose density is $w\delta_0 + (1-w)\mathrm{N}(\gamma; 0, \psi_\gamma^{-1})(1-\delta_0)$, and the proposal

$$q(\gamma, \gamma') = \begin{cases} \delta_0(\gamma') & \text{if} \quad \gamma \neq 0 \\ g(\gamma') & \text{if} \quad \gamma = 0, \end{cases}$$

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 (12); the two formulations are equivalent. Thinking about the problem with a common dominating measure allows us to use the Metropolis-Hastings algorithm in a natural way. It also allows us to use the Gibbs sampler.

Usually, in reversible jump algorithms, there are two main types of moves: between-model moves and within-models 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 a reversible jump algorithm. In Example 2, where it was easy to construct an efficient proposal, the reversible jump performed slightly better than the Gibbs sampler in terms of mixing, but took substantially more computer time for the same number of iterations. 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,

$$\begin{aligned}
\mathcal{M}_1 &: \gamma_1 = \gamma_2 = \gamma_3 \\
\mathcal{M}_2 &: \gamma_1 \neq \gamma_2 = \gamma_3 \\
\mathcal{M}_3 &: \gamma_1 = \gamma_3 \neq \gamma_2 \\
\mathcal{M}_4 &: \gamma_1 = \gamma_2 \neq \gamma_3 \\
\mathcal{M}_5 &: \gamma_1 \neq \gamma_2 \neq \gamma_3,
\end{aligned}$$

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. In each model, the parameter vector can be written as

$$\begin{aligned}
\boldsymbol{\theta}_1 &= (\gamma_{123}) \\
\boldsymbol{\theta}_2 &= (\gamma_1, \gamma_{23}) \\
\boldsymbol{\theta}_3 &= (\gamma_2, \gamma_{13}) \\
\boldsymbol{\theta}_4 &= (\gamma_3, \gamma_{12}) \\
\boldsymbol{\theta}_5 &= (\gamma_1, \gamma_2, \gamma_3),
\end{aligned}$$

and the sample space is given by $S = \cup_i \{i\} \times \boldsymbol{\theta}_i$. Since ψ is common to all models, we can update $\boldsymbol{\theta}$ conditionally on ψ.

Similarly to Example 3, if g as given by (11) is the proposal used in the Metropolis-Hastings algorithm, 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 induce 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,

$$\boldsymbol{x'} = \begin{cases} (1, \frac{\gamma_1 + \gamma_{23}}{2}) & \text{if} \quad k = 2 \\ (2, \gamma_{123} - u, \gamma_{123} + u) & \text{if} \quad k = 1, \end{cases}$$

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

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

where

$$r(\boldsymbol{x}, \boldsymbol{x'}) = \frac{f(\boldsymbol{y}|\psi, \gamma_{123}) w_1 \mathrm{N}(\gamma_{123}; 0, \psi_{\gamma_{123}}^{-1})}{f(\boldsymbol{y}|\psi, \gamma_1, \gamma_{23}) w_2 \mathrm{N}(\gamma_1; 0, \psi_{\gamma_1}^{-1}) \mathrm{N}(\gamma_{23}; 0, \psi_{\gamma_{23}}^{-1})} \frac{p_{12} g((\gamma_1 - \gamma_{23})/2)}{p_{21}} \frac{1}{2},$$

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,

$$\boldsymbol{x'} = \begin{cases} (2, \gamma_1, \frac{\gamma_2 + \gamma_3}{2}) & \text{if} \quad k = 5 \\ (5, \gamma_1, \gamma_{23} - u, \gamma_{23} + u) & \text{if} \quad k = 2, \end{cases}$$

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

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

where

$$r(\boldsymbol{x}, \boldsymbol{x'}) = \frac{f(\boldsymbol{y}|\psi, \gamma_1, \gamma_{23}) w_2 \mathrm{N}(\gamma_1; 0, \psi_{\gamma_1}^{-1}) \mathrm{N}(\gamma_{23}; 0, \psi_{\gamma_{23}}^{-1})}{f(\boldsymbol{y}|\psi, \gamma_1, \gamma_2, \gamma_3) w_5 \mathrm{N}(\gamma_1; 0, \psi_{\gamma_1}^{-1}) \mathrm{N}(\gamma_2; 0, \psi_{\gamma_2}^{-1}) \mathrm{N}(\gamma_3; 0, \psi_{\gamma_3}^{-1})} \frac{p_{25} g((\gamma_2 - \gamma_3)/2)}{p_{52}} \frac{1}{2},$$

and p_{ij} is the probability of choosing a move from model i to model j. The acceptance ratio for a move between 5 and 3 (or 4) would be the same with obvious changes in notation.

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 (4). However, the improvement is not great and the Gibbs sampler 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 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 singular distributions. We have analyzed four 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, Giudici, and Roberts 2003). Using a simple example we have shown the relation 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 is a matter of taste. 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 with thousands of genes. Both computation and tuning would be a serious problem. In the past few years there has been some progress towards automatic reversible jump algorithms (Green 2003).

On the other hand, in some highly complex model selection problems where there is a large difference of dimension between the models, it might be hard to write the prior as a mixture of singular distributions. 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 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, Robert, and Ryden 2003; Brooks, Giudici, and Roberts 2003). To facilitate the computations, almost all Bayesian mixture models use (semi) conjugate priors.

Finally, we would like to stress out that our formulation is different from the product space approach (Carlin and Chib 1995; Besag 1997; Godsill 2001; Dellaportas, Forster, and Ntzoufras 2002). In the product space approach, one keeps the number of parameters fixed and uses pseudo-priors to update the parameters that are not in the model currently visited. In our approach we also keep the number of parameters fixed but we allow the (Hausdorff) dimension of the support of the distribution to vary. As a consequence there is no need for pseudo-priors. 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 Hausdorff dimension of the support of the distribution.

7 Acknowledgments

The author wish to thank Julian Besag, Peter Hoff and Matthew Stephens for helpful discussions. The author also thank Charlie Geyer for pointing out that the proof in Tierney (1994) did not apply to the one-variable-at-a-time Metropolis-Hastings algorithm.

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 $\pi_i \ll \nu_i$, 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 in the three way comparison

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

$$\begin{aligned}
(\boldsymbol{\gamma}|\boldsymbol{\psi}_{\boldsymbol{\gamma}}, \boldsymbol{w}) \;\propto\; & w_1 k_{123} \mathrm{N}(\gamma_1; \gamma^*_{123}, {\psi^*_{123}}^{-1}) \mathbf{1}_{[\gamma_1=\gamma_2=\gamma_3]} \\
+\; & w_2 k_1 k_{23} \mathrm{N}(\gamma_1; \gamma^*_1, \psi^{*-1}_1) \mathrm{N}(\gamma^*_2; {\gamma_{23}}^*, {\psi^*_{23}}^{-1}) \mathbf{1}_{[\gamma_1\neq\gamma_2=\gamma_3]} \\
+\; & w_3 k_2 k_{13} \mathrm{N}(\gamma_2; \gamma^*_2, \psi^{*-1}_2) \mathrm{N}(\gamma_1; {\gamma_{13}}^*, {\psi^*_{13}}^{-1}) \mathbf{1}_{[\gamma_1=\gamma_3\neq\gamma_2]} \\
+\; & w_4 k_3 k_{12} \mathrm{N}(\gamma_3; \gamma^*_3, \psi^{-1}_3) \mathrm{N}(\gamma_1; \gamma^*_{12}, {\psi^*_{12}}^{-1}) \mathbf{1}_{[\gamma_1=\gamma_2\neq\gamma_3]} \\
+\; & w_5 k_1 k_2 k_3 \mathrm{N}(\gamma_1; \gamma^*_1, \psi^{*-1}_1) \mathrm{N}(\gamma_2; \gamma^*_2, \psi^{*-1}_2) \mathrm{N}(\gamma_3; \gamma^*_3, \psi^{*-1}_3) \mathbf{1}_{[\gamma_1\neq\gamma_2\neq\gamma_3]}
\end{aligned}$$

where

$$\psi^*_c = n_c\psi + \psi_{\gamma_c}, \quad \gamma^*_c = \psi {\psi^*_c}^{-1} \sum_{i=1}^{n_c} y_{ci},$$

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

and

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

The constants k's are given by,

$$k_i = \sqrt{\frac{\psi_{\gamma_i}}{\psi^*_i}} \exp\{-0.5\psi \sum_{j=1}^{n_i} y^2_{ij} + 0.5{\psi^*_i}^{-1} (\psi \sum_{j=1}^{n_i} y_{ij})^2\},$$

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

and

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

References

Atkinson, A. (1985). *Plots, transformations, and regression.* Clarendon Press, Oxford.

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 B 59*, 774.

Brooks, S., P. Giudici, and G. Roberts (2003). Efficient construction of reversible jump mcmc proposal distributions. *Journal of the Royal Statistical Society B 65*, 3–55.

Brownlee, K. (1965). *Statistical theory and methodology in science and engineering, 2nd edn.* Wiley, New York.

Cappé, O., C. Robert, and T. Ryden (2003). Reversible jump, birth-and-death and more general continuous time Markov chain Monte Carlo samplers. *J Royal Statistical Soc B 65*, 679–679.

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

Chan, K. S. and C. J. Geyer (1994). Comment on "Markov chains for exploring posterior distributions". *The Annals of Statistics 22*, 1747–1758.

Daniel, C. and F. Wood (1980). *Fitting equations to data*. Wiley, New York.

Dellaportas, P., J. Forster, and I. Ntzoufras (2002). On Bayesian model and variable selection using MCMC. *Statistics and Computing 12*, 27–36.

Escobar, M. D. and M. West (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association 90*, 577–588.

Falconer, K. (2003). *Fractal Geometry: Mathematical Foundations and Applications, 2nd Edition*. Halsted Press.

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

Geman, S. and D. Geman (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 R. E. McCulloch (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association 88*, 881–889.

George, E. I. and R. E. McCulloch (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.

Green, P. and A. Mira (2001). Delayed rejection in reversible jump metropolis-hastings. *Biometrika 4*, 1035–1053.

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

Green, P. J. (2003). Trans-dimensional Markov chain Monte Carlo. In P. J. Green, N. L. Hjort, and S. Richardson (Eds.), *Highly Structured Stochastic Systems*. Oxford.

Grenander, U. and M. I. Miller (1994). Representations of knowledge in complex systems (Disc: p581-603). *Journal of the Royal Statistical Society, Series B, Methodological 56*, 549–581.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika 57*, 97–109.

Hedenfalk, I., D. Duggan, Y. Chen, M. Radmacher, M. Bittner, R. Simon, P. Meltzer, B. Gusterson, M. Esteller, O. Kallioniemi, B. Wilfond, A. Borg, and J. Trent (2001). Gene-expression profiles in hereditary breast cancer. *The New England Journal of Medicine 344*(8), 539–548.

Hoeting, J. A., A. E. Raftery, and D. Madigan (1996). A method for simultaneous variable selection and outlier identification in linear regression. *Computational Statistics & Data Analysis 22*, 251–270.

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

Madigan, D. and J. York (1995). Bayesian graphical models for discrete data. *International Statistical Review 63*, 215–232.

Metropolis, N., A. W. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys. 21*, 1087–91.

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 L. Tardella (2003). A geometric approach to transdimentional Markov chain Monte Carlo. *The Canadian Journal of Statistics 31*(4), 469–482.

Phillips, D. B. and A. F. M. Smith (1995). Bayesian model comparison via jump diffusions. In *Markov chain Monte Carlo in Practice*, Chapter 13. Chapman and Hall.

Raftery, A. E., D. Madigan, and J. A. Hoeting (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association 92*, 179–191.

Richardson, S. and P. J. Green (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, Methodological 59*, 731–758.

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

Roberts, G. O. and R. L. Tweedie (1996). Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika 83*, 95–110.

Smith, M. and R. Kohn (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*(1), 40–74.

Tierney, L. (1994). Markov Chains for exploring posterior distributions (disc: P1728-1762). *The Annals of Statistics 22*, 1701–1728.

Tierney, L. (1998). A note on Metropolis-Hastings kernels for general state spaces. *Annals of Applied Probability 8*, 1–9.