A geometric approach to transdimensional MCMC

Giovanni Petris and Luca Tardella

University of Arkansas and Università di Roma "La Sapienza"

January 13, 2003

Abstract

In this paper we present some theoretical results which show how the simulation from a mixture distribution with components supported by subspaces of different dimension can be reformulated in terms of drawing from an auxiliary continuous distribution on the larger subspace and appropriately transforming the simulated auxiliary deviates. Motivated by the importance of enlarging the available Monte Carlo Markov chain (MCMC) techniques for coping with inference on varying dimensional parametric spaces, we show how our results can be fruitfully employed in some relevant types of inferential problems such as (a) model selection/averaging of nested models and (b) regeneration of Markov chains for evaluating standard deviations of estimated expectations derived from MCMC simulations. We illustrate the effectiveness of our approach and some of its appealing features compared to other currently available techniques.

1 Introduction

The use of Monte Carlo Markov chain (MCMC) algorithms has become more and more important in the last decades as the main tool to cope with statistical problems where a distribution is known up to a proportionality constant and some approximation of its features is sought for. In particular the large availability of easy-to-implement algorithms such as the Gibbs sampler and some of its variants allowed for an exponential growth of the routine application of simulation-based Bayesian inference. The theoretical development of these techniques and the enlargement of their scope of application have been boosted by their extension to distributions supported on subspaces of variable dimension starting from Carlin and Chib (1995) and the Reversible Jump

(RJ) of Green (1995). Since then a lot of research efforts have been devoted to enlarging the availability of transdimensional simulation techniques as well as the applicability of the available ones and also to trying to overcome some of their intrinsic difficulties; another crucial point still open to improvements is related to finding effective diagnostics to monitor appropriate convergence of the chain. Currently there is a substantial presence in the literature of applications of Bayesian inference via MCMC methods in multimodel transdimensional settings, such as those arising in model selection/averaging. However, it is apparent that a lot of expertise with these techniques is needed for a safe and successful implementation and often problem-specific difficulties can arise (Godsill and Troughton; 1998; Dellaportas et al.; 2002; Rotondi; 2002). There is a clear need in the Statistics community for easy-to-implement, generic approaches to Bayesian multimodel inference. This is also testified by current ongoing research mainly aimed at proposing improved variants of RJ (Al-Awhadhi et al.; 2001; Brooks et al.; 2003; Green; 2003).

The most popular approach to multimodel MCMC is currently RJ. Although the method is certainly very flexible and some attempts have recently been made towards devising effective proposal distributions (Brooks et al.; 2003), it is widely recognized that efficient implementation of RJ is problem-specific (Rotondi; 2002) and convergence diagnostics of the simulated chain deserve special care (Brooks and Giudici; 1999, 2000; Castelloe and Zimmerman; 2002). In fact, the more recent Birth-and-Death (BD) approach, proposed by Stephens (2000), was motivated, at least in part, with the need for a strategy which simplifies the problem of specifying well calibrated moves together with the corresponding need of evaluating the Jacobians of the related one-to-one tranforms. When cleverly implemented, as in the examples provided by Stephens, BD is effective and as efficient as RJ. A detailed comparison of the two approaches in the context of mixture model inference is contained in Cappé et al. (2001). However, also the BD approach suffers from the lack of efficient natural ways of specifying the birth proposals in a general context.

In this paper we develop the theoretical basis for an original, more automatic approach to multimodel MCMC for the nested-model case. The paper focuses on the problem of generating a sample from a distribution $\bar{\mu}$ – typically the posterior distribution of a parameter of interest – supported by a sequence of nested hyperplanes of $\mathbb{R}^K$. It is first shown how to construct, in a natural fashion, an absolutely continuous distribution $\bar{\tau}$ on $\mathbb{R}^K$ and a transformation ϕ from $\mathbb{R}^K$ to itself such that $\bar{\tau}\phi^{-1}$ is equal to the target distribution $\bar{\mu}$. In this way, even if one is not able to simulate directly from $\bar{\tau}$, one can easily generate a finite realization $\zeta_1, \ldots, \zeta_N$ of an ergodic Markov chain having limiting distribution $\bar{\tau}$, and then approximate a sample from $\bar{\mu}$ with $\phi(\zeta_1), \ldots, \phi(\zeta_N)$. One of the advantages of this approach is that it allows to monitor convergence and mixing properties of the simulation output through the analysis of the untransformed sample $\zeta_1, \ldots, \zeta_N$ using standard tools of fixed dimension MCMC. We also show how to use the same basic device as a building block to construct a Markov kernel with limiting distribution $\bar{\mu}$ directly in the original parameter space and how this can be useful within a Gibbs sampler (Section 3) and for simulating a regenerative chain (Section 4).

The distinguishing feature of our proposal is the automatic transformation of the mixture of the posterior densities appearing in the nested models into a comprehensive global density which has the advantage of offering a geometric intuition of the jumps between different models which are to be simulated. One of the pragmatic motivations of proposing a radically different simulation strategy stems from realizing that RJ often lacks an automatic way of being implemented and we believe the availability of other effective automatic alternatives might help more and more researchers to entertain nested models in a fully Bayesian fashion through MCMC approximations. Furthermore, with the approach we propose, there is no need to evaluate Jacobians of involved transformations.

The layout of the paper is as follows. In Section 2 we present the main theoretical results that justify the proposed method. Section 3 illustrates, in the controlled context of nested linear model setting the performance of our proposal for a simulation-based Bayesian inference. The choice of this relatively simple model for illustrative purposes is suggested by the availability, when using a natural conjugate prior, of posterior model probabilities. This allows to compare the true posterior with the simulation results based on our proposed method as well as on alternative methods. For the data analyzed in this example, our method gives a reliable estimate of the posterior distribution, while RJ performed worse, at least in terms of precision. Comparison in terms of computational time are more in favor of RJ, but we believe that a satisfactory precision is of primary importance. In Section 4 we explore a different kind of application of the theoretical results developed in Section 2: starting from an idea by Brockwell and Kadane (2002), we show an alternative algorithm for simulating a regenerative Markov chain with a prescribed limiting distribution.

2 Main results

We start by illustrating the core idea in the simple case of an unnormalized probability distribution for θ, having an absolutely continuous component and a component degenerate at one point. In what follows, for sake of notational simplicity, we assume that point to be the origin. Let η_K and δ_K denote K-dimensional Lebesgue measure and Dirac measure (concentrated on the origin of $\mathbb{R}^K$), and let f_0 be a measurable positive function defined on $\mathbb{R}^K$ and f_K be a constant. Consider the measure on $\mathbb{R}^K$

$$\mu(d\theta) := f_0(\theta)\eta_K(d\theta) + f_K\delta_K(d\theta), \tag{1}$$

that we assume to be finite, but not necessarily a probability measure. We use the notation $\bar{\mu}$ for the probability proportional to μ, i.e.

$$\bar{\mu}(\cdot) = \frac{\mu(\cdot)}{\mu(\mathbb{R}^K)}.$$

Our first goal is to determine an absolutely continuous measure τ on $\mathbb{R}^K$ and a function $\phi\colon \mathbb{R}^K \to \mathbb{R}^K$ having the property that $\tau\phi^{-1}(B) = \mu(B)$ for every Borel set B. Note that this implies that $\mu(\mathbb{R}^K) = \tau(\mathbb{R}^K)$. Moreover, if a random vector $\tilde{\zeta}$ has distribution proportional to τ, then the distribution of $\tilde{\theta} = \phi(\tilde{\zeta})$ is proportional to μ. In order to define the function ϕ, let $B_K(r) := \{\zeta \in \mathbb{R}^K : |\zeta| \leq r\}$ be the K-dimensional closed ball of radius r, centered at the origin, and consider the radial contraction

$$\psi_K(\zeta; r) := \frac{\zeta}{|\zeta|}\big(|\zeta|^K - r^K\big)^{1/K}, \qquad \zeta \in \mathbb{R}^K,\ \zeta \notin B_K(r).$$

The inverse function, defined for $\theta \neq 0$, is the radial expansion

$$\psi_K^{-1}(\theta; r) := \frac{\theta}{|\theta|}\big(|\theta|^K + r^K\big)^{1/K}.$$

It is easy to check, considering polar coordinates in $\mathbb{R}^K$, that for any r, both ψ_K and ψ_K^{-1} preserve Lebesgue measure. Thus, loosely speaking, one can use ψ_K^{-1} to move the absolutely continuous part of μ away from the origin, leaving an empty ball $B_K(r)$, and then spread the mass f_K corresponding to the degenerate component of μ into this emptied ball. More formally, define

$$g(\zeta) := \begin{cases} cf_K & \text{if } \zeta \in B_K(r), \\ f_0(\psi_K(\zeta; r)) & \text{if } \zeta \notin B_K(r), \end{cases}$$

with c and r positive constants satisfying $c\eta_K(B_K(r)) = 1$. Then, taking $\tau(d\zeta) := g(\zeta)\eta_K(d\zeta)$ and

$$\phi(\zeta) := \begin{cases} 0 & \text{if } \zeta \in B_K(r), \\ \psi_K(\zeta; r) & \text{if } \zeta \notin B_K(r), \end{cases}$$

we have exacly what we were looking for, namely an absolutely continuous

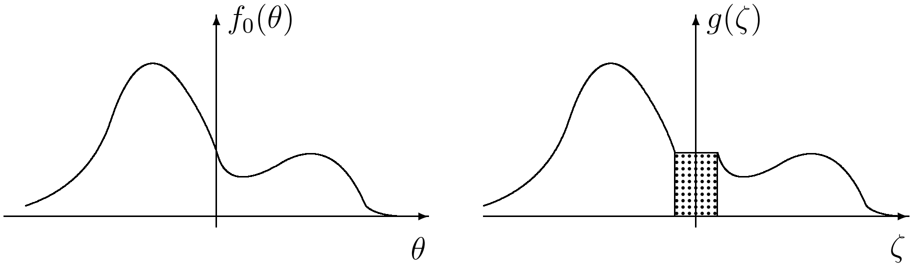

Figure 1: *Plots of f_0 and g in the one-dimensional case ($K = 1$). The area of the dotted region is equal to f_K.*

measure τ and a function ϕ such that $\tau\phi^{-1} = \mu$. A convenient choice for c and r is $c = f_0(0)/f_K$ and $r = \frac{1}{\sqrt{\pi}}\left(\frac{f_K\Gamma(\frac{K}{2}+1)}{f_0(0)}\right)^{\frac{1}{K}}$, for in this way the continuity of f_0 is inherited by g. A graphical illustration of the procedure when $K = 1$ is provided in Figure 1.

We now consider formally the more general case of a mixture of two nested components, one having full support, and the second supported by a hyperplane defined by a certain number of the last coordinates being equal to zero. We will write $\theta = (\theta_{k,1}, \theta_{k,2})$ for the generic point of $\mathbb{R}^K$, with $\theta_{k,1}$ and $\theta_{k,2}$ being, respectively, the first $h = K - k$ and last k components. Consider, for a fixed k in $\{1, \ldots, K\}$, the finite measure

$$\mu(d\theta) := f_0(\theta)\eta_K(d\theta) + f_k(\theta_{k,1})\eta_h(d\theta_{k,1})\delta_k(d\theta_{k,2}), \tag{2}$$

where f_0 and f_k are measurable positive functions defined on $\mathbb{R}^K$ and $\mathbb{R}^h$, respectively. To clarify the notation, we note that the integer indexing each density is the dimension of the degenerate part of the corresponding component. The next theorem shows how the function ψ_k can be used to transform the measure (2) into an absolutely continuous measure. The reader will realize that the idea sketched at the beginning of this section corresponds to the special case $h = 0$, $k = K$.

Theorem 1. *Let $c(\cdot)$ and $r(\cdot)$ be two measurable positive functions defined on $\mathbb{R}^h$ such that $c(\cdot)\eta_k(B_k(r(\cdot))) \equiv 1$. Define the transformation of $\mathbb{R}^K$ in itself*

$$\phi(\zeta) := \begin{cases} (\zeta_{k,1}, 0) & \textit{if } \zeta_{k,2} \in B_k(r(\zeta_{k,1})), \\ (\zeta_{k,1}, \psi_k(\zeta_{k,2}; r(\zeta_{k,1}))) & \textit{if } \zeta_{k,2} \notin B_k(r(\zeta_{k,1})). \end{cases}$$

Consider the density g defined on $\mathbb{R}^K$ by

$$g(\zeta) := \begin{cases} f_k(\zeta_{k,1})c(\zeta_{k,1}) & \textit{if } \zeta_{k,2} \in B_k(r(\zeta_{k,1})), \\ f_0(\phi(\zeta)) & \textit{if } \zeta_{k,2} \notin B_k(r(\zeta_{k,1})), \end{cases}$$

and let $\tau(d\zeta) := g(\zeta)\eta_K(d\zeta)$. Then $\tau\phi^{-1} = \mu$. Moreover, if f_0 and f_k are continuous and $c(\zeta_{k,1})f_k(\zeta_{k,1}) = f_0(\zeta_{k,1}, 0)$ for every ζ, then g is continuous.

Remark 1. When one has to deal with a target measure μ that has more than two components, supported by a family of nested hyperplanes, one can make use of Theorem 1 repeatedly. Suppose, for instance, that μ is given by

$$\begin{aligned} f_0(\theta)\eta_K(d\theta) + f_1(\theta_{1,1})\eta_{K-1}(d\theta_{1,1})\delta_1(d\theta_{1,2}) + \dots \\ \dots + f_{K-1}\eta_1(d\theta_{K-1,1})\delta_{K-1}(d\theta_{K-1,2}) + f_K\delta_K(d\theta). \end{aligned}$$

One can first apply the theorem to the first two components, obtaining an absolutely continuous measure and a transformation of $\mathbb{R}^K$ in itself, say τ_1 and ϕ_1. If g_1 is the density of τ_1, one can then apply the theorem to the measure

$$g_1(\zeta)\eta_K(d\zeta) + f_2(\zeta_{2,1})\eta_{K-2}(d\zeta_{2,1})\delta_2(d\zeta_{2,2}),$$

to construct a measure τ_2 with density g_2 and a transformation ϕ_2. Proceeding in this way, one finally obtains an absolutely continuous measure $\tau = \tau_K$ with density $g = g_K$ and a function $\phi := \phi_1 \circ \phi_2 \circ \ldots \circ \phi_K$ such that $\tau_K\phi^{-1} = \mu$. The pseudo-code of a simple algorithm that efficiently evaluates g_K and ϕ can be found in Petris and Tardella (2000). □

Remark 2. The new density g of Theorem 1 can be imagined as a sort of reshaping of the density f_0 with largest support determined by the "embedding" of another density supported on a restricted space. This reshaping effect is less pronounced whenever the opening is made around a point with high f_0 density – keeping constant the value of the other density – or when the other density is small relatively to f_0. Geometrically this is explained by the fact that the radius function $r(\cdot)$ of the opening determined by $\psi_k(\cdot, r(\cdot))$ is related to the value of the two densities according to

$$\frac{f_0(\zeta_{k,1}, 0)}{f_k(\zeta_{k,1})}\eta_k(B_k(r(\cdot))) \equiv 1.$$

This suggests that in order to obtain a more regular shape of the resulting density g, one can reparameterizing f_0 and f_k so that their maximum is attained at or near the origin. In the particular case of $k = K$, when f_K is a constant, this is particularly evident. □

Suppose now that, instead of using a transformation ϕ of a simulated Markov chain, one wishes to directly define a Markov kernel H with stationary distribution equal to the target distribution $\bar{\mu}$. A typical situation where the availability of such a transition kernel is desirable is within a hybrid MCMC sampler, where μ is proportional to one of the full-conditional distributions – see Section 3 for an example. An additional interesting motivation is provided in Section 4, where such a kernel is used to produce a regenerative Markov chain. It turns out that the auxiliary absolutely continuous measure τ and the transformation ϕ introduced above can be used to construct a transition kernel with the required property. In fact we show how the sought for H can be constructed by combining a kernel K which has $\bar{\tau}$ as its stationary distribution together with an extra kernel J, which is essentially a conditional version of $\bar{\tau}$, given ϕ. In order to give the result in a general form, let $(Z, \mathcal{S}_Z, \tau)$ and $(\Theta, \mathcal{S}_\Theta, \mu)$ be probability spaces, and let ϕ be a measurable function from Z onto Θ such that $\tau\phi^{-1} = \mu$. Let K be a transition kernel on $(Z, \mathcal{S}_Z)$ for which τ is invariant:

$$\tau(B) = \int_Z \tau(d\zeta)K(\zeta; B) \qquad \forall B \in \mathcal{S}_Z.$$

Consider $\tau^*(B \mid \phi(\zeta) = \theta)$, a regular version of τ given $\phi^{-1}\mathcal{S}_\Theta$, and define a transition kernel J from Θ to Z by setting

$$J(\theta; B) = \tau^*(B \mid \theta).$$

Theorem 2. *Consider a Markov chain $\tilde{\theta}_0, \tilde{\theta}_1, ..., \tilde{\theta}_t, ...$ on $(\Theta, \mathcal{S}_\Theta)$, whose transitions are described by the following steps:*

1. *draw $\tilde{\zeta}_{t,0}$ according to $J(\tilde{\theta}_t; \cdot)$;*

2. draw $\tilde{\zeta}_{t,1}$ according to $K(\tilde{\zeta}_{t,0}; \cdot)$;

3. set $\tilde{\theta}_{t+1} = \phi(\tilde{\zeta}_{t,1})$.

Let H denote the corresponding transition kernel, i.e.

$$H(\theta; A) = \int_Z J(\theta; d\zeta) K(\zeta; \phi^{-1} A) \qquad \forall \theta \in \Theta,\ A \in \mathcal{S}_\Theta.$$

Then μ is an invariant measure for H.

As a special case, consider the setting of Theorem 1. In that case, it is easy to show that a regular conditional probability $\tau^*(B \mid \theta)$ is the distribution degenerate at $(\theta_{k,1}, \psi_k^{-1}(\theta_{k,2}; r(\theta_{k,1})))$, when $\theta_{k,2} \neq 0$, and the product of the distribution degenerate at $\theta_{k,1}$ for the first components and the uniform distribution over $B_k(r(\theta_{k,1}))$ for the last components, when $\theta_{k,2} = 0$.

3 Application to Bayesian model selection

From a Bayesian point of view, any inferential problem involving more than one model – e.g. model selection, model averaging, variable selection in regression etc. – can be approached by assigning a prior probability to each model and, conditionally on each specific model, a prior distribution for the parameters of the model. Any inference can then be based on the posterior probabilities of the models and on the posterior distributions of the parameters of each model. Conceptually, the different models under consideration are simply subregions of a unique inclusive model. Although straightforward from a theoretical point of view, deriving the posterior from a prior and a data set may not be easy. The source of most troubles is that the prior, and therefore the posterior, does not have a density with respect to Lebesgue measure on $\mathbb{R}^K$ nor with respect to the counting measure on a countable set. For example, if one wants to consider two distinct models only, with n and m continuous parameters respectively, then the support of the prior and posterior distributions is given by the disjoint union of $\mathbb{R}^n$ and $\mathbb{R}^m$. An important special case is that of nested models. In this case there is a maximal model and all the others are obtained by setting one or more of its parameters to specific fixed values. Many research contexts are often modelled through flexible nested model: polynomial regression, autoregressive time series with unknown order, finite mixture distributions with unknown number of components, just to mention some.

In this section we present a simple numerical illustration of how the theory developed so far can be usefully applied to Bayesian model selection, when the models under consideration are nested.

We consider a linear model setting, with $n \times (p+1)$ design matrix X, where the first column of X is a column of ones. We will denote by M_k the submodel corresponding to a design matrix obtained by eliminating the last k columns of X $(k = 0, \dots, p)$. The reduced design matrix will be denoted by X_k. Note that X_k is an $n \times (p-k+1)$ matrix – the subscript k refers to the number of regressors dropped from the full model. The full model can be written as

$$y \sim \mathcal{N}(X\beta, I\sigma^2),$$

and submodel M_k is obtained by setting the last k components of β to zero. Within a Bayesian framework one needs to specify a prior probability that charges all the $p+1$ events $H_k = \{\beta_1 \neq 0, \dots, \beta_{p-k} \neq 0, \beta_{p-k+1} = \ldots = \beta_p = 0\}$, $k = 0, \dots, p$. These events correspond to a sequence of nested hyperplanes in the parameter space. We define a prior distribution in a hierarchical way as follows:

- given H_k and σ^2, $\beta_{k,1} = (\beta_0, \dots, \beta_{p-k})'$ has a $\mathcal{N}(0, \sigma^2 V_k)$ distribution;
- given H_k, σ^2 has an $\mathcal{IG}(d/2, a/2)$ distribution;
- $P(H_k) = \alpha_k$.

Thus, the global prior distribution can be written as

$$\begin{aligned} d\pi(\beta, \sigma^2) = {} & \frac{(a/2)^{d/2}}{\Gamma(d/2)} (\sigma^2)^{-(d+2)/2} \exp\bigl(-a/(2\sigma^2)\bigr) \\ & \sum_{k=0}^{p} \alpha_k (2\pi)^{-(p-k+1)/2} |V_k|^{-1/2} (\sigma^2)^{-(p-k+1)/2} \\ & \exp\bigl(-\beta'_{k,1} V_k^{-1} \beta_{k,1}/(2\sigma^2)\bigr) \\ & \eta_{p-k+1}(d\beta_{k,1}) \delta_k(d\beta_{k,2}) \eta(d\sigma^2). \end{aligned}$$

Introducing the indicator functions

$$R_k(\beta) = \begin{cases} 1 & \text{if} \quad \beta \in H_k, \\ 0 & \text{if} \quad \beta \notin H_k, \end{cases}$$

and the σ-finite measure

$$d\nu(\beta,\sigma^2) = \sum_{k=0}^{p} \eta_{p-k+1}(d\beta_{k,1})\delta_k(d\beta_{k,2})\eta(d\sigma^2),$$

one can define the prior density

$$\begin{aligned} f(\beta,\sigma^2) = \frac{d\pi}{d\nu} = \frac{1}{\Gamma(d/2)} \sum_{k=0}^{p} & R_k(\beta)\alpha_k(a/2)^{d/2}(2\pi)^{-(p-k+1)/2}|V_k|^{-1/2} \\ & (\sigma^2)^{-(d+p-k+3)/2}\exp\bigl(-Q_k/(2\sigma^2)\bigr), \end{aligned} \tag{3}$$

with

$$Q_k = \beta_{k,1}' V_k^{-1}\beta_{k,1} + a.$$

The likelihood of the parameters is the standard one from linear model theory, namely

$$L(\beta,\sigma^2) = (\sigma^2)^{-n/2}\exp\bigl(-(y-X\beta)'(y-X\beta)/(2\sigma^2)\bigr).$$

The posterior density, which can be found using Bayes theorem, is therefore equal to

$$\begin{aligned} f(\beta,\sigma^2 \mid y) &\propto L(\beta,\sigma^2)\cdot f(\beta,\sigma^2) \\ &\propto \sum_{k=0}^{p} R_k(\beta)\alpha_k(2\pi)^{-(p-k+1)/2}|V_k|^{-1/2} \\ &\qquad (\sigma^2)^{-(d+n+p-k+3)/2}\exp\bigl(-Q_k^*/(2\sigma^2)\bigr) \\ &\propto \sum_{k=0}^{p} R_k(\beta)\left[\frac{|V_k^*|^{1/2}\alpha_k}{|V_k|^{1/2}(a_k^*/2)^{d^*/2}}\right](a_k^*/2)^{d^*/2}(2\pi)^{-(p-k+1)/2}|V_k^*|^{-1/2} \\ &\qquad (\sigma^2)^{-(d^*+p-k+3)/2}\exp\bigl(-Q_k^*/(2\sigma^2)\bigr) \end{aligned} \tag{4}$$

with

$$\begin{aligned} Q_k^* &= (y - X_k\beta_{k,1})'(y - X_k\beta_{k,1}) + \beta_{k,1}'V_k^{-1}\beta_{k,1} + a \\ &= (\beta_{k,1} - m_k^*)'(V_k^*)^{-1}(\beta_{k,1} - m_k^*) + a_k^* \\ V_k^* &= (X_k'X_k + V_k^{-1})^{-1} \\ m_k^* &= V_k^* X_k' y \\ a_k^* &= a + y'y - (m_k^*)'(V_k^*)^{-1}m_k^* \\ d^* &= d + n. \end{aligned}$$

Comparing (3) and (4) one can deduce that, in this simple conjugate context, posterior model probabilities can be derived in closed form and are proportional to the expressions in square brackets in (4), i.e.

$$P(M_k \mid y) \propto \frac{|V_k^*|^{1/2}\alpha_k}{|V_k|^{1/2}(a_k^*/2)^{d^*/2}}.$$

Hence, in this setting, we have the opportunity to compare exact posterior model probabilities with MCMC estimates derived through different methods.

Let us start by describing a new computational strategy based on the results of the previous section. An approximation of the posterior distribution can be obtained by running a Gibbs sampler, sampling σ^2 and β in turn, each from its full conditional distribution. One easily obtains that the full conditional distribution of σ^2 given β when $\beta \in H_k$ (i.e. model M_k is visited) is $\mathcal{IG}((d^*+p-k+1)/2, Q_k^*/2)$. On the other hand, the full conditional density of β, with respect to $\sum_{k=0}^p \eta_{p-k+1}(d\beta_{k,1})\delta_k(d\beta_{k,2})$, is proportional to

$$\sum_{k=0}^{p} R_k(\beta)\alpha_k(2\pi\sigma^2)^{-(p-k+1)/2}|V_k|^{-1/2}\exp\bigl(-Q_k^*/(2\sigma^2)\bigr).$$

The repeated use of Theorem 1 together with Theorem 2 allows to reduce the problem of sampling from this distribution to that of sampling from an absolutely continuous distribution on $\mathbb{R}^{p+1}$. This is basically accomplished by replacing the full conditional distribution $f(\cdot|\sigma^2)$ with an absolutely continuous distribution $\tau(\cdot|\sigma^2)$, according to Remark 1 and then making use of the appropriate corresponding transformation ϕ. More precisely one proceeds as follows: according to Theorem 2 the current β is first randomly transformed through $J(\beta, \cdot)$ into, say, a current $\tilde{\zeta}$ then a draw $\tilde{\zeta}_{\text{new}}$ from a transition kernel $K(\cdot,\cdot)$ preserving τ is realized using ARMS (Adaptive Rejection Metropolis Sampling, see Gilks et al.; 1995) along a randomly selected straight line passing through $\tilde{\zeta}$ and, finally, $\tilde{\zeta}_{\text{new}}$ is back-transformed to β_{new} by applying the function ϕ, i.e. essentially undoing the sequence of hyperplane inflations needed to construct τ. A particularly appealing feature of this sampler is that it completely avoids the need to specify a proposal distribution and also to compute any Jacobians.

For comparison, we also implemented a Reversible Jump sampler, since Reversible Jump is probably the most widely used approach to perform

Model	Exact	RJ Sampler	Our Sampler	Our Sampler Accellerated
M_0	0.0099	0.014	0.0129	0.0099
M_1	0.0249	0.032	0.0244	0.0212
M_2	0.0940	0.122	0.0995	0.0929
M_3	0.3466	0.350	0.3516	0.3508
M_4	0.4943	0.461	0.4892	0.4939
M_5	0.0302	0.022	0.0223	0.0313

Table 1: *Simulation A: exact posterior model probabilities are compared to those estimated by our sampler and by an implementation of the Reversible Jump sampler. Models M_6 through M_9 have exact posterior probability less than* 10^{-4}.

MCMC computations in varying dimensional models. Here are some implementation details. At each sweep the sampler: (a) updates a randomly selected β_j within the current model, using Metropolis-Hastings algorithm; (b) updates σ^2, drawing it from its full conditional distribution and, (c) jumps – provided the proposed move is accepted – to a model having one more regressor, or one less regressor. In case a move to a larger model is proposed, say from M_k to M_{k-1}, the β_js already included in M_k are kept fixed, as well as σ^2, and a proposal for the "new" parameter β_{p-k+1} is drawn from the conditional distribution of β_{p-k+1} given $\beta_0, \ldots, \beta_{p-k}$ derived from the distribution of the Maximum Likelihood estimates of $(\beta_0, \ldots, \beta_{p-k}, \beta_{p-k+1})$ in M_{k-1}.

In order to assess the performance of our computational strategy we used two simulated data sets. In both we set $p = 9$ and we used an $n \times p$ matrix of covariates and n observations from the linear model, using $\beta = (6, 5, 4, 3, 2, 1, 0, 0, 0, 0)'$ and $\sigma^2 = 1$. The true model is therefore M_4. The only difference between the two data sets is that in simulation A we chose $n = 200$ as sample size, while in simulation B we chose $n = 100$. In the prior distribution we took m_k to be zero, V_k to be the identity matrix of the appropriate order, $a = d = 0.01$. Finally, we assumed the models to be a priori equally likely. Table 1 and Table 2 report the MCMC estimates of posterior model probabilities based on runs of 100000 iterations of our sampler; Figure 2 shows the ergodic means of model probabilities.

We decided to implement Reversible Jump as a benchmark for a compar-

Model	Exact	RJ Sampler	Our Sampler	Our Sampler Accellerated
M_0	0.039	0.087	0.0730	0.0378
M_1	0.077	0.122	0.1052	0.0960
M_2	0.181	0.220	0.2016	0.1751
M_3	0.186	0.195	0.1782	0.1761
M_4	0.420	0.322	0.3839	0.4225
M_5	0.097	0.055	0.0730	0.0924

Table 2: *Simulation B: exact posterior model probabilities are compared to those estimated by our sampler and by an implementation of the RJ sampler. Models M_6 through M_9 have exact posterior probability less than* 10^{-4}.

ative analysis. Of course every computational comparison has its pitfalls and it may give different answers according to the chosen perspective. Anyway, we believe this is a useful objective starting point to give an idea of possible competitive features of a new simulation strategy when compared with existing techniques.

If one looks at the computation time for each iteration as a criterion, our method definitely run in a longer time. Note, however, that this does not take into account that the two methods can reach convergence in a different length of time. The relative slowness was of course expected in so far as the simulation target is always supported by a subspace of maximal dimension. But the main reason accounting for this time consumption is that our sampling strategy didn't exploit the structure of the model as in fact the proposed RJ sampler did. Anyway, computational time should not be the only concern when the precisions of the methods under investigation differ significantly.

In fact, as one can grasp from Table 2, the RJ sampler – somewhat surprisingly – didn't perform satisfactorily in Simulation B; for instance the largest exact posterior probability corresponding the true model M_4 is grossly underestimated (about 25%) after 100000 iterations. We originally thought that this was due to the fact that convergence was not achieved and that the chain should run for more iterations. This was not confirmed since we performed different chains with 10^6 iterations obtainig actually the same results and no diagnostic evidence of convergence trouble according to the procedure proposed in Brooks and Giudici (1999, 2000); Castelloe and Zimmerman (2002).

If we have to sum up the comparative performance with the two simulated data sets fixing the same amount of iterations we can conclude that our method shows a faster convergence and a better precision.

In the attempt to see if our strategy could be speeded up we devised the following modification aimed to approximate the exact posterior probabilities: if one overweight artificially the full model density f_0, say $f_0'(\cdot) = W \cdot f_0(\cdot)$, so that the corresponding estimated probability is relatively high (say about 0.75) then one could exploit the conjugacy property of the single full model to make a natural proposal distribution for the modified density g corresponding to τ. Hence with a correct Metropolis-Hasting acceptance probability this produces a faster simulation scheme with the reweighted target as invariant distribution. The geometric intuition behind this idea is that the transformed density g should then be a not-so-radical reshaping of the original f_0 and this explains why the original full-conditional can represent a good proposal. The simulation scheme using this reweighted f_0' has as limiting distribution that differs from the desired one only in the artificially changed weighting system. Of course one can easily underweight the resulting estimates of all the visited submodels to get correct estimates of the original posterior probabilities. In fact, this strategy proved itself effective in giving precise estimates and saving a good deal of computational time with a very little tuning effort to guess an appropriate weighting constant W as shown in the last column of Table 1 and Table 2.

Let us stress that the most important distinguishing feature of our basic strategy is the absence of researcher expertise to get the method work. This can well pay-off some extra computational time at least when this turns out to be available. Also, with this automatic procedure requiring basically only the density functions and, possibly, locations with high density, a considerable amount of programming and debugging time can be saved.

4 Application to regenerative Markov chain

Suppose that an MCMC scheme has been implemented for approximating a target distribution

$$\pi(A) \propto \int_A f_0(\theta)\eta_K(d\theta)$$

through the realization of an ergodic Markov chain with π as stationary distribution. Let us denote with Q_w the corresponding Markov kernel. Here we exploit the result of Brockwell and Kadane (2002, Theorem 2.1) where it is shown how to construct a regenerative Markov chain with the same stationary distribution π simply relying on a Harris recurrent Markov chain with limiting distribution

$$\pi^*(A) = (1-\lambda)\pi(A) + \lambda I_{\{0\}}(A). \tag{5}$$

The regenerative chain is then exploited to give an estimation of the standard errors of the MCMC estimates of features of π. We refer to their paper for implementation details. Here we show how the results of Theorem 1 and Theorem 2 can be combined to obtain a new final algorithm, different from that of Brockwell and Kadane, for constructing the basic ingredient, i.e. a Markov chain with (5) as limiting distribution.

One immediately realizes that the mixture π^* in (5) is actually of the form (1), i.e. proportional to a finite measure that can be written as

$$\mu(d\theta) = f_0(\theta)\eta_K(d\theta) + f_K\delta_K(d\theta).$$

Of course the constant f_K and the mixing weight λ are functionally related by

$$\lambda = \frac{f_K}{\int_{\mathbb{R}_K} f_0(\theta)\eta_K(d\theta) + f_K}$$

so that, fixing either one, the other is automatically determined.

Hence Theorem 1 can be used to define the appropriatey measure τ and the function ϕ so that $\mu = \tau\phi^{-1}$. Let us denote with $K(\cdot,\cdot)$ an appropriate Markov kernel so that τ is invariant for $K(\cdot,\cdot)$. Keeping this notation we can use Theorem 2 to get a Markov kernel $H(\cdot,\cdot)$ so that μ is the invariant distribution.

The only thing to discuss at this point is about guidelines for a working procedure. Notice that we have degrees of freedom in choosing λ or, equivalently, f_K. Also, no mention has been made so far about how to construct an appropriate $K(\cdot,\cdot)$ to approximate τ.

In our limited experience the following suggestion is likely to be effective. From a pilot run of the original working kernel Q_w used for approximating π, one can get an idea where the "center" x^* of π is located and how large is the total mass $\int_{\mathbb{R}_K} f_0(\theta)\eta_K(d\theta)$ so that one can set an appropriate mass f_K in order to have a corresponding small weight λ, let us say approximately equal to 10^{-3}. Also, without loss of generality let us admit that f_0 is such that

x^* is located at the origin, otherwise reparameterize accordingly through a translation. In order to easily get a Markov kernel to simulate from a distribution proportional to τ one can rely on a Metropolis-Hasting scheme using Q_w as a basis to construct an appropriate proposal. In fact having fixed f_K so that λ is approximately equal to 10^{-3} then one can expect that the measure τ is not very different from the original f_0. Geometrically only a relatively "small" opening around the origin (a point corresponding to a high density f_0) has to be made to accomodate for the mass f_K. Let us denote with $K(\cdot, \cdot)$ the Markov kernel just derived through this Metropolized scheme.

	mc-estimate	rmc-s.e.	low	upp
$\log \alpha$	0.979318	0.000440	0.978438	0.980198
$\log \beta$	-0.022342	0.000010	-0.022362	-0.022321
logit γ	1.895405	0.000854	1.893696	1.897113
σ^2	0.008617	0.000004	0.008609	0.008625

Table 3: *Petris & Tardella Method: from about* 10^7 *original iterations run in about 7,5 hours 3193 tours were obtained with average tour length of about 3128; the estimated coefficient of variation of tour length divided by total length was 0.000331*

Now we describe in detail how we implemented this idea to the Dugong dataset used in Brockwell and Kadane (2002), originally taken from Ratkowsky (1983). The data consist in measurements of length (Y) and age (X) of 27 dugongs, and the following regression model is considered:

$$\begin{aligned} Y_i &\sim \mathcal{N}(\mu_i, \sigma^2) \\ \mu_i &= \alpha - \beta\gamma^{X_i}, \end{aligned}$$

with unknown parameters $\alpha > 0$, $\beta > 0$, $\gamma \in (0, 1)$ and $\sigma^2 > 0$. Prior specification for a Bayesian analysis has been chosen as

$$\begin{aligned} \alpha &\sim \mathcal{N}(0, 10000) \\ \beta &\sim \mathcal{N}(0, 10000) \\ \gamma &\sim \mathcal{U}(0, 1) \\ \sigma^2 &\sim \mathcal{IG}(0.001, 0.001) \end{aligned}$$

For approximating posterior quantities releative to

$$\pi(\alpha, \beta, \gamma, \sigma^2|\text{data})$$

Brockwell and Kadane implemented a simple MCMC strategy consisting of an hybrid Gibbs-Metropolis with sequential draws from (respectively) the full-conditionals of α, β and σ^2, while for the γ component a simple Metropolis step with independent uniform proposal is used instead of its (unavailable) full-conditional. This corresponds in our previous notation to the working kernel Q_w used to simulate a Markov chain with $\pi(\alpha, \beta, \gamma, \sigma^2|\text{data})$ as limiting distribution. Similarly, to get the Markov kernel $K(\cdot, \cdot)$ mentioned in Theorem 2 we use the same full conditionals of f_0 (here corresponding to π) for α, β and σ^2 and an independent uniform, for γ, as Metropolis proposals to simulate from the corresponding full-conditionals of the auxiliary density g corresponding to τ and accept the proposed draws with the probability that ensures that τ is stationary for $K(\cdot, \cdot)$.

	mc-estimate	rmc-s.e.	low	upp
$\log \alpha$	0.978697	0.000440	0.977816	0.979578
$\log \beta$	-0.022591	0.000010	-0.022611	-0.022570
logit γ	1.888902	0.000859	1.887184	1.890620
σ^2	0.008615	0.000004	0.008608	0.008623

Table 4: *Brockwell & Kadane Method: from* 10^7 *original iterations, run in about 10 hours, 3422 tours were obtained with average tour length of about 2921; the estimated coefficient of variation of tour length divided by total length was 0.000398*

In order to compare our strategy with that of Brockwell and Kadane – the results are displayed in Table 4 and Table 4 – we calibrated the tuning weight λ of their method to get approximately the same expected tour length. Both procedures were written in R (Ihaka and Gentleman; 1996). The parameter estimates and estimated standard errors are very close for the two methods, as was to be expected, and computing time is reduced by about 25% using our approach. From a general point of view, an appealing feature of our strategy is that it relies on a geometric understanding of the modified chain and avoids the use of a multinormal distribution to guess the shape of f_0 as suggested in Brockwell and Kadane.

5 Concluding remarks

In this paper we have introduced some theoretical results which form the basis for a new strategy for simulating from mixture distributions with components being supported on subspaces of different dimension. The potential of these results has been illustrated in two relevant situations: (1) Bayesian inference for nested linear models; (2) contruction of regenerative Markov chains for evaluating the standard error of MCMC estimates. In the first example the attention is focused on comparing the behaviour of the proposed techniques with an alternative RJ strategy in a controlled context where the posterior analysis has an exact analytical answer. The numerical results stress once again the importance of enlarging the pool of available computational techniques for coping with inference on varying dimensional parametric spaces.

Rather than illustrating the effectiveness of our proposal in other more sophisticated nested models, we preferred to justify the usefulness of the theoretical results in the different context of building regenerative Markov chains. In the same spirit of the approach of Brockwell and Kadane (2002), we conceived indirectly a Markov chain on $\mathbb{R}^K$ with an artificial atom in a different way from the well known splitting technique of Nummelin (1978), just by adjusting its invariant distribution and restating it in equivalent form. In fact we have started from a mixture distribution on $\mathbb{R}^K$, to be interpreted as the invariant measure of the ergodic chain, and through techniques standard in MCMC we derived a Markov chain where the artificial atom is phisically visualized within the original space. The main difference with the original splitting technique is that no minorization condition is used in the construction, which in fact is also true for Brockwell and Kadane's construction.

It has been shown elsewhere (Petris and Tardella; 2000) that other less trivial nested models, such as autoregressive time series, can be addressed with the technique presented here; other interesting applications to normal and binomial mixture models will be explored in a forthcoming paper.

As far as comparison with RJ or other existing alternatives, we didn't aim at proving the absolute superiority of our method in terms of computational time and efficiency. We showed that our method lends itself more easily to an automatic implementation which may be slower in terms of computational time but sometimes can reveal itself safer than a RJ implementation which is not carefully designed and monitored. We also showed that we can improve on computational time just working on a faster MCMC strategy for simulating from a distribution which is absolutely continuous with respect to the Lebesgue measure, and for which a geometric intuition of its shape can be derived.

Appendix

Proof of Theorem 1. Writing ψ_k in polar coordinates, it is easy to check that, for any fixed r, it preserves Lebesgue measure. As a consequence, the Jacobian of the transformation ϕ is one. Consider two Borel sets A_1 and A_2 in $\mathbb{R}^h$ and $\mathbb{R}^k$ respectively. If $0 \notin A_2$, then

$$\begin{aligned}\tau\phi^{-1}(A_1 \times A_2) &= \tau(\{\zeta : \phi(\zeta) \in A_1 \times A_2\}) \\ &= \int_{\{\zeta:\phi(\zeta)\in A_1\times A_2\}} g(\zeta)\eta_K(d\zeta) \\ &= \int_{\{\zeta:\phi(\zeta)\in A_1\times A_2\}} f_0(\phi(\zeta))\,\eta_K(d\zeta) \\ &= \int_{A_1\times A_2} f_0(\theta)\,\eta_K(d\theta) \\ &= \mu(A_1 \times A_2).\end{aligned}$$

On the other hand, if $A_2 = \{0\}$, then

$$\begin{aligned}\tau\phi^{-1}(A_1 \times A_2) &= \tau(\{\zeta : \phi(\zeta) \in A_1 \times A_2\}) \\ &= \int_{\{\zeta:\phi(\zeta)\in A_1\times A_2\}} g(\zeta)\eta_K(d\zeta) \\ &= \int_{\{\zeta:\zeta_{k,1}\in A_1,\zeta_{k,2}\in B_k(r(\zeta_{k,1}))\}} f_k(\zeta_{k,1})c(\zeta_{k,1})\,\eta_K(d\zeta) \\ &= \int_{A_1} f_k(\zeta_{k,1})c(\zeta_{k,1})\Big(\int_{B_k(r(\zeta_{k,1}))} \eta_k(d\zeta_{k,2})\Big)\,\eta_h(d\zeta_{k,1}) \\ &= \int_{A_1} f_k(\zeta_{k,1})c(\zeta_{k,1})\eta_k(B_k(r(\zeta_{k,1})))\,\eta_h(d\theta_{k,1}) \\ &= \int_{A_1} f_k(\theta_{k,1})\,\eta_h(d\theta_{k,1}) \\ &= \mu(A_1 \times A_2).\end{aligned}$$

The two equalities imply that for any Borel set A in $\mathbb{R}^K$, $\tau\phi^{-1}(A) = \mu(A)$. Continuity of g, under the specified additional assumptions, is straightforward. □

Proof of Theorem 2. For every $B \in \mathcal{S}_Z$ one has

$$\int_\Theta \mu(d\theta)J(\theta; B) = \int_Z \tau(d\zeta)J(\phi(\zeta); B) = \tau(B).$$

Therefore,

$$\begin{aligned}
&\int_\Theta \mu(d\theta)H(\theta; A) \\
&\quad = \int_\Theta \mu(d\theta) \int_Z J(\theta, d\zeta)K(\zeta, \phi^{-1}A) \\
&\quad = \int_Z \tau(d\zeta)K(\zeta; \phi^{-1}A) \\
&\quad = \tau(\phi^{-1}A) = \mu(A).
\end{aligned}$$

□

References

Al-Awhadhi, F., Hurn, M. A. and Jennison, C. (2001). Improving the acceptance rate of reversible jump MCMC proposals, *Technical report*, University of Bath.

Brockwell, A. E. and Kadane, J. B. (2002). Identification of regeneration times in MCMC simulation, with application to adaptive schemes, *Technical report*, Department of Statistics, Carnegie Mellon University.

Brooks, S. P. and Giudici, P. (1999). Convergence assessment for reversible jump MCMC simulations, *BayesStat6*, pp. 733–742.

Brooks, S. P. and Giudici, P. (2000). Markov chain Monte Carlo convergence assessment via two-way analysis of variance, *Journal of Computational and Graphical Statistics* **9**(2): 266–285.

Brooks, S. P., Giudici, P. and Roberts, G. O. (2003). Efficient construction of reversible jump MCMC proposal distributions (with discussion), *Journal of the Royal Statistical Society B.* To appear .

Cappé, O., Robert, C. and Rydén, T. (2001). Reversible jump MCMC converging to birth-and-death MCMC and more general continuous time samplers, *Technical report*, CREST, INSEE, Paris.

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

Castelloe, J. M. and Zimmerman, D. M. (2002). Convergence assessment for reversible jump MCMC samplers, *Technical report*, Department of Statistics and Actuarial Science, University of Iowa.

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

Gilks, W. R., Best, N. G. and Tan, K. K. C. (1995). Adaptive rejection Metropolis sampling within Gibbs sampling (Corr: 97V46 p541-542 with R. M. Neal), *Applied Statistics* **44**: 455–472.

Godsill, S. and Troughton, P. (1998). A reversible jump sampler for autoregressive time series, *ICASSP '98*, Vol. 4 of *1998 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 2257–2260.

Green, P. (2003). Trans-dimensional MCMC, *in* P. Green, N. Hjort and S. Richardson (eds), *Highly Structured Stochastic Systems*, Oxford University Press.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika* **82**(4): 711–732.

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics, *Journal of Computational and Graphical Statistics* **5**(3): 299–314.

Nummelin, E. (1978). A splitting technique for Harris recurrent Markov chains, *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **43**(4): 309–318.

Petris, G. and Tardella, L. (2000). A new strategy for simulating from mixture distributions with applications to bayesian model selection, *Technical report*, Department of Statistics, University of Rome "La Sapienza".

Ratkowsky, D. A. (1983). *Nonlinear regression modeling: A unified practical approach*, Marcel Dekker Inc.

Rotondi, R. (2002). On the influence of the proposal distributions on a reversible jump MCMC algorithm applied to the detection of multiple change-points, *Computational Statistics and Data Analysis* **40**(3): 633–653.

Stephens, M. (2000). Bayesian analysis of mixture models with an unknown number of components—an alternative to reversible jump methods, *Ann. Statist.* **28**(1): 40–74.

Giovanni Petris
Department of Mathematical Sciences
University of Arkansas
Fayetteville, AR 72701
USA
E-mail: GPetris@uark.edu

Luca Tardella
Dipartimento di Statistica, Probabilità e Statistiche Applicate
Università di Roma "La Sapienza"
Piazzale Aldo Moro, 5
00185 Roma
Italy
E-mail: luca.tardella@uniroma1.it

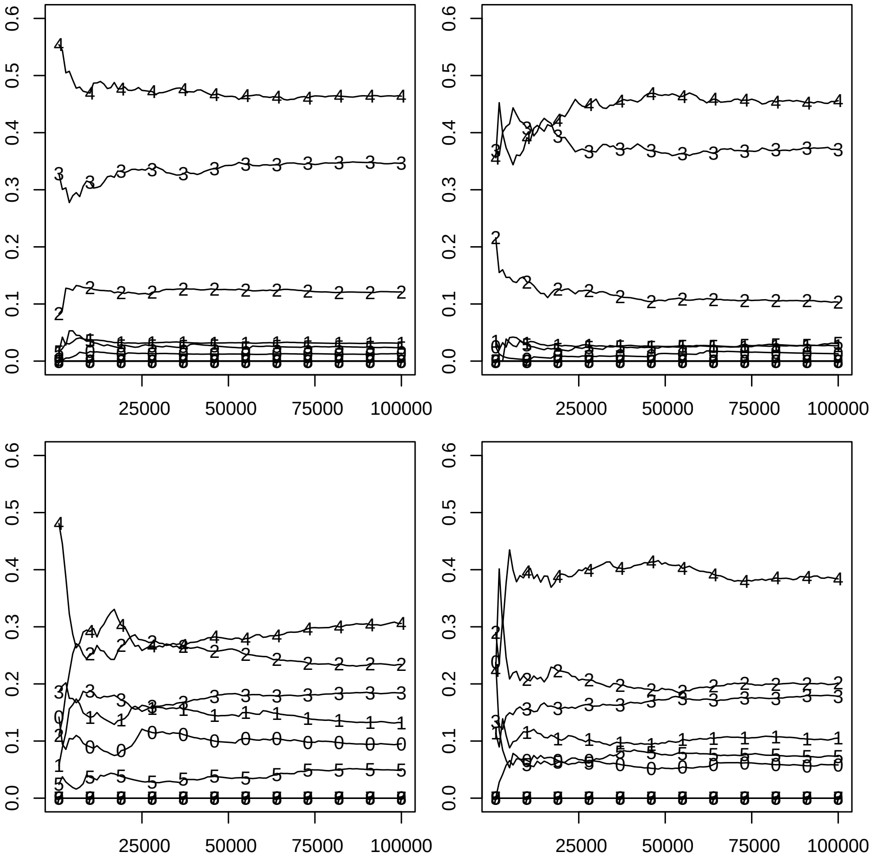

Figure 2: *Ergodic means of model probabilities. Top row: simulation A; bottom row: simulation B. Left column: RJ; right column: our sampler.*