Model Choice using Reversible Jump Markov Chain Monte Carlo,

David I. Hastie[*]
Imperial College London.

Peter J. Green[†]
University of Bristol.

May 11, 2011

Abstract

We review the across-model simulation approach to computation for Bayesian model determination, based on the reversible jump Markov chain Monte Carlo method. Advantages, difficulties and variations of the methods are discussed. We also discuss some limitations of the ideal Bayesian view of the model determination problem, for which no computational methods can provide a cure.

Some key words: Across-model sampling, Bayes factors, Bayesian model determination, posterior model probabilities, transdimensional inference, variable dimension problems,

1 Introduction

Problems where 'the number of things you don't know is one of the things you don't know' seem to be ubiquitous in statistical modelling, both in traditional modelling situations, such as variable selection in regression, and in more novel methodologies, such as object recognition, signal processing, and Bayesian nonparametrics. A feature of all such problems is that they can be addressed using the basic formulation of model determination or choice.

This article serves as an introduction to the Bayesian approach for model determination problems, emphasising the computation of posterior model probabilities, specifically using reversible jump Markov chain Monte Carlo (MCMC) methods.

1.1 Ideal Bayes model determination

What we choose to call the 'ideal Bayes' approach to model determination treats uncertainty about a statistical model and uncertainty about its parameters in a unified way: all unknowns are modelled by random variables, and inference is based on conditional (posterior) distributions induced by an assumed joint probability model for unknowns and observed data.

Given a countable set of models, the Bayesian model choice problem generically involves joint inference about a model indicator k and a parameter vector θ_k, where the model indicator determines the dimension n_k of the parameter and this dimension may vary from model to model. In a frequentist setting, inference about the two kinds of unknown, k and θ_k, is almost invariably based

[*] Department of Epidemiology and Public Health, Imperial College School of Medicine, Norfolk Place, London W2 1PG, UK.
Email: `d.hastie@imperial.ac.uk`

[†] School of Mathematics, University of Bristol, Bristol BS8 1TW, UK.
Email: `P.J.Green@bristol.ac.uk`.

on different logical principles. In contrast, the ideal Bayesian treats (k, θ_k) as a joint unknown and to make inference only the joint posterior $p(k, \theta_k|Y)$ is needed.

Joint inference for the generic Bayesian model choice problem can be set naturally in the form of a simple Bayesian hierarchical model. We suppose that a prior $p(k)$ is specified over models k in a countable set $\mathcal{K}$, and for each k we are given a prior distribution $p(\theta_k|k)$, along with a likelihood $p(Y|k, \theta_k)$ for the data Y. In some settings, $p(k)$ and $p(\theta_k|k)$ are not separately available, even up to multiplicative constants; this applies for example in many point process models. However it will be clear that what follows requires specification only of the product $p(k, \theta_k) = p(k) \times p(\theta_k|k)$ of these factors, up to a multiplicative constant.

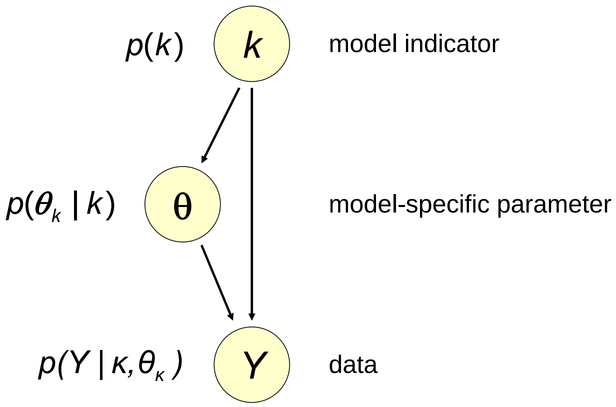

Figure 1: The basic hierarchical model for Bayesian model choice, as a directed acyclic graph.

Throughout this article all probability distributions are proper. Furthermore, for simplicity of exposition (rather than as a necessity of the method), we suppose that $p(\theta_k|k)$ is a probability density with respect to the n_k-dimensional Lesbegue measure. Where there are parameters common to all models, perhaps in additional layers of hierarchy, these are subsumed into each $\theta_k \in \mathcal{X}_k \subset \mathcal{R}^{n_k}$. In many models there are discrete unknowns as well as continuously distributed ones. Such unknowns, whether fixed or variable in number, cause no additional difficulties; only discrete-state Markov chain notions are needed to handle them, and formally speaking, the variable k can be augmented to include these variables; such problems then fit into the above framework.

The joint posterior

$$p(k, \theta_k|Y) = \frac{p(k, \theta_k)p(Y|k, \theta_k)}{\sum_{k' \in \mathcal{K}} \int p(k', \theta'_{k'})p(Y|k', \theta'_{k'})\mathrm{d}\theta'_{k'}}$$

can always be factorised as

$$p(k, \theta_k|Y) = p(k|Y)p(\theta_k|k, Y),$$

that is, as the product of posterior model probabilities and model-specific parameter posteriors. This identity is very often the basis for reporting the inference and in particular, for model determination, the (marginal) posterior model probabilities $p(k|Y)$ are commonly of interest.

Before proceeding, it is important to appreciate the generality of the above model determination formulation. In particular, note that it embraces not only genuine model choice situations, where the variable k indexes the collection of discrete models under consideration, but also settings where there is really a single model, but one with a variable-dimension parameter, for example a functional representation such as a series whose number of terms is not fixed. In the latter case, arising sometimes in Bayesian nonparametrics, for example, k is unlikely to be of direct inferential interest.

These problems really form a continuous spectrum rather than a sharp dichotomy: a point made well by considering curve-fitting to noisy data using families of smoothing splines or polynomials of different complexity. We might naturally think that choosing the degree of a polynomial is a choice of model while choosing the tuning constant multiplying a roughness functional is estimating a hyperparameter; but in reality it is hard to argue that these are really distinct kinds of problem.

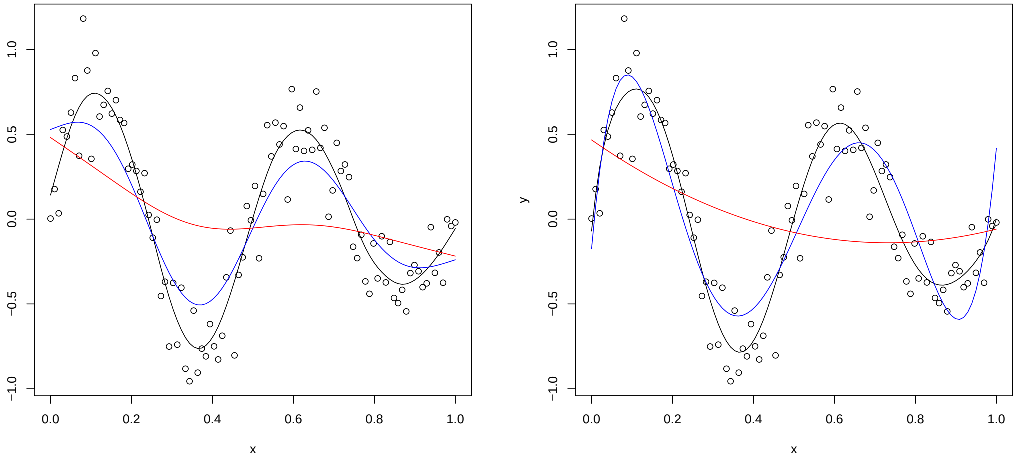

Figure 2: Three curves fitted to the same artificial data, using (left panel) cubic smoothing splines and (right panel) polynomials. In each case the three curves use 3, 6 and 11 degrees of freedom respectively.

1.2 Simulation-based computation for model determination

In practice, the posterior probabilities required to conduct Bayesian model determination are almost invariably not available analytically. Thus to make inference we must rely on numerical methods, with the consequence that the resulting computations can be quite intensive. The only generic methods available currently that produce results that are 'exact' (that is, up to simulation error only) are Monte Carlo methods of various kinds.

A favoured Monte Carlo method is Markov chain Monte Carlo (MCMC). The basis for this approach is to construct a Markov chain which has as its invariant distribution the (posterior) distribution of interest. There are two main approaches using MCMC for model determination problems: *across*-model simulation, in which there is a single MCMC simulation with states of the form $(k, \theta_k) \sim p(k, \theta_k|Y)$; and *within*-model simulation, in which there are separate simulations of $\theta_k \sim p(\theta_k|k, Y)$ for each k. The former is the primary subject of this article. We omit further details of the latter as they are covered in brief in section 4 of Green (2003) and in more detail by Friel (this volume).

Reversible jump Markov chain Monte Carlo (Green 1995) is a method for *across*-model simulation of posterior distributions of the form introduced in the previous section. More generally, reversible jump is a technique for simulating from a Markov chain whose state is a vector whose dimension is not fixed. The majority of the remainder of this article is focussed on how to implement this method for the model determination problem and the associated challenges that arise.

We note that other simulation methods that are not MCMC include *particle filters*, using sequential Monte Carlo in place of MCMC (Jasra *et al.* 2008) and *ABC* – 'likelihood-free' methods where $Y|k, \theta_k$ can be simulated but $p(Y|k, \theta_k)$ cannot be evaluated (Didelot *et al.* 2011). These methods are outside the scope of this article.

1.3 Structure of the paper

For the remainder of this article some understanding of standard MCMC methods is assumed, and the unfamiliar reader is referred to Gamerman (1997) or Brooks (1998) for an introductory tutorial.

In section 2, reversible jump MCMC is presented and discussed, and an illustrative example is given in section 3, along with a brief look at past literature citing the method. Section 4 discusses some methodological extensions aimed particularly at construction of efficient proposals. We then consider the idea of a fully-automated reversible jump sampler in Section 5. In Section 6 we present some recent methodologies exploiting reversible jump and briefly review other model-jumping approaches. Leaving reversible jump MCMC behind, in Section 7 we return to the more general question of Bayesian model determination and consider some of the outstanding philiosophical difficulties, before offering some brief words of conclusion.

This paper updates Green (2003) but omits, in particular, coverage of within-model sampling approaches to trans-dimensional sampling problems, while including extra material on more recent developments, and on statistical methodologies built on reversible jump and related across-model sampling methods.

2 Metropolis–Hastings in a more general light

In the direct approach to computation of the joint posterior $p(k, \theta_k|Y)$ via MCMC we construct a single Markov chain simulation, with states of the form $(k, \theta_k) = (k, \theta_{k,1}, \theta_{k,2}, \ldots, \theta_{k,n_k})$; we might call this an *across-model* simulation. The state space for such an across-model simulation is $\mathcal{X} = \bigcup_{k \in \mathcal{K}}(\{k\} \times \mathcal{X}_k)$, where for each k, $\mathcal{X}_k \subset \mathcal{R}^{n_k}$. The point of defining $\mathcal{X}$ in this way is that even in cases where the dimensions $\{n_k\}$ are all different, we often wish to have direct inferential access to the 'model indicator' k; in cases where the $\{n_k\}$ are not all different, this becomes essential. Mathematically, $\mathcal{X}$ is not a particularly awkward object, and our construction involves no especially challenging novelties. However, such a state space is at least a little non-standard!

Formally, our task is to construct a Markov chain on this general state space with a specified limiting distribution. Reversible jump MCMC is one means of achieving this goal, using the Metropolis–Hastings paradigm to build a suitable reversible chain, as is usual in Bayesian MCMC for complex models.

We begin by presenting an introduction to reversible jump, considering how transitions between different states in $\mathcal{X}$ might practically be achieved by a computer program. We build first upon a fixed-dimensional case and then demonstrate how this is immediately extended to the trans-dimensional case. The aim of this perspective is to show that the generalisation of the Metropolis–Hastings algorithm is straightforward and dispel the myth that reversible jump is technically challenging.

2.1 A constructive representation in terms of random numbers

Our aim is to construct a Markov chain on a general state space $\mathcal{X}$ with invariant distribution p. (Note that neither $\mathcal{X}$ nor p need refer to the Bayesian model-choice problem formulated in the previous section). As is common in MCMC we will consider only reversible chains, so that, put simply, we require the equilibrium probability that the state of the chain is in a general set A and moves to a general set B to be the same with A and B reversed. This is known as the *detailed balance condition.*

Suppose initially that we have a simpler state space, $\mathcal{X} \subset \mathcal{R}^n$. As usual with the Metropolis–Hastings algorithm, we can satisfy the detailed balance condition by applying a protocol that proposes a new state for the chain and then accepts this proposed state with an appropriately derived probability. This probability is obtained by considering a transition and its reverse simultaneously. Let the density of the invariant distribution p also be denoted by p. At the current state x, we generate, say, r random numbers u from a known joint density g. The proposed new state of the chain x' is then constructed by some suitable deterministic function h such that $(x', u') = h(x, u)$. Here, u' are the r-dimensional random numbers, generated from a known joint density g' that would be required for the reverse move from x' to x, using the inverse function h' of h. If the move from x to x' is accepted with probability $\alpha(x, x')$ and likewise, the reverse move is accepted with probability $\alpha(x', x)$, the detailed balance requirement can be written as

$$\int_{(x,x')\in A\times B} p(x)g(u)\alpha(x,x')\mathrm{d}x\,du = \int_{(x,x')\in A\times B} p(x')g'(u')\alpha(x',x)\mathrm{d}x'\,du'. \tag{1}$$

If the transformation h from (x, u) to (x', u') and its inverse h' are differentiable, then we can apply the standard change-of-variable formula to the right hand side of equation (1). We then see that the $(n + r)$-dimensional integral equality (1) holds if

$$p(x)g(u)\alpha(x,x') = p(x')g'(u')\alpha(x',x)\left|\frac{\partial(x',u')}{\partial(x,u)}\right|,$$

where the last factor is the Jacobian of the transformation from (x, u) to (x', u'). Thus, a valid choice for α is

$$\alpha(x,x') = \min\left\{1, \frac{p(x')g'(u')}{p(x)g(u)}\left|\frac{\partial(x',u')}{\partial(x,u)}\right|\right\}, \tag{2}$$

involving only ordinary joint densities.

While this formalism is perhaps a little indirect for the fixed-dimensional case, it proves a flexible framework for constructing quite complex moves using only elementary calculus. In particular, the possibility that $r < n$ covers the case, typical in practice, that given $x \in \mathcal{X}$, only a lower-dimensional subset of $\mathcal{X}$ is reachable in one step. (The Gibbs sampler is the best-known example of this, since in that case only some of the components of the state vector are changed at a time, although the formulation here is more general as it allows the subset not to be parallel to the coordinate axes.) Separating the generation of the random innovation u and the calculation of the proposal value through the deterministic function h is deliberate; it allows the proposal distribution

$$q(x,B) = \int_{\{u:h(x,u)\in B\times\mathcal{R}^r\}} g(u)\mathrm{d}u$$

to be expressed in many different ways, for the convenience of the user.

2.2 The trans-dimensional case

The main benefit of this formalism is that expression (2) applies, without change, in a variable-dimension context. Consider now allowing $\mathcal{X}$ to be a more complex space, such that x has different dimension in different parts of $\mathcal{X}$. (We use the same symbol $p(x)$ for the target density whatever the dimension of x.) Provided that the transformation from (x, u) to (x', u') remains a diffeomorphism, the individual dimensions of x and x' can be different. The dimension-jumping has become essentially ‘invisible’.

In this setting, suppose the dimensions of x, x', u and u' are n, n', r and r' respectively, then we have functions $h : \mathcal{R}^n \times \mathcal{R}^r \to \mathcal{R}^{n'} \times \mathcal{R}^{r'}$ and

$h' : \mathcal{R}^{n'} \times \mathcal{R}^{r'} \to \mathcal{R}^n \times \mathcal{R}^r$, used respectively in $(x', u') = h(x, u)$ and $(x, u) = h'(x', u')$. For the transformation from (x, u) to (x', u') to be a diffeomorphism requires that $n + r = n' + r'$, so-called 'dimension-matching'; if this equality failed, the mapping and its inverse could not both be differentiable. We note, however, that one or both of r, r' might be 0.

2.3 Multiple move types and the model-choice problem

Returning to our generic model-choice problem, we wish to use these reversible jump moves to sample the space $\mathcal{X} = \bigcup_{k \in \mathcal{K}} (\{k\} \times \mathcal{X}_k)$ with invariant distribution p, which here is $p(k, \theta_k | Y)$.

Just as in ordinary MCMC, although each move is a transition kernel reversible with respect to p, we typically need multiple types of moves to traverse the whole space $\mathcal{X}$. Again, as in ordinary MCMC, we can scan through the available moves according to various deterministic or random schedules. Here we consider the case of move types chosen independently for each sweep of the MCMC run, and extend conventional Metropolis–Hastings by allowing the probabilities of each move type to depend on the current state.

Indexing the move types by m in a countable set $\mathcal{M}$, a particular move type m consists of both the forwards move from $x = (k, \theta_k)$ to $x' = (k', \theta'_{k'})$ and the reverse, taking x' to x, for a specific pair (k, k'). For the forwards move, r_m random numbers u are generated from known joint distribution g_m, and the new state $\theta'_{k'} \in \mathcal{R}^{n_{k'}}$ is constructed as $(\theta'_{k'}, u') = h_m(\theta_k, u)$. Here u' are the r'_m random numbers from joint distribution g'_m needed for the reverse move, to move from $\theta'_{k'}$ to θ_k, using the inverse function h'_m of h_m.

Letting $j_m(x)$ denote the probability that move m is attempted at state x, the move-type specific equivalent to equation (1) is

$$\begin{aligned} &\int_{(x,x') \in A \times B} p(x) j_m(x) g_m(u) \alpha_m(x, x') \mathrm{d}x\, du \\ &\quad = \int_{(x,x') \in A \times B} p(x') j_m(x') g'_m(u') \alpha_m(x', x) \mathrm{d}x'\, du'. \end{aligned} \tag{3}$$

Since the complete transition kernel is obtained by summing over $m \in M$, ensuring that the detailed balance equation (3) holds for for each move type m is sufficient to ensure that the detailed balance condition holds. Thus, a sufficient choice for the acceptance probability α_m associated with move type m is given by $\alpha_m(x, x') = \min\{1, A_m(x, x')\}$ where

$$A_m(x, x') = \frac{p(x')}{p(x)} \frac{j_m(x')}{j_m(x)} \frac{g'_m(u')}{g_m(u)} \left| \frac{\partial(\theta'_{k'}, u')}{\partial(\theta_k, u)} \right|. \tag{4}$$

Here the Jacobian factor is from the transformation from (θ_k, u) to $(\theta'_{k'}, u')$, and is obviously dependent upon the move type m. In order for this transformation to be a diffeomorphism we again require the dimension matching to hold, so that $n_k + r_m = n_{k'} + r'_m$.

Finally we note, when at $x = (k, \theta_k)$, only a limited number of moves m will typically be available, namely those for which $j_m(x) > 0$. With probability $1 - \sum_{m \in M} j_m(x)$ no move is attempted.

2.4 Some remarks and ramifications

To summarise, 'reversible jump' MCMC is just Metropolis–Hastings, formulated to allow for sampling from a distribution on a union of spaces of differing dimension, and permitting state-dependent choice of move type. In understanding the framework, it may be helpful to stress the key role played by the joint state–proposal equilibrium distributions. In fact, detailed balance is explicitly characterised as the invariance of these distributions to time-reversal. The fact that the degrees of freedom in these joint distributions are unchanged when x and x' are interchanged allows the possibility of reversible jumps across dimensions, and these distributions directly determine the move acceptance probabilities. Contrary to some accounts that connect it with the jump in dimension, the Jacobian comes into the acceptance probability simply through the fact that the proposal destination x' is specified indirectly through $h(x, u)$.

Note that the framework gives insights into Metropolis–Hastings that apply quite generally. State-dependent mixing over a family of transition kernels in general infringes detailed balance, but is permissible if, as here, the move probabilities $j_m(x)$ enter properly into the acceptance probability calculation. Note also the contrast between this *randomised* proposal mechanism, and the related idea of *mixture* proposals, where the acceptance probability does not depend on the move actually chosen; see the discussion in Appendix 1 of Besag *et al.* (1995).

To properly ascertain the theoretical validity of the general state space Metropolis–Hastings algorithm, we require a measure-theoretic approach, defining dominating measures and Radon–Nikodym derivatives. In practice however, as is demonstrated by the presentation above, the measure theory becomes essentially invisible and can be safely ignored. To avoid getting distracted by details that are peripheral to this article, we exclude further discussion of these aspects. For a detailed consideration, the interested reader is referred to the original presentation in Green (1995) or the alternative, and we trust improved, discussion in Green (2003).

For most purposes, theoretical study of such Markov chains simply replicates the corresponding study of ordinary Metropolis–Hastings. There are exceptions: for example, verification of Harris recurrence for a chain seems to demand analysis of the structure of the space $\mathcal{X}$ and details of the transitions. See Roberts and Rosenthal (2006).

Finally, note that in a large class of problems involving nested models, the only dimension change necessary is the addition or deletion of a component of the parameter vector (think of polynomial regression, or autoregression of variable order). In such cases, omission of a component is often equivalent to setting a parameter to zero. These problems can be handled in a seemingly more elementary way, through allowing proposal distributions with an atom at zero: the usual Metropolis–Hastings formula for the acceptance probability holds for densities with respect to arbitrary dominating measures, so the reversible jump formalism is not explicitly needed. Nevertheless, it leads to exactly the same algorithm.

2.5 Alternative presentations and related methods

Reversible jump has been presented in several different ways by other authors. Of note is the tutorial by Waagepetersen and Sorensen (2001) which follows the lines of Green (1995) but in more detail and minimising the measure theoretic notation.

Sisson (2005) provides an excellent review of trans-dimensional MCMC in the decade since Green (1995), with good coverage of the literature. Emphasis is placed on model choice applications, the efficient construction of samplers, and convergence diagnostics, but the review also covers other relevant work. Particularly helpful is the collation of freely available software for implementing various reversible jump algorithms.

Related to reversible jump, several authors have introduced different perspectives on trans-dimensional sampling. Keith *et al.* (2004) have proposed a novel broad framework for many samplers, including RJMCMC, in the guise of their 'generalised Markov sampler'. The idea is to augment the state space $\mathcal{X}$ by a discrete set, whose role is to provide an index to the type of the next transition, that is, to the move m in the language of section 2.3 above. This formalism provides an alternative view of state-dependent mixing over moves that the authors have found useful in applications, notably in phylogenetic inference. In contrast, Besag (1997) and Besag (2000) give a novel formulation in which variable-dimension notation is circumvented by embedding all θ_k within one compound vector. We consider the related product-space formulations in section 6, along with other approaches.

There are also a number of alternative sampling methods to reversible jump. One example is presented by Petris and Tardella (2003), who propose a formalism, directed primarily at situations where all models are nested within each other but possibly capable of generalisation, in which the variable-dimension character of a model choice problem is finessed. All models are embedded into the largest subspace, and the probability atoms induced by smaller models are smeared out across a neighbourhood. The original models can be recovered by transformation.

3 A simple example and existing literature

We highlight the ideas of section 2 with an illustrative example, chosen for its simplicity, allowing us to avoid the complexities that exist in many problems. However, we note that for such a simple example, the use of within-model approaches (Green 2003) may be more appropriate than reversible jump MCMC.

3.1 Poisson versus negative binomial

When modelling count data a question that is often of interest is whether the data is over-dispersed relative to a Poisson distribution. In such cases, data may be better modelled by a negative binomial distribution.

For data Y of length N, the likelihood under an independent identically distributed Poisson model with parameter $\lambda > 0$ is

$$p(Y|\lambda) = \prod_{i=1}^{N} \frac{\lambda^{Y_i}}{Y_i!} \exp(-\lambda),$$

whereas under an independent identically distributed negative binomial model with parameters $\lambda > 0$ and $\kappa > 0$ it is

$$p(Y|\lambda, \kappa) = \prod_{i=1}^{N} \frac{\lambda^{Y_i}}{Y_i!} \frac{\Gamma(1/\kappa + Y_i)}{\Gamma(1/\kappa)(1/\kappa + \lambda)^{Y_i}} (1 + \kappa\lambda)^{-1/\kappa}.$$

For both distributions the mean is given by λ. For the negative binomial distribution the parameter κ characterises the over-dispersion relative to a Poisson distribution, such that the variance is given by $\lambda(1 + \kappa\lambda)$.

Newton and Hastie (2006) consider a question along these lines in the context of tumour counts in genetically-engineered mice. To avoid the complexities intrinsic in their problem, we consider an example applied to total goals data from $1,040$ English Premiership soccer matches for the seasons 2005/06 to 2007/08, treated simplistically as if this was a simple random sample.

Adopting the framework above our problem is a very simple model choice problem. When $k = 1$, we suppose $Y_i \sim \text{Poisson}(\lambda)$, for $i = 1, 2, \ldots, N$. Using the notation introduced above, $\theta_1 = \lambda$. For $k = 2$, the data is allowed to follow a negative binomial distribution so that $Y_i \sim \text{NegBin}(\lambda, \kappa)$, meaning $\theta_2 = (\lambda, \kappa)$. Over-dispersion in model 2 may be indicative of other effects, such as team effects, that are not captured by a global mean parameter λ.

For our Bayesian approach, our priors on each model are such that $p(k = 1) = p(k = 2) = 0.5$. For θ_1 and $\theta_{2,1}$ (corresponding to λ in models 1 and 2 respectively) we use a $\text{Gamma}(\alpha_\lambda, \beta_\lambda)$ prior. For $\theta_{2,2}$ we adopt a $\text{Gamma}(\alpha_\kappa, \beta_\kappa)$ prior. This results in a posterior distribution of

$$p(k, \theta_k|Y) \propto \begin{cases} \frac{1}{2} p(\theta_1|k = 1) p(Y|\theta_1) & \text{for } k = 1 \\ \frac{1}{2} p(\theta_{2,1}, \theta_{2,2}|k = 2) p(Y|\theta_{2,1}, \theta_{2,2}) & \text{for } k = 2. \end{cases}$$

where

$$p(\theta_1|k=1) = \gamma(\theta_1, \alpha_\lambda, \beta_\lambda), \quad p(\theta_{2,1}, \theta_{2,2}|k=2) = \gamma(\theta_{2,1}, \alpha_\lambda, \beta_\lambda) \times \gamma(\theta_{2,2}, \alpha_\kappa, \beta_\kappa).$$

and $\gamma(\cdot, \alpha, \beta)$ is the density of the Gamma(α, β) distribution.

We choose $\alpha_\lambda = 25$ and $\beta_\lambda = 10$, giving a mean value of 2.5, which is typical for total goals in a football match. In addition, we choose $\alpha_\kappa = 1$ and $\beta_\kappa = 10$. These priors result in an average of around 25% extra variance for the negative binomial distribution.

Despite being an integral part of our MCMC sampler, we do not illustrate within-model moves as these are straightforward fixed-dimensional Metropolis–Hastings moves. However, the sampler also needs to be able to jump between models 1 and 2, and noting that these are of different dimensions, reversible jump methodology must be applied.

Consider the move from model 1 to model 2. Let $x = (1, \theta)$ be the current state of the chain. Since there is no equivalent to the parameter κ in model 1, we proceed using an independence approach. Specifically, we generate u from a $N(0, \sigma)$ distribution, where σ is fixed, so that g is the density of this distribution. We then set $x' = (2, \theta')$, where $\theta' = (\theta'_1, \theta'_2) = h(\theta, u) = (\theta, \mu \exp(u))$, for some fixed μ. In words, the parameter λ is maintained between models, but the new parameter κ is a log-normal random variable, multiplicatively centred around μ.

It is trivial to calculate the Jacobian factor, giving

$$|J| = \begin{vmatrix} \frac{\partial \theta'_1}{\partial \theta_1} & \frac{\partial \theta'_1}{\partial u} \\ \frac{\partial \theta'_2}{\partial \theta_1} & \frac{\partial \theta'_2}{\partial u} \end{vmatrix} = \mu \exp(u)$$

The reverse move, from model 2 to 1, requires no random variable u' (i.e. $r' = 0$), instead just setting $(\theta, u) = h'(\theta') = (\theta'_1, \log(\theta'_2/\mu))$. This means the acceptance probability for the move from model 1 to 2 is $\min\{1, A_{1,2}\}$, where

$$A_{1,2} = \frac{p(2, \theta'|Y)}{p(1, \theta|Y)} \left\{ \frac{1}{\sqrt{2p\sigma^2}} \exp\left[\frac{-u^2}{2\sigma^2} \right] \right\}^{-1} \mu \exp(u)$$

and from model 2 to 1 is $\min\{1, A_{2,1}\}$, where

$$A_{2,1} = \frac{p(1, \theta|Y)}{p(2, \theta'|Y)} \frac{1}{\sqrt{2p\sigma^2}} \exp\left[\frac{-(\log(\theta'_2/\mu))^2}{2\sigma^2} \right] \frac{1}{\theta'_2}.$$

Note that, as must be the case, these are reciprocals after change of notation.

Importantly, our specification of g and h was not restricted; any choice of g and h is valid, but different choices will lead to algorithms that perform differently. As an example, we might alternatively have chosen $u \sim \text{Exp}(\beta)$, for some fixed β, and $h(\theta, u) = (\theta, u)$.

For our proposal, the parameters μ and σ are crucial to the success of the algorithm; poorly chosen values may lead to slow convergence and ultimately even non-convergence during a run of the sampler. In this example, μ can be chosen naturally: by considering $\text{Var}(Y)/\mathbb{E}(Y)$ and approximating $\mathbb{E}(Y)$ by $\bar{y}$ and $\text{Var}(Y)$ by the sample variance, we set $\mu = 0.015$. Note that for a poorer choice of $\mu = 1.0$, no trans-dimensional moves were accepted in our runs, so that the sampler remained in the model it had been initialised in. The choice of σ is less sensitive, although we discuss this a little further below.

We ran our sampler for 50 000 sweeps, with an additional burn-in of 5 000 sweeps. At each sweep a trans-dimensional move was attempted, along with within-model moves for each parameter. The posterior probability of the models were $p(k = 1|Y) = 0.708$ and $p(k = 2|Y) = 0.292$. Figure 3.1

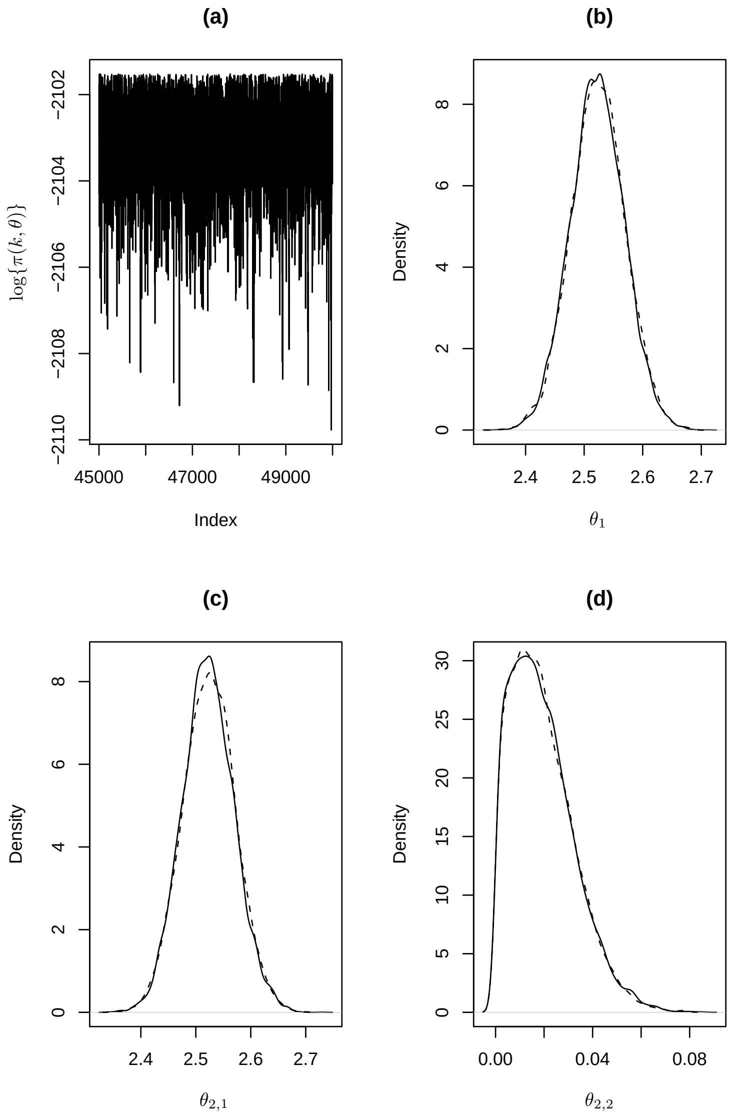

Figure 3: MCMC output for example problem: (a) Trace of log posterior for last 5 000 sweeps of sampler; (b) density estimate for θ_1 when $k = 1$; (c) density estimate for $\theta_{2,1}$ when $k = 2$; and (d) density estimate for $\theta_{2,2}$ when $k = 2$. Solid lines are problem specific sampler of section 3.1, dashed lines are AutoMix sampler of section 5.

shows the trace plot of the log posterior for the last 5 000 sweeps, along with density estimates for θ_1, $\theta_{2,1}$ and $\theta_{2,2}$ (solid lines).

Although there appears to be some support for model 2, the Poisson distribution of model 1 has higher posterior probability. Because the data exhibits only slight over-dispersion relative to the Poisson distribution, our prior specification for κ impacts on the posterior support for model 2.

We did not attempt to optimise the choice of σ but note that for this example, setting $\sigma = 1.5$ gave an acceptance rate of 58%, compared to 8% when $\sigma = 0.05$. This lower acceptance rate leads to higher autocorrelation in the k chain, although both samplers converged within the burn-in period. For harder problems, sub-optimal choice of proposal scaling parameter cause convergence difficulties.

We return briefly to this example in section 5.1. We refer the reader to Chapter 3 of Hastie (2005), and Newton and Hastie (2006) for a related more complex example.

3.2 The literature on reversible jump

In order to get an idea of the relevance of reversible jump MCMC, we might consider the citations of the original reversible jump paper in other publications. In fact, according to the ISI *Web of Knowledge*, at the time of writing, there are over 1 400 citations in over 500 journals, although many of these may be simply mentions in passing. Some of the non-trivial citations are in reviews or tutorials, or prove mathematical properties, others propose alternative approaches. However, the majority are implementations of reversible jump, typically presenting statistical methodologies using the method, recipes of generic methodological significance, or applications to specific analyses of data. Even a superficial review of this work would take up far more space than is available here, but the reader is urged to consult this literature, using search engines, etc., before starting on a purportedly novel application.

Some idea of the balance between methodology and different broad application domains can be obtained by noting that around 45% of the articles citing Green (1995) are in statistics and probability, about 28% in biology, genetics and medicine (Sisson (2005) notes that "one in every five citations . . . can be broadly classified as genetics-based research"), about 20% in computer science and engineering, and about 15% in other disciplines ranging from archaeometry through management science to water resources research.

Later sections in this article highlight some of the more important methodological extensions to reversible jump.

4 Challenges of implementation

Although applications of reversible jump have been diverse, much of the reported work has been carried out by MCMC "experts". With wider adoption, the method might yield promising analysis of many more problems.

Part of the apparent reluctance to adopt reversible jump methods, is a belief that such samplers are difficult to employ. This perception may in part be fueled by the often complex and formal language used to present the method. In truth, reversible jump at the practical level is quite simple and it is not necessary to fully understand the underlying technicalities in order to apply the method. Nonetheless, while the method itself may not be complicated, for many applications the complexity of the space $\mathcal{X}$ may present challenging issues.

Specifically, the construction of across-model proposals between the state spaces $\mathcal{X}_k$ may appear difficult, as natural ideas of proximity and neighbourhood that help guide the design of within-model proposals may no longer be intuitive. Heikkinnen (2003) demonstrates that in some extreme instances, this may lead to difficulty in designing valid proposals. More commonly, designing valid proposals is not the challenge, but difficulty lies in ensuring that the chosen proposals are efficient.

Inefficient proposal mechanisms result in Markov chains that are slow to explore the state space, and hence demonstrate slow convergence to the stationary distribution p; this leads to the Markov chain having high autocorrelation, which increases the asymptotic variance of Monte Carlo estimators.

In fixed-dimensional MCMC, proposed new states will be accepted with high probability if they are very close to the current state. Inefficiency can be caused by not proposing large moves away from the current state of the chain or by proposing bolder moves that have associated acceptance probabilities that are prohibitively small. For the random-walk Metropolis and Langevin algorithms and simple target distributions this comes down to a well studied question of optimal scaling of a proposal variance parameter (Roberts *et al.* 1997; Roberts and Rosenthal 1998).

For reversible jump, within-model proposals are no different than fixed-dimensional MCMC so identical principles apply. For across-model proposals the lack of a concept of closeness means that frequently it is the problem of low acceptance probabilities that makes efficient proposals hard to design; it is usual for across-model moves to display much lower acceptance probabilities than within-model moves.

Generally, the intuitive principle behind efficient across-model proposal design is to ensure that our proposed new state $(k', \theta'_{k'})$ will have similar posterior support to our existing state (k, θ_k). This ensures that the move and its reverse will both have a good chance of being accepted. While this may be easier said than done, for common move types which appear in a number of applications, such as the *split–merge* move type introduced by Richardson and Green (1997), general ideas such as moment matching can help achieve this aim. See Richardson and Green (1997) and Green and Richardson (2001) for details.

In order to achieve an efficient reversible jump algorithm for a specific problem of interest, many aspects of the proposal mechanism need to be carefully specified and then tuned using pilot runs. Examples for tuning include scaling, blocking and re-parameterisation. This process is often arduous and has motivated researchers to concentrate efforts on providing more general techniques for efficient proposal design, to help unlock the full potential of reversible jump MCMC. We dedicate the remainder of this section to a review of a selection of the advances that have been made in this area.

4.1 Efficient proposal choice for reversible jump MCMC

Perhaps the most substantial recent methodological contribution to the general construction of proposal distributions is work by Brooks *et al.* (2003b). The authors propose several new methods, falling into two main classes. Their methods are implemented and compared on examples including choice of autoregressive models, graphical gaussian models, and mixture models.

Order methods, which are the first class of methods, focus mainly on the quantitative question of efficiently parameterising a proposal density ($g(u)$ in section 2.1), having already fixed the transformation ($(\theta', u') = h(\theta, u)$) into the new space. This is achieved by imposing various constraints on the acceptance ratio (4), for jumps between the existing state θ_k in model k and an appropriately chosen "centring point" $c_{k,k'}(\theta_k)$ in k'. The centring point is chosen with the aim of being the equivalent state in model k' of θ_k in model k (in a sense defined within the paper).

The detail of the order methods is in the specific constraints that are used. The constraint imposed by the zeroth-order method is that

$$A((k, \theta_k), (k', c_{k,k'}(\theta_k))) = 1,$$

where A is the acceptance ratio for a particular move-type, as in equation (4). By scaling the proposal so that the acceptance ratio is 1 for a jump to the chosen centring point, frequent across-model moves are encouraged. The authors present a simple motivating example, wherein the method results in transition probabilities that are optimal (in a sense explained within the paper).

Constraints imposed for the first-order (and higher-order) methods set the first (and higher) order derivatives of the acceptance ratio (with respect to the random numbers u) equal to 0, so for example,

$$\nabla A((k, \theta_k), (k', c_{k,k'}(\theta_k))) = \mathbf{0}.$$

First- and higher-order methods are inspired by the Langevin algorithm (Roberts and Rosenthal (1998)), with the attractive property that the acceptance probability remains high in a region around the centring point. The number of order methods that can be used depends upon the number of parameters to be determined in the proposal distribution g, but numerical support for first and higher order methods is strong, leading to significant performance increases.

The second class of methods, named the *saturated space approach* work in a product-space formulation somewhat like that in section 6.2.

Essentially, the idea is to augment the state space $\mathcal{X}$ with auxiliary variables, to ensure that all models share the same dimension $n_{\max}$ as that of the "largest" model. MCMC is then used to create a chain with stationary distribution equal to an augmented target distribution, which combines the target distribution p and the distributions of the auxiliary variables.

Inclusion of auxiliary variables aids across-model moves, essentially rendering them fixed-dimensional. In the updating mechanism introduced by Brooks *et al.* (2003b), conditional upon the selection of one of a finite number of transforms, the proposal to a state in a different model is deterministic. Randomness is achieved by within-model updates, applied to both model parameters and the auxiliary variables, allowing temporal memory and possible dependency in the auxiliary variables. In essence, this allows the chain to have some memory of states visited in other models, resulting in more efficient proposals.

Ehlers and Brooks (2008) extend this work for time series models, looking at more flexible reversible jump moves. Godsill (2003) suggests further developments, possibly using only a randomly selected subset of the auxiliary variables when proposing the new state.

4.2 Adaptive MCMC

Another area of recent research offering efficiency gains is adaptive sampling. The underlying idea is that under suitable conditions the proposal mechanisms may be allowed to depend on past realisations of the chain, not just the current state, without invalidating the ergodicity of the resulting process. In other words, the resulting chain may still be used to make inference about the target distribution. This observation means that questions such as optimal location and scaling of proposals can be determined online during the run of the algorithm, eliminating the need for tuning and pilot runs.

Research into adaptive sampling has taken two distinct directions known as *diminishing adaptation* and *adaptation through regeneration*, which differ in how the adaptation of proposal distributions occur.

Diminishing adaptation is the most popular of these and allows adaptation to continue indefinitely, but at a rate that is decreasing as the chain progresses. Rosenthal, this volume, provides an introduction to this approach, including the assumptions required for the validity of the approach. Comprehensive references for further reading are also provided.

Erland (2003) provides a review of adaptation through regeneration, which requires the existence of parts of the state space where the Markov chain regenerates (i.e. the sub-chains separated by visits to these regeneration areas are independent of each other with some probability ψ). It is then valid to adapt the proposal distribution upon visits to these regeneration states.

Little work has yet been done to extend adaptive MCMC to the more general moves of reversible jump. For within-model moves, adaptive proposals could be applied, however for across-model moves the situation is more difficult. Hastie (2005) discusses adapting the probabilities $j_m(x)$, in the particular case where the probabilities of proposing a jump from one model to another do not depend on either k or θ_k. (Note that the new state $\theta'_{k'}$ in model k' is still allowed to depend on $x = (k, \theta)$). Two methods are suggested for adaptation, the most promising being a diminishing adaptation algorithm. Although not all of assumptions that guarantee convergence are confirmed, numerical results are encouraging and we hope that subsequent research will extend the methods to more general across-model moves.

4.3 Other ways of improving proposals

An interesting modification to Metropolis–Hastings is the splitting rejection idea of Tierney and Mira (1999), extended to the reversible jump setting by Green and Mira (2001), who call it *delayed rejection*.

Using this algorithm, if a proposal to x' is rejected (with the usual probability $1 - \alpha(x, x')$), instead of immediately taking the new state of the chain to be the existing state x, a secondary proposal to x'' is attempted. This is accepted with a probability that takes into account the rejected first proposal, in a way that the authors show maintains detailed balance for the compound transition.

Numerical results demonstrate efficiency improvements, but the benefits of more accepted across-model moves needs to be weighed against the increased computational cost of the two stage proposal. Combining a “bold” first proposal with a conservative second proposal upon rejection, might lead to a sampler that better explores the state space but the question of sensible proposal design for general reversible jump problems remains difficult.

Other authors have also tried to adapt the reversible jump algorithm to improve across-model acceptance rates. Al-Awadhi *et al.* (2004) propose across-model moves that make clever use of an intermediate within-model chain, which maintains detailed balance with respect to an alternative distribution p^*. By making the p^* a flatter version of p, the aim is to encourage moves in situations where the conditional distributions $p(\theta_k|k)$ are multi-modal; such cases often have near zero acceptance rates for across-model moves. While the algorithm increases the rates, they remain small, at a non-negligible increase in computational cost. A similarly motivated idea by Tjelmeland and Hegstad (2001), modifies the acceptance probability by considering pairs of proposal distributions, each one optimised at each iteration to locally approximate a mode of the posterior distribution. Again, acceptance rate gains are realised but the optimisation would be prohibitively expensive in many problems, especially those with high dimensional spaces.

While the research above highlights the progress being made, important questions such as the choice of g and h remain largely unaddressed. Furthermore, with the possible exception of adaptive MCMC, there remains the need for tuning runs. In section 5 we present a sampler designed to address these issues.

4.4 Diagnostics for reversible jump MCMC

Monitoring of MCMC convergence on the basis of empirical statistics of the sample path is important, while not of course a substitute for a good theoretical understanding of the chain. There has been some concern that across-model chains are intrinsically more difficult to monitor, perhaps almost amounting to this being a reason to avoid their use.

In truth, the degree of confidence that convergence has been achieved provided by ‘passing’ a diagnostic convergence test declines very rapidly as the dimension of the state space increases. In more than, say, a dozen dimensions, it is difficult to believe that a few, even well-chosen, scalar statistics give an adequate picture of convergence of the multivariate distribution. It is high, rather than variable, dimensions that are the problem.

In most trans-dimensional problems in Bayesian MCMC it is easy to find scalar statistics that retain their definition and interpretation across models, typically those based on fitted and predicted values of observations, and these are natural candidates for diagnostics, requiring no special attention to the variable dimension.

However, recognising that there is often empirical evidence that a trans-dimensional simulation stabilises more quickly within models than it does across models, there has been recent work on diagnostic methods that address the trans-dimensional problem more specifically. The promising approach by Brooks and Giudici (2000), following Brooks and Gelman (1998), is based on analysis of sums of squared variation in sample paths from multiple runs of a sampler. This is decomposed into terms attributable to between- and within-run, and between- and within-model variation.

More recently, Sisson and Fan (2007) have extended this idea to propose a specific distance-based diagnostic for trans-dimensional chains, applicable whenever the unknown x can be given a point process interpretation, essentially, that is, whenever the variable-dimension aspect of the parameter vector consists of exchangeable sub-vectors. Examples include change-point problems and mixture models.

5 Automatic RJMCMC

One idea aimed at eliminating the intricacies of sampler design is that of an automatic reversible jump MCMC sampler that can be applied to any given target distribution. The first steps in this direction were taken by Green (2003), motivated by the fixed-dimensional random-walk Metropolis sampler.

The assumption is made that an across-model move from model k to model k' is proposed with some probability $q(k, k')$ that does not depend on θ_k. Under this set up, the central idea is that in order to maximise the acceptance probability for the move, $\theta'_{k'}$ would ideally be sampled from the conditional distribution $p(\theta'_{k'}|k')$. Although typically these conditional distributions are not known, Green (2003) suggests using Normal distributions that crudely approximate these conditionals as proposal distributions. Hastie (2005) introduces the *AutoMix* sampler, extending this approach by exploring the possibility that for each k, a mixture approximation to $p(\theta_k|k)$ could be used instead.

Hastie (2005) supposes that for model k there are L_k components in the mixture, indexed by l, each with weight λ_k^l, fixed n_k-dimensional mean-vector μ_k^l, and fixed $n_k \times n_k$-matrix B_k^l such that $B_k^l[B_k^l]^T$ is the covariance matrix. By allocating the existing state θ_k to a component l_k in model k with probability $p_{k,\theta_k}(l_k)$, and choosing a component $l'_{k'}$ in model k' with probability $\lambda_{k'}^{l'_{k'}}$ the proposed new state $\theta'_{k'}$ depends on dimensions n_k and $n_{k'}$ as follows:

$$\theta'_{k'} = \begin{cases} \mu_{k'}^{l'_{k'}} + B_{k'}^{l'_{k'}}[(B_k^{l_k})^{-1}(\theta_k - \mu_k^{l_k})]_1^{n_{k'}} & n_{k'} < n_k \\ \mu_{k'}^{l'_{k'}} + B_{k'}^{l'_{k'}}(B_k^{l_k})^{-1}(\theta_k - \mu_k^{l_k}) & n_{k'} = n_k \\ \mu_{k'}^{l'_{k'}} + B_{k'}^{l'_{k'}} \begin{pmatrix} (B_k^{l_k})^{-1}(\theta_k - \mu_k^{l_k}) \\ u \end{pmatrix} & n_{k'} > n_k \end{cases} . \tag{5}$$

Here, $[\cdot]_1^m$ denotes the first m elements of a vector and u is an $(n_{k'} - n_k)$-vector of random numbers drawn from density g', which are taken to be that of independent standard Normal distributions or independent Student t distributions.

Following simple arguments, Hastie (2005) shows that detailed balance is preserved if the move is accepted with probability $\alpha(x, x') = \min\{1, A(x, x')\}$ where

$$A_(x, x') = \frac{p(k', \theta'_{k'})}{p(k, \theta_k)} \frac{p_{k',\theta'_{k'}}(l'_{k'})}{p_{k,\theta_k}(l_k)} \frac{q(k', k)}{q(k, k')} \frac{\lambda_k^{l_k}}{\lambda_{k'}^{l'_{k'}}} \frac{|B_{k'}^{l'_{k'}}|}{|B_k^{l_k}|} G_{k,k'}(u)$$

and

$$G_{k,k'}(u) = \begin{cases} g'(u) & n_{k'} < n_k \\ 1 & n_{k'} = n_k \\ [g'(u)]^{-1} & n_{k'} > n_k \end{cases} .$$

Having allocated a state to a component l_k in the existing model, the state is standardised by using $\mu_k^{l_k}$ and $B_k^{l_k}$. If the proposal is to a model with a higher dimension, new standard random variables are appended to this standardised vector; if a model with a lower dimension is proposed, the appropriate number of elements are discarded. The new standardised vector is then transformed to $\theta'_{k'}$ using the mean, $\mu_{k'}^{l'_{k'}}$, and (matrix square root of the) covariance matrix, $B_{k'}^{l'_{k'}}$, corresponding to a randomly selected mixture component $l'_{k'}$. Notice also that if $q(k, k) > 0$, then the above mechanism might also be used for within-model moves, potentially allowing moves to states well separated from the current state, by jumping between different components of the mixture.

Central to the AutoMix sampler is the specification of the mixture distributions for each model. Hastie (2005) suggests looping over the models, performing preliminary adaptive random-walk Metropolis (RWM) pilot runs to obtain a sample from each posterior conditional and then fitting the mixtures using the EM-like algorithm of Figueiredo and Jain (2002). The increased cost of fitting mixtures compared to computing a mean vector and covariance matrix (as required by the automatic sampler introduced by Green (2003)) should not be overlooked; if the conditional posteriors appear to be largely unimodal then the more simple sampler may be preferable, although adaptive sampling at the initial RWM stage appears prudent.

As Hastie (2005) observes, the inclusion of within-model pilot runs and mixture fitting increases run-time considerably when compared to a reversible jump sampler designed for a particular problem. However, one should not discount the fact that an automatic sampler may be implemented with minimal user input, saving on the sampler design time. In addition, computational savings could in theory be made by replacing within-model pilot-runs with adaptive fitting of mixtures throughout the reversible jump stage.

5.1 A simple example revisited

Primarily automatic samplers are designed to be easy to apply and relatively broad in their applicability. As such, it is easy to apply such an approach to the problem we considered in section 3.1. By downloading the C program that implements the AutoMix software[1], we need only to specify a function which computes $\log p(k, \theta_k|y)$, along with simple other functions setting the maximum number of models, the dimension of each model and initial values for the chain.

Applying the AutoMix sampler for 50 000 reversible jump sweeps gives posterior model probabilities of 0.707 for model 1 and 0.293 for model 2. Density estimates for θ_1, $\theta_{2,1}$ and $\theta_{2,2}$ are included (dashed lines) in figure 3.1, demonstrating good agreement with the problem specific sampler.

Hastie (2005) applies the sampler to a number of non-trivial problems including the tumour count problem studied by Newton and Hastie (2006), and change-point processes applied to coal-mining disaster data as studied by Green (1995). Following a similar approach, Spirling (2007) uses the sampler to consider civilian casualty rates in the Iraq conflict. Furthermore, using the sampler

[1] Package including code, instructions and example files are freely available from http://www.davidhastie.me.uk/AutoMix

for a model choice problem for 2 mixed effects model, applied to data from an AIDS clinical trial, Hastie (2005) avoids issues of implementation, tuning and marginalisation as encountered by Han and Carlin (2001) who studied the problem in a comparison of reversible jump with other trans-dimensional approaches.

6 Subsequent and alternative methodologies

In section 3 we cited a handful of applications that have benefited from reversible jump. We have no doubt that the future will provide many other interesting problems for which reversible jump may yield important conclusions. Furthermore, we anticipate that future methodologies may be built using reversible jump methods as foundations. In the following sub-section we briefly note a selection of methods that fall into this category.

6.1 Methodologies exploiting RJMCMC

Based on sequential Monte Carlo (SMC) (see Doucet *et al.* (2001) for a review), Jasra *et al.* (2008) introduce a method they call *Interacting sequential Monte Carlo samplers* (ISMC). The key to ISMC is that several SMC samplers are run in parallel, initially on separate subspaces. For each sampler, at time $t < T$, particles are updated using MCMC moves (including reversible jump moves for trans-dimensional problems) so that they are samples from p_t, which is typically a version of p that facilitates mixing, for example by tempering. Importantly, $p_T = p$. When some predetermined time $t^* < T$ is reached, the separate samplers are combined and a single sampler is implemented, moving across all models. Jasra *et al.* (2008) take advantage of this formulation by using the separate samplers to provide samples for each model k, which are then used to fit a mixture distribution to approximate $p(\theta_k|k)$. The single SMC sampler then uses reversible jump moves very similar to those in the AutoMix sampler (see equation (5)), extended to include an identifiability constraint that is necessary for their application.

The authors present their work for an example in population genetics, demonstrating a marked improvement of between-model mixing over regular SMC methods, albeit at a cost of increased computational time.

We note the similarities between the ISMC method and the population reversible jump MCMC method introduced by Jasra *et al.* (2007). Extending population MCMC (Cappé *et al.* 2004) to the trans-dimensional case, and drawing on the ideas of evolutionary Monte Carlo (Liang and Wong 2000), this algorithm also employs tempered distributions (again to encourage mixing) but this time in parallel. Markov chains are constructed using reversible jump methods to sample from each distribution, but the parallel chains also interact by including moves that allow the states to be swapped or combined. The authors prove the ergodicity of the resulting algorithm, and show for a particular hard genetic example, mixing between models is improved.

Tempering based ideas have also been used by other authors. Gramacy *et al.* (2008) detail a further related method which combines simulated tempering using RJMCMC and importance sampling, allowing samples from p to be recovered when using the tempered distributions. Simulated tempering and reversible jump are also combined by Brooks *et al.* (2006), who use the ideas to create a perfect simulation algorithm to provide exact samples from the target distribution p.

A common use for reversible jump is to process the output from the chain to assess support for the various models by calculating the *Bayes factor*, $B_{k,k'} = p_k(Y)/p_{k'}(Y)$, where $p_k(Y) = \int p(Y|\theta_k, k)p(\theta_k|k)\mathrm{d}\theta_k$ is the marginal likelihood of model k. Alternatively this can be written as the ratio of posterior and prior odds of models k and k'. Assuming equal prior probabilities on models k and k', this motivates the simple estimate of $B_{k,k'}$ as $J_k/J_{k'}$, where J_k is the number of visits (out of a chain of length J) to model k.

Applying the concept of Rao–Blackwellisation, Bartolucci *et al.* (2006) propose an improved estimate (in terms of reduced variance) by using the bridge sampling identity (Meng and Wong 1996), given by

$$B_{k,k'} = \frac{\mathbb{E}_{k'}[\phi(\theta_k)p(Y|\theta_k,k)p(\theta_k|k)]}{\mathbb{E}_k[\phi(\theta'_{k'})p(Y|\theta'_{k'},k')p(\theta'_{k'}|k')]} \tag{6}$$

for a general function ϕ, where $\mathbb{E}_k$ is the expectation with respect to $p(\theta_k|Y) \propto p(Y|\theta_k,k)p(\theta_k|k)$.

Bartolucci *et al.* (2006) extend equation (6) to the trans-dimensional case and suggest a choice of the function ϕ that requires no extra computational cost. For models k, k', where a jump is proposed between these models at each sweep, the resulting estimate is:

$$B_{k,k'} = \frac{\sum_{i=1}^{J_{k'}} \alpha_{k',k}((\theta'_{k'})^i, \theta_k^i)/J_{k'}}{\sum_{i=1}^{J_k} \alpha_{k,k'}(\theta_k^i, (\theta'_{k'})^i)/J_k},$$

where θ_k^i is the value of θ_k at the i^{th} visit to model k and $\alpha_{k,k'}(\theta_k, \theta'_{k'})$ is the reversible jump acceptance probability of moving from (k, θ_k) to $(k', \theta'_{k'})$. For the examples considered, the improvements are marked.

Probabilistic inference is not the only use of MCMC methodology. A specific example, is the *simulated annealing* (Geman and Geman 1984) algorithm for function optimisation, recently extended for trans-dimensional problems where an optimal model may need to be determined, see Brooks *et al.* (2003a) and Andrieu *et al.* (2000). For a particular function $f(k, \theta_k)$, then it is possible to construct the Boltzmann distribution with parameter T, with density $b_T(k, \theta_k) \propto \exp(-f(k, \theta_k)/T)$. The function f is the quantity that we wish to minimise, perhaps with some penalisation term, for example to mimic the AIC or BIC (Andrieu *et al.* (2000)). Trans-dimensional simulated annealing proceeds by using reversible jump moves, to construct a Markov chain where the invariant distribution for phase i is the Boltzmann distribution with parameter T_i. Once equilibrium has been reached, the temperature T_i is decreased, and a new phase is started from the state the chain ended in. By decreasing T_i in this manner, we are left with a distribution with all its weight in the global minima, resulting in an effective optimisation algorithm.

6.2 Alternatives to reversible jump MCMC

It is important to observe that there are several alternative formalisms for across-model simulations. While full coverage of these methods falls outside the scope of this paper, considering reversible jump as one of a wider class of methods can be instructive for developing a better understanding and guiding future research into RJMCMC methods. We now reference a few of the most relevant alternatives.

Predating reversible jump, Grenander and Miller (1994) proposed a sampling method they termed *jump diffusion*, involving between-model jumps and within-model diffusion according to a Langevin stochastic differential equation. Had the sampler been corrected for time discretisation by using a Metropolis–Hastings accept/reject decision, this would have been an example of reversible jump.

Various trans-dimensional statistical models can be viewed as abstract *marked point processes* (Stephens 2000). In these problems, the items of which there are a variable number are regarded as marked points. For example in a normal mixture model the points represent the (mean, variance) pairs of the components, marked with the component weights. Stephens (2000) borrows the birth-and-death simulation idea of Preston (1977) and Ripley (1977) to develop a methodology for finite mixture analysis. The key feature that allows the approach to work for a particular application is the practicability of integrating out latent variables so that the likelihood is fully available.

Extending the point process idea, Cappé *et al.* (2003) have recently given a rather complete analysis of the relationship between reversible jump and continuous time birth-and-death samplers.

Unlike reversible jump, the birth-death process accepts all across-model moves, but maintains detailed balance through the length of time spent in each model. The authors conclude that little benefit is gained from formulating a problem one way or another, as low acceptance rates in reversible jump are just replaced by significant phases where the point process approach does not move between models. Nonetheless, as mentioned in section 4, this alternative formulation can be useful in other respects, such as the convergence diagnostic proposed by Sisson and Fan (2007).

Another class of alternative methods is termed the *product space* approach and was first used to consider trans-dimensional problems by Carlin and Chib (1995). Since then, work has been done to extend the method (Green and O'Hagan 1998; Dellaportas *et al.* 2002), leading to the more general *composite model space* framework of Godsill (2001). Sisson (2005) provides a review.

As in the saturated state space of Brooks *et al.* (2003b) (see section 4), the idea is to work on a more general state space, where the simulation keeps track of all θ_k rather than only the current one. Thus the state vector is of fixed dimension, circumventing the trans-dimensional nature of the problem.

Letting θ_{-k} denote the composite vector consisting of all $\theta_l, l \neq k$ concatenated together, the joint distribution of $(k, (\theta_l : l \in \mathcal{K}), Y)$ can be expressed as

$$p(k)p(\theta_k|k)p(\theta_{-k}|k,\theta_k)p(Y|k,\theta_k). \tag{7}$$

The third factor $p(\theta_{-k}|k,\theta_k)$ has no effect on the joint posterior $p(k,\theta_k|Y)$; the choice of these conditional distributions, which Carlin and Chib (1995) call 'pseudo-priors', is entirely a matter of convenience. However, the efficiency of the resulting sampler depends entirely on these quantities, effective meaning that the choice of efficient proposal distribution for reversible jump is replaced by the specification of appropriate pseudo-priors.

Godsill (2001)'s formulation extends equation (7) to allow the parameter vectors θ_k to overlap arbitrarily, and embraces both product space and reversible jump methods, facilitating comparisons between them. The framework also provides useful insight into some of the important factors governing the performance of reversible jump. Godsill (2003) discusses these issues in some detail, including using retained information from past visits to other models, to help design effective across-model moves.

Whether or not jumping between parameter subspaces benefits sampler performance has been a question of some debate. Han and Carlin (2001) suggest that MCMC samplers that avoid a model space search may result in estimates with improved precision whereas Richardson and Green (1997) present an example that suggests the contrary. In fact, there is no one answer, and in some instances trans-dimensional moves will help samplers, whereas in others they will be unnecessary. Green (2003) considers this question in more detail, using a simple example to provide insight.

Little research has comprehensively compared the performance of reversible jump and product space methods. Dellaportas *et al.* (2002) study the methods in the context of model choice, along with some less generally applicable approaches. There is little to differentiate the results from alternative approaches, and both approaches perform adequately. However, neither reversible jump proposal design or product space pseudo-prior specification appear particularly hard for the examples they consider. More research would be welcome in this area, but we believe that for difficult problems, implementation of both approaches will involve complex practical issues; which method yields the better results may come down to which method the researcher has most experience with.

7 Practical Bayesian model determination

Having considered *how* we might make inference for 'ideal Bayesian' model determination, we set aside our review of methodology and return to some of the more fundamental issues that face the practitioner wishing to perform model choice for a real Bayesian problem.

In many instances it may be that that the analyst does not wish to report the marginal model posterior probabilities $p(k|Y)$. Of course, using these probabilities we could alternatively report Bayes Factors

$$B_{kl} = \frac{p(Y|k)}{p(Y|l)} = \frac{p(k|Y)}{p(l|Y)} \div \frac{p(k)}{p(l)}$$

for pairwise comparison of models. For some, the marginal likelihood, $p(Y|k)$, itself has an intrinsic meaning and interpretation. In either case, the same methods for posterior computation will yield results.

A more philosophical question is that of model choice versus model averaging. With model choice we may be interested in determining a best model $k^\star$ (understanding the uncertainty of this choice), and restricting our inference about the model parameters $\theta_{k^\star}$ (and functions of these parameters) conditional upon this model. The benefit of this approach is that reported parameters summaries retain natural interpretations.

On the other hand, to properly account for model uncertainty, we should average over models. Again, the same posterior computation allows model averaging

$$E(F|Y) = \sum_k \int F(k, \theta_k) p(k, \theta_k|Y) \mathrm{d}\theta_k$$

for any function F with the same interpretation in each model. One example is prediction where

$$p(Y^+|Y) = \sum_k p(Y^+|k, Y) p(k|Y)$$

is a posterior-weighted mixture of the within-model-k predictions

$$p(Y^+|k, Y) = \int p(Y^+|k, \theta_k) p(\theta_k|k, Y) \mathrm{d}\theta_k.$$

When probabilities are computed using reversible jump MCMC methods, the expectation can be estimated simply by averaging F along the entire run, essentially ignoring the model indicator k.

Some would argue that it is only responsible to adopt a Bayesian hierarchical model of the kind introduced above when there is compatibility between models, that is, when the parameter priors $p(\theta_k|k)$ are such that inference about functions of parameters that are meaningful in several models should be approximately invariant to k. Such compatibility could in principle be exploited in the construction of MCMC methods, although we are not aware of general methods for doing so.

Even taking these considerations into account, we have remained focused on the ‘ideal Bayes’ approach. In reality there are a number of reasons why this simple idealised view fails to reflect practical applications.

7.1 Prior model probabilities may be fictional

The ideal Bayesian has real prior probabilities (perhaps imparted by colleagues) reflecting scientific judgement or belief across the model space. In practice, however, such priors may not be commonly available. For example, we know a lot more about how to elicit scientific judgements about model parameters than we do about the models themselves. Arbitrariness in prior model probabilities may not affect Bayes factors (since prior probabilities cancel) but it sabotages Bayesian model averaging.

7.2 No chance of passing the test of a sensitivity analysis

In ordinary parametric problems we commonly find that inferences are rather insensitive to moderately large variations in prior assumptions, except when there are very few data. In fact, the opposite case, of high sensitivity, poses a greater challenge to the non-Bayesian as perhaps the data carry less information than hoped. However, it is clear that a test of sensitivity to model probabilities

$$\frac{p^\star(k|Y)}{p^\star(l|Y)} = \frac{p(k|Y)}{p(l|Y)} \times \left(\frac{p^\star(k)}{p^\star(l)} \div \frac{p(k)}{p(l)}\right)$$

will always fail, due to the fact that $\mathcal{K}$ is discrete and thus the second factor on the left hand side can be arbitrarily changed depending on the prior distributions assumed. This dependence on prior model probabilities is of course exactly the situation that use of Bayes factors avoids.

7.3 Improper parameter priors problems

In ordinary parametric problems it is commonly true that it is safe to use improper priors, specifically when posterior distributions are well-defined as limits of a sequence of approximating proper priors (without sensitivity to what that sequence is). However, when comparing models, improper parameter priors make Bayes factors indeterminate (since improper priors can only be defined up to arbitrary normalising constants, which persist into marginal likelihoods). Using proper but vague (or diffuse) priors alleviates this problem, but only partially, as the Bayes factors will then depend on the arbitrary degree of vagueness used.

In certain circumstances, ideas such as Intrinsic or Fractional Bayes factors, or Expected Posterior priors, can be applied, essentially based on tying together improper priors across different models. These ideas lose much of the appeal of ideal Bayes arguments, have arbitrary aspects, and are not widely accepted.

8 Conclusion

In conclusion, model uncertainty is a fact of life. As statistical scientists, we are still learning about when it can be quantified, eliminated or accommodated. The computational challenges posed by model determination can now often be met, for example using the methods in this article although there are still many directions in which future research might be directed. Amongst these, the guidance for efficient proposal design and the search for generic samplers remain elusive and challenging questions. Equally crucial is the area of reversible jump diagnostics: if we wish to encourage the wider adoption of the technique, then it is vital that we equip the user with tools for ascertaining that the conclusions that they reach are valid.

In many ways, the broad applicability of reversible jump, clearly a strength of the method, is also an obstacle. A panacea for the above questions is unlikely to be found, as no one method will be suitable for all problems. Rather, it is probable that RJMCMC will continue to evolve slowly, with researchers adding to the collection of existing methods and extensions, building upon contributions from many different perspectives

Beyond the methodological challenges, more philosophical problems remain about model determination. It is much easier to recognise uncertainty about models than it is to do objectively-justified quantification of that uncertainty. *Choosing* a single model is fraught with difficulties. Furthermore, there is probably no single best model, since the criteria by which we choose may be directed by different reasons. The real objective of inference may be prediction, for example, or it may be scientific understanding. We may be forced to choose a model for external reasons, such as presentation to a lay audience, or a policy maker, and led to conceal doubt about models for 'defensive' reasons.

In this article, we have not considered the alternatives to basing model determination simply on posterior distributions. There are many other criteria and developed methodologies of Bayesian hypothesis testing including complexity criteria such as AIC, BIC, DIC, DIC+, MDL, and C_p, decision theory, Bayesian p-values and posterior predictive checks. This diversity of approach reflects the different flavours of model determination question that statisticians face.

In fact, the very term 'model' may be too much of a catch-all. The ease with which many different tasks (such as choice between different scientific mechanisms, selection of predictors in regression, the number of components in a mixture, the order of an AR model, or the complexity of polynomial or spline) can be cast as problems of model determination, obscures the differences in character between these tasks. While there is certainly an obvious advantage in developing generic methods that are appropriate for all flavours of the problem, it may be that this task is too challenging. Greater progress may be made by focussing on application specific developments and subsequently adapting these for different problems where appropriate.

Acknowledgements

We are grateful to all our collaborators on reversible jump MCMC and related research problems: Graeme Ambler, Christophe Andrieu, Julian Besag, Stuart Coles, Carmen Fernández, Philip Giles, Paolo Giudici, Miles Harkness, Matthew Hodgson, Merrilee Hurn, Antonietta Mira, Michael Newton, Agostino Nobile, David Nott, Tony O'Hagan, Marco Pievatolo, Sylvia Richardson, Luisa Scaccia, Scott Sisson, Claudia Tarantola and Valerie Viallefont.

References

Al-Awadhi, F., Hurn, M., and Jennison, C. (2004). Improving the acceptance rate of reversible jump MCMC proposals. *Statistics & Probability Lett.*, **69**, (2), 189–98.

Andrieu, C., de Freitas, N., and Doucet, A. (2000). Reversible jump MCMC simulated annealing for neural networks. In *Uncertainty in Artificial Intelligence*, pp. 11–8. Morgan Kauffman, San Francisco, CA.

Bartolucci, F., Scaccia, L., and Mira, A. (2006). Efficient Bayes factor estimation from the reversible jump output. *Biometrika*, **93**, (1), 41–52.

Besag, J. E. (1997). Contribution to the discussion of paper by Richardson and Green. *Journal of the Royal Statistical Society, B*, **59**, 774.

Besag, J. E. (2000). Markov chain Monte Carlo for statistical inference. Technical Report Working paper no. 9, Center for Statistics and the Social Sciences, University of Washington.

Besag, J. E., Green, P. J., Higdon, D., and Mengersen, K. (1995). Bayesian computation and stochastic systems (with discussion). *Statistical Science*, **10**, 3–66.

Brooks, S. and Gelman, A. (1998). Some issues in monitoring convergence of iterative simulations. In *In Proceedings of the Section on Statistical Computing. ASA.*

Brooks, S. P. (1998). Markov chain Monte Carlo method and its application. *The Statistician*, **47**, 69–100.

Brooks, S. P., Fan, Y., and Rosenthal, J. S. (2006). Perfect forward simulation via simulated tempering. *Comm. In Statistics-simulation Computation*, **35**, (3), 683–713.

Brooks, S. P., Friel, N., and King, R. (2003a). Classical model selection via simulated annealing. *J. Royal Statistical Soc. Series B-statistical Methodology*, **65**, 503–20.

Brooks, S. P. and Giudici, P. (2000). Markov chain Monte Carlo convergence assessment via two-way analysis of variance. *J. Computational Graphical Statistics*, **9**, (2), 266–85.

Brooks, S. P., Giudici, P., and Roberts, G. O. (2003b). Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions. *J. Royal Statistical Soc. Series B-statistical Methodology*, **65**, 3–39.

Cappé, O., Guillin, A., Marin, J. M., and Robert, C. P. (2004). Population Monte Carlo. *J. Computational Graphical Statistics*, **13**, (4), 907–29.

Cappé, O., Robert, C. P., and Ryden, T. (2003). Reversible jump, birth-and-death and more general continuous time Markov chain Monte Carlo samplers. *J. Royal Statistical Soc. Series B-statistical Methodology*, **65**, 679–700.

Carlin, B. P. and Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo. *Journal of the Royal Statistical Society, B*, **57**, 473–84.

Dellaportas, P., Forster, J. J., and Ntzoufras, I. (2002). On Bayesian model and variable selection using MCMC. *Statistics and Computing*, **12**, (1), 27–36.

Didelot, X., Everitt, R., Johansen, A., and Lawson, D. (2011). Likelihood-free estimation of model evidence. *Bayesian Analysis*, **6**, 49–76.

Doucet, A., De Frietas, J. F. G., and Gordon, N. J. (2001). *Sequential Monte Carlo Methods in Practice*. Springer, New York.

Ehlers, R. S. and Brooks, S. P. (2008). Adaptive proposal construction for reversible jump mcmc. *Scandinavian Journal of Statistics*, **35**, 677–90.

Erland, S. (2003). *On adaptivity and Eigen-decompositions of Markov chains*. PhD thesis, Department of Mathematical Sciences, Norwegian University of Science and Technology.

Figueiredo, M. A. T. and Jain, A. K. (2002). Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24**, 381–96.

Gamerman, D. (1997). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman & Hall/CRC, New York.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattn Anal. Mach. Intell.*, **6**, 721–41.

Godsill, S. J. (2001). On the relationship between Markov chain Monte Carlo methods for model uncertainty. *J. Computational Graphical Statistics*, **10**, (2), 230–48.

Godsill, S. J. (2003). *Proposal densities and product-space methods*, pp. 199–203. OUP, Oxford.

Gramacy, R., Samworth, R., and King, R. (2008). Importance tempering. Technical report, Statistical Laboratory, University of Cambridge.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, (4), 711–32.

Green, P. J. (2003). *Trans-dimensional Markov chain Monte Carlo*, pp. 179–98. OUP, Oxford.

Green, P. J. and Mira, A. (2001). Delayed rejection in reversible jump Metropolis-Hastings. *Biometrika*, **88**, (4), 1035–53.

Green, P. J. and O'Hagan, A. (1998). Model choice with MCMC on product spaces without using pseudo-priors. Technical report, Department of Mathematics, University of Nottingham.

Green, P. J. and Richardson, S. (2001). Modelling heterogeneity with and without the Dirichlet process. *Scandinavian J. Statistics*, **28**, (2), 355–75.

Grenander, U. and Miller, M. I. (1994). Representations of knowledge in complex systems (with discussion). *Journal of the Royal Statistical Society, B*, **56**, 549–603.

Han, C. and Carlin, B. P. (2001). Markov chain Monte Carlo methods for computing Bayes factors: A comparative review. *J. Am. Statistical Association*, **96**, (455), 1122–32.

Hastie, D. I. (2005). *Towards automatic reverible jump Markov chain Monte Carlo*. PhD thesis, Department of Mathematics, University of Bristol.

Heikkinnen, J. (2003). *Trans-dimensional Bayesian nonparametrics with spatial point processes*, pp. 203–6. OUP, Oxford.

Jasra, A., Doucet, A., Stephens, D. A., and Holmes, C. C. (2008). Interacting sequential Monte Carlo samplers for trans-dimensional simulation. *Computational Statistics & Data Analysis*, **52**, (4), 1765–91.

Jasra, A., Stephens, D. A., and Holmes, C. C. (2007). Population-based reversible jump Markov chain Monte Carlo. *Biometrika*, **94**, (4), 787–807.

Keith, J. M., Kroese, D. P., and Bryant, D. (2004). A generalized Markov sampler. *Methodology Computing In Appl. Probability*, **6**, (1), 29–53.

Liang, F. M. and Wong, W. H. (2000). Evolutionary Monte Carlo: Applications to C-p model sampling and change point problem. *Statistica Sinica*, **10**, (2), 317–42.

Meng, X.-L. and Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, **6**, 831–60.

Newton, M. A. and Hastie, D. I. (2006). Assessing Poisson variation of intestinal tumour multiplicity in mice carrying a Robertsonian translocation. *J. Royal Statistical Soc. Series C-applied Statistics*, **55**, 123–38.

Petris, G. and Tardella, L. (2003). A geometric approach to transdimensional Markov chain Monte Carlo. *Canadian J. Statistics-revue Canadienne De Statistique*, **31**, (4), 469–82.

Preston, C. J. (1977). Spatial birth-and-death processes. *Bulletin of the International Statistical Institute*, **46**, (2), 371–91.

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society, B*, **59**, 731–92.

Ripley, B. D. (1977). Modelling spatial patterns (with discussion). *Journal of the Royal Statistical Society, B*, **39**, 172–212.

Roberts, G. O., Gelman, A., and Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk metropolis algorithms. *Annals of Applied Probability*, **7**, 110–20.

Roberts, G. O. and Rosenthal, J. S. (1998). Optimal scaling of discrete approximations to Langevin diffusions. *Journal of the Royal Statistical Society, B*, **60**, 255–68.

Roberts, G. O. and Rosenthal, J. S. (2006). Harris recurrence of Metropolis-within-Gibbs and trans-dimensional Markov chains. *Annals Appl. Probability*, **16**, (4), 2123–39.

Sisson, S. A. (2005). Transdimensional Markov chains: a decade of progress and future perspectives. *J. Am. Statistical Association*, **100**, (471), 1077–89.

Sisson, S. A. and Fan, Y. (2007). A distance-based diagnostic for trans-dimensional Markov chains. *Statistics and Computing*, **17**, (4), 357–67.

Spirling, A. (2007). "Turning points" in the Iraq conflict: Reversible jump Markov chain Monte Carlo in political science. *Am. Statistician*, **61**, (4), 315–20.

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

Tierney, L. and Mira, A. (1999). Some adaptive Monte Carlo methods for Bayesian inference. *Statistics in Medicine*, **18**, 2507–15.

Tjelmeland, H. and Hegstad, B. K. (2001). Mode jumping proposals in MCMC. *Scandinavian J. Statistics*, **28**, (1), 205–23.

Waagepetersen, R. and Sorensen, D. (2001). A tutorial on reversible jump MCMC with a view toward applications in QTL-mapping. *Int. Statistical Rev.*, **69**, (1), 49–61.