Reversible jump Markov chain Monte Carlo computation and Bayesian model determination

Peter J. Green
Department of Mathematics
University of Bristol
Bristol BS8 1TW, UK.

Summary

Markov chain Monte Carlo methods for Bayesian computation have until recently been restricted to problems where the joint distribution of all variables has a density with respect to some fixed standard underlying measure. They have therefore not been available for application to Bayesian model determination, where the dimensionality of the parameter vector is typically not fixed. This article proposes a new framework for the construction of reversible Markov chain samplers that jump between parameter subspaces of differing dimensionality, which is flexible and entirely constructive. It should therefore have wide applicability in model determination problems. The methodology is illustrated with applications to multiple change-point analysis in one and two dimensions, and to a Bayesian comparison of binomial experiments.

Some key words: Change-point analysis, Image segmentation, Jump diffusion, Markov chain Monte Carlo, Multiple binomial experiments, Multiple shrinkage, Step function, Voronoi tessellation.

1 Introduction

There are a number of challenging statistical problems, often involving inference about curves, surfaces or images, where the dimension of the object of inference is not fixed. One example discussed in detail later in this article concerns the multiple change-point problem for Poisson processes, where it is assumed that the rate is piecewise constant, but changes an unknown number of times. The times of change and the different rates are unknown. The object of inference is therefore a step function.

There are many problems of broadly similar vein, with the same general ingredients: a discrete choice between a set of models, a parameter vector with an interpretation depending on the model in question, and data, influenced by the model and parameter values, to be used as a basis for inference. Some examples are:

a. factorial experiments, with a prior allowing factor effects to tie;

b. variable selection in regression;

c. non-nested regression models;

d. mixture deconvolution, with an unknown number of components;

e. Bayesian choice between models with different numbers of parameters;

f. multiple change-point problems;

g. image segmentation, the two-dimensional analogue of the change-point problem;

h. object recognition, approached via marked spatial point processes.

Model criticism, model choice, model selection, model averaging, etc., all require the same basic computational tasks, and it is a technology for these tasks that is the focus here. The aim of this article is to add further weight to the assertions (a) that a Bayesian approach is attractive for such problems, and (b) that the computations for such inference can be handled by Markov chain Monte Carlo methods. In particular, in Section 3 we introduce a novel class of such methods capable of jumping between subspaces of differing dimensionality. This considerably extends the scope of Metropolis-Hastings methods, and applies to very many varying-dimension problems.

2 Bayesian model choice as a hierarchical model

Suppose that we have a countable collection of candidate models $\{\mathcal{M}_k, k \in \mathcal{K}\}$. Model $\mathcal{M}_k$ has a vector $\theta^{(k)}$ of unknown parameters, assumed to lie in $\mathcal{R}^{n_k}$, where the dimension n_k may vary from model to model. With obvious changes, our methods would apply to an arbitrary collection of parameter subspaces. We observe data y. There is a natural hierarchical structure expressed by modelling the joint distribution of $(k, \theta^{(k)}, y)$ as

$$p(k, \theta^{(k)}, y) = p(k)p(\theta^{(k)}|k)p(y|k, \theta^{(k)}),$$

that is, the product of model probability, prior and likelihood. It will be convenient to abbreviate the pair $(k, \theta^{(k)})$ by x. For given k, x lies in $\mathcal{C}_k = \{k\} \times \mathcal{R}^{n_k}$; generally, x varies over $\mathcal{C} = \cup_{k\in\mathcal{K}}\mathcal{C}_k$.

As a concrete example, consider a change-point problem in which there is an unknown number of change-points in a piecewise constant regression function on the interval $[0, L]$. For $k \in \mathcal{K} = \{0, 1, 2, \ldots\}$, model $\mathcal{M}_k$ says that there are exactly k change-points. To parametrise the resulting step function, we need to specify the position of each change-point, and the value of the function on each of the $(k+1)$ subintervals into which $[0, L]$ is divided. Thus $\theta^{(k)}$ is a vector of length $n_k = 2k + 1$.

Bayesian inference about k and $\theta^{(k)}$ will be based on the joint posterior $p(k, \theta^{(k)}|y)$, which is the target of the Markov chain Monte Carlo computations described below. It will often be appropriate to factorise this as

$$p(k, \theta^{(k)}|y) = p(k|y)p(\theta^{(k)}|k, y),$$

and to interpret the two terms separately, thus avoiding any "model averaging". Inference about the model indicator may sometimes be phrased in terms, not of $p(k|y)$, but of the Bayes factor for one model relative to another:

$$\frac{p(k_1|y)}{p(k_0|y)} \div \frac{p(k_1)}{p(k_0)},$$

which does not depend on the hyperprior $p(k)$. All these quantities are readily estimated from the Markov chain Monte Carlo sample obtained by the methods below; if Bayes factors are all that are required, $p(k)$ must nevertheless be specified to implement the computation, but it can be chosen on grounds of convenience. Note that regarding the posterior $p(k, \theta^{(k)})$ as the objective of the computation does not preclude model selection or prediction being ultimately based on a non-coherent principle such as that advocated by Madigan and Raftery (1994); thus the methods of the present paper would be applicable to their analysis.

Recent work on Markov chain Monte Carlo computation with application to aspects of Bayesian model determination includes Phillips and Smith (1995), based on the jump-diffusion samplers of Grenander and Miller (1994), Carlin and Chib (1995) who effectively work with the product space $\prod_{k \in \mathcal{K}} \mathcal{C}_k$, and unpublished work of Piccioni and Jona-Lasinio, who devise an embedding method in which the $\{\mathcal{C}_k\}$ are mapped onto subsets of a single parameter space. Each of these approaches has its merits and its disadvantages. In jump-diffusion, there is a conflict between minimising the distortion caused by using a positive time increment, and improving Monte Carlo efficiency. Further, although the jump-diffusion principle is really rather general, the range of jump transitions discussed by Grenander and Miller, and used by Phillips and Smith, is somewhat limited, amounting to conditional versions of Gibbs kernels, and Hastings kernels based on proposals generated from the prior. While these moves seem adequate for Grenander and Miller's applications, they are perhaps too restricted for general Bayesian computation. The product space approach of Carlin and Chib requires that irrelevant parameters, the $\theta^{(k')}$ for k' different from the current k, need to be continually updated, which apparently limits the approach to a small set of models $\mathcal{K}$. In recent unpublished work, A. O'Hagan and the author have pointed out that there is no need to update the irrelevant parameters to ensure the proper limiting distribution of the chain, but performance of the modified method is not very encouraging. The embedding method seems cumbersome and inexplicit in use.

3 Markov chain Monte Carlo using reversible jumps

3.1 Introduction

Let $\pi(dx)$ denote a *target distribution* of interest. In Bayesian inference, this is the posterior distribution for the parameters given the data, and in the present context of model determination, "parameters" include the indicator k for the model itself, as well as the parameter vector $\theta^{(k)}$ specific to that model. In Markov chain Monte Carlo computation, we construct a Markov transition kernel $P(x, dx')$ that is aperiodic and irreducible, and satisfies detailed balance:

$$\int_A \int_B \pi(dx)P(x, dx') = \int_B \int_A \pi(dx')P(x', dx), \tag{1}$$

for all appropriate A, B, and then simulate this chain to obtain a dependent, approximate, sample from $\pi(dx)$. Although detailed balance is more than is needed for ergodicity and the correct limiting distribution, in practical design of samplers it is a convenient restriction to impose.

In straightforward cases, $\pi(dx)$ is either a discrete probability distribution, or has a joint density with respect to some simple measure, usually Lebesgue; then methods for constructing suitable transition kernels are familiar. The two most popular methods are the Gibbs sampler (Geman and Geman, 1984), and the Metropolis-Hastings method (Metropolis, *et al.*, 1953; Hastings, 1970). A full description and some comparisons can be found in Tierney (1994), Besag *et al.* (1995), and elsewhere. Briefly, each method proceeds by sweeping around all the variables $x = (x_1, x_2, \ldots, x_n)$, visiting subsets of the indices in turn, either randomly or systematically. When a subset T of $\{1, 2, \ldots, n\}$ is visited, the variables $x_T := \{x_i : i \in T\}$ are updated. In the Gibbs sampler, the new values are drawn from the full conditional distributions $\pi(x_T | x_{-T})$, where $x_{-T} := \{x_i : i \notin T\}$. In the Hastings method, proposed new values x'_T for these variables are drawn from an essentially arbitrary distribution $q_T(x'_T; x)$. Then, with probability

$$\alpha(x, x') = \min\left\{1, \frac{\pi(x'_T | x_{-T})q_T(x_T; x')}{\pi(x_T | x_{-T})q_T(x'_T; x)}\right\}$$

the proposed values are accepted; otherwise, the existing values are retained.

The Gibbs sampler hardly even makes sense when x has a length that is not fixed, and elements which need not have a fixed interpretation across all models; to resample some components conditional on the remainder would rarely be meaningful. We therefore concentrate on adapting the wider class of Hastings algorithms to the present situation, following the approach outlined by Green (1994), in discussion of Grenander and Miller (1994). This gives a framework for dealing with the case where there is no simple underlying measure.

3.2 The general case

In a typical application with multiple parameter subspaces $\{\mathcal{C}_k\}$ of different dimensionality, it will be necessary to devise different types of move between the subspaces. These will be combined to form what Tierney (1994) calls a hybrid sampler, by random choice between available moves at each transition, in order to traverse freely across the combined parameter space $\mathcal{C}$. We restrict attention to Markov chains in which detailed balance is attained within each move type.

When the current state is x, we propose a move of type m, that would take the state to dx', with probability $q_m(x, dx')$. For the moment, this is an arbitrary sub-probability measure on m and x'. Thus $\sum_m q_m(x, \mathcal{C}) \leq 1$, and with probability $1 - \sum_m q_m(x, \mathcal{C})$, no change to the present state is proposed. Not all moves m will be available from all starting states x, so for each x, $q_m(x, \mathcal{C}) = 0$ for some, perhaps many, m.

As usual with Hastings algorithms, the proposal is not automatically accepted. The probability of acceptance will be denoted by $\alpha_m(x, x')$, and is left undefined at present; the objective of the following analysis is to derive an expression for $\alpha_m(x, x')$ which achieves the stated aim of attaining detailed balance within each move type.

The transition kernel we have defined can be written

$$P(x, B) = \sum_m \int_B q_m(x, dx')\alpha_m(x, x') + s(x)I[x \in B] \tag{2}$$

for Borel sets B in $\mathcal{C}$, where $I[.]$ denotes the indicator function, and

$$s(x) := \sum_m \int_{\mathcal{C}} q_m(x, dx')\{1 - \alpha_m(x, x')\} + 1 - \sum_m q_m(x, \mathcal{C})$$

is the probability of not moving from x, either through a proposed move being rejected, or because no move is attempted.

The detailed balance relation (1) requires the equilibrium probability of moving from A to B to equal that from B to A, for all Borel sets A, B in $\mathcal{C}$. Substituting (2), we need

$$\sum_m \int_A \pi(dx) \int_B q_m(x, dx')\alpha_m(x, x') + \int_{A \cap B} \pi(dx)s(x)$$
$$= \sum_m \int_B \pi(dx') \int_A q_m(x', dx)\alpha_m(x', x) + \int_{B \cap A} \pi(dx')s(x'). \tag{3}$$

For this to hold, it is sufficient that

$$\int_A \pi(dx) \int_B q_m(x, dx')\alpha_m(x, x') = \int_B \pi(dx') \int_A q_m(x', dx)\alpha_m(x', x)$$

for each m, A, B, and to achieve this we choose $\alpha_m(x, x')$ as follows.

Suppose that

$\pi(dx)q_m(x, dx')$ has a finite density $f_m(x, x')$ with respect to a symmetric measure ξ_m on $\mathcal{C} \times \mathcal{C}$. (*)

Then

$$\begin{aligned}\int_A \pi(dx) \int_B q_m(x, dx')\alpha_m(x, x') &= \int_A \int_B \xi_m(dx, dx')f_m(x, x')\alpha_m(x, x') \\ &= \int_B \int_A \xi_m(dx', dx)f_m(x', x)\alpha_m(x', x) \\ &= \int_B \pi(dx') \int_A q_m(x', dx)\alpha_m(x', x),\end{aligned}$$

as required, with the middle equality holding, by the assumed symmetry of ξ_m, provided that

$$\alpha_m(x, x')f_m(x, x') = \alpha_m(x', x)f_m(x', x). \tag{4}$$

As shown by Peskun(1973) with a proof only for the finite state space case, it is optimal, in the sense of reducing autocorrelation in the realised chain, to make the acceptance probability as large as possible subject to retaining detailed balance. Thus we take

$$\alpha_m(x, x') = \min\left\{1, \frac{f_m(x', x)}{f_m(x, x')}\right\} \tag{5}$$

which satisfies (4). The possibility that the denominator of the ratio above is zero is not of concern, since for such x, dx', there is zero probability of proposing such a move, by definition of f; the ratio can therefore safely be set to an arbitrary value. Less formally, but more transparently, we could write this expression using a ratio of measures

$$\alpha_m(x, x') = \min\left\{1, \frac{\pi(dx')q_m(x', dx)}{\pi(dx)q_m(x, dx')}\right\}. \tag{6}$$

For straightforward cases, the dimension-matching requirement can be imposed fairly simply, by following a standard "template". We give further details in the following subsection, but in the meantime add a few remarks.

Remark 1. The definition of the sampling method is entirely constructive. No integration, by simulation or otherwise, is needed to set up the transition mechanism.

Remark 2. The method allows great flexibility to the algorithm designer to exploit the structure of the problem at hand. Intuition can be used to choose moves that plausibly induce good mixing behaviour, while not imposing a heavy burden of algebraic and analytic work to establish validity.

Remark 3. Although as usual with Hastings methods, the distribution π need not be normalised, relative normalising constants between different subspaces are needed. Specifically, while it is not necessary that the prior distributions $p(\theta^{(k)}|k)$ are properly normalised, there must be only one unknown multiplicative constant among all such priors, unless only posteriors conditional on k are needed. Detailed balance between different subspaces could not be achieved otherwise, a point apparently missed by Grenander and Miller (1994).

Remark 4. Our general framework includes various familiar special cases. When there is only one parameter subspace, with a single dominating measure, it is just the random scan Hastings method. Our framework provides a natural generalisation of Hastings methods to general parameter spaces. In the case of point processes, the method is closely related to the spatial birth and death process studied by Preston (1977). Recently, Geyer and Møller (1994) have developed a Hastings sampler for point processes, which is a special case of our construction; they derive likelihood inference procedures for point patterns based on this, and prove results on convergence. The jump-diffusion processes of Grenander and Miller (1994), proposed for Bayesian computation in certain computer vision problems, also provide a special case of our method, but one in which within-parameter-subspace moves are made by a continuous-time diffusion process, which, when discretised temporally for computational purposes, only approximately maintains detailed balance. The range of jump transitions presented by Grenander and Miller is also somewhat restricted.

3.3 Switching between two simple subspaces

The rather obscure "dimension-matching" assumption (*) above deserves interpretation in more intuitive terms. Suppose first that there are just two subspaces $\mathcal{C}_1 = \{1\} \times \mathcal{R}$ and $\mathcal{C}_2 = \{2\} \times \mathcal{R}^2$, with π having proper densities on each subspace conditional on $k = 1$ and

2. The context might suggest, for example, that from a point $(2, \theta_1, \theta_2) \in \mathcal{C}_2$, a good move might be to $\{1, \frac{1}{2}(\theta_1 + \theta_2)\}$. For this move type, the equilibrium joint proposal probability

$$\int_B \pi(dx) \int_A q_m(x, dx'),$$

where $A \subset \mathcal{C}_1$ and $B \subset \mathcal{C}_2$, must have a density with respect to a singular measure on $\mathcal{R} \times \mathcal{R}^2$ placing all of its mass on $\{(\theta, \theta_1, \theta_2) : \theta = \frac{1}{2}(\theta_1 + \theta_2)\}$, instead of Lebesgue measure on $\mathcal{R}^3$. For detailed balance to be attainable, therefore, it is necessary that the reverse move from A to B should be defined via a proposal distribution $q_m(x, dx')$ that for each $x = (1, \theta)$ is singular, with all its probability on $\{(2, \theta_1, \theta_2) : \theta = \frac{1}{2}(\theta_1 + \theta_2)\}$. For example, we might draw a random variable u from some distribution, independently of the current state θ, and set $\theta_1 = \theta + u, \theta_2 = \theta - u$. All that (*) does is to ensure that singularities of the sort arising above are self-consistent.

To describe in detail how to implement the dimension-matching requirement in many standard cases, we consider a setup a little more general than the example just described. Suppose there are two subspaces, given by $k = 1$ and 2, and that $p(\theta^{(1)}|k = 1)$ and $p(\theta^{(2)}|k = 2)$ are proper densities in $\mathcal{R}^{n_1}$ and $\mathcal{R}^{n_2}$. Consider just one move type, which always switches subspaces, so that $q(x, \mathcal{C}_1) = 0$ for $x \in \mathcal{C}_1$, and $q(x, \mathcal{C}_2) = 0$ for $x \in \mathcal{C}_2$; the subscript m is being suppressed. The probability of choosing this move will be denoted by $j(x)$. A typical way of accomplishing a transition from $\mathcal{C}_1$ to $\mathcal{C}_2$ will be by generating a vector of continuous random variables $u^{(1)}$ of length m_1, independently of $\theta^{(1)}$, and then setting $\theta^{(2)}$ to be some deterministic function of $\theta^{(1)}$ and $u^{(1)}$. Similarly, to switch back, $u^{(2)}$ of length m_2 will be generated and $\theta^{(1)}$ set to some function of $\theta^{(2)}$ and $u^{(2)}$. For dimension-matching, there must be a bijection between $(\theta^{(1)}, u^{(1)})$ and $(\theta^{(2)}, u^{(2)})$. In particular, the lengths of $u^{(1)}$ and $u^{(2)}$ must satisfy $n_1 + m_1 = n_2 + m_2$. The proposal distribution $q(x, dx')$ can now be defined by the distributions of $u^{(1)}$ and $u^{(2)}$, which we suppose given by proper densities q_1 and q_2 with respect to Lebesgue measure in $\mathcal{R}^{m_1}$ and $\mathcal{R}^{m_2}$, respectively.

We can now be explicit about the condition (*) in this context. For $A \subset \mathcal{C}_1$ and $B \subset \mathcal{C}_2$, set

$$\xi(A \times B) = \xi(B \times A) = \lambda\{(\theta^{(1)}, u^{(1)}) : \theta^{(1)} \in A, \theta^{(2)}(\theta^{(1)}, u^{(1)}) \in B\}$$

where λ denotes $(n_1 + m_1)$-dimensional Lebesgue measure. For general $A, B \subset \mathcal{C}$, put

$$\xi(A \times B) = \xi\{(A \cap \mathcal{C}_1) \times (B \cap \mathcal{C}_2)\} + \xi\{(A \cap \mathcal{C}_2) \times (B \cap \mathcal{C}_1)\}.$$

This is symmetric, as required. Then for $x = (1, \theta^{(1)}) \in \mathcal{C}_1$ and $x' = (2, \theta^{(2)}) \in \mathcal{C}_2$, let

$$f(x, x') = p(1, \theta^{(1)}|y) j(1, \theta^{(1)}) q_1(u^{(1)}),$$

$$f(x', x) = p(2, \theta^{(2)}|y) j(2, \theta^{(2)}) q_2(u^{(2)}) \left| \frac{\partial(\theta^{(2)}, u^{(2)})}{\partial(\theta^{(1)}, u^{(1)})} \right|.$$

and otherwise set $f(x, x') = 0$. Then for all $x, x' \in \mathcal{C}$, $f(x, x')$ is the density with respect to ξ of the equilibrium joint proposal distribution $\pi(dx)q(x, dx')$.

According to (5), the appropriate acceptance probability for the proposed transition from $x = (1, \theta^{(1)})$ to $x' = (2, \theta^{(2)})$ is

$$\min \left\{ 1, \frac{p(2, \theta^{(2)}|y) j(2, \theta^{(2)}) q_2(u^{(2)})}{p(1, \theta^{(1)}|y) j(1, \theta^{(1)}) q_1(u^{(1)})} \left| \frac{\partial(\theta^{(2)}, u^{(2)})}{\partial(\theta^{(1)}, u^{(1)})} \right| \right\}, \tag{7}$$

which restores the anti-symmetry that was lost in the particular representation of ξ used above.

In practice, such moves will often be set up so that m_1 or m_2 is zero. In one direction, then, there is no need to generate the corresponding $u^{(i)}$, and the expression for the acceptance probability simplifies. For example, with $m_2 = 0$, it becomes

$$\min \left\{ 1, \frac{p(2, \theta^{(2)}|y) j(2, \theta^{(2)})}{p(1, \theta^{(1)}|y) j(1, \theta^{(1)}) q_1(u^{(1)})} \left| \frac{\partial(\theta^{(2)})}{\partial(\theta^{(1)}, u^{(1)})} \right| \right\}. \tag{8}$$

Finally, this example is somewhat simplified compared with many real applications, and appropriate modifications may need to be made. For example, $u^{(1)}$ may be generated dependently on $\theta^{(1)}$, in which case $q_1(u^{(1)})$ is replaced by the conditional density. If other discrete variables are generated in making the proposals, the probability functions of their realised values are multiplied into the move probabilities $j(x)$. With this latter change, (8) is used repeatedly in the applications later in this article.

4 Application to one-dimensional multiple change-point problems

4.1 Coal mining disasters

As our first application of the general construction of the previous section, we present a new Bayesian model for multiple change-point analysis, and develop a reversible jump Markov chain Monte Carlo sampler to compute the posterior distribution.

A data set that has been frequently used in illustrating new methods for change-point analysis is the point process of dates of serious coal-mining disasters between 1851 and 1962, given by Raftery and Akman (1986). In contrast to some other previous analyses of these data, we will work in continuous time, with the points recorded in days rather than years. Figure 1 displays the dates of the 192 disasters in these 112 years = 40907 days as a jittered dot plot, together with the cumulative counting process, shown as a dotted line. For data points $\{y_i, i = 1, 2, \ldots, n\} \in [0, L]$ from a Poisson process with rate given by the function $x(t)$, the log-likelihood is

$$\sum_{i=1}^{n} \log\{x(y_i)\} - \int_0^L x(t)dt. \tag{9}$$

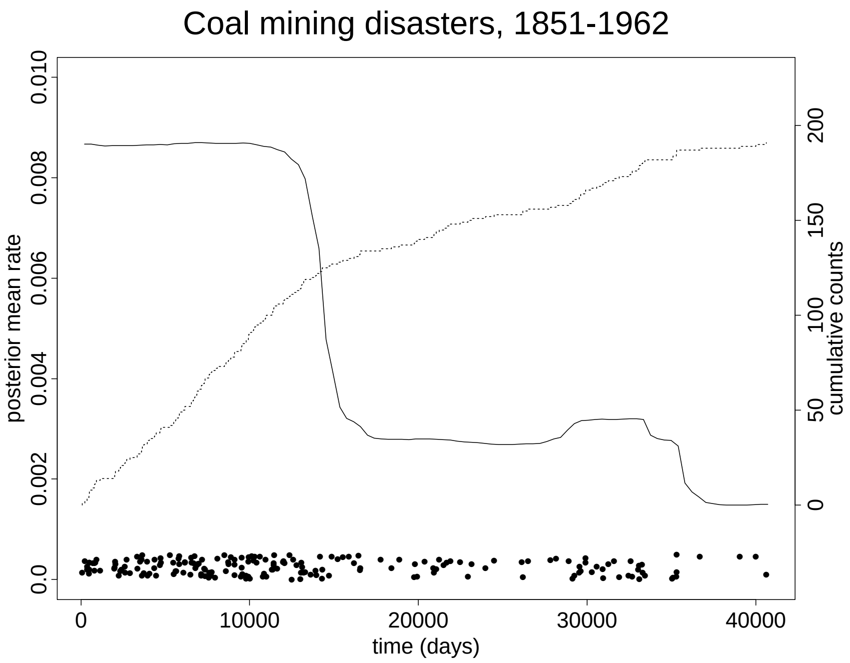

Figure 1: Coal mining disaster data: dates of disasters, cumulative counting process (dotted curve) and posterior mean rate of occurrence (solid curve)

4.2 A prior model for step functions

We develop a Bayesian multiple change-point analysis of point process data, by assuming that the rate function $x(.)$ on $[0, L]$ is a step function. In this section, we formulate a prior distribution for x.

Suppose that there are k steps, at positions $0 < s_1 < s_2 < \ldots < s_k < L$, and that the step function takes the value h_j, which we call its height, on the subinterval $[s_j, s_{j+1})$, $j = 0, 1, 2, \ldots, k$ (writing $s_0 = 0, s_{k+1} = L$ for convenience). The prior model is specified by supposing that k is drawn from the Poisson distribution

$$p(k) = e^{-\lambda} \frac{\lambda^k}{k!},$$

but conditioned on $k \leq k_{\max}$. Given k, the step positions $s_1, s_2, \ldots, s_k$ are distributed as the even-numbered order statistics from $2k+1$ points uniformly distributed on $[0, L]$, and the heights $h_0, h_1, \ldots, h_k$ are independently drawn from the $\Gamma(\alpha, \beta)$ density

$$\beta^\alpha h^{\alpha-1} e^{-\beta h} / \Gamma(\alpha) \quad \text{for} \quad h > 0.$$

This prior model for step functions is intended to be close to "uninformative". It is not appropriate to select an improper gamma distribution $\Gamma(0, 0)$ for the heights, because that causes insurmountable difficulties with normalisation across differing numbers of steps; all of the probability in the posterior would be assigned to the simplest model. It would perhaps have been more natural to take the step positions independently uniformly distributed on $[0, L]$ before sorting. However, this allows too many "short" steps, with $s_{j+1} - s_j$ small. Since there may be no data in the interval (s_j, s_{j+1}), such short intervals are barely penalised by the likelihood and so survive in the posterior, giving a more complicated picture of the true step function than is really justified by the data. The modification used here has the effect of probabilistically spacing out the step positions.

4.3 Using reversible jumps for step functions

In developing a reversible jump Monte Carlo sampler for the change-point problem, we are guided by intuition in designing appropriate moves, coupled with the requirements that the dimensions can be balanced properly, that the moves can be simulated conveniently, and that the acceptance ratio can be computed economically. As always with Hastings methods, there is flexibility in this process, and we are not constrained by fine details of the model in question. We make no claim of optimality for the particular choices made.

When the object x is a step function on $[0, L]$, some possible transitions are: (a) a change to the height of a randomly chosen step, (b) a change to the position of a randomly chosen step, (c) "birth" of a new step at a randomly chosen location in $[0, L]$, and (d) "death" of a randomly chosen step. Note that (c) and (d) involve changing the dimension of x, so that standard Markov chain Monte Carlo theory does not apply. In the general framework of Section 3 these transitions can be attained with a countable set of moves, which we denote by $\{H, P, 0, 1, 2, \ldots\}$. Here H means a height change, P a position change, and $m = 0, 1, 2, \ldots$ denotes the birth-death pair that increases the number of steps from m to $m+1$ steps, or reduces it from $m+1$ to m.

In some applications, the number of steps would be fixed in advance; often, change-point analysis assumes exactly one step. Nevertheless, there are clear advantages for efficient Monte Carlo computation in allowing k to vary, but to condition on k when drawing information from the realisation. This will allow much better mixing.

We now describe these transitions in more detail. At each transition, an independent random choice is made between attempting each of the at most four available move types $(H, P, k, k-1)$, signifying height change, position change, birth or death respectively. These have probabilities η_k for H, π_k for P, b_k for k and d_k for $k-1$, depending only on the current number of steps k, and satisfying $\eta_k + \pi_k + b_k + d_k = 1$. Naturally, $d_0 = \pi_0 = 0$, and $b_{k_{\max}} = 0$ to impose the preassigned upper limit $k_{\max}$ on the number of steps. Apart from these constraints, these probabilities were chosen so that $b_k = c \min\{1, p(k+1)/p(k)\}$ and $d_{k+1} = c \min\{1, p(k)/p(k+1)\}$ with the constant c as large as possible subject to $b_k + d_k \leq 0.9$ for all $k = 0, 1, \ldots, k_{\max}$. This choice ensures that $b_k p(k) = d_{k+1} p(k+1)$, which is the condition on b_k and d_k that would guarantee certain acceptance in the corresponding, but much simpler, Hastings sampler for the number of steps alone. Finally for $k \neq 0$, we took $\eta_k = \pi_k$.

If a move of type H or P is chosen, the remaining details are straightforward. A change to a height is attempted by first choosing one of $h_0, h_1, \ldots, h_k$ at random, obtaining h_j say, then proposing a change to h'_j such that $\log(h'_j/h_j)$ is uniformly distributed on the interval $[-\frac{1}{2}, +\frac{1}{2}]$; this choice is made from convenience, the proposal density ratio taking a simple form. The acceptance probability for this move is found to be

$$\min\left[1, (\text{likelihood ratio}) \times (h'_j/h_j)^{\alpha} \exp\{-\beta(h'_j - h_j)\}\right]$$

in the usual way. Here and later, "likelihood ratio" means $p(y|x')/p(y|x)$, where x and x'

stand for the current and proposed new values of all parameters. For a position change move, one of $s_1, s_2, \ldots, s_k$ is drawn at random, obtaining say s_j. The proposed replacement value is s'_j, drawn uniformly on $[s_{j-1}, s_{j+1}]$, and the acceptance probability turns out to be

$$\min\left\{1, (\text{likelihood ratio}) \times \frac{(s_{j+1}-s'_j)(s'_j-s_{j-1})}{(s_{j+1}-s_j)(s_j-s_{j-1})}\right\}.$$

The details for a birth of a step are more complicated, and follow the prescription in Section 3.3. We first choose a position s^* for the proposed new step, uniformly distributed on $[0, L]$. This must lie, with probability 1, within an existing interval (s_j, s_{j+1}), say. If accepted, s'_{j+1} will be set to s^*, and $s_{j+1}, s_{j+2}, \ldots, s_k$ will be relabelled as $s'_{j+2}, s'_{j+3}, \ldots, s'_{k+1}$, with corresponding changes to the labelling of step heights. We wish to propose new heights h'_j, h'_{j+1} for the step function on the subintervals (s_j, s^*) and (s^*, s_{j+1}) which recognise that the current height h_j on the union of these two intervals is typically well-supported in the posterior distribution, and should therefore not be completely discarded. Thus the new heights h'_j, h'_{j+1} should be perturbed in either direction from h_j in such a way that h_j is a compromise between them. To preserve positivity and maintain simplicity in the acceptance ratio calculations, we use a weighted geometric mean for this compromise, so that

$$(s^*-s_j)\log(h'_j) + (s_{j+1}-s^*)\log(h'_{j+1}) = (s_{j+1}-s_j)\log(h_j),$$

and define the perturbation to be such that

$$\frac{h'_{j+1}}{h'_j} = \frac{1-u}{u}$$

with u drawn uniformly from $[0, 1]$.

Following the analysis of Section 3.3, the acceptance probability for this proposal has to be calculated to achieve detailed balance with the corresponding death move, which we must therefore first specify. Dimension-matching is achieved by reversing the above calculation, so that if s_{j+1} is removed, the new height over the interval $(s'_j, s'_{j+1}) = (s_j, s_{j+2})$ is h'_j, the weighted geometric mean satisfying

$$(s_{j+1}-s_j)\log(h_j) + (s_{j+2}-s_{j+1})\log(h_{j+1}) = (s'_{j+1}-s'_j)\log(h'_j).$$

The s_{j+1} that is proposed for removal is simply drawn at random from $s_1, s_2, \ldots, s_k$.

The pair of birth and death moves thus defined satisfies the dimension-matching requirement. The birth increases the dimensionality from $2k+1$ to $2k+3$, the difference being accounted for by two continuous variables, the new position s^* and the u used to separate h'_j and h'_{j+1}.

In deriving an expression for the acceptance probability of the birth proposal, it is helpful to re-write (8) in the form

$$\min\{1, (\text{likelihood ratio}) \times (\text{prior ratio}) \times (\text{proposal ratio}) \times (\text{Jacobian})\},$$

noting that $p(x|y) = p(y|x)p(x)/p(y)$. In the present context, the likelihood ratio is straightforward, using (9); the prior ratio, which was previously $p(2, \theta^{(2)})/p(1, \theta^{(1)})$, becomes

$$\frac{p(k+1)}{p(k)} \frac{2(k+1)(2k+3)}{L^2} \frac{(s^* - s_j)(s_{j+1} - s^*)}{s_{j+1} - s_j}$$

$$\times \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{h'_j h'_{j+1}}{h_j}\right)^{\alpha-1} \exp\{-\beta(h'_j + h'_{j+1} - h_j)\};$$

the proposal ratio, which was $j(2, \theta^{(2)})/j(1, \theta^{(1)})q_1(u^{(1)})$, becomes

$$\frac{d_{k+1}L}{b_k(k+1)},$$

and the Jacobian is

$$\frac{(h'_j + h'_{j+1})^2}{h_j}.$$

The acceptance probability for the corresponding death step has the same form with the appropriate change of labelling of the variables, and the ratio terms inverted.

There have been at least two previous proposals for dealing with step functions with a variable number of steps by Markov chain Monte Carlo methods. Newton, Guttorp and Abkowitz (1992) build a model for a biological process using a hidden continuous-time Markov chain, and Arjas and Gasbarra (1994) develop a nonparametric approach to survival analysis assuming a step function form for the hazard rate. In both of these applications, the step function is not tied down at the right hand end of the observation interval, so that it can be encoded in a way that side-steps the varying dimensionality problem.

4.4 Analysis of the coal mining disaster data

Presentation of conclusions from Bayesian inference about any reasonably complicated object such as a function has to be partial. The displays given here should not be taken as examples

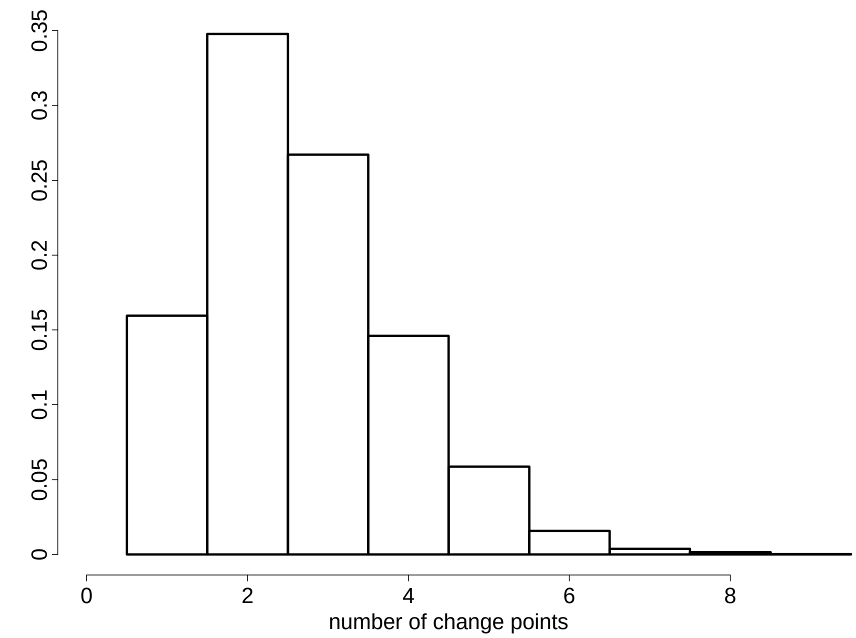

Figure 2: Coal mining disaster data: posterior distribution of k, the number of change-points

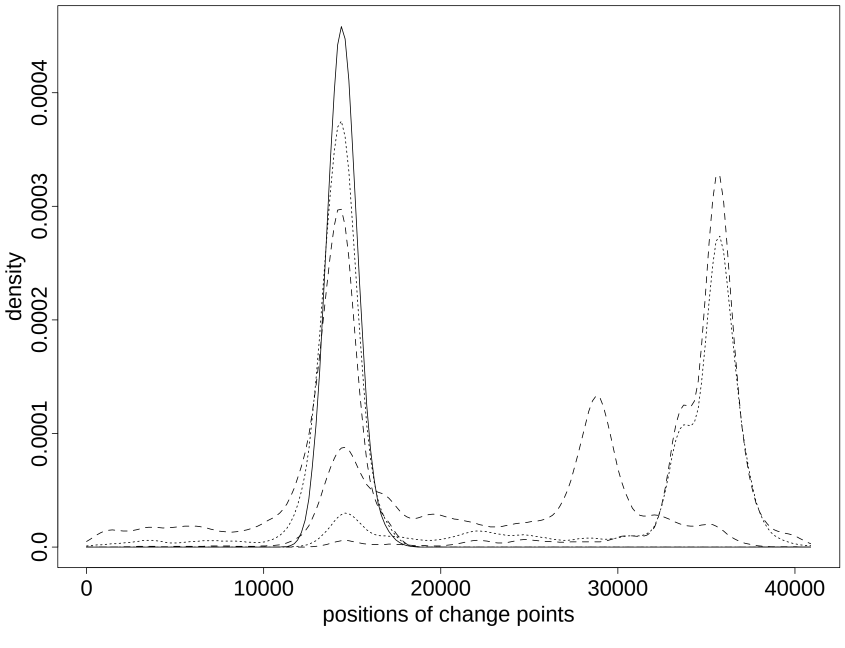

Figure 3: Coal mining disaster data: posterior density estimates of positions of change-points, conditional on number of change-points $k = 1$ (solid curve), $k = 2$ (dotted curves) and $k = 3$ (broken curves).

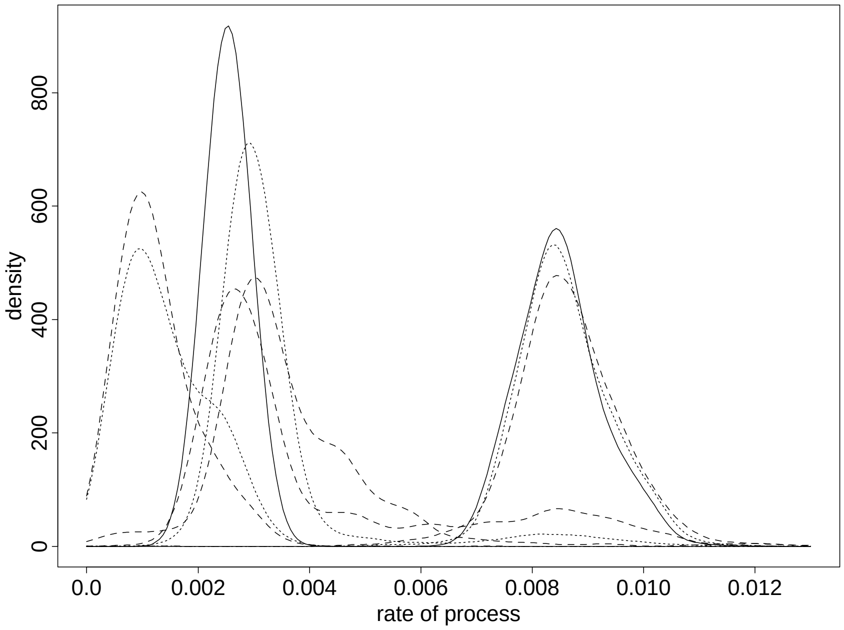

Figure 4: Coal mining disaster data: posterior density estimates of heights of segments of rate step function, conditional on number of change-points $k = 1$ (solid curves), $k = 2$ (dotted curves) and $k = 3$ (broken curves).

of the last word, either about this particular data set, or about how to present inference for step functions in general. Figures 1 to 4 show different aspects of one particular analysis, in which the hyperparameters are fixed as $\lambda = 3$, $k_{\max} = 30$, $\alpha = 1$ and $\beta = 200$. The Monte Carlo simulation was run for 40000 updates, after a burn-in period of 4000 updates. A pilot run established that one could have confidence that convergence had taken place by this point. The computation took 45 seconds on a Sun Sparc 2 workstation. In Figure 1, the solid curve shows the estimated posterior mean curve $E\{x(t)|y\}$, which is not a step function. Figure 2 shows the posterior distribution of k, the number of steps. In Figure 3, we show the posterior densities of the step positions, conditional on values $k = 1, 2$ and 3; the graphs become confusing to interpret with more than this many superimposed. The density estimates are obtained using a Gaussian kernel with standard deviation 625 days. Similarly, Figure 4 shows the corresponding conditional posterior density estimates of step height, using kernel standard deviation 0.0003 days^{-1}.

Some comparisons and contrasts with previous analyses of these data can be made. Raftery and Akman (1986) assume a single change-point, with location τ assumed a priori to be uniform on the interval $[0, L]$. The step heights are drawn independently from the improper Gamma distribution $\Gamma(\frac{1}{2}, 0)$. They use the point process likelihood, and calculate the posterior density of τ and of the relative change in step height, and the Bayes factor for comparing the hypothesis of a change versus no change, all by numerical integration. The Bayes factor turns out to be over 10^{13}, overwhelming evidence for a change. The posterior mode of the time of change is 10 March 1890 = day 14313 and a 95% credible interval is [15 May 1887, 3 August 1895] = [13283,16285] in days, which compare with a mode of 25 June 1890 = day 14420 and an interval of [24 May 1887, 7 May 1896] = [13292,16563] for our analysis, conditional on $k = 1$. Raftery and Akman also give a substantive interpretation of their inference in the context of the historical circumstances underlying the data. Carlin, Gelfand and Smith (1992) develop a hierarchical Bayesian approach for the single change-point problem for regression. They apply this to Poisson process data such as the coal mining disaster data by discretising into counts in annual intervals. The position of change is taken as a discrete variable; the step heights are drawn independently from the Gamma distribution $\Gamma(\alpha, \beta)$ in our notation, with $\alpha = 0.5$ and β^{-1} drawn from the third stage prior $\Gamma(0, 1)$. They produce posterior densities of step heights and of the position of change, all based on Gibbs sampling. The posterior modal year for change is 1891. Barry and Hartigan (1992, 1993) analyse change-point problems using product-partition models; again Markov chain Monte Carlo methods are used, but the change-points are coded discretely, so that they can be handled using a fixed set of indicator variables. Stephens (1994) and Phillips and Smith (1995) develop Bayesian analyses for the multiple change-point regression problem, with the positions of change taken as discrete variables, and computations performed by Gibbs sampling and jump-diffusion sampling respectively; however, they do not adapt these methodologies for the point process problem. None of these approaches treats the multiple change-point problem in genuinely continuous time, as does our proposed methodology. We see no difficulty with introducing a hierarchical structure into our modelling, if desired.

5 Image segmentation via Voronoi Tessellation

There are various two-dimensional analogues of change-point analysis. The problem discussed briefly in this section is intended to give an idea of one possibility.

Image segmentation is the process of subdividing a digital image into homogeneous regions, generally as a prelude to further analysis; see Sonka, Hlavac and Boyle (1993). What should be regarded as "homogeneous" depends on context; often, for example, it involves texture more than intensity. However, here we consider only the simplest version of the problem, in which we wish to subdivide a noisy image, i.e. observations arranged on a regular rectangular grid, into regions of homogeneous mean intensity. With additive noise, occurring independently and without blur at each pixel, it is natural to specify a regression model with a piecewise constant mean function, a form of two-dimensional step function.

For computational tractability, we consider here only step functions of this form in which the regions of constancy are polygonal, and we are thus concerned with a polygonal tessellation of that part of the plane that is within the field of view. For a flexible and convenient tessellation, we use the Voronoi, or Dirichlet, tessellation, in which each individual polygon, or tile, is defined to be that region of the plane nearer to that tile's generating point than to any other. The tessellation is thus specified by the coordinates $(u_i, v_i), i = 1, 2, \ldots, k$ of the k generating points, and the entire step function by these points and the heights h_i of the function within the i^{th} tile. The step function x therefore satisfies

$$x(u,v) = h_i \quad \text{where} \quad i = \text{argmin}\{(u-u_i)^2 + (v-v_i)^2\}.$$

For a general discussion of the Voronoi tessellation, and an algorithm for its computation, see Green and Sibson (1978). The basic algorithm described there, and its subsequent development in the TILE4 package by Sibson and co-workers at the University of Bath are ideally suited to the birth-death Markov chain Monte Carlo simulation methodology used in Section 4 for the one-dimensional change-point problem, appropriately modified.

In our general notation, the candidate models are indexed by $k \in \mathcal{K} = 1, 2, \ldots$, and the parameter vector for model k is $\theta^{(k)} = (u_i, v_i, h_i)_{i=1}^{k}$, with dimension $n_k = 3k$. The likelihood assumed here will be that based on independent Gaussian noise:

$$p(y|k, \theta^{(k)}) \propto \exp[-\frac{1}{2\sigma^2} \sum \{y(u,v) - x(u,v)\}^2],$$

where $y(u,v)$ denotes the observed intensity in the pixel centred at (u,v), and the sum is over all pixels.

The prior model used in the illustration below is again an uninformative one. The number of tiles k is modelled to have a Poisson distribution with parameter λ, truncated to $k = 1, 2, \ldots, k_{\max}$. Given k, the locations (u_i, v_i) of the generating points are independently and uniformly distributed over the unit square representing the field of view, and the heights h_i are drawn independently from the $\Gamma(\alpha, \beta)$ distribution.

The move types used in this problem correspond closely to H, and $m = 0, 1, 2, \ldots$ of Section 4.3; it is not computationally convenient to perform the analogue of P, to move a generating point. However, the TILE4 package includes routines for adding and deleting generating points, corresponding to birth and death of a step, and changing the height h_i in one tile under detailed balance is entirely straightforward. To explain the birth and death transitions in more detail, some further notation is needed. Let the probabilities of proposing a birth or death when the current number of steps, viz. tiles, is k be b_k,d_k respectively. Consider a proposed birth which would increase the number of steps from k to $k+1$, and suppose that the new generating point is labelled k^*. Its location (u_{k^*}, v_{k^*}) is drawn uniformly from the unit square, and the tessellation modified by the addition of this point; this modification is done on a trial basis, as this birth may not be accepted. In the updated tessellation the new point has "neighbours" (Green and Sibson, 1978), which we label as $i \in \mathcal{I}$. We compute the old and new areas of these tiles, and denote them by $s_i + t_i$ and t_i respectively. The total reduction $\sum_{i\in\mathcal{I}} s_i$ gives the area of the tile of the new point k^*. The height assigned to the new point is given by $h^* = \tilde{h}v$, where $\tilde{h}$ is the weighted geometric mean of the original heights for the neighbouring tiles:

$$\tilde{h} = \left(\prod_{i\in\mathcal{I}} h_i^{s_i}\right)^{1/\sum_i s_i};$$

and v is drawn independently with density function $f(v) = 5v^4/(1+v^5)^2$, so that $\log(v)$ has a distribution symmetric about 0. Finally, the new heights for those tiles modified by the addition are given by

$$h_i' = \left\{h_i^{s_i+t_i}(h^*)^{-s_i}\right\}^{1/t_i}.$$

The motivation for making these particular assignments is that the integral of $\log(h)$ over the whole unit square is thereby left unchanged, while the height assigned to the new tile is a compromise between the heights previously assigned to points in that tile, modified by a small multiplicative perturbation. For the death transition corresponding to this birth, a randomly chosen generating point is deleted, and the points in its tile re-assigned to neighbours. Using t_i and $s_i + t_i$ to denote the old and new areas for neighbouring tile i, its height is changed to

$$\left\{h_i^{t_i}(h^*)^{s_i}\right\}^{1/(s_i+t_i)},$$

which has the effect of reversing the birth move exactly.

With this pair of proposal mechanisms, it turns out after some straightforward algebra that the acceptance ratio for the birth is $\min(1, R)$, and for the death $\min(1, R^{-1})$ where

$$R = (\text{likelihood ratio}) \times \lambda \frac{\beta^\alpha}{\Gamma(\alpha)}(h^*)^{\alpha-1} \prod_{i\in\mathcal{I}} \left(\frac{h_i'}{h_i}\right)^{\alpha-1} \exp[-\beta\{h^* + \sum_{i\in\mathcal{I}}(h_i' - h_i)\}]$$
$$\times \frac{d_{k+1}}{b_k(k+1)f(v)} \times \tilde{h} \prod_{i\in\mathcal{I}} \left\{\frac{(s_i+t_i)h_i'}{t_i h_i}\right\},$$

using (8).

Figure 5 displays results from one simple example testing this methodology, based on synthetic data. A "true" image consisting of a disc of intensity 2.0 against a background of a lower intensity 0.5 was degraded with additive Gaussian noise, independently at each pixel

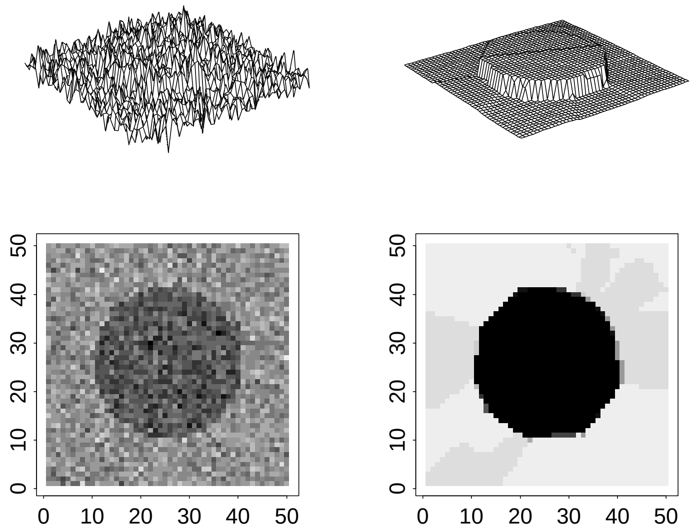

Figure 5: Synthetic segmentation problem: on the left, noisy data; on the right, estimated posterior mean. Upper plots show perspective views of the same surfaces displayed as images below.

on a 50×50 grid, with standard deviation $\sigma = 0.7$. Note that a disc cannot be perfectly fitted by a finite union of Voronoi polygons. The hyperparameters in the prior were fixed at $\lambda = 15$, $k_{\max} = 30$, $\alpha = 1.0$ and $\beta = 1.0$. The Figure shows, on the left, the data $y(u, v)$ and, on the right, the posterior mean surface $E\{x(u, v)|y\}$, estimated from a run of the sampling method described above, using 20000 sweeps after a burn-in period of 4000 sweeps.

Notwithstanding the apparent complexity of the geometrical calculations to maintain the tessellation and its modifications, and of the computations described in the paragraphs above, the entire sampler runs quite quickly. On a Sun Sparc 2 workstation, the run described above takes approximately 260 seconds.

6 Partition models

6.1 A hierarchical model for binomial probabilities

Several of the contexts listed in the introduction, viz. factorial experiments, variable selection in regression and mixture deconvolution, have the common feature that the discrete model-choice problem is equivalent to determining a partition, either of the original data units or of some other labels applying to the data, for example factor levels. Here we describe a general partition sampler, and its application to an ANOVA-like problem for binomial data discussed by Consonni and Veronese (1995).

A partition of a set $I = \{1, 2, \ldots, n\}$ is a collection $g = \{S_1, S_2, \ldots, S_d\}$ of subsets of I, which we call groups, where the S_j are disjoint with union I. The number d of groups into which I is divided by g will be called the degree of g, and written $d(g)$. To emphasise dependence on g, we also write $S_j(g)$, etc.

Suppose we have n responses $y_1, y_2, \ldots, y_n$, assumed drawn independently from binomial distributions: $y_i \sim \mathrm{Bin}(w_i, \theta_i)$, where the index parameters $\{w_i\}$ are known, and the probabilities $\{\theta_i\}$ unknown. We construct a prior distribution for $\{\theta_i\}$ that acknowledges that these parameters may have similar values within groups defined by a partition g of $I = \{1, 2, \ldots, n\}$. Within each group $S_j(g)$, the θ_i are drawn independently from Beta distributions:

$$\theta_i \sim \mathrm{Beta}\{q\alpha_j, q(1-\alpha_j)\} \quad \text{for} \quad i \in S_j(g), \quad j = 1, 2, \ldots, d(g).$$

The group mean parameters $\{\alpha_j\}$ are in turn drawn independently from the uniform distribution U(0,1), while the group precision parameter q is either fixed at a known value, or drawn from a hyperdensity $p(q)$. This is essentially the model proposed by Consonni and Veronese, except that they took a more general Beta distribution than U(0,1) for the α_j, and allowed separate q_j in each group, but took these to be fixed constants only. It would be routine to modify what follows to deal with this situation. Consonni and Veronese used conventional numerical techniques to fit their model, and so were constrained to use conjugate distributions, for which these techniques were practicable. With reversible jump Markov chain Monte Carlo computation, such constraints need not have been imposed.

Following Consonni and Veronese, the prior distribution for g is taken as

$$p(g) \propto \frac{d(g)^{-1}}{\#\{g' : d(g') = d(g)\}},$$

giving equal probability to all partitions of the same degree, and placing probability $\propto d^{-1}$ on the set of g with degree d. Calculation with this prior is straightforward. It is necessary to count the number of partitions of degree d of a set of n items: this count $c(n, d)$ is the solution of the recurrence relation $c(n, d) = dc(n-1, d) + c(n-1, d-1)$. Such counts become very large with n, and some care is needed to avoid overflow. An alternative model for the partitions that could have been used is Hartigan's product-partition model (see Barry and Hartigan, 1992); for given $d(g)$, this favours a more unequal distribution of the items into groups.

The joint distribution of all variables is now determined as

$$\begin{aligned}
p(g, \alpha, q, \theta, y) &= p(g)p(\alpha, q|g)p(\theta|g, \alpha, q)p(y|g, \alpha, q, \theta) \\
&= p(g)p(\alpha|g)p(q)p(\theta|g, \alpha, q)p(y|\theta) \\
&= p(g) \times \prod_{j=1}^{d(g)} 1 \times p(q) \times \prod_{j=1}^{d(g)} \left[\prod_{i \in S_j(g)} \frac{\theta_i^{q\alpha_j - 1}(1-\theta_i)^{q(1-\alpha_j)-1}}{B\{q\alpha_j, q(1-\alpha_j)\}} \right] \\
&\quad \times \prod_{i=1}^{n} \binom{w_i}{y_i} \theta_i^{y_i}(1-\theta_i)^{w_i - y_i},
\end{aligned}$$

where $B(.,.)$ is the Beta function. In the general notation of Section 2, the model indicator k is g, while the parameter vector $\theta^{(k)}$ is $(\alpha_1, \ldots, \alpha_{d(g)}, q, \theta_1, \ldots, \theta_n)$, of dimension $n_g = n + d(g) + 1$.

6.2 Reversible jump Markov chain Monte Carlo for partition problems

Much of the following discussion would apply, with few changes, to other partition problems. First we deal with updating the elements of $\theta^{(k)}$. The full conditionals for $\theta_i, i = 1, 2, \ldots, n$ are independent Beta distributions

$$\theta_i | \cdots \sim \text{Beta}\{q\alpha_j + y_i, q(1-\alpha_j) + w_i - y_i)\} \quad \text{for} \quad i \in S_j(g),$$

where, here and below, we use "$\cdots$" to denote all other variables among $\{g, \alpha_1, \ldots, \alpha_{d(g)}, q,$ $\theta_1, \ldots, \theta_n\}$. Therefore each θ_i can be updated with a Gibbs kernel. For q, we find

$$p(q|\cdots) \propto p(q) \times \prod_{j=1}^{d(g)} \left\{ \prod_{i \in S_j(g)} \theta_i^{q\alpha_j - 1}(1-\theta_i)^{q(1-\alpha_j)-1} \right\},$$

which is not a standard distribution but is easily evaluated, and so we use it in a Hastings step, with a proposal that, on the log scale, is uniformly distributed about the current value. The group mean parameters are also conditionally independent:

$$p(\alpha_j|\cdots) \propto \frac{\prod_{i \in S_j(g)} \theta_i^{q\alpha_j - 1}(1-\theta_i)^{q(1-\alpha_j)-1}}{B\{q\alpha_j, q(1-\alpha_j)\}^{\#S_j(g)}}.$$

Application of Stirling's formula shows that this full conditional has a normal approximation, for large q:

$$\alpha_j|\cdots \approx \mathrm{N}\left\{\mu, \frac{\mu(1-\mu)}{q\#S_j(g)}\right\},$$

where μ is such that $\mu/(1-\mu)$ is the geometric mean of $\theta_i/(1-\theta_i)$, $i \in S_j(g)$. This approximation could have been used explicitly in an approximate Gibbs sampler, but we choose to use it as a proposal distribution for a Hastings step.

Turning now to the step updating the partition g to g', say, we note that with the prior $p(g)$ specified above all partitions have positive probability, and so a process that jumps between partitions making only the modest changes of splitting a group, a "birth", and combining two groups, a "death", will be irreducible. It would have been quite natural to have included a move that changed the partition by reallocation of items while fixing the number of groups, but that was not implemented here. We have found the following mechanisms for the partition moves effective in practice, applied to partitions of up to a few dozen objects.

For a birth, which is attempted with probability b_g when the current partition is g, we first choose a group to split, uniformly among those with at least two items. This group is then split at random "binomially", i.e. each item is assigned to one of the two daughter subgroups independently, with probability one-half for each, but conditional on neither subgroup being empty. For a death, attempted with probability d_g, we simply choose two groups at random to be combined into one.

Jumping to a new partition necessitates a change also to the vector α, since its length has to increase or decrease by 1. Our proposal for the additional component is gaussian on a logit scale, and takes account of the numbers of binary responses influenced by each of the relevant α_j. Specifically, suppose that a proposed birth splits S_j into subgroups S_{j1} and S_{j2}. Let α_j be the current value, and α_{j1}, α_{j2} the new values for the two subgroups. Then we set

$$\alpha_{j1} = \frac{\alpha_j e^{\sigma z/W_1}}{1-\alpha_j+\alpha_j e^{\sigma z/W_1}} \quad \text{and} \quad \alpha_{j2} = \frac{\alpha_j e^{-\sigma z/W_2}}{1-\alpha_j+\alpha_j e^{-\sigma z/W_2}}$$

where $W_r = \sum_{i\in S_{jr}} w_i$, $r=1,2$, z is an independent standard gaussian random variable, and σ is a spread parameter to be chosen later. For the corresponding death move, α_{j1} and α_{j2} are merged to form the α_j that solves these simultaneous equations.

This completes the specification of the jump proposal; its acceptance probability is necessarily somewhat complicated in form, but is calculated as usual from (8). For the birth and death, the probabilities are respectively $\min(1,R)$ and $\min(1,R^{-1})$, where

$$\begin{aligned}
R \;&=\; \frac{B\{q\alpha_j, q(1-\alpha_j)\}^{\#S_j}}{B\{q\alpha_{j1}, q(1-\alpha_{j1})\}^{\#S_{j1}} B\{q\alpha_{j2}, q(1-\alpha_{j2})\}^{\#S_{j2}}} \\
&\times \prod_{i\in S_{j1}} \left(\frac{\theta_i}{1-\theta_i}\right)^{q(\alpha_{j1}-\alpha_j)} \prod_{i\in S_{j2}} \left(\frac{\theta_i}{1-\theta_i}\right)^{q(\alpha_{j2}-\alpha_j)} \times \frac{p(g')}{p(g)} \\
&\times \frac{d_{g'}}{b_g} \#\{j : \#S_j(g)\geq 2\} \frac{2}{d(g)\{d(g)+1\}} (2^{\#S_j-1}-1) \\
&\times \frac{\alpha_{j1}(1-\alpha_{j1})\alpha_{j2}(1-\alpha_{j2})}{\alpha_j(1-\alpha_j)} \sigma(W_1^{-1}+W_2^{-1}) \div (2\pi)^{-\frac{1}{2}} \exp(-\tfrac{1}{2}z^2).
\end{aligned}$$

6.3 Application to pine seedling mortality data

We apply the methodology described above to a small data set, one of those analysed by Consonni and Veronese (1995). This concerns 4 binomial responses $y = (59, 89, 88, 95)$, each based on $w_i = 100$ trials. The data arise from a 2×2 factorial experiment, comparing two treatments (H: planting too high; D: planting too deep) on two varieties of pine seedling (L: longleaf; S: slash). The responses are indexed in the order (LH, LD, SH, SD). Consonni and Veronese compare various statistical methods for analysing these data, including a Bayesian method based on their model described above, which has an "adaptive multiple shrinkage" property; see also George (1986). The data determine a partition of the 4 responses into groups that are similar, and estimates of probabilities θ_i within such a group S_j borrow strength by shrinking towards a common value α_j. Alternative estimators considered include the maximum likelihood estimators for both a saturated model and for an additive

Table 1: Mortality of pine seedlings: posterior means and standard deviations, in parentheses, of $\{\theta_i\}$

		Consonni & Veronese			Reversible jump MCMC			
Experiment	y_i	$q = 100$	$q = 200$	$q = 300$	$q = 100$	$q = 200$	$q = 300$	random q
LH	59	0.589	0.588	0.588	0.587	0.585	0.586	0.588
		(0.059)	(0.056)	(0.054)	(0.049)	(0.050)	(0.047)	(0.049)
LD	89	0.893	0.894	0.895	0.892	0.893	0.894	0.893
		(0.031)	(0.028)	(0.027)	(0.027)	(0.026)	(0.025)	(0.026)
SH	88	0.886	0.889	0.891	0.886	0.890	0.890	0.888
		(0.032)	(0.029)	(0.028)	(0.029)	(0.027)	(0.026)	(0.026)
SD	95	0.929	0.924	0.922	0.930	0.926	0.921	0.926
		(0.027)	(0.026)	(0.026)	(0.023)	(0.025)	(0.025)	(0.024)

logistic regression, a parametric empirical Bayes estimator which shrinks all θ_i together, and a nonparametric empirical Bayes estimator, which again has the multiple shrinkage property.

We refer the reader to Consonni and Veronese for further background, including discussion of some of the philosophical issues that arise in the modelling.

Our analysis has been confined to repeating that of Consonni and Veronese, but obtained using reversible jump Markov chain Monte Carlo instead of their analytic approximations. We extended their results very slightly by allowing q to be random, as well as fixed at each of the values they use (100, 200 and 300). This adaptation made use of a prior $p(q)$ under which $\log q$ is uniform on the interval $[\log 100, \log 300]$; the proposal for updating q described in the previous section was interpreted as wrapped periodically onto this interval. There were no other unspecified hyperparameters in the model defined above.

The samplers were also completely specified above, except for the scale factor σ, which we took as 50, after a little experimentation, and the probabilities assigned to each move type. We took the birth and death rates b_g and d_g each to be 0.3 for all g, except for the extreme partitions where $d(g) = 1$ or $n(= 4)$, where b_g and d_g were taken as $(0.6, 0)$ and $(0, 0.6)$. At each transition, θ was updated with probability 0.2, and similarly for the pair (α, q).

Results are presented in Table 1, based on run lengths of 40000 attempted updates, after

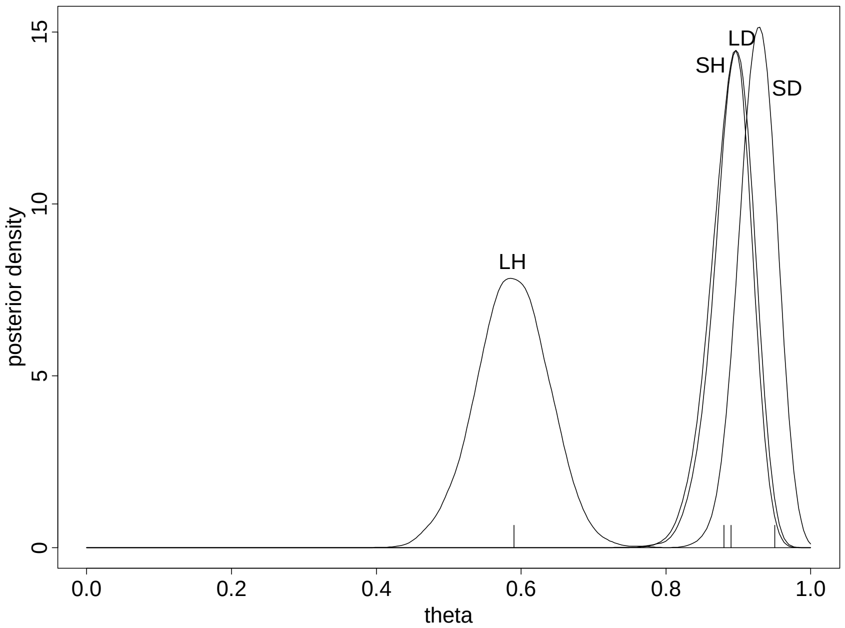

Figure 6: Posterior density estimates of $\{\theta_i\}$ for the pine seedling mortality data, together with raw data plotted as tick marks

burn-in periods of 4000; such runs took 36 seconds on a Sun Sparc 2. Posterior expectations of $\{\theta_i\}$ are close to those obtained by Consonni and Veronese. For the case where q was taken as random, with the hyperprior specified above, its posterior mean and standard deviation were estimated as 181 and 58. The sampling-based computation permits other information to be extracted and displayed. In Figure 6, we show posterior density estimates for the $\{\theta_i\}$, under the random q version of the model, together with the raw data, plotted with tick marks at the points y_i/w_i. The adaptive multiple shrinkage is evident here: note that the estimates for factor combinations (LD, SH, SD) are shrunk together, and correspondingly have smaller posterior variance. The data suggest that treatment H increases mortality, but only on seedlings of type L: a more subtle conclusion than from the logistic regression analysis, which simply concludes that both treatment and variety factors have significant effects.

7 Discussion

The theory and applications presented in this article have demonstrated that the advantages of Markov chain Monte Carlo computation can be extended to new classes of problems, where the object of inference has a dimension that is not fixed, including difficult Bayesian model-determination problems.

We have presented three applications of a new Markov chain Monte Carlo methodology; other implementations have also been developed. For example, jointly with Dr S. Richardson, the author is investigating Bayesian mixture estimation, with an unknown number of components, Ph. D. students at Bristol are applying the methods to various image analysis problems, and in his Ph. D. thesis at Cambridge University Dr R. Morris has developed a new method of removal of scratches from movie film.

There remain a number of questions about the methodology, to be resolved in future work. One concerns the development of understanding about moves that are likely to be effective generically, to aid intuition about the design of moves. Secondly, in situations where the collection of candidate models is restricted by practical or statistical considerations, there is the question of whether inventing additional models and corresponding parameter subspaces may facilitate mixing, and if so, how to do it effectively. In problems involving partitions of larger sets of items than those arising in Section 6, we need new jump proposal mechanisms. The proposals used in the pine seedling mortality study were completely "blind" in that they made no reference to the current values of any of the other variables. It might be anticipated that taking account of the $\{\alpha_j\}$ would allow the construction of much more efficient proposals, and indeed this is borne out in our recent experience with mixture estimation. Finally, the complications of multiple parameter subspaces of differing dimensionality make the problems of assessing convergence yet more difficult, and there is an urgent need for research on effective diagnostics of broad applicability.

Acknowledgements

I wish particularly to thank Sylvia Richardson for stimulating discussions about this work, and for making many valuable suggestions. I also acknowledge comments, connections, corrections and correspondence from Julian Besag, Andrew Gelman, Charlie Geyer, Paolo Giudici, Vincent Granville, Andrew Lawson, Jesper Møller, Tony O'Hagan, Marco Pievatolo, Renata Rotondi, David Stephens, Mike Titterington, and the referee and associate editor.

References

Arjas, E. and Gasbarra, D. (1994) Nonparametric Bayesian inference from right censored survival data using the Gibbs sampler. *Statistica Sinica*, **4**, 505–524.

Barry, D. and Hartigan, J. (1992) Product partition models for change point problems. *Ann. Statist.*, **20**, 260–279.

Barry, D. and Hartigan, J. (1993) A Bayesian analysis of change point problems. *J. Amer. Statist. Assoc.*, **88**, 309–319.

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

Carlin, B. P. and Chib, S. (1995) Bayesian model choice via Markov chain Monte Carlo. *J. Roy. Statist. Soc.*, B, **57**, 473–484.

Carlin, B. P., Gelfand, A. E. and Smith, A. F. M. (1992) Hierarchical Bayesian analysis of changepoint problems. *App. Statist.*, **41**, 389–405.

Consonni, G. and Veronese, P. (1995) A Bayesian method for combining results from several binomial experiments. To appear in *J. Amer. Statist. Assoc.*, **90**.

George, E. I. (1986) Combining minimax shrinkage estimators. *J. Amer. Statist. Assoc.*, **81**, 437–445.

Geyer, C. J. and Møller, J. (1994) Simulation procedures and likelihood inference for spatial point processes. *Scand. J. Statist.*, **21**, 359–373.

Green, P. J. (1994) Contribution to the discussion of paper by Grenander and Miller (1994). *J. Roy. Statist. Soc.*, B, **56**, 589–590.

Green, P. J. and Sibson, R. (1978) Computing Dirichlet tessellations in the plane. *Computer J.*, **21**, 168–173.

Grenander, U. and Miller, M. (1994) Representations of knowledge in complex systems (with Discussion). *J. Roy. Statist. Soc.*, B, **56**, 549–603.

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

Madigan, D. and Raftery, A. E. (1994) Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. Amer. Statist. Assoc.*, **89**, 1335–1346.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953) Equations of state calculations by fast computing machines. *J. Chem. Phys.*, **21**, 1087–1091.

Newton, M. A., Guttorp, P. and Abkowitz, J. L. (1992) Bayesian inference by simulation in a stochastic model from hematology. *Computing Science and Statistics*, **24**, 449–455. H. J. Newton, editor. Interface Foundation of North America.

Peskun, P. H. (1973) Optimum Monte-Carlo sampling using Markov chains. *Biometrika*, **60**, 607–612.

Phillips, D. B. and Smith, A. F. M. (1995) Bayesian model comparison via jump diffusions. In *Markov chain Monte Carlo in practice* (Ed. W. R. Gilks, S. T. Richardson and D. J. Spiegelhalter), chapter 13, London: Chapman and Hall.

Preston, C. J. (1977) Spatial birth-and-death processes. *Bull. Int. Statist. Inst.* (2) **46**, 371–391.

Raftery, A. E. and Akman, V. E. (1986) Bayesian analysis of a Poisson process with a change point. *Biometrika*, **73**, 85–89.

Sonka, M., Hlavac, V. and Boyle, R. (1993) *Image processing, analysis and machine vision.* London: Chapman and Hall.

Stephens, D. A. (1994) Bayesian retrospective multiple-changepoint identification. *Appl. Statist.*, **43**, 159–178.

Tierney, L. (1994) Markov chains for exploring posterior distributions, *Ann. Statist.*, **22**, 1701–1728.