Stat Comput (2006) 16: 203–213
DOI 10.1007/s11222-006-8450-8

Exact and efficient Bayesian inference for multiple changepoint problems

Paul Fearnhead

Received: September 2004 / Accepted: January 2006

Abstract We demonstrate how to perform direct simulation from the posterior distribution of a class of multiple changepoint models where the number of changepoints is unknown. The class of models assumes independence between the posterior distribution of the parameters associated with segments of data between successive changepoints. This approach is based on the use of recursions, and is related to work on product partition models. The computational complexity of the approach is quadratic in the number of observations, but an approximate version, which introduces negligible error, and whose computational cost is roughly linear in the number of observations, is also possible. Our approach can be useful, for example within an MCMC algorithm, even when the independence assumptions do not hold. We demonstrate our approach on coal-mining disaster data and on well-log data. Our method can cope with a range of models, and exact simulation from the posterior distribution is possible in a matter of minutes.

Keywords Bayes factor · Forward-backward algorithm · Model choice · Perfect simulation · Reversible jump MCMC · Well-log data

1. Introduction

Many time-series models incorporate one, or multiple, changepoints. Some examples include Poisson processes with a piece-wise constant rate parameter (Raftery and Ak-man, 1986; Yang and Kuo, 2001; Ritov *et al.*, 2002), changing linear regression models (Carlin *et al.*, 1992; Lund and Reeves, 2002), Gaussian observations with varying mean (Worsley, 1979) or variance (Chen and Gupta, 1997; Johnson *et al.*, 2003), and Markov models with time-varying transition matrices (Braun and Muller, 1998). Such models have been used for modelling stock prices, muscle activation, climatic time-series, DNA sequences and neuronal activity in the brain, amongst many other applications.

In this paper we consider Bayesian analysis for a class of multiple changepoint problems. We call a period of time between two consecutive changepoints a *segment.* This class of models assumes that the parameter values associated with each segment are independent from each other. Yang and Kuo (2001) comment that calculating the Bayes factors for models with different numbers of changepoints is "essentially infeasible for a large model with many changepoints". Our aim is to show that calculation of Bayes factors, and perfect sampling from the posterior distribution of changepoint locations, is both possible, and computationally inexpensive for the class of models we consider. While this class of models may seem restrictive, recent examples of work on such models can found in Johnson *et al.* (2003), Punskaya *et al.* (2002), and Braun *et al.* (2000).

Although we use the phrase "perfect simulation", we do not use coupling-from-the past (Propp and Wilson, 1996), or related ideas, which have become synonymous with this phrase. Instead, the work we present is closely related to work by Yao (1984), Barry and Hartigan (1992) and Barry and Hartigan (1993). These papers present efficient recursions that allow the posterior probabilities of different numbers of changepoints, and the posterior mean of the parameters to be calculated. Despite the desirability of exact solutions, and the simplicity and computational efficiency of the recursions, these methods are currently underused. We extend these methods to allow for direct simulation from the posterior distribution of the number and position of the changepoints, and to also perform inference conditional on

P. Fearnhead (✉)
Department of Mathematics and Statistics, Lancaster University

the number of changepoints. Our approach is a generalisation of that suggested by Liu and Lawrence (1999).

Much recent research for changepoint models is based on the use of MCMC. For inference in the presence of a known number of changepoints, Stephens (1994) and Chib (1996) both propose MCMC methods. For models with an unknown number of changepoints, a common approach is that of Green (1995): a set of models, each incorporating a different number of changepoints, are introduced, and reversible jump MCMC is used to explore the joint space of model and parameters. An alternative approach, based on analysing the different models separately is given by Chib (1998); with the different models being compared based on their evidence (also known as marginal likelihood), which can be estimated using ideas from Chib (1995). Potential difficulties of these approaches include designing moves, particular ones between different models, which enable the MCMC algorithm to mix well (for guidelines on designing reversible jump MCMC algorithms see Brooks *et al.*, 2003), and being able to detect convergence of the algorithm. For example, in the analysis of the coal-mining disaster data in Green (1995), the reversible jump MCMC algorithm had not converged. The reanalysis of the data in Green (2003), using a reversible jump MCMC algorithm run for 25 times as long, does fully explore the posterior distribution. The exact simulation method we describe here avoids any problems of needing to diagnose convergence of an MCMC algorithm.

We consider two classes of prior for the changepoint process. One, that of Green (1995), involves a prior on the number of changepoints, and then a conditional prior on their position. The other is based on modelling the changepoint process by a point process (Pievatolo and Green, 1998), and is a special case of a product-partion model (Hartigan, 1990). This indirectly specifies a joint prior on the number and position of the change-points. In both cases we assume that, conditional on the realisation of the changepoint process, the joint posterior distribution of the parameters is independent across the segments of the time series. We also assume a conjugate prior for the parameters associated with each segment. Under these two assumptions we derive a set of recursions to perform exact inference.

The recursions are similar to those of the Forward-Backward algorithm (see Scott, 2002, for a review). Recent work has shown how such recursions can be used to perform exact inference for a range of problems (Fearnhead and Meligkotsidou, 2004; Fearnhead, 2005a). The assumption of independence between segments ensures the necessary Markov property that is required for Forward-Backward type recursions. For a data set consisting of observations at discrete times, $1, \ldots, n$, the recursions are based on calculating the probability of the data from time t to time n, given a changepoint at time t, in terms of the equivalent probabilities at times $t+1, \ldots, n$. Once these probabilities have been calculated for all time-points, it is possible to directly simulate from the posterior distribution of the time of the first changepoint, and then the conditional distribution of the time of the second changepoint, given the first, and so on. The recursions can also be used to perform exact inference conditional on the number of changepoints, and in some cases to calculate the posterior distribution of the parameters that govern the point process model for the changepoints.

The computational cost of the recursions increases quadratically with n. However an approximate version, which introduces negligible error, is possible. In limiting situations where the length of time series increases, and the number of changepoints is increasing linearly with the number of observations, the computational cost increases roughly linearly with n. (In the alternative limiting regime of more frequent observations, the computational cost remains quadratic in n.) The assumption of conjugate priors can potentially be relaxed, but with an increase in the computational cost. Essentially, low-dimensional integrals that can be calculated analytically under conjugate priors would need to be calculated numerically (for example see Section 4.2). Relaxation of the independence assumption is more difficult, but our algorithm can still be used as a useful tool for analysing such data. For example, the algorithm can be embedded in an MCMC algorithm, and we demonstrate such an approach on some real data.

The outline of the paper is as follows. In Section 2 we introduce the two classes of changepoint model that we consider. The recursions are derived and detailed in Section 3. The resulting algorithm is demonstrated on two data sets in Section 4. For the second, we show how our method can be applied within an MCMC scheme to analyse the data under a model where there is dependence between the parameters for each segment. The paper concludes with a discussion.

2. Models and notation

We consider the following class of multiple changepoint models. Consider a sample of size n, $y_1, \ldots, y_n$. Observation y_i is obtained at time i, and we let $y_{i:j}$ denote the observations from time i to time j inclusive.

Firstly condition on m integer-valued changepoints, at points $0 < \tau_1 < \tau_2 < \cdots < \tau_m < n$. We let $\tau_0 = 0$ and $T_{m+1} = n$. Then the jth segment consists of the observations from time $\tau_{j-1} + 1$ to time τ_j. We associate a (possibly vector-valued) parameter θ_j with the jth segment for $j = 1, \ldots, m+1$. Conditional on the change-points and parameter values, the observations are independent; observation y_i being drawn from a density $f(y_i|\theta_j)$ if time i is in the jth segment.

We assume independent priors for the parameters associated with each segment. The prior for θ_j is denoted by $\pi(\theta_j)$.

Here, and throughout, we use $\pi(\cdot)$ solely to denote a prior density; the argument making it clear as to which parameter the prior is for.

We assume that the changepoints occur at discrete time points, and consider two priors for the changepoints. The first prior is based on a prior for the number of changepoints, and then a conditional prior on their positions. We will define this conditional prior on the positions in terms of $\pi_m(\tau_m)$ the prior for the last change point, and, for $j = 1, \ldots, m-1$, $\pi_m(\tau_j|\tau_{j+1})$, the prior for the position of the jth changepoint, given the position of the $(j+1)$st.

The second prior is obtained from a point process on the positive and negative integers. The point process is specified by the probability mass function $g(t)$ for the time between two succesive points. We assume that this time must be a strictly positive integer. We observe the point process on the interval $[1, n-1]$, and assume that changepoints occur at the positions of points in the point process. This prior is an example of a product-partition model.

If $G(t) = \sum_{s=1}^{t} g(s)$, is the distribution function of the distance between two succesive points, and $g_0(t)$ is the mass function of the first point after 0, then the probability of m changepoints occuring at $\tau_1, \ldots, \tau_m$ is

$$g_0(\tau_1)\left(\prod_{j=2}^{m} g(\tau_j - \tau_{j-1})\right)(1 - G(\tau_{m+1} - \tau_m)).$$

Natural choices for the distribution of the time between successive points are from the negative binomial family. For a negative binomial distribution with parameters k, a positive integer, and p we have

$$g(t) = \binom{t-k}{k-1} p^k (1-p)^{t-k}$$

$$g_0(t) = \sum_{i=1}^{k} \binom{t-i}{i-1} p^i (1-p)^{t-i}/k.$$

The negative binomial distribution can be thought of as a discrete version of the gamma distribution (especially if p is small). If $k = 1$ then the negative binomial distribution is the geometric distribution, and the point process is Markov. Larger values of k can reduce the number of very short segments.

3. Filtering recursions

We first derive the recursions for analysing data under the point process prior for the changepoints. We later derive recursions to perform inference conditional on the number of changepoints, and show how these can be used to perform inference under the other prior, and to perform inference about the parameters of the point process prior.

3.1. Basic recursions

For times $s \geq t$, define

$$\begin{aligned} P(t,s) &= \Pr(y_{t:s}|t, s \text{ in the same segment}) \\ &= \int \prod_{i=t}^{s} f(y_i|\theta)\pi(\theta)\mathrm{d}\theta. \end{aligned} \tag{1}$$

We will assume that the probabilities $P(t,s)$ can be calculated for all t and s. In practice this will require conjugate priors on θ, or, if θ is low-dimensional, that the required integration can be calculated numerically.

We next define for $t = 2, \ldots, n$

$$Q(t) = \Pr(y_{t:n}|\text{changepoint at } t-1),$$

with $Q(1) = \Pr(y_{1:n})$. A set of recursions for calculating these probabilities are given by the following theorem.

Theorem 1. *Define the probabilities $Q(t)$ and $P(t,s)$ as above. Then for $t = 2, \ldots, n$*

$$\begin{aligned} Q(t) &= \sum_{s=t}^{n-1} P(t,s)Q(s+1)g(s+1-t) \\ &\quad + P(t,n)(1 - G(n-t)), \end{aligned} \tag{2}$$

and

$$\begin{aligned} Q(1) &= \sum_{s=1}^{n-1} P(1,s)Q(s+1)g_0(s) + P(1,n) \\ &\quad \times (1 - G_0(n-1)), \end{aligned} \tag{3}$$

where $G_0(t) = \sum_{s=1}^{t} g_0(s)$.

Proof: We only prove Equation 2. Equation 3 can be derived similarly.

For notational convenience we drop the explicit conditioning on a changepoint at $t-1$ in the following. Thus,

$$\begin{aligned} Q(t) &= \Pr(y_{t:n}) \\ &= \sum_{s=t}^{n-1} \Pr(y_{t:n}, \text{ next changepoint at } s) \\ &\quad + \Pr(y_{t:n}, \text{ no further changepoints}). \end{aligned}$$

Now these probabilities can be calculated by the product of the prior probability on the changepoints, and the probabilities of the observations from a single segment, $P(t, s)$. Thus

$$\begin{aligned}
&\Pr(y_{t:n}, \text{ next changepoint at } s)\\
&\quad = \Pr(\text{next changepoint at } s)\\
&\qquad \Pr(y_{t:s}, y_{s+1:n} | \text{next changepoint at } s)\\
&\quad = g(s+1-t)\Pr(y_{t:s}|t, s \text{ in same segment})\\
&\qquad \Pr(y_{s+1:n}|\text{changepoint at } s)\\
&\quad = g(s+1-t)P(t,s)Q(s+1)
\end{aligned}$$

Similarly

$$\begin{aligned}
\Pr(y_{t:n}, \text{ no further changepoints}) = P(t,n)\\
\cdot(1-G_0(n-t)),
\end{aligned}$$

as required. □

Equations 2 and 3 give recursions that can be used to calculate $Q(t)$ in turn for $t = n, \ldots, 1$. The evidence of the model is just $Q(1)$. These equations are equivalent to those of Barry and Hartigan (1992), and are based on the same idea as recursions of Yao (1984).

The computational complexity of the resulting algorithm is quadratic in n. However often only a small proportion of the terms on the right-hand side of (2) make an appreciable contribution to $Q(t)$. This can happen when the data makes it almost certain that a changepoint occurs before a given time-point. Thus the summation can often be truncated with negligible error. We propose truncating the sum at term k when

$$\frac{P(t,k)Q(s+1)g(k+1-t)}{\sum_{s=t}^{k} P(t,s)Q(s+1)g(s+1-t)} \tag{4}$$

is less than some predetermined value, for example 10^{-10}.

In the limiting regime of analysing a process over a longer time period, so that the number of changepoints will increase roughly linearly with the number of observations, n, the computational complexity of the resulting approximate set of recursions will be linear in n. Essentially the average number of terms required in the right-hand side of (2) will be constant with t. Thus the average computational cost of one of the n recursions will be independent of n.

3.2. Perfect simulation of changepoints

Given the values of $Q(t)$ for $t = 1, \ldots, n$ it is straightforward to simulate from the posterior distribution of the changepoints as follows.

The posterior distribution of the first changepoint is given by

$$\begin{aligned}
\Pr(\tau_1|y_{1:n}) &= \Pr(y_{1:n}, \tau_1)/\Pr(y_{1:n})\\
&= \Pr(\tau_1)\Pr(y_{1:\tau_1}|\tau_1)\Pr(y_{\tau_1+1:n}|\tau_1)Q(1)\\
&= P(1,\tau_1)Q(\tau_1+1)g_0(\tau_1)/Q(1),
\end{aligned}$$

for $\tau_1 = 1, \ldots, n-1$. The probability of no further changepoint being $P(1,n)(1-G_0(n-1))/Q(1)$.

Similarly the posterior distribution of the τ_j given τ_{j-1} is

$$\begin{aligned}
\Pr(\tau_j|\tau_{j-1}, y_{1:n}) = P(\tau_{j-1}+1, \tau_j)Q(\tau_j+1)\\
\times g(\tau_j-\tau_{j-1})/Q(\tau_{j-1}+1),
\end{aligned}$$

for $\tau_j = \tau_{j-1}+1, \ldots, n-1$, and the probability of no further breakpoint is $P(\tau_{j-1}+1, n)(1-G_0(n-\tau_{j-1}-1))/Q(\tau_{j-1}+1)$.

Efficient simulation of large samples of changepoints from the posterior distribution can be done by simulating the samples concurrently, using the following algorithm. We denote the generic posterior distribution of the next changepoint, given a changepoint at t by $\Pr(\tau|y_{1:n}, t)$, which can be calculated as above.

(1) For a sample of size M, initiate each of the M samples with a changepoint at $t = 0$.
(2) For $t = 0, \ldots, n-2$:
   (i) Calculate n_t the number of samples whose last changepoint was at time t.
   (ii) If $n_t > 0$ calculate the probability distribution $\Pr(\tau|y_{1:n}, t)$.
   (iii) Sample n_t times from $\Pr(\tau|y_{1:n}, t)$ using Algorithm 1 of Carpenter *et al.* (1999) (see the Appendix). Use these values to update the n_t samples of changepoints which have a changepoint at t.

There are two advantages of this algorithm. The first is that the probability mass function $\Pr(\tau|y_{1:n}, t)$ need only be calculated once regardless of the number of samples required from it. If changepoints are sampled one at a time, then either these densities will, potentially, need to be calculated for each sample, or they will need to be stored. Storing these mass functions can place large burdens on computational memory. The storage requirements will be quadratic in n; by comparison the above algorithm has storage requirements that are linear in n.

The second is that simulating a sample of size M from a general discrete mass function can be achieved more efficiently than sampling M samples of size 1. Algorithm 1 of Carpenter *et al*. (1999) allows a sample of size M to be simulated with order $n + M$ effort, rather than the nM effort of sampling M samples of size 1.

3.3. Conditioning on the number of changepoints

Now consider inference conditional on m changepoints. As in Section 2 we define the prior for the changepoints via $\pi_m(\tau_m)$ and conditional probabilities of the form $\pi_m(\tau_j|\tau_{j+1})$. We define $P(s,t)$ as before, and for $j = 1, \ldots m$, and $t = j+1, \ldots, n-m-1+j$,

$$Q_j^{(m)}(t) = \Pr(y_{t:n}|\tau_j = t-1, m \text{ changepoints}).$$

We can derive the following set of recursions. For $t = m+1, \ldots, n-1$,

$$Q_m^{(m)}(t) = P(t,n)\pi_m(\tau_m = t-1).$$

For $j = 1, \ldots, m-1$, and $t = j+1, \ldots, n-m-1+j$

$$Q_j^{(m)}(t) = \sum_{s=t}^{n-m+j} P(t,s)Q_{j+1}^{(m)}(s+1) \times \pi_m(\tau_j = t-1|\tau_{j+1} = s).$$

Finally

$$\Pr(y_{1:n}|m \text{ changepoints}) = \sum_{s=1}^{n-m} P(1,s)Q_1^{(m)}(s+1). \qquad (5)$$

These can be proved in a similar way to Theorem 1.

If the number of changepoints is unknown, with prior $\pi(m)$, then the posterior distribution of m can be calculated as

$$\Pr(m|y_{1:n}) \propto \pi(m)\Pr(y_{1:n}|m \text{ changepoints}),$$

with the last term, the evidence for m changepoints, being calculated, for each m, using the recursions.

Simulation from the joint posterior distribution is possible by first simulating M samples from $\Pr(m|y_{1:n})$. If the value m is sampled N_m times, then N_m samples from the posterior distribution of the changepoint positions, conditional on m changepoints, can be obtained as described in Section 3.2. The only difference is that the conditional distribution of τ_j given τ_{j-1} is now

$$\Pr(\tau_j|\tau_{j-1}, y_{1:n}, m) = P(\tau_{j-1}+1, \tau_j)Q_j^{(m)}(\tau_j+1) \times \pi_m(\tau_{j-1}|\tau_j)/Q_{j-1}^{(m)}(\tau_{j-1}+1).$$

Finally, in the case of the Markov point process prior (that is, a geometric distribution for the distance between changepoints), exact inference is possible even if the probability of a changepoint at any timepoint, p, is unknown. This is because, conditional on the number of changepoints, the positions are distributed uniformly along the interval, independent of p. We can thus perform inference conditional on m changepoints. Furthermore, under the Markov point process prior for the changepoints, the conditional prior for the number of changepoints, m, $\pi(m|p)$ has a Binomial distribution with paramters $n-1$ and p. Thus the marginal prior for m is obtained by averaging this conditional prior, $\pi(m|p)$, with respect to the prior for p.

4. Examples

4.1. Coal mining disaster data

As our first example, we consider fitting multiple changepoints to the coal mining disaster data of Jarrett (1979). This is a standard data set for testing methods for inferring changepoints. Raftery and Akman (1986) and Carlin *et al*. (1992) fit single changepoint models. Green (1995) fits a multiple changepoint model using reversible jump MCMC, and Yang and Kuo (2001) infer multiple changepoints using binary segmentation. Our analysis is based on the model of Green (1995), except that we discretize time into weekly time units, and allow changepoints to only occur at these discrete time points. By comparison, Green (1995) uses a continuous time model.

The data consists of the dates of 191 coal mining disasters between 1851 and 1962, a period of 5844 weeks. We assume the number of disasters in any week has a Poisson distribution, and the underlying Poisson mean, μ, is piecewise constant through time. For a given segment, μ has a $\Gamma(\alpha, \beta)$ prior density, $\beta^\alpha \mu^{\alpha-1} e^{-\beta\mu}/\Gamma(\alpha)$, for $\mu > 0$. We thus obtain from (1)

$$P(t,s) = \frac{\Gamma(\alpha + \sum_{i=t}^{s} y_i)\beta^\alpha}{\Gamma(\alpha)(\beta + s - t + 1)^{\alpha + \sum_{i=t}^{s} y_i}}.$$

Our prior distribution for the number of changepoints is Poisson with mean 3. We consider a prior distribution for the changepoint positions, conditional on m changepoints, which is specified by the even order statistics of $2m+1$ uniform draws from the numbers $\{1, \ldots, n-1\}$ without replacement. This is a discrete version of the prior of Green (1995), and has the advantage of penalising very short segments. This prior is specified by

$$\pi_m(\tau_1, \ldots, \tau_m) = K_m^{-1} \prod_{i=0}^{m} (\tau_{i+1} - \tau_i - 1),$$

where K_m is the normalising constant, the number of combinations of picking $2m + 1$ from $n - 1$ numbers.

We calculate the $Q_j^{(m)}$s using $\pi_m(\tau_j = t|\tau_{j+1} = s) = s - t - 1$. The correct value of $\Pr(y_{1:n}|m$ changepoints) is obtained by dividing through by K_m in (5). The advantage of this approach is that the $Q_j^{(m)}$s, for different values of m, are related by

$$Q_j^{(m)}(t) = Q_{j+k}^{(m+k)}(t),$$

for any integer k.

Our first analysis uses an equivalent prior for the Poisson mean to that of Green (1995), with $\alpha = 1$ and $\beta = 200/7$. The difference in the scale parameter is due to the different time units that we use: weeks rather than days. The results for the posterior distribution of the number of changepoints, and their positions conditional on 2 changepoints, are shown in Figure 1 (a)–(b).

The posterior distribution for the number of changepoints differs from that shown in Green (1995); in the production of those previous results the Monte Carlo simulation was not run for long enough. A reanalysis of the data in Green (2003), using the same model and MCMC algorithm, but run for 25 times as many iterations, gives almost identical results to those shown here. The only difference between these previous analyses and ours is that we have discretised time. This comparison, together with a simple analysis of the data under a continuous time model with $m = 1$ and $m = 2$ using importance sampling (results not shown), suggest that the difference introduced by discretising time is negligible. Our algorithm took an order of magnitude longer to analyse this data than the analysis of Green (2003): 5 minutes on a 900MHz Pentium PC rather than half a minute on an 800MHz PC. However analysis of the data under a point process prior (with the distribution of the time between changepoints being negative binomial with parameters 2 and 0.001) produces almost identical results and takes around half a minute to run. Further computational savings may be possible using the approximation described in Section 3.1.

Figure 1 (c)–(d) shows results for analysing the data with the diffuse prior of Yang and Kuo (2001) for the Poisson means ($\alpha = 0.5$ and $\beta = 10^{-7}$). The choice of prior

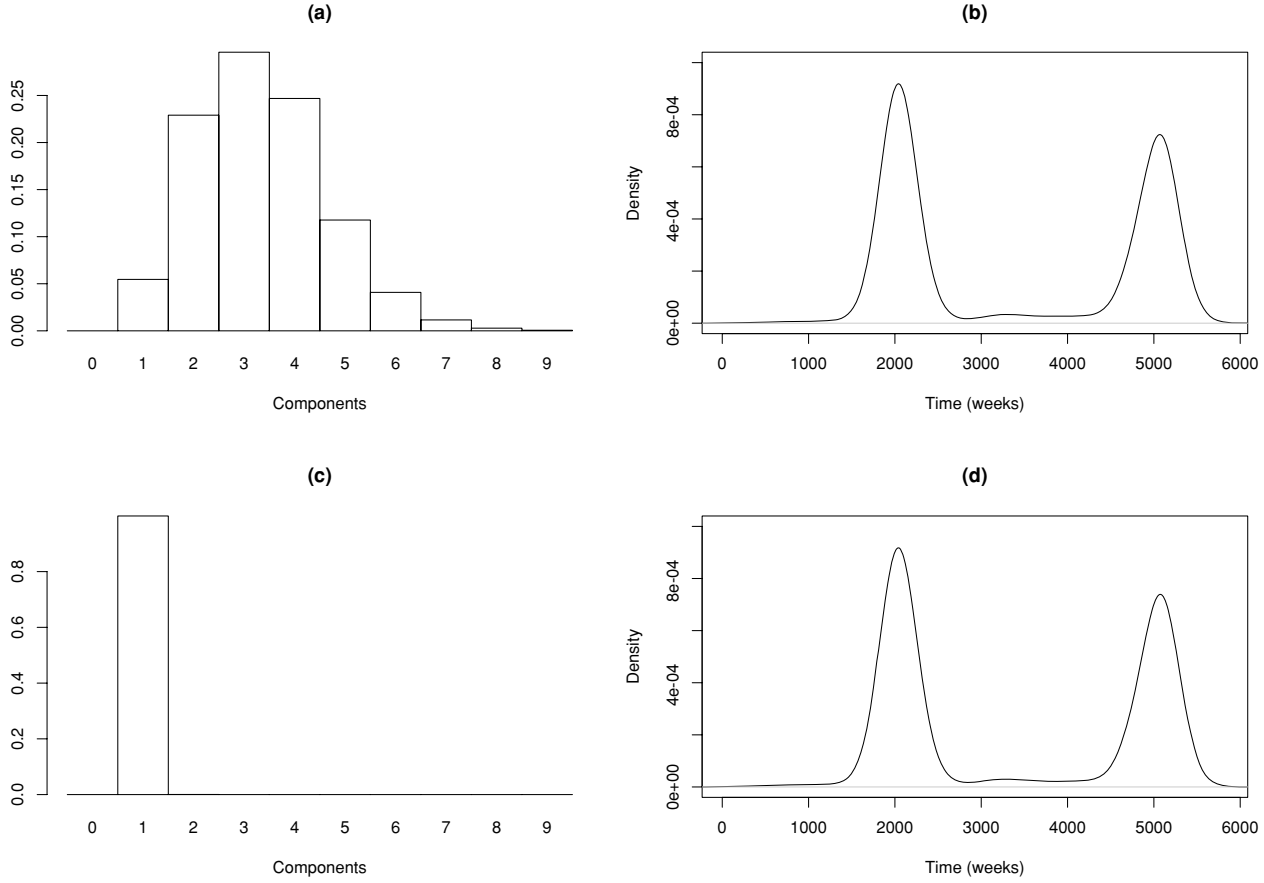

Fig. 1 Analysis of the Coal-mining data for two different priors for the Poisson means for each segment. Left-hand plots are the posterior distribution of the number of changepoints, and right-hand plots are the conditional posterior distributions of the position of the changepoints, conditional on there being 2, and are obtained from perfect samples of size 10,000 from the corresponding posterior distributions. The prior for the Poisson means were $\Gamma(\alpha, \beta)$, with (a)–(b) $\alpha = 1$ and $\beta = 200/7$; and (c)–(d) $\alpha = 0.5$ and $\beta = 10^{-7}$

distribution greatly affects the posterior distribution for the number of changepoints, but has negligible effect on the distribution of the changepoint positions (this is still true if we condition on the presence of a different number of changepoints). The choice of prior effectively controls how much penalty is incurred for introducing additional changepoints, and thereby additional parameters. The more diffuse the prior, the larger this penalty is, and the approach of Yang and Kuo (2001) of introducing a prior to mimic an improper prior on the parameters is inappropriate.

4.2. Well-log data

We now consider the problem of detecting changepoints in well-log data. An example of well-log data, which comes from Ó Ruanaidh and Fitzgerald (1996), is given in Figure 2(a). The data consist of 4050 measurements of the nuclear-magnetic response of underground rocks. The data were obtained by lowering a probe into a bore-hole. Measurements were taken at discrete timepoints by the probe as it was lowered through the hole. The underlying signal is roughly piecewise constant, with each constant segment relating to a single rock type (that has constant physical properties). The changepoints in the signal occur each time a new rock type is encountered. Detecting the change-points is important in oil-drilling; see the introduction of Fearnhead and Clifford (2003) for more details.

These data have been previously analysed by Ó Ruanaidh and Fitzgerald (1996), who used MCMC to fit a change-point model with a fixed number of changepoints; and by Fearnhead and Clifford (2003) who considered online analysis of the data using particle filters. We performed a batch analysis of the data, but allowed for multiple changepoints.

Piecewise constant model

Initially we consider an analysis based on a model taken Fearnhead and Clifford (2003). We assume a Markov point process prior for the changepoints. There are a number of outliers in the data which were removed before the data was

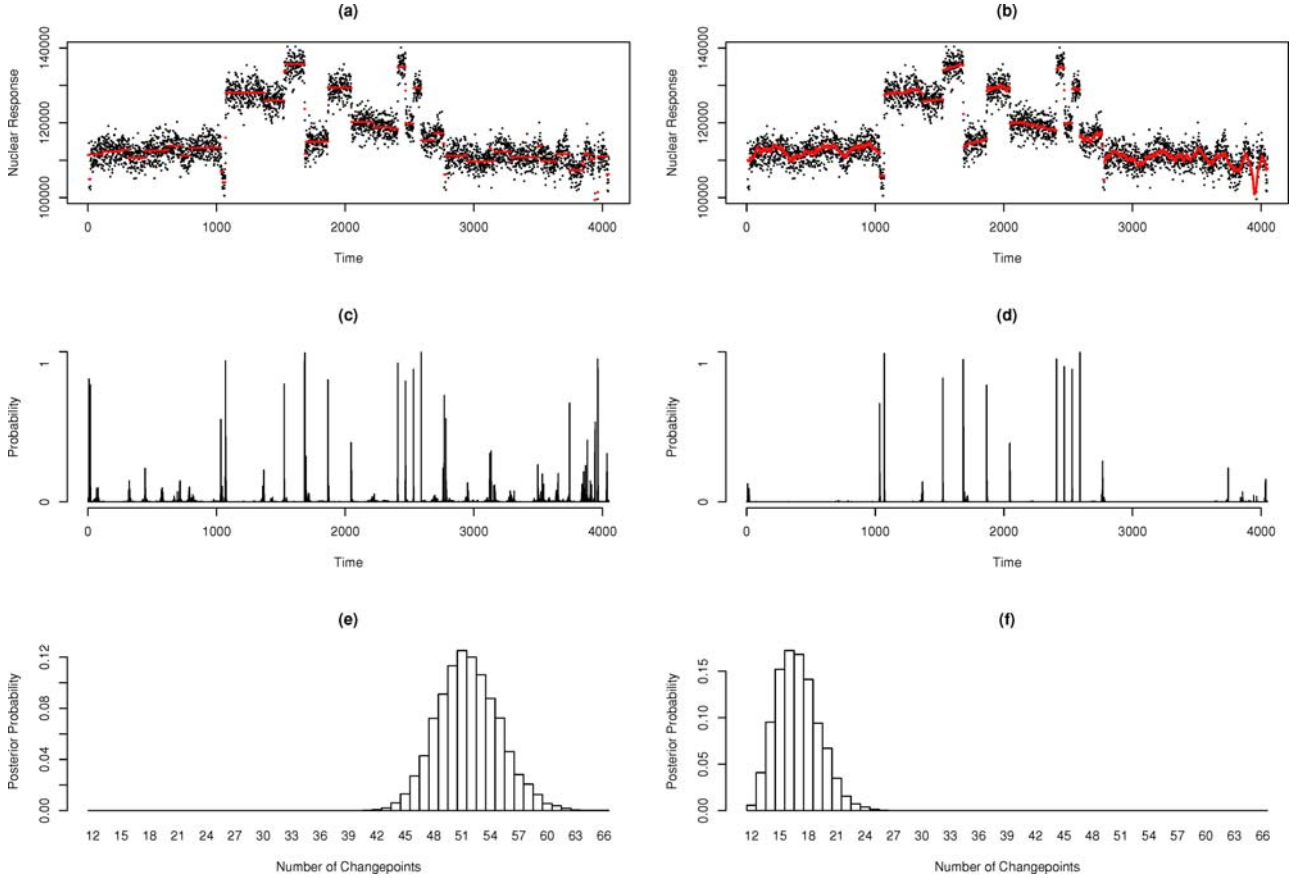

Fig. 2 Well-log data, and analysis under piecewise constant model (left-hand column) and random walk model (right-hand column). Plots (a) and (b) show the well-log data together with a realisation from the posterior distribution of the signal mean (in red). Plots (c) and (d) give marginal probabilities of changepoints at each time-point; and plots (e) and (f) give histograms for the posterior distribution of the number of change points

analysed. For a time t which belongs to segment i, we model a non-outlying observation, y_t, by

$$y_t \sim \mathrm{N}(\mu_i, \sigma^2),$$

where μ_i is the mean associated with the ith segment, and we assume a common known variance, $\sigma^2 = 2500^2$. We assume that the segment means have independent normal priors with mean $\eta = 115,000$ and variance $\tau^2\sigma^2 = 10,000^2$ (so $\tau = 4$). Conditional on the segment means, the observations are independent. For this model, we have from (1)

$$P(t,s) = (k\tau^2+1)^{-1/2} \times \exp\left\{-\frac{1}{2\sigma^2}\left(\frac{k}{k\tau^2+1}(\eta - s_1/k)^2 + \left(s_2 - s_1^2/k\right)\right)\right\},$$

where k is the number of (non-outlier) observations between times t and s (inclusive), and s_1 and s_2 are respectively the sum of these observations and the sum of the square of these observations. Using the methods of Section 3.3 we performed inference for p, assuming a uniform prior for p. This gave a posterior mode of $p = 0.013$, and the results we present are for p fixed to this value.

In the left-hand column of Figure 2 we present the results of our analysis. The plots are for the posterior distribution of the number of changepoints and their positions (based on 10,000 independent draws from the posterior), and one realisation of the underlying signal. It took 26 seconds on a 3.4GHz PC to perform this analysis.

We repeated our analysis using the approximate algorithm suggested in Section 3.1. We truncated the sums in Equations (2) and (3), used to calculate the $Q(t)$s, when the value of (4) was less than 10^{-10}. The resulting algorithm on average required sums of 222 terms to be calculated for each $Q(t)$; which compares with average sums of 2025 terms for the exact algorithm. This is a nine-fold reduction in the complexity of the algorithm. The resulting approximation of the log evidence was correct to 4 decimal places, which suggests that negligible errors were introduced.

Comparison with MCMC

For comparison we analysed the same data set under the same model with the MCMC method of Lavielle and Lebarbier (2001), which has been specifically designed for the Gaussian changepoint model used for the well-log data.

This method integrates out the segment parameters, and uses MCMC to mix over the number and position of changepoints. Let $\mathbf{r} = (r_1, r, \ldots r_{n-1})$ be a binary vector with $r_t = 1$ if and only if there is a change-point at time t. At each iteration of the algorithm, we successively use one of three possible moves to update $\mathbf{r}$. These moves propose a new vector $\mathbf{r}'$, and the move is accepted with a probability that ensures the MCMC algorithm has the correct stationary distribution. The possible moves are

(A) Propose a new set of changepoint positions, $\mathbf{r}'$ from the prior.
(B) Choose a time-point, t, uniformly at random from $1, 2, \ldots, n-1$; if $r_t = 0$ propose $r'_t = 1$, else propose a $r'_t = 0$, with $r'_s = r_s$ for $s \neq t$.
(C) Choose uniformly at random two time-points, t and t' such that $r_t = 1$ and $r_{t'} = 0$. Propose a move to $r'_t = 0$, $r'_{t'} = 1$ and $r_s = r'_s$ for $s \neq t, t'$.

We ran this MCMC algorithm for a total of 500,000 iterations (where one iteration involves an application of each of moves (A)-(C)). Of the moves which change the number of changepoints, move (A) had negligible acceptance probability (such a move was never accepted in our run), and move (B) had an acceptance probability of 0.4%. The autocorrelation for the number of changepoints was 0.97 at lag 100, and the estimated autocorrelation time was around 6,000. This suggests that 6×10^7 iterations would be required to have the same amount of information as 10,000 independent draws from the posterior. Such a run would take around two orders of magnitude longer than the direct simulation method.

Inclusion of hyperpriors

We now consider an extension of the above model where all parameters in our model were unknown, and we introduce hyperpriors for them. This introduces dependence between the segments, and our direct simulation algorithm has to be used within an MCMC scheme.

We used a uniform prior for p, and an improper prior for σ, $\pi(\sigma) \propto 1/\sigma$. We parameterised the prior for the segment means as

$$\mu_i \sim \mathrm{N}(\eta, \tau^2\sigma^2),$$

and used improper hyperpriors on η and τ : $\pi(\eta) \propto 1$ and $\pi(\tau) \propto 1/\tau$.

We analysed this model using MCMC. The MCMC algorithm used the following three updates:

(1) Update the changepoints conditional on σ, p, η, and τ. We used an independent proposal from the true posterior distribution conditional on $\sigma = 2,330$, $p = 0.013$, $\eta = 115000$, and $\tau = 4.3$.
(2) Update σ, p and the μ_is from their full conditional distribution given the changepoints and η and τ.
(3) Update η and τ from their full-conditionals given the μ_is.

Each of these moves satisfies detailed balance. Steps (2) and (3) are Gibbs steps, and thus the proposed values are always accepted. Step (1) is not a Gibbs step. Although it would be possible to make it so, there is a substantial overhead to calculating the posterior distribution of the changepoints at each iteration. Thus while this algorithm may mix more slowly, a single iteration will be substantially quicker, and hence we hope it will be more efficient. In updating the changepoints in step (1) we throw away the segment means. This is an example of collapsing (Liu, 2001, pages 146–151), which usually improves the mixing of the Markov Chain.

We ran this Markov chain for 10,000 iterations. The acceptance probability of step (1) was 61.8%. The 1-lag autocorrelation for each of the parameters was less than 0.03, which suggests that the chain is mixing extremely quickly. The advantage of using the direct simulation method for updating the number and position of changepoints rather than proposing these from the prior (as in step (A) of the MCMC algorithm Lavielle and Lebarbier, 2001) is significant (see above).

The reason why this MCMC algorithm performs so well is because the posterior probability of the parameters is concentrated in a small region of the parameter space. Over this small region, the parameters are almost independent; the maximum absolute value of the correlation between any pair of parameters is 0.01. Furthermore, the conditional distribution of the changepoints changes little over this range of parameter values, which means that the average acceptance probability in step (1) of the algorithm is high. This situation is likely to occur in other situations where there is a large and informative data set with many changepoints.

Alternative models

The model used in the previous Section is based around those previously used in the literature for this data. However the realisations from the posterior distribution have many more changepoints, and thus suggest many more rock strata, than is realistic. It appears that the piecewise constant model used is overly simplistic for the data, and that this has resulted in the need for too many changepoints in order to fit the data.

We have considered numerous extensions to the model. Two possibilities are: (i) to allow different noise variances for different segments; and (ii) to model each segment using a mean-shifted AR(1) model (Albert and Chib, 1993). Both of these models can be analysed via our direct simulation method, though for (ii) we need to numerically integrate out the autoregressive coefficient (this can be done in a similar way to that described below). However neither of these extensions enable the data to be fit with substantially fewer change points (results not shown).

Instead we consider the following state-space model for the data within a segment, where if $t-1$ and t both lie within segment i

$$\mu_t \sim \mathrm{N}\left(\mu_{t-1}, \tau_i^2\right)$$
$$y_t \sim \mathrm{N}(\mu_t, \sigma^2)\ .$$

The initial μ value for each segment is drawn from the same independent normal priors as before. This is an extension of the piecewise constant model which allows the signal within a segment to perform a random walk. We allow the variance of the random walk to vary among segments, and assume a Gamma prior for τ_i with parameters 2 and $1/40$. This prior places most probability mass on values of τ_i which lie in the interval [0,150].

The idea of this model is that the random walk element can fit the small-scale variation in the underlying signal without the need to infer changepoints.

If τ_i were known for each segment then it would be straightforward to apply our direct simulation method, using the Kalman Filter (Harvey, 1989) to integrate out the underlying signal. To incorporate a prior on τ_i we resort to numerical integration to calculate the $P(t, s)$ values required by our algorithm. A simple, but adequate, approach to numerical integration is based on using a grid of τ_i values, and we obtained such a grid as follows. For a grid with K points, first simulate for $k = 1, \ldots, K$, a realisation, u_k, of a uniform random variable on $[(k-1)/K, k/K]$; then fix the kth grid point to be the u_kth quantile from the prior for τ_i.

In practice we found that a grid of 100 points produced accurate results; and for such a grid it took less than 19 minutes to simulate 10,000 draws from the joint posterior distribution of changepoint positions on a 3.4GHz PC. The results of the analysis (assuming $\sigma = 2,500$ and $p = 0.013$) are shown in Figure 2. This model gives more realistic inferences about the number and positions of the changepoints.

5. Discussion

We have described ways in which recursions, based on the Forward-Backward algorithm, can be used to perform Bayesian analysis of multiple changepoint problems. As mentioned previously, the work we present is closely related to work by Barry and Hartigan (1992). The main novelty of what we propose is that we demonstrate how the recursions can be used for perfect simulation from the posterior distribution of the number and position of change-points, and hence from the posterior distribution of the parameters. Presenting results from a Bayesian analysis via simulations from the posterior distribution is both quicker than calculating the

posterior means (as done by Barry and Hartigan, 1992, where the cost is cubic in the number of observations), and also encapsulates information about uncertainty about parameters, which is one of the advantages of Bayesian inference. We have also extended the use of recursions to inference conditional on the number of changepoints. Further extensions to allow for model-choice within segments (Fearnhead, 2005b) and online inference (Fearnhead and Liu, 2005) are also possible.

The ability to simulate from posterior distributions also enables the algorithms we present to be used in analysing more complex models, for example by embedding our algorithm within an MCMC algorithm (see Section 4.2). While it may seem natural in such cases just to use standard MCMC algorithms, the use of direct simulation enabled us to construct an MCMC algorithm for the Well-log data that had exceptional mixing properties.

Acknowledgements I would like to thank Peter Green and an anonymous referee for helpful comments on an earlier version of this paper; and to thank Bill Fitzgerald for sending me the Well-log data.

Appendix

To simulate in linear time a sample of size n from a discrete distribution $\Pr(\tau)$, which takes values of $\tau = 1, 2, \ldots$:

1(a) for $i = 1, \ldots, n+1$, simulate x_i a realisation from an exponential distribution with rate parameter 1;
1(b) Calculate $S = \sum_{i=1}^{n+1} x_i$;
1(c) Set $u_1 = x_1/S$ and for $i = 2, \ldots, n$ $u_i = u_{i-1} + x_i/S$.
2 Set $Q = 0, U = u_1, j = 1$ and $i = 1$.
3 If $U < Q + \Pr(\tau = j)$ then output j and set $U = u_{i+1}$ and $i = i + 1$; otherwise set $Q = \Pr(\tau = j)$ and $j = j + 1$. Repeat until $i = n + 1$.

References

Albert J. H. and Chib S. 1993. Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts. Journal of Business and Economic Statistics 11: 1–15.

Barry D. and Hartigan J. A. 1992. Product partition models for change point problems. The Annals of Statistics 20: 260–279.

Barry D. and Hartigan J. A. 1993. A Bayesian analysis for change point problems. Journal of the American Statistical Society 88: 309–319.

Braun J. V. and Muller H. G. 1998. Statistical methods for DNA sequence segmentation. Statistical Science 13: 142–162.

Braun J. V., Braun R. K., and Muller H. G. 2000. Multiple changepoint fitting via quasilikelihood, with application to DNA sequence segmentation. Biometrika 87: 301–314.

Brooks S. P., Giudici P., and Roberts G. O. 2003. Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions. Journal of the Royal Statistical Society, series B 65: 3–39.

Carlin B. P., Gelfand A. E., and Smith A. F. M. 1992. Hierarchical Bayesian analysis of changepoint problems. Applied Statistics 41: 389–405.

Carpenter J., Clifford P., and Fearnhead P. 1999. An improved particle filter for non-linear problems. IEE proceedings-Radar, Sonar and Navigation 146: 2–7.

Chen J. and Gupta A. K. 1997. Testing and locating changepoints with application to stock prices. Journal of the American Statistical Association 92: 739–747.

Chib S. 1995. Marginal likelihood from the Gibbs output. Journal of the American Statistical Association 90: 1313–1321.

Chib S. 1996. Calculating posterior distributions and modal estimates in Markov mixture models. Journal of Econometrics 75: 79–98.

Chib S. 1998. Estimation and comparison of multiple change-point models. Journal of Econometrics 86: 221–241.

Fearnhead P. 2005a. Direct simulation for discrete mixture distributions. Statistics and Computing 15: 125–133.

Fearnhead P. 2005b. Exact Bayesian curve fitting and signal segmentation. IEEE Transactions on Signal Processing 53: 2160–2166.

Fearnhead P. and Clifford P. 2003. Online inference for hidden Markov models. Journal of the Royal Statistical Society, Series B 65: 887–899.

Fearnhead P. and Liu Z. 2005. Online inference for multiple change-point problems. Submitted Available from www.maths.lancs.ac.uk/∼fearnhea/publications.

Fearnhead P. and Meligkotsidou L. 2004. Exact filtering for partially-observed continuous-time Markov models. Journal of the Royal Statistical Society, series B 66: 771–789.

Green P. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82: 711–732.

Green P. J. 2003. Transdimensional Markov chain Monte Carlo. In: Highly Structured Stochastic Systems (eds. Green P. J., Hjort N. L., and Richardson S.), Oxford University Press.

Hartigan J. A. 1990. Partition models. Communications in Statistics 19: 2745–2756.

Harvey A. C. 1989. Forecasting, stuctural time series and the Kalman filter. Cambridge University Press, Cambridge, UK.

Jarrett R. G. 1979. A note on the intervals between coal-mining disasters. Biometrika 66: 191–3.

Johnson T. D., Elashoff R. M., and Harkema S. J. 2003. A Bayesian change-point analysis of electromyographic data: detecting muscle activation patterns and associated applications. Biostatistics 4: 143–164.

Lavielle M. and Lebarbier E. 2001. An application of MCMC methods for the multiple change-points problem. Signal Processing 81: 39–53.

Liu J. S. 2001. Monte Carlo strategies in scientific computing. New York: Springer.

Liu J. S. and Lawrence C. E. 1999. Bayesian inference on biopolymer models. Bioinformatics 15: 38–52.

Lund R. and Reeves J. 2002. Detection of undocumented changepoints: A revision of the two-phase regression model. Journal of Climate 15: 2547–2554.

Ó Ruanaidh J. J. K. and Fitzgerald W. J. 1996. Numerical Bayesion Methods Applied to Signal Processing. New York: Springer.

Pievatolo A. and Green P. J. 1998. Boundary detection through dynamic polygons. Journal of the Royal Statistical Society, Series B 60: 609–626.

Propp J. G. and Wilson D. B. 1996. Exact sampling with coupled Markov chains and applications to statistical mechanics. Random Structures and Algorithms 9: 223–252.

Punskaya E., Andrieu C., Doucet A., and Fitzgerald W. J. 2002. Bayesian curve fitting using MCMC with applications to signal segmentation. IEEE Transactions on Signal Processing 50: 747–758.

Raftery A. E. and Akman V. E. 1986. Bayesian analysis of a Poisson process with a change-point. Biometrika 73: 85–89.

Ritov Y., Raz A., and Bergman H. 2002. Detection of onset of neuronal activity by allowing for heterogeneity in the change points. Journal of Neuroscience Methods 122: 25–42.

Scott S. L. 2002. Bayesian methods for hidden Markov models: Recursive computing in the 21st century. Journal of the American Statistical Association 97: 337–351.

Stephens D. A. 1994. Bayesian retrospective multiple-changepoint identification. Applied Statistics 43: 159–178.

Worsley K. J. 1979. On the likelihood ratio test for a shift in location of normal populations. Journal of the American Statistical Association 74: 363–367.

Yang T. Y. and Kuo L. 2001. Bayesian binary segmentation procedure for a Poisson process with multiple changepoints. Journal of Computational and Graphical Statistics 10: 772–785.

Yao Y. 1984. Estimation of a noisy discrete-time step function: Bayes and empirical Bayes approaches. The Annals of Statistics 12: 1434–1447.