Efficient Bayesian Analysis of Multiple Changepoint Models with Dependence across Segments

Paul Fearnhead and Zhen Liu

Department of Mathematics and Statistics, Lancaster University

Summary: We consider Bayesian analysis of a class of multiple changepoint models. While there are a variety of efficient ways to analyse these models if the parameters associated with each segment are independent, there are few general approaches for models where the parameters are dependent. Under the assumption that the dependence is Markov, we propose an efficient online algorithm for sampling from an approximation to the posterior distribution of the number and position of the changepoints. In a simulation study, we show that the approximation introduced is negligible. We illustrate the power of our approach through fitting piecewise polynomial models to data, under a model which allows for either continuity or discontinuity of the underlying curve at each changepoint. This method is competitive with, or out-performs, other methods for inferring curves from noisy data; and uniquely it allows for inference of the locations of discontinuities in the underlying curve.

1 Introduction

Changepoint models are commonly used for time-series, to allow for abrupt changes in the underlying model or structure for the data. Some example applications areas include genetics (Liu and Lawrence, 1999; McVean *et al.*, 2004), environmental time-series (Dobigeon and Toumeret, 2007; Seidou and Ouarda, 2007), and signal processing (Punskaya *et al.*, 2002), amongst many others.

We consider Bayesian inference for changepoint models. Existing methods for such inference are either based on MCMC approaches (e.g. Stephens, 1994; Chib, 1996, 1998; Lavielle and Lebarbier, 2001), or methods for direct simulation from the posterior (see e.g. Yao, 1984; Barry and Hartigan, 1992; Liu and Lawrence, 1999; Fearnhead, 2008). The methods for direct simulation have the advantage over MCMC of producing iid draws from the posterior, and they can also be implemented efficiently so that their computational cost is linear in the number of observations (Fearnhead and Liu, 2007). However they are limited in terms of the class of models that can be considered. If we call the period of time between two successive changepoints a segment, then direct simulation methods require the parameters associated with each segment to be independent of each other, and that the marginal likelihood for the data within each segment can be calculated analytically (or numerically, see Fearnhead, 2006).

One implementation of the direct simulation methods is based on solving filtering recursions (Fearnhead and Liu, 2007). We process the observations one at a time, and when processing the observation at a time t say, we calculate the posterior distribution of the time of the most recent changepoint prior to t. Here, we extend this direct simulation methods to models where there is dependence across segments. We assume that the dependence is Markov, so that parameters in the current segment depend only on the parameters in the previous segment. The assumption of dependence across segments greatly increases the complexity of calculating the posterior distribution, and to develop a computationally efficient algorithm we introduce a simple approximation. At time t we approximate the distribution of the parameters associated with a new segment, conditional on a changepoint at t. While this conditional distribution is a mixture distribution, with the number of terms in the mixture increasing exponentially with t, we approximate the mixture by a single distribution. This approximation leads to an efficient algorithm, but one that produces iid samples from an approximation to the posterior distribution of interest.

We demonstrate our new method on the problem of fitting piece-wise polynomial models. Here dependence across segments arises due to assumptions of continuity of the underlying curve. Existing methods for this problem include the MCMC methods of Denison *et al.* (1998) and DiMatteo *et al.* (2001), who also sample from an approximation to the posterior of interest, from approximating the marginal likelihood associated with each segment. Our model extends existing models that are considered, by allowing for the possibility of either continuity or discontinuity of the curve at each changepoint. Our approach also allows for online analysis of time-series.

The outline of the paper is as follows. Firstly we introduce the class of changepoint models we consider. Then in Section 3 we develop out algorithm for Bayesian inference for these models. Section 4 then analyses the resulting algorithm for the specific application of fitting piece-wise polynomial models. We first show that the approximation introduces negligible error when analysing simulated data from the true model. We also compare the resulting method with both wavelet-based methods and the MCMC method of Denison *et al.* (1998), and look at the power of the method for detecting discontinuities in the underlying signal. Section 5 applies our method to analysing well-log data. Here the focus of inference is in detecting changepoints where the underlying signal is discontinuous. Finally the paper ends with a discussion.

2 Changepoint model

We consider the following hierarchical model for observations $y_{1:n} = (y_1, \ldots, y_n)$. Firstly we introduce a model for the number, l, and position, $0 < \tau_1 < \cdots < \tau_l < n$, of the

changepoints. This is based on a distribution for the distance between two successive changepoints

$$p(\tau_k - \tau_{k-1} = d) = g(d),\qquad(1)$$

for some discrete distribution $g(\cdot)$ on the positive integers. We define $\tau_0 = 0$ and $\tau_{l+1} = n$, and we let $G(s) = \sum_{d=1}^{s} g(d)$ be the corresponding cumulative distribution function. We assume independence of the distance between different pairs of successive changepoints, so that the joint probability of l specific changepoints is

$$\Pr(\tau_1, \ldots, \tau_l) = \left(\prod_{k=1}^{l} g(\tau_k - \tau_{k-1})\right)(1 - G(n - \tau_l)).$$

The changepoints split the data into $l + 1$ segments, with the kth segment containing observations $\mathbf{y}_{\tau_k+1:\tau_{k+1}}$, for $k = 0, \ldots, l$. For segment k we associate a model M_k and a vector of parameters θ_k. The model is drawn from a finite set of possible models, $\mathcal{M}$ and we assume that there is independence of the choice of model across different segments. (Extension to the case where the model of a segment depends on the model of the previous segment is possible, see Fearnhead and Vasileiou, 2008).

For $k \geq 1$ we allow the distribution of θ_k to depend on the position of segment $k - 1$, τ_{k-1} and τ_k, and its parameter θ_{k-1}. Thus we have that the conditional probability of the model and parameters for the segment can be factorised as

$$\Pr(M_k = m)p_m(\theta_k|\theta_{k-1}\tau_k, \tau_{k-1}).$$

For the first segment we assume a prior for θ_0. Note that this framework includes models where there are common parameters across segments. In this case some components of θ_k are equal to the equivalent components of θ_{k-1} and the conditional probability $p_m(\theta_k|\theta_{k-1}\tau_k, \tau_{k-1})$ in only non-zero for parameter combinations that obey this constraint.

Given a segment defined by changepoints at positions s and t, and with model m and parameter θ we have a likelihood model

$$p_m(\mathbf{y}_{s+1:t}|\theta).\qquad(2)$$

We assume that conditional on the changepoints, segment models and parameters, the observations within each segment are independent of each other.

Finally we assume that there exists a family of conjugate priors for θ, $p_m(\theta|\varsigma)$. Thus for all m, ς and y_t and s, t, we can calculate

$$P_s(t, m, \varsigma) = \int p_m(y_t|\theta, s)p_m(\theta|\varsigma)\mathrm{d}\theta,\qquad(3)$$

3

where

$$p_m(y_t|\theta, s) = \frac{p_m(\mathbf{y}_{s+1:t}|\theta)}{p_m(\mathbf{y}_{s+1:t-1}|\theta)}$$

is the probability density of y_t given a segment that started with observation y_{s+1}. Furthermore, there exists a ζ' such that

$$p_m(\theta|\zeta') \propto p_m(y_t|\theta, s)p_m(\theta|\zeta), \qquad (4)$$

where the constant of proportionality is defined so that the right-hand side integrates to 1 (with respect to θ). We denote the value of ζ' defined by (4) by an update function u_s:

$$\zeta' = u_s(t, m, \zeta). \qquad (5)$$

We now give an example of such a changepoint model, which will be used throughout the paper to demonstrate and make concrete the ideas we present.

Example: Piecewise Polynomial Regression

We consider filtering a piecewise polynomial regression model to bi-variate data (x_i, y_i) for $i = 1, \ldots, n$, with the data ordered so that $x_1 < x_2 < \cdots < x_n$. For concreteness we will focus on piecewise quadratic models, but the extension to polynomials of different order is straightforward.

If the observations $\mathbf{y}_{s+1:t}$ are in the kth segment, we specify the model of (2) by:

$$\mathbf{y}_{s+1:t} = \mathbf{H}_k\beta_k + \varepsilon_k, \qquad (6)$$

where the design matrix $\mathbf{H}_k$ is of form

$$\mathbf{H}_k = \begin{pmatrix} 1 & 0 & 0 \\ 1 & x_{s+2} - x_{s+1} & (x_{s+2} - x_{s+1})^2 \\ \vdots & \vdots & \vdots \\ 1 & x_t - x_{s+1} & (x_t - x_{s+1})^2 \end{pmatrix},$$

ε_k is a vector of noises that are independently drawn from a $N(0, \sigma^2)$ distribution, and $\beta_k = (\beta_{k,0}, \beta_{k,1}, \beta_{k,2})$ is a vector-valued regression parameter.

For simplicity, we model the distance between successive changepoints as geometric with mean $1/p$, so $g(d) = p(1-p)^{d-1}$. For each segment except the first we allow for one of two models: $M = 1$ refers to the underlying curve being discontinuous at the changepoint that starts the segment, and $M = 2$ refers to the curve being continuous at this changepoint. Our prior is that the model of each segment is equally likely to be either possibility. Note

that if $M = 2$ then $\beta_{k,0}$ will be determined by the length and parameters of the previous segment.

We assume that σ^2 is common to all the segments. However, to be consistent with the above framework, we introduce σ_k^2 to denote its value in the kth segment. Thus we have that $\theta_k = (\sigma_k, \beta_k)$ and θ_k depends on θ_{k-1} as $\sigma_k = \sigma_{k-1}$, and if $M = 2$ through the dependence of $\beta_{k,0}$ on β_{k-1}.

We use the following standard conjugate priors for the variance σ_k^2 and the regression parameter β_k for both $M = 1, 2$:

$$
\begin{aligned}
\sigma_k^2 &\sim \text{IG}(\nu/2, \gamma/2), \\
\beta_k | \sigma_k^2 &\sim \text{N}(\mu, \sigma_k^2 \mathbf{D}),
\end{aligned}
\tag{7}
$$

where IG denotes the inverse Gamma distribution and N denotes the Gaussian distribution. With the notation above, we have $\zeta = (\nu, \gamma, \mu, \mathbf{D})$. For the first segment, for which $M_0 = 1$, we have prior parameter $\zeta_{0,1} = (\nu_0, \gamma_0, \mathbf{0}, \mathbf{D}_0)$, with $\mathbf{D}_0 = \text{diag}(\delta_0, \delta_1, \delta_2)$. For a future segment k with $M_k = 1$, we have the distribution for β_k given by (7) with $\mu = (0, 0, 0)$ and $\mathbf{D} = \mathbf{D}_0$. For a segment k with $M_k = 2$, and the previous segment starting with observation x_{r+1} and ends with observation x_s, the distribution for β_k is given by (7) with $\mu = (\beta_{k-1,0} + \Delta\beta_{k-1,1} + \Delta^2\beta_{k-1,2}, 0, 0)$, where $\Delta = (x_{s+1} - x_{r+1})$, and $\mathbf{D} = \text{diag}(0, \delta_1, \delta_2)$. This prior distribution ensures continuity of the underlying curve.

We can calculate $P_s(t, m, \zeta)$ and $u_s(t, m, \zeta)$ (see Equations 3 and 5) using standard updates for dynamic linear models (West and Harrison, 1989); details are given in the Appendix. Given the changepoint positions and segment models, we have a linear model for our data, and due to the choice of priors we can simulate directly from the posterior distribution of the parameters. The difficulty with Bayesian inference for this model is due to the intractability of the posterior distribution for changepoint positions and segment models.

3 Approximate Inference

We now describe our method for drawing, approximately, from the posterior distribution of the number and position of changements, and model and parameter values for each segment. The approach is based on recursive filtering and smoothing algorithms, which we will describe in turn. Throughout our description we will introduce a (potentially artificial) time, with observation y_t arriving at time t. For ease of presentation it will be useful to refer to the model and parameter values associated with the segment to which y_t belongs. Hence, for the rest of the paper we will slightly change notation, with θ_t and M_t refering to the parameter and model value at this time t. That is we will *subscript by time rather*

than by segment. We also introduce a new variable, C_t, which will denote the position of the most recent changepoint prior to time t.

3.1 Filtering Algorithm

To simplify the following exposition we will first derive the filtering algorithm for the case of a geometric segment length, $g(d) = p(1-p)^{d-1}$.

First note that (C_t, M_t, θ_t) are a Markov process; and in particular the marginal dynamics for C_t, M_t are given by

$$p(C_{t+1} = j, M_{t+1} = m | C_t = i, M_t = m') = \begin{cases} 1 - p & \text{if } j = i \text{ and } m = m', \\ p \Pr(M = m) & \text{if } j = t, \ m \in \mathcal{M}, \\ 0 & \text{otherwise.} \end{cases}$$

The top probability refers to there not being a changepoint between y_t and y_{t+1}, and the middle probability refers to the event that there is.

Now we wish to recursively approximate

$$p(C_t, M_t, \theta_t | y_{1:t}) = p(C_t, M_t | y_{1:t}) p(\theta_t | y_{1:t}, C_t, M_t).$$

The first term on the right-hand side is a discrete distribution, and we approximate $p(C_t = s, M_t = m | y_{1:t}) \approx w_t^{(s,m)}$. Whereas for given $C_t = s$ and $M_t = m$ we will approximate $p(\theta_t | y_{1:t}, C_t, M_t)$ by $p_m(\theta_t | \zeta_t^{(s,m)})$, for some $\zeta_t^{(s,m)}$. Our approximation is specified by the set of probabilities $w_t^{(s,m)}$ and parameters $\zeta_t^{(s,m)}$ for $s = 0, \ldots, t-1$ and $m \in \mathcal{M}$ the set of possible models.

We initiate our algorithm using the model prior, with $w_0^{(0,m)} = \Pr(M = m)$ for $m \in \mathcal{M}$, and prior for the parameters $\zeta_0^{(0,m)} = \zeta_{0,m}$. For $t = 1, \ldots, n$ we have the following set of recursions. Firstly for $s \in \{0, \ldots, t-1\}$, $C_{t+1} = s$ means that there is no changepoint at time t. Thus we have $M_{t+1} = M_t$ and $\theta_{t+1} = \theta_t$, and

$$p(C_{t+1} = s, M_{t+1} = m, \theta_{t+1} = \theta | y_{1:t+1}) = $$
$$Kp(C_t = s, M_t = m, \theta_t = \theta | y_{1:t}) \Pr(C_{t+1} = s | C_t = s) p_m(y_{t+1} | \theta),$$

for some normalising constant K. Now we substite our approximation, $p(C_t = s, M_t = m, \theta_{t+1} | y_{1:t}) \approx w_t^{(s,m)} p(\theta | \zeta_t^{(s,m)})$. Integrating with respect to θ gives $\Pr(C_t = s, M_t = m | y_{1:t+1})$, and thus

$$w_{t+1}^{s,m} = K w_t^{s,m} (1-p) P_s \left(t+1, m, \zeta_t^{(s,m)} \right).$$

While, using the updates for the conjugate distribution for θ we get

$$p(\theta_{t+1}|y_{1:t+1}, C_{t+1} = s, M_{t+1} = m) \propto p_m(\theta_{t+1}|\zeta_t^{(s,m)})p_m(y_{t+1}|\theta_{t+1}) = p_m(\theta_{t+1}|\zeta_{t+1}^{(s,m)}),$$

for $\zeta_{t+1}^{(s,m)} = u_s(t + 1, m, \zeta_t^{(s,m)})$.

Now consider $C_{t+1} = t$. This corresponds to a changepoint at time t, and C_t can take any value in $\{0, \ldots, t-1\}$. We derive an approximate recursion by considering

$$p(C_{t+1} = t, M_{t+1} = m, \theta_{t+1}|y_{1:t}) =$$
$$\sum_{s=0}^{t-1} \sum_{m' \in \mathcal{M}} p(C_t = s, M_t = m', \theta_t|y_{1:t}) \Pr(C_{t+1} = t|C_t = s) \Pr(M = m)p(\theta_{t+1}|\theta_t, t, s),$$

where $p(\theta_{t+1}|\theta_t, t, s)$ denotes the conditional distribution of θ_{t+1} given θ_t and that the previous segment contained observations $y_{s+1:t}$. Now, substituing our approximations to $p(C_t = s, M_t = m', \theta_t|y_{1:t})$ we have

$$p(\theta_{t+1}|y_{1:t}, C_{t+1} = t, M_{t+1} = m) \propto \sum_{s=0}^{t-1} \sum_{m' \in \mathcal{M}} pw_t^{(s,m')}p_m(\theta_t|\zeta_t^{(s,m')})p(\theta_{t+1}|\theta_t, t, s). \quad (8)$$

Our approach is to approximate this by $p_m(\theta_{t+1}|\zeta_t^{(t,m)})$ for some suitable choice of $\zeta_t^{(t,m)}$. Thus as

$$p(C_{t+1} = t, M_{t+1} = m, \theta_{t+1}|y_{1:t+1}) = Kp(C_{t+1} = t, M_{t+1} = m, \theta_{t+1}|y_{1:t})p_m(y_{t+1}|\theta_{t+1})$$
$$= Kp(C_{t+1} = t, M_{t+1} = m|y_{1:t})p(\theta_{t+1}|y_{1:t}, C_{t+1} = t, M_{t+1} = m)p_m(y_{t+1}|\theta_{t+1}),$$

we get the approximate recursion

$$w_{t+1}^{(t,m)} = K \Pr(M = m)P_t(t + 1, m, \zeta_t^{(t,m)}) \sum_{s=0}^{t-1} \sum_{m' \in \mathcal{M}} w_t^{(s,m')}p,$$

and $\zeta_{t+1}^{(t,m)} = u_t(t + 1, m, \zeta_t^{(t,m)})$.

Note that the only approximation in our filtering recursions is in the approximation of (8). There are various ways of choosing $\zeta_t^{(t,m)}$ for this approximation, and in practice we use a simple method of moments approach (see below). Note that this approximation is required to avoid the exponentially increasing computational cost of the exact filtering recursions. Similar approximations have been used in the Generalised Pseudo-Bayes algorithm (Tugnait, 1982), or the Interacting Multiple Model filter (Blom and Bar-Shalom, 1988).

Algorithm 1 Filtering Algorithm

Initiate Set $w_1^{(0,m)} = \Pr(M = m)P_s(1, m, \zeta_0^{(0,m)})$ and $\zeta_1^{(0,m)} = u_s(1, m, \zeta_0^{(0,m)})$ for $m \in \mathcal{M}$.
   Normalise weights, $w_1^{(0,m)}$ and let $t = 1$.

While $t < n$ (i) For $s = 0, \ldots, t - 1$ and $m \in \mathcal{M}$, set

$$w_{t+1}^{(s,m)} = \frac{1 - G(t + 1 - s)}{1 - G(t - s)} w_t^{(s,m)} P_s\left(t + 1, m, \zeta_t^{(s,m)}\right),$$

   and $\zeta_{t+1}^{(s,m)} = u_s(t + 1, m, \zeta_t^{(s,m)})$.
   (ii) For $m \in \mathcal{M}$, calculate $\zeta_t^{(t,m)}$ to produce the approximation to (8).
   (iii) For $m \in \mathcal{M}$, set

$$w_t^{(t,m)} = \Pr(M = m)P_t\left(t + 1, m, \zeta_t^{(t,m)}\right) \sum_{s=0}^{t-1} \sum_{m' \in \mathcal{M}} w_t^{(t,m')} \left(\frac{G(t + 1 - s) - G(t - s)}{1 - G(t - s)}\right),$$

   and $\zeta_{t+1}^{(t,m)} = u_t(t + 1, m, \zeta_t^{(t,m)})$.
   (iv) Normalise weights, $w_{t+1}^{(s,m)}$.

8

The full filtering algorithm, allowing for a general distribution of segment lengths and prior distribution for models is described in Algorithm 1.

Example Revisited

We now give details of step (ii) of the algorithm for the piecewise polynomial regression model. Remember $\zeta_t = (\nu_t, \gamma_t, \mu_t, \mathbf{D}_t)$. For a new segment with $M_{t+1} = 1$ we have $\mu_t = \mathbf{0}$ and $\mathbf{D}_t = \mathbf{D}_0$. We choose ν_t and γ_t to match moments of the predictive distribution of σ_{t+1}^{-2}.

Assume $\nu_t^{(s,m')}$ and $\gamma_t^{(s,m')}$ are the first two components of $\zeta_t^{(s,m')}$. Then we solve

$$\mathrm{E}(\sigma_{t+1}^{-2}) = \sum_{s=0}^{t-1} \sum_{m'=1}^{2} w_t^{(s,m')} \frac{\nu_t^{(s,m')}}{\gamma_t^{(s,m)}} = \frac{\nu_t^{(t,1)}}{\gamma_t^{(t,1)}}.$$

and

$$\mathrm{E}(\sigma_{t+1}^{-4}) = \sum_{s=0}^{t-1} \sum_{m'=1}^{2} w_t^{(s,m')} \frac{\nu_t^{(s,m')}(2 + \nu_t^{(s,m')})}{(\gamma_t^{(s,m)})^2} = \frac{\nu_t^{(t,1)}(2 + \nu_t^{(t,1)})}{(\gamma_t^{(t,1)})^2}.$$

for $\nu_t^{(t,1)}$ and $\gamma_t^{(t,1)}$.

For a new segment with $M_{t+1} = 2$, we have identical calculations for $\nu_t^{(t,2)}$ and $\gamma_t^{(t,2)}$. However, in this case we have $\mu_{t+1} = (\eta, 0, 0)$ and $\mathbf{D}_{t+1} = \mathrm{Diag}(\tau, \delta_1, \delta_2)$ for some η and τ to be calculated. Again we choose values based on matching moments, this time of $\beta_{t+1,0}$.

Let $\Delta_s = (x_{t+1} - x_{s+1})$, and $\mathbf{a}_s = (1, \Delta_s, \Delta_s^2)^T$, then

$$\mathrm{E}(\beta_{t+1,0}) = \sum_{s=0}^{t-1} \sum_{m'=1}^{2} w_t^{(s,m')} \mu_t^{s,m'} \mathbf{a}_s = \eta,$$

and

$$\mathrm{E}(\beta_{t+1,0}^2) = \sum_{s=0}^{t-1} \sum_{m'=1}^{2} w_t^{(s,m')} \left[\mathbf{a}_s^T \mathbf{D}_t^{(s,m')} \mathbf{a}_s + (\mu_t^{s,m'} \mathbf{a}_s)^2 \right] = \eta^2 + \tau.$$

3.2 Smoothing

Once we have calculated the filtering distributions for all t, we can simulate, backwards in time, the number and position of changepoints, the segment models and parameters, given the full data $y_{1:n}$.

Firstly, we can simulate (C_n, M_n, θ_n) from (our approximation to) $p(C_n, M_n, \theta_n | y_{1:n})$. These will give us the start of the final segment, together with its model and parameter values. Assume we simulate $C_n = t$, then we will next simulate (C_t, M_t, θ_t) from

$$p(C_t, M_t, \theta_t | y_{1:n}, C_{t+1} = t, C_{t+2:n}, M_{t+1:n}, \theta_{t+1:n}).$$

This will give us the start of the penultimate segment, its model and parameter values. We can then repeat this backwards in time until we simulate the first segment for our data.

To perform the simulation we use the fact that

$$p(C_t, M_t, \theta_t | y_{1:n}, C_{t+1:n}, M_{t+1:n}, \theta_{t+1:n}) = p(C_t, M_t, \theta_t | y_{1:t}, C_{t+1}, M_{t+1}, \theta_{t+1}),$$

by the conditional independence structure of the model. Thus we have

$$
\begin{aligned}
&p(C_t = s, M_t = m, \theta_t | y_{1:n}, C_{t+1} = t, C_{t+2:n}, M_{t+1} = m', M_{t+2:n}, \theta_{t+1:n}) \\
&= \; p(C_t = s, M_t = m, \theta_t | y_{1:t}, C_{t+1} = t, M_{t+1} = m', \theta_{t+1}) \\
&\propto \; p(C_t = s, M_t = m, \theta_t | y_{1:t}) p(C_{t+1} = t, M_{t=1} = m', \theta_{t+1} | C_t = s, M_t = m, \theta_t, y_{1:t}) \\
&\propto \; p(C_t = s, M_t = m, \theta_t | y_{1:t}) \Pr(C_{t+1} = t | C_t = s) p(\theta_{t+1} | C_{t+1} = t, M_{t+1} = m', C_t = s, M_t = m, \theta_t),
\end{aligned}
$$

where in the final step we have used that the model of a new segment is independent of the model of the preceeding segment.

To simplify notation, let $\mathcal{F}_t = \{y_{t+1:n}, C_{t+1:n}, M_{t+1:n}, \theta_{t+1:n}\}$ denote the future of the process after time t. Now substituting $p(C_t = s, M_t = m, \theta_t | y_{1:t}) = w_t^{(s,m)} p(\theta_t | \zeta_t^{(s,m)})$ we get

$$\Pr(C_t = s, M_t = m | y_{1:t}, \mathcal{F}_t) \propto w_t^{(s,m)} \Pr(C_{t+1} = t | C_t = s) \tag{9}$$

$$\int p(\theta_t | \zeta_t^{(s,m)}) p(\theta_{t+1} | C_{t+1} = t, M_{t+1} = m', C_t = s, M_t = m, \theta_t) \mathrm{d}\theta_t, \tag{10}$$

and

$$p(\theta_t | C_t = s, M_t = m, y_{1:t}, \mathcal{F}_t) \propto p(\theta_t | \zeta_t^{(s,m)}) p(\theta_{t+1} | C_{t+1} = t, M_{t+1} = m', C_t = s, M_t = m, \theta_t). \tag{11}$$

We need to be able calculate (or approximate) the integral in (10) and simulate from (11) to perform the smoothing. The full smoothing algorithm is given by Algorithm 2. The smoothing algorithm simulates the number and position of the changepoints, and the segment models and parameters. Often more accurate results can be obtained by throwing away the simulated parameter values, and re-simulating these from their conditional distribution given the changepoints and segment models (assuming this distribution is tractable). Such an approach is possible for our piecewise polynomial regression example, and is what we used in the simulation studies later.

Algorithm 2 Smoothing Algorithm

Initiate   1. Simulate (C_n, M_n) from the discrete distribution that gives probability $w_n^{(s,m)}$ to the value (s, m). Assuming $(C_n, M_n) = (s, m)$, then simulate θ_n from $p(\theta_n | \zeta_n^{(s,m)})$.

2. Set $t = s, m' = m$ and $\theta = \theta_n$.

While $t > 0$ 1. For $s = 0, \ldots, t-1$ and $m \in \mathcal{M}$ calculate

$$\tilde{w}^{(s,m)} = w_t^{(s,m)} \Pr(C_{t+1} = t | C_t = s)$$
$$\times \int p(\theta_t | \zeta_t^{(s,m)}) p(\theta_{t+1} = \theta | C_{t+1} = t, M_{t+1} = m', C_t = s, M_t = m, \theta_t) \mathrm{d}\theta_t$$

2. Simulate (C_t, M_t) from the discrete distribution that gives probability proportional to $\tilde{w}^{(s,m)}$ to the value (s, m).

3. Assume $(C_t, M_t) = (s, m)$. Simulate θ_t from the distribution proportional to

$$p(\theta_t | \zeta_t^{(s,m)}) p(\theta_{t+1} = \theta | C_{t+1} = t, M_{t+1} = m', C_t = s, M_t = m, \theta_t).$$

4. Set $t = s$, $m' = m$ and $\theta = \theta_{t+1}$.

We now give details of the calculations involved in the smoothing algorithm for our example.

Example Revisited

For our example $\theta_t = (\sigma_t, \beta_t)$. Consider a changepoint at t, and $C_t = s$. Define $\mathbf{h} = (1, \Delta, \Delta^2)$ and $\Delta = (x_{t+1} - x_{s+1})$.

Firstly consider calculating an integral of the form

$$\int p(\theta_t | \zeta) p(\theta_{t+1} | C_{t+1} = t, M_{t+1} = m', C_t = s, M_t = m, \theta_t) d\theta_t,$$

where $\zeta = (\nu, \gamma, \mu, \mathcal{D})$, for step 1 of Algorithm 2. For $m' = 1$ this becomes

$$\mathrm{IG}(\sigma_{t+1}; \nu/2, \gamma/2) \mathrm{N}(\beta_{t+1}; \mathbf{0}, \sigma_{t+1}^2 \mathbf{D}_0),$$

where $\mathrm{IG}(x; a, b)$ denotes the probability density function (pdf) of an inverse-gamma distribution with parameter a and b, evaluated at x; and $\mathrm{N}(\mathbf{x}; \eta, \Sigma)$ denotes the pdf of a multivariate normal distribution with mean μ and variance Σ, evaluated at $\mathbf{x}$. The first term comes from the fact that $\sigma_{t+1} = \sigma_t$, and second due to the independence of β_{t+1} and β_t. For $m' = 2$, the integral becomes

$$\mathrm{IG}(\sigma_{t+1}; \nu/2, \gamma/2) \mathrm{N}(\beta_{t+1}; \eta, \sigma_{t+1}^2 \Sigma),$$

where $\eta = (\mathbf{h}\mu^T, 0, 0)$, and $\Sigma = \mathrm{Diag}(\mathbf{h}^T \mathbf{D} \mathbf{h}, \delta_1, \delta_2)$. Here we the term has changed as now $\beta_{t+1,0} = \mathbf{h}\beta_t$ due to continuity.

Now consider simulating θ_t from a density proportional to

$$p(\theta_t | \zeta) p(\theta_{t+1} | C_{t+1} = t, M_{t+1} = m', C_t = s, M_t = m, \theta_t),$$

in step 3 of Algorithm 2. For $m' = 1$ we set $\sigma_t = \sigma_{t+1}$, and simulate β_t from a multivariate normal distribution with mean μ and variance $\sigma_{t+1}^2 \mathbf{D}$. For $m' = 2$ we again set $\sigma_t = \sigma_{t+1}$, but now simulate β_t from a multivariate normal distribution with mean μ and variance $\sigma_{t+1}^2 \mathbf{D}$ conditional on $\mathbf{h}\beta_t^T = \beta_{t+1,0}$. Standard results (see e.g. Rue and Held, 2005), gives that we simulate β_t from a multivariate normal with mean

$$\mu - \mathbf{D}\mathbf{h}^T (\mathbf{h}\mathbf{D}\mathbf{h}^T)^{-1} (\mathbf{h}\mu^T - \beta_{t+1,0}),$$

and variance

$$\sigma_{t+1}^2 \left(\mathbf{D} - \mathbf{D}\mathbf{h}^T (\mathbf{h}\mathbf{D}\mathbf{h}^T)^{-1} \mathbf{h}^T \mathbf{D} \right).$$

3.3   Resampling

Simulating from the posterior distribution of the number and position of changepoints, and the segment models and parameters, using the filtering and smoothing algorithms has a complexity which is quadratic in n. This is due to the number of support points of (C_t, M_t) increasing linearly with t.

At the expense of further approximation, we can develop an algorithm whose total computational cost is linear in n via using particle-filter resampling algorithms (e.g. Liu *et al.*, 1998; Fearnhead and Clifford, 2003) to approximate the distributions of (C_t, M_t) by discrete distributions with fewer support points. (The resampling procedures ensure that the number of support points in the resulting approximation is bounded by a constant for all t.) This was investigated in Fearnhead and Liu (2007), who propose two optimal resampling algorithms for changepoint models, and show that substantial computational savings can be obtained with negligible approximation error.

4   Simulation Study

We now evaluate out method through a simulation study using the piecewise quadratic model introduced within our example. We first look at the accuracy of our filtering and smoothing method for simulating from the posterior distribution, and then compare the accuracy of our method to other approaches for curve-fitting. Finally we look at the accuracy of our method at inferring discontinuities in the underlying curve.

In implementing our method we used the filter and smoothing algorithms with the stratified rejection control resampling method of Fearnhead and Liu (2007). The threshold parameter within the resampling algorithm was set to 10^{-6} (see Fearnhead and Liu, 2007, for details). We used the filter and smoothing algorithms to simulate the number and position of changepoints, the value of the observation variance and the model for each segment. Conditioned on these, we then simulated the β values associated with each segment from their conditional distribution.

The filtering and smoothing algorithms were implemented within `C++` and `R`. The computational cost of the algorithms is roughly linear in the number of observations, and to run them on a data set with 4000 data points took of the order of 10 seconds on a desktop PC.

4.1   Accuracy of the Simulation Method

To test the accuracy of the filtering and smoothing algorithms at drawing samples from the true posterior distribution, we ran a simulation study where we simulated data under the exact model that we used for analysis. We then calculated the posterior quantiles of the true value for σ and the true value of the underlying curve at each time point. The rationale is that if we could draw from the true posterior, then these posterior quantiles should be uniformly distributed on $[0, 1]$. Any inaccuracies in our simulation method will be demonstrated through deviations of the posterior quantiles from such a uniform distribution.

We simulated data for the piecewise-quadratic model with $\mathbf{D}_0 = \text{Diag}(1, 10^2, 40^2)$, $p = 4/n$ and $\sigma^2 = 1$. We analysed the data under the model with the same value for $\mathbf{D}_0$ and p, but with an improper prior for σ^2 (equivalent to $\nu = \gamma = 0$). To detect any affect that the amount of data had on the performance of our method we simulated 100 data sets for each of $n = 256$, 512 and 1024. In each case we used equally spaced x_t points in $[0, 1]$.

Plots of the posterior quantiles are shown in Figure 1. In both cases they are close to that expected if they were drawn from the true posterior distribution. The extra smoothness in the plot of posterior quantiles of the underlying curve is due to the larger number of quantiles obtained in this case, $100n$ as we obtain one quantile for each data point. For the posterior quantiles of σ we are able to construct confidence intervals, as the posterior quantiles are independent. We notice that the observed quantiles generally lie within the plotted 90% confidence interval. Taken together, these results suggest that negligible error is being introduced by the approximations in our method for simulating from the posterior distribution.

4.2   Comparison for curve-fitting

We now look at the accuracy of our piecewise quadratic regression model, together with the new simulation method, for curve-fitting. Firstly, in order to implement our method we need to choose the prior parameter values. As above we will use the default uninformative prior for σ. The values of $\mathbf{D}_0$ and p we used for inference were chosen in an iterative procedure (as suggested in Fearnhead, 2005). We did a preliminary analysis of the data (using default choices for $\mathbf{D}_0$ and p), and then estimated $\mathbf{D}_0$ and p from the posterior distribution of the βs and the number of changepoints. If necessary this could be repeated, with simulation from the posterior given the latest estimates for $\mathbf{D}_0$ and p, and new estimates of $\mathbf{D}_0$ and p obtained. In practice we did not find it necessary to repeat the procedure.
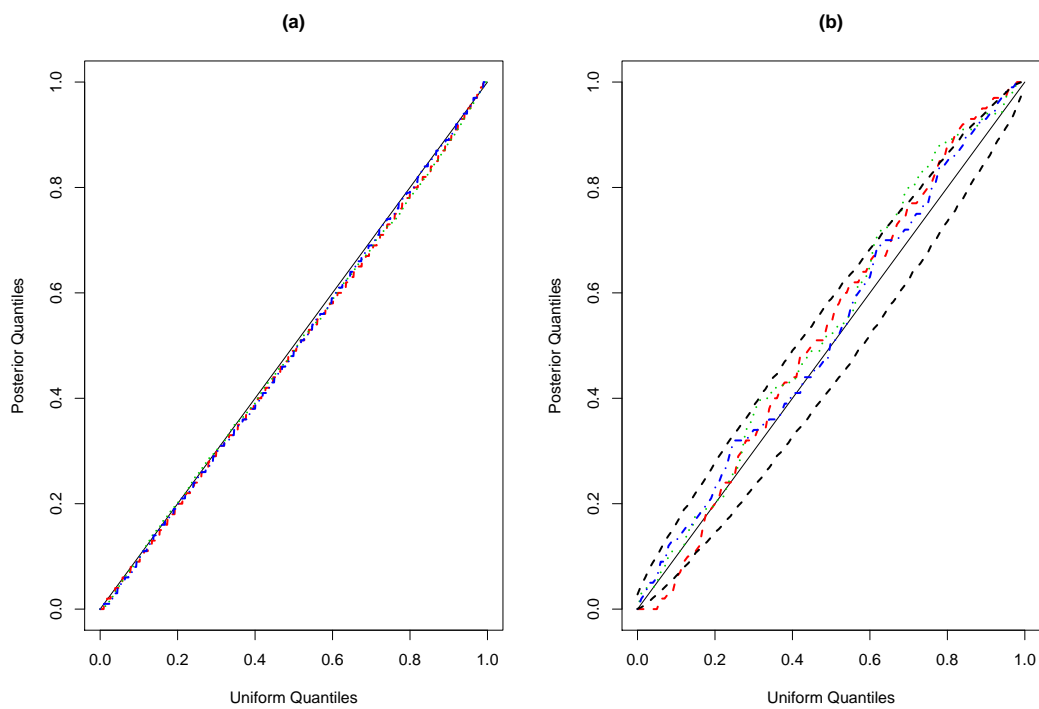
Figure 1: Posterior quantile plots of (a) the underlying curve and (b) σ^2 for different values of n: 256 (red, dashed line), 512 (green dotted line) and 1024 (blue dot-dashed line). For (b) we give 90% confidence intervals obtained through simulation (black dashed line).

n	MSE			Coverage	
	New	BAYES.THR	cthresh	New	wave.band
256	0.056	0.215	0.15	0.87	0.79
512	0.027	0.138	0.093	0.87	0.79
1024	0.014	0.087	0.056	0.89	0.79

Table 1: Mean square error (MSE) and coverage of putative 90% confidence/credible intervals for our new method, and wavelet based methods.

We first quantify the accuracy of our method for analysing the same simulated data sets that were used in Section 4.1. For a given data set let z_t denote the value of the underlying curve at time t (so observations are $y_t = z_t + \sigma \epsilon_t$ where ϵ_t is a standard normal random variable). Denote by $\hat{z}_t$ an estimate of z_t, then we estimate the accuracy of an estimate of the curve $z_{1:n}$ by the average mean square error

$$\text{MSE} = \frac{1}{n} \sum_{t=1}^{n} (z_t - \hat{z}_t)^2.$$

For our method we use the posterior mean as our estimate of z_t. We also look at the mean point-wise coverage of 90% credible (or confidence) intervals for z_t.

For comparison we estimate the underlying curve using wavelets. We implement two wavelet methods, that of Abramovich *et al.* (1998) implemented using the BAYES.THR function in R, and one using complex wavelest (Barber and Nason, 2004) implemented using the cthresh function in R. We also constructed wavelet-based confidence intervals (Barber *et al.*, 2002) using the wave.band function in R. (See http://www.stats.bris.ac.uk/~wavethresh/ for details of these functions; we used default settings for the R functions in all cases.)

Results for the simulated data described in Section 4.1 are given in Table 1. We notice that the MSE for estimates of the underlying curve is substantially smaller for our new approach than for either wavelet method. Of the two wavelet methods, the one using complex wavelets gives superior performance. The MSE of our new method halves each time n is doubled, whereas the MSE of the wavelet methods decreases by a smaller proportion each time. Finally, the coverage of our 90% credible intervals are close to 90% in each case. The fact that the coverage of the intervals is less than their putative size is likely to be down to errors in estimating the hyperparameters.

The above results are not surprising, in the sense that the data was simulated under the model assumed by our new method. To test robustness of this method to data being simulated from an alternative model, we repeated our simulation study but with data simulated under a piecewise cubic model. For this model we set $\mathbf{D}_0 = \text{diag}(1, 10^2, 40^2, d^2)$,

		MSE			Coverage
d	New	BAYES.THR	cthresh	New	wave.band
100	0.06	0.34	0.16	0.86	0.80
200	0.07	0.69	0.17	0.86	0.82
400	0.11	2.45	0.18	0.84	0.86

Table 2: Mean square error (MSE) and coverage of putative 90% confidence/credible intervals for our new method, and wavelet based methods. Data simulated under a piecewise cubic model, with d affecting the size of the cubic co-efficients. All data sets were simulated with $n = 100$.

and considered the effect that d had. Note that the expected value of the modulus of the cubic co-efficient is $d(2/\pi)^{1/2}$. For simplicity we fixed $n = 256$ for all simulations that we carried out.

Results are given in Table 2, again based on 100 simulated data sets for each set of parameters. As expected, as d increases, which corresponds to an increasingly non-quadratic components of the underlying curve, the performance of the new method deteriorates. This is both in terms of the coverage properties of the credible intervals, and the mean square error of estimates of the underlying curve. However for all values of d we considered, the new method still substantially out-performs both wavelet methods in terms of estimating the underlying curve.

As a final comparison, we applied our new method to various test data sets from the literature, and compare our method with the published results of Denison *et al.* (1998) (henceforth DMS). The test data sets used are shown in Figure 2, and consist of the Heavisine, Blocks, Bumps and Doppler signals of Donoho and Johnstone (1994); and the smooth function (a) and (b) from Denison *et al.* (1998) (denoted DMS A amd DMS B). The method of Denison *et al.* (1998) uses a reversible jump MCMC to fit a piecewise cubic function, under continuity and differentiability constraints. The MCMC algorithm samples from an approximation to the posterior, based on approximating the marginal likelihood for each segment. The MCMC procedure takes up to about an order of magnitude longer to analyse the data than our approach.

We compare methods based on MSE as before. Results are given in Table 3. Our method does considerably better at estimating the curves which contain discontinuities, as our model allows for discontinuities in the underlying curve. While we do similarly or better on DMS A and DMS B, our method is substantially worse for the Bumps and Doppler data sets. This is due to errors in estimating the peaks in the Bumps data set, and the initial part of the curve in the Doppler data set. In both cases these are where the underlying curve
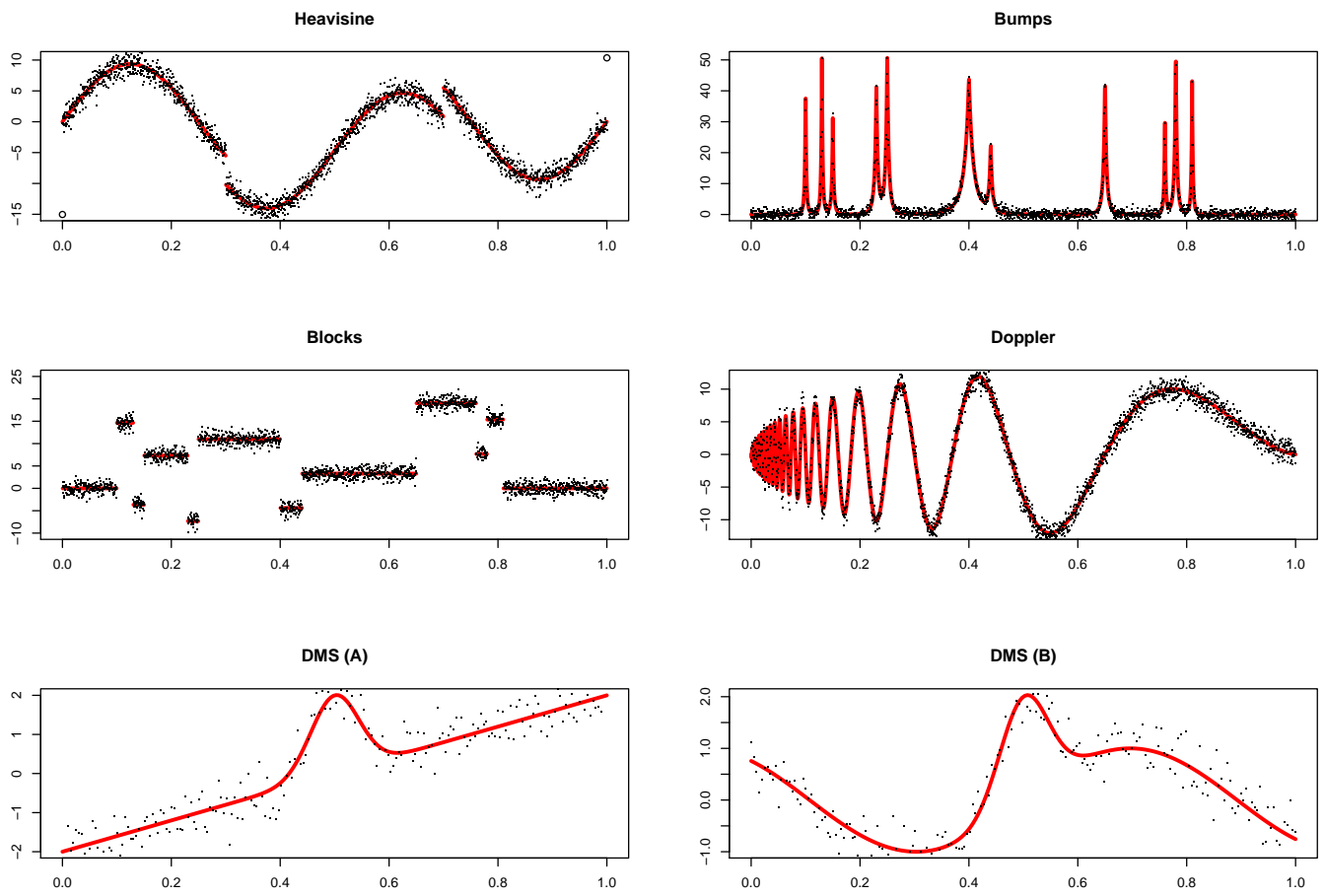
Figure 2: Simulated data sets used for comparison with method of Denison *et al.* (1998)

	n	σ	SNR	DMS	NEW
Heavisine	2048	1.0	7	0.033	0.022
Blocks	2048	1.0	7	0.170	0.016
Bumps	2048	1.0	7	0.167	0.318
Doppler	2048	1.0	7	0.135	0.198
DMS A	200	0.4	3	0.010	0.010
DMS B	200	0.3	3	0.009	0.006

Table 3: MSE results for 7 test data sets (see Figure 2). For each data set we give the number of data points, n, the observation error, σ, and the signal-to-noise ratio. MSE results for DMS are taken from Denison *et al.* (1998).

changes most rapidly. One explanation for this is that using only quadratic polynomials, rather than cubic, makes it harder for our model to fit these parts of the curve.

4.3 Power at detecting discontinuities

Finally we look at the power of our method for detecting discontinuities in the underlying curve. Note that it is only our method that can potentially distinguish between change-points at which the underlying curve may be either continuous or discontinuous. We focus on this feature of our method due to the application of the method we consider in Section 5.

We used as a basis the continuous curve in DMS B (see Figure 2). We then introduced a discontinuity into the curve. If we denote the underlying DMS B curve by $f(x)$ for $x \in [0, 1]$, then we introduce a changepoint of size c at point x_c to produce the curve:

$$f(x; c, x_c) = \left\{ \begin{array}{ll} f(x) - c\sigma & \text{for } x < x_c, \\ f(x) & \text{for } x \geq x_c, \end{array} \right.$$

where σ^2 is the variance of the observations. We then simulated data centered on this curve, and look at the posterior probability of a discontinuous changepoint at between $[x_c - 0.01, x_c + 0.01]$. We repeated this for different values of c, x_c and sample size n.

Results are given in Figure 3. As expected the posterior probability of a changepoint increases with both n and c, and to a lesser extent by the position of the changepoint. The lowest posterior probability of a changepoint occurs when $x_c = 0.45$, which is the point at which the gradient of the signal is greatest, and this makes jumps in the signal harder to infer. In general an average posterior probability of a changepoint of greater than 0.5 can occurs with $c > 3$ when n is 200 or more; and when $c > 2$ and $n = 800$.
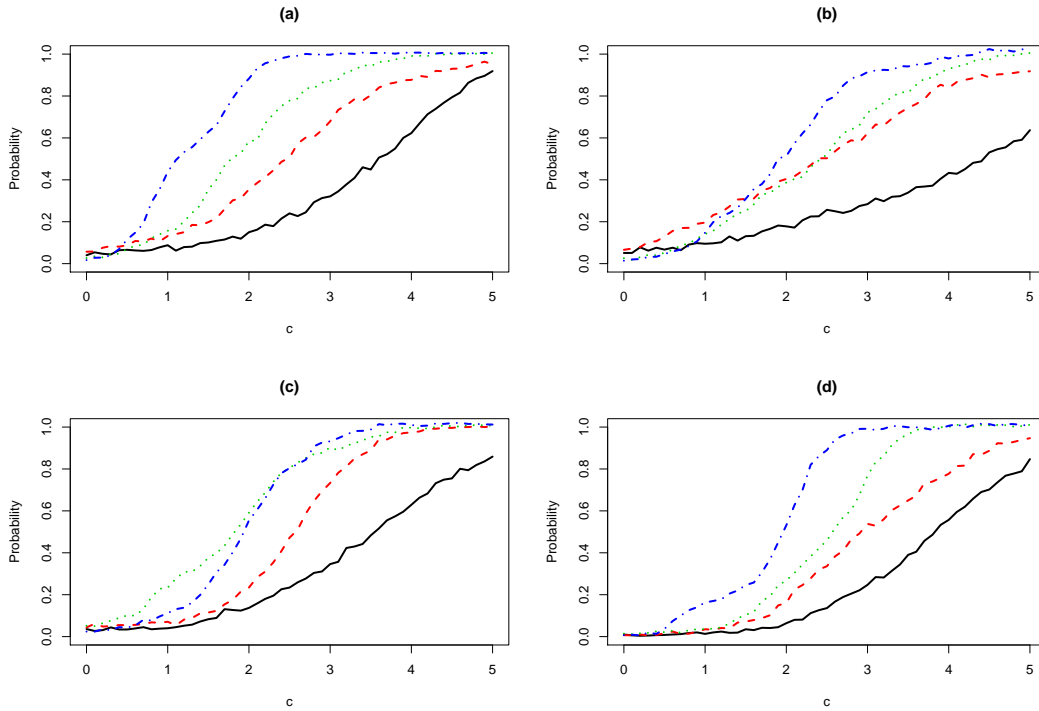
19

Figure 3: The posterior probability of a changepoint within $[x_c - 0.01, x_c + 0.01]$ for DMS A, for different changepoint positions x_c, size of changepoint c and number of observations n. Figure (a) is $x_c = 0.3$, (b) is $x_c = 0.45$, (c) is $x_c = 0.6$ and (d) is $x_c = 0.7$. For each plot the lines correspond to different values of n: $n = 100$ (black full line); $n = 200$ (red dashed line); $n = 400$ (green dotted line); and $n = 800$ (blue dot-dashed line).

5 Well-log Data

We now apply our method to analyse the well-log data of Ó Ruanaidh and Fitzgerald (1996). The data is shown in Figure 4, and consists of a time-series of measurements of rock as a probe is lowered through a bore-hole in the earth's surface. We have scaled time so that time-series is over the interval $[0, 1]$. The underlying signal has a number of abrupt changes, due to the changes in rock strata. It is of interest to locate these abrupt changes in the signal. See Ó Ruanaidh and Fitzgerald (1996) and Fearnhead and Clifford (2003) for further discussion of this data set, and the practical importance of detecting changes in rock strata. Furthermore Fearnhead and Clifford (2003) discuss the need for online methods for analysing data of this type.

Both Ó Ruanaidh and Fitzgerald (1996) and Fearnhead and Clifford (2003) fit a piecewise constant signal to the data and assume observation error is independent over time. However, Fearnhead (2006) suggests that such a model is inappropriate as it ignores local variation within segments, and fitting such a model results in the detection of too many changepoints. Thus here we will consider analysing the data under our model. The idea is that our model is flexible to allow for variation within rock strata through changepoints at which the underlying signal is continuous. Changes in rock strata will correspond to changepoints at which the underlying signal is discontinuous. Our interest is thus in detecting the position of these discontinuous changepoints.

As in Ó Ruanaidh and Fitzgerald (1996) we first remove outliers from the data, and then analyse the data in batch. We consider two analyses, one allowing for the possibility of changepoints at which the underlying signal is either continuous of discontinuous; and the other which only allows changepoints where the underlying signal is discontinuous. The latter mimics the models of Ó Ruanaidh and Fitzgerald (1996) and Fearnhead and Clifford (2003). We call these models, model A and model B respectively.

Results are given in Figure 4. For each model we plot the posterior probability of a discontinuity of the signal in an interval $[t - 0.001, t + 0.001]$ for different values of t. For simplicity we infer a discontinuity whenever this probability is greater than 0.5, and plot the inferred changepoints for the two models. Model B appears to overfit discontinuities in the data (posterior mean number of discontinuities, 30, is nearly twice that for model A), and using our simple procedure for highlighting changepoints, infers an extra three discontinuities in the data – which by eye look spurious.
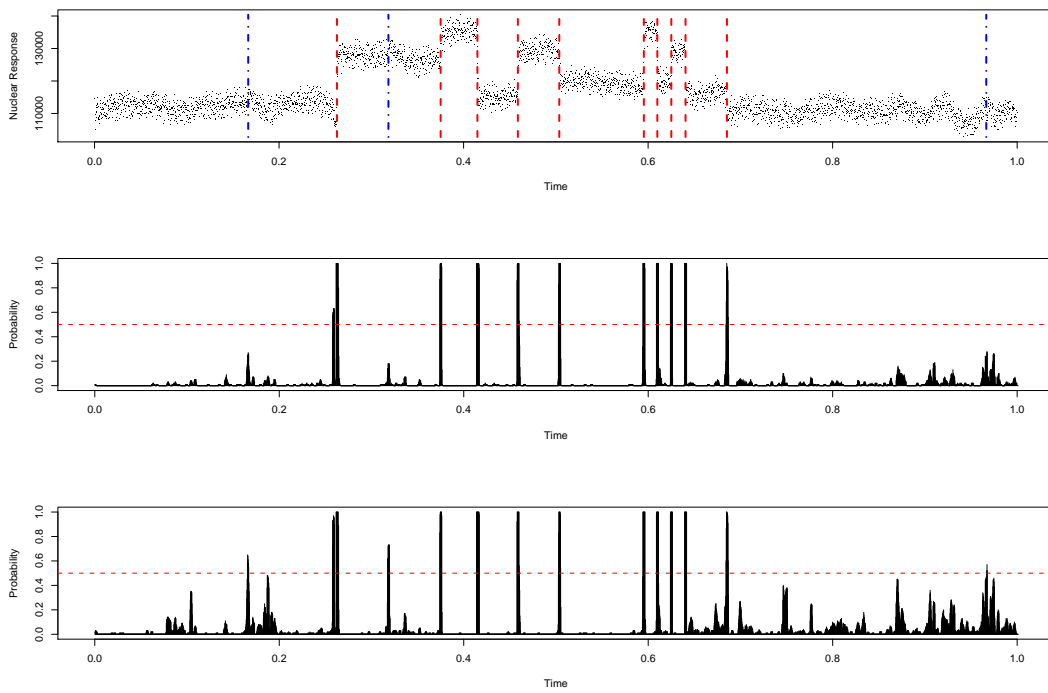
Figure 4: (Top) Raw well-log data, with inferred discontinuities (red dashed vertical lines: both models; blue dot-dashed line: model B only). (Middle and Bottom) Posterior probability of discontinuity at time $[t - 0.001, t + 0.001]$ for model A and B respectively.

6 Discussion

We have presented a novel and computationally efficient procedure for Bayesian inference for changepoint models, where there is Markov dependence in the segment parameters. The method is approximate, in that it is based on an approximation to the filtering distribution of parameters associated with a new segment. When used with the resampling idea of Section 3.3 the resulting algorithm has computational and storage costs that are linear in the number of observations. The simulation results in Section 4.1, showed that, for the examples we considered, the error introduced by our approximations were negligible.

We demonstrated the potential of this new procedure through the fitting of piece-wise quadratic functions. The model we fit allowed for both the possibility of continuity of discontinuity at changepoints. Our simulation studies showed that this model is more accurate at fitting curves that contain discontinuities than the related method of Denison *et al.* (1998), and can also perform better at estimating the underlying curve than wavelet procedures. Further advantages of our approach is that it can allow for online inference, and also can allow for inference about discontinuities in the underlying signal.

Appendix

Here we give details of $P_s(t, m, \zeta)$ and $u_s(t, m, \zeta)$ for the piecewise polynomial regression. Now denote $\mathbf{h} = (1, \Delta, \Delta^2)$ where $\Delta = (x_t - x_{s+1})$ (suppressing the dependence on s and t). Then given the most recent changepoint is at time s, the mean of the observation at time t is $\mathbf{h}\beta_t^T$.

Remember $\zeta = (\nu, \gamma, \mu, \mathbf{D})$, and define $\zeta' = (\nu', \gamma', \mu', \mathbf{D}')$. Define $e = y_t - \mathbf{h}\mu^T$, $Q = \mathbf{h}\mathbf{D}\mathbf{h}^T + 1$, and $\mathbf{A} = \mathbf{D}\mathbf{h}^T/Q$. Then if $\zeta' = u_s(t, m, \zeta)$, we get

$$\nu' = \nu + 1,$$
$$\gamma' = \gamma + e^2/Q,$$
$$\mu' = \mu + \mathbf{A}e,$$
$$\mathbf{D}' = \mathbf{D} - \mathbf{A}^T\mathbf{A}Q.$$

Furthermore, let $T_d(x; a, R)$ denote the density of a student's t random variable d degrees of freedonm, and with mean a and scale parameter R. Then we have

$$P_s(t, m, \zeta) = T_\nu(y_t; \mathbf{h}\mu^T, \gamma Q/\nu).$$

References

Abramovich, F., Sapatinas, T. and Silverman, B. W. (1998). Wavelet thresholding via a Bayesian approach. *Journal of the Royal Statistical Society, Series B* **60**, 725–749.

Barber, S. and Nason, G. P. (2004). Real nonparametric regression using complex wavelets. *Journal of the Royal Statistical Society, Series B* **66**, 927–939.

Barber, S., Nason, G. P. and Silverman, B. W. (2002). Posterior probability intervals for wavelet thresholding. *Journal of the Royal Statistical Society, Series B* **64**, 189–205.

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

Blom, H. A. P. and Bar-Shalom, Y. (1988). The interacting multiple model algorithm for systems with Markovian switching coefficients. *IEEE Transactions on Automatic Control* **33**, 780–783.

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.

Denison, D. G. T., Mallick, B. K. and Smith, A. F. M. (1998). Automatic Bayesian curve fitting. *Journal of the Royal Statistical Society, series B* **60**, 333–350.

DiMatteo, I., Genovese, C. R. and Kass, R. E. (2001). Bayesian curve-fitting with free-knot splines. *Biometrika* **88**, 1055–1071.

Dobigeon, N. and Toumeret, J. Y. (2007). Joint segmentation of wind speed and direction using a hierarchical model. *Computational Statistics and Data Analysis* **51**, 5603–5621.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425–455.

Fearnhead, P. (2005). Exact Bayesian curve fitting and signal segmentation. *IEEE Transactions on Signal Processing* **53**, 2160–2166.

Fearnhead, P. (2006). Exact and efficient inference for multiple changepoint problems. *Statistics and Computing* **16**, 203–213.

Fearnhead, P. (2008). Computational methods for complex stochastic systems: A review of some alternatives to MCMC. *Statistics and Computing* **18**, 151–171.

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. (2007). Online inference for multiple changepoint problems. *Journal of the Royal Statistical Society Series B* **69**, 589–605.

Fearnhead, P. and Vasileiou, D. (2008). Bayesian analysis of isochores. *To appear in Journal of the American Statistical Association* Available from `www.maths.lancs.ac.uk/~fearnhea/publications`.

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. and Lawrence, C. E. (1999). Bayesian inference on biopolymer models. *Bioinformatics* **15**, 38–52.

Liu, J. S., Chen, R. and Wong, W. H. (1998). Rejection control and sequential importance sampling. *Journal of the American Statistical Society* **93**, 1022–1031.

McVean, G. A. T., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R. and Donnelly, P. (2004). The fine-scale structure of recombination rate variation in the human genome. *Science* **304**, 581–584.

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

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.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. CRC Press/Chapman and Hall.

Seidou, O. and Ouarda, T. B. M. J. (2007). Recursion-based multiple changepoint detection in multiple linear regression and application to river streamflows. *Water Resources Research* **43**, W07404.

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

Tugnait, J. K. (1982). Detection and estimation for abruptly changing systems. *Automatica* **18**, 607–615.

West, M. and Harrison, J. (1989). *Bayesian forecasting and dynamic models*. Springer-Verlag, New York.

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