Exact Bayesian Curve Fitting and Signal Segmentation

Paul Fearnhead

***Abstract*—We consider regression models where the underlying functional relationship between the response and the explanatory variable is modeled as independent linear regressions on disjoint segments. We present an algorithm for perfect simulation from the posterior distribution of such a model, even allowing for an unknown number of segments and an unknown model order for the linear regressions within each segment. The algorithm is simple, can scale well to large data sets, and avoids the problem of diagnosing convergence that is present with Monte Carlo Markov Chain (MCMC) approaches to this problem. We demonstrate our algorithm on standard denoising problems, on a piecewise constant AR model, and on a speech segmentation problem.**

***Index Terms*—Changepoints, denoising, forward-backward algorithm, linear regression, model uncertainty, perfect simulation.**

I. Introduction

A. Overview

REGRESSION problems are common in signal processing. The aim is to estimate, from noisy measurements, a functional relationship between a response and a set of explanatory variables. We consider the approach of [1], who model this functional relationship as a sequence of linear regression models on disjoint segments. Both the number and position of the segments and the order and parameters of the linear regression models are to be estimated. A Bayesian approach to this inference is taken.

In [1], Bayesian inference is performed via the reversible jump MCMC methodology of [2]. We consider applying recently developed perfect simulation ideas [3] to this problem. These ideas are closely related to the Forward–Backward algorithm [4] and methods for product partition models [5], [6]. To define segments, we need to assume that the response can be ordered linearly through "time." (Whereas time may be artifical, in estimating a polynomial relationship between a response and an explanatory variable, time can be defined so that the order that responses are observed is in increasing value of the explanatory variable.) The perfect simulation algorithm requires independence between segments and utilizes the Markov property of changepoint models in such cases. It involves a recursion for the probability of the data from time t onwards, conditional on a changepoint immediately before time t, given similar quantities for all times after t. Once these probabilities have been calculated for all t, simulating from the posterior distribution of the number and position of the changepoints is straight forward.

Manuscript received August 13, 2003; revised July 27, 2004. The associate editor coordinating the review of this manuscript and approving it for publication was Zixiang Xiong.

The author is with the Department of Mathematics and Statistics, Lancaster University, Lancaster, LA1 4YF U.K. (e-mail: p.fearnhead@lancaster.ac.uk).

Digital Object Identifier 10.1109/TSP.2005.847844

This approach to perfect simulation is different from the more common approach based on coupling from the past [7], which has been used, for example, on the related problem of reconstructing signals via wavelets [8].

We develop the existing methodology by allowing for model uncertainty within each segment. We also implement a Viterbi version of the algorithm to perform maximum *a posteriori* (MAP) estimation. The advantages of our approach over MCMC are that we have the following.

i) The perfect simulation algorithm draws independent samples from the true posterior distribution and avoids the problems of diagnosing convergence that occur with MCMC.
ii) The recursions of the algorithm are generic and simpler than designing efficient MCMC moves.
iii) The computational cost of the algorithm can scale linearly with the length of data analyzed, thus making it applicable for large data sets.

The first advantage is particularly important. There are a number of examples of published results from MCMC analyses that have later proven to be inaccurate because the MCMC algorithm had not been run long enough (for example, compare the results of [9] and [10] or those of [2] with those of [11]). Our approach avoids this problem by enabling iid draws from the true posterior (the goal of Bayesian inference). Thus, it can be viewed as enabling "exact Bayesian" inference for these problems.

In Bayesian inference, the posterior distribution depends on the choice of prior distribution. When inference is made conditional on a specific model, uninformative priors can then be chosen so that the posterior reflects the information in the data. The regression problem we address involves model choice, and for such problems, uninformative priors do not exist, as the choice of prior affects the Bayes factor between different competing models. The use of uninformative priors for the parameters can severely penalize models with larger numbers of parameters (see [12] for more details).

One approach to choosing priors for inference problems that include model uncertainty is to let the data inform the choice of prior [1], [13], for example, by using a hierarchical model with hyperpriors on the prior parameters [14]. However, the inclusion of hyperpriors on the regression parameters violates the independence assumption required for perfect simulation. We suggest two possible approaches for choosing the prior parameters. The simpler is based on a recursive use of the perfect simulation algorithm, with the output of a preliminary run of the algorithm being used to choose the prior parameters. Alternatively, if hyperpriors are used, the perfect simulation algorithm can then be

incorporated within a simple MCMC algorithm, which mixes over the hyperparameters.

The outline of the paper is as follows. In Section II we describe our modeling assumptions together with the methodology for exact simulation and MAP estimation; we also describe how to use the exact simulation algorithm within MCMC for the case of Bayesian inference with hyperpriors. In Section III, we demonstrate our method on standard denoising problems and on speech segmentation.

II. Model and Method

A. Model

Our model is based on that of [1]. We assume we have n observations $\mathbf{y}_{1:n} = (y_1, \ldots, y_n)$. Throughout, we use the notation $\mathbf{y}_{i:j}$ to denote $(y_i, \ldots, y_j)$, which is the ith to jth entries of the vector $\mathbf{y}$. Given m segments, defined by the ordered changepoints $\tau_0, \tau_1, \ldots, \tau_m$, with $\tau_0 = 0$ and $\tau_m = n$, we model the observations $\mathbf{y}_{(\tau_i+1):\tau_{i+1}}$, which are associated with the ith segment by a linear regression of order p_i. Denote the p_i parameters by the vector $\boldsymbol{\beta}_i$, and the matrix of basis functions by $\mathbf{G}_i^{(p_i)}$. Then, we have

$$\mathbf{y}_{(\tau_i+1):\tau_{i+1}} = \mathbf{G}_i^{(p_i)}\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_{(\tau_i+1):\tau_{i+1}}$$

where $\boldsymbol{\epsilon}_{(\tau_i+1):\tau_{i+1}}$ is a vector of independent and identically distributed (iid) Gaussian random variables with mean 0 and variance σ_i^2. For examples, see the polynomial regression model of Section III-A or the auto regression model of Section III-B.

The number and positions of the changepoints and the order, parameters, and variance of the regression model for each segment are all assumed to be unknown. We introduce conjugate priors for each of these.

The prior on the changepoints is given by

$$p(m, \tau_1, \ldots, \tau_{m-1}) = \lambda^{m-1}(1-\lambda)^{n-m}$$

for $m = 0, \ldots, n-1$ and for $0 < \tau_1 < \cdots < \tau_{m-1} < n$, for some probability λ. For the jth regression parameter of the ith segment $\beta_{i,j}$, we have a normal prior with mean 0 and variance $\sigma_i^2\delta_j^2$, independent of all other regression parameters. We assume an Inverse-Gamma prior for the noise variances σ_i^2 with parameters $\nu/2$ and $\gamma/2$. The priors are independent for different segments. Finally, we constrain $p_i \leq \bar{p}$ and introduce an arbitrary discrete prior $p(p_i)$, again assuming independence between segments.

We now describe how perfect simulation can be performed for this model. We then discuss how the data can be used to choose the prior parameters of the model $(\lambda, \nu, \gamma, \delta_1, \ldots, \delta_{\hat{p}})$.

B. Perfect Simulation

1) Recursions: Define for $t = 2, \ldots, n$

$$Q(t) = \Pr(\mathbf{y}_{t:n} | \text{changepoint at } t-1)$$

and $Q(1) = \Pr(\mathbf{y}_{1:n})$. The model of Section II-A has a Markov property that enables $Q(t)$ to be calculated in terms of $Q(s)$, for $s > t$, by averaging over the position of the next changepoint after t.

Consider a segment with observations $\mathbf{y}_{t:s}$ for $s \geq t$ and a linear regression model order q. Let G be the $(s-t+1) \times q$ matrix of basis vectors for the qth-order linear regression model on this segment. Let $\mathbf{D} = \text{Diag}\left(\delta_1^2, \ldots, \delta_q^2\right)$ be the prior variance on the regression parameters for this segment, and let $\mathbf{I}$ be the $(s-t+1) \times (s-t+1)$ identity matrix. Define

$$\mathbf{M} = (\mathbf{G}^T\mathbf{G} + \mathbf{D})^{-1}\mathbf{P} = (\mathbf{I} - \mathbf{G}\mathbf{M}\mathbf{G}^T)$$

and

$$\|\mathbf{y}\|_{\mathbf{P}}^2 = \mathbf{y}^T\mathbf{P}\mathbf{y}$$

Finally, define

$$\begin{aligned} P(s,t,q) &= \Pr(\mathbf{y}_{t:s} | t : s \text{ is a segment, model order } q) \\ &= |\mathbf{M}|^{1/2}\left(\gamma + \|\mathbf{y}_{t:s}\|_{\mathbf{P}}^2\right)^{-(\nu+s-t+1)/2} \\ &\quad \times \frac{\Gamma\left(\frac{(\nu+s-t+1)}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \prod_{j=1}^{q} \delta_j^{-1} \end{aligned} \tag{1}$$

where (1) is obtained by integrating out the regression parameters and variance.

Then, for $t = 1, \ldots, n$

$$\begin{aligned} Q(t) = \sum_{s=t}^{n-1}\sum_{q=1}^{\bar{p}} P(t,s,q)Q(s+1)p(q)\lambda(1-\lambda)^{s-t} \\ + \sum_{q=1}^{\bar{p}} P(t,n,q)p(q)(1-\lambda)^{n-t}. \end{aligned} \tag{2}$$

The intuition behind this recursion is that (suppressing the conditioning on a changepoint at $t-1$ for notational convenience)

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

The respective joint probabilities are given by the two sets of sums over the model order that appears on the right-hand side of (2). See [3] for a formal proof of this recursion.

2) Simulation: Once the $Q(t)$s have been calculated for $t = 1, \ldots, n$, it is straightforward to recursively simulate the changepoints and linear regression orders. To simulate the changepoints, we set $\tau_0 = 0$, and then recursively simulate τ_j, given τ_{j-1} for $j = 1, 2, \ldots$, until $\tau_m = n$ for some value of m. The conditional posterior distribution of $\tau_j = s$, given $\tau_{j-1} = t-1$, is

$$\begin{aligned} \Pr(\tau_j = s | \tau_{j-1} = t-1, \mathbf{y}_{1:n}) \\ \propto \sum_{q=1}^{\bar{p}} P(t,s,q)p(q)Q(s+1)\lambda(1-\lambda)^{s-t} \end{aligned}$$

for $s = t, \ldots, n-1$, and

$$\Pr(\tau_j = n | \tau_{j-1} = t-1, \mathbf{y}_{1:n}) \propto \sum_{q=1}^{\bar{p}} P(t,n,q)p(q)(1-\lambda)^{n-t}.$$

For the ith segment, the posterior distribution of the model order is given by

$$p(q|\tau_{i-1}=t-1,\tau_i=s,\mathbf{y}_{1:n}) \propto P(t,s,q)p(q).$$

3) Viterbi Algorithm: A Viterbi algorithm [15] can be used to calculate the MAP estimate of the changepoint positions and model orders. Define

$$Q^*(t) = \Pr(\mathbf{y}_{t:n}|\text{changepoint at } t-1, \text{ MAP estimate for } t:n)$$

$I(s \neq n) = 1$ if $s \neq n$ and 0 otherwise, and $Q^*(n+1) = 1$. Then

$$Q^*(t) = \max_{s,q}\{P(t,s,q)Q^*(s+1)p(q)\lambda^{I(s\neq n)}(1-\lambda)^{s-t}\}$$

where the maximum is taken over $s = t,\ldots,n$, and $q = 1,\ldots,\bar{p}$. Define $s^*(t)$ and $q^*(t)$ to be the values of s and q, which achieve the maximum. Then, the MAP estimate of the changepoints $\tau_0^*,\ldots,\tau_m^*$ and the model orders $p_1^*,\ldots,p_m^*$ can be obtained recursively by the following:

a) Set $\tau_0^* = 0$ and $j = 0$.
b) While $\tau_j^* < n$: i) set $\tau_{j+1}^* = s^*(\tau_j^*+1)$; ii) set $p_{j+1}^* = q^*(\tau_j^*+1)$; iii) set $j = j+1$, and go to (b).

The MAP estimate produced by this algorithm may have a different number of segments than the MAP estimate of the number of segments (see Section III-B). An alternative approach to MAP estimation is to fix the number of segments to the MAP estimate $\hat{m}$, say, and calculate the MAP estimate of the position of the changepoints and the model orders conditional on $\hat{m}$ segments.

A simple adaptation of the above algorithm can perform such a conditional MAP estimation of the changepoints and model orders. Define

$$Q_k^*(t) = \Pr(\mathbf{y}_{t:n}|\text{changepoint at } t-1, \; k \text{ segments after } t-1, \text{ MAP estimate for } t:n).$$

Then

$$Q_1^*(t) = \max_{q=1,\ldots,\bar{p}}\{P(t,n,q)p(q)(1-\lambda)^{n-t}\}$$

and for $k = 2,\ldots,\hat{m}$

$$Q_k^*(t) = \max_{q,s}\{P(t,s,q)Q_{k-1}^*(s+1)p(q)\lambda(1-\lambda)^{s-t}\}$$

where the maximum is taken over $q = 1,\ldots,\bar{p}$, and $s = t,\ldots,n+1-k$. Define $s_k^*(t)$ and $q_k^*(t)$ as the values of s and q that achieve the maximum (with $s_1^*(t) = n$). Then, the MAP estimate of the changepoints and model orders can be obtained by the following.

a) Set $\tau_0^* = 0$, $k = \hat{m}$, and $j = 0$.
b) While $\tau_j^* < n$, i) set $\tau_{j+1}^* = s_k^*(\tau_j^*+1)$, ii) set $p_{j+1}^* = q_k^*(\tau_j^*+1)$, iii) set $j = j+1$, $k = k-1$, and go to (b).

4) Implementation: As written, (2) suffers from numerical instabilities. This can be overcome by calculating $\log Q(t)$ recursively, using

$$\log Q(t) = \log Q(t+1) + \log\left(\frac{Q(t)}{Q(t+1)}\right)$$

and

$$\frac{Q(t)}{Q(t+1)} = \sum_{q=1}^{\bar{p}} p(q) \sum_{s=t}^{n-1} \lambda \times \exp\{\log P(t,s,q) + \log Q(s+1) - \log Q(t+1) + (s-t)\log(1-\lambda)\} + \sum_{q=1}^{\bar{p}} p(q)\exp\{\log P(t,n,q) - \log Q(t+1) + (n-t)\log(1-\lambda)\}.$$

Evaluating $Q(t)/Q(t+1)$ can be achieved in computational time, which is $O(n-t)$, as the matrix multiplications involved in calculating the $P(t,s,q)$s can be done recursively. For example, the $\mathbf{G}^T\mathbf{G}$ term required for $P(t,s,q)$ can be calculated from the equivalent term required for $P(t,s-1,q)$. This is also true for the $\mathbf{y}^T\mathbf{y}$ and $\mathbf{G}^T\mathbf{y}$ terms required in $||\mathbf{y}||_{\mathbf{P}}$.

When calculating the $Q(t)$ values, we store the values for $P(t,s,q)$ that we have calculated. These stored values can then be used in the simulation stage of the algorithm. An efficient algorithm for simulating large samples from the posterior distribution once the $Q(t)$ values have been calculated is given in [3]. The main computational cost of the perfect simulation is that of evaluating the recursions to calculate the $Q(t)$s; once calculated, simulating samples are computationally very cheap.

The computational complexity of the recursion for the $Q(t)$s is $O(n^2)$. However, computational savings can be made in general as the terms in (2) tend to become negligible for sufficiently large s. We suggest truncating the sum in (2) at term k when

$$\frac{\sum_{q=1}^{\bar{p}} P(t,k,q)Q(k+1)p(q)\lambda(1-\lambda)^{k-t}}{\sum_{s=t}^{k}\sum_{q=1}^{\bar{p}} P(t,s,q)Q(s+1)p(q)\lambda(1-\lambda)^{s-t}} \tag{3}$$

becomes smaller than some predetermined value, for example, 10^{-12}.

In a limiting regime, where data is observed over longer time periods (as opposed to observations being made more frequently) such that as n increases, the number of changepoints increases linearly with n, this simplification is likely to make the algorithm's complexity $O(n)$. See Section III for empirical evidence of this.

C. Hyperpriors and MCMC

We have described an algorithm for perfect simulation from the model of Section II-A. While this algorithm produces iid draws from the true posterior distribution of the model, the usefulness of the approach and the model will depend on the choice of prior parameters

$$\boldsymbol{\theta} = \{\lambda,\nu,\gamma,\delta_1,\ldots,\delta_{\bar{p}}\}.$$

The choice of these parameters defines the Bayes factors for the competing models and, hence, the posterior distribution from which it is sampled.

The approach of [13], which is also used by [1], lets the data choose the prior parameters. In these two papers, this is achieved by introducing uninformative hyperpriors on the prior parameters. Unfortunately, using hyperpriors introduces dependence

between the segments, such that the approach of Section II-B is no longer directly applicable.

One solution is to use the results from a preliminary analysis of the data to choose the prior parameters. Thus, we implement the perfect simulation algorithm using a default choice of the prior parameters $\boldsymbol{\theta}$. New values of $\boldsymbol{\theta}$ can be chosen based on the perfect samples of the number of changepoints as well as the regression variance and parameters. For example, the $\boldsymbol{\theta}$ values could be chosen so that the prior means of the regression variance, parameters, and number of changepoints are close to the posterior means from the preliminary analysis. We denote such estimated values of the prior parameters by $\hat{\boldsymbol{\theta}}$. If necessary, this approach could be iterated a number of times until there is little change in the posterior means. We call this a recursive approach.

A less *ad hoc* approach, which mimics that of [1], is to use a hyperprior for $\boldsymbol{\theta}$. A simple MCMC algorithm in this case is as follows.

a) Update the number of segments m, the changepoints, and model orders conditional on $\boldsymbol{\theta}$.
b) Update the $\boldsymbol{\beta}_i$s and σ_i^2s for $i = 1, \ldots, m$ conditional on m, the changepoints, the model orders, and $\boldsymbol{\theta}$.
c) Update $\boldsymbol{\theta}$ conditional on m, the changepoints, the model orders, and, for $i = 1, \ldots, m$, the $\boldsymbol{\beta}_i$s and σ_i^2s.

If conjugate hyperpriors are used (for example, those of [1] or of Section III), then Gibbs updates can be used in steps b) and c). The perfect simulation algorithm can be used in step a) to simulate the changepoints and model orders from the full conditional, given $\boldsymbol{\theta}$. However, we advocate a more efficient (in terms of computing time) approach, which is to use an independence proposal from the posterior distribution conditional on the prior parameters being $\hat{\boldsymbol{\theta}}$.

We test and compare the accuracy and efficiency of both the recursive and MCMC approaches on a number of examples in Section III.

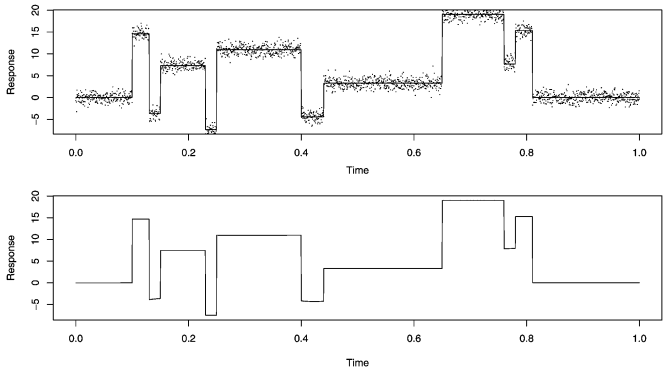

Fig. 1. (Top) Blocks function and observations and (bottom) estimates (posterior means) based on the recursive and MCMC approaches. The two estimates are almost exact and indistinguishable on the plot. The average square error of the two estimates are 0.0045 and 0.0043, respectively.

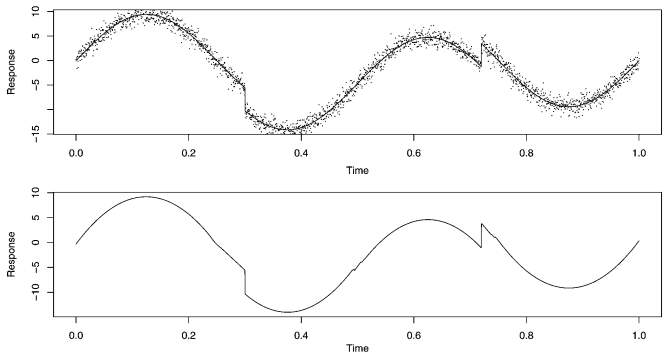

Fig. 2. (Top) Heavisine function and observations and estimates (posterior means) based on the (bottom) recursive and MCMC approaches. The two estimates are almost exact and indistinguishable on the plot. The average square error of the two estimates are 0.0266 and 0.0264, respectively.

III. Examples

A. Polynomial Regression

For our first class of examples, we assume that for each segment, there is a polynomial relationship between the response and the explanatory variable. Here, we assume that the response is either constant, linear, or quadratic. For a segment consisting of observations $y_{t:s}$ and explanatory variables $x_{t:s}$, the $(s-t+1) \times 3$ matrix of basis vectors for the quadratic relationship is taken to be

$$G_{t:s}^{(3)} = \begin{pmatrix} 1 & x_t - m_1 & x_t^2 - ax_t - b \\ 1 & x_{t+1} - m_1 & x_{t+1}^2 - ax_{t+1} - b \\ \vdots & \vdots & \vdots \\ 1 & x_s - m_1 & x_s^2 - ax_s - b \end{pmatrix}$$

where

$$a = \frac{m_3 - m_1 m_2}{m_2 - m_1^2}, \quad b = m_2 + a m_1$$

and for $j = 1, 2, 3$

$$m_j = \frac{1}{s-t+1} \sum_{i=t}^{s} x_i^j$$

which is the mean value of the jth power of the explanatory variables $x_{t:s}$.

The reason for this choice of model is that the basis vectors are orthogonal, and for a given segment, the regression parameters are independent under the posterior distribution. This helps with the interpretability of the parameter estimates and slightly reduces the computational cost of perfect simulation. The first- and second-order models are obtained by taking the first and first two columns of $G_{t:s}^{(3)}$, respectively.

We tested our algorithm on the four test data sets of [16]. These have been previously analyzed by [17] and [1], among others. Each data set consists of 2048 equally spaced observations of an underlying functional relationship (for example, see Figs. 1 and 2). The noise variance was 1.0 throughout, which gives a signal-to-noise ratio of 7 in each case. In our simulation study, we focused primarily on the computational aspects of our approach. The accuracy of inference from a related model to the one we use for these test data sets is given in [1].

We set $\nu = 2$ as in [1]. Initial parameter values were $\lambda = 0.01$, $\gamma = 2$, and $\boldsymbol{\delta^2} = (10, 10n, 10n^2)$. We first tried the recursive approach, with two preliminary runs being used to obtain an estimate of the prior parameter values $\hat{\boldsymbol{\theta}}$. Second, we implemented the MCMC approach, assuming an Inverse-Gamma prior on the δ_j^2s, a uniform prior on λ, and an improper Jeffreys' prior on γ. In implementing step a) of the MCMC algorithm of

TABLE I
RESULTS OF ANALYSIS OF THE FOUR TEST DATA SETS

	k_{ave}	τ	Acc Prob
Blocks	140	1.2	0.97
Heavisine	310	1.1	0.94
Bumps	130	3.6	0.68
Doppler	240	1.8	0.80

Section II-C, we used proposals from the posterior distribution conditional on $\hat{\boldsymbol{\theta}}$.

In calculating the $Q(t)$s for each data set, we truncated the sum in (2) when (3) was less than 10^{-12}. By varying this cutoff, we could tell that any inaccuracies introduced were negligible. For each data set, evaluating the recursions to calculate the $Q(t)$s took less than 10 sec on a 900-MHz Pentium PC.

Summaries of the computational aspects of the perfect simulation and the MCMC algorithms are given in Table I. These include the average number of terms calculated in the sum of (2) k_{ave}; the autocorrelation time of the MCMC algorithm τ, and the acceptance probability of the independence sampler. The autocorrelation time was calculated as the maximum estimated time for all the prior parameters.

The average number of terms calculated for the sums in (2) was much less in all cases than the roughly 1000 that would occur if no truncation of the sum was used. The average number of terms depends primarily on the average length of the segments for realizations that have non-negligible posterior probability. For example, the Heavisine function had fewest segments (as few as 5) and, hence, the most terms, whereas, for example, Bumps had many more segments (at least 35) and, thus, fewer terms.

The MCMC algorithm mixed extremely well in all cases. The acceptance probabilities of the independence sampler in part a) of the algorithm were high (close to 1 for Blocks and Heavisine). The autocorrelation times were also low because they were close to 1 for Blocks and Heavisine, which suggests near iid samples from the target posterior distribution.

For each dataset, the estimates based on the recursive approach and those based on the MCMC approach were almost identical. For example, the two estimates for the Blocks and the Heavisine data sets are shown in Figs. 1 and 2. The average mean square errors of the estimates were also almost identical in all cases. As can be seen from these figures, the reconstruction of the Blocks and Heavisine functions are very good.

B. Auto Regressive Processes

Our second example is based on an analyzing data from a piecewise constant AR process. Such models are used for speech data [18]. We considered models of order up to 3. For a segment consisting of observation $y_{t:s}$, the matrix of basis vectors for the third-order model is

$$G_{t:s}^{(3)} = \begin{pmatrix} y_{t-1} & y_{t-2} & y_{t-3} \\ y_t & y_{t-1} & y_{t-2} \\ \vdots & \vdots & \vdots \\ y_{s-1} & y_{s-2} & y_{s-3} \end{pmatrix}.$$

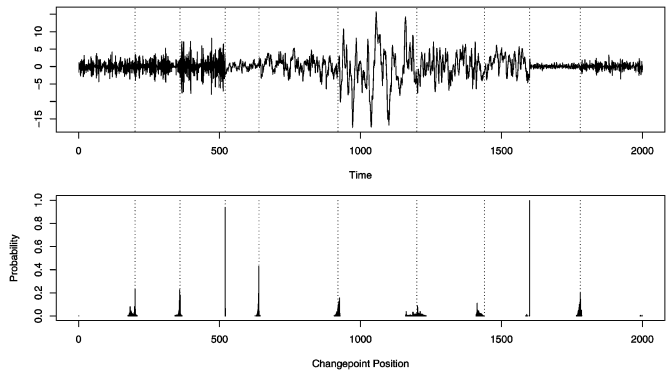

Fig. 3. (Top) Simulated AR process and (bottom) the posterior distribution of the position of the changepoints. The true changepoint positions are denoted by dashed lines.

TABLE II
MAP ESTIMATES OF CHANGEPOINTS AND MODEL ORDER

Segment i	τ_i	$\hat{\tau}_i$	p_i	$\hat{p}_i$
1	200	200	1	1
2	360	359	2	2
3	520	520	3	3
4	640	639	1	1
5	920	926	1	1
6	1200	1202	2	2
7	1440	1414	3	2
8	1600	1600	3	2
9	1780	1780	3	3
10	-	-	2	2

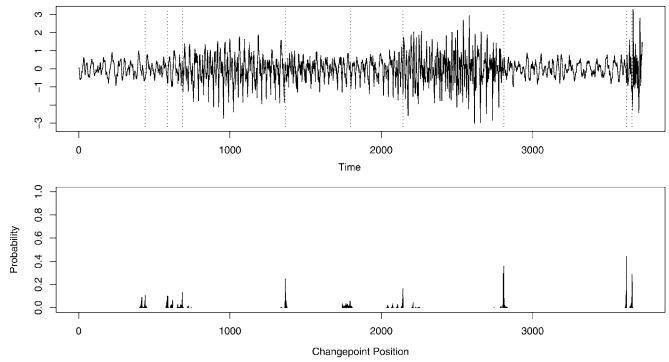

Fig. 4. (Top) Speech data with conditional MAP estimates of the changepoints given by vertical dashed lines and (bottom) posterior distribution of the changepoint positions.

The matrices for the first- and second-order models consist of the first and first two columns of $G_{t:s}^{(3)}$, respectively.

We simulated 2000 data points from a model with nine breakpoints. The data is shown in Fig. 3. We only used the recursive approach to analyze this data, as the MCMC approach had produced a negligible difference for the polynomial regression examples.

As above, we implemented the recursions for the $Q(t)$s by truncating the sums when (3) was less than 10^{-12}. On average, 250 terms were required in the summation. For simplicity, we summarize our results in terms of the MAP estimate of the changepoint positions and the model orders (see Table II) and the posterior distribution of the changepoint positions (see Fig. 3).

TABLE III
CHANGEPOINT POSITIONS FOR DIFFERENT METHODS

Method	AR order	Estimated Changepoints									
Divergence [20]	16	445	-	645	1550	1800	2151	2797	-	3626	-
GLR [21]	16	445	-	645	1550	1800	2151	2797	-	3626	-
GLR [21]	2	445	-	645	1550	1750	2151	2797	3400	3626	-
Approx ML [18]	2	445	-	626	1609	-	2151	2797	-	3627	-
MCMC [1]	estimated	448	-	624	1377	-	2075	2807	-	3626	-
Conditonal MAP	estimated	439	585	685	1365	1795	2144	2811	-	3621	3657
Unconditional MAP	estimated	439	-	620	1365	1795	2144	2811	-	3621	3657

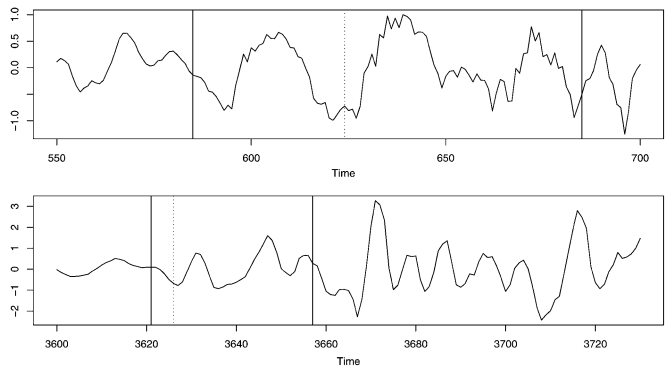

Fig. 5. Changepoint positions of the conditional MAP estimate (solid lines) and those estimated by Punskaya *et al.* (dashed lines).

The MAP estimate of the changepoint positions and model orders consists of only eight changepoints, whereas the MAP estimate of the number of changepoints is 9 (posterior probability 0.48). The MAP estimates given in Table II are conditional on there being nine changepoints. For the unconditional MAP estimate, the seventh changepoint is missed, but otherwise, the estimates of the changepoint positions and model orders are unchanged. The MAP estimate incorrectly infers the model order for segments 7 and 8. In each case, the MAP estimate is one less than the true model order, and the AR coefficients that are incorrectly estimated as 0 are both small (0.1 and 0.2).

C. Speech Segmentation

We also used our method to analyze a real speech signal [19], which has also previously been analyzed in the literature [1], [18], [19]. We analyzed the data under a piecewise AR model that allowed the AR orders of between 1 and 6 for each segment. We implemented the recursive approach, where an initial run of the exact simulation algorithm is used to construct suitable prior parameters.

The signal, MAP estimates of the changepoints, and posterior distribution of the changepoints are given in Fig. 4. A comparison of our estimates of the changepoint positions to previous esimates are shown in Table III, where we give our MAP estimates, both conditional and unconditional, on the MAP estimate for the number of changepoints.

Our MAP estimates are similar to those of Punskaya *et al.* [1], except for the inclusion by the conditional MAP estimate of an extra changepoint near the beginning of the signal and the inclusion by both MAP estimates of an extra changepoint near the end of the signal. Fig. 5 shows these two regions and the estimated changepoints from the different methods.

IV. CONCLUSION

We have presented a novel algorithm for performing exact Bayesian inference for regression models, where the underlying function relationship consists of independent linear regressions on disjoint segments. The algorithm is both scalable and easy to implement. It avoids the problems of diagnosing convergence that are common with MCMC methods.

We have focused on models suggested by [1], but the algorithm can be applied more generally. The main requirement is that of independence between the parameters associated with each segment.

The regression problem we have addressed involves model uncertainty. In practice, the accuracy of Bayesian inference for such model-choice problems depends on the choice of prior. We considered two approaches to choosing these priors, both based on letting the data inform the choice of prior parameters. In our examples, we found that the simpler of the two (the recursive approach) performs as well as the approach based on introducing hyperparameters, and we would suggest such an approach in practice.

We have also demonstrated how MAP estimates of the changepoints can be obtained. There are two ways of defining the MAP estimate, depending on whether or not the MAP estimate of the number of changepoints is first calculated, and then, the changepoints are inferred, conditional on this number of changepoints. In some cases, these different approaches can give different estimates for the number and position of the changepoints: For example, when there is a likely changepoint in some period of time but there is a lot of uncertainty over when this changepoint occured, conditioning on the MAP number of changepoints will pick up a changepoint during this period of time, but it may be omitted otherwise (see Section III-B). Note that for the related problem of inferring changepoints in continuous time, it would clearly be correct to conditon on the number of changepoints, as it is inappropriate to compare joint densities of positions of changepoints that are of different dimension.

ACKNOWLEDGMENT

The author would like to thank E. Punskaya for providing the speech data.

REFERENCES

[1] E. Punskaya, C. Andrieu, A. Doucet, and W. J. Fitzgerald, "Bayesian curve fitting using MCMC with applications to signal segmentation," *IEEE Trans. Signal Process.*, vol. 50, no. 3, pp. 747–758, Mar. 2002.

[2] P. Green, "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, vol. 82, pp. 711–732, 1995.
[3] P. Fearnhead. (2004) Exact and Efficient Inference for Multiple Changepoint Problems. [Online]. Available: http://www.maths.lancs.ac.uk/~fearnhea/
[4] L. R. Rabiner and B. H. Juang, "An introduction to hidden Markov models," *IEEE Acoust., Speech, Signal Process. Mag.*, vol. ASSP-34, no. 1, pp. 4–15, Jan. 1986.
[5] D. Barry and J. A. Hartigan, "Product partition models for change point problems," *Ann. Statist.*, vol. 20, pp. 260–279, 1992.
[6] ——, "A Bayesian analysis for change point problems," *J. Amer. Statist. Soc.*, vol. 88, pp. 309–319, 1993.
[7] J. G. Propp and D. B. Wilson, "Exact sampling with coupled Markov chains and applications to statistical mechanics," in *Random Structures Algorithms*, 1996, vol. 9, pp. 223–252.
[8] C. C. Holmes and D. G. T. Denison, "Perfect sampling for the wavelet reconstruction of signals," *IEEE Trans. Signal Process.*, vol. 50, no. 2, pp. 337–344, Feb. 2002.
[9] P. D. O'Neill and G. O. Roberts, "Bayesian inference for partially observed stochastic epidemics," *J. R. Statist. Soc.*, ser. A, vol. 162, pp. 121–129, 1999.
[10] P. Fearnhead and L. Meligkotsidou, "Exact filtering for partially-observed continuous-time Markov models," *J. R. Statist. Soc.*, ser. B, vol. 66, pp. 771–789, 2004.
[11] P. J. Green, "Trans-dimensional Markov chain Monte Carlo," in *Highly Structured Stochastic Systems*, P. J. Green, N. L. Hjort, and S. Richardson, Eds. Oxford, U.K.: Oxford Univ. Press, 2003.
[12] J. M. Bernardo and A. F. M. Smith, *Bayesian Theory*. Chichester, U.K.: Wiley, 1994.
[13] S. Richardson and P. J. Green, "On Bayesian analysis of mixtures with an unknown number of components," *J. R. Statist. Soc.*, ser. B, vol. 59, pp. 731–792, 1997.
[14] B. P. Carlin, A. E. Gelfand, and A. F. M. Smith, "Hierarchical Bayesian analysis of changepoint problems," *Appl. Statist.*, vol. 41, pp. 389–405, 1992.
[15] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. Inf. Theory*, vol. IT-13, no. 2, pp. 260–269, Apr. 1967.
[16] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, pp. 425–455, 1994.
[17] D. G. T. Denison, B. K. Mallick, and A. F. M. Smith, "Automatic Bayesian curve fitting," *J. R. Statist. Soc.*, ser. B, vol. 60, pp. 333–350, 1998.
[18] F. Gustafsson, *Adaptive Filtering and Change Detection*. New York: Wiley, 2000.
[19] M. Basseville and I. V. Nikiforov, *Detection of Abrupt Changes: Theory and Applications*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[20] M. Basseville and A. Benveniste, "Design and comparative study of some sequential jump detection algorithms for digital signals," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-31, no. 2, pp. 521–535, Apr. 1983.
[21] U. Appel and A. V. Brandt, "Adaptive sequential segmentation of piecewise stationary time series," *Inf. Sci.*, vol. 29, pp. 27–56, 1983.

Paul Fearnhead received the B.A. and D.Phil. degrees in mathematics from the University of Oxford, Oxford, U.K.

Since 2001, he has been a Lecturer in statistics at Lancaster University, Lancaster, U.K. Prior to this, he was a research associate with the Mathematical Genetics group, University of Oxford. His research interests include computational statistical methods and, in particular, particle filters and the use of filters within Markov chain Monte Carlo methods. He is also interested in modeling and inference in population genetics.