Bayesian Curve Fitting Using MCMC With Applications to Signal Segmentation

Elena Punskaya, Christophe Andrieu, Arnaud Doucet, and William J. Fitzgerald

***Abstract*—We propose some Bayesian methods to address the problem of fitting a signal modeled by a sequence of piecewise constant linear (in the parameters) regression models, for example, autoregressive or Volterra models. A joint prior distribution is set up over the number of the changepoints/knots, their positions, and over the orders of the linear regression models within each segment if these are unknown. Hierarchical priors are developed and, as the resulting posterior probability distributions and Bayesian estimators do not admit closed-form analytical expressions, reversible jump Markov chain Monte Carlo (MCMC) methods are derived to estimate these quantities. Results are obtained for standard denoising and segmentation of speech data problems that have already been examined in the literature. These results demonstrate the performance of our methods.**

***Index Terms*—Bayesian model, curve fitting, Markov chain Monte Carlo methods, signal segmentation.**

I. Introduction

A. Problem Statement

REGRESSION problems are among the most common problems in signal processing. The aim is to estimate an assumed functional relationship between a response and some explanatory variables given noisy measurements. Many parametric and semi-parametric methods have been proposed in the literature in order to solve these problems, including smoothing splines and kernel methods. We adopt here a standard model where the regression function is assumed to be a function made up of low-order pieces that are standard linear regression models within some segments, where the number and position of the segments are parameters to estimate.

More formally, let us denote for any generic sequence κ_t, $\boldsymbol{\kappa}_{i:j} \triangleq (\kappa_i, \kappa_{i+1}, \ldots, \kappa_j)^{\mathrm{T}}$, and let $\mathbf{y}_{0:T-1}$ be a vector of T real observations. The elements of $\mathbf{y}_{0:T-1}$ may be represented by one of the models $\mathcal{M}_{k,\mathbf{p}_k}$, corresponding to the case when the signal is in the form of the linear regression model with piecewise constant parameters and k ($k = 0, \ldots, k_{\max}$) changepoints. That is, one has

$$\mathcal{M}_{k,\mathbf{p}_k}\colon \mathbf{y}_{\tau_i:\tau_{i+1}-1} = \mathbf{G}_i^{(p_i)}\boldsymbol{\beta}_i^{(p_i)} + \mathbf{n}_{\tau_i:\tau_{i+1}-1} \quad i = 0, \ldots, k \tag{1}$$

Manuscript received January 31, 2001; revised October 15, 2001. The associate editor coordinating the review of this paper and approving it for publication was Dr. Petar M. Djuric.

E. Punskaya and W. J. Fitzgerald are with the Signal Processing Laboratory, Department of Engineering, University of Cambridge, Cambridge, U.K. (e-mail: op205@eng.cam.ac.uk; wjf@eng.cam.ac.uk).

C. Andrieu is with the Department of Mathematics, Statistics Group, University of Bristol, Bristol, U.K. (e-mail: C.Andrieu@bristol.ac.uk).

A. Doucet is with the Department of Electrical Engineering, The University of Melbourne, Parkville, Victoria, Australia (e-mail: a.doucet@ee.mu.oz.au).

Publisher Item Identifier S 1053-587X(02)02163-3.

where $\boldsymbol{\beta}_i^{(p_i)}$ is a vector of p_i model parameters for the ith ($i = 0, \ldots, k$) segment,[1] and $\mathbf{n}_{\tau_i:\tau_{i+1}-1}$ is a vector of i.i.d. Gaussian noise samples of variance σ_i^2 associated with the ith model. The changepoints of the model $\mathcal{M}_{k,\mathbf{p}_k}$ are arranged in the vector $\boldsymbol{\tau}_k \triangleq \boldsymbol{\tau}_{1:k}$, and we adopt the convention $\tau_0 = 0$ and $\tau_{k+1} = T - 1$ for notational convenience. We also denote $\boldsymbol{\sigma}_k^2 \triangleq \boldsymbol{\sigma}_{0:k}^2$ and $\mathbf{p}_k \triangleq \mathbf{p}_{0:k}$, where $p_i = 0, \ldots, p_{\max}$. The matrix $\mathbf{G}_i^{(p_i)}$ is the matrix of basis functions for the ith segment ($i = 0, \ldots, k$). For example, for the piecewise polynomial model, $\mathbf{G}_i^{(p_i)}$ is given by

$$\mathbf{G}_{\mathrm{poly}\,i}^{(p_i)} = \begin{bmatrix} 1 & x_{\tau_i} & x_{\tau_i}^2 & \cdots & x_{\tau_i}^{p_i-1} \\ 1 & x_{\tau_i+1} & x_{\tau_i+1}^2 & \cdots & x_{\tau_i+1}^{p_i-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{\tau_{i+1}-1} & x_{\tau_{i+1}-1}^2 & \cdots & x_{\tau_{i+1}-1}^{p_i-1} \end{bmatrix}$$

and for a piecewise constant autoregressive (AR) process, $\mathbf{G}_i^{(p_i)}$ is of the following form:

$$\mathbf{G}_{\mathrm{AR}\,i}^{(p_i)} = \begin{bmatrix} y_{\tau_i-1} & y_{\tau_i-2} & \cdots & y_{\tau_i-p_i} \\ y_{\tau_i} & y_{\tau_i-1} & \cdots & y_{\tau_i+1-p_i} \\ \vdots & \vdots & \ddots & \vdots \\ y_{\tau_{i+1}-2} & y_{\tau_{i+1}-3} & \cdots & y_{\tau_{i+1}-1-p_i} \end{bmatrix}.$$

Typically, the orders of the different linear regression models $\mathbf{p}_{0:k}$ are assumed equal and known, that is, $p_i = p_j = p_0$ for any $(i, j) \in \{0, \ldots, k\}$. However, in practice, there are numerous applications (speech processing, for example) where different model orders should be considered for different segments and estimated from the data. Thus, in the general case, the number of changepoints k and the associated parameters $\boldsymbol{\Psi}_{k,\mathbf{p}_k} \triangleq (\boldsymbol{\tau}_k, \mathbf{p}_k, \{\boldsymbol{\beta}_i^{(p_i)}\}_{i=0}^k, \boldsymbol{\sigma}_k^2)$ are unknown. Given $\mathbf{y}_{0:T-1}$, our aim is to estimate k and $\boldsymbol{\Psi}_{k,\mathbf{p}_k}$.

B. Background

This model allows for a wide range of applications from curve fitting of noisy data [1] to changepoint detection and signal segmentation [2]. For example, the general piecewise linear model and its extension to study multiple changepoints in non-Gaussian impulsive noise environments is studied in [3]. In [4] and [5], it is shown that piecewise constant autoregressive (AR) processes excited by white Gaussian noise have proved useful for processing real signals, such as speech data.

In general, this class of models is very flexible with a large number of parameters to be estimated; therefore, one needs to prevent overfitting in some way. We adopt a Bayesian approach

[1]For notational simplicity, index k is suppressed here and later.

and set a prior (which also works as a penalty against overfitting) on all the unknown parameters. Bayesian curve fitting/signal segmentation for related models has been studied by several authors recently, including [1]–[3] and [6]. Gustafsson [2] and Djurić [6] have proposed to perform MAP (maximum *a posteriori*) changepoint estimation using deterministic algorithms. Although these methods are fast and can give good results, one cannot compute any confidence intervals or perform Bayesian model averaging. Moreover, it seems difficult to generalize them to the case where the model order p_i within each segment is unknown.

C. Resolution

We favor a "full" Bayesian approach where the complete posterior distribution and any posterior feature of interest is estimated using MCMC. Bayesian approaches for multiple changepoint detection based on MCMC for different models are proposed, for example, in [7] or [8]. The closest work to the one presented here is the technique followed in [1]; see also [9]. Our methodology is, however, different in many respects.

Our model is more general as it allows not only for an unknown number of segments [1] but for an unknown model order within each segment as well, if necessary [9], that is, we face a "double" model selection problem. We also adopt hierarchical prior distributions where the hyperparameters are assumed random with a vague prior distribution; a similar approach was adopted in [10]. This has the effect of increasing robustness of the Bayesian models in comparison with the standard approach, where these parameters are fixed [1], [2], which was also demonstrated by a simulation study. We propose efficient algorithms in order to sample from the posteriors based on reversible jump MCMC [11].

D. Plan

The rest of the paper is organized as follows. For the sake of clarity, as the "double" selection problem is quite complex, we have chosen to begin in Section II with the case where the orders of the different linear regression models p_i are all equal and known; then, in Section III, the case where they can be different and are unknown is treated. In Section IV, we apply our methods to standard denoising problems [1] and speech segmentation [2], [4], [5].

II. Bayesian Inference for Fixed Model Dimensions

We assume that the model order for each segment is fixed and known *a priori*, i.e., $p_i = p_0$, for $i = 0, \ldots, k$, p_0 is given, and for notational convenience, we will denote the unknown parameters in this case $\boldsymbol{\Psi}_k \triangleq (\boldsymbol{\tau}_k, \{\boldsymbol{\beta}_i^{(p_0)}\}_{i=0}^k, \boldsymbol{\sigma}_k^2)$.

A. Bayesian Model and Estimation Objectives

We follow a Bayesian approach where $\boldsymbol{\Psi}_k$ are regarded as random with a known prior that reflects our degree of belief in the different values of these quantities. In order to increase robustness of the prior, the hyperparameters are assumed random with a vague distribution [12], that is, we adopt a hierarchical Bayesian model.

1) Bayesian Hierarchical Model: In our case, it is natural to introduce a binomial distribution as a prior distribution for the number of changepoints and their positions (as in [2])

$$p(k, \boldsymbol{\tau}_k|\lambda) = \lambda^k(1-\lambda)^{T-2-k}\mathbb{I}_{\boldsymbol{\Upsilon}_k}(\tau_k), \qquad 0<\lambda<1 \quad (2)$$

where $\boldsymbol{\Upsilon}_k \triangleq \{\boldsymbol{\tau}_{1:k} \in \{1, \ldots, T-2\}^k$ such that $\tau_1 \neq \tau_2 \neq \cdots \neq \tau_k\}$, and $\mathbb{I}_{\boldsymbol{\Upsilon}_k}(\tau_k)$ is an indicator function of the set $\boldsymbol{\Upsilon}_k$ (1 if $\boldsymbol{\tau}_k \in \boldsymbol{\Upsilon}_k$, 0 otherwise). We assign a normal distribution to the parameters of the models ($p_i = p_0$ here)

$$\boldsymbol{\beta}_i^{(p_i)}\Big|\left(\sigma_i^2, \delta_i^2\right) \sim \mathcal{N}\left(0, \sigma_i^2\delta_i^2\mathbf{I}_{p_i}\right), \qquad i=0, \ldots, k \quad (3)$$

with the same hyperparameter δ_i^2 for all segments when the model order is assumed known, $\delta_i^2 = \delta_0^2$, $i = 0, \ldots, k$, and a conjugate Inverse-Gamma distribution to the noise variances

$$\sigma_i^2\Big|\left(\frac{\nu_\sigma}{2}, \frac{\gamma_\sigma}{2}\right) \sim \mathcal{IG}\left(\frac{\nu_\sigma}{2}, \frac{\gamma_\sigma}{2}\right), \qquad i=0, \ldots, k \quad (4)$$

with $\nu_\sigma = 2$. This choice of prior [see (3) and (4)], given the Gaussian noise model, allows the marginalization of the parameters[2] $(\{\boldsymbol{\beta}_i^{(p_0)}\}_{i=0}^k, \boldsymbol{\sigma}_k^2)$.

The algorithm requires the specification of λ, δ_0^2 and γ_σ. It is clear that these parameters play an important role in model selection. Indeed, the Bayes factors are dependent on them. Thus, in order to increase the robustness of the prior, we propose to estimate λ, δ_0^2, γ_σ from the data (as it is done, for example, in [10]), i.e., we consider λ, δ_0^2, γ_σ to be random. We assign a vague conjugate Inverse-Gamma distribution to the scale hyperparameter δ_0^2

$$\delta_0^2|(\alpha_\delta, \theta_{\delta_0}) \sim \mathcal{IG}(\alpha_\delta, \theta_{\delta_0}), \qquad \alpha_\delta>0,\, \theta_{\delta_0}>0,\, i=0, \ldots, k$$

and set $\alpha_\delta = 1$. We also choose a uniform prior distribution for λ, $\lambda \sim \mathcal{U}_{(0,1)}$ and a noninformative improper Jeffreys' prior for γ_σ.

Thus, the following hierarchical structure is assumed for the prior of the parameters.

$$p\left(k, \boldsymbol{\Psi}_k, \lambda, \delta_0^2, \gamma_\sigma\right) = p\left(k, \boldsymbol{\tau}_k|\,\lambda\right)p(\lambda) \times \prod_{i=0}^{k}\left[p\left(\boldsymbol{\beta}_i^{(p_0)}\Big|\,\sigma_i^2, \delta_0^2\right)p\left(\sigma_i^2\big|\,\gamma_\sigma\right)\right]p(\delta_0^2)p(\gamma_\sigma)$$

which can be visualized with a directed acyclic graph (DAG), as shown in Fig. 1 (for convenience, we do not show the fixed parameters ν_σ, α_δ, and θ_{δ_0}).

For our problem, the overall parameter space can be written as a finite union of subspaces $\boldsymbol{\Theta} \triangleq \bigcup_{k=0}^{k_{\max}}\{k\} \times \boldsymbol{\Upsilon}_k \times \prod_{i=0}^k \boldsymbol{\Phi}_{p_0} \times (0,1) \times \boldsymbol{\Xi}_k$, where $\boldsymbol{\Phi}_{p_0}$ denotes the space of the parameters $\boldsymbol{\beta}_i^{(p_0)}$, σ_i^2 for the ith segment, i.e., $\boldsymbol{\Phi}_0 \triangleq \mathbb{R}^+$, $\boldsymbol{\Phi}_{p_0} \triangleq (\mathbb{R}^{p_0} \times \mathbb{R}^+)$, $\boldsymbol{\Xi}_k$ denotes the hyperparameter $\boldsymbol{\xi}_k \triangleq (\delta_0^2, \gamma_\sigma)$ space, which is given by $\boldsymbol{\Xi}_k \triangleq \mathbb{R}^+ \times \mathbb{R}^+$, and $k_{\max} = T-2$ ($\boldsymbol{\Upsilon}_k$ is defined in Section II-A1).

[2]It is worth noticing that from the algorithmic point of view, this model allows for the faster updating of the Markov chain due to conditional independence between the coefficient and regression variance parameters on the hyperparameters.

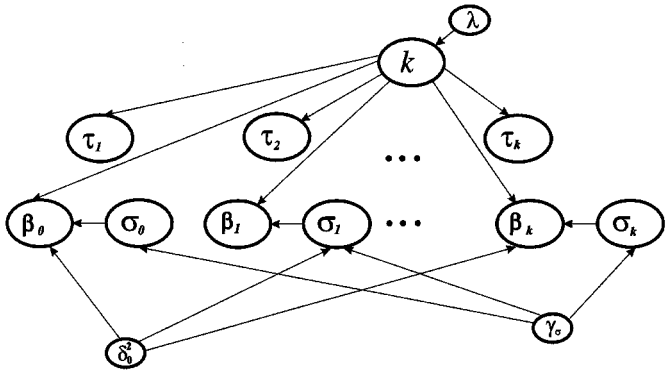

Fig. 1. DAG for the prior distribution.

There is a natural hierarchical structure for this setup, which we can formalize by modeling the joint distribution of all variables as

$$p(k, \boldsymbol{\Psi}_k, \lambda, \boldsymbol{\xi}_k, \mathbf{y}_{0:T-1}) = p(k, \boldsymbol{\Psi}_k, \lambda, \boldsymbol{\xi}_k)\, p(\mathbf{y}_{0:T-1} | k, \boldsymbol{\Psi}_k).$$

As the noise is assumed to be i.i.d. Gaussian (Section I), the likelihood takes the form

$$p(\mathbf{y}_{0:T-1} | k, \boldsymbol{\Psi}_k) = \prod_{i=0}^{k} (2\pi\sigma_i^2)^{-((\tau_{i+1}-\tau_i)/2)} \times \exp\left(-\frac{\left\| \mathbf{y}_{\tau_i:\tau_{i+1}-1} - \mathbf{G}_i^{(p_0)} \boldsymbol{\beta}_i^{(p_0)} \right\|^2}{2\sigma_i^2} \right)$$

where $\|\cdot\|$ is the Euclidean norm.

2) Bayesian Detection and Estimation: Any Bayesian inference on k and $\boldsymbol{\Psi}_k$, λ, $\boldsymbol{\xi}_k$ is based on the following posterior obtained using Bayes' theorem:

$$p(k, \boldsymbol{\Psi}_k, \lambda, \boldsymbol{\xi}_k | \mathbf{y}_{0:T-1}) \propto p(\mathbf{y}_{0:T-1} | k, \boldsymbol{\Psi}_k)\, p(k, \boldsymbol{\Psi}_k, \lambda, \boldsymbol{\xi}_k).$$

Our aim is to estimate this posterior distribution and, more specifically, some of its features such as the marginal distributions. In our case, however, it is not possible to obtain these quantities analytically, as it requires the evaluation of high-dimensional integrals of nonlinear functions in the parameters. Therefore, we apply MCMC methods and a reversible jump MCMC method in particular (see Section II-B for details). The key idea is to build an ergodic Markov chain $(k^{(j)}, \boldsymbol{\Psi}_k^{(j)}, \lambda^{(j)}, \boldsymbol{\xi}_k^{(j)})_{j\in\mathbb{N}}$ whose equilibrium distribution is the desired posterior distribution. Under weak additional assumptions, the $P \gg 1$ samples generated by the Markov chain are asymptotically distributed according to the posterior distribution and can thus be used to easily evaluate all posterior features of interest.

The proposed Bayesian model allows for the integration of the nuisance parameters $(\{\boldsymbol{\beta}_i^{(p_0)}\}_{i=0}^k, \boldsymbol{\sigma}_k^2)$ and hyperparameter λ. The resultant expression for $p(k, \boldsymbol{\tau}_k, \boldsymbol{\xi}_k | \mathbf{y}_{0:T-1})$ up to a normalizing constant is (here $\|\mathbf{y}\|_{\mathbf{A}}^2 \triangleq \mathbf{y}^{\mathrm{T}}\mathbf{A}\mathbf{y}$)

$$\begin{aligned} p(k, \boldsymbol{\tau}_k, \boldsymbol{\xi}_k | \mathbf{y}_{0:T-1}) &\propto \prod_{i=0}^{k} \left[\left| \mathbf{M}_i^{(p_0)} \right|^{1/2} \times \left(\gamma_\sigma + \left\| \mathbf{y}_{\tau_i:\tau_{i+1}-1} \right\|_{\mathbf{P}_i^{(p_0)}}^2 \right)^{-((\nu_\sigma+\tau_{i+1}-\tau_i)/2)} \right] \\ &\times \prod_{i=0}^{k} \left[2\pi^{-((\tau_{i+1}-\tau_i)/2)} \times \frac{\Gamma\left(\frac{\nu_\sigma+\tau_{i+1}-\tau_i}{2}\right)}{\Gamma\left(\frac{\nu_\sigma}{2}\right)} (\gamma_\sigma)^{\nu_\sigma/2} \right] \mathbb{I}_{\boldsymbol{\Upsilon}_k}(\boldsymbol{\tau}_k) \\ &\times \frac{\Gamma(k+1)\Gamma(T-k-1)}{\gamma_\sigma^{-1}} (\delta_0^2)^{-\alpha_\delta-(kp_0/2)-1} \times \exp\left(-\frac{\theta_{\delta_0}}{\delta_0^2}\right) \end{aligned} \tag{5}$$

with

$$\begin{aligned} \mathbf{M}_i^{(p_i)} &= \left[\mathbf{G}_i^{(p_i)\mathrm{T}} \mathbf{G}_i^{(p_i)} + \frac{1}{\delta_i^2} \mathbf{I}_{p_i} \right]^{-1} \\ \mathbf{m}_i^{(p_i)} &= \mathbf{M}_i^{(p_i)} \mathbf{G}_i^{(p_i)\mathrm{T}} \mathbf{y}_{\tau_i:\tau_{i+1}-1} \\ \mathbf{P}_i^{(p_i)} &= \mathbf{I}_{\tau_i:\tau_{i+1}-1} - \mathbf{G}_i^{(p_i)} \mathbf{M}_i^{(p_i)} \mathbf{G}_i^{(p_i)\mathrm{T}} \end{aligned} \tag{6}$$

where, again, $p_i = p_0$ and $\delta_i^2 = \delta_0^2$ for the fixed model order case.

It has already been pointed out that this posterior distribution is complex in the parameters $(k, \boldsymbol{\tau}_k, \boldsymbol{\xi}_k)$ and that the posterior model probability $p(k|\mathbf{y}_{0:T-1})$ cannot be determined analytically. In the next section, we develop a method to estimate $p(k, \boldsymbol{\tau}_k, \boldsymbol{\xi}_k|\mathbf{y}_{0:T-1})$ or, if needed

$$p\left(k, \boldsymbol{\tau}_k, \left\{ \boldsymbol{\beta}_i^{(p_0)} \right\}_{i=0}^k, \boldsymbol{\sigma}_k^2, \lambda, \boldsymbol{\xi}_k \middle| \mathbf{y}_{0:T-1} \right).$$

Once the approximation of $p(k, \boldsymbol{\tau}_k, \boldsymbol{\xi}_k|\mathbf{y}_{0:T-1})$ is obtained, the number of changepoints and their positions can be easily estimated according to the MAP criterion

$$(\hat{k}, \hat{\boldsymbol{\tau}}_k, \hat{\boldsymbol{\xi}}_k) = \underset{(k, \boldsymbol{\tau}_k, \boldsymbol{\xi}_k)}{\arg\max}\ \hat{p}(k, \boldsymbol{\tau}_k, \boldsymbol{\xi}_k | \mathbf{y}_{0:T-1})$$

where $\hat{k}, \hat{\boldsymbol{\tau}}_k$ and $\hat{\boldsymbol{\xi}}_k$ is the corresponding estimates. Alternatively, one can compute the minimum mean square error (MMSE) estimate of the regression function using Bayesian model averaging.

B. MCMC Algorithm

The problem addressed here is, in fact, a model uncertainty problem of variable dimensionality in terms of the number of changepoints. It can be treated efficiently using reversible jump MCMC method [11]. This method extends the traditional Metropolis–Hastings algorithm to the case where moves

from one dimension to another are proposed and accepted with some probability. This probability should be designed in a special way in order to preserve reversibility and thus ensure that $p(k, \tau_k, \boldsymbol{\xi}_k|\mathbf{y}_{0:T-1})$ is the invariant distribution of the Markov chain (MC). In general, if we propose a move from the model (k) with parameters $(\tau_k, \boldsymbol{\xi}_k)$ to the model (k') with parameters $(\boldsymbol{\tau}_{k'}, \xi_{k'})$ using a proposal distribution $q(k', \boldsymbol{\tau}_{k'}, \boldsymbol{\xi}_{k'}|k, \boldsymbol{\tau}_k, \boldsymbol{\xi}_k)$, the acceptance probability is given by

$$\alpha = \min\left\{1, \frac{p(k', \boldsymbol{\tau}_{k'}, \boldsymbol{\xi}_{k'}|\mathbf{y}_{0:T-1})}{p(k, \boldsymbol{\tau}_k, \boldsymbol{\xi}_k|\mathbf{y}_{0:T-1})} \times \frac{q(k, \tau_k, \boldsymbol{\xi}_k|k', \boldsymbol{\tau}_{k'}, \boldsymbol{\xi}_{k'})}{q(k', \boldsymbol{\tau}_{k'}, \boldsymbol{\xi}_{k'}|k, \boldsymbol{\tau}_k, \boldsymbol{\xi}_k)}\right\}. \quad (7)$$

Here, the proposal is made directly in the new parameter space rather than via "dimensional" matching random variables [11], and the Jacobian term is equal to 1; see [13] and [14] for a detailed introduction.

In fact, a particular choice of the moves will only affect the convergence rate of the algorithm. To ensure a low level of rejection, we want the proposed "jumps" to be small; therefore, the following moves have been selected:

- birth of a changepoint (proposing a new changepoint at random);
- death of a changepoint (removing a changepoint chosen randomly);
- update of the changepoint positions (proposing a new position for each of the existing changepoints).

At each iteration, one of the moves described above is randomly chosen with probabilities b_k, d_k, and u_k such that $b_k + d_k + u_k = 1$ for all $0 \leq k \leq k_{\max}$. For $k = 0$, the death of a changepoint is impossible, and for $k = k_{\max}$, the birth is impossible; thus, $d_0 \triangleq 0$, $b_{k_{\max}} \triangleq 0$. Otherwise, we choose $b_k = d_k = u_k$. After each move, an update of the hyperparameters is performed. We now describe the main steps of the algorithm.

Reversible Jump MCMC Algorithm (Main Procedure)

1) Initialize $(k^{(0)}, \boldsymbol{\tau}_k^{(0)}, \boldsymbol{\xi}_k^{(0)}) \in \boldsymbol{\Theta}$. Set $j = 1$.
2) If $(u \sim \mathcal{U}_{(0,1)}) \leq b_{k^{(j)}}$ then birth of a new changepoint (Section II-B1);
   else if $u \leq b_{k^{(j)}} + d_{k^{(j)}}$ then death of a changepoint (Section II-B1);
   else update the changepoints positions (Section II-B2).
3) Update of the hyperparameters (Section II-B3).
4) $j \leftarrow j + 1$ and goto 2.

We now detail the steps of the algorithm. To simplify the notation, we drop the superscript (j) from all variables at iteration j.

1) Death/Birth of the Changepoints: First, let the current state of the MC be $(k+1, \boldsymbol{\tau}_{k+1}, \delta_0^2, \gamma_\sigma)$, and consider the death move, which implies a modification of the dimension of the model, respectively, from $k+1$ to k. Our proposal begins by choosing a changepoint to be removed among $k+1$ existing ones. If the move is accepted, then two segments $(l-1)$th and lth will be merged, thus reducing $k+1$ by 1, and a new segment will be created (see Fig. 2).

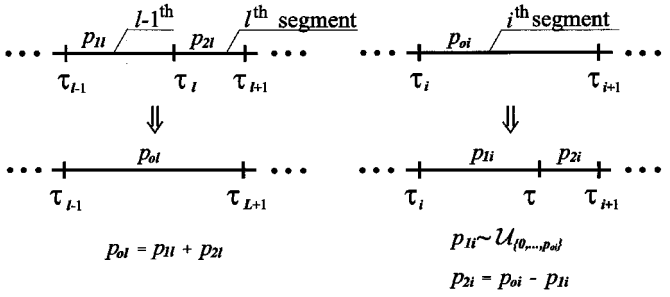

Fig. 2. Death (left) and birth (right) moves.

Algorithm for the Death Move

- Choose a changepoint to be removed: $l \sim \mathcal{U}_{\{1,\ldots,k+1\}}$.
- Evaluate α_{death}; see (8).
- If $(u_d \sim \mathcal{U}_{(0,1)}) \leq \alpha_{death}$, the new MC state is accepted.

For the birth move $(k \rightarrow k+1)$, again, the position of a new changepoint τ is first proposed, which means that the ith segment (for $\tau_i < \tau < \tau_{i+1}$) should be split into two if the move is accepted. Assuming that the current state of the MC is $(k, \tau_k, \delta_0^2, \gamma_\sigma)$, we have the following ($\{1, \ldots, T-2\}\backslash\{\boldsymbol{\tau}_k\} \triangleq \bigcup_{j=0}^{k}\{\tau_j + 1, \ldots, \tau_{j+1} - 1\}$ here).

Algorithm for the Birth Move

- Propose a new changepoint position $\tau \sim \mathcal{U}_{\{1,\ldots,T-2\}\backslash\{\boldsymbol{\tau}_k\}}$.
- Evaluate α_{birth}; see (8).
- If $(u_b \sim \mathcal{U}_{(0,1)}) \leq \alpha_{birth}$, the new state of the MC is accepted.

The acceptance ratio of the birth and death (of a changepoint) moves are deduced from the general expression (7), and the corresponding acceptance probabilities are

$$\alpha_{birth} = \min\{1, r_{birth}\}, \quad \alpha_{death} = \min\{1, r_{birth}^{-1}\} \quad (8)$$

$$r_{birth} = \frac{p(k+1, \boldsymbol{\tau}_{k+1}, \delta_0^2, \gamma_\sigma|\mathbf{y}_{0:T-1})}{p(k, \boldsymbol{\tau}_k, \delta_0^2, \gamma_\sigma|\mathbf{y}_{0:T-1})} \frac{q(k, \boldsymbol{\tau}_k|k+1, \boldsymbol{\tau}_{k+1})}{q(k+1, \tau_{k+1}|k, \boldsymbol{\tau}_k)}.$$

For the birth of the changepoint τ, $(\tau_i \leq \tau < \tau_{i+1})$, we obtain from (5)

$$r_{birth} = \frac{f(\tau_i, \tau, p_0) f(\tau, \tau_{i+1}, p_0)}{f(\tau_i, \tau_{i+1}, p_0)} \frac{2(\gamma_\sigma)^{\nu_\sigma/2}(\delta_0^2)^{-(p_0/2)}}{\Gamma\left(\frac{\nu_\sigma}{2}\right)} \frac{d_{k+1}}{b_k}$$

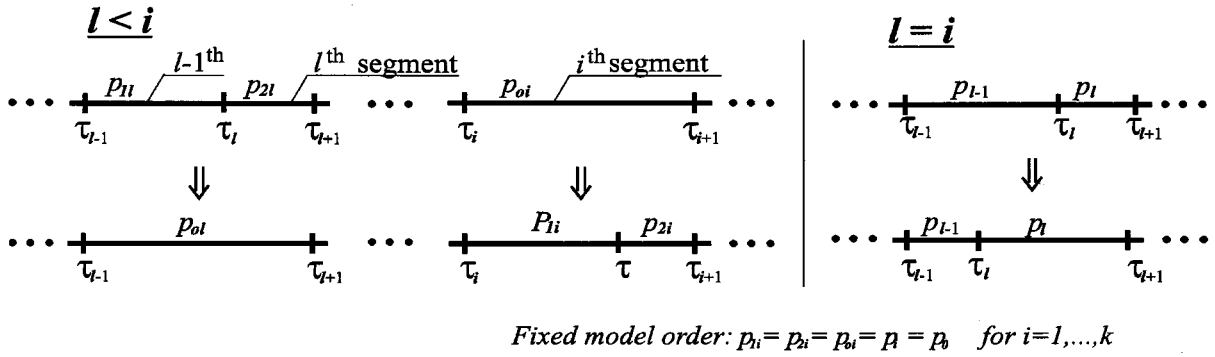

Fig. 3. Update of the changepoint positions.

where, for convenience, we denote for the segment between the changepoints τ_i and τ_{i+1}

$$f(\tau_i, \tau_{i+1}) \triangleq \left|\mathbf{M}_i^{(p_i)}\right|^{1/2} \Gamma\left(\frac{\nu_\sigma + \tau_{i+1} - \tau_i}{2}\right) \times \left[\gamma_\sigma + \left\|\mathbf{y}_{\tau_i:\tau_{i+1}-1}\right\|^2_{\mathbf{P}_i^{(p_i)}}\right]^{-((\nu_\sigma+\tau_{i+1}-\tau_i)/2)}. \quad (9)$$

2) Update of the Changepoint Positions: Although the update of the changepoint positions does not involve a change in dimension, it is somewhat more complicated than the birth/death moves. In fact, updating the position of changepoint τ_l means removing the lth changepoint and proposing instead a new one τ (this approach also facilitates the extension of the algorithm to the more complex case of unknown model orders for each segment). We determine i such that $\tau_i < \tau < \tau_{i+1}$, and it is worth noticing that if $i \neq l$, the update move may actually be described as a combination of the birth of the changepoint τ and the death of the changepoint τ_l (see Fig. 3). Otherwise, we just update the position within the same segment. This process is repeated for all existing changepoints $l = 1, \ldots, k$ and is described later in more detail.

Algorithm for the Update of the Changepoint Positions

For $l = 1, \ldots, k$

- Propose a new position for the lth changepoint $\tau \sim \mathcal{U}_{\{1,\ldots,T-2\}\setminus\{\boldsymbol{\tau}_k\}}$; determine i such that $\tau_i < \tau < \tau_{i+1}$.
- Evaluate α_{update}, if $l \neq i$ then see (10) else see (11).
- If $(u_u \sim \mathcal{U}_{(0,1)}) \leq \alpha_{update}$ then the new state of the MC is accepted.

Since for $i \neq l$ the update of the positions of changepoints combines the birth of the ith changepoint and death of the lth changepoint at the same time, the acceptance ratio for the proposed move is given by

$$\alpha_{update} = \min\{1, r_{update}\} = \min\{1, r^i_{birth} r^l_{death}\}. \quad (10)$$

If $l = i$, it becomes

$$r_{update} = \frac{f(\tau_{l-1}, \tau) f(\tau, \tau_{l+1})}{f(\tau_{l-1}, \tau_l) f(\tau_l, \tau_{l+1})} \quad (11)$$

where $f(\cdot)$ is defined in (9).

We have also used a Metropolis–Hasting update with random walk proposal to perform a local exploration of the space.

3) Update of the Hyperparameters: The algorithm developed requires the simulation of the hyperparameters δ_0^2 and γ_σ. This can be done according to standard Gibbs moves [15] so that δ_0^2 and γ_σ are sampled from Inverse-Gamma and Gamma distributions, respectively.

$$\delta_0^2 \sim \mathcal{IG}\left(\alpha_\delta + \frac{kp_0}{2}, \theta_\delta + \sum_{i=0}^{k} \frac{\boldsymbol{\beta}_i^{(p_0)\mathrm{T}} \boldsymbol{\beta}_i^{(p_0)}}{2\sigma_i^2}\right) \quad (12)$$

$$\gamma_\sigma \sim \mathcal{G}a\left(\frac{\nu_\sigma(k+1)}{2}, \sum_{i=1}^{k} \frac{1}{2\sigma_i^2}\right). \quad (13)$$

The probability distribution allowing the update of δ_0^2 requires the simulation of the nuisance parameters $\boldsymbol{\beta}_i^{(p_0)}$, σ_i^2, which are, in turn, sampled as

$$\sigma_i^2 \sim \mathcal{IG}\left(\frac{\nu_\sigma + \tau_{i+1} - \tau_i}{2}, \frac{\gamma_\sigma + \left\|\mathbf{y}_{\tau_i:\tau_{i+1}-1}\right\|^2_{\mathbf{P}_i^{(p_i)}}}{2}\right) \quad (14)$$

$$\boldsymbol{\beta}_i^{(p_i)} \sim \mathcal{N}\left(\mathbf{m}_i^{(p_i)}, \sigma_i^2 \mathbf{M}_i^{(p_i)}\right) \quad (15)$$

with $p_i = p_0$, $\mathbf{P}_i^{(p_i)} = \mathbf{P}_i^{(p_0)}$ (as we are considering the fixed model order case). Thus, assuming that the current state of the MC is $(k, \boldsymbol{\tau}_k, \delta_0^2, \gamma_\sigma)$, the update of the hyperparameters is performed according to the following algorithm.

Algorithm for the Update of the Hyperparameters

- Update of δ_0^2
  sample $\sigma_i^2|(k, \boldsymbol{\Psi}_k, \boldsymbol{\xi}_k, \mathbf{y}_{0:T-1})$ and $\boldsymbol{\beta}_i^{(p_0)}|(k, \boldsymbol{\Psi}_k, \boldsymbol{\xi}_k, \mathbf{y}_{0:T-1})$, see (14) and (15).
  sample $\delta_0^2|(k, \tau_k, \mathbf{y}_{0:T-1}, \boldsymbol{\beta}_1^{(p_0)}, \ldots \boldsymbol{\beta}_k^{(p_0)}, \boldsymbol{\sigma}_k^2)$ see (12).
- Sample $\gamma_\sigma|(k, \tau_k, \mathbf{y}_{0:T-1}, \boldsymbol{\sigma}_k^2)$ see (13).

III. Bayesian Inference for Unknown Model Dimensions

In the previous section, we addressed the problem of segmentation under the assumption that $p_i = p_0$, for $i = 0, \ldots, k$, with known p_0. However, in many applications, different model orders should be considered for different segments, and these model orders should also be estimated from the data available. We now address this difficult problem.

A. Extended Bayesian Model

In Section II-A, an original Bayesian model was proposed for the case of fixed model orders for each segment. The steps analogous to those taken in that section can yield an extended Bayesian model whereby the unknown parameters, including the orders of the models for different segments, are regarded as being drawn from appropriate prior distributions.

1) Hierarchical Structure for the Prior: Here, we adopt a truncated Poisson distribution for the model order[3] p_i

$$p(p_i|\psi) = \frac{\psi^{p_i}}{C_{p_{\max}} p_i!}\, \mathbb{I}_{\{0,\ldots,p_{\max}\}}(p_i), \quad C_{p_{\max}} = \sum_{p_i=0}^{p_{\max}} \frac{\psi^{p_i}}{p_i!}$$

where the mean ψ is interpreted as the expected (approximate as $p_{\max} < \infty$) number of model parameters.

For the parameters k, $\boldsymbol{\tau}_k$, $\boldsymbol{\beta}_i^{(p_i)}$, and σ_i^2, we assign priors similar to the ones introduced in Section II-A1 [see (2)–(4)], with the only exception that the hyperparameter δ_i^2 is now different for the different segments $\boldsymbol{\delta}_k^2 \triangleq \boldsymbol{\delta}_{0:k}^2$, although it is still assumed to be drawn from the Inverse-Gamma distribution

$$\delta_i^2 \big|\, (\alpha_\delta, \theta_\delta) \sim \mathcal{IG}(\alpha_\delta, \theta_\delta), \qquad i = 0, \ldots, k$$

with $\alpha_\delta = 1$. However, since in our particular case the Bayes factors depends on the hyperparameter θ_δ [see (22)], we assume that θ_δ is also randomly distributed according to a conjugate prior Gamma distribution to make the prior more robust

$$\theta_\delta \big|\, (\zeta_\theta, \chi_\theta) \sim \mathcal{G}a(\zeta_\theta, \chi_\theta)$$

with $\zeta_\theta = 1$ and fixed $\chi_\theta = \epsilon_\theta$, ($\epsilon_\theta \ll 1$). Similarly, we assign a conjugate prior Gamma density to ψ:

$$\psi \big|\, (\zeta_\psi, \chi_\psi) \sim \mathcal{G}a(\zeta_\psi, \chi_\psi)$$

where $\zeta_\psi = 1$ and $\chi_\psi = \epsilon_\psi$, ($\epsilon_\psi \ll 1$). Again, a uniform prior distribution and a noninformative Jeffreys' priors are chosen for λ and γ_σ, correspondingly.

As a result, the following extended hierarchical structure is assumed for the prior of the parameters.

$$\begin{aligned} &p\left(k, \boldsymbol{\Psi}_{k,\mathbf{p}_k}, \lambda, \psi, \boldsymbol{\delta}_k^2, \gamma_\sigma, \theta_\delta\right) \\ &\quad = \prod_{i=0}^{k} \left[p\left(p_i \big|\, \psi\right)\right] p(\psi) \\ &\qquad \times \prod_{i=0}^{k} \left[p\left(\boldsymbol{\beta}_i^{(p_i)} \Big|\, \sigma_i^2, \delta_i^2\right) p\left(\sigma_i^2 \big|\, \gamma_\sigma\right) p\left(\delta_i^2 \big|\, \theta_\delta\right)\right] \\ &\qquad \times p\left(k, \boldsymbol{\tau}_k \big|\, \lambda\right) p(\lambda) p(\gamma_\sigma) p(\theta_\delta) \end{aligned} \tag{16}$$

[3]In fact, any other discrete probability distribution may be adopted as a prior for p_i. In addition, the priors dependent on the number of changepoints can be introduced.

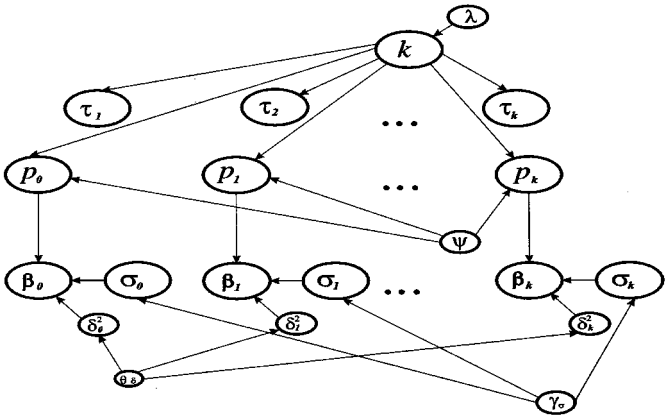

Fig. 4. DAG for the prior distribution.

which can be visualized with a DAG, as shown in Fig. 4 (for convenience, we do not show fixed parameters ν_σ, α_δ, ζ_ψ, χ_ψ, ζ_θ, χ_θ).

2) Bayesian Inference: As was mentioned in Section II-A2, the Bayesian inference on the unknown parameters k, $\boldsymbol{\Psi}_{k,\mathbf{p}_k}$, λ, and $\boldsymbol{\xi}_{k,\mathbf{p}_k}$ [where $\xi_{k,\mathbf{p}_k} \triangleq (\psi, \boldsymbol{\delta}_k^2, \gamma_\sigma, \theta_\delta)$] is based on the posterior probability distribution $p(k, \boldsymbol{\Psi}_{k,\mathbf{p}_k}, \lambda, \boldsymbol{\xi}_{k,\mathbf{p}_k} | \mathbf{y}_{0:T-1})$

$$\begin{aligned} &p\left(k, \boldsymbol{\Psi}_{k,\mathbf{p}_k}, \lambda, \boldsymbol{\xi}_{k,\mathbf{p}_k} \big|\, \mathbf{y}_{0:T-1}\right) \\ &\quad \propto p\left(\mathbf{y}_{0:T-1} \big|\, k, \boldsymbol{\Psi}_{k,\mathbf{p}_k}\right) p\left(k, \boldsymbol{\Psi}_{k,\mathbf{p}_k}, \lambda, \boldsymbol{\xi}_{k,\mathbf{p}_k}\right). \end{aligned} \tag{17}$$

As in the fixed model order case, the parameters $(\{\boldsymbol{\beta}_i^{(p_i)}\}_{i=0}^k, \boldsymbol{\sigma}_k^2)$ and hyperparameter λ can be integrated out giving the marginalized expression for $p(k, \boldsymbol{\tau}_k, \mathbf{p}_k, \xi_{k,\mathbf{p}_k} | \mathbf{y}_{0:T-1})$

$$\begin{aligned} &p\left(k, \boldsymbol{\tau}_k, \mathbf{p}_k, \boldsymbol{\xi}_k \big|\, \mathbf{y}_{0:T-1}\right) \\ &\quad \propto \Gamma(k+1)\Gamma(T-k-1) \\ &\qquad \times \prod_{i=0}^{k} \left[\left|\mathbf{M}_i^{(p_i)}\right|^{1/2} \right. \\ &\qquad\qquad \left. \times \left(\gamma_\sigma + \left\|\mathbf{y}_{\tau_i:\tau_{i+1}-1}\right\|^2_{\mathbf{P}_i^{(p_i)}}\right)^{-((\nu_\sigma+\tau_{i+1}-\tau_i)/2)} \right] \\ &\qquad \times \prod_{i=0}^{k} \left[\Gamma\left(\frac{\nu_\sigma + \tau_{i+1} - \tau_i}{2}\right) \frac{\psi^{p_i}}{C_{p_{\max}} p_i!} \right] \psi^{\zeta_\psi - 1} \\ &\qquad \times \exp\left(-\chi_\psi \psi\right) \\ &\qquad \times \prod_{i=0}^{k} \left[2\pi^{-((\tau_{i+1}-\tau_i)/2)} \frac{(\gamma_\sigma)^{\nu_\sigma/2}}{\Gamma\left(\frac{\nu_\sigma}{2}\right)} \right. \\ &\qquad\qquad \left. \times \frac{\theta_\delta^{\alpha_\delta} \left(\delta_i^2\right)^{-\alpha_\delta - (p_i/2) - 1}}{\Gamma(\alpha_\delta)} \exp\left(-\frac{\theta_\delta}{\delta_i^2}\right) \right] \\ &\qquad \times \theta_\delta^{\zeta_\theta - 1} \exp\left(-\chi_\theta \theta_\delta\right) \gamma_\sigma^{-1} \mathbb{I}_{\boldsymbol{\Upsilon}_k}(\boldsymbol{\tau}_k) \mathbb{I}_{\mathbf{F}_k}(\mathbf{p}_k) \end{aligned} \tag{18}$$

where $\mathbf{F}_k \triangleq \{0, \ldots, p_{\max}\}^k$ [see also (6) for $\mathbf{M}_i^{(p_i)}$, $\mathbf{m}_i^{(p_i)}$, and $\mathbf{P}_i^{(p_i)}$].

The resulting posterior distribution again appears highly nonlinear in its parameters, thus precluding analytical calculations,

and MCMC methods must be employed in order to evaluate the posterior features of interest.[4]

B. MCMC Algorithm

In the case where the orders of the models for each segment are unknown, Bayesian computation for the estimation of the joint posterior distribution $p(k, \boldsymbol{\tau}_k, \mathbf{p}_k, \boldsymbol{\xi}_{k,\mathbf{p}_k} | \mathbf{y}_{0:T-1})$ is even more complex. Here, a "double" model selection, in terms of both the number of changepoints k and the model orders p_i, should be performed. Therefore, an MCMC sampler capable of "jumping" between subspaces of variable dimensionality in terms of both k and p_i, $i = 0, \ldots, k$ should be constructed. In order to be able to sample directly from the joint distribution on $\boldsymbol{\Theta} \triangleq \bigcup_{k=0}^{k_{\max}} \{k\} \times \boldsymbol{\Theta}_k$ with $\boldsymbol{\Theta}_k \triangleq \boldsymbol{\Upsilon}_k \times \prod_{i=0}^{k} \bigcup_{p_i=0}^{p_{\max}} \{p_i\} \times \boldsymbol{\Xi}_k$ ($\boldsymbol{\Xi}_k$ denotes the hyperparameter space), we propose a reversible jump MCMC method [11].

The procedure is similar to the one described in Section II-B. The moves from the model $(k, \mathbf{p}_k)$ with parameters $(\boldsymbol{\tau}_k, \boldsymbol{\xi}_{k,\mathbf{p}_k})$ to the model $(k', \mathbf{p}_{k'})$ with parameters $(\boldsymbol{\tau}_{k'}, \boldsymbol{\xi}_{k',\mathbf{p}_{k'}})$ are generated using a proposal distribution $q(k', \boldsymbol{\tau}_{k'}, \mathbf{p}_{k'}, \boldsymbol{\xi}_{k',\mathbf{p}_{k'}} | k, \boldsymbol{\tau}_k, \mathbf{p}_k, \boldsymbol{\xi}_{k,\mathbf{p}_k})$ and are randomly accepted according to the acceptance probability $\alpha_{k,\mathbf{p}_k}$

$$\alpha_{k,\mathbf{p}_k} = \min\left\{1, \frac{p\left(k', \boldsymbol{\tau}_{k'}, \mathbf{p}_{k'}, \boldsymbol{\xi}_{k',\mathbf{p}_{k'}} \middle| \mathbf{y}_{0:T-1}\right)}{p\left(k, \boldsymbol{\tau}_k, \mathbf{p}_k, \boldsymbol{\xi}_{k,\mathbf{p}_k} \middle| \mathbf{y}_{0:T-1}\right)} \times \frac{q\left(k, \boldsymbol{\tau}_k, \mathbf{p}_k, \boldsymbol{\xi}_{k,\mathbf{p}_k} \middle| k', \boldsymbol{\tau}_{k'}, \mathbf{p}_{k'}, \boldsymbol{\xi}_{k',\mathbf{p}_{k'}}\right)}{q\left(k', \boldsymbol{\tau}_{k'}, \mathbf{p}_{k'}, \boldsymbol{\xi}_{k',\mathbf{p}_{k'}} \middle| k, \boldsymbol{\tau}_k, \mathbf{p}_k, \boldsymbol{\xi}_{k,\mathbf{p}_k}\right)}\right\}. \quad (19)$$

In particular, moves relative to birth, death of the changepoints, or update of their positions are randomly chosen with probabilities b_k, d_k, and u_k such that $b_k + d_k + u_k = 1$ for all $0 \le k \le k_{\max}$; b_k, d_k, and u_k are the same as in Section II-B. In addition, due to "double" variable dimensionality, an update of the model order for each of the segments is performed after each of the "changepoint" moves. Thus, the algorithm proceeds as follows.

Reversible Jump MCMC Algorithm (Main Procedure)

1) Initialize $(k^{(0)}, \boldsymbol{\tau}_k^{(0)}, \mathbf{p}_k^{(0)}, \boldsymbol{\xi}_{k,\mathbf{p}_k}^{(0)}) \in \boldsymbol{\Theta}$. Set $j = 1$.
2) If $(u \sim \mathcal{U}_{(0,1)}) \le b_{k^{(j)}}$ then birth of a new changepoint (Section III-B1);
   else if $u \le b_{k^{(j)}} + d_{k^{(j)}}$ then death of a changepoint (Section II-B1);
   else update the changepoints positions (Section III-B1).
3) Update of the model orders and hyperparameters (Section II-B2).
4) $j \leftarrow j+1$ and goto 2.

■

[4]We could also integrate out the parameter β_δ, but this increases significantly the computational complexity of the resulting MCMC sampler.

The different steps of the algorithm are described in the following [the superscript (j) from all variables at iteration j is dropped].

1) Changepoint Moves: The algorithms for the birth, death of the changepoints, and update of their positions presented in Section II-B-1 can be easily extended to the case where the p_i are unknown. The main difficulty here is to choose the proposal for the new model orders. We employ the following approach. If two segments $(l-1)$th and lth are to be merged, the model order for the new segment p_{ol} is the sum of the model orders of the two original segments, i.e., $p_{ol} = p_{1l} + p_{2l}$, where p_{1l}, p_{2l} are the model orders of the existing $(l-1)$th and lth segments. If the ith segment is to be split, one of the new model orders is selected randomly, $p_{1i} \sim \mathcal{U}_{\{0,\ldots,p_{oi}\}}$, and another one is set equal to $p_{2i} = p_{oi} - p_{1i}$, where p_{oi} is the order of the original ith model (see Fig. 2). The latter ensures that the birth/death moves are reversible ($p_{oi} = p_{1i} + p_{2i}$). The update move, as was mentioned before, is performed as a combination of the birth and death of changepoints (see Fig. 3). However, if, during the update, the position of a changepoint with respect to the other changepoints does not change (it stays between the same changepoints as it was before), we do not update the order of the models.

In addition, we should also sample the hyperparameter δ_i^2 for the new segments created when removing or adding a changepoint (recall that δ_i^2 is different from segment to segment). We select as proposal distribution for δ_i^2

$$\mathcal{IG}\left(\delta_i^2; \tilde{\alpha}_\delta, \tilde{\theta}_\delta\right); \quad \tilde{\alpha}_\delta = \alpha_\delta + \frac{p_i}{2}, \quad \tilde{\theta}_\delta = \theta_\delta + \frac{\overline{\boldsymbol{\beta}}_i^{(p_i)\mathrm{T}} \overline{\boldsymbol{\beta}}_i^{(p_i)}}{2\overline{\sigma}_i^{\,2}} \quad (20)$$

where $\overline{\boldsymbol{\beta}}_i^{(p_i)}$, $\overline{\sigma}_i^{\,2}$ are the means of the distributions given by (14) and (15) but with matrices $\mathbf{M}_i^{(p_i)*}$, $\mathbf{P}_i^{(p_i)*}$ and $\mathbf{m}_i^{(p_i)*}$ corresponding to the value of the hyperparameter $\delta_i^{2*} = \theta_\delta/(\alpha_\delta + p_i/2 - 1)$ [the mean of the distribution $\mathcal{IG}(\alpha_\delta + p_i/2, \theta_\delta)$] (see [16] for details)

$$\overline{\boldsymbol{\beta}}_i^{(p_i)} = \mathbf{m}_i^{(p_i)*}, \quad \overline{\sigma}_i^{\,2} = \frac{\gamma_\sigma + \mathbf{y}_{\tau_i:\tau_{i+1}-1}^{\mathrm{T}} \mathbf{P}_i^{(p_i)*} \mathbf{y}_{\tau_i:\tau_{i+1}-1}}{\nu_\sigma + \tau_{i+1} - \tau_i - 2}. \quad (21)$$

The acceptance probabilities for the birth and death moves are as in (19)

$$\alpha_{birth} = \min\{1, r_{birth}\}, \; \alpha_{death} = \min\{1, r_{birth}^{-1}\} \quad (22)$$

where from (18) for the birth of the changepoint τ, $(\tau_i \le \tau < \tau_{i+1})$, we obtain

$$r_{birth} = 2\frac{(\gamma_\sigma)^{\nu_\sigma/2}}{\Gamma\left(\frac{\nu_\sigma}{2}\right)} \frac{d_{k+1}}{b_k} (p_{oi} + 1) \times \frac{f(\tau_i, \tau, p_{1i}, \delta_{1i}^2) f(\tau, \tau_{i+1}, p_{2i}, \delta_{2i}^2)}{f(\tau_{i+1}, \tau_i, p_i, \delta_i^2)} \quad (23)$$

with

$$f(\tau_i, \tau_{i+1}, p_i, \delta_i^2) = f(\tau_i, \tau_{i+1}, p_i) \times \frac{\psi^{p_i}}{C_{p_{\max}} p_i!} \frac{\theta_\delta^{\alpha_\delta}}{\theta_\delta^{\tilde{\alpha}_\delta}} \frac{\Gamma(\tilde{\alpha}_\delta)}{\Gamma(\alpha_\delta)} \exp\left(-\frac{\theta_\delta - \tilde{\theta}_\delta}{\delta_i^2}\right). \quad (24)$$

The acceptance probability for the update of the changepoint positions is given by (10) if $l \neq i$. For $l = i$, it becomes

$$r_{update} = \frac{f(\tau_{l-1}, \tau, p_{l-1}, \delta_{1l}^2) f(\tau, \tau_{l+1}, p_l, \delta_{2l}^2)}{f(\tau_{l-1}, \tau_l, p_{l-1}, \delta_{l-1}^2) f(\tau_l, \tau_{l+1}, p_l, \delta_l^2)}. \quad (25)$$

The algorithms for these moves are presented in more detail in [16].

2) Model Orders Update: The update of the model orders for each segment does not involve changing the number of changepoints or their positions. However, we still have to perform "jumps" between the subspaces of different dimensions p_i and will therefore continue using the reversible jump MCMC method, although it is formulated now in a less complicated form. Similarly, the moves are chosen to be

1) birth of the model parameter ($p_i \rightarrow p_i + 1$);
2) death of the model parameter ($p_i \rightarrow p_i - 1$);
3) update of the hyperparameter δ_i^2.

The probabilities for choosing these moves are defined in exactly the same way as for changepoint moves: $b_{p_i} + d_{p_i} + u_{p_i} = 1$; $d_0 \triangleq 0$, $b_{p_{\max}} \triangleq 0$; otherwise, $b_{p_i} = d_{p_i} = u_{p_i}$ for $i = 0, \ldots, k$. The procedure is performed for each segment and the main steps are described as follows.

Algorithm for the Update of the Model Orders and Hyperparameters

- For $i = 1, \ldots, k$,
  - if $(u_p \sim \mathcal{U}_{(0,1)}) \leq b_{p_i}$ then propose $p_i' = p_i + 1$;
    else if $u_p \leq b_{p_i} + d_{p_i}$ then propose $p_i' = p_i - 1$;
  - if $(u_{p_d} \sim \mathcal{U}_{(0,1)}) \leq \alpha_{(p_i \rightarrow p_i')}$ [see (26)], the new state of the MC is accepted.
  - sample $\delta_i^2 | (k, \boldsymbol{\tau}_k, \mathbf{p}_k, \mathbf{y}_{0:T-1}, \boldsymbol{\beta}_i, \sigma_i^2)$; see (29),
    $\sigma_i^2 | (k, \boldsymbol{\tau}_k, \mathbf{p}_k, \boldsymbol{\delta}_k^2, \mathbf{y}_{0:T-1})$,
    $\boldsymbol{\beta}_i | (k, \boldsymbol{\tau}_k, \mathbf{p}_k, \boldsymbol{\delta}_k^2, \mathbf{y}_{0:T-1}, \sigma_i^2)$ are sampled from (14) and (15).
- Propose $\psi' | (k, \mathbf{p}_k)$ [see (27)]; if $(u_\psi \sim \mathcal{U}_{(0,1)}) \leq \alpha_\psi$ [see (28)] then $\psi = \psi'$.
- Sample hyperparameters $\gamma_\sigma, \theta_\delta$ see (13) and (30).

■

The acceptance probability for the different types of moves (in terms of the model orders) is given by

$$\alpha_{(p_i \rightarrow p_i')} = \min\left\{1, r_{(p_i \rightarrow p_i')}\right\} \quad (26)$$

where from (19)

$$r_{(p_i \rightarrow p_i')} = \frac{\left|\mathbf{M}_i^{(p_i')}\right|^{1/2} \psi^{p_i'} p_i! \delta_i^{-((p_i'-p_i)/2)}}{\left|\mathbf{M}_i^{(p_i)}\right|^{1/2} \psi^{p_i} p_i'!} \times \left(\frac{\gamma_\sigma + \mathbf{y}_{\tau_i:\tau_{i+1}-1}^{\mathrm{T}} \mathbf{P}_i^{(p_i')} \mathbf{y}_{\tau_i:\tau_{i+1}-1}}{\gamma_\sigma + \mathbf{y}_{\tau_i:\tau_{i+1}-1}^{\mathrm{T}} \mathbf{P}_i^{(p_i)} \mathbf{y}_{\tau_i:\tau_{i+1}-1}}\right)^{-((\nu_\sigma + \tau_{i+1} - \tau_i)/2)}.$$

For the birth move ($p_i \rightarrow p_i + 1$), the acceptance ratio is $\alpha_{birth}^p = \min\{1, r_{birth}\}$, where $r_{birth} = r_{(p_i \rightarrow p_i+1)}$. Assuming that the current model order is $(p_i + 1)$, one obtains the acceptance ratio for the death move ($p_i + 1 \rightarrow p_i$) as $\alpha_{death}^p = \min\{1, r_{birth}^{-1}\}$. Thus, the birth/death moves are, indeed, reversible.

Taking into account the series representation of the exponential function, we adopt the following proposal distribution for the parameter ψ:

$$\mathcal{G}a\left(\psi; \zeta_\psi + \sum_{i=0}^{k} p_i, \chi_\psi + (k+1)\right) \quad (27)$$

and sample ψ according to a Metropolis–Hastings step with the acceptance probability equal to

$$\alpha_\psi = \min\left\{1, \left[\frac{\sum_{p=0}^{p_{\max}} \psi^p}{\sum_{p=0}^{p_{\max}} (\psi')^p} \frac{\exp(-\psi)}{\exp(-\psi')}\right]^{(k+1)}\right\}. \quad (28)$$

The hyperparameters δ_i^2 are sampled using a standard Gibbs move by analogy with (12).

$$\delta_i^2 \sim \mathcal{IG}\left(\alpha_\delta + \frac{p_i}{2}, \theta_\delta + \frac{\boldsymbol{\beta}_i^{(p_i)\mathrm{T}} \boldsymbol{\beta}_i^{(p_i)}}{2\sigma_i^2}\right). \quad (29)$$

Similarly, we sample γ_σ as in (13) and θ_δ according to

$$\theta_\delta \sim \mathcal{G}a\left(\alpha_\delta(k+1), \sum_{i=1}^{k} \frac{1}{\delta_i^2}\right). \quad (30)$$

IV. Simulations

In the first set of simulations, we address the standard problem of denoising smooth and unsmooth test functions [1], [17]. To compare our results with [1], we have used a fixed model order p_0. Subsequently, we apply our algorithm with unknown model orders to the segmentation of signals modeled as piecewise constant AR processes and a speech signal [2], [4], [5].

A. Curve Fitting: Fixed Model Order p_0

1) Smooth Function: First, we assessed the performance of the algorithm proposed in Section II by applying it to synthetic piecewise polynomials with $k = 4$, $p_0 = 3$, and $T = 500$ (model parameters and noise variances are presented in Table I).

The estimates of the number of changepoints and their positions were obtained using the MAP criterion (see Section II-A2) after 50 000 iterations of the algorithm and a burn-in period of 10 000 (further iterations yielded no appreciable difference). The estimated number of changepoints was equal to $\hat{k} = 4$, and Table II gives the estimated changepoint positions. Fig. 5 show the original noisy and estimated curves. In Fig. 6, the estimation of the marginal posterior distributions of the number of changepoints $\hat{p}(k|\mathbf{y}_{0:T-1})$ is presented. The MMSE estimate of the regression function obtained by making use of Bayesian model averaging was 0.026.

TABLE I
PARAMETERS OF THE POLYNOMIAL MODEL AND NOISE VARIANCE FOR EACH SEGMENT

ith segment	σ_i^2	$\beta_i^{(p_i)}$		
0	0.09	0	10	−40
1	0.09	1.985	−10	10
2	0.09	0.83	−5	5
3	0.09	−1.27	2	−0.8
4	0.09	−2.104	4	−2

TABLE II
REAL AND ESTIMATED VALUES FOR CHANGEPOINT POSITIONS

ith segment	0	1	2	3	4
τ_i (true value)	-	90	180	320	410
$\hat{\tau}_i$	-	91	179	320	412

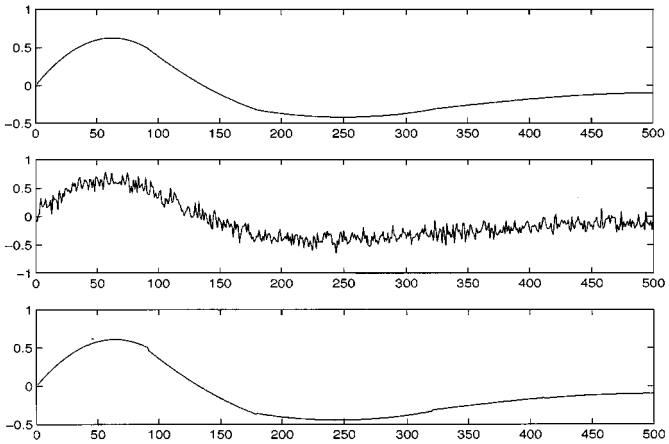

Fig. 5. Piecewise polynomials. (Top) Original curve. (Middle) Curve with noise added. (Bottom) Estimate.

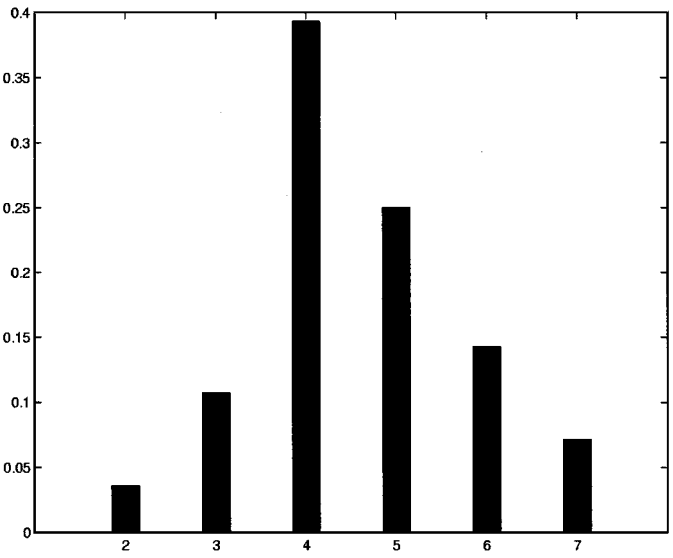

Fig. 6. Estimation of the marginal posterior distribution of the number of changepoints for the piecewise polynomial model.

The algorithm was coded using Matlab, and the simulations were performed on a 500 MHz Intel Pentium III PC. Processing of 1000 iterations required on average 95 s.

TABLE III
ESTIMATED NUMBER OF CHANGEPOINTS AND AVERAGE MMSE

Function	$\hat{k}_{[1]}$	$\hat{k}$	$\text{MMSE}_{[1]}$	MMSE
Blocks	35	11	0.170	0.021
Heavisine	17	7	0.033	0.019
Bumps	62	40	0.167	0.121
Doppler	37	18	0.135	0.098

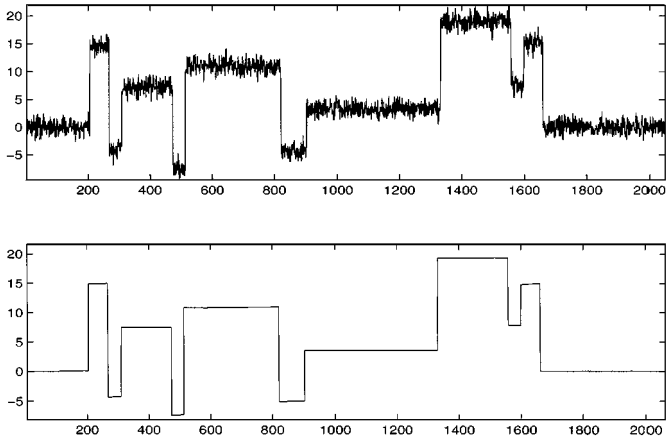

Fig. 7. Blocks test curve. (Top) True function with noise added. (Bottom) Estimate of the function.

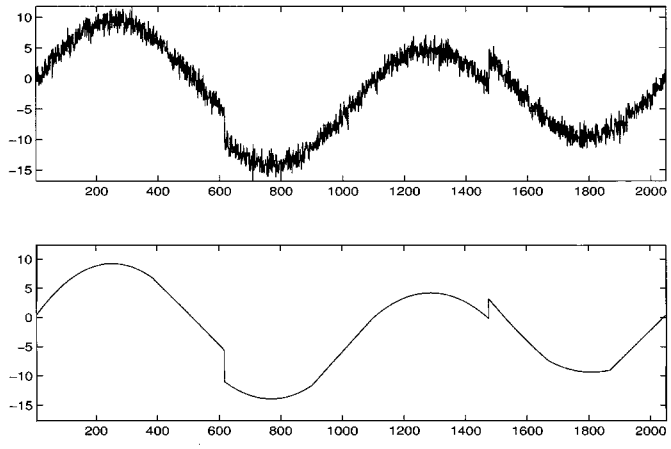

Fig. 8. Heavisine test curve. (Top) True function with noise added. (Bottom) Estimate of the function.

2) Unsmooth Functions: In the second example, we applied our algorithm to some common curves (such as "Blocks," "Heavisine") previously used in the literature as a test [1], [17]. Following [1], the number of grid points was taken to be 2048, the standard noise deviation was set equal to $\sigma = 1$ for all segments, and the model order for each polynomial was set to be $p_0 = 3$. The results for the number of changepoints and average MMSE compared with those obtained by [1] are presented in Table III. Figs. 7 and 8 show the original functions with the noise added and the estimates obtained. It is worth noticing that although the polynomial order equal to 3 was adopted throughout, the reconstructions of both "Blocks" and "Heavisine" curve are almost perfect. The estimated number of changepoints and the average MMSE for "Bumps" and

TABLE IV
PARAMETERS OF THE AR MODEL AND NOISE VARIANCE FOR EACH SEGMENT

ith segment	σ_i^2	$\boldsymbol{\beta}_i^{(p_i)}$			
0	1.6	−2.3000	−2.6675	−1.8437	−0.5936
1	0.8	1.3000	−0.9200	0.2600	
2	1.7	0.8000	−0.5200		
3	0.5	2.0000	−1.6350	0.5075	
4	0.6	−1.7000	−0.7450		
5	1.8	−0.5000	0.6100	0.5850	

TABLE V
REAL AND ESTIMATED VALUES FOR CHANGEPOINT AND MODEL ORDER

ith segment	0	1	2	3	4	5
τ_i (true value)	-	90	160	250	365	430
$\hat{\tau}_i$	-	91	162	249	366	434
p_i (true value)	4	3	2	3	2	3
$\hat{p}_i = \max \hat{p}\left(p_i \mid \hat{k}, \mathbf{y}_{0:T-1}\right)$	4	3	2	3	2	3

"Doppler" obtained in additional experiments are also shown in Table III. (The results for the wavelet methods in [17] are given in [1]; the method in [1] performs significantly better).

Our hierarchical model allows the "automatic" determination of hyperparameter values, contrary to the one in [1]. This has the effect of providing both a more sparse representation of the regression function and a reduced MMSE as we prevent overfitting.

B. Signal Segmentation: Unknown Model Orders p_i

1) Piecewise Constant Autoregressive Processes: We now illustrate the performance of the segmentation method proposed above by applying it to synthetic data ($T = 500$) , which can be described as a piecewise constant autoregressive (AR) process with $k = 5$ changepoints. The parameters of the AR models $\{\boldsymbol{\beta}_i^{(p_i)}\}_{i=0}^{5}$ and noise variances $\boldsymbol{\sigma}_5^2$, drawn at random, are given in Table IV. We interpret the first $p_{\max}$ samples as the initial conditions and proceed with analysis on the remaining $T - p_{\max}$ data points.

The number of iterations of the algorithm was 100 000 (the results for a higher number of iterations are indistinguishable), and as was described in Section II-A2, we adopt the MAP as a detection criterion, from which one, indeed, finds $\hat{k} = 5$ changepoints. Then, for fixed $k = \hat{k}$, the model order for each segment p_i is estimated by MAP $\hat{p}_i = \max \hat{p}(p_i|\hat{k}, \mathbf{y}_{0:T-1})$, $i = 0, \ldots, \hat{k}$. The results are presented in Table V. Fig. 9 shows the segmented signal and the estimation of the marginal posterior distributions of the changepoint positions $\hat{p}(\tau_i|\mathbf{y}_{0:T-1})$.

Then we estimated the mean and the associated standard deviation of the marginal posterior distributions $p(k\mathbf{y}_{0:T-1})$ for 50 realizations of the experiment with fixed model parameters and changepoint positions. The results are presented in Fig. 10, and it is worth noticing that they are very stable with respect to the fluctuations in the excitation noise realization.

2) Speech Segmentation: We also implemented the proposed algorithm for processing a real speech signal which was previously examined in the literature [2], [4], [5]. It was recorded inside a car by the French National Agency for Telecommunications for testing and evaluating speech recognition algorithms, as described in [5]. According to [2], the sampling frequency was 12 kHz, and a highpass filtered version of the signal with cut-off frequency 150 Hz and resolution of 16 bits is presented in Fig. 11.

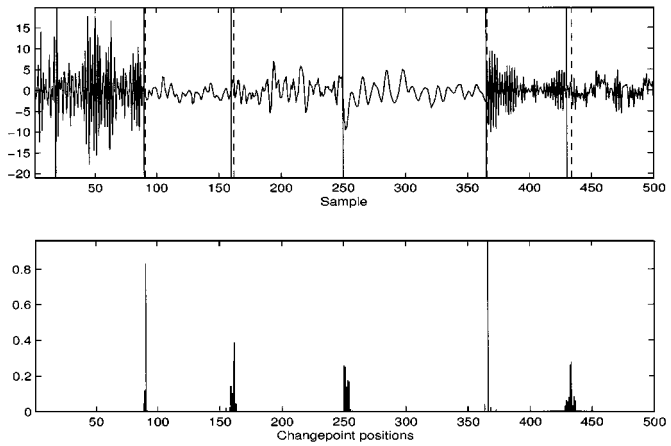

Fig. 9. (Top) Segmented signal (the original changepoints are shown as a solid line, and the estimated changepoints are shown as a dotted line). (Bottom) Estimation of the marginal posterior distribution of the changepoint positions.

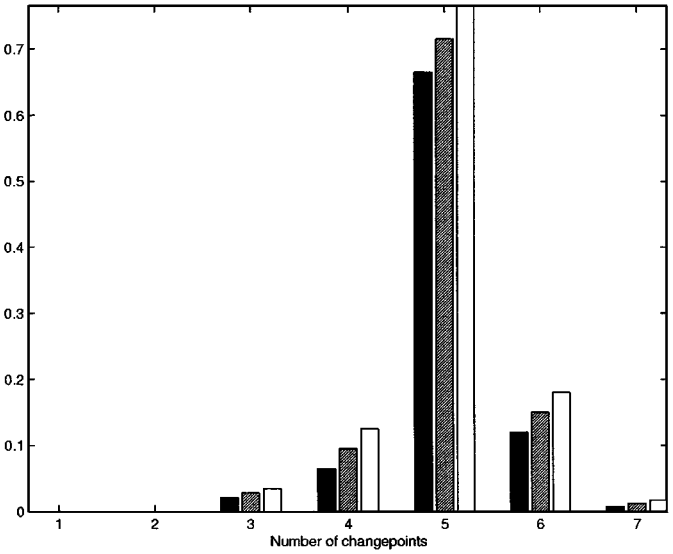

Fig. 10. Mean and standard deviation for 50 realizations of the posterior distribution.

Different segmentation methods [2], [4], [5] were applied to the signal, and a summary of the results can be found in [2]. We show these results in Table VI in order to compare them to the ones obtained using our proposed method (see also Figs. 11 and 12). The estimated orders of the AR models are presented in Table VII, and as one can see, they are quite different from segment to segment. This resulted in different positions for the changepoints, which is especially crucial in the case of the third and fourth changepoint. Its position changed significantly due to the estimated model orders for the second ($\hat{p}_2 = 19$) and third segments ($\hat{p}_3 = 17$). As it is illustrated in Fig. 12, the changepoints obtained by the proposed method visually seem preferable.

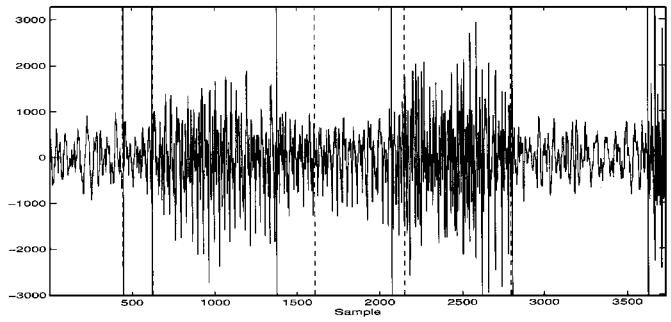

Fig. 11. Segmented speech signal (the changepoints estimated by Gustafsson are shown as a dotted line and ones estimated using the proposed method are shown as a solid line).

TABLE VI
CHANGEPOINT POSITIONS FOR DIFFERENT METHODS

Method	AR order	Estimated changepoints							
Divergence ([18])	16	445	645	1550	1800	2151	2797	-	3626
Brand's GLR([19])	16	445	645	1550	1800	2151	2797	-	3626
Brand's GLR([19])	2	445	645	1550	1750	2151	2797	3400	3626
Approx. ML ([2])	2	445	626	1609	-	2151	2797	-	3627
Proposed method	estimated	448	624	1377	-	2075	2807	-	3626

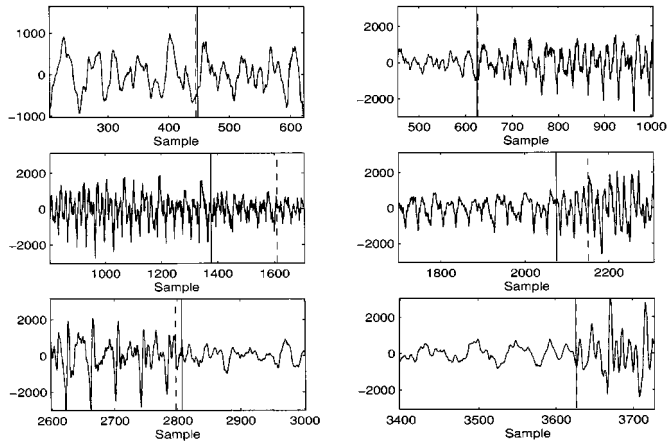

Fig. 12. Changepoint positions (the changepoints estimated by Gustafsson are shown as a dotted line and the ones estimated using the proposed method are shown as a solid line).

TABLE VII
ESTIMATED MODEL ORDERS

Segment	0	1	2	3	4	5	6
Model order	6	5	19	27	16	9	11

V. CONCLUSION

In this paper, we have presented some Bayesian methods for a variety of challenging functions modeled as a piecewise constant linear regression. Our model and algorithms appear quite flexible and have applications in denoising and in signal segmentation. They can also be used for prediction, as one can straightforwardly evaluate the predictive distribution using the MCMC samples.

There are many possible extensions to this work. Among others, one could extend the algorithm to multivariate signals, the statistical assumptions on the noise distribution could be relaxed, or an on-line version of the algorithm could be developed based on particle filtering methods.

REFERENCES

[1] D. Denison, B. Mallick, and A. F. M. Smith, "Automatic Bayesian curve fitting," *J. R. Stat. Soc. B*, vol. 60, pp. 333–350, 1998.
[2] F. Gustafsson, *Adaptive Filtering and Change Detection*. New York: Wiley, 2000.
[3] J. J. K. O'Ruanaidh and W. J. Fitzgerald, *Numerical Bayesian Methods Applied to Signal Processing*. New York: Springer-Verlag, 1996.
[4] R. André-Obrecht, "A new statistical approach for automatic segmentation of continuous speech signals," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 29–40, Jan. 1988.
[5] M. Basseville and I. V. Nikiforov, *Detection of Abrupt Changes: Theory and Application*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[6] P. Djurić, "A MAP solution to off-line segmentation of signals," in *Proc. Conf. IEEE ICASSP*, 1994, pp. 505–508.
[7] D. Barry and J. Hartigan, "A Bayesian analysis for changepoint problems," *J. Amer. Stat. Assoc.*, vol. 88, pp. 309–319, 1993.
[8] D. A. Stephens, "Bayesian retrospective multiple-changepoint identification," *Appl. Stat.*, vol. 43, pp. 159–178, 1994.
[9] B. K. Mallick, "Bayesian curve estimation by polynomials of random order," *J. Statist. Plan. Inform.*, vol. 70, pp. 91–109, 1997.
[10] S. Richardson and P. J. Green, "On Bayesian analysis of mixtures with unknown number of components," *J. Roy. Stat. Soc. B*, vol. 59, no. 4, pp. 731–792, 1997.
[11] P. J. Green, "Reversible jump MCMC computation and Bayesian model determination," *Biometrika*, vol. 82, pp. 711–732, 1995.
[12] C. P. Robert, *The Bayesian Choice. A Decision-Theoretic Motivation*, 2nd ed. New York: Springer-Verlag, 2001.
[13] C. Andrieu, P. M. Djurić, and A. Doucet, "MCMC computation for Bayesian model selection," *Signal Process.*, vol. 81, no. 1, pp. 19–37, 2001.
[14] S. J. Godsill, "On the relationship between MCMC model uncertainty methods," *J. Comput. Graph. Stat.*, to be published.
[15] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*. New York: Springer-Verlag, 1999.
[16] E. Punskaya, C. Andrieu, A. Doucet, and W. J. Fitzgerald, "Bayesian segmentation of piecewise constant autoregressive processes using MCMC," Cambridge Univ. Eng. Dept., Tech. Rep. CUED/F-INFENG/TR.344, 1999.
[17] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, pp. 425–455, 1994.
[18] M. Basseville and A. Benveniste, "Design and comparative study of some sequential jump detection algorithms for digital signals," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-31, pp. 521–535, 1983.
[19] U. Appel and A. V. Brandt, "Adaptive sequential segmentation of piecewise stationary time series," *Inform. Sci.*, vol. 29, no. 1, pp. 27–56, 1983.

Elena Punskaya received the M.Sc. degree in mechanical engineering from the Bauman Moscow State Technical University, Moscow, Russia, in 1997 and the M.Phil. degree in information engineering from the University of Cambridge, Cambridge, U.K., in 1999. Currently, she is pursuing the Ph.D. degree with the Signal Processing Laboratory, Department of Engineering, University of Cambridge.

Her research interests focus on Bayesian inference, Markov chain Monte Carlo methods, and sequential Monte Carlo methods and their application to wireless communication problems.

Christophe Andrieu was born in France in 1968. He received the M.S. degree from the Institut National des Télécommunications, Paris, France, in 1993 and the D.E.A. degree in 1994 and the Ph.D. degree in 1998 from the University of Paris XV.

From 1998 to September 2000, he was a research associate with the Signal Processing Group, Cambridge University, Cambridge, U.K. He is now a lecturer with the Department of Mathematics, Bristol University, Bristol, U.K. His research interests include spectral analysis, source separation, Markov chain Monte Carlo methods, and stochastic optimization.

Arnaud Doucet was born in Melle, France, on November 2, 1970. He received the M.S. degree from the Institut National des Telecommunications, Paris, France, in 1993 and the Ph.D. degree from University Paris XI, Orsay, France, in 1997.

During 1998, he was a Visiting Scholar with the Signal Processing Group, Cambridge University, Cambridge, U.K. From 1999 to 2001, he was a research associate in the same group. Since March 2001, he has been a Senior Lecturer with the Department of Electrical Engineering, Melbourne University, Parkville, Victoria, Australia. His research interests include Markov chain Monte Carlo methods, sequential Monte Carlo methods, Bayesian statistics, and hidden Markov models. He has co-edited, with J. F. G. de Freitas and N. J. Gordon, *Sequential Monte Carlo Methods in Practice* (New York: Springer-Verlag, 2001).

William J. Fitzgerald received the B.Sc., M.Sc., and Ph.D. degrees in physics from the University of Birmingham, Birmingham, U.K.

He is the Head of the Statistical Modeling Group within the Signal Processing Laboratory, Department of Engineering, University of Cambridge, Cambridge, U.K., were he is also University Reader in statistics and signal processing. Prior to joining Cambridge, he worked at the Institut Laue Langevin, Grenoble, France for seven years and was then a Professor of physics at the Swiss Federal Institute of Technology (ETH), Zürich, for five years. He is a Fellow of Christ's College, Cambridge, where he is Director of Studies in Engineering. He works on Bayesian inference applied to signal and data modeling. He is also interested in nonlinear signal processing, image restoration, and extreme value statistics.