Bayesian Computational Methods for Inference in Multiple Change-points Models

Nick Whiteley*, Christophe Andrieu* and Arnaud Doucet†

Abstract

Multiple change-point models provide a flexible and interpretable framework for representation of temporal heterogeneity in data. In addition to the locations of change-points, these models typically involve parameters which specify the distributions of data between change-points and other quantities. However, the values of these parameters are usually unknown and need to be inferred from the data. We develop new Markov chain Monte Carlo algorithms which provide an efficient means for full Bayesian inference in the presence of parameter uncertainty. Performance is demonstrated on various examples.

1 Introduction

Time series data often exhibit temporal heterogeneity. In multiple change-point problems, the task is to segment a sequence of observations $y_1, y_2, ..., y_T$ by choosing a sequence of change-point locations $0 < \tau_1 < \tau_2 < ... < \tau_k < T$ such that the observations are, in some sense, homogeneous within segments and heterogeneous across segments. Statistical analysis of change–point problems has a long history; as of 1992 the literature on this topic was "enormous" [Carlin et al., 1992]. In terms of application, multiple change-point problems are common in biology [Fearnhead and Liu, 2007], finance [Chopin, 2007], signal processing [Fearnhead and Clifford, 2005], and other areas, arising quite naturally when there is increasing availability of long data sequences. A diverse of set of non-, semi- and fully parametric methods for change-point models has been developed. A full survey is beyond the scope of the present article, the focus here is on Bayesian methods. In this approach, both the prior over the number and location of change-points [Barry and Hartigan, 1992, 1993], and the likelihood function depend on an unknown and typically multi-dimensional parameter θ. The values taken by this parameter can dramatically influence the properties of the model.

However, performing Bayesian inference for multiple change-points and the parameter θ is a challenging problem. Even when θ is assumed known, exact computation of the posterior distribution over change-point configurations is intractable for large data sets and Markov

*Department of Mathematics, University of Bristol, University Walk, Bristol, BS8 1TW, UK.
†Department of Statistics, University of Oxford, UK.

chain Monte Carlo (MCMC) techniques are typically employed. Unfortunately, MCMC algorithms which update change-point locations in a one-at-a-time manner [Stephens, 1994, Lavielle and Lebarbier, 2001], or condition on latent variables associated with each segment [Chib, 1998], can be slow mixing due to the strong correlations between the change-point locations and latent variables. Alternatives, which sample all the change-points in one block have a cost per MCMC iteration which is of the order T^2 [Fearnhead, 2006]. For real-world data sets of several thousand observations, this T^2 method can be prohibitively expensive.

Particle MCMC (PMCMC) algorithms, recently introduced by Andrieu et al. [2010], are a class of MCMC algorithms which allow sequential Monte Carlo (SMC) [Doucet et al., 2001], to be used in building high-dimensional proposals within a MCMC scheme. In the context of change-point models, an efficient SMC method has been proposed by Fearnhead and Liu [2007]. However, its structure differs very significantly from the SMC algorithms treated in Andrieu et al. [2010]. The main contribution of this paper is to show how this SMC algorithm can be used and manipulated within MCMC schemes to obtain efficient samplers for multiple-change point models. We derive two original PMCMC algorithms, whose cost per MCMC iteration is of the order $N\,T$, where N is the number of particles in the SMC approximation, and whose experimental performance compares remarkably well to T^2 algorithms for $N \ll T$. This is a further development of a conceptual approach introduced by Whiteley et al. [2010] in the context of PMCMC algorithms for switching state-space models. However the multiple change-point models considered in the present paper have a more specific conditional independence structure, which is reflected and exploited through the specific form of the proposed computational algorithms. Hence the algorithms presented here are markedly different from those proposed in Whiteley et al. [2010].

Section 2 specifies the change-point model of interest. Section 3 describes existing MCMC methods, and the sampling techniques of Fearnhead and Liu [2007]. The new PMCMC algorithms are introduced in section 4 and their theoretical validity is established in section 5. Performance is demonstrated on two real data sets in section 6.

2 Multiple Change-point Model

2.1 Bayesian Model

The change-point model we consider is essentially the same as that treated by Fearnhead and Liu [2007] except that the parameter θ, valued in some space Θ, is here assumed unknown

and is assigned a suitable prior $p(\theta)$. In the following model specification, it is important to note that the parameter θ is global, in the sense that it is common to all segments defined by any change-point configuration (specific examples are given in section 6). From now on, by convention we fix $\tau_0 = 0$, $\tau_{k+1} = T$, where k is the effective number of change-points, and for some generic sequence $\{z_n\}$ we adopt the notation $z_{i:j} := (z_i, z_{i+1}, ..., z_j)$.

A collection of change-points $\tau_{1:k}$ is an increasing sequence of integers $0 < \tau_1 < \tau_2 < ... < \tau_k < T$ for some integer $k \geq 0$ which defines $k+1$ segments where for $j = 1, 2, ..., k+1$, the j^{th} segment is $\tau_{j-1} + 1 : \tau_j$. The collection $\tau_{1:k}$ therefore allows us to partition the data $y_{1:T}$ into segments $y_{\tau_{j-1}+1:\tau_j}$. Similarly to Barry and Hartigan [1992], Fearnhead and Liu [2007], we assume that given the location of a change-point and θ, the observations before that change-point are independent of those after; that is the likelihood factorizes

$$p_\theta(y_{1:T}|\tau_{1:k}) = \prod_{j=1}^{k+1} p_\theta(y_{\tau_{j-1}+1:\tau_j}|\tau_{j-1:j}).$$

As in [Fearnhead and Liu, 2007], this likelihood may arise from summing over a number of possible models for the data in each segment and/or integrating out local latent variables associated with each segment under conjugate priors. It is assumed here that we can evaluate this likelihood exactly. Examples with this structure are given in section 6.

The sequence $\tau_0, \tau_1, \tau_2, ...$ is increasing and is assumed Markov with homogeneous transition probabilities given for $j = 1, 2, ...$ by $p_\theta(\tau_j|\tau_{j-1}) = h_\theta(\tau_j - \tau_{j-1})$. We are concerned with the space of all change–point configurations over $\{1, ..., T-1\}$ which we denote by the disjoint union $\mathcal{T}_T := \bigcup_{k=0}^{T-1} \mathcal{T}_{T,k}$, where $\mathcal{T}_{T,k} := \{\tau_{1:k} \in \{1, ..., T-1\}^k; \tau_1 < ... < \tau_k\}$ for $k \geq 1$ and $\mathcal{T}_{T,0}$ is the empty configuration. The joint prior probability of exactly k change-points and their locations is then given by

$$p_\theta(\tau_{1:k}) = [1 - H_\theta(T - \tau_k)] \prod_{j=2}^{k} h_\theta(\tau_j - \tau_{j-1}),$$

where H_θ is the c.d.f. associated with h_θ, i.e.

$$H_\theta(n) = \sum_{i=1}^{n} h_\theta(i).$$

For simplicity of presentation, we do not make explicit in the notation $p_\theta(\tau_{1:k})$ the fact that

k is a random variable - this does not lead to ambiguities in what follows. We are interested in the joint posterior distribution

$$p(\theta, \tau_{1:k}|y_{1:T}) \propto p_\theta(y_{1:T}|\tau_{1:k})p_\theta(\tau_{1:k})p(\theta). \tag{1}$$

2.2 A Canonical Reparameterization

As advocated by Chib [1998], Chopin [2007], Fearnhead and Liu [2007], for computational purposes it is fruitful to consider the following reparameterization of the change-point model. For each $n = 1, 2, ..., T$, let X_n be the random variable valued in $E_n := \{0, ..., n-1\}$ which is the location of the latest change-point before time n. More precisely, given $\tau_{1:k}$, for any $n = 1, \ldots, T$ we define $x_n := \max\{\tau_j : \tau_j \leq n-1\}$. Immediate properties of this reparametrization which will be repeatedly used in what follows are that for any $j = 0, \ldots, k$, $x_{\tau_j+1} = \tau_j$, and $x_{\tau_j+1} = x_{\tau_j+2} = \cdots = x_{\tau_{j+1}}$. It is straightforward to verify that this definition implies a one-to-one correspondence between paths $x_{1:T} \in \prod_{n=1}^T E_n$ and change-point configurations $\tau_{1:k} \in \mathcal{T}_T$. Furthermore, the above prior on $\tau_{1:k}$ is equivalent to the sequence $X_{1:T}$ being Markov with, for $0 \leq x_{n-1} \leq x_n < n$, transition probabilities $f_n^\theta(x_n|x_{n-1})$ given by

$$f_n^\theta(x_n|x_{n-1}) = \begin{cases} \dfrac{1 - H_\theta(n - x_{n-1} - 1)}{1 - H_\theta(n - x_{n-1} - 2)} & \text{if } x_n = x_{n-1}, \\ \dfrac{H_\theta(n - x_{n-1} - 1) - H_\theta(n - x_{n-1} - 2)}{1 - H_\theta(n - x_{n-1} - 2)} & \text{if } x_n = n - 1, \\ 0 & \text{otherwise,} \end{cases}$$

and initial distribution specified by the convention $f_1^\theta(0|0) = 1$. For $n > x_n$ the predictive likelihood is denoted

$$g_n^\theta(x_n) := p_\theta(y_n|y_{1:n-1}, x_n),$$

with the dependence on the data suppressed for notational convenience. The interest in this reparameterization is that computing the posterior in (1) is equivalent to computing the joint posterior over $x_{1:T}$ and θ:

$$p(x_{1:T}, \theta|y_{1:T}) \propto f_1^\theta(x_1)g_1^\theta(x_1) \prod_{n=2}^T f_n^\theta(x_n|x_{n-1})g_n^\theta(x_n)p(\theta), \tag{2}$$

and the manner in which (2) factorizes makes it particularly amenable to approximation using SMC methods. More specifically, in section 4 we will see that the tasks of approximating

$p_\theta(y_{1:T})$ and sampling from $p_\theta(x_{1:T}|y_{1:T})$ using SMC are central to the construction of new MCMC algorithms targeting $p(\theta, \tau_{1:k}|y_{1:T})$.

In order to simplify presentation, throughout this paper we assume that for all θ, n, x_n and x_{n-1} such that $x_n = x_{n-1}$ or $x_n = n-1$, $f_n^\theta(x_n|x_{n-1}) > 0$ and $g_n^\theta(x_n) > 0$. This assumption is satisfied for the majority of change–point models of practical interest and relaxing it requires only cosmetic changes to the proposed algorithms.

We next briefly review some existing approaches to posterior sampling in multiple-change point models using standard MCMC methods. This is not just for completeness. As we shall see, the proposed methods admit some of these existing techniques as special cases.

3 Computational Methods

3.1 Review of MCMC Methods

The design of MCMC algorithms for posterior sampling in change–point models has been a topic of interest for some years. Carlin et al. [1992] devised a Gibbs sampler for a Bayesian change–point model in which the number of change–points was fixed to 1. This Gibbs sampler was extended to models with multiple change–points by Stephens [1994]. The latter approach involves one-at-a-time sampling of the change-point locations and associated latent variables from their respective full conditional distributions. It is well known that one-at-a-time Gibbs samplers can suffer from slow mixing which arises from strong correlations in the target distribution. Empirical evidence shows that similar algorithms which employ Metropolis–Hastings steps to sample one-at-a-time in change–point models [Lavielle and Lebarbier, 2001] suffer from similar problems [Fearnhead, 2006].

For multiple change–point models, Chib [1998] presented a Gibbs sampling algorithm in which the entire change–point configuration $\tau_{1:k}$ is sampled from its full conditional distribution, given θ and the latent variables associated with each segment. This is an instance of blocking in a Gibbs sampler: a technique which often improves mixing significantly Liu et al. [1994]. One step further is to sample $\tau_{1:k}$ from $p_\theta(\tau_{1:k}|y_{1:T})$, i.e. with the latent variables associated with each segment integrated out. This can improve mixing even further and is achievable using the exact forward-backward recursions of Fearnhead and Clifford [2005], Fearnhead [2006], Fearnhead and Liu [2007]. However, these schemes have a computational cost per MCMC iteration which grows quadratically in T, the length of the data record. In modern applications, T may be of the order of thousands and so the cost of these methods

can be prohibitive. In the next section we review the exact forward-backward methods for sampling from $p_\theta(\tau_{1:k}|y_{1:T})$, computing $p_\theta(y_{1:T})$, and the corresponding SMC methods.

3.2 Exact Filtering and Backward Sampling

Fearnhead and Clifford [2005], Fearnhead [2006], Fearnhead and Liu [2007], proposed methods for obtaining exact samples from $p_\theta(\tau_{1:k}|y_{1:T})$ and computing $p_\theta(y_{1:T})$. We focus on the algorithm of Fearnhead and Liu [2007], which involves two steps. In the *forward filtering* pass, for $n = 1, ..., T$, each filtering distribution $p_\theta(x_n|y_{1:n})$ is computed recursively and stored. The likelihood $p_\theta(y_{1:T})$ can be obtained from quantities computed in this filtering pass. In the *backward sampling* pass, a sequence of change-point locations $\tau_{1:k}$ is obtained by sampling back through the stored filtering distributions, re-weighted appropriately. We refer to Fearnhead and Liu [2007] for specific details.

In the change-point model of section 2, recall that the support of $p_\theta(x_n|y_{1:n})$ is $E_n = \{0, ..., n-1\}$ which obeys the trivial recursion

$$E_1 = \{0\}, \quad E_{n+1} = E_n \cup \{n\}, \quad n = 1, 2, ... \tag{3}$$

Hence the support of the filtering distributions is increasing with time. It follows that the cost of computing $p_\theta(y_{1:T})$ and the storage requirements of sampling from $p_\theta(\tau_{1:k}|y_{1:T})$ grow quadratically in T. In many practical problems, T may be of the order of thousands and so these exact methods can be prohibitively expensive. Fearnhead and Clifford [2005], Fearnhead [2006] suggested to reduce computational complexity by performing a deterministic truncation of certain quantities which arise in the filtering recursion, but this introduces a bias which is difficult to quantify.

3.3 SMC

Fearnhead and Liu [2007] also proposed an algorithm which employs SMC ideas in order to approximate the filtering distributions. This allows the cost per filtering iteration to be upper–bounded uniformly in time and is of order N, where N is an algorithmic parameter. The storage requirements of obtaining a sample from the corresponding approximation of $p_\theta(\tau_{1:k}|y_{1:T})$ and computing the corresponding approximation of $p_\theta(y_{1:T})$ then grow only linearly with T.

SMC methods are a class of stochastic algorithms which allow approximation of a sequence of probability distributions [Doucet et al., 2001]. Standard SMC methods yield a collection of N weighted samples, termed *particles* $\{X^i, W(X^i)\}_{i=1}^N$ which define a random probability distribution,

$$\sum_{i=1}^{N} W(X^i)\delta_{X^i}, \quad \sum_{i=1}^{N} W(X^i) = 1, \tag{4}$$

where in full generality, the weight $W(\cdot)$ is some possibly random function of the particle location X^i. The structure of the change-point model is significantly different from that of the state-space models to which standard SMC algorithms are usually applied. For each n, $|E_n|$ is finite; in standard state-space models, the filtering distribution is usually defined on some general state space. The SMC algorithm of Fearnhead and Liu [2007] exploits the fact that $|E_n|$ is finite and avoids the random proposal/importance sampling step present in standard SMC. Furthermore, it has the desirable property of avoiding particle duplication. This is a further development of the ingenious optimal resampling particle filter introduced in Fearnhead and Clifford [2003]. We refer to [Fearnhead and Liu, 2007] for a detailed discussion of efficiency and applicability of this method relative to alternative SMC schemes for sampling change-points; see for example [Chopin, 2007]. For brevity, we treat the variant of [Fearnhead and Liu, 2007] in which resampling is applied at every iteration, but other variants can be considered in the same framework.

The remainder of this section is dedicated to a sketch of the SMC filtering algorithm's probabilistic operation, followed by a statement of the algorithm itself. In particular we introduce a non–standard approach to describing the SMC algorithm of [Fearnhead and Liu, 2007] and the random probability distributions it generates. Our system of notation is central to the understanding of the new MCMC algorithms described in section 4. In this section, the notational ideas are introduced somewhat incrementally, a more precise probabilistic formulation is given in section 5.1. The first step in this non-standard approach is to move away from the system of indexing displayed in (4) via the trivial equality

$$\sum_{i=1}^{N} W(X^i)\delta_{X^i} = \sum_{X \in \mathbb{S}} W(X)\delta_X, \tag{5}$$

where $\mathbb{S} := \{X^1, X^2, ..., X^N\}$ is a random support set. In standard SMC algorithms, $\mathbb{S}$ is propagated from one iteration to the next by resampling followed by random proposal of new support points. In the resampling step, particles with large weights are duplicated and

those with small weights are discarded, according to some stochastic rule. The stochasticity of the proposal and resampling steps inevitably contributes to the fluctuations of (5) and the variance of associated estimators.

In the case of the change–point model outlined above, the state space $|E_n|$ is finite and so the random proposal step can be avoided, and replaced with a systematic exploration. Empirical studies, see for example [Fearnhead and Clifford, 2003], show that this leads to a dramatic decrease in Monte Carlo error. In expressing the algorithm of Fearnhead and Liu [2007], for each n, we consider $\hat{p}_\theta(x_n|y_{1:n})$, an approximation of the filtering distribution, as having random support $\mathbb{S}_n \subset E_n$, corresponding to the non-zero members of a collection of random weights $\mathbf{W}_n^\theta := \{W_n^\theta(x_n)\}_{x_n \in E_n}$, which will be precisely defined below. We write the approximation $\hat{p}_\theta(x_n|y_{1:n})$ in the form

$$\sum_{x_n \in \mathbb{S}_n} W_n^\theta(x_n)\delta_{x_n}, \quad \sum_{x_n \in \mathbb{S}_n} W_n^\theta(x_n) = 1, \tag{6}$$

where now the members of each $\mathbb{S}_n$ are not obtained by a random proposal mechanism and are thus written in lower case. We are aiming for algorithms with upper bounded the cost per time instant by ensuring that $|\mathbb{S}_n| = (N+1) \wedge n$ for some parameter N. We now describe how the random support $\mathbb{S}_n$ is propagated from one iteration to the next in the algorithm of Fearnhead and Liu [2007]. In order to do so, for each n we specify a collection of binary-valued random variables, $\mathbf{S}_n := \{S_n(x_n)\}_{x_n \in E_n}$, which we refer to as survival indicator variables, and are precisely defined below. The propagation of the random support occurs according to a recursion:

$$\mathbb{S}_1 = \{0\}\,, \ \mathbb{S}_{n+1} = \{x_n \in \mathbb{S}_n : S_n(x_n) = 1\} \cup \{n\}, \ \ n = 1, 2, ... \tag{7}$$

Note that from (7), once a support point is lost, it is never subsequently recovered and the weights associated with it need never be computed. Discarding support points in this manner implies that the computational cost per SMC filtering iteration is bounded uniformly in time. Thus in the SMC forward filtering algorithm below it is implicit that, for any n and x_n, once $S_n(x_n)$ is set to zero, all subsequent survival indicators and weights associated with this point are also set to zero. We return to this issue in section 5.

The SMC forward filtering scheme of Fearnhead and Liu [2007] is presented below using our system of notation. A component of this algorithm is the stratified resampling procedure

[Carpenter et al., 1999], which is described in the appendix. As noted in [Fearnhead and Liu, 2007], in this context the stratified resampling procedure has the property of assigning either 1 or 0 offspring to each particle, thus avoiding duplicates.

SMC Forward Filtering

At time $n = 1$

• Set $\overline{W}_1^\theta(0) = g_1^\theta(0)$, $W_1^\theta(0) = 1$, $\mathbb{S}_1 = \{0\}$ and $S_1(0) = 1$.

At time $n = 2, 3, ..., T$

• If $n - 1 \leq N$ set $C_{n-1} = \infty$, otherwise set C_{n-1} to the unique solution of

$$\sum_{x_{n-1} \in \mathbb{S}_{n-1}} 1 \wedge C_{n-1} W_{n-1}^\theta(x_{n-1}) = N.$$

• Set $\mathcal{I}_{n-1} = \{x_{n-1} \in \mathbb{S}_{n-1} : W_{n-1}^\theta(x_{n-1}) \leq 1/C_{n-1}\}$.

• Maintain the $L_{n-1} := |\mathbb{S}_n| - |\mathcal{I}_{n-1}|$ support points $x_{n-1} \in \mathbb{S}_{n-1} \backslash \mathcal{I}_{n-1}$ by setting $S_{n-1}(x_{n-1}) = 1$. If $|\mathcal{I}_{n-1}| > 0$ resample $N - L_{n-1}$ times from $\mathcal{I}_{n-1}$ using the stratified resampling mechanism.

• For each x_{n-1} having survived the previous resampling steps, set $S_{n-1}(x_{n-1}) = 1$ and otherwise set $S_{n-1}(x_{n-1}) = 0$.

• Set for $x_n = n - 1$

$$\overline{W}_n^\theta(n-1) = g_n^\theta(n-1) \sum_{x_{n-1} \in \mathbb{S}_{n-1}} f_n^\theta(n-1|x_{n-1}) W_{n-1}^\theta(x_{n-1}).$$

• For $x_{n-1} \in \mathbb{S}_{n-1}$, set $\overline{W}_n^\theta(x_{n-1}) = 0$ if $S_{n-1}(x_{n-1}) = 0$ and otherwise set

$$\overline{W}_n^\theta(x_{n-1}) = g_n^\theta(x_{n-1}) f_n^\theta(x_{n-1}|x_{n-1}) \frac{W_{n-1}^\theta(x_{n-1})}{1 \wedge C_{n-1} W_{n-1}^\theta(x_{n-1})}.$$

• Update the support

$$\mathbb{S}_n = \{x_{n-1} \in \mathbb{S}_{n-1} : S_{n-1}(x_{n-1}) = 1\} \cup \{n-1\}$$

• For $x_n \in \mathbb{S}_n$,

$$W_n^\theta(x_n) \propto \overline{W}_n^\theta(x_n), \quad \sum_{x_n \in \mathbb{S}_n} W_n^\theta(x_n) = 1.$$

The computational complexity of the algorithm scales as $N \times T$ in contrast with that of the exact procedure which scales as T^2. This can be recovered by noticing that if $n-1 < N = T$ in the forward SMC filtering algorithm, then the resampling operation is never applied, in which case the SMC filtering algorithm performs exact filtering as described in section 3.2. In any case, the forward SMC filtering algorithm yields an approximation of the filtering distributions $\{p_\theta(x_n|y_{1:n})\}_{n=1}^T$ of the form (6) and an estimate of the likelihood $p_\theta(y_{1:T})$ given by

$$\hat{p}_\theta(y_{1:T}) := \prod_{n=1}^{T} \left(\sum_{x_n \in \mathbb{S}_n} \overline{W}_n^\theta(x_n) \right). \tag{8}$$

Given the sequence of approximate filtering distributions $\{\widehat{p}_\theta(x_n|y_{1:n})\}_{n=1}^T$, we can obtain an approximate sample from $p_\theta(\tau_{1:k}|y_{1:T})$ using a backward sampling recursion [Fearnhead and Liu, 2007] and use the already noticed property $x_{\tau_j+1} = \tau_j$. Again, if $N = T$, the forward filtering is exact and the below scheme then yields an exact sample from $p_\theta(\tau_{1:k}|y_{1:T})$.

SMC Backward Sampling

- Sample ι_1 from the distribution on $\mathbb{S}_T$ defined by $\{W_T^\theta(x_T)\}_{x_T \in \mathbb{S}_T}$ and set $k = 1$.
- While $\iota_k > 0$, sample ι_{k+1} from the distribution proportional to $f_{\iota_k+1}^\theta(\iota_k|\iota_{k+1}) W_{\iota_k}^\theta(\iota_{k+1})$ on $\mathbb{S}_{\iota_k}$ and

  set $k = k + 1$.
- Output the sampled sequence $\tau_{1:k}$ such that $\tau_j = \iota_{k-j+1}$.

---

4 Particle MCMC

In this section we introduce two new PMCMC algorithms. Each PMCMC algorithm is the analogue of a standard MCMC algorithm which uses the exact forward filtering/backward sampling methods of section 3.2 to compute $p_\theta(y_{1:T})$ and sample exactly from $p_\theta(\tau_{1:k}|y_{1:T})$ at cost T^2 per MCMC iteration. In section 5.1 we establish the validity of the new PMCMC algorithms: even though they employ SMC approximations of the exact forward filtering and backward sampling schemes, they are valid MCMC algorithms for sampling from $p(\theta, \tau_{1:k}|y_{1:T})$.

4.1 Particle Marginal Metropolis Hastings

A standard marginal Metropolis-Hastings (MMH) algorithm samples from $p(\theta, \tau_{1:k}|y_{1:T})$ using the joint proposal given by

$$q\left((\theta^*, \tau_{1:k}^*)|\,(\theta, \tau_{1:k})\right) = q\left(\theta^*|\,\theta\right) p_{\theta^*}\left(\tau_{1:k}^*|y_{1:T}\right) \ .$$

Note that in order to be consistent the notation for the proposed set of change-points should be $(\tau_{1:k})^*$ or $\tau_{1:k^*}^*$ instead of $\tau_{1:k}^*$ but that we adopt the latter in order to alleviate notation. This does not lead to possible confusion in what follows. In this scenario $\tau_{1:k}^*$ is proposed conditionally upon the proposed θ^*, and the resulting acceptance ratio is given by

$$\frac{p_{\theta^*}\left(y_{1:T}\right)\ p\left(\theta^*\right)}{p_{\theta}\left(y_{1:T}\right)\ p\left(\theta\right)} \frac{q\left(\theta|\theta^*\right)}{q\left(\theta^*|\theta\right)}. \tag{9}$$

If T is large, it is too expensive to compute the likelihood terms appearing in this ratio and sample from $p_{\theta^*}\left(\tau_{1:k}^*|y_{1:T}\right)$. We propose the following particle MMH (PMMH) sampler.

PMMH Sampler

Initialisation, $i = 0$

- Set $\theta(0)$ arbitrarily.
- Run the SMC forward filtering and backward sampling algorithms with parameter $\theta(0)$, yielding an approximation $\hat{p}_{\theta(0)}\left(y_{1:T}\right)$ and a change–point configuration $\tau_{1:k}\left(0\right)$.

For iteration $i \geq 1$

- Sample $\theta^* \sim q\left(\cdot|\,\theta(i-1)\right)$.
- Run the SMC forward filtering and backward sampling algorithms with parameter θ^*, yielding an approximation $\hat{p}_{\theta^*}\left(y_{1:T}\right)$ and a change–point configuration $\tau_{1:k}^*$.
- Set $\left(\theta(i), \tau_{1:k}(i), \hat{p}_{\theta(i)}(y_{1:T})\right) = (\theta^*, \tau_{1:k}^*, \hat{p}_{\theta^*}(y_{1:T}))$ with probability

$$1 \wedge \frac{\hat{p}_{\theta^*}\left(y_{1:T}\right)\ p\left(\theta^*\right)}{\hat{p}_{\theta(i-1)}\left(y_{1:T}\right)\ p\left(\theta(i-1)\right)} \frac{q\left(\theta(i-1)|\theta^*\right)}{q\left(\theta^*|\theta(i-1)\right)},$$

otherwise set $\left(\theta(i), \tau_{1:k}(i), \hat{p}_{\theta(i)}(y_{1:T})\right) = \left(\theta(i-1), \tau_{1:k}(i-1), \hat{p}_{\theta(i-1)}(y_{1:T})\right)$.

4.2 Particle Gibbs Sampler

It is possible to implement a block Gibbs algorithm which samples from $p(\theta, \tau_{1:k}|y_{1:T})$ using draws from $p_\theta(\tau_{1:k}|y_{1:T})$ and $p(\theta|y_{1:T}, \tau_{1:k})$. If T is large, it is too expensive to sample

exactly from $p_\theta(\tau_{1:k}|y_{1:T})$, so we propose the following particle Gibbs (PG) sampler. The corresponding conditional resampling algorithm is given in the appendix.

PG Sampler

Initialisation, $i = 0$

• Set $\theta(0), X_{1:T}(0)$ arbitrarily.

For iteration $i \geq 1$

• With parameter $\theta(i-1)$ and given $\tau_{1:k}(i-1)$, run the conditional SMC forward filtering algorithm and then run the SMC backward sampling algorithm to obtain a change–point configuration $\tau_{1:k}(i)$.

• Sample $\theta(i) \sim p(\cdot|y_{1:T}, \tau_{1:k}(i))$.

This PG algorithm relies on the conditional forward filtering algorithm given below and the conditional stratified resampling scheme which is described in the appendix.

5 Validity of the Algorithms

5.1 SMC and Target Distributions

To establish the validity of the PMCMC algorithms, we need to express precisely the probability law of the SMC approximation, which turns out to take a simple form when expressed in terms of the survival variables $\{\mathbf{S}_n\}_{n=1}^{T-1}$ introduced in section 3.3. Before proceeding we note that the SMC forward filtering algorithm has the following key properties, which can be verified by inspection of the algorithm:

1. under functional relationships between $\{\mathbb{S}_n\}_{n=1}^{T}$, $\{\mathbf{S}_n\}_{n=1}^{T-1}$ and $\{\mathbf{W}_n^\theta\}_{n=1}^{T}$ implicitly specified by the algorithm, all information relevant to the random measures as in (5) for $n = 1, ..., T$, is carried by $\{\mathbf{S}_n\}_{n=1}^{T-1}$,

2. under the probability law implicitly specified by the algorithm, for each n, $\mathbf{S}_n$ is conditionally independent of the history of the algorithm, given $\mathbf{W}_n^\theta$, and

3. for $n \leq N$, $|\mathbb{S}_n| = n$, and for $n > N$, $|\mathbb{S}_n| = N + 1$ with probability (w.p.) 1 (c.f. the support of the exact filtering distributions which grows deterministically over time (3)).

Conditional SMC Forward Filtering

Input: change-point locations $\tau_{1:k}$, and θ.

At time $n = 1$

• Set $\overline{W}_1^\theta(0) = g_1^\theta(0)$, $W_1^\theta(0) = 1$, $\mathbb{S}_1 = \{0\}$, $S_1(0) = 1$ and $\kappa = 0$.

At time $n = 2, ..., T$

• If $n - 1 \leq N$ set $C_{n-1} = \infty$, otherwise set C_{n-1} to the unique solution of

$$\sum_{x_{n-1} \in \mathbb{S}_{n-1}} 1 \wedge C_{n-1} W_{n-1}^\theta(x_{n-1}) = N.$$

• If $\tau_{\kappa+1} < n - 1$ set $\kappa = \kappa + 1$.

• Set $\mathcal{I}_{n-1} = \{x_{n-1} \in \mathbb{S}_{n-1} : W_{n-1}^\theta(x_{n-1}) \leq 1/C_{n-1}\}$.

• If $\tau_\kappa \notin \mathcal{I}_{n-1}$, maintain the $L_{n-1} = |\mathbb{S}_{n-1}| - |\mathcal{I}_{n-1}|$ support points $x_{n-1} \in \mathbb{S}_{n-1} \backslash \mathcal{I}_{n-1}$ (which includes τ_κ) by setting $S_{n-1}(x_{n-1}) = 1$. Then if $|\mathcal{I}_{n-1}| > 0$ resample $N - L_{n-1}$ times from $\mathcal{I}_{n-1}$ using the stratified resampling mechanism.

• If $\tau_\kappa \in \mathcal{I}_{n-1}$, maintain the $L_{n-1} = n - 1 - |\mathcal{I}_{n-1}|$ support points $x_{n-1} \in \mathbb{S}_{n-1} \backslash \mathcal{I}_{n-1}$ (which does not include τ_κ) by setting $S_{n-1}(x_{n-1}) = 1$. Then resample $N - L_{n-1}$ times from $\mathcal{I}_{n-1}$ using the conditional stratified resampling mechanism.

• For each x_{n-1} having survived the previous resampling steps, set $S_{n-1}(x_{n-1}) = 1$ otherwise $S_{n-1}(x_{n-1}) = 0$.

• Set

$$\overline{W}_n^\theta(n-1) = g_n^\theta(n-1) \sum_{x_{n-1} \in \mathbb{S}_{n-1}} f_n^\theta(n-1|x_{n-1}) W_{n-1}^\theta(x_{n-1}).$$

• For $x_{n-1} \in \mathbb{S}_{n-1}$, set $\overline{W}_n^\theta(x_{n-1}) = 0$ if $S_{n-1}(x_{n-1}) = 0$ and otherwise

$$\overline{W}_n^\theta(x_{n-1}) = g_n^\theta(x_{n-1}) f_n^\theta(x_{n-1}|x_{n-1}) \frac{W_{n-1}^\theta(x_{n-1})}{1 \wedge C_{n-1} W_{n-1}^\theta(x_{n-1})}.$$

• Update the support

$$\mathbb{S}_n = \{x_{n-1} \in \mathbb{S}_{n-1} : S_{n-1}(x_{n-1}) = 1\} \cup \{n-1\}.$$

• For $x_n \in \mathbb{S}_n$, set

$$W_n^\theta(x_n) \propto \overline{W}_n^\theta(x_n), \quad \sum_{x_n \in \mathbb{S}_n} W_n^\theta(x_n) = 1.$$

We first write an expression for the distribution of the random variables $\mathbf{S}_1, \mathbf{S}_2, ..., \mathbf{S}_{T-1}$ generated through the SMC forward filtering algorithm. By construction, we have

$$\mathbf{S}_n | \left(\mathbf{W}_n^\theta = \mathbf{w}_n^\theta \right) \sim r_{n,\theta}^N(\cdot | \mathbf{w}_n^\theta). \tag{10}$$

This density is parameterized by N and for all $n \geq N$,

$$\sum_{x_n \in \mathbb{S}_n} S_n(x_n) = N, \quad \text{w.p. 1.}$$

We will not need an explicit expression for the density (10), but from the definition of the stratified resampling mechanism [Fearnhead and Clifford, 2003], we know that it has the following marginal property: for all $x_n \in \{0, 1, ..., n-1\}$,

$$r_{n,\theta}^N(S_n(x_n) = 1 | \mathbf{w}_n^\theta) = 1 \wedge c_n w_n^\theta(x_n), \tag{11}$$

(where c_n is the value counterpart of C_n used in the SMC forward filtering algorithm) which implies that

$$r_{n,\theta}^N(S_n(x_n) = 1 | w_n^\theta(x_n) = 0) = 0,$$

so combined with Eq. (10) we see that for any $n > 0$ and $x \in \{1, ..., n-1\}$, conditional on the event that, $W_{n-1}^\theta(x) = 0$, at any subsequent $k \geq n$, $W_k^\theta(x) = 0$ w.p. 1. Thus the corresponding subsequent survival indicators and weights need never be simulated or stored. To summarize the law of the SMC forward filtering algorithm, we can write the distribution of $\mathbf{S}_1, \mathbf{S}_2, ..., \mathbf{S}_{T-1}$ on $\prod_{n=1}^{T-1} [\{0,1\}^n]$ as

$$\psi_\theta^N(\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_{T-1}) = \prod_{n=1}^{T-1} r_{n,\theta}^N(\mathbf{s}_n | \mathbf{w}_n^\theta). \tag{12}$$

The weights $\mathbf{W}_n^\theta$ being just a deterministic function of $\mathbf{S}_1, \ldots, \mathbf{S}_{n-1}$, it is not necessary to introduce them as arguments of ψ_θ^N.

The key to establishing the validity of the PMCMC algorithms is to define the artificial joint probability density for θ, $\tau_{1:k}$ and $\mathbf{S}_1, ..., \mathbf{S}_{T-1}$ on $\Theta \times \mathcal{T}_T \times \prod_{n=1}^{T-1} \{0,1\}^n$ in 13, with the convention that the products are equal to unity when $k = 0$. By construction (13) admits

$p(\theta, \tau_{1:k}|\, y_{1:T})$ as a marginal.

$$\begin{aligned}\pi^N(\theta, \tau_{1:k}, \mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_{T-1}) := p(\theta, \tau_{1:k}|y_{1:T})\psi_\theta^N(\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_{T-1}) \\ \cdot \frac{\prod_{j=1}^{k}\prod_{n=\tau_{j-1}+1}^{\tau_j}\mathbb{I}[s_n(\tau_{j-1})=1]}{\prod_{j=1}^{k}\prod_{n=\tau_{j-1}+1}^{\tau_j} r_{n,\theta}^N(s_n(\tau_{j-1})=1|\mathbf{w}_n^\theta)} \\ \cdot \frac{\prod_{n=\tau_k+1}^{T-1}\mathbb{I}[s_n(\tau_k)=1]}{\prod_{n=\tau_k+1}^{T-1} r_{n,\theta}^N(s_n(\tau_k)=1|\mathbf{w}_n^\theta)},\end{aligned} \tag{13}$$

5.2 Convergence Results

We show in the following theorems that the PMMH and PG algorithms are just standard MCMC updating schemes targeting (13). Furthermore, the convergence to $p(\theta, \tau_{1:k}|y_{1:T})$ of the distribution of the samples of $\theta, \tau_{1:k}$ they generate is inherited from the corresponding standard MCMC schemes. Proofs are given in the appendix.

We first deal with the PMMH algorithm and consider the following assumption.

(**A1**) The MMH sampler of target density $p\left(\theta, \tau_{1:k}|\, y_{1:T}\right)$ and proposal density $q\left(\theta^*|\,\theta\right) p_{\theta^*}\left(\tau_{1:k}^*|y_{1:T}\right)$ is irreducible and aperiodic (and hence converges for almost all starting points).

With $\|\cdot\|_{tv}$ the total variation distance, we have the following results.

Theorem 1.

1. For any $N \geq 1$, the PMMH algorithm is a Metropolis-Hastings (MH) sampler on the space
$\Theta \times \mathcal{T}_T \times \prod_{n=1}^{T-1}\left[\{0,1\}^n\right]$ *with target density $\pi^N(\theta, \tau_{1:k}, \mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_{T-1})$ defined in Eq. (13) and using the proposal:*

$$q(\theta^*|\theta)\cdot\psi_{\theta^*}^N(\mathbf{s}_1^*, ..., \mathbf{s}_{T-1}^*)\cdot w_T^{\theta^*}(\tau_k^*)\prod_{j=1}^{k^*}\left\{\frac{f_{\tau_j^*+1}^{\theta^*}(\tau_j^*|\tau_{j-1}^*)w_{\tau_j^*}^{\theta^*}(\tau_{j-1}^*)}{\sum_{x\in\mathbb{S}_{\tau_j^*}} f_{\tau_j^*+1}^{\theta^*}(\tau_j^*|x)w_{\tau_j^*}^{\theta^*}(x)}\right\}. \tag{14}$$

*2. If additionally **(A1)** holds, the PMMH sampler generates a sequence $\{\theta(i), \tau_{1:k}(i)\}$ whose marginal distributions $\mathcal{L}^N((\theta(i), \tau_{1:k}(i)) \in \cdot)$ satisfies*

$$\|\mathcal{L}^N((\theta(i), \tau_{1:k}(i)) \in \cdot) - p(\cdot|y_{1:T})\|_{tv} \to 0 \;\; as \;\; i \to \infty.$$

We now turn to the PG sampler and consider the following assumption on the corresponding standard Gibbs sampler.

(**A2**) The block Gibbs sampler which samples from $p_\theta(\tau_{1:k}|\, y_{1:T})$ and $p(\theta|\tau_{1:k}, y_{1:T})$ is irreducible and aperiodic (and hence converges for almost all starting points).

We can establish the following result.

Theorem 2.

1. For any $N \geq 2$, the PG sampler defines a transition kernel on the space $\Theta \times \mathcal{T}_T \times \prod_{n=1}^{T-1}[\{0,1\}^n]$ of invariant density $\pi^N(\theta, \tau_{1:k}, \mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_{T-1})$. One PG iteration is equivalent to sampling from the sequence of conditional distributions $\pi^N(\mathbf{s}_1, ..., \mathbf{s}_{T-1}|\theta, \tau_{1:k})$, $\pi^N(\tau_{1:k}|\theta, \mathbf{s}_1, ..., \mathbf{s}_{T-1})$ and $\pi^N(\theta|\tau_{1:k})$.

*2. If additionally **(A2)** holds, then the PG sampler generates a sequence $\{\theta(i), \tau_{1:k}(i)\}$ whose marginal distributions $\mathcal{L}^N((\theta(i), \tau_{1:k}(i)) \in \cdot)$ satisfies*

$$\|\mathcal{L}^N((\theta(i), \tau_{1:k}(i)) \in \cdot) - p(\cdot|y_{1:T})\|_{tv} \to 0 \;\; as \;\; i \to \infty.$$

6 Examples

6.1 Well-Log Data

We consider a piecewise constant model which was used in [Fearnhead, 2006] to analyze well-log data. The prior distribution h_θ is geometric with unknown success probability p, over which a flat Beta hyper-prior is placed. Given change-points $\tau_{j-1:j}$ the observations $y_{\tau_{j-1}+1:\tau_j}$ in the j^{th} segment are i.i.d. $\mathcal{N}(\mu_j, \sigma^2)$, where μ_j is the mean associated with segment j. Given $\tau_{1:k}$ and hyper parameters η and α, the segment means, $\{\mu_j\}_{j=1}^{k+1}$ are i.i.d. $\mathcal{N}(\eta, \sigma^2\alpha^2)$. Finally, uninformative, improper priors are placed over the hyper-parameters: $p(\eta) \propto 1$, $p(\sigma) \propto 1/\sigma$ and $p(\alpha) \propto 1/\alpha$. The parameter vector is then $\theta = [\eta \; p \; \sigma \; \alpha]^{\mathrm{T}}$. We refer to [Fearnhead, 2006] for the expression for the likelihood of observations in one segment, given values of η, α and σ, with the corresponding value of μ_j integrated out.

The data, originating from [Ó Ruanaidh and Fitzgerald, 1996], are shown in figure 1 and consist of $T = 4050$ measurements of the nuclear magnetic response of underground rocks. As in [Fearnhead, 2006], a small number of outlying observations were manually removed from the data set. For comparison with PMCMC, we considered the block Gibbs sampler which samples from $p_\theta(\tau_{1:k}|y_{1:T})$ (using the exact method of Fearnhead and Liu

[2007]), $p(\sigma, p, \{\mu_j\}_{j=1}^{k+1} | y_{1:T}, \tau_{1:k}, \eta, \alpha)$ and $p(\eta, \alpha | y_{1:T}, \tau_{1:k}, \sigma, p, \{\mu_j\}_{j=1}^{k+1})$. In Fearnhead [2006] it was shown that for the same data set and model, an algorithm of this form significantly outperformed the MCMC scheme of Lavielle and Lebarbier [2001].

The above block Gibbs sampler was compared with the PG algorithm obtained by substituting in place of the draw from $p_\theta(\tau_{1:k}|y_{1:T})$ a conditional SMC forward filtering/backward sampling step. Figure 2 shows autocorrelation plots for each component of θ, for the block Gibbs and PG algorithm with $N = 50$. For this data record $T = 4050$, so the computational cost of the PG algorithm is an order of magnitude less than that of the block Gibbs sampler. Again the complexity of the SMC forward filtering algorithm scales as $N \times T$ while the exact algorithm scales as T^2. The auto-correlation plots indicate that the block Gibbs sampler is extremely efficient for this model, with the autocorrelation dropping to zero well before lag 10. The performance of the PG algorithm is almost identical to that of the block Gibbs. Using other values of N resulted in auto-correlation curves which were too similar to be displayed clearly on the same axes. For $N < 50$ it was found that there was a small difference in the corresponding approximations of the posterior marginal for the parameter p as the PG scheme did not fully explore its support (not shown).

Figure 1 shows estimated marginal posterior probabilities of change-point occurrences, from the block Gibbs and the PG algorithm with $N = 50$. These posterior probabilities are indistinguishable to the eye and the results are very similar to those reported in [Fearnhead, 2006] with any discrepancies likely to be attributable to which outlying observations were removed (Fearnhead [2006] does not state how this was performed).

Figure 2 also shows approximations of the marginal posterior distributions for each of the four parameters obtained with the PG algorithm. Identical histograms were obtained from the block Gibbs output. The modes of these posterior marginals coincide with the results reported in Fearnhead [2006].

6.2 Coal Mining Disasters

We consider a piecewise constant Poisson intensity model for the classic Coal Mining Disaster data set of Jarrett [1979]. This data set consists of the dates of 191 disasters between 1851 and 1962. The data were analysed by Green [1995] via a reversible–jump MCMC algorithm for a continuous-time model and in [Fearnhead, 2006] using the exact sampling methods.

Following [Fearnhead, 2006] we discretize time and form observations by counting the

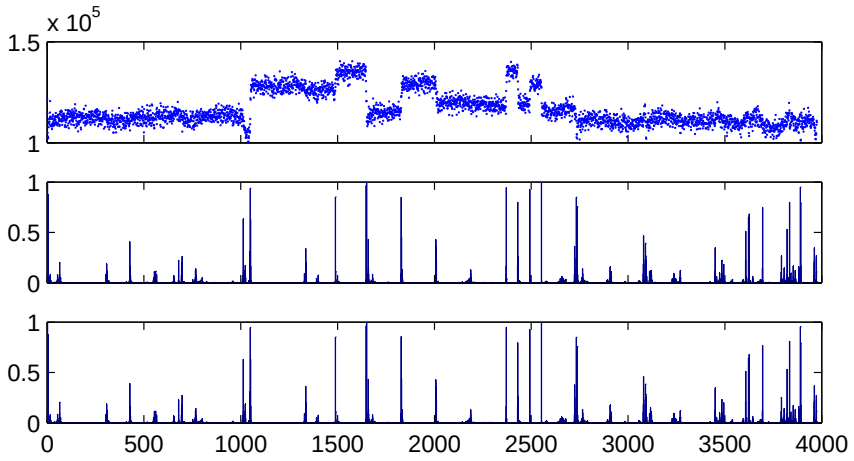

Figure 1: Well-log example. Top: well-log data. Approximate marginal posterior probabilities of a change-point at each time index obtained from PG sampler with $N = 50$ (middle) and the Block Gibbs sampler (bottom).

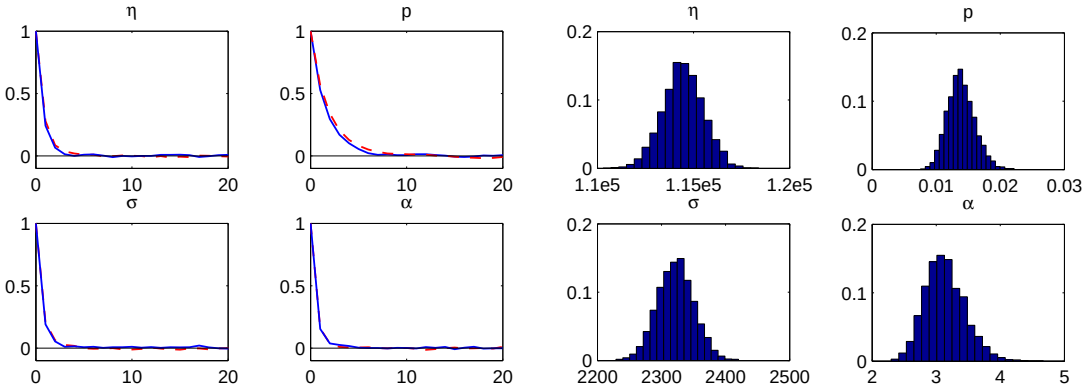

Figure 2: Well-log example. The left pane shows auto-correlation plots. Solid: Block Gibbs. Dashed: PG with $N = 50$. The right pane shows approximate posterior marginals obtained from the PG algorithm. Identical histograms were obtained using the Block Gibbs sampler.

number of disasters each week. This yields $T = 5844$ observations. Given change-points $\tau_{j-1:j}$ the observations $y_{\tau_{j-1}+1:\tau_j}$ in the j^{th} segment are i.i.d. $\mathcal{P}o(\lambda_j)$, where λ_j is the intensity associated with segment j. Given $\tau_{1:k}$ the segment intensities, $\{\lambda_j\}_{j=1}^{k+1}$ are i.i.d. $\mathcal{G}(1, 200/7)$. As opposed to choosing the prior distribution on inter-change-point times, in [Fearnhead, 2006] a Poisson prior was placed over the number of change-points. The intensity parameter of this Poisson prior was treated as fixed and chosen to give a mean of 3 change-points over the duration of the observation record. Given the number of change–points k, the change–point locations were a priori given by the even order statistics of $2k + 1$ uniform draws on $\{1, ..., T - 1\}$. We investigate a different approach via the distribution on inter-change-point

times, with parameters to be inferred from the data. In our model the prior density h_θ is negative binomial with parameters r and p, over which gamma and beta hyper-priors were placed, respectively. The parameters of the hyper-priors were chosen so as to provide some penalisation of very short segments whilst still allowing flexibility: for the gamma hyper-prior on r the shape parameter was 10 and the scale parameter 1; for the beta prior on p the parameters were $(1, 10)$. In this model the parameter is $\theta = [r\ p]^{\mathrm{T}}$.

We compared the PMMH algorithm with a standard MMH algorithm employing exact computation of $p_\theta(y_{1:T})$ over $30,000$ MCMC iterations. Random walk proposals were made over r and p with standard deviations of 1 and 0.005, respectively. These values were chosen after a preliminary run and resulted in an average acceptance probability of 0.48 for the MH algorithm.

Figure 3 shows approximations of the marginal and joint posterior distributions of the two parameters from the output of the PMMH algorithm with $N = 200$. Identical histograms were obtained using the standard MMH algorithm. The figure also shows auto–correlation plots for the two parameters. In these plots, the solid line corresponds to the exact MMH algorithm. The dashed and dash–dot lines correspond to the PMMH algorithm with $N = 200$ and $N = 50$, respectively. In terms of auto-correlation, there is very little difference between the exact MMH and PMMH with $N = 200$. For the parameter r, the difference between the dash-dot, dashed and solid lines is more noticeable, but still not very large. Figure 4 shows the acceptance rate in the PMMH algorithm against N. Only as the number of particles falls below 200 does the acceptance rate fall significantly. For this data set, $N = T = 5844$, corresponds to the exact MMH algorithm; with N an order of magnitude less than this value the performance of the PMMH algorithm is very good.

Figure 4 also shows the approximate marginal posterior distributions over the number of change–points and over the location of change–points, given that there are 2, obtained from the PMMH algorithm with $N = 200$. Again, identical results were obtained with the MMH algorithm. There are some similarities and interesting differences between these results and those obtained in [Fearnhead, 2006] under the Poisson change–point prior with fixed parameter. Firstly, the marginal over the number of change–points shown in figure 4 exhibits a stronger mode at 2 than that obtained in [Fearnhead, 2006], but is otherwise similar. Secondly, the conditional posterior for the change–point locations, given there are 2, exhibits two strong modes in the same locations as found in [Fearnhead, 2006]. However the results in figure 4, obtained using the PMMH algorithm with the negative binomial

inter-change-point prior, exhibit a third, weaker mode around time 4900. This third mode was not reported in [Fearnhead, 2006]. In [Fearnhead, 2006] it was not reported what kernel bandwidth was used, so precise comparisons are difficult. However, even with a relatively large bandwidth the third mode shown in figure 4 was evident. This shows how the values of parameters can affect inferences drawn about change–point locations.

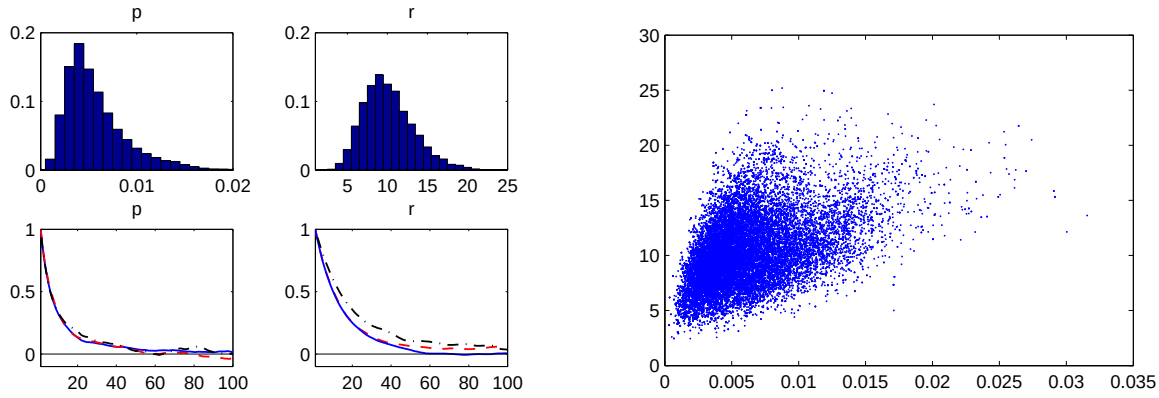

Figure 3: Coal Mining disasters example. Left: Approximate posterior marginals using PMMH and auto–correlation plots for the two parameters. In auto-correlation plots, solid: Exact MMH. dashed: PMMH with $N = 200$, dash-dot: PMMH with $N = 50$.

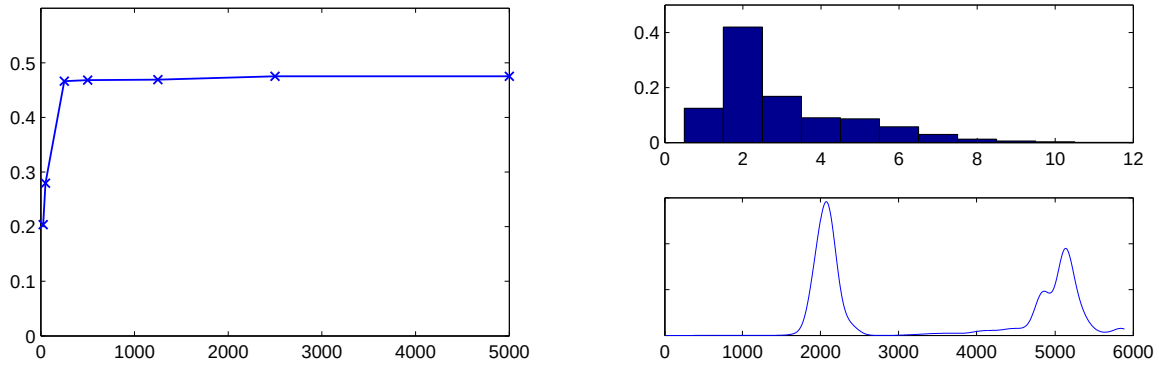

Figure 4: Coal Mining Disasters example. Left: Acceptance rate vs. N. Right top: Approximate posterior over number of change–points from PMMH algorithm. Right bottom: kernel smoothed posterior marginal for change–point locations conditional on 2 change–points.

7 Discussion

We have proposed PMCMC algorithms for multiple change-point models which rely on the efficient SMC method proposed by Fearnhead and Liu [2007] to approximate the filtering distributions and likelihood. These PMCMC algorithms have a cost per MCMC iteration of the order $N\,T$, where N is the number of particles in the SMC algorithm, compared to T^2

for the "exact" MCMC algorithms relying on the exact filtering distributions and likelihood. We have demonstrated experimentally that these PMCMC algorithms perform remarkably well compared to the "exact" algorithms even for N an order of magnitude smaller than T. This is an attractive feature for performing Bayesian inference in multiple change–point models with long data records.

8 Appendix

8.1 Resampling algorithms

Stratified Resampling

Input: $\mathcal{I}_{n-1}$ and the corresponding weights.

• Normalise the weights by setting for $x_{n-1} \in \mathcal{I}_{n-1}$

$$\widehat{W}^{\theta}_{n-1}(x_{n-1}) \propto W^{\theta}_{n-1}(x_{n-1}), \sum_{x_{n-1}\in\mathcal{I}_{n-1}} \widehat{W}^{\theta}_{n-1}(x_{n-1}) = 1$$

and construct the corresponding c.d.f,

$$Q^{\theta}_{n-1}(x_{n-1}) = \sum_{\{x'_{n-1}\in\mathcal{I}_{n-1}:x'_{n-1}\leq x_{n-1}\}} \widehat{W}^{\theta}_{n-1}(x'_{n-1}).$$

• Sample U_1 uniformly on $[0, 1/(N - L_{n-1})]$, then set $U_p = U_{p-1} + \dfrac{1}{N - L_{n-1}}$ for $p = 2, .., N - L_{n-1}$.

• Each particle in $\mathcal{I}_{n-1}$ is assigned $O_{x_{n-1}} \in \{0, 1\}$ offspring where

$$O_{x_{n-1}} = \#\left\{U_p : Q^{\theta}_{n-1}(x_{n-1} - 1) \leq U_p \leq Q^{\theta}_{n-1}(x_{n-1})\right\}.$$

Conditional Stratified Resampling

Input: $\mathcal{I}_{n-1}$, the corresponding weights and τ_κ.

- Normalise the weights by setting for $x_{n-1} \in \mathcal{I}_{n-1}$

$$\widehat{W}^{\theta}_{n-1}(x_{n-1}) = \frac{W^{\theta}_{n-1}(x_{n-1})}{\sum_{x'_{n-1}\in\mathcal{I}_{n-1}} W^{\theta}_{n-1}(x'_{n-1})}$$

and construct the corresponding c.d.f,

$$Q^{\theta}_{n-1}(x_{n-1}) = \sum_{\{x'_{n-1}\in\mathcal{I}_{n-1}:x'_{n-1}\leq x_{n-1}\}} \widehat{W}^{\theta}_{n-1}(x'_{n-1}).$$

- Sample U^* uniformly on $\left[Q^{\theta}_{n-1}(\tau_\kappa - 1), Q^{\theta}_{n-1}(\tau_\kappa)\right]$, set $U_1 = U^* - \frac{\lfloor (N - L_{n-1})U^* \rfloor}{N - L_{n-1}}$ and $U_p = U_{p-1} + \frac{1}{N - L_{n-1}}$ for $p = 2, .., N - L_{n-1}$.
- Each particle is assigned $O_{x_{n-1}} \in \{0, 1\}$ offspring where

$$O_{x_{n-1}} = \#\left\{U_p : Q^{\theta}_{n-1}(x_{n-1} - 1) \leq U_p \leq Q^{\theta}_{n-1}(x_{n-1})\right\}.$$

8.2 Proofs

Proof of Theorem 1.

1. The PMMH algorithm makes proposals from (14): the left-hand term is the proposal distribution for the new parameter θ^*; the middle term is the law of the SMC forward filtering algorithm with this parameter; the right-hand term is the conditional probability of obtaining a sequence of change-points from the SMC backward sampling algorithm. Furthermore (14) is a distribution over the same space as (13). From the definition of the change-point model in section 2, we may write equation (15) with, whenever $\tau_j > \tau_{j-1} + 1$, $\phi(\tau_{j-1}, \tau_j) := \prod_{n=\tau_{j-1}+2}^{\tau_j} g^{\theta}_n(\tau_{j-1}) f^{\theta}_n(\tau_{j-1}|\tau_{j-1})$. From the definition of the SMC forward filtering algorithm, for two consecutive change-point locations $\tau_j > \tau_{j-1} > 0$, on the event $\prod_{n=\tau_{j-1}+1}^{\tau_j} S_n(\tau_{j-1}) = 1$, we have the expansion

of the weight $w^{\theta}_{\tau_j}(\tau_{j-1})$ in equation (16).

$$
\begin{aligned}
&p_\theta(\tau_{1:k}, y_{1:T}) \\
=&\left\{\left[\mathbb{I}\left[T > \tau_k + 1\right]\left(\prod_{n=\tau_k+2}^{T} g^{\theta}_n(\tau_k) f^{\theta}_n(\tau_k|\tau_k)\right) + \mathbb{I}[T = \tau_k + 1]\right] g^{\theta}_{\tau_k+1}(\tau_k) f^{\theta}_{\tau_k+1}(\tau_k|\tau_{k-1})\right\} \\
\cdot &\prod_{j=2}^{k}\left\{\left[\mathbb{I}\left[\tau_j > \tau_{j-1} + 1\right]\phi(\tau_{j-1}, \tau_j) + \mathbb{I}\left[\tau_j = \tau_{j-1} + 1\right]\right] g^{\theta}_{\tau_{j-1}+1}(\tau_{j-1}) f^{\theta}_{\tau_{j-1}+1}(\tau_{j-1}|\tau_{j-2})\right\} \\
&\cdot \left\{\left[\mathbb{I}\left[\tau_1 > 1\right]\left(\prod_{n=2}^{\tau_1} g^{\theta}_n(0) f^{\theta}_n(0|0)\right) + \mathbb{I}\left[\tau_1 = 1\right]\right] g^{\theta}_1(0) f^{\theta}_1(0|0)\right\}. \quad (15)
\end{aligned}
$$

$$
\begin{aligned}
&w^{\theta}_{\tau_j}(\tau_{j-1}) \\
&= \left\{\mathbb{I}\left[\tau_j > \tau_{j-1} + 1\right]\left(\prod_{n=\tau_{j-1}+2}^{\tau_j} \frac{g^{\theta}_n(\tau_{j-1}) f^{\theta}_n(\tau_{j-1}|\tau_{j-1})}{1 \wedge c_{n-1} w^{\theta}_{n-1}(\tau_{j-1})} \frac{1}{\sum_{x\in\mathbb{S}_n} \overline{w}^{\theta}_n(x)}\right) + \mathbb{I}\left[\tau_j = \tau_{j-1} + 1\right]\right\} \\
&\quad \cdot \left(g^{\theta}_{\tau_{j-1}+1}(\tau_{j-1}) \sum_{x\in\mathbb{S}_{\tau_{j-1}}} f^{\theta}_{\tau_{j-1}+1}(\tau_{j-1}|x) w^{\theta}_{\tau_{j-1}}(x)\right)\left(\frac{1}{\sum_{x\in\mathbb{S}_{\tau_{j-1}+1}} \overline{w}^{\theta}_{\tau_{j-1}+1}(x)}\right). \quad (16)
\end{aligned}
$$

By using completely analogous expansions of $w^{\theta}_T(\tau_k)$ and $w^{\theta}_{\tau_1}(0)$ on the corresponding events and by using (15), (8) and (11) it follows by elementary manipulations that the acceptance probability given in the PMMH algorithm is precisely that of an MH sampler targeting the extended target distribution (13) and proposing from (14).

2. Proof of the second component of the theorem is a direct consequence of the arguments in Theorem 1 in [Andrieu and Roberts, 2009] and **(A1)**.

Proof of Theorem 2.

1. For the first sampling step stated in the theorem, it is straightforward to check that running the conditional SMC forward filtering algorithm is equivalent to sampling from $\pi^N(\mathbf{s}_1, ..., \mathbf{s}_{T-1}|\theta, \tau_{1:k})$. Now consider the second step. Recalling the definition of the extended target distribution (13), using again on the event $\prod_{n=\tau_{j-1}+1}^{\tau_j} S_n(\tau_{j-1}) = 1$, the expansion (16), and using (15) and (11) we conclude that

$$
\pi^N(\tau_{1:k}|\theta, \mathbf{s}_1, ..., \mathbf{s}_{T-1}) = w^{\theta}_T(\tau_k) \prod_{j=1}^{k} \frac{f^{\theta}_{\tau_j+1}(\tau_j|\tau_{j-1}) w^{\theta}_{\tau_j}(\tau_{j-1})}{\sum_{x\in\mathbb{S}_{\tau_j}} f^{\theta}_{\tau_j+1}(\tau_j|x) w^{\theta}_{\tau_j}(x)},
$$

where the right-hand term is the conditional probability of sampling the change–point configuration $\tau_{1:k}$ using the SMC backward sampling algorithm, given the outcome of the forward sampling. For the third sampling step it is direct that $p(\theta|y_{1:T}, \tau_{1:k}) = \pi^N(\theta|\tau_{1:k})$.

2. We establish irreducibility and aperiodicity of the transition probability of the PG algorithm. We denote by $\mathcal{L}_G$ the law of the standard Gibbs sampler to which assumption **(A2)** applies and $\mathcal{L}_{PG}^N$ the law of the PG sampler using N particles.

Let $A \times B \times C \in \mathcal{B}(\Theta) \times \mathcal{B}(\mathcal{T}_T) \times \mathcal{B}\left(\prod_{n=1}^{T-1}\{0,1\}^n\right)$ be such that $\pi^N(\theta \in A, \tau_{1:k} \in B, \mathbf{S}_1, ..., \mathbf{S}_{T-1} \in C) > 0$. It follows from (13) that $p((\theta, \tau_{1:k}) \in A \times B|y_{1:T}) > 0$ and then from irreducibility of the corresponding block Gibbs sampler (assumption **(A2)**) there exists a finite i such that $\mathcal{L}_G((\theta(i), \tau_{1:k}(i)) \in A \times B) > 0$.

From the specification of the conditional SMC forward filtering scheme, for any $\theta \in \Theta$, $N \geq 2$, given any $\tau_{1:k}$ and for any time step, any particle which has positive weight immediately before resampling has a positive probability of surviving that resampling step. It follows that for any $n = 1, ..., T-1$, any point in the support of $p_\theta(x_n|y_{1:n})$ has positive probability of being assigned a positive weight at time n. It follows from the definition of $X_{1:T}$ that any point in the support of $p_\theta(\tau_{1:k}|y_{1:n})$ has positive probability of being selected in the backward sampling. Then the $A \times B$ from above is marginally an accessible set of the PG sampler for the same i: i.e. $\mathcal{L}_{PG}^N((\theta(i), X_{1:T}(i)) \in A \times B) > 0$. Furthermore, as the conditional forward SMC filtering corresponds to drawing from $\pi^N(\mathbf{s}_1, ..., \mathbf{s}_{T-1}|\theta, \tau_{1:k})$,

$$\mathcal{L}_{PG}^N\left((\theta(i+1), \tau_{1:k}(i+1), \mathbf{S}_1(i+1), ..., \mathbf{S}_{T-1}(i+1)) \in A \times B \times C\right) > 0$$

and irreducibility follows. Furthermore, aperiodicity of the PG sampler holds by contradiction: if the PG sampler were periodic, then the Gibbs sampler would be too; this violates **(A2)**.

References

C. Andrieu and G.O. Roberts. The pseudo-marginal approach for efficient computation. *Ann. Statist.*, 37:697–725, 2009.

C. Andrieu, A. Doucet, and R. Holenstein. Particle Markov chain Monte Carlo methods (with discussion). *J. Roy. Stat. Soc. B*, 72:1–33, 2010.

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

D. Barry and J.A. Hartigan. A Bayesian analysis for change point problems. *J. Am. Stat. Soc.*, 88 (421):309–319, 1993.

B.P. Carlin, A.E. Gelfand, and Smith A.F.M. Hierarchical Bayesian analysis of change-point problems. *Appl. Statist.*, 41:389–405, 1992.

J. Carpenter, P. Clifford, and P. Fearnhead. An improved particle filter for non-linear problems. *IEE Proc. F*, 146:2–7, 1999.

S. Chib. Estimation and comparison of multiple change-point models. *J. Econometr.*, 98:221–241, 1998.

N. Chopin. Dynamic detection of change points in long time series. *Ann. Inst. Statist. Math.*, 59: 349–366, 2007.

A. Doucet, J.F.G. de Freitas, and N.J. Gordon, editors. *Sequential Monte Carlo Methods in Practice*, New York, 2001. Springer-Verlag.

P. Fearnhead. Efficient and exact Bayesian inference for multiple changepoint problems. *Statistics and Computing*, 16:203–213, 2006.

P. Fearnhead and P. Clifford. On-line inference for hidden Markov models via particle filters. *J. Roy. Stat. Soc. B*, 65:887–899, 2003.

P. Fearnhead and P. Clifford. Exact Bayesian curve fitting and signal segmentation. *IEEE Trans. Sig. Proc.*, 53(6):2160–2166, 2005.

P. Fearnhead and Z. Liu. On-line inference for multiple changepoint problems. *J. R. Statist. Soc. B*, 69(4):589–605, 2007.

P.J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

R.G. Jarrett. A note on the intervals between coal-mining disasters. *Biometrika*, 66:191–193, 1979.

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

J.S. Liu, W.H. Wong, and A. Kong. Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika*, 81(1):27–40, 1994.

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

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

N. Whiteley, C. Andrieu, and A. Doucet. Efficient Bayesian inference for switching state-space models using discrete particle Markov chain Monte Carlo methods. *ArXiv Preprint*, 1011.2437, 2010.