An Online Expectation-Maximization Algorithm for Changepoint Models

Sinan Yıldırım[1, 2], Sumeetpal S. Singh[2], and Arnaud Doucet[3]

[1]Statistical Laboratory, DPMMS, University of Cambridge, UK
[2]Department of Engineering, University of Cambridge, UK
[3]Department of Statistics, University of Oxford, UK

Abstract

Changepoint models are widely used to model the heterogeneity of sequential data. We present a novel sequential Monte Carlo (SMC) online Expectation-Maximization (EM) algorithm for estimating the static parameters of such models. The SMC online EM algorithm has a cost per time which is linear in the number of particles and could be particularly important when the data is representable as a long sequence of observations, since it drastically reduces the computational requirements for implementation. We present an asymptotic analysis for the stability of the SMC estimates used in the online EM algorithm and demonstrate the performance of this scheme using both simulated and real data originating from DNA analysis.

1 Introduction

Consider a sequence of observations $\{y_1, y_2, \ldots\}$ collected sequentially in time. A changepoint model is a particular model for heterogeneity of sequential data that postulates the existence of a strictly increasing time sequence $t_1, t_2, \ldots$ with $t_1 = 1$, that partitions the data into disjoint segments

$$\{y_{t_1}, \ldots, y_{t_2-1}\}, \{y_{t_2}, \ldots, y_{t_3-1}\}, \ldots$$

and that the data is correlated within a segment but are otherwise independent across segments. The time instances $t_1, t_2, \ldots$ are known as the *changepoints* and constitute a random unobserved sequence. This segmental structure is both an intuitive and versatile model for heterogeneity and it is the reason why changepoint models have enjoyed a wide appeal in a variety of disciplines such as Biological Science (Braun and Muller, 1998; Johnson et al., 2003; Fearnhead and Vasileiou, 2009; Caron et al., 2011), Physical Science (Ó Ruanaidh and Fitzgerald, 1996; Lund and Reeves, 2002) Signal Processing (Punskaya et al., 2002; Cemgil et al., 2006), and Finance (Dias and Embrechts, 2004).

In a Bayesian approach to inferring changepoints, one adopts a prior distribution on their locations and a likelihood function for the observed process given these changepoints. However, both of these laws typically depend on a finite dimensional real parameter vector $\theta \in \Theta$ where Θ denotes the set of permissible parameter vectors. In all realistic applications, the static parameter θ is unknown and needs to be estimated from the data as well. A fully Bayesian approach would assign a prior distribution to θ. However the resulting posterior distribution is intractable. Several Markov chain Monte Carlo (MCMC) schemes have been proposed in this context (Stephens, 1994; Chib, 1998; Lavielle and Lebarbier, 2001; Fearnhead, 2006). Unfortunately these algorithms are far too computationally intensive when dealing with very large datasets. Alternative to an MCMC based full Bayesian analysis is sequential Monte Carlo (SMC); however, SMC methods to perform online Bayesian static parameter estimation suffer from the well-known particle path degeneracy problem and can provide unreliable estimates; see Andrieu et al. (2005), Olsson et al. (2008) for a discussion of this issue. This is why we focus here on estimating the parameter θ using a maximum likelihood approach; i.e. the Maximum Likelihood Estimate (MLE) of interest is the parameter vector from Θ that maximizes the probability density of the observed data sequence $p_\theta(y_1, \ldots, y_n)$. This is a challenging problem as computing the likelihood $p_\theta(y_1, \ldots, y_n)$ requires a computational cost increasing super-linearly with n (Chopin, 2007; Fearnhead and Liu, 2007).

Our main contribution is a novel online EM algorithm to compute the MLE of the static parameter θ for changepoint models. We remark that standard batch EM algorithms for a restricted class of changepoint models have been proposed before, e.g. see Gales and Young (1993), Barbu and Limnios (2008), Fearnhead and Vasileiou (2009). The main reason why an online algorithm is desirable is that huge computational and memory savings are possible.

For a long data sequence, a standard EM algorithm requires a complete browse through the entire data set at each iteration to update the MLE of θ; and many such iterations are needed until the estimate of θ converges. This not only requires storing the entire data sequence but also the probability laws that are needed in the intermediate computations done in each EM iteration, which can be impractical. For this reason, there has been a strong interest in online methods which make parameter estimation possible by browsing through the data only once and hence circumventing the need to store it in its entirety (see Kantas et al. (2009) for a review). The only other work on computing the MLE of θ for a more restrictive class of changepoint models in an online manner that we are aware of is Caron et al. (2011), where the authors used a recursive gradient algorithm. If the model permits an EM implementation then it is fair to say that the EM is generally preferred by practitioners as no algorithm tuning is required whereas it can be difficult to properly scale the components of the computed gradient vector.

For finite state-space Hidden Markov Models (HMM) (Mongillo and Deneve, 2008; Cappé, 2011) and linear Gaussian state-space models (Elliott et al., 2002), it is possible to implement exactly the online EM algorithm. A detailed study of this algorithm in the finite state-space case can be found in Cappé (2011). For changepoint models, it is necessary to approximate numerically certain expectations sequentially over time with respect to (w.r.t.) the conditional law of the changepoints and other latent random variables of the model given the available observations up to that point in time. We present SMC estimates of these expectations and establish the stability (via the variance) of these estimates w.r.t. time n and the number of particles N both theoretically and with numerical examples. Stability of the SMC estimates of the expectations is important for assessing the performance and reliability of the EM algorithm and is not to be taken for granted because these expectations are computed w.r.t. a probability law whose dimension increases linearly with time n. We note that the computational cost of the proposed SMC online EM algorithm is $\mathcal{O}(N)$ per-time whereas a $\mathcal{O}(N^2)$ per-time algorithm is required to obtain similar stability results for general state-space HMMs (Del Moral et al., 2009). Cappé (2011), remarked that "although the online EM algorithm resembles a classical stochastic approximation algorithm, it is sufficiently different to resist conventional 'analysis of convergence'. We believe that limited results similar to those discussed in Cappé (2011, Section 4) identifying the potential accumulation points of the online EM procedure could be established but this is beyond the scope of this paper. In

the numerical studies reported in this paper, and indeed in all the ones we have conducted, the SMC online EM algorithm converges, and to a very close vicinity of the correct values when these are known, e.g. in synthetic examples. Moreover, we observed that online EM converged significantly quicker than the batch EM implementation.

The organization of the paper is as follows. In Section 2, we describe a general changepoint model. In Section 3, we present the associated online EM algorithm and its SMC implementation. Theoretical results on the stability of the SMC estimates used in the online EM algorithm are given in Section 4. In Section 5, we demonstrate the performance of the SMC online EM algorithm on both simulated and real data. We finish with a discussion in Section 6 and finally, some detailed model specific derivations as well as mathematical proofs are given in Appendix.

2 The changepoint model

In this paper a changepoint model is defined to be comprised of two discrete-time stochastic processes which are $\{(X_k, Z_k)\}_{k\geq 1}$ and $\{Y_k\}_{k\geq 1}$. $\{(X_k, Z_k)\}_{k\geq 1}$ is an unobserved time-homogeneous Markov chain taking values in $\mathcal{X} \times \mathcal{Z}$ where $\mathcal{X} = \{1, 2, \ldots\} \times \{1, \ldots, R\}$ and $\mathcal{Z} \subseteq \mathbb{R}^p$. (While the definition of $\mathcal{X}$ in this manner is necessary for the resulting model to be a changepoint model, the definition of $\mathcal{Z}$ can change depending on the application domain.) We denote realizations of the first component of this chain by $x_k = (d_k, m_k)$. The variable m_k takes values in the index set $\{1, \ldots, R\}$ and indicates the (generative) model the chain is in at that time while d_k indicates the duration the chain has spent in model m_k. The transition law of $\{(X_k, Z_k)\}_{k\geq 1}$ is

$$X_1 \sim \mu, \qquad X_k \,|(x_{k-1} = (d, m), z_{k-1}) = \begin{cases} (d+1, m) & \text{w.p. } 1 - \lambda_{\theta,m}(d) \\ (1, m') & \text{w.p. } \lambda_{\theta,m}(d) \times P_\theta(m, m') \end{cases},$$

$$Z_k \,|(x_k = (d', m'), x_{k-1}, z_{k-1}) \sim \begin{cases} f_{\theta,m'}(z|z_{k-1})dz & \text{if } d' \neq 1 \\ \pi_{\theta,m'}(z)dz & \text{if } d' = 1 \end{cases}, \tag{1}$$

where $\lambda_{\theta,m}(d) \in [0, 1]$ for all $\theta \in \Theta$ and $(d, m) \in \mathcal{X}$; P_θ is an $R \times R$ row stochastic matrix; for each θ and m, $f_{\theta,m}(z|z_{k-1})$ is the density of a Markov transition kernel on $\mathcal{Z}$ w.r.t. a suitable dominating measure which is denoted by dz; and for each θ and m, $\pi_{\theta,m}$ is a probability density on $\mathcal{Z}$. The transition kernel of the Markov chain $\{(X_k, Z_k)\}_{k\geq 1}$ is assumed to be parametrised by the finite dimensional parameter $\theta \in \Theta$. Without loss of generality, it is

assumed that the probability distribution of the initial state of the chain $\{X_k\}_{k\geq 1}$, denoted μ, has all its mass on $\{(1,1),\ldots,(1,R)\}$, e.g. the uniform distribution on $\{(1,1),\ldots,(1,R)\}$.

For a sequence $\{a_k\}_{k\geq 1}$ and integers i, j, let $a_{i:j}$ denote the set $\{a_i, a_{i+1}, ..., a_j\}$, which is empty if $j < i$, and $a_{i:\infty} = \{a_i, a_{i+1}, ...\}$. The process $\{Y_k\}_{k\geq 1}$ is a $\mathcal{Y}$-valued observed process which satisfies the following conditional independence property:

$$Y_k \left| \left(\{x_k, z_k\}_{k\geq 1}, y_{1:k-1}, y_{k+1:\infty} \right) \sim g_{\theta,m_k}(y|z_k)dy \right. \tag{2}$$

where for each θ and m, $g_{\theta,m}$ is a probability density on $\mathcal{Y}$ with respect to the dominating measure dy. In this work $\mathcal{Y} \subseteq \mathbb{R}^q$ although the definition of $\mathcal{Y}$ may be altered depending on the application. Equations (1) and (2), now define the law of $(X_{1:n}, Z_{1:n}, Y_{1:n})$.

Note that $\{X_k\}_{k\geq 1}$ itself is a Markov chain and we denote its transition matrix by $p_\theta(x_k|\, x_{k-1})$. Secondly, it is useful to visualize a realization of $\{X_k\}_{k\geq 1}$ as a labelled contiguous partition of $\{1,2,\ldots\}$, $\{[t_1,t_2),[t_2,t_3),\ldots\}$ and $t_{i+1} > t_i$, where each set $[t_i, t_{i+1})$ of the partition, which we call a *segment*, is accompanied by m_{t_i}, the model number during that segment. The variables t_i are the instances $\{X_k\}_{k\geq 1}$ visits the set $\{1\} \times \{1,\ldots,R\}$ and are called as the changepoints. As $\{Z_k\}_{k\geq 1}$ forgets its past at times of changepoints, within the segment $[t_i, t_{i+1})$, $\{(Z_k, Y_k)\}_{t_i \leq k < t_{i+1}}$ is a HMM with initial, state transition, and observation densities $\pi_{\theta,m_{t_i}}, f_{\theta,m_{t_i}}$, and $g_{\theta,m_{t_i}}$ respectively. In this sense, our model is general enough to encompass both *hidden semi-Markov models* ((Murphy, 2002; Barbu and Limnios, 2008) and *segmented hidden semi-Markov models* (Gales and Young, 1993; Dong and He, 2007). Below, we give an example of a changepoint model, which we will use in our experiments throughout the paper.

Example 1. *Consider the following changepoint model presented in Fearnhead and Vasileiou (2009), where $Z_k = (Z_{k,1}, Z_{k,2}) \in \mathbb{R} \times \mathbb{R}^+$, and $\mathcal{Y} = \mathbb{R}$. The model satisfies*

$$X_1 \sim \mathcal{U}_{\{1\}\times\{1,\ldots,R\}}, \qquad X_k\,|(x_{k-1} = (d,m)) = \begin{cases} (d+1, m) & w.p.\ \ (1-\lambda_m) \\ (1, m') & w.p.\ \ \lambda_m \times P(m,m') \end{cases},$$

$$Z_k\,|(x_k = (d',m'), z_{k-1}) \sim \begin{cases} \delta_{z_{k-1}} & \text{if } d' \neq 1 \\ \mathcal{N}\Gamma^{-1}(\xi_m, \kappa_m, \alpha, \beta) & \text{if } d' = 1 \end{cases},$$

$$Y_k\,|z_k \sim \mathcal{N}(z_{k,1}, z_{k,2}),$$

where $\mathcal{N}\Gamma^{-1}(\cdot)$ denotes the normal-inverse gamma distribution and $\mathcal{U}_A$ is the uniform distribution over the set A. In relation to (1) *and* (2)*, we have $\lambda_{\theta,m}(d) = \lambda_m$, $f_{\theta,m}(z|z_{k-1})dz =$*

$\delta_{z_{k-1}}(dz)$, $\pi_{\theta,m} = \mathcal{N}\Gamma^{-1}(\xi_m, \kappa_m, \alpha, \beta)$, *and* $g_\theta(y|z_k) = \mathcal{N}(y; z_{k,1}, z_{k,2})$. *Therefore, the parameters of interest are* $\theta = (\xi_{1:R}, \kappa_{1:R}, \lambda_{1:R}, \alpha, \beta, P)$. *In this model, the observations in each segment are i.i.d. Gaussian random variables whose mean and variance change from segment to segment and are drawn from the normal-inverse gamma distribution.*

The following important conditional independence property, which follows from (1) and (2), will be frequently used in the derivations to follow: for any $k' \geq k$,

$$p_\theta(y_k|x_{1:k'}, y_{1:k-1}) = p_\theta(y_k|x_k, y_{1:k-1}) = p_\theta(y_k|x_k, y_{k-d_k+1:k-1}).$$

(Recall that d_k is the first component of x_k.) This equation may be interpreted to mean that y_k only depends statistically on the past observations that are received since the most recent changepoint and not on the observations before that. For the models considered in this work we assume that $p_\theta(y_k|x_k, y_{1:k-1})$ can be evaluated for any x_k and $y_{1:k}$ (whenever the conditional law is well defined). This assumption is satisfied by some important models (e.g. Fearnhead and Vasileiou (2009); Whiteley et al. (2009); Caron et al. (2011)), and allows us to focus inference on $X_{1:n}$ and θ given $Y_{1:n}$ as $Z_{1:n}$ may be integrated out.

For a given realization of observations $\{y_k\}_{k\geq 1}$, we define the potential function $G_{\theta,k} : \mathcal{X} \to [0, \infty)$ as

$$G_{\theta,k}(x_k) = \frac{\int \pi_{\theta,m_k}(z_j) \prod_{i=j+1}^{k} f_{\theta,m_k}(z_i|z_{i-1}) \prod_{i=j}^{k} g_{\theta,m_k}(y_i|z_i) dz_{j:k}}{\int \pi_{\theta,m_k}(z_j) \prod_{i=j+1}^{k-1} f_{\theta,m_k}(z_i|z_{i-1}) \prod_{i=j}^{k-1} g_{\theta,m_k}(y_i|z_i) dz_{j:k-1}}, \quad j = \max(k-d_k+1, 1).$$

($G_{\theta,k}$ is introduced for brevity.) Note that $G_{\theta,k}(x_k)$ is precisely $p_\theta(y_k|x_k, y_{1:k-1})$ at values of x_k where the latter is well defined. We can now express the probability density of the observed process, or likelihood, succinctly as

$$p_\theta(y_{1:n}) = \mathbb{E}_\theta \left[\prod_{k=1}^{n} G_{\theta,k}(X_k) \right].$$

3 EM algorithms for changepoint models

Our main aim is to estimate the static parameter θ of the changepoint model in an online manner using the EM algorithm. We first introduce the batch EM algorithm and then explain how it can be modified to obtain the online EM version.

3.1 Batch EM

Given $Y_{1:n} = y_{1:n}$, the EM algorithm for maximizing $p_\theta(y_{1:n})$ is given by the following iterative procedure: if θ_i is the estimate of the maximizer at the ith iteration, then at iteration $i+1$ we first calculate the following intermediate optimization criterion,

$$\begin{aligned} Q(\theta_i, \theta) &= \mathbb{E}_{\theta_i}\left[\log p_\theta(y_{1:n}, Z_{1:n}, X_{1:n}) \middle| y_{1:n}\right] \\ &= \mathbb{E}_{\theta_i}\left[\log p_\theta(X_{1:n}) + \log p_\theta(y_{1:n}, Z_{1:n}|X_{1:n}) \middle| y_{1:n}\right] \\ &= \mathbb{E}_{\theta_i}\left[\log p_\theta(X_{1:n}) + \mathbb{E}_{\theta_i}\left\{\log p_\theta(y_{1:n}, Z_{1:n}|X_{1:n}) \middle| y_{1:n}, X_{1:n}\right\} \middle| y_{1:n}\right]. \end{aligned} \tag{3}$$

This step is known as the expectation (E) step. The inner expectation in (3) is w.r.t. the law of $Z_{1:n}$ conditioned on $y_{1:n}$ and $X_{1:n}$ under θ_i, that is $p_{\theta_i}\left(z_{1:n} \middle| y_{1:n}, x_{1:n}\right)$, whereas the outer expectation is w.r.t. the law of $X_{1:n}$ conditioned on $y_{1:n}$ under θ_i, that is $p_{\theta_i}\left(x_{1:n} \middle| y_{1:n}\right)$. The updated estimate is then computed in the maximization (or M) step

$$\theta_{i+1} = \arg\max_\theta Q(\theta_i, \theta).$$

This procedure is repeated until θ_i converges (or ceases to change significantly).

Let us define the integrand of the outer expectation in (3) as the function $H_k : \mathcal{X}^k \times \mathcal{Y}^k \times \Theta^2 \to \mathbb{R}$, $k = 1, \ldots, n$,

$$H_k(x_{1:k}, y_{1:k}, \theta_i, \theta) := \log p_\theta(x_{1:k}) + \mathbb{E}_{\theta_i}\left[\log p_\theta(y_{1:k}, Z_{1:k}|x_{1:k}) \middle| y_{1:k}, x_{1:k}\right]$$

We can exploit the following three properties of H_k and $Q(\theta_i, \theta)$. Firstly, H_k has an additive structure (see Appendix A.1 for a derivation):

$$H_k(x_{1:k}, y_{1:k}, \theta_i, \theta) = H_{k-1}(x_{1:k-1}, y_{1:k-1}, \theta_i, \theta) + h_k(x_{k-1}, x_k, y_{k-d_k+1:k}, \theta_i, \theta) \tag{4}$$

where the incremental term h_k is a function of $(x_{k-1}, x_k, y_{k-d_k+1}, \ldots, y_k, \theta_i, \theta)$. Secondly, when the transition laws of the changepoint model given in (1)-(2) belong to the exponential family then the incremental terms can be expressed as

$$h_k(x_{k-1}, x_k, y_{k-d_k+1:k}, \theta_i, \theta) = v_\theta^{\mathrm{T}} s_k(x_{k-1}, x_k, y_{k-d_k+1:k}, \theta_i) \tag{5}$$

where v_θ is a $r \times 1$ vector depending only on θ, s_k is a $r \times 1$ vector valued function of $(x_{k-1}, x_k, y_{k-d_k+1}, \ldots, y_k, \theta_i)$. (From now on, we omit the dependency of H_k, h_k, and s_k on $y_{1:k}$ for the sake of conciseness.) Thirdly, $Q(\theta_i, \theta) = v_\theta^{\mathrm{T}} \mathbb{E}_{\theta_i}\left[S_n(X_{1:n}, \theta_i) \middle| y_{1:n}\right]$ where

$$S_n(x_{1:n}, \theta_i) = \sum_{j=1}^{n} s_j(x_{j-1}, x_j, \theta_i), \tag{6}$$

with $s_1(x_0, x_1, \theta) = s_1(x_1, \theta)$ by convention, and its maximizer is explicitly characterized by a function $\Lambda : \mathbb{R}^r \to \Theta$

$$\arg\max_{\theta \in \Theta} Q(\theta_i, \theta) = \Lambda\left(\mathbb{E}_{\theta_i}\left[S_n(X_{1:n}, \theta_i) \middle| y_{1:n}\right]\right). \tag{7}$$

Hence from a practical point of view, it is necessary to compute the expectation of additive functionals (6) w.r.t. $p_{\theta_i}(x_{1:n}|\, y_{1:n})$. As for a standard HMM, this can be achieved using a forward-backward type algorithm; see Gales and Young (1993), Barbu and Limnios (2008), Fearnhead and Vasileiou (2009). However in a general scenario the computational complexity is quadratic in n and approximations are necessary when n is very large. In Fearnhead and Vasileiou (2009) a Monte Carlo EM (MCEM) algorithm was proposed for a specific changepoint model (see Section 5) where the expectations in the E-step are computed using a backward Monte Carlo sampling procedure.

3.2 Online EM

The development of an online version of the EM rests on the following key fact (Del Moral et al., 2009; Cappé, 2011). The quantity $\mathbb{E}_\theta\left[S_n(X_{1:n}, \theta) \middle| y_{1:n}\right]$ when S_n has the additive structure in (6) can be evaluated sequentially with the following recursion which we will refer to as the *forward smoothing recursion*:

$$\begin{aligned} T_n(x_n, \theta) &:= \sum_{x_{1:n-1} \in \mathcal{X}^{n-1}} S_n(x_{1:n}, \theta) p_\theta(x_{1:n-1} | y_{1:n-1}, x_n) \\ &= \sum_{x_{n-1} \in \mathcal{X}} \left[T_{n-1}(x_{n-1}, \theta) + s_n(x_{n-1}, x_n, \theta)\right] p_\theta(x_{n-1} | y_{1:n-1}, x_n) \end{aligned}$$

with $T_1(x_1, \theta) = s_1(x_1, \theta)$. The second line follows from (6) and the decomposition

$$p_\theta(x_{1:n-1} | y_{1:n-1}, x_n) = p_\theta(x_{1:n-2} | y_{1:n-2}, x_{n-1}) p_\theta(x_{n-1} | y_{1:n-1}, x_n) \tag{8}$$

due to the fact that given x_{n-1}, $x_{1:n-2}$ do not depend on $x_n, x_{n+1}, \ldots, y_{n-1}, y_n, \ldots$, which follows from (1) and (2). The function $T_n(\cdot, \theta) : \mathcal{X} \to \mathbb{R}^r$ can be computed in an online manner and hence so can

$$\mathbb{E}_\theta\left[S_n(X_{1:n}, \theta) \middle| y_{1:n}\right] = \sum_{x_n \in \mathcal{X}} T_n(x_n, \theta) p_\theta(x_n | y_{1:n}).$$

It is possible to use this recursion to implement the batch EM algorithm. Compared to the standard forward-backward implementation, this approach does not require a backward pass to compute the expectations of interest and hence requires far less memory to implement.

The online EM algorithm is a variation over the batch EM where the parameter is re-estimated each time a new observation is collected. In this approach running averages of $\mathbb{E}_\theta\left[S_n(X_{1:n},\theta)\middle|\, y_{1:n}\right]$ are computed (Elliott et al., 2002; Mongillo and Deneve, 2008; Cappé, 2009, 2011), (Kantas et al., 2009, Section 3.2.). Let $\gamma = \{\gamma_n\}_{n\geq 1}$, called the step-size sequence, be a positive decreasing sequence satisfying $\sum_{n\geq 1}\gamma_n = \infty$ and $\sum_{n\geq 1}\gamma_n^2 < \infty$. A common choice is $\gamma_n = n^{-a}$ for $0.5 < a \leq 1$. Let θ_1 be the initial guess of θ^* before having made any observations and let $\theta_{1:n}$ be the sequence of parameter estimates of the online EM algorithm computed sequentially based on $y_{1:n-1}$. When y_n is received, online EM computes

$$T_{\gamma,n}(x_n) = \sum_{x_{n-1}\in\mathcal{X}} \left[(1-\gamma_n)\, T_{\gamma,n-1}(x_{n-1}) + \gamma_n s_n(x_{n-1},x_n,\theta_n)\right] p_{\theta_{1:n}}(x_{n-1}|y_{1:n-1},x_n), \tag{9}$$

$$\mathcal{S}_n = \sum_{x_n\in\mathcal{X}} T_{\gamma,n}(x_n) p_{\theta_{1:n}}(x_n|y_{1:n}) \tag{10}$$

and then sets $\theta_{n+1} = \Lambda\left(\mathcal{S}_n\right)$. The subscript $\theta_{1:n}$ on $p_{\theta_{1:n}}(x_{n-1}|y_{1:n-1},x_n)$ and $p_{\theta_{1:n}}(x_n|y_{1:n})$ indicates that these laws are being computed sequentially using the parameter θ_k at time k, $k \leq n$. (See Algorithm 1 for details.) In practice, the maximization step is not executed until a burn-in time n_b for added stability of the estimators as discussed in Cappé (2009).

The online EM algorithm can be implemented exactly for a linear Gaussian state-space model (Elliott et al., 2002) and for finite state-space HMM's. (Mongillo and Deneve, 2008; Cappé, 2011). An exact implementation is not possible for changepoint models in general, therefore we now investigate SMC implementations of the online EM algorithm.

3.3 SMC implementations of the online EM algorithm

Let $\mathbb{Q}_{\theta,n}(x_{1:n}) = p_\theta\left(x_{1:n}|y_{1:n-1}\right)$ denote the law of $X_{1:n}$ conditioned on the sequence of observed variables $y_{1:n-1}$, and let $\eta_{\theta,n}(x_n) = p_\theta(x_n|y_{1:n-1})$ denote the time n marginal of $\mathbb{Q}_{\theta,n}$. $\eta_{\theta,n}$ is also known as the predicted filter but we refer to it simply as the filter. In order to

execute (9) and (10) at time n, we need to calculate the following probability distributions:

$$p_\theta(x_{n-1}|x_n, y_{1:n-1}) = \frac{\eta_{\theta,n-1}(x_{n-1})G_{\theta,n-1}(x_{n-1})p_\theta(x_n|x_{n-1})}{\sum_{x'_{n-1}} \eta_{\theta,n-1}(x'_{n-1})G_{\theta,n-1}(x'_{n-1})p_\theta(x_n|x'_{n-1})} \tag{11}$$

$$p_\theta(x_n|y_{1:n}) = \frac{\eta_{\theta,n}(x_n)G_{\theta,n}(x_n)}{\sum_{x'_n} \eta_{\theta,n}(x'_n)G_{\theta,n}(x'_n)} \tag{12}$$

Note that to calculate these probability distributions we only need $\eta_{\theta,n-1}$ and $\eta_{\theta,n}$ at time n. Besides, $\eta_{\theta,n}$ may be computed recursively using Bayes' formula:

$$\eta_{\theta,n}(x_n) = \frac{\sum_{x_{n-1}} \eta_{\theta,n-1}(x_{n-1})\, G_{\theta,n-1}(x_{n-1})\, p_\theta(x_n|\, x_{n-1})}{\sum_{x_{n-1}} \eta_{\theta,n-1}(x_{n-1})\, G_{\theta,n-1}(x_{n-1})}, \quad n > 1, \tag{13}$$

However, the computational cost of the filtering recursion in (13) at time n is $\mathcal{O}(nR)$; this follows since $p_\theta(x'|x)$ is non-zero for at most $R+1$ values of x'. For the analysis of large amounts of data, exact filtering is computationally infeasible and SMC methods have been introduced as a viable alternative (Chopin, 2007; Fearnhead and Liu, 2007).

One way to obtain the SMC approximation to $\eta_{\theta,n}$ is via the *path space* particle approximation of $\mathbb{Q}_{\theta,n}$. This is the empirical measure corresponding to a set of $N \geq 1$ random samples termed particles (Del Moral, 2004):

$$\mathbb{Q}^{\mathrm{p},N}_{\theta,n}(x_{1:n}) = \frac{1}{N}\sum_{i=1}^{N} \delta_{X^{(i)}_{1:n}}(x_{1:n})\,. \tag{14}$$

where $\delta_a(\cdot)$ is the probability mass function concentrated at a. These particles are then propagated in time using importance sampling and resampling steps; see Doucet et al. (2001) and Cappé et al. (2005) for a review of the literature. Specifically, $\mathbb{Q}^{\mathrm{p},N}_{\theta,n}$ is the empirical measure constructed from N independent samples from

$$\frac{\mathbb{Q}^{\mathrm{p},N}_{\theta,n-1}(x_{1:n-1})\, G_{\theta,n-1}(x_{n-1})\, p_\theta(x_n|\, x_{n-1})}{\sum_{x_{1:n-1}} \mathbb{Q}^{\mathrm{p},N}_{\theta,n-1}(x_{1:n-1})\, G_{\theta,n-1}(x_{n-1})}. \tag{15}$$

The particle approximation of $\eta_{\theta,n}$ can now be obtained from $\mathbb{Q}^{\mathrm{p},N}_{\theta,n}$ by marginalization

$$\eta^N_{\theta,n}(x_n) = \frac{1}{N}\sum_{i=1}^{N} \delta_{X^{(i)}_n}(x_n)\,. \tag{16}$$

Other than the one in (16), there are other ways to sequentially update $\eta^N_{\theta,n-1}$ so that $\eta_{\theta,n}$ is approximated at N distinct particles. Given $\eta^N_{\theta,n-1}$, at time n the distribution

$$\frac{\sum_{x_{n-1}} \eta^N_{\theta,n-1}(x_{n-1})\, G_{\theta,n-1}(x_{n-1})\, p_\theta(x_n|\, x_{n-1})}{\sum_{x_{n-1}} \eta^N_{\theta,n-1}(x_{n-1})\, G_{\theta,n-1}(x_{n-1})}$$

with support at $N + R$ points is calculated exactly and then $\eta^N_{\theta,n}$ is obtained by sampling this distribution independently N times (see Algorithm 1). Caron et al. (2011) propose truncating to the N support points with the highest weights. Fearnhead and Liu (2007) propose an unbiased resampling scheme that retains the maximum number of unique parti cles in the reduced representation of size N. In the same work, and in Fearnhead and Vasileiou (2009), resampling schemes that allow changing number of particles in time are proposed.

The online EM algorithm in Section 3.2 can be approximated with $\mathcal{O}(N)$ cost per time using the SMC approximation of the densities in (11) and (12). The resulting algorithm, presented as Algorithm 1, will be referred to as the *SMC-FS online EM* algorithm.

Algorithm 1. ***SMC-FS online EM algorithm for changepoint models***

- ***E-step:*** *If $n = 1$, initialize θ_1; sample $\widetilde{X}_1^{(i)} \sim \mu$, set $\widetilde{T}_1^{(i)} = s_1(\widetilde{X}_1^{(i)}, \theta_1)$, $i = 1, \ldots, N$.*

  If $n \geq 2$

  - *For $i = 1, \ldots, N$, set $\widetilde{X}_n^{(i)} = (d_{n-1}^{(i)} + 1, m_{n-1}^{(i)})$, where $X_{n-1}^{(i)} = (d_{n-1}^{(i)}, m_{n-1}^{(i)})$*
  - *For $m = 1, \ldots, R$, set $\widetilde{X}_n^{(N+m)} = (1, m)$.*
  - *For $i = 1, \ldots, N + R$, compute $\widetilde{W}_n^{(i)} = \sum_{j=1}^N G_{\theta_{n-1},n}(X_{n-1}^{(j)}) p_{\theta_n}(\widetilde{X}_n^{(i)} | X_{n-1}^{(j)})$ and*

  $$\widetilde{T}_n^{(i)} = \frac{1}{\widetilde{W}_n^{(i)}} \sum_{j=1}^{N} G_{\theta_{n-1},k}(X_{n-1}^{(j)}) p_{\theta_n}(\widetilde{X}_n^{(i)} | X_{n-1}^{(j)}) \left[(1 - \gamma_n) T_{n-1}^{(j)} + \gamma_n s_n(X_{n-1}^{(j)}, \widetilde{X}_n^{(i)}, \theta_n) \right]$$

  Resample $\{\widetilde{X}_n^{(i)}, \widetilde{T}_n^{(i)}\}_{i=1,\ldots,N+R}$ according to the weights $\{\widetilde{W}_n^{(i)}\}_{i=1,\ldots,N+R}$ to get resampled particles $\{X_n^{(i)}, T_n^{(i)}\}_{i=1,\ldots,N}$ each with weight $1/N$.

- ***M-step:*** *If $n < n_b$, set $\theta_{n+1} = \theta_n$ else, calculate using the particles before resampling*

  $$\mathcal{S}_n = \frac{\sum_{i=1}^{N+R} \widetilde{T}_n^{(i)} \widetilde{W}_n^{(i)} G_{\theta_n,n}(\widetilde{X}_n^{(i)})}{\sum_{i=1}^{N+R} \widetilde{W}_n^{(i)} G_{\theta_n,n}(\widetilde{X}_n^{(i)})},$$

  update the parameter $\theta_{n+1} = \Lambda(\mathcal{S}_n)$.

3.4 Comparison with the path space online EM

As shown in Section 3.1, the EM algorithm requires certain expectations w.r.t. the measure $\mathbb{Q}_{\theta,n}$, and the online EM algorithm in Section 3.2 relies on the running averages of these

expectations. Consider the following backward representation of $\mathbb{Q}_{\theta,n}$

$$\mathbb{Q}_{\theta,n}(x_{1:n}) = \eta_{\theta,n}(x_n) \prod_{k=n}^{2} p_\theta(x_{k-1}|x_k, y_{1:k-1}).$$

Then a corresponding particle approximation, different from the path-space one, is given by

$$\mathbb{Q}^N_{\theta,n}(x_{1:n}) = \eta^N_{\theta,n}(x_n) \prod_{k=n}^{2} p^N_\theta(x_{k-1}|x_k, y_{1:k-1}). \tag{17}$$

where $p^N_\theta(x_{k-1}|x_k, y_{1:k-1})$ is (11) with $\eta_{\theta,k-1}$ replaced with $\eta^N_{\theta,k-1}$. One can then show that the online EM algorithm using the SMC approximation to the forward smoothing recursion relies on the particle approximation $\mathbb{Q}^N_{\theta,n}$ described above. More precisely, in Algorithm 1, if $\gamma_i = 1/i$, $n < n_b$ (see the M-step), $\theta_1 = \cdots = \theta_{n+1} = \theta$, and $s_{n+1}(x_n, x_{n+1}, \theta) = 0$, then

$$\mathcal{S}_{n+1} = \mathbb{Q}^N_{\theta,n+1}((n+1)^{-1} S_n).$$

This observation will be useful for analyzing the stability properties of the sufficient statistics calculated SMC-FS online EM algorithm in Section 4.

As an alternative to SMC-FS online EM, we could have proposed an SMC online EM algorithm relying on the particle approximation $\mathbb{Q}^{\mathrm{p},N}_{\theta,n}$ defined in (14)-(15). In that case (using the short-hand notation in Algorithm 1) the approximation to (9) and (10) becomes

$$\widetilde{T}^{(i)}_n = (1-\gamma_n) T^{(i)}_{n-1} + \gamma_n s_n(X^{(i)}_{n-1}, \widetilde{X}^{(i)}_n, \theta_n)$$

for each $i = 1, \ldots, N$, and then calculating the estimates of sufficient statistics as

$$\mathcal{S}_n = \frac{\sum_{i=1}^N \widetilde{T}^{(i)}_n G_{\theta_n,n}(\widetilde{X}^{(i)}_n)}{\sum_{i=1}^N G_{\theta_n,n}(\widetilde{X}^{(i)}_n)}.$$

Recall that each $\widetilde{X}^{(i)}_n$ is sampled from $p_{\theta_n}(x_n|X^{(i)}_{n-1})$. $\{\widetilde{X}^{(i)}_n, \widetilde{T}^{(i)}_n\}_{i=1,\ldots,N}$ are then resampled to obtain $\{X^{(i)}_n, T^{(i)}_n\}_{i=1,\ldots,N}$ according to the weights $\{G_{\theta_n,n}(\widetilde{X}^{(i)}_n)\}_{i=1,\ldots,N}$. Based on the path space approximation, we will hereafter call this algorithm the *SMC-PS online EM* algorithm. In the context of general state-space HMM, this was proposed in Cappé (2009) and only requires $\mathcal{O}(N)$ computations per time step. However, it is a well-known fact that $\mathbb{Q}^{\mathrm{p},N}_{\theta,n}$ becomes progressively impoverished as n increases because of the successive resampling steps (Del Moral and Doucet, 2003; Olsson et al., 2008). That is, the number of distinct particles representing the marginal $\mathbb{Q}^{\mathrm{p},N}_{\theta,n}(x_{1:k})$ for any fixed $k < n$ diminishes as n increases

until it eventually collapses to a single particle – this is known as the *particle path degeneracy* problem. Whereas, in the backward particle approximation $\mathbb{Q}_{\theta,n}^N$, we do not have this problem since it relies on the SMC approximations to the filters $\eta_{\theta,n}$ only. Therefore, we expect that the resulting SMC estimates in the SMC-PS online EM algorithm have higher variances than those in the SMC-FS online EM algorithm (Del Moral et al., 2009). For a numerical illustration of this fact, see Section 5.

4 Theoretical results

Recall that the M-step of the exact online EM algorithm applies a mapping Λ which maps expectations of sufficient statistics $\mathbb{Q}_{\theta,n+1}(n^{-1}S_n) = \mathbb{E}_\theta\left[n^{-1}S_n(X_{1:n})|y_{1:n}\right]$ to a parameter estimate in Θ; see (9) and (10) with $\gamma_n = n^{-1}$. It follows from the discussion in Section 3.4 that the reliability of the SMC online EM algorithm described in Section 3.2 depends on how stable the estimates of expectations of the type $\mathbb{Q}_{\theta,n}^N(S_n)$ are. One convenient way of assessing the stability is to check how the asymptotic (in particle number) variance of $\sqrt{N}\left(\mathbb{Q}_{\theta,n}^N - \mathbb{Q}_{\theta,n}\right)(S_n)$ changes with time n. The asymptotic analysis will give us an idea about what will happen when we use a large number of particles. We would like the order of the variance to grow less than quadratically in time n; since then the variance of $\sqrt{N}\left(\mathbb{Q}_{\theta,n}^N - \mathbb{Q}_{\theta,n}\right)(n^{-1}S_n)$, which is the variance of the estimates in the M-step, is not only time uniformly bounded but also vanishes. This should result in the variability of the EM's parameter update step to particle realization also diminishing over time. Before proceeding further we shall make clear that our analysis is for the approximation $\mathbb{Q}_{\theta,n}^N$ defined in (17) for a fixed θ. That is, our results are *only* indicative of the stability of the sufficient statistics calculated in the SMC-FS online EM algorithm, which actually uses a changing sequence of θ's. In summary, our main result in this section establishes that (under certain assumptions) the asymptotic (in particle number) variance of $\sqrt{N}\left(\mathbb{Q}_{\theta,n}^N - \mathbb{Q}_{\theta,n}\right)(S_n)$ is upper bounded by a term $\mathcal{O}(n)$ or $\mathcal{O}(n\log^2 n)$. The tighter $\mathcal{O}(n)$ bound is for finite duration models while the looser $\mathcal{O}(n\log^2 n)$ bound is for infinite duration models.

The results in this section are phrased for any fixed θ and any sequence of observations $y = \{y_n\}_{n\geq 1}$. Also, to keep the notation "light" θ is omitted from the subscripts. Some basic definitions are provided first. For a real valued functions $\varphi : \mathcal{X} \to \mathbb{R}$, let $\|\varphi\|_A = \sup_{x\in A}|\varphi(x)|$ for $A \subseteq \mathcal{X}$. Let $\mathcal{B}(\mathcal{X})$ denote the space of bounded real valued functions

on $\mathcal{X}$. For a probability measure ν on $\mathcal{X}$, let $\nu(\varphi) = \sum_{x\in\mathcal{X}} \nu(x)\ \varphi(x)$, and for $A \in \mathcal{X}$, $\nu(A) = \nu(I_A)$ where I_A is the indicator function for the set A such that $I_A(x) = 1$ if $x \in A$, 0 otherwise. Denote the support of ν by $\mathrm{supp}(\nu) = \{x \in \mathcal{X} : \nu(x) > 0\}$. If $M(x, x')$ is a transition probability (from x to x') on $\mathcal{X}$, let $(M\varphi)(x) = M(\varphi)(x) = \sum_{x'} M(x, x')\varphi(x')$. For $\varphi \in \mathcal{B}(\mathcal{X})$ and $A \subseteq \mathcal{X}$, let $\mathrm{osc}_A(\varphi) = \sup_{x,x'\in A} |\varphi(x) - \varphi(x')|$ be the oscillation of the function over A and $\mathrm{osc}(\varphi) = \mathrm{osc}_{\mathcal{X}}(\varphi)$. The complement of a set A is $\overline{A}$.

We will require the following result concerning the asymptotic variance of particle smoothing (Del Moral et al., 2010).

Theorem 1. *Given $y = \{y_n\}_{n\geq 1}$, assume there exists finite constants c_n such that $c_n^{-1} \leq G_n \leq c_n$ for all n. For any $n \geq 1$, $F_n \in \mathcal{B}(\mathcal{X}^n)$, $\sqrt{N}\left(\mathbb{Q}_n^N - \mathbb{Q}_n\right)(F_n)$ converges in law, as $N \to \infty$, to a centered Gaussian random variable with variance*

$$\sum_{i=1}^{n} \eta_i \left([G_{i,n}\ D_{i,n}(F_n - \mathbb{Q}_n(F_n))]^2 \right). \tag{18}$$

where, for $1 \leq i \leq n$, the potential function $G_{i,n}$ and the bounded integral operator $D_{i,n}$ are

$$G_{i,n}(x_i) := \frac{p(y_{i:n-1}|x_i, y_{1:i-1})}{p(y_{i:n-1}|y_{1:i-1})}, \quad D_{i,n}(F_n)(x_i) := \mathbb{E}\left[F_n(X_{1:n}) | \, y_{1:n-1}, x_i \right].$$

The assumption that the potentials G_n are uniformly bounded below by c_n^{-1} is not overly restrictive as it is satisfied when $g_m(y|z) > 0$ for all m, y and z. The latter is a typical assumption in the context of the analysis of particle filters to avoid the possibility of all the particles having weight zero (Del Moral, 2004).

In order to discuss the rate of growth of the asymptotic variance (18) as a function of time n, we need to quantify the sensitivity of the forward and backward smoothers to their initializations. For a given sequence of observations $y_{1:n}$, the *forward smoother* is defined as the Markov chain on $\mathcal{X}$ with transition kernel $p(x_{k+1}|x_k, y_{1:n})$, $k = 1, \ldots, n-1$. Similarly, the *backward smoother* is the reverse time Markov chain with transition kernel $p(x_k|x_{k+1}, y_{1:n})$, $k = n-1, n-2, \ldots, 1$. Each term of the sum in (18) is an integral over $\mathcal{X}^n$ and will typically grow linearly with n unless both the forward and backward smoother forget their initializations quick enough (e.g. with geometric rate) and the class of functions F_n is restricted. Indeed the E-step of the EM algorithm computes the expectation for not an arbitrary F_n but one that has a specific additive structure; see Section 3.1, also Proposition 1. A definition of geometric rate is as follows. Given $\{y_i\}_{i\geq 1}$, if for some integer $L > 0$ there

exists a finite constant $c(L) \geq 1$ such that for all $m - k \geq L$, $n \geq m$,

$$|\mathbb{E}\left[s(X_m)|\, x_k, y_{1:n}\right] - \mathbb{E}\left[s(X_m)|\, x'_k, y_{1:n}\right]| \leq \mathrm{osc}(s)(1 - c(L)^{-2})^{\left\lfloor \frac{m-k}{L} \right\rfloor} \tag{19}$$

irrespective of (x_k, x'_k) provided both conditional expectations are well defined, then the forward smoother is said to forget its initialization with geometric rate. (A similar definition applies for the backward smoother; see (32)). Henceforth, when we say *forward forgetting* we mean that the forward smoother forgets its initial condition in the sense of (19) but without any specific reference to a rate. By backward forgetting, similarly, we will mean the insensitivity of the backward smoother to its initialization.

A typical route to establish forward and backward forgetting is to exploit the fact that the Markov chain $\{X_k\}_{k\geq 1}$ satisfies a majorization and minorization condition: that is there exists a probability measure $m(x)$, positive integer l and positive constant c such that $c^{-1}m(x_k) \leq p(x_k|x_{k-l}) \leq c\, m(x_k)$ for all $(x_{k-l}, x_k) \in \mathcal{X}^2$. When this condition is satisfied it may be shown that the backward and forward smoothers forget their initializations at geometric rate, which is quick enough such each term of the sum (18) is uniformly bounded over time. For changepoint models however, the majorization-minorization condition is not satisfied in general. Consider the following example: let $R = 1$ (in which case we drop the variable m_k from x_k, i.e. $x_k = d_k$) and

$$X_k = \begin{cases} x_{k-1} + 1 & \text{w.p.} \quad 1 - \lambda \\ 1 & \text{w.p.} \quad \lambda \end{cases} \tag{20}$$

Furthermore, given $X_k = d$ then it must be that $X_{k-i} = d - i$ for $i < d$. Thus the distance between the probability distributions $\Pr(X_{k-i}|X_k = d)$ and $\Pr(X_{k-i}|X_k = d')$ will not decrease at geometric rate and the same cannot be expected for the backward smoother (which is essentially these laws but with additional conditioning on $y_{1:k-1}$.) In this paper, we analyze the asymptotic variance for changepoint models using a slightly refined approach.

We analyze two types of changepoint models separately, namely *finite duration* changepoint models and *infinite duration* changepoint models. We distinguish between the two models as follows. In a finite duration changepoint model, for each $m \in \{1, \ldots, R\}$ there exists some finite $\bar{d}_m$ such that $\lambda_m(d) = 1$ for all $d \geq \bar{d}_m$, and smallest such $\bar{d}_m$ is the maximum duration length for model m. If, for at least one $m \in \{1, \ldots, R\}$, $\lambda_m(d) < 1$ for all $d > 0$, then the model is called an infinite duration model.

Given $\{y_n\}_{n\geq 1}$, for positive integers $k \geq 1$, (lag) l and set $A \subseteq \mathcal{X}$, let

$$c_{k,l}(A) = \sup_{\substack{x_{k+l}\in A,\\ x_k, x'_k \in \mathrm{supp}(\eta_k)}} \frac{p(x_{k+l}, y_{k:k+l-1}|x_k, y_{1:k-1})}{p(x_{k+l}, y_{k:k+l-1}|x'_k, y_{1:k-1})} \tag{21}$$

where $c_{k,l}$ is taken to be infinity if the denominator can be made zero while the numerator is not. By convention $0/0 = 1$. The variables x_k and x'_k range over $\mathrm{supp}(\eta_k)$ to ensure the conditional expectations in the numerator and denominator are well defined. Also, we abbreviate $c_{k,l}(\mathcal{X})$ to $c_{k,l}$. The variance result is now stated for additive functions of the form $S_k(x_{1:k}) = \sum_{i=1}^k s_i(x_i)$ and may be extended to the case where $S_k(x_{1:k}) = \sum_{i=1}^k s_i(x_{i-1}, x_i)$. The proof of the result is based on some supporting results and is given in Appendix A.3.

Proposition 1. *Assume $S_n(x_{1:n}) = \sum_{k=1}^n s_k(x_k)$ where $\mathrm{osc}(s_k) \leq 1$.*

- *If $\{X_k\}$ is a finite duration changepoint model which is irreducible and aperiodic; and there exists a finite constant c such that $c^{-1} \leq G_n \leq c$ for all n, then the asymptotic variance of $\sqrt{N}\left(\mathbb{Q}_n^N - \mathbb{Q}_n\right)(S_n)$ given in (18) is upper bounded by a term $\mathcal{O}(n)$.*

- *Assume $\{X_k\}$ is an infinite duration changepoint model whose forward smoother forgets its initialization at geometric rate in the sense of (19). Furthermore, let $A = \{1,\ldots,L\}\times\{1,\ldots,R\}$. If there exist a finite positive constant c such that $c^{-1} \leq G_n \leq c$ for all n and finite positive constants C, $\gamma \in (0,1)$ and c' such that for all n and L*

$$\sup_{i\geq 1} \eta_i(\overline{A}) \leq C\gamma^L, \qquad \text{and} \qquad \sup_{i\geq 1} c_{i,L}(A) \leq c', \tag{22}$$

then the asymptotic variance of $\sqrt{N}\left(\mathbb{Q}_n^N - \mathbb{Q}_n\right)(S_n)$ is upper bounded by $\mathcal{O}(n\log^2 n)$.

The first condition in (22) is a uniform tightness condition on the probabilities η_i, whereas the second condition means that if a changepoint occurs between times k and $k+L$, the observations up to the last changepoint prior to time $k+L$ do not favor one x_k over another too much. Proposition 1 is now shown to be applicable to the infinite duration model in (20) with the following example whose verification is shown in Appendix A.3.1.

Example 2. *For the infinite duration model in (20), recall that Z_k (see Section 2) is a Markov process that "resets" itself when X_k returns to state 1, i.e.*

$$Z_k\,|(x_{1:k}, z_{1:k-1}) \sim \begin{cases} \pi(z_k)dz_k & \text{if } x_k = 1 \\ f(z_k|z_{k-1})dz_k & \text{otherwise} \end{cases}$$

We will assume that the process $\{Z_k\}_{k\geq 1}$ assumes values from a compact space and that there exists some positive constant c such that for all (z_{k-1}, z_k)

$$c^{-1/2} \leq \pi(z_k) \leq c^{1/2}, \qquad c^{-1/2} \leq f(z_k|z_{k-1}) \leq c^{1/2}. \tag{23}$$

Furthermore, assume $g(y_k|z_k) > 0$ for all z_k, y_k. For example, a changepoint model satisfying these assumptions could be the changepoint model in Example 1 in Section 2 with $R = 1$ and instead of a static $\{Z_k\}_{k\geq 1}$ process, a slowly moving one which is "mixing". Note that a slowly moving $\{Z_k\}_{k\geq 1}$ process permits a more parsimonious representation of the data.

5 Numerical examples

5.1 Simulated experiments

For the experiments in this section, we will use the infinite duration changepoint model in Example 1 in Section 2, where $\theta = (\xi_{1:R}, \kappa_{1:R}, \lambda_{1:R}, \alpha, \beta, P)$. The constituent distributions of this model belong to the exponential family and so (5) holds; see Appendix A.2 for details.

5.1.1 Online EM applied to long data sequence

We applied Algorithm 1 to a data sequence of length 500000 generated by the model in Example 1 with $R = 2$ and parameter values $\alpha = 10, \beta = 0.1, \xi_1 = 1.445, \xi_2 = -0.214, \kappa_1 = 1.588, \kappa_2 = 0.379, \lambda_1 = 0.12, \lambda_2 = 0.09, P_{ij} = 0.5, i, j = 1, 2$. The M-step was not executed for the first 2000 points (i.e. $n_b = 2000$). The step-size sequence was $\gamma_n = n^{-0.8}$. Figure 1 shows the trace parameter estimates over time. We observe that the algorithm converges towards the true values. We also did multiple runs to check that the algorithm would not only converge to a local maximum.

5.1.2 Comparison between online and batch EM for a short data sequence

Figure 1 also suggests that online EM requires a long data sequence for convergence. Therefore, for short data sequences the algorithm may not converge and its potential use is questionable. One can of course use the batch EM algorithm in such cases but another solution might be to apply online EM to the concatenated sequence $\{y_{1:K}, y_{1:K}, \ldots\}$. By doing so, the online EM solution is not 'online' anymore. However, it can still be significantly faster than

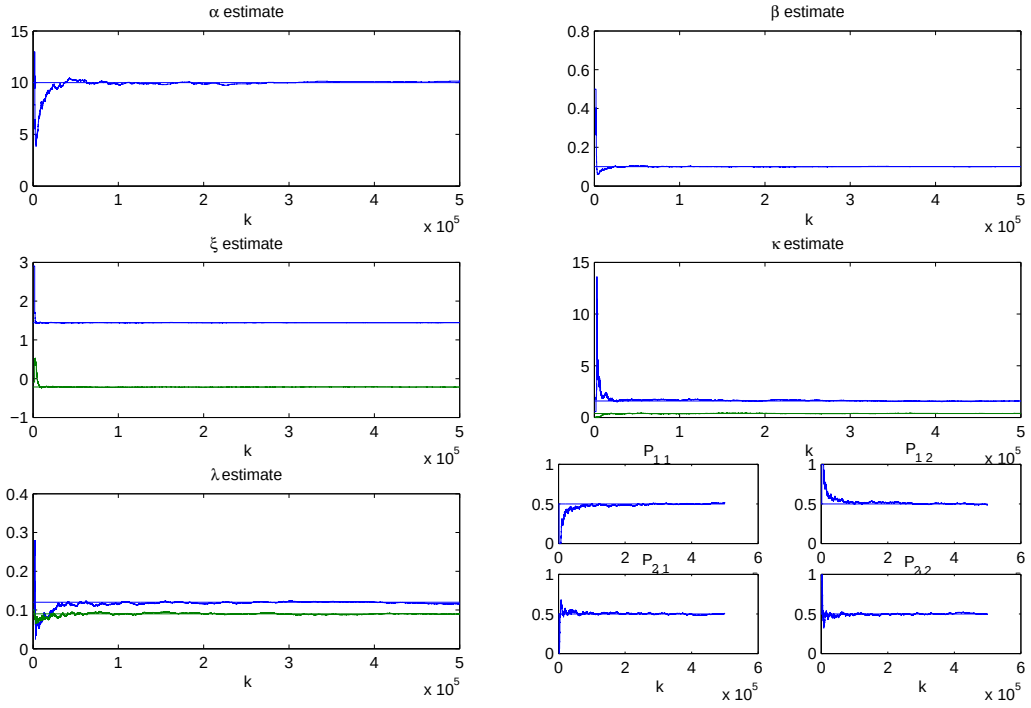

Figure 1: SMC-FS online EM estimates vs time for a long simulated data sequence. The true parameter values are indicated with a horizontal line.

the offline version as we demonstrate below. Figures 2 and 3 show results for such a scenario for 2000 data points. We used Algorithm 1 to obtain the results in Figure 2 by replicating $y_{1:K}$ 100 times and the SMC-FS batch EM algorithm (the batch version of SMC-FS online EM) to obtain the results in Figure 3. The true parameter values are $\alpha = 10$, $\beta = 0.1$, $\xi_1 = 1.78, \xi_2 = 3.56, \kappa_1 = 0.30, \kappa_2 = 0.03, \lambda_1 = \lambda_2 = 0.1, P_{i,j} = 0.5, i, j = 1, 2$.

There are two main outcomes to be stressed from the results in Figures 2 and 3. First, the online EM algorithm in this example is much faster since it converges after around 50 passes, whereas the batch EM algorithm needs over 1000 iterations for convergence. Notice that the computational cost of one pass over the data in the online case and one iteration in the batch case are almost the same and therefore the comparison makes sense. Second, the parameter estimates of both algorithms converge to almost the same points. This empirically validates the potential benefit of the online EM algorithm even in the offline setting.

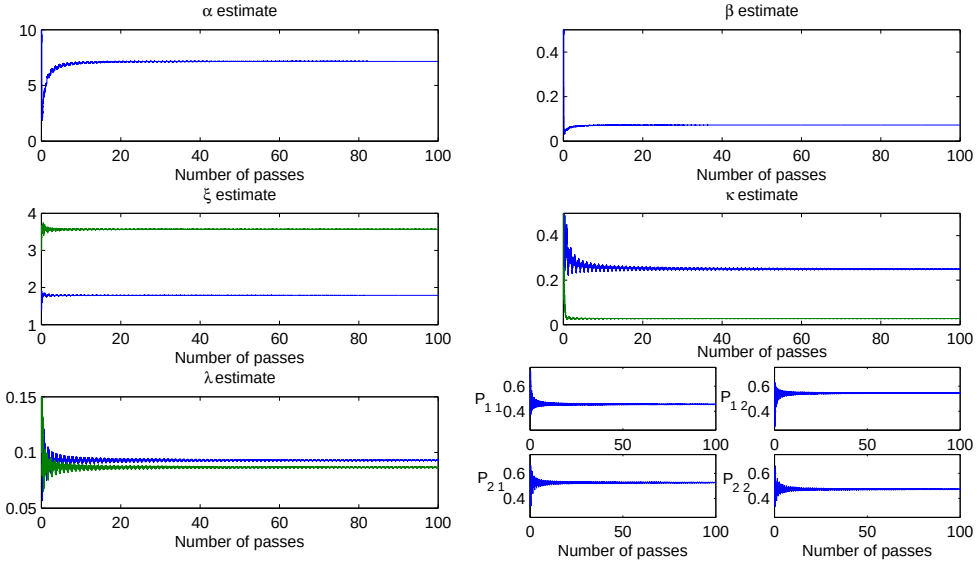

Figure 2: SMC-FS online EM estimates vs number of passes for the concatenated data set $\{y_{1:2000}, y_{1:2000}, \ldots\}$ where each pass is one complete browse of $y_{1:2000}$. The true parameter values: $\alpha = 10$, $\beta = 0.1$, $\xi_1 = 1.78, \xi_2 = 3.56, \kappa_1 = 0.30, \kappa_2 = 0.03, \lambda_1 = \lambda_2 = 0.1, P_{i,j} = 0.5$.

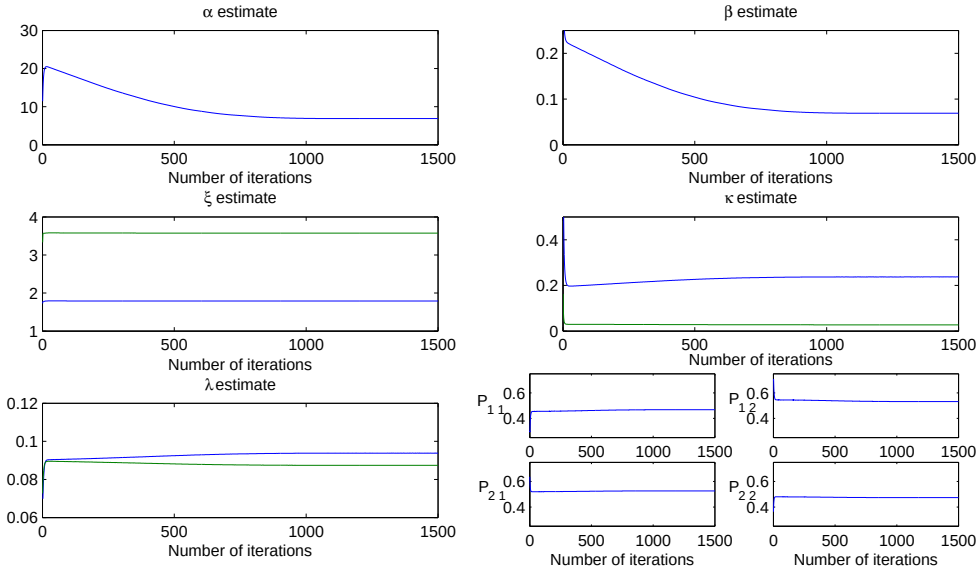

Figure 3: SMC-FS batch EM estimates vs number of iterations for for the same $y_{1:2000}$ used to produce the results in Figure 2.

5.1.3 Comparison with the path space method

As stated in Section 3.3, other than the SMC-FS online EM algorithm, it is possible to devise an online EM algorithm using $\mathbb{Q}_{\theta,n}^{p,N}$ (SMC-PS online EM), but it suffers from higher variance. In the following, we compare the performances of these two online EM algorithms.

In the first experiment, we compare the variability in the estimates of the sufficient statistics of the changepoint model defined above when the SMC-FS online EM algorithm and the SMC-PS online EM algorithm (see Section 3.4) are used with θ_n frozen to θ. We show the results for only one of the statistics, $S_{6,n}^1$, required for the EM algorithm (see Appendix A.2) in Figure 4. The figures are obtained after running 100 Monte Carlo simulations for the same sequence of observation data. For illustration purposes, while the box plots show the estimates up to time 10000, we show the relative variance along 100000 time steps. We can deduce from the box-plots and relative variance that there is much less variability in the estimates obtained by using forward smoothing and the SMC-FS method always outperforms the SMC-PS method in time and thus should be favored. Note that, using a finite number of particles, these SMC estimates are biased and will result in a loss of accuracy in the EM algorithms. To assess this bias, studies in the context of Feynman-Kac formulae are helpful. For example, the result in Del Moral et al. (2009) suggests that the bias of SMC-FS estimate of S_n/n for finite duration models is bounded by a term $\mathcal{O}(1/N)$.

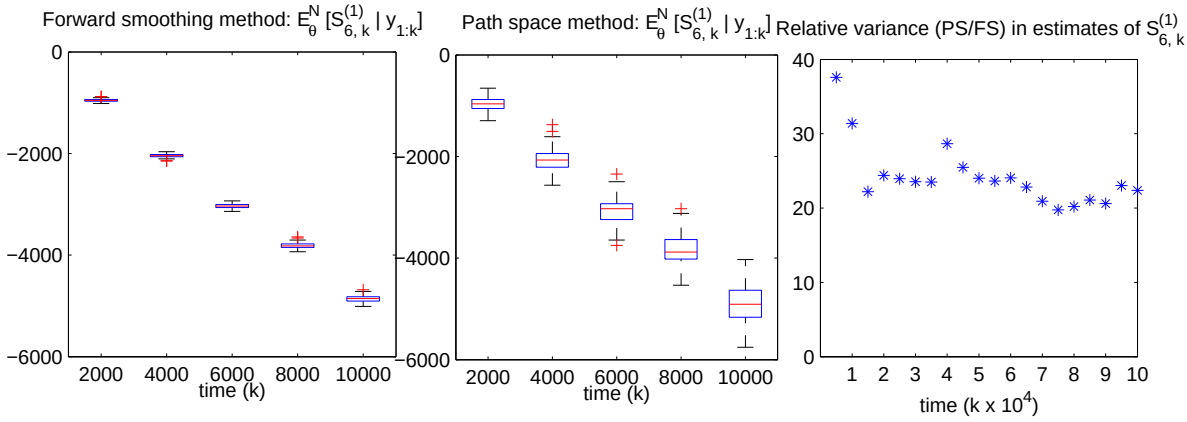

Figure 4: Comparison of the forward smoothing and the path space methods in terms of the variability in the estimates of $S_{6,n}^1$. The box plots and the relative variance plot are generated from 100 Monte Carlo simulations using the same observation data.

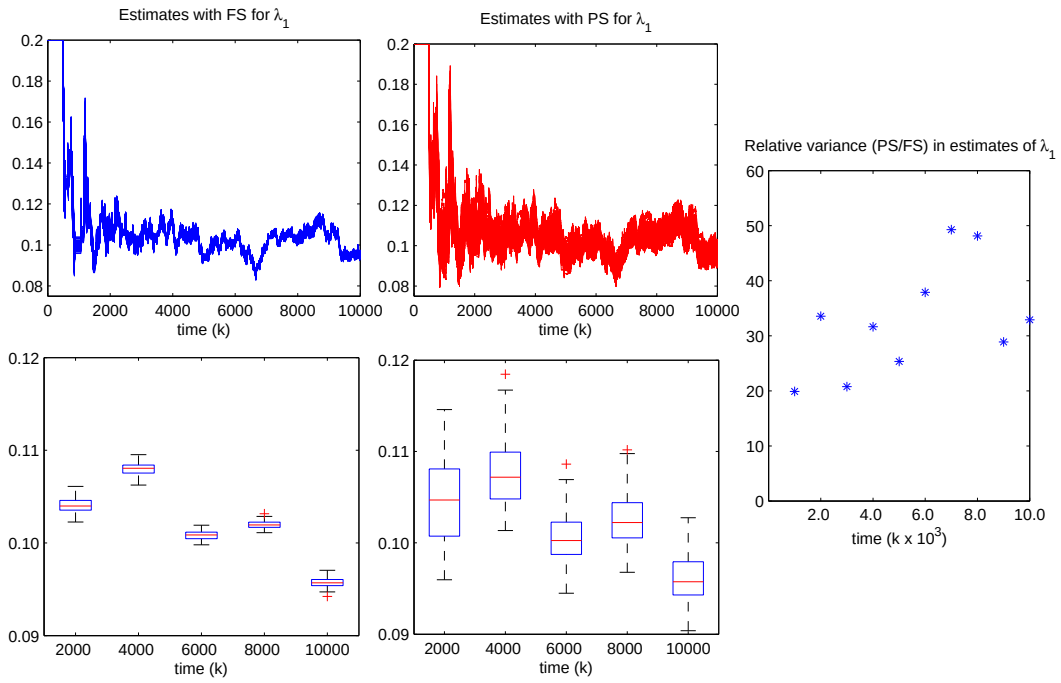

Figure 5: Comparison of SMC-FS online EM and SMC-PS online EM in terms of the variability in their estimates of $\lambda_1 = 0.1$. The two plots at the top generated by superimposing different estimates, the box plots, and the relative variance plot are generated from estimates out of 100 different Monte Carlo runs using the same observation data.

The second experiment compares the variability in the parameter estimates of the SMC-FS online EM and the SMC-PS online EM algorithms. Figure 5 shows the estimation results for the parameter λ_1 when the two algorithms are used. The results are obtained from 100 Monte Carlo simulations using the same sequence of observation data of length 10000. It is interesting to observe that the trends of estimates over time are similar for both algorithms; however, it is obvious from the box plots as well as the relative variance over time that the SMC-FS online EM estimates have less variance than the SMC-PS online EM estimates.

5.2 GC content in the DNA of Human Chromosome no. 2

We applied our online EM method to estimate the parameters of a changepoint model used for modeling the Guanine+Cytosine (GC) content along human chromosome. It appears that many features of the genome are correlated with GC content, such as gene density, repeat density, substitution rates, and recombination rates; see Fearnhead and Vasileiou (2009)

and the references therein for further explanation. It is assumed that the chromosome is separated into successive segments by changepoints and the GC content during each segment is constant. However, as the signal is obscured by small scale noise, a statistical approach may be used to uncover the sequence of changepoints. There is a commonly used binary segmentation approach implemented within the program IsoFinder (Oliver et al., 2004). Fearnhead and Vasileiou (2009) proposed the changepoint model described in Example 1. Regarding the model variables, $Z_k = (Z_{k,1}, Z_{k,2})$ were interpreted as the mean and variance of the GC content during the segment at window k, and Y_k was taken to be the observed GC content of the k'th window. The authors estimated the model parameters by using a MCEM approach and their results outperformed the ones obtained using IsoFinder.

In our experiments we used human chromosome 2, which can be downloaded via the link `http://hgdownload.cse.ucsc.edu/goldenPath/hg17`. The raw data was preprocessed as follows. The raw data consists of a single contiguous stretch of DNA data containing only four different letters: A, C, G, and T. We summarized the DNA data by partitioning the 24 Megabase (Mb) region, which is nearly the whole data set, into 80000 windows, each 3.0 kb long, and for each window recording the proportion of letters within that window that are G or C. Some parts of the DNA sequence could not be measured leading to missing parts. The noisy GC content with missing parts, which we use as the observation sequence, is shown in Figure 6. We assumed two generative models ($R = 2$) to represent segments of high and low

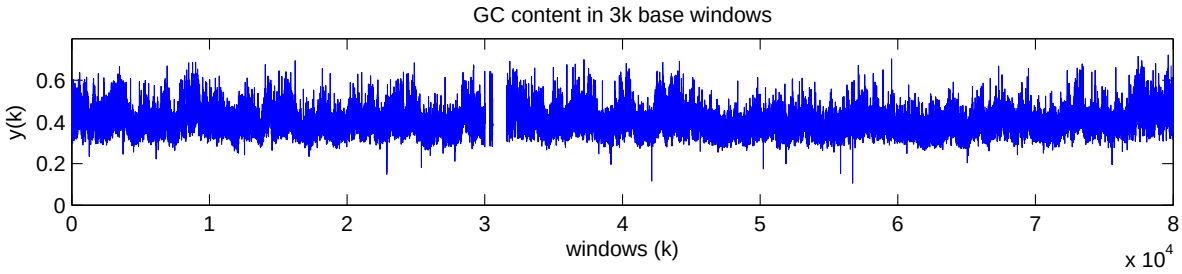

Figure 6: Noisy GC content over 3 kb windows in human DNA chromosome 2.

GC contents. The missing data problem is straightforward to handle, e.g. see Fearnhead and Vasileiou (2009). Figure 7 shows the online EM parameter estimates versus number of passes over the data obtained with Algorithm 1. One can see that most of the parameter estimates converge after 10 passes, whereas for convergence of the rest 30 passes are enough.

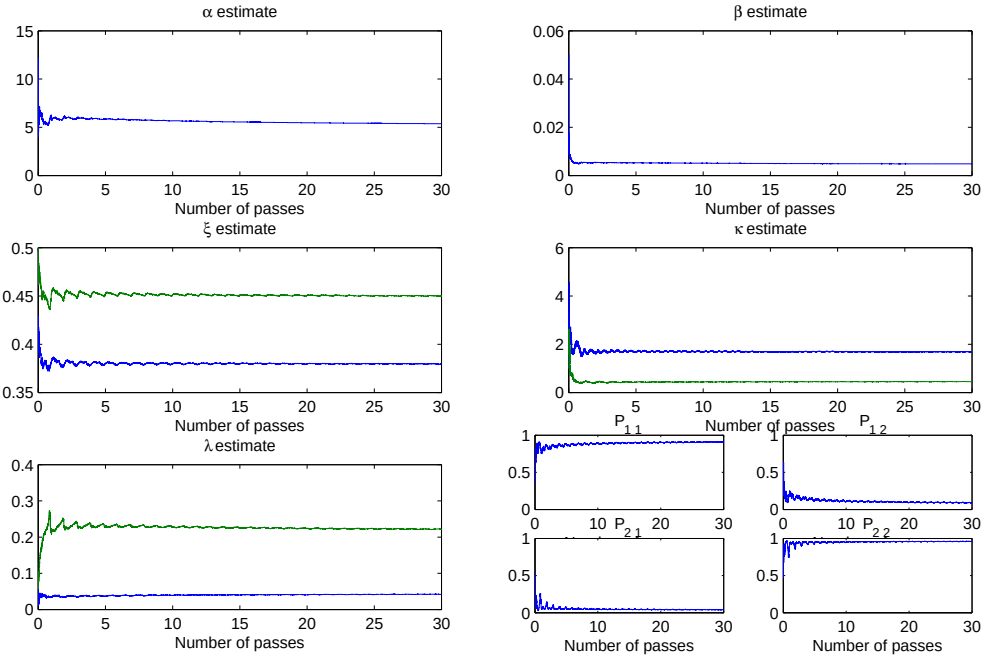

Figure 7: Online EM estimates vs number of passes over the data sequence in Figure 6.

6 Discussion

We have presented a novel SMC online EM algorithm for changepoint models and we have studied the stability of the associated SMC estimates. The proposed EM algorithm does not require the filters to be stored and has memory requirements independent of the size of the dataset. We have shown that it is practical for very long data sequences, and it can outperform the batch EM even when the data length is not so long that batch EM is impractical (in terms of memory requirement to store the filters and the entire data set).

From a Monte Carlo point of view, our SMC implementation of the forward smoothing recursion at the core of the online EM algorithm is essentially an online implementation of the forward-filtering backward-smoothing algorithm of Doucet et al. (2000) where the filtering densities are approximated using SMC and then backward smoothing is executed exactly. This method is more efficient than using the path space method as demonstrated in Section 5.1.3. Since we need only the SMC approximation of the filters, we could even use more effective SMC routines that are not applicable to a path space method; see for example the SMC algorithm in Fearnhead and Vasileiou (2009). Besides, unlike the general

state-space model case (Del Moral et al., 2009), the computational cost of our algorithm is of the same order as the cost of using a path space method in changepoint models.

Even though the numerical examples were presented for one specific changepoint model, our online EM algorithm is also applicable to the changepoint models studied in Whiteley et al. (2009) and Caron et al. (2011). More generally, the proposed online EM algorithm is applicable when the constituent laws of the changepoint model given in (1)-(2) belong to the exponential family and the latent variable $\{Z_k\}_{k\geq 1}$ can be integrated out analytically.

References

Andrieu, C., Doucet, A., and Tadić, V. B. (2005). On-line parameter estimation in general state-space models. In *Proc. 44th IEEE Conf. on Decision and Control*, pages 332–337.

Barbu, V. and Limnios, N. (2008). *Semi-Markov Chains and Hidden Semi-Markov Models toward Applications: Their Use in Reliability and DNA Analysis*. Springer.

Braun, J. V. and Muller, H. G. (1998). Statistical methods for DNA sequence segmentation. *Statistical Sciences*, 13:142–162.

Cappé, O. (2009). Online sequential Monte Carlo EM algorithm. In *Proc. IEEE Workshop on Statistical Signal Processing*.

Cappé, O. (2011). Online EM algorithm for hidden Markov models. *Journal of Computational and Graphical Statistics*, 20(3):728–749.

Cappé, O., Moulines, E., and Rydén, T. (2005). *Inference in Hidden Markov Models*. Springer.

Caron, F., Doucet, A., and Gottardo, R. (2011). On-line changepoint detection and parameter estimation with application to genomic data. *Statistics and Computing*, pages 1–17. 10.1007/s11222-011-9248-x.

Cemgil, A. T., Kappen, H. J., and Barber, D. (2006). A generative model for music transcription. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(2):679–694.

Chib, S. (1998). Estimation and comparison of multiple change-point models. *Journal of Econometrics.*, 86:221241.

Chopin, N. (2007). Dynamic detection of change points in long time series. *Annals of the Institute of Statistical Mathematics*, 59(2):349–366.

Del Moral, P. (2004). *Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications*. Springer-Verlag, New York.

Del Moral, P. and Doucet, A. (2003). On a class of genealogical and interacting Metropolis models. *Lecture Notes in Mathematics*.

Del Moral, P., Doucet, A., and Singh, S. (2009). Forward smoothing using sequential Monte Carlo. Technical Report 638, Cambridge University, Engineering Department.

Del Moral, P., Doucet, A., and Singh, S. (2010). A backward particle interpretation of Feynman-Kac formulae. *Ann. Inst. Stat. Math.ESAIM - Mathematical Modelling and Numerical Analysis*, 44:947–975.

Dias, A. and Embrechts, P. (2004). *Change-point analysis for dependence structures in finance and insurance*, chapter 16, pages 321–335. Wiley Finance Series.

Dong, M. and He, D. (2007). A segmental hidden semi-Markov model (HSMM)-based diagnostics and prognostics framework and methodology. *Mechanical Systems and Signal Processing*, 21(5):2248–2266.

Doucet, A., De Freitas, J., and Gordon, N. (2001). *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, New York.

Doucet, A., Godsill, S., and Andrieu, C. (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, 10:197–208.

Elliott, R. J., Ford, J. J., and Moore, J. B. (2002). On-line almost-sure parameter estimation for partially observed discrete-time linear systems with known noise characteristics. *International Journal of Adaptive Control and Signal Processing*, 16:435–453.

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

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

Fearnhead, P. and Vasileiou, D. (2009). Bayesian analysis of isochores. *Journal of the American Statistical Association*, 104(485):132–141.

Gales, M. J. F. and Young, S. J. (1993). The theory of segmental hidden Markov models. Technical report, Cambridge Univ. Eng. Dept.

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

Kantas, N., Doucet, A., Singh, S. S., and Maciejowski, J. M. (2009). An overview of sequential Monte Carlo methods for parameter estimation in general state-space models. In *Proceedings IFAC System Identification (SysId) Meeting.*

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

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

Mongillo, G. and Deneve, S. (2008). Online learning with hidden Markov models. *Neural Computation*, 20(7):1706–1716.

Murphy, K. P. (2002). Hidden semi-Markov models (hsmms). Technical report, UBC.

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

Oliver, J. L., Carpena, P., Hackenberg, M., and Bernaola-Galvan, P. (2004). Isofinder: Computational prediction of isochores in genome sequences. *Nucleic Acids Research*, 32:W287W29.

Olsson, J., Cappé, O., Douc, R., and Moulines, E. (2008). Sequential Monte Carlo smoothing with application to parameter estimation in nonlinear state space models. *Bernoulli*, 14:155–179.

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

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

Whiteley, N., Doucet, A., and Andrieu, C. (2009). Particle MCMC for multiple change-point models. Technical report, University of Bristol, Department of Mathematics.

A Appendix

A.1 Derivation of H_k in (4)

Given $\{x_k\}_{k\geq 1}$, consider the partition of $\{1, 2, \ldots\}$ $\{[t_1, t_2), [t_2, t_3), \ldots\}$ where t_i is the i'th time when $d_k = 1$. Each set $[t_n, t_{n+1})$ is called a segment. To emphasize the segmented structure of the changepoint model, we define $a_k = \sum_{i=1}^{k} I_{\{1\}}(d_k)$ to be the number of segments up to time k, $l_n = t_{n+1} - t_n$ to be the length of the n'th segment, and $\bar{m}_n = m_{t_n}$ to be the model number in the n'th segment. Also, we define $\bar{Z}_n = Z_{t_n:t_{n+1}-1}$ and $\bar{Y}_n = Y_{t_n:t_{n+1}-1}$ to group the variables Z_k and Y_k that belong to the same segment with shorthand notation. Recall that

$$H_k(x_{1:k}, y_{1:k}, \theta_i, \theta) = \log p_\theta(x_{1:k}) + \mathbb{E}_{\theta_i}\left[\log p_\theta(y_{1:k}, Z_{1:k}|x_{1:k}) \middle| y_{1:k}, x_{1:k}\right]. \tag{24}$$

Proposition 2. *For any changepoint model defined as in Section 2, we have*

$$H_k(x_{1:k}, \theta', \theta) = H_{k-1}(x_{1:k-1}, \theta', \theta) + h_k(x_{k-1}, x_k, \theta', \theta)$$

Proof. Since $\{X_k\}_{k\geq 1}$ is a Markov chain, so $\log p_\theta(x_{1:k}) = \log p_\theta(x_{1:k-1}) + \log p_\theta(x_k|x_{k-1})$, and we are done for the first term in (24). For the second term in (24), due to the conditional independence of $(\bar{Z}_n, \bar{Y}_n)$ given the model number at the segment n, which is $\bar{m}_n$, we have

$$p_{\theta'}(z_{1:k}|y_{1:k}, x_{1:k}) = \left[\prod_{n=1}^{a_k-1} p_{\theta'}(\bar{z}_n|\bar{y}_n, \bar{m}_n)\right] p_{\theta'}(z_{k-d_k+1:k}|y_{k-d_k+1:k}, m_k) \tag{25}$$

$$\log p_\theta(y_{1:k}, z_{1:k}|x_{1:k}) = \left[\sum_{n=1}^{a_k-1} \log p_\theta(\bar{y}_n, \bar{z}_n|\bar{m}_n)\right] + \log p_\theta(y_{k-d_k+1:k}, z_{k-d_k+1:k}|m_k) \tag{26}$$

Combining (25) and (26), we have

$$\begin{aligned} H_k(x_{1:k}, \theta', \theta) = {} & \log p_\theta(x_{1:k}) + \mathbb{E}_{\theta'}\left[\sum_{n=1}^{a_k-1} \log p_\theta(\bar{y}_n, \bar{Z}_n|\bar{m}_n) \middle| \bar{y}_n, \bar{m}_n\right] \\ & + \mathbb{E}_{\theta'}\left[\log p_\theta(y_{k-d_k+1:k}, Z_{k-d_k+1:k}|m_k) \middle| y_{k-d_k+1:k}, m_k\right] \end{aligned}$$

Now consider H_{k-1}. Given d_{k-1}, there are two possibilities for d_k, either $d_k = 1$, $d_k = d_{k-1}+1$.

- If $d_k = 1$, it means a new segment starts at time k. Therefore, $a_k = a_{k-1} + 1$ and the a_{k-1}'th segment ends at time $k-1$. This gives $H_{k-1}(x_{1:k-1}, \theta', \theta)$ being equal to

$$\log p_\theta(x_{1:k-1}) + \mathbb{E}_{\theta'}\left[\sum_{n=1}^{a_k-1} \log p_\theta(\bar{y}_n, \bar{Z}_n|\bar{m}_n)\middle| \bar{y}_n, \bar{m}_n\right]$$

- If $d_k = d_{k-1} + 1$, then we are still at the segment at which we were at time $k-1$. Therefore, we have $a_k = a_{k-1}$, $m_k = m_{k-1}$, and $H_{k-1}(x_{1:k-1}, \theta', \theta)$ is equal to

$$\begin{aligned}\log p_\theta(x_{1:k-1}) \quad &+ \quad \mathbb{E}_{\theta'}\left[\sum_{n=1}^{a_k-1} \log p_\theta(\bar{y}_n, \bar{Z}_n|\bar{m}_n)\middle| \bar{y}_n, \bar{m}_n\right] \\ &+ \quad \mathbb{E}_{\theta'}\left[\log p_\theta(y_{k-d_k:k-1}, Z_{k-d_k:k-1}|m_k)\middle| y_{k-d_k+1:k-1}, m_k\right]\end{aligned}$$

Therefore, we have $H_k(x_{1:k}, \theta', \theta) = H_{k-1}(x_{1:k-1}, \theta', \theta) + h_k(x_{k-1}, x_k, \theta', \theta)$ where

$$\begin{aligned}&h_k(x_{k-1}, x_k, \theta', \theta) = \log p_\theta(x_k|x_{k-1}) \\ &\quad + \begin{cases} \mathbb{E}_{\theta'}\left[\log p_\theta(y_k, Z_k|m_k)\middle| y_k, m_k\right], & \text{if } d_k = 1 \\ \mathbb{E}_{\theta'}\left[\log p_\theta(y_{k-d_k+1:k}, Z_{k-d_k+1:k}|m_k)\middle| y_{k-d_k+1:k}, m_k\right] & \\ -\mathbb{E}_{\theta'}\left[\log p_\theta(y_{k-d_k+1:k-1}, Z_{k-d_k+1:k-1}|m_k)\middle| y_{k-d_k+1:k-1}, m_k\right], & \text{if } d_k = d_{k-1}+1 \end{cases}\end{aligned}$$

which does not depend on the values of x_1 to x_{k-2}. □

A.2 Derivation of the EM algorithm for the model in Section 5

We write $(Z_1, Z_2) \sim \mathcal{N}\Gamma^{-1}(\xi, \kappa, \alpha, \beta)$ to mean $Z_2 \sim \Gamma^{-1}(\alpha, \beta)$ and $Z_1|z_2 \sim \mathcal{N}(\xi, \frac{z_2}{\kappa})$. If $Y_k|\,(z_1, z_2) \sim \mathcal{N}(z_1, z_2)$ for $k = 1, \ldots, n$, the marginal likelihood and the posterior are:

$$p(y_{1:n}) = \frac{\pi^{-n/2}\,(2\beta)^\alpha\,\Gamma\,(\alpha + n/2)}{\left(2\beta + \sum_{k=1}^n y_k^2 + \xi^2\kappa - \frac{\sum_{k=1}^n y_k + \xi^2\kappa}{n+\kappa}\right)^{n/2+\alpha}}$$

$$(Z_1, Z_2)\,|\,(y_{1:n}) \sim \mathcal{N}\Gamma^{-1}\left(\frac{\kappa\xi + n\bar{y}}{\kappa+n}, \kappa + n, \alpha + \frac{n}{2}, \beta + \frac{1}{2}\sum_{k=1}^n (y_k - \bar{y})^2 + \frac{n\kappa}{n+\kappa}\frac{(\bar{y}^2 - \xi)^2}{2}\right)$$

where $\bar{y} = \frac{1}{n}\sum_{k=1}^n y_k$. Also, the required expectations are analytically available:

$$\mathbb{E}\left[1/Z_2\right] = \alpha/\beta,\ \mathbb{E}\left[Z_1/Z_2\right] = \xi\alpha/\beta,\ \mathbb{E}\left[Z_1^2/Z_2\right] = 1/\kappa + \xi^2\alpha/\beta,\ \mathbb{E}\left[\log Z_2\right] = \log\beta - \Psi(\alpha)$$

For the EM algorithm, we estimate the following functionals for $m, m_1, m_2 = 1, \ldots, R$:

$$S^m_{1,k}(x_{1:k}, \theta_i) = \sum_{n:\bar{m}_n = m}^{a_k} 1, \quad S^m_{2,k}(x_{1:k}, \theta_i) = \sum_{n:\bar{m}_n = m}^{a_k - 1} (l_n - 1) + I_{\{m\}}(m_k)\,(d_k - 1)\,,$$

$$S^{m_1,m_2}_{3,k}(x_{1:k}, \theta_i) = \sum_{n:\bar{m}_n = m_1, \bar{m}_{n+1} = m_2}^{a_k - 1} 1$$

$$S^m_{4,k}(x_{1:k}, \theta_i) = \sum_{n:\bar{m}_n = m}^{a_k - 1} \mathbb{E}_{\theta_i}\left[\log Z_{t_n,2} \middle| \bar{y}_n, m\right] + I_{\{m\}}(m_k)\mathbb{E}_{\theta_i}\left[\log Z_{k,2} \middle| y_{k-d_k+1:k}, m\right],$$

$$S^m_{5,k}(x_{1:k}, \theta_i) = \sum_{n:\bar{m}_n = m}^{a_k - 1} \mathbb{E}_{\theta_i}\left[1/Z_{t_n,2} \middle| \bar{y}_n, m\right] + I_{\{m\}}(m_k)\mathbb{E}_{\theta_i}\left[1/Z_{t_n,2} \middle| y_{k-d_k+1:k}, m\right],$$

$$S^m_{6,k}(x_{1:k}, \theta_i) = \sum_{n:\bar{m}_n = m}^{a_k - 1} \mathbb{E}_{\theta_i}\left[Z_{t_n,1}/Z_{t_n,2} \middle| \bar{y}_n, m\right] + I_{\{m\}}(m_k)\mathbb{E}_{\theta_i}\left[Z_{t_n,1}/Z_{t_n,2} \middle| y_{k-d_k+1:k}, m\right],$$

$$S^m_{7,k}(x_{1:k}, \theta_i) = \sum_{n:\bar{m}_n = m}^{a_k - 1} \mathbb{E}_{\theta_i}\left[Z^2_{t_n,1}/Z_{t_n,2} \middle| \bar{y}_n, m\right] + I_{\{m\}}(m_k)\mathbb{E}_{\theta_i}\left[Z^2_{t_n,1}/Z_{t_n,2} \middle| y_{k-d_k+1:k}, m\right].$$

The corresponding additive functions are

$$s^m_{1,k}(x_{k-1}, x_k, \theta_i) = I_{\{m\}}(m_k)I_{\{1\}}(d_k) \quad s^m_{2,k}(x_{k-1}, x_k, \theta_i) = I_{\{m\}}(m_k)I_{\{d_{k-1}+1\}}(d_k),$$

$$s^{m_1,m_2}_{3,k}(x_{k-1}, x_k, \theta_i) = I_{\{1\}}(d_k)I_{\{m_1\}}(m_{k-1})I_{\{m_2\}}(m_k),$$

$$\begin{aligned} s^m_{4,k}(x_{k-1}, x_k, \theta_i) &= I_{\{m\}}(m_k)\left\{I_{\{1\}}(d_k)\mathbb{E}_{\theta_i}\left[\log Z_{k,2} \middle| y_k, m\right]\right. \\ &\quad \left. + I_{\{d_{k-1}+1\}}(d_k)\left(\mathbb{E}_{\theta_i}\left[\log Z_k \middle| y_{k-d_k+1:k}, m\right] - \mathbb{E}_{\theta_i}\left[\log Z_{k,2} \middle| y_{k-d_k+1:k-1}, m\right]\right)\right\}, \\ s^m_{5,k}(x_{k-1}, x_k, \theta_i) &= I_{\{m\}}(m_k)\left\{I_{\{1\}}(d_k)\mathbb{E}_{\theta_i}\left[1/Z_{k,2} \middle| y_k, m\right]\right. \\ &\quad \left. + I_{\{d_{k-1}+1\}}(d_k)\left(\mathbb{E}_{\theta_i}\left[1/Z_{k,2} \middle| y_{k-d_k+1:k}, m\right] - \mathbb{E}_{\theta_i}\left[1/Z_{k,2} \middle| y_{k-d_k+1:k-1}, m\right]\right)\right\}, \\ s^m_{6,k}(x_{k-1}, x_k, \theta_i) &= I_{\{m\}}(m_k)\left\{I_{\{1\}}(d_k)\mathbb{E}_{\theta_i}\left[Z_{k,1}/Z_{k,2} \middle| y_k, m\right]\right. \\ &\quad \left. + I_{\{d_{k-1}+1\}}(d_k)\left(\mathbb{E}_{\theta_i}\left[Z_{k,1}/Z_{k,2} \middle| y_{k-d_k+1:k}, m\right] - \mathbb{E}_{\theta_i}\left[Z_{k,1}/Z_{k,2} \middle| y_{k-d_k+1:k-1}, m\right]\right)\right\}, \\ s^m_{7,k}(x_{k-1}, x_k, \theta_i) &= I_{\{m\}}(m_k)\left\{I_{\{1\}}(d_k)\mathbb{E}_{\theta_i}\left[Z^2_{k,1}/Z_{k,2} \middle| y_k, m\right]\right. \\ &\quad \left. + I_{\{d_{k-1}+1\}}(d_k)\left(\mathbb{E}_{\theta_i}\left[Z^2_{k,1}/Z_{k,2} \middle| y_{k-d_k+1:k}, m\right] - \mathbb{E}_{\theta_i}\left[Z^2_{k,1}/Z_{k,2} \middle| y_{k-d_k+1:k-1}, m\right]\right)\right\}. \end{aligned}$$

The maximization step is as follows: Letting $\widehat{S}^m_{j,n}(\theta) = \mathbb{E}_\theta\left[S^m_{j,n}(X_{1:n},\theta)\middle|\, y_{1:n}\right]$,

$$\alpha^{(i+1)} = \Psi^{-1}\left(\frac{\log\beta^{(i)}\sum_{m=1}^R \widehat{S}^m_{1,n}(\theta_i) + \sum_{m=1}^R \widehat{S}^m_{4,n}(\theta_i)}{\sum_{m=1}^R \widehat{S}^m_{1,n}(\theta_i)}\right),\ \beta^{(i+1)} = \alpha^{(i+1)}\frac{\sum_{m=1}^R \widehat{S}^m_{1,n}(\theta_i)}{\sum_{m=1}^R \widehat{S}^m_{5,n}(\theta_i)}$$

$$\xi^{(i+1)}_m = \widehat{S}^m_{6,n}(\theta_i)/\widehat{S}^m_{5,n}(\theta_i),\ \kappa^{(i+1)}_m = \widehat{S}^m_{1,n}(\theta_i)/\left(\widehat{S}^m_{7,n}(\theta_i) - 2\xi^{(i+1)}_m\widehat{S}^m_{6,n}(\theta_i) + \xi^{(i+1)2}_m\widehat{S}^m_{5,n}(\theta_i)\right)$$

$$\lambda^{(i+1)}_m = \widehat{S}^m_{1,n}(\theta_i)/\left(\widehat{S}^m_{2,n}(\theta_i) + \widehat{S}^m_{1,n}(\theta_i)\right),\ P^{(i+1)}_{m_1,m_2} = \widehat{S}^{m_1,m_2}_{3,n}(\theta_i)/\sum_{m=1}^R \widehat{S}^{m_1,m}_{3,n}(\theta_i)$$

where $\Psi(x) = d\log\Gamma(x)/dx$ is the derivative of the log-gamma function.

A.3 Proof of Proposition 1

We will first establish a weaker form of backward forgetting for the infinite duration model with the aid for the following lemma, whose proof is straightforward and is omitted.

Lemma 1. *Let $M(x,x')$ be a Markov transition kernel (from x to x') on $\mathcal{X}$, c a constant and m a probability measure on $\mathcal{X}$. If $c^{-1}\,m(x') \le M(x,x') \le c\,m(x')$ for all $x \in A$, where $A \subseteq \mathcal{X}$, then for any $B \subseteq \mathcal{X}$ and $\varphi \in \mathcal{B}(\mathcal{X})$ such that $\operatorname{osc}(\varphi) \le 1$,*

$$\operatorname{osc}_A(M(\varphi)) \le (1-c^{-1})\operatorname{osc}_B(\varphi) + 2c\,m\left(\overline{B}\right),$$

Corollary 1. *Assume $\{y_i\}_{i\ge1}$ is given with $p(y_{1:n}) > 0$ for all n. Let $\varphi_n(x_n) = \mathbb{E}\left[\,s(X_1)\middle|\, x_n, y_{1:n-1}\right]$. For any $L > 0$, $n - L > 0$, $\operatorname{osc}(s) \le 1$, $A \subseteq \operatorname{supp}(\eta_n)$, $B \subseteq \operatorname{supp}(\eta_{n-L})$*

$$\operatorname{osc}_A(\varphi_n) \le \left(1 - c_{n-L,L}(A)^{-1}\right)\operatorname{osc}_B(\varphi_{n-L}) + 2c_{n-L,L}(A)\eta_{n-L}(\overline{B}). \tag{27}$$

Furthermore, let $A = \{1,\dots,L\}\times\{1,\dots,R\}$. If there exist finite positive constants C, $\gamma \in (0,1)$ and $c(L)$ such that for all L

$$\sup_{i\ge1}\eta_i(\overline{A}) \le C\gamma^L \qquad \text{and} \qquad \sup_{i\ge1} c_{i,L}(A) \le c(L) \tag{28}$$

then for all L large enough, for all n,

$$\operatorname{osc}_{A\cap\operatorname{supp}(\eta_n)}(\varphi_n) \le (1-c(L)^{-1})^{\left\lfloor\frac{n-1}{L}\right\rfloor} + 2c(L)^2C\gamma^L. \tag{29}$$

Proof. Substituting $l = L$ and $k = n - L$ in (21), it can be shown that

$$c_{n-L,L}(A)^{-1}p\left(x_{n-L}|y_{1:n-L-1}\right) \le p\left(x_{n-L}|x_n, y_{1:n-1}\right) \le c_{n-L,L}(A)p\left(x_{n-L}|y_{1:n-L-1}\right)$$

for all $x_n \in A$. The bound (27) now follows from Lemma 1 with $c = c_{n-L,L}(A)$, $m(x_{n-L}) = p\,(x_{n-L}|y_{1:n-L-1})$, $M(x_n, x_{n-L}) = p(x_{n-L}|y_{1:n-1}, x_n)$, and $\varphi = \varphi_{n-L}$. The second bound (29) follows from (27) by iterating the backward kernels with $B = A \cap \mathrm{supp}(\eta_{n-L})$, and using the tail behavior of the minorization measure in (28). □

The first condition in (28) is a uniform tightness condition on the probabilities η_i. This bound for the tail probabilities can be loosened but only at the expense of a weaker bound in Proposition 1. It is clear that (29) is weaker than backward forgetting at geometric rate.

Corollary 1 presents a weaker form of backward forgetting for the infinite duration model. The following lemma establishes that the finite duration models posses the geometric forward forgetting and geometric backward forgetting properties; both of which are necessary in order to establish linear growth of the variance.

Lemma 2. *For a finite duration changepoint model, let* $\bar{d}_m = \min\{d' : \lambda_m(d) = 1, d \geq d'\}$ *be the maximum duration length in model* m *and let* $\mathcal{X}_f = \bigcup_{m=1}^{R} \{(1,m), \ldots, (\bar{d}_m, m)\}$. *Assume that the transition matrix* $\{p(x_k|x_{k-1}) : x_k, x_{k-1} \in \mathcal{X}_f\}$ *is irreducible and aperiodic; and that for the given* $\{y_n\}_{n\geq 1}$ *there exist finite positive constants* c_n *such that* $c_n^{-1} \leq G_n \leq c_n$ *for all* n. *(i) Then there exists a positive integer* L *such that* $c_{k,l}$ *defined in (21) is finite for all* $l \geq L$, $k \geq 1$. *(ii) It now follows that for all* $l \geq L$, $n \geq k + l$, *and* $x_{k+l} \in \mathcal{X}_f$,

$$p(x_{k+l}|x_k, y_{1:n}) \geq c_{k,l}^{-2}\, p(x_{k+l}|x'_k, y_{1:n}) \tag{30}$$

and the inequality holds irrespective of (x_k, x'_k) *provided both conditional probabilities are well defined. (iii) Furthermore, the Markov chain on* $\mathcal{X}$ *with transition kernel* $p(x_{k+1}|x_k, y_{1:n})$, $k = 1, \ldots, n-1$, *forgets its initialization in the following sense: for all* $n \geq m \geq k \geq 1$,

$$|\mathbb{E}\,[\,s(X_m)|\,x_k, y_{1:n}] - \mathbb{E}\,[\,s(X_m)|\,x'_k, y_{1:n}]| \leq \mathrm{osc}(s) \prod_{i=1}^{\lfloor \frac{m-k}{L} \rfloor} (1 - c_{k+(i-1)L,L}^{-2}) \tag{31}$$

irrespective of (x_k, x'_k) *provided both conditional expectations are well defined. If* $c_n \leq c < \infty$ *for all* n, *then (iv)* $c_{k,l} \leq c(l) < \infty$ *for* $l \geq L$, $k \geq 1$ *and the rate in (31) is geometric, and (v) letting* $\varphi_n(x_n) = \mathbb{E}\,[\,s(X_1)|\,x_n, y_{1:n-1}]$, *for all* $l \geq L$, *for all* n, *all* $A \subseteq \mathrm{supp}(\eta_n)$

$$\mathrm{osc}_A(\varphi_n) \leq \mathrm{osc}(s)(1 - c(l)^{-1})^{\lfloor n/l \rfloor}. \tag{32}$$

Proof. (Outline only) Property (i) is a consequence of some well known facts for finite state Markov chains. We use the fact that, under the stated assumptions, the Markov chain

restricted to $\mathcal{X}_f$ has a stationary distribution, say $\nu(x)$, and we have $\nu(\mathcal{X}_f) = 1$ and $\nu > 0$ on $\mathcal{X}_f$. This ensures the ratio $p(x_{k+l}|x_k)/p(x_{k+l}|x'_k)$ is close to 1 uniformly in its arguments and k, provided l is large enough. The result now follows from the the fact that G_n is bounded from below and above. Property (ii) follows from (i) while the forgetting property in (31) is a simple consequence of (30), e.g. see Del Moral (2004). Property (iv) is proved similarly to (i) using instead the uniform bound on G_n. To verify (v) use (iv) and (27), i.e. iterate the backward kernels starting with $B = \text{supp}(\eta_{n-l})$ □

Finally, we will need the following lemma to prove Proposition 1

Lemma 3. *Given $\{y_n\}_{n\geq 1}$, assume there exists a finite constant c such that $c^{-1} \leq G_n \leq c$ for all n and that (19) holds then, for all n, $1 < k \leq n$,*

$$\sup_{(x_k, x'_k) \in \text{supp}(\eta_k)} \frac{p(y_{k:n}|x_k, y_{1:k-1})}{p(y_{k:n}|x'_k, y_{1:k-1})} < \infty.$$

Proof. Using $|\log(b) - \log(a)| \leq \frac{|b-a|}{\min(a,b)}$,

$$\begin{aligned} \log \frac{p(y_{k:n}|x_k, y_{1:k-1})}{p(y_{k:n}|x'_k, y_{1:k-1})} &= \sum_{i=k}^{n} \log p(y_i|x_k, y_{1:i-1}) - \log p(y_i|x'_k, y_{1:i-1}) \\ &\leq \sum_{i=k}^{n} \frac{|p(y_i|x_k, y_{1:i-1}) - p(y_i|x'_k, y_{1:i-1})|}{\min(p(y_i|x_k, y_{1:i-1}), p(y_i|x'_k, y_{1:i-1}))}. \end{aligned}$$

Since $p(y_i|x_k, y_{1:i-1}) = \mathbb{E}\left[G_i(X_i)\middle| x_k, y_{1:i-1}\right]$, each ratio can be bounded using (19) and constant c, which then results in a geometric sum and gives the desired uniform bound. □

We can now present the proof of Proposition 1.

Proof. ***(Proposition 1):*** The asymptotic variance is

$$\sum_{i=0}^{n} \eta_i \left([G_{i,n}\ D_{i,n}(S_n - \mathbb{Q}_n(S_n))]^2 \right). \tag{33}$$

Consider the infinite duration model. Consider the ith term: For any $A \subseteq \mathcal{X}$,

$$\begin{aligned} \eta_i \left([G_{i,n}\ D_{i,n}(S_n - \mathbb{Q}_n(S_n))]^2 \right) &\leq \|G_{i,n}\|^2_{\text{supp}(\eta_i)}\, \eta_i \left([D_{i,n}(S_n - \mathbb{Q}_n(S_n))]^2 \right) \\ &\leq \|G_{i,n}\|^3_{\text{supp}(\eta_i)} \int \eta_i(dx_i)\eta_i(dx'_i) \left([D_{i,n}(S_n)(x_i) - D_{i,n}(S_n)(x'_i)]^2 \right) \\ &\leq \|G_{i,n}\|^3_{\text{supp}(\eta_i)} \left(\left[\text{osc}_{A \cap \text{supp}(\eta_i)} D_{i,n}(S_n)\right]^2 + 2n^2 \eta_i(\overline{A}) \right) \end{aligned} \tag{34}$$

Now let $A = \{1, \ldots, L\} \times \{1, \ldots, R\}$. It follows from (19) that for some integer L',

$$\sup_{x_i, x_i' \in \mathrm{supp}(\eta_i)} \left| \mathbb{E}\left[\sum_{k=i}^{n} s_k(X_k) \middle| x_i, y_{1:n} \right] - \mathbb{E}\left[\sum_{k=i}^{n} s_k(X_k) \middle| x_i', y_{1:n} \right] \right| \leq L' c(L')^2,$$

and from (29) that

$$\sup_{x_i, x_i' \in A \cap \mathrm{supp}(\eta_i)} \left| \mathbb{E}\left[\sum_{k=1}^{i-1} s_k(X_k) \middle| x_i, y_{1:n-1} \right] - \mathbb{E}\left[\sum_{k=1}^{i-1} s_k(X_k) \middle| x_i', y_{1:n-1} \right] \right| \leq (i-1) 2c(L)^2 C \gamma^L I_{[i \geq L]} + Lc(L).$$

Thus using Lemma 3 to uniformly bound $\|G_{i,n}\|_{\mathrm{supp}(\eta_i)}$ and the fact that the bounds in (28) are satisfied for all L large enough with $c(L) < c' < \infty$, (33) can be upper bounded by

$$\begin{aligned} & C' \sum_{i=1}^{n} \left((i-1)^2 \gamma^{2L} I_{[i \geq L]} + L^2 + (L')^2 + n^2 \eta_i(\overline{A}) \right) \\ & \leq C' n^3 \gamma^{2L} + C' n L^2 + C' n (L')^2 + n^3 C \gamma^L \end{aligned}$$

where C' is independent of L and n. Setting $L = k \log n$ for some fixed constant k we see that (33) is upper bounded by a term $\mathcal{O}(n \log^2 n)$.

The proof for the finite duration model follows the same lines where Lemma 2 is used instead of Corollary 1, hence it is omitted. □

A.3.1 Verification of Example 2 satisfying the conditions of Proposition 1

The first condition of Theorem 1 is satisfied since $g(y_k|z_k) > 0$ for all z_k, y_k. It follows from (23) that

$$c^{-1} \leq \frac{\int \prod_{i=1}^{n} f(z_i''|z_{i-1}'') g(y_i|z_i'') \, dz_{1:n}''}{\int \prod_{i=1}^{n} f(z_i'|z_{i-1}') g(y_i|z_i') \, dz_{1:n}'} \leq c$$

for all $n \geq 1$, $y_{1:n}$, z_0' ,z_0''. This, together with (20) implies

$$c^{-1} \leq \frac{p(y_{k:n}|x_k, y_{1:k-1})}{p(y_{k:n}|x_k', y_{1:k-1})} \leq c \tag{35}$$

for all $(x_k, x_k') \in \mathrm{supp}(\eta_k)$, $k \leq n$. (35) now implies the term $\|G_{i,n}\|_{\mathrm{supp}(\eta_i)}$ in (34) is also uniformly bounded by the constant c. (Note that the condition $c^{-1} \leq G_n \leq c$ for all n in Proposition 1 is used to verify the term $\|G_{i,n}\|_{\mathrm{supp}(\eta_i)}$ in (34) is uniformly bounded in n and is now no longer needed for this example as we have direct verification via (35).)

Since

$$\begin{aligned} p(x_k|x_{k-1}, y_{1:n}) &\propto p(y_{k:n}|x_k, x_{k-1}, y_{1:k-1})p(x_k|x_{k-1}, y_{1:k-1}) \\ &= p(y_{k:n}|x_k, y_{1:k-1})p(x_k|x_{k-1}), \end{aligned}$$

we have that

$$p(x_k|x_{k-1}, y_{1:n}) \geq c^{-1}p(x_k|x_{k-1}) \geq c^{-1}\lambda\,\delta_1(x_k) \tag{36}$$

for all $k \leq n$, and obviously for $k > n$ too. To establish forward forgetting, it follows from the minorization condition in (36) that

$$\mathbb{E}\left[s_k(X_k)|\, x_1, y_{1:n}\right] - \mathbb{E}\left[s_k(X_k)|\, x_1', y_{1:n}\right] \leq \mathrm{osc}(s_k)\left(1 - c^{-1}\lambda\right)^{k-1}.$$

Let $A = \{1, \ldots, L\}$. For $x_{k+L} \in A$, $x_k \in \mathrm{supp}(\eta_k)$ and $x_k' \in \mathrm{supp}(\eta_k)$,

$$\frac{p(x_{k+L}, y_{k:k+L-1}|x_k, y_{1:k-1})}{p(x_{k+L}, y_{k:k+L-1}|x_k', y_{1:k-1})} = \frac{p(y_{k:k+L-1}|x_{k+L}, x_k, y_{1:k-1})}{p(y_{k:k+L-1}|x_{k+L}, x_k', y_{1:k-1})}\frac{p(x_{k+L}|x_k)}{p(x_{k+L}|x_k')}.$$

By (35), the first ratio is bounded by c. The second ratio is 1. Thus $c_{k,L}(A) \leq c$. Using (36), $\sup_{i \geq L+1}\mathbb{E}\left[I_{\overline{A}}(X_i)|\, y_{1:i-1}\right] \leq \gamma^L$ where $\gamma = 1 - c^{-1}\lambda$. Hence the bounds in (22) apply with constants independent of L and n.

A.4 Supplementary materials

Code and data for the experiments: A MATLAB package containing the codes and the real data set for the experiments in Section 5 is available online.