Quantifying the Uncertainty in Change Points

Christopher F. H. Nam, John A. D. Aston and Adam M. Johansen,
Department of Statistics, University of Warwick, Coventry, CV4 7AL, UK
{c.f.h.nam | j.a.d.aston | a.m.johansen}@warwick.ac.uk

May 31, 2011

Abstract

Quantifying the uncertainty in the location and nature of change points in time series is important in a variety of applications. Many existing methods for estimation of the number and location of change points fail to capture fully or explicitly the uncertainty regarding these estimates, whilst others require explicit simulation of large vectors of dependent latent variables.

This paper proposes methodology for approximating the full posterior distribution of various change point characteristics in the presence of parameter uncertainty. The methodology combines recent work on evaluation of exact change point distributions conditional on model parameters via Finite Markov Chain Imbedding in a Hidden Markov Model setting, and accounting for parameter uncertainty and estimation via Bayesian modelling and Sequential Monte Carlo. The combination of the two leads to a flexible and computationally efficient procedure, which does not require estimates of the underlying state sequence.

We illustrate that good estimation of posterior distributions regarding change point characteristics is provided for simulated and functional magnetic resonance imaging data. We use the methodology to show that the modelling of relevant physical properties of the scanner can influence detection of change points and their uncertainty.

Keywords: Change Points; Finite Markov Chain Imbedding; Functional Magnetic Resonance Imaging; Hidden Markov Models; Sequential Monte Carlo; Segmentation

1 Introduction

Detecting and estimating the number and location of change points in time series is becoming increasingly important as both a theoretical research problem and a necessary part of applied data analysis. Originating in the 1950s in a quality control setting (Page, 1954), there are numerous existing approaches, both parametric and non-parametric, often requiring strong assumptions upon the type of changes that can occur and the distribution of the data. We refer the reader to Chen and Gupta (2000); Eckley et al. (2011) for good overviews of some of these existing methods. It is also worth noting that change point problems appear under various names including segmentation, novelty detection, structural break identification, and disorder detection. These approaches however, typically fail to fully capture uncertainty in the number and location of these change points. For example, model selection and optimal segmentation based techniques (for example Yao (1988); Davis et al. (2006)) rely on asymptotic arguments on providing consistent estimates of the number of change points present, whilst others assume the number of change points to be known, in order to consider the uncertainty regarding the locations of these change points (see Chib (1998); Stephens (1994)). Those methods which do fully characterise the uncertainty involved typically require simulation of large vectors of correlated latent variables. This paper proposes a methodology which fully quantifies the uncertainty of change points for an observed time series, without estimating or simulating the unobserved state sequence.

Our proposed methodology is based upon three areas of existing work. We model our observed time series and consider change points in a Hidden Markov Model (HMM) framework. HMMs and the general use of dependent latent state variables are widely used in change point estimation (Chib, 1998; Fearnhead, 2006; Fearnhead and Liu, 2007). In these approaches, each state of the underlying chain represents a segment of data between change points and thus a change point is said to occur when there is a change in state in the underlying chain. The underlying chain is constructed so that there are only two possible moves; either stay in the same state (no change point has occurred), or move to the next state in the sequence, corresponding to a new segment and thus a change point has occurred. Interest now lies predominantly in determining the latent state sequence (usually through simulation, by Markov Chain Monte Carlo (MCMC) for example), in order to determine the relevant change point characteristics. We note that under the framework of Chib (1998), the number of change points is assumed to be known since this is related to the number of states of the imposed HMM. However, this is quite restrictive and makes sense only in those settings in which return to a previously visited segment and state is regarded as impossible.

We consider an alternative approach by using HMMs in their usual context, where each state represents different data generating mechanisms (for example the “good” and “bad” states when using a Poisson HMM to model the number of daily epileptic seizure counts Albert (1991)) and returning to previously visited states is possible. This allows the number of change points to be unknown a priori and inferred from the data. We do at present assume that the number of different states is known although the method can be extended to the more general case. This latter point seems less restrictive in a change point context than assuming the number of change points to be known given the quantities of interest.

We also consider a generalised definition of change points corresponding to a *sustained* change in the underlying state sequence. This means that we are looking for runs of particular states in the underlying state sequence: determining that a change point to a particular regime has occurred when a particular sequence of states is observed. We employ Finite Markov Chain Imbedding (FMCI) (Fu and Koutras, 1994; Fu and Lou, 2003), an elegant framework which allows distributions regarding run and pattern statistics to be efficiently calculated exactly in that they are not subject to sampling or approximation error.

The above techniques allow exact change point distributions to be computed. However, these distributions are conditional upon the model parameters. In practice, it is common for these parameters to be treated as known, with maximum likelihood estimates being used. However, in most applications where parameters are estimated from the data itself, it is desirable to account for parameter uncertainty in change point estimates. If a Bayesian approach to the characterisation of changes is employed, then it would also seem desirable to take a Bayesian approach to the characterisation of parameter uncertainty. Recent Bayesian change point approaches have dealt with model parameter uncertainty by integrating the parameters out in some fashion in order to ultimately sample from the joint posterior of the location and number of change points, usually achieved by also sampling the aforementioned latent state sequence (Fearnhead, 2006; Chib, 1998). However, this introduces additional sampling error into the change point estimates and requires the simulation of the underlying state sequence which is often long and highly correlated — and thus hard to sample efficiently. We consider model parameter uncertainty by sampling from the the posterior distribution of the model parameters via Sequential Monte Carlo, without simulating the latent state sequences. This approach introduces sampling error only in the model parameters and retains, conditionally, the exact change point distributions: we will show that this amounts to a Rao-Blackwellized form of the estimator.

Quantifying the uncertainty in change point problems is often overlooked but nevertheless an important aspect of inference. Whilst, quite naturally, more emphasis has typically been placed on detection and estimation in problems, quantifying the uncertainty of change points can lead to a better understanding of the data and the system generating the data. Whenever estimates are provided for the location of change points, we should be interested in determining how confident we can be about these estimates, and whether other change point configurations are plausible. In many situations it may be desirable to average over models rather than choosing a most probable explanation. Alternatively, we may want to assess the confidence we have in the estimate of the number of change points and if there is any substantial probability of any other number of change points having occurred. In addition, different change point approaches can often lead to different estimates when applied to the same time series; this motivates the assessment of the performance and plausibility of these different approaches and their estimates. Quantifying the uncertainty provides a means of so doing.

The exact change point distributions computed via FMCI methodology (Aston et al., 2009) already quantify the residual uncertainty given both the model parameters and the observed data. However, this conditioning on the model parameters is typically difficult to justify. It is important to consider also parameter uncertainty because the use of different model parameters can give quite different change point results and thus conclusions. This effect becomes more important when there are several different competing model parameter values which provide equally-plausible explanations of the data. By considering model parameter uncertainty within the quantification of uncertainty for change points, we are able to account for all types of change point behaviour under a variety of model parameter scenarios and thus fully quantifies the uncertainty regarding change points. This will be seen to be especially true in the analysis of functional Magnetic Resonance Imaging (fMRI) time series.

When analysing fMRI data, it is common to assume that the data arises from a known experimental design (Worsley et al., 2002). However, this assumption is very restrictive particularly in experiments common in psychology where the exact timing of the expected reaction is unknown, with different subjects reacting at different times and in different ways to an equivalent stimulus (Lindquist et al., 2007). Change point methodology has therefore been proposed as a possible solution to this problem, where the change points effectively act as a latent design for each time series. Significant work has been done in designing methodology for these situations for the at-most-one-change situation using control chart type methods (Lindquist et al., 2007; Robinson et al., 2010). Using the methodology developed in this paper, we are able to define an alternative approach based on HMMs that allows not only multiple change points to be taken into account, but also the inclusion of an autoregressive (AR) error process assumptions and detrending within a unified analysis. These features need to be accounted for in fMRI time series (Worsley et al., 2002) and will be shown to have an effect on the conclusions that can be drawn from the associated analysis.

The remainder of this paper has the following structure: Section 2 details the statistical background of the methodology which is proposed in Section 3. This methodology is applied to both simulated and fMRI data in Section 4. We conclude in Section 5 with some discussion of our findings.

2 Background

Let $y_1, y_2, \ldots, y_n$ be an observed non-stationary time series with respect to a varying second order structure. One particular framework for modelling such a time series is via Hidden Markov Models (HMMs) where the observation process $\{Y_t\}_{t>0}$ is conditionally independent given an unobserved underlying Markov chain $\{X_t\}_{t>0}$. The states of the underlying chain correspond to different data generating mechanisms, with each state characterised by a collection of parameter values. The methods presented in this paper can be applied to general finite state Hidden Markov Models (including Markov switching models) with finite dependency on previous states of the underlying chain. This class of HMMs are of the form:

$$\begin{aligned} y_t|y_{1:t-1}, x_{1:t} &\sim f(y_t|x_{t-r:t}, y_{1:t-1}, \theta) && \text{(Emission)} \qquad (1)\\ p(x_t|x_{1:t-1}, y_{1:t-1}, \theta) &= p(x_t|x_{t-1}, \theta) \quad t = 1, \ldots, n && \text{(Transition).} \end{aligned}$$

Given the set of model parameters θ, the observation at time $t = 1, \ldots, n$, y_t has emission density dependent of previous observations $y_{1:t-1}$ and previous r states of the underlying states $x_{t-r}, \ldots, x_{t-1}$. We use the common shorthand notation of $y_{t_1:t_2} = (y_{t_1}, y_{t_1+1}, \ldots, y_{t_2})$ and analogously for $x_{t_1:t_2}$. The underlying states are assumed to follow a first order Markov chain (although standard embedding arguments would in principle allow generalisation to an mth order Markov chain) and takes values in the finite state space Ω_X. The components of θ are dependent on the particular general Hidden Markov model but typically consist of transition probabilities for the underlying Markov chain, and parameters relating to the emission density. For good overviews of HMMs, we refer the reader to MacDonald and Zucchini (1997); Cappé et al. (2005).

A common definition within an HMM framework is that a change point has occurred at time t whenever there is a change in the underlying chain, that is $x_{t-1} \neq x_t$. This definition is currently adopted in existing works such as Chib (1998); Hamilton (1989); Durbin et al. (1998); Fearnhead (2006). However, we consider a slightly more general definition; a change point to a regime occurs at time t when the change in the underlying chain persists for at least k time periods. That is $x_{t-1} \neq x_t = \ldots = x_{t+j}$ where $j \geq k-1$, and the term "regime" refers more specifically to observations from the sustained movement in the underlying chain. Although this definition can be interpreted as an instance of the simpler definition defined on a suitably expanded space, it is both easier to interpret and computationally convenient to make use of this explicit form. The motivation for this generalised definition is that there are several applications and scenarios in which a sustained change is required before a change to a new regime is said to have occurred. Typical examples include Economics where a recession is said to have occurred when there are at least two consecutive negative growth (contraction) states and thus $k = 2$, or in Genetics where a specific genetic phenomena, for example a CpG island (Aston and Martin, 2007), is at least a few hundred bases long (for example $k = 1000$) before being deemed in progress. The standard change point definition can be recovered by setting $k = 1$.

Interest often lies in determining the time of a change point and the number of change points occurring within a time series. Let $M^{(k)}$ and $\tau^{(k)} = (\tau_1^{(k)}, \ldots, \tau_{M^{(k)}}^{(k)})$ be variables denoting the number and times of change points, respectively. Given a vector $\tau^{(k)}$ we use $t \in \tau^{(k)}$ to indicate that one of the elements of $\tau^{(k)}$ is equal to t: if $t \in \tau^{(k)}$, then $\exists j \in \{1, \ldots, M^{(k)}\}$ such that $\tau_j^{(k)} = t$. The goal of this paper is quantifying the uncertainty of these characteristics by estimating:

$$P(M^{(k)} = m|y_{1:n}) \qquad\qquad m = 0, 1, 2, 3, \ldots, \tag{2}$$

$$\text{and } P\left(\tau^{(k)} \ni t|y_{1:n}\right) \tag{3}$$

where $P(\tau^{(k)} \ni t|y_{1:n}) = \sum_m P(M^{(k)} = m|y_{1:n}) \sum_{i=1}^m P(\tau_i^{(k)} = t|y_{1:n}, M^{(k)} = m)$. That is, the probability distribution of the number of changes and the marginal posterior probability that a change point occurs at any particular time.

2.1 Exact Change Point Distributions using Finite Markov Chain Imbedding

Under this generalised change point setting and conditioned on a particular model parameter setting θ, it is possible to compute exact distributions regarding change point characteristics (Aston et al., 2009). That is, it is possible to compute $P(\tau^{(k)} \ni t|y_{1:n}, \theta)$ and $P(M^{(k)} = m|y_{1:n}, \theta)$ exactly, where exact means that they are not subject to sampling or approximation error.

The generalised definition of a change point consequently motivates that we are looking for runs of a minimum length k in the underlying chain, where a run of length k in state $s \in \Omega_X$ is k consecutive occurrences of s. That is, $x_t = s = x_{t+1} = \ldots = x_{t+k-1}$, and in this instance, if $x_{t-1} \neq s$ the run of desired length k has occurred at time $t + k - 1$. Thus in order to consider whether a change point has occurred by time t, we can reformulate this problem to whether a run of length exactly k has occurred at time $t + k - 1$ in the underlying chain.

One popular approach for analysing behaviour in the underlying state sequence for HMMs is to provide an estimate of the underlying state sequence using techniques such as the Viterbi algorithm (Viterbi, 1967) and posterior decoding (Baum et al., 1970). These provide the most probable state sequence and the sequence which maximises a marginal probability of the states at each time respectively. Subsequent inference is often performed conditioned upon these point estimates which are subsequently assumed to be known — all runs and pattern statistics are derived conditional upon the estimated parameter values and the given sequence. This approach fails to capture the uncertainty arising from the unknown latent state sequence and thus for the run and pattern statistics of interest (as all inference is based on this single state sequence estimate), leading to a systematic underestimation of the attendant uncertainty. In addition, posterior decoding can produce estimates which feature impossible state transitions due to its reliance upon marginal distributions. We consider an alternative approach: in order to quantify fully the uncertainty of change points it is necessary to consider all possible state sequences. This is achieved by computing time-inhomogeneous transition probabilities with respect to the observed time series, $p(x_t|x_{t-1}, y_{1:n}), \forall t = 1, \ldots, n$, which can be obtained from smoothing probabilities. This thus allows us to quantify the uncertainty regarding runs in the underlying Markov chain and ultimately the change points themselves.

Let $\tau_u^{(k)}$ denote the time of the uth change point with $u \geq 1$. We can decompose the change point probability of interest into:

$$P(\tau^{(k)} \ni t|y_{1:n}, \theta) = \sum_m P(M^{(k)} = m) \sum_{u=1}^{m} P(\tau_u^{(k)} = t|M^{(k)} = m, y_{1:n}) \tag{4}$$

$$= \sum_{u=1,2,\ldots} P(\tau_u^{(k)} = t, M^{(k)} \geq u|y_{1:n}, \theta). \tag{5}$$

The event of the uth change point occurring at time t can be re-expressed as a quantity involving runs, namely whether the uth run of minimum length k has occurred at time $t+k-1$. Let $W_s(k,u)$ denote the waiting time for the uth occurrence of a run of minimum length k in state $s \in \Omega_X$. Thus $W_s(k,u) = t$ denotes that at time t, the uth occurrence of a such a run occurs. $W(k,u)$ similarly denotes the waiting time for the uth occurrence of a run in any state $s \in \Omega_X$ of at least length k. If change points into a certain regime were of interest, $W_s(k,u)$ where $s \in \Omega_X$ is the state defining the regime of interest, is of greater interest. By re-expressing the uth change point event as the waiting time for the uth occurrence of a run, it is thus possible to compute the corresponding probabilities:

$$P(\tau_u^{(k)} = t|y_{1:n}, \theta) = P(W(k,u) = t + k - 1|y_{1:n}, \theta). \tag{6}$$

It is possible to compute exactly the distribution of waiting time statistics, namely $P(W(k,u) \leq t|\theta, y_{1:n})$, via Finite Markov Chain Imbedding (FMCI) (Fu and Koutras, 1994; Fu and Lou, 2003).

FMCI introduces several auxiliary Markov processes, $\{Z_t^{(1)}, Z_t^{(2)}, Z_t^{(3)}, \ldots\}$ which are defined over the common state space $\Omega_Z^{(k)} = \Omega_X \times \{-1, 0, 1, \ldots, k\}$. $\Omega_Z^{(k)}$ is an expanded version of Ω_X which consists of tuples (X_t, j) where the new variable $j = -1, 0, 1, 2 \ldots, k$ indicates the progress of any potential runs. The auxiliary processes are constructed such that the uth process corresponds to the conditional Markov chain for finding a run of length k, conditional on the fact that $u - 1$ runs of length at least k have already occurred multiplied by the conditional probability of $u - 1$ runs having occurred.

The states of the auxiliary Markov chains can loosely be categorised into three categories: continuation ($j = -1$), run in progress ($j = 0, 1, 2, \ldots, k - 1$) and absorption ($j = k$). Absorption states denote that the run of required length has occurred, the run in progress states are fairly self explanatory, and continuation states denote when the $(u - 1)$th run is still in progress (its length exceeds the required length of k) and needs to end before the occurrence of the new uth run can be considered. The transition probabilities of these auxiliary Markov chains $\{Z_t^{(1)}, Z_t^{(2)}, Z_t^{(3)}, \ldots\}$ are obtained deterministically from those of the original Markov chain $\{X_t\}$. In an HMM framework, the time-inhomogeneous posterior transition probabilities are used in order to account for all possible state sequences given the observed time series.

Thus in order to determine whether the specific occurrence of a run has occurred by a specific time, we simply need to determine if the corresponding auxiliary Markov chain has reached the absorption set, A, the set of all absorption states, by the specified time. The corresponding probability can thus be computed by standard Markov chain results. This leads to computing the probability of the uth change point probability.

$$P(W(k, u) \leq t + k - 1 | y_{1:n}, \theta) = P(Z_{t+k-1}^{(u)} \in A | y_{1:n}, \theta) \tag{7}$$

$$P(\tau_u^{(k)} = t | y_{1:n}, \theta) = P(W(k, u) = t + k - 1 | y_{1:n}, \theta) \tag{8}$$

$$= P(W(k, u) \leq t + k - 1 | y_{1:n}, \theta) - P(W(k, u) \leq t + k - 2 | y_{1:n}, \theta). \tag{9}$$

The distribution of the number of change points can also be computed from these waiting time distributions:

$$P(M^{(k)} = m | y_{1:n}, \theta) = P(W(k, u) \leq n | y_{1:n}, \theta) - P(W(k, u + 1) \leq n | y_{1:n}, \theta) \tag{10}$$

In general, this FMCI approach allows for exact computation of distributions for other change point characteristics such as the probability of a change within a given time interval and the distribution of the regime durations. This thus provides a flexible methodology in capturing the uncertainty of change point problems.

However, these distributions of change point characteristics are conditioned on the model parameters θ. However, it is typical for θ to be unknown, and subject to error and uncertainty (for example estimation error). Thus, in order fully consider uncertainty in change points, it is necessary to also consider the uncertainty of the parameters. We can account for model parameters via the use of Sequential Monte Carlo samplers.

2.2 Sequential Monte Carlo samplers

In order to deal with parameter uncertainty, we adopt a Bayesian approach by integrating out the model parameters to obtain a marginal posterior distribution on the change point quantities alone. However, it is not feasible to perform this integration analytically for the models of interest.

Sequential Monte Carlo (SMC) methods are a class of simulation algorithms for sampling from a sequence of related distributions, $\{\pi_b\}_{b=1}^B$, via importance sampling and resampling techniques. Common applications of these methods in Statistics, Engineering and related disciplines include sampling from a sequence of posteriors as data becomes available and the particle filter for approximating the optimal filter (to obtain the distribution of the underlying state sequence as observations become available) in general (typically continuous) state space nonlinear and non-Gaussian HMMs (Gordon et al., 1993); see Doucet and Johansen (2011) for a recent survey. We do not use SMC to infer on the underlying state sequence in our particular context because the state sequence is ultimately of little interest to us and we can calculate quantities of interest marginally.

The standard application of SMC techniques requires that the sequence of distributions of interest are defined upon a sequence of increasing state spaces and that one is interested in only particular marginal distributions. Sequential Monte Carlo samplers (Del Moral et al., 2006) are a class of SMC algorithms in which a collection of auxiliary distributions are introduced to allow the SMC technique to be applied to essentially arbitrary sequences of distributions defined over any sequence of spaces. One use of this framework is to allow SMC to be used when one has a sequence of related distributions defined over a common space. The innovation is to expand the space under consideration and introduce auxiliary distributions which admit the distributions of interest as marginals. This is done by the introduction of a collection of Markov kernels, $\{L_b\}$ with distributions of interest $\{\pi_b(x_b)\}$ being formally augmented with these Markov kernels to produce $\{\widetilde{\pi}_b\}$ with $\widetilde{\pi}_b(x_{1:b}) := \pi_b(x_b)\prod_{j=1}^{b-1} L_j(x_{j+1}, x_j)$.

Given a weighted sample $\{W_{b-1}^i, \theta_{b-1}^i\}$ which is properly weighted to target $\pi_{b-1}(\theta_{b-1})$ the SMC sampler with proposal kernel $K_b(\theta_{b-1}^i, \theta_b^i)$ is used leading to a sample $\{W_{b-1}^i, (\theta_{b-1}^i, \theta_b^i)\}$ which is properly weighted for the distribution $\pi_{b-1}(\theta_{b-1}^i)K_b(\theta_{b-1}^i, \theta_b^i)$. Given any backward kernel, $L_{b-1}(\theta_b, \theta_{b-1})$ which satisfies an appropriate absolute continuity requirement, one can adjust the weights of the sample such that it is instead properly weighted to target the distribution $\pi_b(\theta_b)L_{b-1}(\theta_b, \theta_{b-1})$ by multiplying those weights by an appropriate incremental weight (setting $W_b^i \propto W_{b-1}^i \cdot \widetilde{w}_b(\theta_{b-1}^i, \theta_b^i)$). These incremental weights are

$$\widetilde{w}_b(\theta_{b-1}^i, \theta_b^i) = \frac{\pi_b(\theta_b^i)L_{b-1}(\theta_b^i, \theta_{b-1}^i)}{\pi_{b-1}(\theta_{b-1}^i)K_b(\theta_{b-1}^i, \theta_b^i)}, \tag{11}$$

where $L_{b-1}(\theta_b^i, \theta_{b-1}^i)$ is a backwards Markov kernel. Del Moral et al. (2006) established that the optimal choice of backward kernel, if resampling is conducted every iteration, is

$$L_{b-1}^{\text{opt}}(\theta_b, \theta_{b-1}) = \frac{\pi_{b-1}(\theta_{b-1})K_b(\theta_{b-1}, \theta_b)}{\int \pi_{b-1}(\theta'_{b-1})K_b(\theta'_{b-1}, \theta_b)d\theta'_{b-1}}$$

the integral in the denominator is generally intractable and it is necessary to find approximations (the use of which increases the variance of the estimator but does not introduce any further approximation).

When π_b-invariant MCMC kernels are used for K_b a widely-used approximation of this optimal quantity can be obtained by noting that consecutive distributions in the sequence are in some sense similar, $\pi_{b-1} \approx \pi_b$ and by replacing π_{b-1} with π_b in the optimal backward kernel, we obtain:

$$L_{b-1}^{\mathrm{tr}}(\theta_b, \theta_{b-1}) = \frac{\pi_b(\theta_{b-1})K_b(\theta_{b-1}, \theta_b)}{\int \pi_b(\theta'_{b-1})K_b(\theta'_{b-1}, \theta_b)d\theta'_b} = \frac{\pi_b(\theta_{b-1})K(\theta_{b-1}, \theta_b)}{\pi_b(\theta_b)}$$

by the π_b-invariance of K_b. This leads to the convenient incremental weight expression:

$$\widetilde{w}_b(\theta_{b-1}^i, \theta_b^i) = \frac{\pi_b(\theta_{b-1}^i)}{\pi_{b-1}(\theta_{b-1}^i)}. \tag{12}$$

A standard use of this framework is to provide samples from a complex distribution by sampling first from a tractable distribution and then employing mutation and selection operations to provide a sample which is appropriately weighted for approximating a complex, intractable distribution of interest. This particular application, with no selection coincides with the Annealed Importance Sampling algorithm of Neal (2001).

In the change point problems described here, the objective is to approximate the posterior distribution of the model parameters, $p(\theta|y_{1:n})$. This can be done via SMC, sampling initially from the prior $\pi_1 = p(\theta)$ and defining the subsequent distributions as:

$$\pi_b(\theta) \propto p(\theta)p(y_{1:n}|\theta)^{\gamma_b}, \tag{13}$$

where $\{\gamma_b\}_{b=1}^B$ is a non-decreasing sequence with $\gamma_1 = 0$ and $\gamma_B = 1$. This has the effect of introducing the likelihood gradually such that π_1 can be sampled from easily, π_{b+1} is similar to π_b and $\pi_B(\theta) = p(\theta|y_{1:n})$ is the distribution of interest. Algorithm 1 shows a generic SMC sampler for problems of this sort.

Resampling alleviates the problem of weight degeneracy in which the variance of weights becomes too large and the approximation of the distribution does not remain accurate. Intuitively, resampling eliminates samples with small weights and replicates those with larger weights stochastically so as to preserve the expectation of the approximation of the integral of any bounded function. Formally, if $\{W^i, \theta^i\}_{i=1}^N$ is a weighted sample, then resampling consists of drawing a collection $\{\widetilde{\theta}^i\}_{i=1}^N$ such that: $\mathbb{E}[\frac{1}{N}\sum_{i=1}^N \varphi(\widetilde{\theta}^i)|\{W^i, \theta^i\}_{i=1}^N] = \sum_{i=1}^N W^i\varphi(\theta^i)$ for any bounded measurable φ. The simplest approach, termed multinomial resampling (as it is equivalent to drawing the number of replicates of each sample from a multinomial distribution with parameters N and $(W^1, \ldots, W^N)$), simply draws N samples with replacement from the weighted empirical distribution associated with the existing sample set; this approach unnecessarily increases the Monte Carlo variance and several other techniques are preferable. A comparison of resampling schemes is provided by Douc and Cappé (2005).

Whilst resampling is beneficial in the long run, resampling too often is not desired since it introduces unnecessary Monte Carlo variance and thus a dynamic resampling scheme where we only resample when necessary, is often implemented. This can be implemented by determining the Effective Sample Size (ESS) which is associated to the variance of the importance weights, and resampling when the ESS is below a pre-specified threshold T. Obtained via Taylor expansion of the variance of associated estimates (Kong et al., 1994), ESS serves as a proxy for the variance of the importance weights. It is computed

Algorithm 1 SMC Sampler for Bayesian Inference (Del Moral et al., 2006)

```
Step 1: Initialisation Set $b = 1$
for $i = 1, \dots, N$ do
    Draw $\theta_1^i \sim \eta_1$ ($\eta_1$ is a tractable importance distribution for $\pi_1$).
    Compute the corresponding importance weight $\{w_1(\theta_1^i)\} \propto \pi_1(\theta_1^i)/\eta_1(\theta_1^i)$.
end for
Normalise these weights, for each $i$:
    $W_1^i = \frac{w_1(\theta_1^i)}{\sum_{j=1}^N w_1(\theta_1^j)}.$

Step 2: Selection
If degeneracy is too severe (e.g. ESS $< N/2$), then resample and set $W_b^i = 1/N$.

Step 3: Mutation Set $b \leftarrow b + 1$.
for $i = 1, \dots, N$ do
    Draw $\theta_b^i \sim K_b(\theta_{b-1}^i, \cdot)$, (a $\pi_b$-invariant Markov kernel)
    Compute the incremental weights:
        $\left\{ \widetilde{w}_b \left(\theta_{b-1}^i, \theta_b^i\right) = \frac{\pi_b(\theta_{b-1}^i)}{\pi_{b-1}(\theta_{b-1}^i)} \right\}_{i=1}^N.$
end for
Compute the new normalised importance weights:
    $W_b^i = W_{b-1}^i \widetilde{w}_b(\theta_{b-1}^i, \theta_b^i) \Big/ \sum_{j=1}^N W_{b-1}^j \widetilde{w}_b(\theta_{b-1}^j, \theta_b^j). \qquad (14)$
if $b < B$ then
    Go to step 2
end if
```

via $ESS = \{\sum_{i=1}^{N}(W^i)^2\}^{-1}$. The criterion provides an approximation of the number of independent samples from the target distribution, π_b, that would provide an estimate of comparable variance. We resample if the ESS falls below a the threshold, $T = N/2$. Resampling at such stopping times rather than deterministic times is valid and it has recently been demonstrated that convergence results can be extended to this case (Del Moral et al., 2011).

We note that the resampling procedure is usually performed after the mutation and reweighting step of samples. However, given that the incremental weights (Equation 12) are only dependent on the sample from the previous iteration, θ_{b-1}^i, and thus the importance weights of the new particles are independent of the new location, θ_b^i, it is possible to resample prior to the mutation step. Resampling before the mutation step thus ultimately leads to greater diversity of the resulting sample, compared to performing it afterwards.

Of course, other sampling strategies could be employed. These can be divided into two categories: those which simulate the latent state sequence and those which work directly on the marginal distribution of the model parameters. We have found that SMC provides robust estimation in the setting of interest. Markov Chain Monte Carlo (Gilks et al., 1996) provides the most common strategy for approximating complex posterior distributions in Bayesian inference. As MCMC involves constructing an ergodic Markov chain which explores the posterior distribution, it would require the design of a π-invariant Markov transition with good global mixing properties. As our marginal posterior is typical multimodal, we found it difficult to obtain reasonable performance with such a strategy; significant application-specific tuning or the design of sophisticated proposal kernels would be necessary to achieve acceptable performance. In principle, a data augmentation strategy in which the latent variables are also sampled could be implemented, but the correlation of the latent state sequence with itself and the parameter vectors would make it difficult to obtain fast mixing. Particle MCMC (Andrieu et al., 2010) justifies the use of SMC algorithms within MCMC algorithms to provide high-dimensional proposals; its use in change point problems has already been investigated and appears promising (Whiteley et al., 2009). In more general settings than that considered here, in which it is not possible to numerically integrate-out the underlying state sequence (or in situations in which that state sequence is of independent interest) this seems a sensible strategy.

The design of an efficient SMC algorithm for our particular problem is discussed in Section 3 and its application to some real problems in Section 4.

3 Methodology

The main quantities of interest in change point problems are often the posterior probability of a change point occurring at a certain time, $P(\tau^{(k)} \ni t|y_{1:n})$, and the posterior distribution of the number of change points, $P(M^{(k)} = m|y_{1:n})$. In order to obtain these two quantities of interest, it can be seen as integrating out the model parameters, θ, and manipulating as follows:

$$P(\tau^{(k)} \ni t|y_{1:n}) = \int_\theta P(\tau^{(k)} \ni t, \theta|y_{1:n})d\theta = \int_\theta P(\tau^{(k)} \ni t|\theta, y_{1:n})p(\theta|y_{1:n})d\theta, \qquad (15)$$

in the case of the posterior probability of a change point at a specific time. A similar expression can be obtained for the distribution of the number of change point. We focus on the posterior change point probability throughout this section; the number of change points can be dealt with analogously.

Equation 15 highlights that we can replace the joint posterior probability of the change points and model parameters, by the product of two familiar quantities; $P(\tau^{(k)} \ni t|\theta, y_{1:n})$, the change point probability conditioned θ and $p(\theta|y_{1:n})$, the posterior of the model parameters. We have shown in Section 2.1 that it is possible to compute exactly $P(\tau^{(k)} \ni t|\theta, y_{1:n})$ via the use of FMCI in an HMM setting. However, it is not generally possible to evaluate the right hand side of Equation 15 and so numerical and simulation based approaches need to be considered.

Viewing this integral as an expectation under $p(\theta|y_{1:n})$,

$$P(\tau^{(k)} \ni t|y_{1:n}) = \mathbb{E}_{p(\theta|y_{1:n})}[P(\tau^{(k)} \ni t|\theta, y_{1:n})], \qquad (16)$$

reduces estimation for the distribution of interest to a standard Monte Carlo approximation of this expectation and allows standard SMC convergence results can be applied.

We can view this as a Rao-Blackwellised version of the estimator one would obtain by simulating both the latent state sequence and the parameters from their joint posterior distribution. By replacing this estimator with its conditional expectation given the sampled parameters, the variance can only be reduced by the Rao-Blackwell theorem (see, for example, (Lehmann and Casella, 1998, Theorem 7.8)).

Thus, given that we can approximate the posterior of the model parameters $p(\theta|y_{1:n})$ by a cloud of N weighted samples $\{\theta^i, W^i\}_{i=1}^N$ via SMC samplers, we can approximate Equation 15 and 16 by

$$P(\tau^{(k)} \ni t|y_{1:n}) \approx \widehat{P^N}(\tau^{(k)} \ni t|y_{1:n}) = \sum_{i=1}^N W^i P(\tau^{(k)} \ni t|\theta^i, y_{1:n}). \qquad (17)$$

The proposed methodology is to approximate the model parameter posterior via the previously discussed SMC samplers in Section 2.2, before computing the exact change point distributions conditional on each of the parameter samples approximating the model parameter posterior. In order to obtain the general change point distribution of interest, we thus take the weighted average of these exact distributions.

An alternative Monte Carlo approach to the evaluation of Equation 15 is via data augmentation. This involves sampling from the joint posterior distribution of the model parameters and the underlying state sequence (see for example Chib (1998); Fearnhead (2006); Fearnhead and Liu (2007)). However, it is not necessary to sample the entire underlying state sequence in order to compute the change point quantities of interest. In addition, due to the high dimensionality of this state sequence, it is often difficult to design good MCMC moves to ensure that the chain mixes well. Our methodology has the advantage that we do not need to sample this underlying state sequence and has the advantage that we introduce Monte Carlo error only on the model parameters. This thus retains the exactness of the change point distributions when conditioned on model parameters. In addition, parameter estimation is available within our methodology via inference on the approximated version of the parameter posterior. This estimation does not require knowledge of the underlying state sequence, and remains invariant with regards to change point behaviour.

The general procedure of our algorithm is displayed in Algorithm 2.

Algorithm 2 SMC algorithm for quantifying the uncertainty in change points.

```
Approximating $p(\theta|y_{1:n})$
Initialisation: Sample from prior, $p(\theta)$, $b = 1$
for $i = 1, \ldots, N$ do
    Sample $\theta_1^i \sim q_1$.
end for
Compute for each $i$
    $$W_1^i = \frac{w_1(\theta_1^i)}{\sum_{i=1}^N w_1(\theta_1^i)} \qquad \text{where } w_1(\theta_1) = \frac{p(\theta_1)}{q(\theta_1)} \tag{18}$$
if $ESS < T$ then
    Resample
end if
for $b = 2, \ldots, B$ do
    Reweighting:
    For each $i$ compute
        $$W_b^i = \frac{W_{b-1}^i \widetilde{w}_b(\theta_{b-1}^i, \theta_b^i)}{\sum_{i=1}^N W_{b-1}^i \widetilde{w}_b(\theta_{b-1}^i, \theta_b^i)} \tag{19}$$
        $$\text{where } \widetilde{w}_b(\theta_{b-1}^i, \theta_b^i) = \frac{\pi_b(\theta_{b-1}^i)}{\pi_{b-1}(\theta_{b-1}^i)} = \frac{p(y_{1:n}|\theta_{b-1}^i)^{\gamma_b}}{p(y_{1:n}|\theta_{b-1}^i)^{\gamma_{b-1}}}. \tag{20}$$
    Selection:
    if  $ESS < T$ then Resample.
    Mutation:
    For each $i = 1, \ldots, N$
    Sample $\theta_b^i \sim K_b(\theta_{b-1}^i, \cdot)$ where $K_b$ is a $\pi_b$ invariant Markov kernel.
end for

Obtaining the change point estimates of interest using FMCI
Using,
    $$p(\theta|y_{1:n})d\theta = \pi_B(d\theta) \approx \sum_{i=1}^N W_B^i \delta_{\theta_B^i}(d\theta),$$
yields:
    $$\widehat{P}(\tau^{(k)} \ni t|y_{1:n}) = \sum_{i=1}^N W^i P(\tau^{(k)} \ni t|y_{1:n}, \theta_B^i) \tag{21}$$
    $$\widehat{P}(M^{(k)} = m|y_{1:n}) = \sum_{i=1}^N W^i P(M^{(k)} = m|y_{1:n}, \theta_B^i) \tag{22}$$
where $P(\tau^{(k)} \ni t|y_{1:n}, \theta_B^i)$ and $P(M^{(k)} = m|y_{1:n}, \theta_B^i)$ can be computed exactly via FMCI.
```

3.1 Approximating the model parameter posterior $p(\theta|y_{1:n})$

As mentioned previously, we aim to approximate the model parameter posterior $p(\theta|y_{1:n})$ via an SMC sampler and define the sequence of distributions

$$\pi_b(\theta) \propto p(y_{1:n}|\theta)^{\gamma_b} p(\theta), \tag{23}$$

where $p(\theta)$ denotes the prior on the model parameters and $p(y_{1:n}|\theta)$ the likelihood. There is great flexibility in the choice of non-decreasing tempering schedule, $\{\gamma_b\}_{b=1}^{B}$ with $\gamma_1 = 0$ and $\gamma_B = 1$, ranging from a simple linear sequence, where $\gamma_b = \frac{b-1}{B-1}$ for $b = 1, \ldots, B$, to more sophisticated tempering schedules. We approximate each distribution by a weighted cloud of N samples, denoted by $\{\theta_b^i, W_b^i\}_{i=1}^{N}$. As the weighted cloud of samples approximating the posterior is ultimately of interest, we simplify the notation by dropping the subscript as follows, $\{\theta^i, W^i\}_{i=1}^{N} \equiv \{\theta_B^i, W_B^i\}_{i=1}^{N}$

Dependent on the particular class of general HMM considered, the specifics of the SMC algorithm differ. We partition θ into $\theta = (\mathbf{P}, \eta)$ where $\mathbf{P}$ denotes the transition probability matrix and η represents the parameters for the emission distributions. As $\mathbf{P}$ is a standard component in HMMs, we discuss a general implementation for it within our SMC algorithm. We discuss a specific approach to η, the emission parameters, for a particular model in Section 4.

3.1.1 Intialisation

The first stage of our SMC algorithm is to sample from an initial tractable distribution, $\pi_1 = p(\theta)$, either directly or via importance sampling. Following Chopin (2007), we see no reason to assume a dependency structure between the transition and emission parameter sets and assume prior independence amongst the emission parameters and the transition probabilities,

$$p(\theta) = p(\eta)p(\mathbf{P}). \tag{24}$$

We further assume prior independence amongst the rows of the transition probability matrix and impose an independent Dirichlet prior on each such row:

$$p(\mathbf{P}) = \prod_{h=1}^{H} p(p_h) \tag{25}$$

$$p(p_h) \sim \text{Dirichlet}_H(\alpha_h), \qquad h = 1, \ldots H \tag{26}$$

where p_h denotes row h of the transition matrix and $\alpha_h = (\alpha_{h1}, \ldots, \alpha_{hH})$ are the corresponding hyperparameters. As HMMs are often used in scenarios where the underlying chain does not switch underlying states often and thus there is a persistent nature, we typically assume an asymmetric Dirichlet prior on the transition probabilities which favours configurations in which the latent state process remains in the same state for a significant number of iterations. We thus choose our hyperparameters to reflect this.

There is also considerable flexibility when implementing the sampling from the prior of the emission parameters η. In the present work we assume that the components are independent *a priori*. Our general approach when choosing priors and their associated hyperparameters has been to use priors which are not very informative over the range of values which are observed in the applications which we have encountered. The methodology which we develop is flexible and the use of other priors should not present substantial difficulties if this were necessary in another context. In the settings we are investigating, the likelihood typically needs to provide most of the information in the posterior as prior information is often sparse. As ever, informative priors could be employed if they were available; this would require no more than some tuning of the SMC proposal mechanism.

We can sample directly from the prior described above by sampling from standard distributions for each of the components, this consequently means the importance weights of the associated model parameter samples, $\{\theta_1^i\}_{i=1}^N$, are all equally weighted, $W_1^i = \frac{1}{N}, \quad i = 1, \ldots, N$. More generally, we could use importance sampling: if q_1 is the instrumental density that we use during the first iteration of the algorithm, then the importance weights are of the form $W_1^i \propto \frac{p(\theta_1^i)}{q(\theta_1^i)}$. Regardless of how we obtain this weighted sample, we have a weighted cloud of N samples, $\{\theta_1^i, W_1^i\}_{i=1}^N$, which approximates the prior distribution $\pi_1 = p(\theta)$.

3.1.2 Approximating π_b, given weighted samples approximating π_{b-1}

Having obtained a weighted sample approximation of distribution π_{b-1}, $\{\theta_{b-1}^i, W_{b-1}^i\}_{i=1}^N$, it is necessary to mutate and weight it to properly approximate π_b. We can achieve this by mutating existing samples with a π_b-invariant Markov kernel, $K_b(\theta_{b-1}^i, \cdot)$. There is a great deal of of flexibility in this mutation step — essentially any MCMC kernel can be used, including Gibbs and Metropolis Hastings kernels, as well as mixtures and compositions of these.

As in an MCMC setting, it is desirable to update highly dependent components of the parameter vector jointly. We update $\mathbf{P}$ and η, sequentially. The row vectors $p_h, h = 1, \ldots, H$ can be mutated via a Random Walk Metropolis (RWM) strategy on a logit scale (which ensures that the sampled values remain within the appropriate domain). In some settings it may be necessary to block the row vectors together and mutate them simultaneously. This is discussed in section 4.

Having obtained $\theta_b^i, i = 1, \ldots, N$, from θ_{b-1}^i it is necessary to re-weight the sample so that they properly approximate the new distribution π_b. The new unnormalised importance weights can be obtained via the equation

$$w_b(\theta_b^i) = W_{b-1}^i \widetilde{w}_b(\theta_{b-1}^i, \theta_b^i), \tag{27}$$

where $\widetilde{w}_b(\theta_{b-1}^i, \theta_b^i) = \frac{p(y_{1:n}|\theta_{b-1}^i)^{\gamma_b}}{p(y_{1:n}|\theta_{b-1}^i)^{\gamma_{b-1}}}$ by substituting π_{b-1} and π_b into Equation 12. Note that the incremental weights do not depend on the new mutated particle θ_b^i, allowing resampling to be performed before this sampling step.

We have thus obtained a new collection of weighted samples $\{\theta_b^i, W_b^i\}_{i=1}^N$ which approximates the distribution π_b, by using the existing approximation of π_{b-1}.

4 Applications

The following section applies the proposed methodology of Section 3 to simulated and real data. We consider data generated by Hamilton's Markov Switching Autoregressive model of order r, MS-AR(r) (Hamilton, 1989). The model for the observation at time t, y_t, is defined as,

$$y_t = \mu_{x_t} + a_t \tag{28}$$

$$a_t = \phi_1 a_{t-1} + \ldots + \phi_r a_{t-r} + \epsilon_t \qquad \epsilon_t \sim N(0, \sigma^2), \tag{29}$$

where the underlying mean μ, switches according to the underlying hidden state x_t, and y_t is dependent on previous r observations in this autoregressive manner using the associated parameters $\phi_1, \ldots, \phi_r$. ϵ_t is additional Gaussian white noise with mean 0 and variance σ^2. The emission density for this model is thus

$$\begin{aligned} f(y_t|X_{1:t}, y_{1:t-1}, \theta) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(a_t - \left(\sum_{j=1}^{r} \phi_j a_{t-j}\right)\right)\right) \qquad (30) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}((y_t - \mu_{X_t}) - \left(\sum_{j=1}^{r} \phi_j \left(y_{t-j} - \mu_{X_{t-j}}\right)\right)\right). \end{aligned}$$

Notice that Y_t is dependent on the previous r underlying states of the Markov chain, $X_{t-r:t}$, in addition to the observations, $y_{t-r:t-1}$. Hamilton's MS-AR(r) is commonly used in Econometrics in modelling the business cycles within GNP data (Hamilton, 1989) and in Biology for modelling functional Magnetic Resonance Imaging data (fMRI) (Peng et al., 2011) for example. We consider in particular a 2-state Hamilton's MS-AR model of order 1, MS-AR(1) which has been shown to be useful in modelling fMRI data (Peng et al., 2011). The model parameters to be estimated are thus the transition probabilities, state dependent means, global precision and AR parameter, $\theta = (p_{11}, p_{22}, \mu_1, \mu_2, \lambda = 1/\sigma^2, \phi_1)$. It is more convenient to work with the precision than directly with the variance.

4.1 Implementation for 2-state MS-AR(1) model

In the absence of substantial prior knowledge concerning the parameters, we assume that there is no correlation structure between the emission parameters and thus assume independence between the emission parameters themselves. We employ the following prior distributions for the parameters:

$$\begin{aligned} \mu_1 &\sim \text{N}(0, \sigma^2_{\mu_1} = 50) & \mu_2 &\sim \text{N}(-1, \sigma^2_{\mu_2} = 50) \qquad (31) \\ \lambda &\sim \text{Gamma}(\text{shape} = 5, \text{scale} = 2) & \phi_1 &\sim \text{Unif}(-1, 1) \end{aligned}$$

Of course, other priors could be implemented, dependent on one's belief about the parameters. Nevertheless, these prior distributions have been chosen with respect to our belief and the domain of the parameters. To obtain interpretable results we introduce the constraint $\mu_1 < \mu_2$, which can be viewed as specifying a joint prior distribution proportional to $\text{N}(\mu_1; 0, \sigma^2_{\mu_1})\text{N}(\mu_2; -1, \sigma^2_{\mu_2})I_{(\mu_1,\infty)}(\mu_2)$ where $I_A(x)$ denotes the indicator function on set A evaluated at x. We also expect stationarity and invertibility within regimes, in the sense of a constant second order structure, and as no additional information is provided on AR parameter ϕ_1, we consequently assume a uniform prior on the interval $(-1, 1)$ for ϕ_1. This is the default prior as in Huerta and West (1999), and our methodology is flexible enough to permit non-uniform priors on this interval for ϕ_1 if necessary.

As mentioned previously in section 3, we assume an asymmetric Dirichlet prior for the transition probabilities such that transition matrices which lead to periods in which a particular state persists are favoured a priori. Using the benchmark that the majority of mass should be placed in the $(0.5, 1)$ interval similar to that of Albert and Chib (1993), we employed the following priors in this particular case.

$$p_{11} \sim \text{Beta}(3, 1) \qquad p_{22} \sim \text{Beta}(3, 1) \tag{32}$$

We mutate current samples, θ via a RWM proposal applied to components of the sample according to the following mutation strategy:

1. Mutate p_{11}, p_{22} simultaneously via RWM on a logit scale, with some specified correlation structure. That is, proposals for the transition probabilities, $p^\star_{11}, p^\star_{22}$ are performed as follows:

$$\begin{bmatrix} l^\star_{11} = \log\left(\frac{p^\star_{11}}{1-p^\star_{11}}\right) \\ l^\star_{22} = \log\left(\frac{p^\star_{22}}{1-p^\star_{22}}\right) \end{bmatrix} \sim \text{N}\left(\begin{bmatrix} l_{11} = \log\left(\frac{p_{11}}{1-p_{11}}\right) \\ l_{22} = \log\left(\frac{p_{22}}{1-p_{22}}\right) \end{bmatrix}, \Sigma = \begin{bmatrix} \sigma^2_p & \rho_p \\ \rho_p & \sigma^2_p \end{bmatrix}\right), \tag{33}$$

   where σ^2_p is the proposal variance for the transition probabilities, and ρ_p is a specified covariance between l_{11} and l_{22}.

2. Mutate μ_1, μ_2 independently via RWM on the standard scale. That is, proposals, $\mu^\star_i$ are performed by

$$\mu^\star_i \sim \text{N}(\mu, \sigma^2_\mu) \qquad i = 1, 2, \tag{34}$$

   where σ^2_μ is the specified proposal variance for the means.

3. Mutate λ via RWM on a log-normal scale. Proposals, $\lambda^\star$ are thus performed via

$$\log(\lambda^\star) \sim \text{N}(\log(\lambda), \sigma^2_\lambda), \tag{35}$$

   where σ^2_λ is the specified proposal variance for the precision.

4. Mutate ϕ_1 by transforming onto the interval $(0, 1)$ and then performing RWM on a logit scale. That is, proposals $\phi^\star_1$ are obtained by sampling from interval $(-1, 1)$.

$$l^\star = \log\left(\frac{\phi^\star_1 + 1}{1 - \phi^\star_1}\right) \sim \text{N}\left(l = \log\left(\frac{\phi_1 + 1}{1 - \phi_1}\right), \sigma^2_{\phi_1}\right), \tag{36}$$

   where $\sigma^2_{\phi_1}$ is the proposal variance for the AR parameter.

We perform the mutation on subcomponents of θ independently of each other, using the most recent values of other subcomponents of θ. Note that this fits into the SMC framework described above with the proposal kernels K_b corresponding to the composition of a sequence of Metropolis-Hastings kernels and the associated backward kernel. We note that the RWM mutations are performed on different scales due to the differing domains of the parameters. To ensure good mixing, we mutated the transition probabilities simultaneously because we believe that their is a significant degree of *a posteriori* correlation between them.

As the values of p_{11} and p_{22} are closely related to the probable relative occupancy of the two regimes, it is expected that for given values of the other parameters there will be significant posterior correlation between these parameters (and also between l_{11} and l_{22}). In the current context, the two values were updated concurrently using a bivariate Gaussian random walk on the logit scale, with a positive correlation of $\rho_p = 0.75$.

In selecting proposal variances for each group of sub components, we have attempted to encourage good global exploration at the beginning, and then more localised exploration in any possible modes, towards the end of the algorithm and as we approach the target posterior distribution. This has been implemented by decreasing the effective proposal variance with respect to the iteration. The initial proposal variances used for each of the considered components are $\sigma_p^2 = 10, \sigma_\mu^2 = 10, \sigma_\lambda^2 = 5, \sigma_{\phi_1}^2 = 10$. We note that these proposal variances are not optimal and performance would be improved by further tuning (see Roberts et al. (1997) and related work for guidelines on optimal acceptance rates). However, these convenient choices demonstrate that adequate performance can be obtained *without* careful application-specific tuning.

The following results, both simulated and real, are obtained using $500 = N$ samples and $100 = B$ time steps taken to move from the initial prior distribution $\pi_1 = p(\theta)$ to the target posterior distribution $\pi_B = p(\theta|y_{1:n})$. A simple linear tempering schedule, $\gamma_b = \frac{b-1}{B-1}, \quad b = 1, \ldots, B$ was used to define the sequence of distributions. Systematic resampling (Carpenter et al., 1999) was was carried out whenever the ESS fell below $T = N/2$.

4.2 Simulated Data

The following results consider a variety of data where the AR parameter, ϕ_1, varies in value. We fix however the underlying state sequence and the values of the remaining parameters as follows: $p_{11} = 0.99, p_{22} = 0.99, \mu_1 = 0, \mu_2 = 1, \lambda = 16$. We consider a sequence of 200 observations and consider a variety of AR parameter values ranging from 0.5 to 0.9 where the location and number of change points becomes increasing less obvious.

Figure 1 displays the various simulated time series and the state sequence of the underlying Markov chain in addition to plots for the change point probabilities (left column) and the distribution of the number of change points (right column) obtained via our proposed SMC based algorithm. The latent state sequence is common to all of the simulated time series and is denoted by the dashed line superimposed on the simulated time series plot.

Our change point results consider changes into and out of regime 1 which is that with smaller mean for at least 2 time periods ($k = 2$ and $s = 1$ with an ordering constraint placed upon the mean parameters) . The change point probability (CPP) plots display the probability of switching into and out of this regime. In all simulated time series, there are two occurrences of this regime, starting at times of approximately 20 and 120, and ending at time 100 and continuing to the end of the data, respectively.

In all three time series considered, our results indicate that our proposed methodology works well with good detection and estimation for the change point characteristics of interest. Change point probabilities are centred around the true locations of the starts and ends of the regime of interest with a degree of concentration dependent upon the information contained in the data. The true number of regimes is the most probable in all three of the time series considered.

As ϕ_1 increases, the distribution of the change point characteristics becomes more diffuse. This is what would be expected as the data becomes less informative as ϕ_1 increases. This uncertainty is a feature of the model, not a deficiency of the inferential method, and it is important to account for it when performing estimation and prediction of related quantities. The proposed methodology is able to do this.

We also observe that the probability that there are no change points is not negligible for $\phi_1 = 0.75$ and for $\phi_1 = 0.90$. These results illustrate the necessity of accounting for parameter uncertainty in change point characterisation.

Table 1 displays the posterior means of the model parameter samples obtained via the SMC sampler. These are calculated by taking the weighted average of the weighted cloud of samples approximating the model parameter posterior distribution. In addition, we provide the weighted standard deviation of the cloud of samples. We observe that the posterior values are reasonably close to the true values used to generate the time series. We note that as ϕ_1 increases and consequently the data becomes less informative, less accurate estimates are provided with greater deviation from the true values and commensurate increase in standard deviation. Nevertheless, we observe that the model parameter posterior has been reasonably well approximated.

As a comparison, we also consider the exact change point distributions obtained by conditioning on these posterior means. We observe from the corresponding plots in Figure 1 that quite different results can be achieved. We observe that some of the uncertainty concerning the possible additional change points has been eliminated (see, for example, the two CPP plots when $\phi_1 = 0.75$). In addition, as illustated by the righthand column of the figure, the distribution of the number of switches to the regime of interest has substantially more mass on two switches having occurred. This apparently improved confidence could be dangerously misleading in real applications.

The $\phi_1 = 0.9$ case in particular illustrates the importance of accounting for parameter uncertainty when considering change points. We observe in the exact calculations that only one switch to the regime of interest is the most probable which occurs at the beginning of the data, and the second occurrence to the regime is generally not accounted for. Thus the true behaviour of the underlying system is not correctly identified in this instance. Thus obtaining results by conditioning on model parameters may

	p_{11}	p_{21}	μ_1	μ_2	λ	ϕ_1
True	0.99	0.01	0	1	16	–
Posterior Means						
$\phi_1 = 0.5$	0.982 (0.010)	0.086 (0.046)	0.006 (0.033)	0.975 (0.074)	15.314 (1.538)	0.414 (0.073)
$\phi_1 = 0.75$	0.958 (0.093)	0.121 (0.123)	-0.057 (1.117)	1.201 (1.666)	14.764 (1.854)	0.731 (0.086)
$\phi_1 = 0.9$	0.891 (0.161)	0.190 (0.178)	-0.039 (1.856)	1.718 (2.606)	14.038 (1.916)	0.905 (0.044)

Table 1: Estimated posterior means and posterior standard deviations of parameters for the three simulated time series.

provide misleading change point conclusions and accounting for model parameter uncertainty is able to provide an general overview with regards to different types of possible change point behaviours that may be occurring. In Bayesian inference one should, whenever possible, base all inference upon the full posterior distribution, marginalising out any nuisance variables and that is exactly what the proposed method allows us to do.

4.3 fMRI Data

Functional magnetic resonance imaging (fMRI) allows the quantification of neuronal activity *in-vivo* through the surrogate measurement of blood flow changes in the brain. The ability to measure these blood flow changes relates to the so-called BOLD (blood oxygenation level dependent) effect (Ogawa et al., 1990) where hemoglobin changes its magnetic properties dependent on whether it is carrying oxygen or not (oxyhemoglobin and deoxyhemoglobin are diamagnetic and paramagnetic respectively). By examining the small magnetic field changes, it is possible to quantify the relative changes in the oxygen concentrations in the blood, which are a downstream product of neuronal activation. More information regarding fMRI and its many inherent statistical problems can be found in Lindquist (2008).

As mentioned above, most analysis of fMRI experiments is conducted by assuming a postulated experimental design (see Worsley et al. (2002) for example) and using standard linear modelling techniques, usually accounting for an AR component in the model. However, in many situations, it is not easy to determine an appropriate form for the design and there is no reason to suppose a direct temporal alignment of the stimulus and the response. This has been shown to be particularly an issue in psychology studies such as those on anxiety (Lindquist et al., 2007). Indeed, it will be the data from one such experiment, previously analysed in Lindquist et al. (2007) which will be of interest here. In particular, we will examine the manner in which making particular time series assumptions can affect the experimental conclusions of the experiment and their associated uncertainty.

The data analysed in this paper comes from an anxiety inducing experiment. Below is the task description as given in (Lindquist et al., 2007):

> The design was an off-on-off design, with an anxiety-provoking speech preparation task

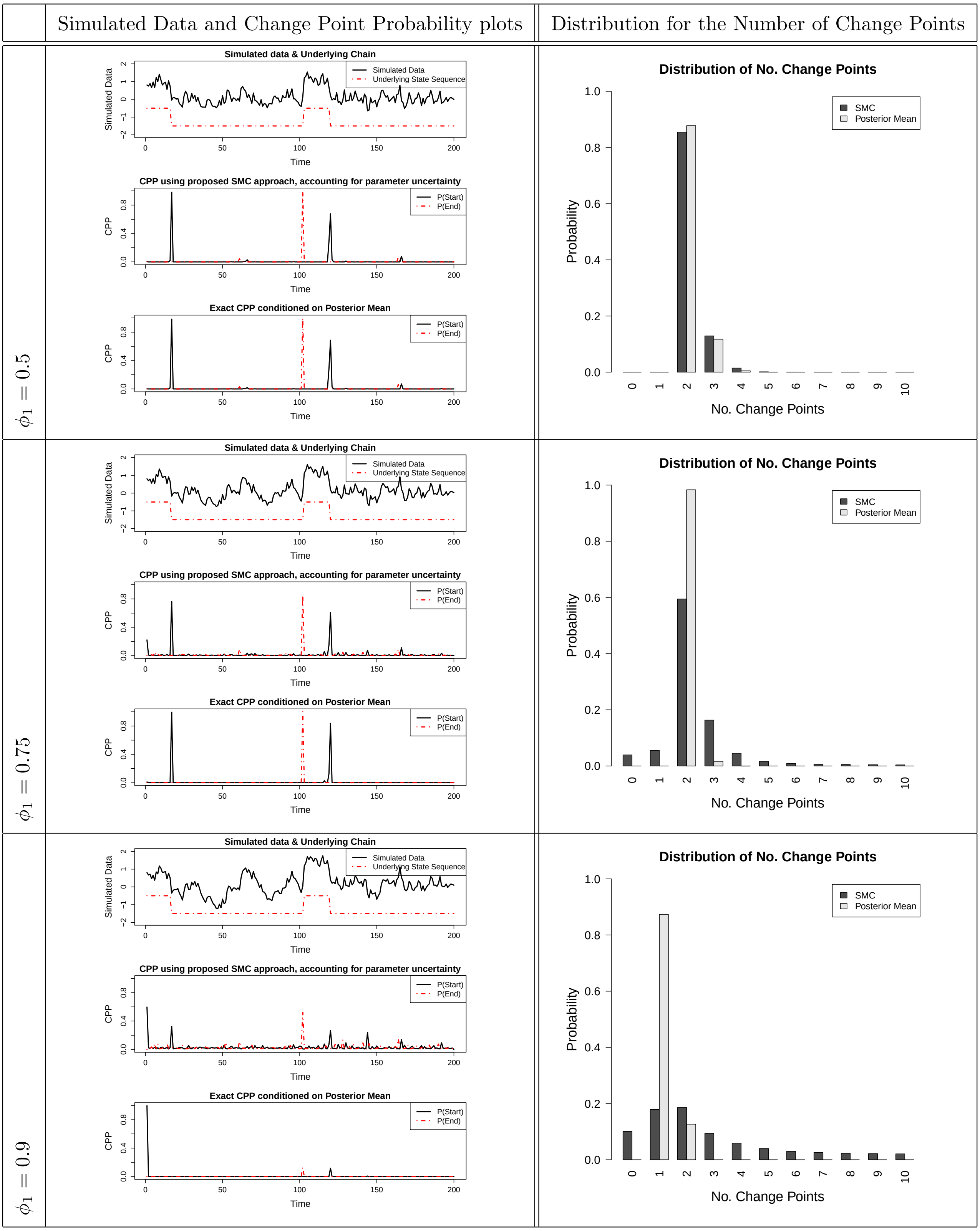


Figure 1: Results on Simulated Data generated from a Hamilton's MS-AR(1) model. We consider a variety of data and display the change point probability plots and distribution of number of change points obtained by implementing our proposed SMC based methodology. As a comparison, we also consider the exact change point distributions when conditioned on posterior means.

> occurring between lower-anxiety resting periods. Participants were informed that they were to be given two minutes to prepare a seven-minute speech, and that the topic would be revealed to them during scanning. They were told that after the scanning session, they would deliver the speech to a panel of expert judges, though there was "a small chance" that they would be randomly selected not to give the speech.
>
> After the start of fMRI acquisition, participants viewed a fixation cross for 2 min (resting-baseline). At the end of this period, participants viewed an instruction slide for 15 s that described the speech topic, which was to speak about "why you are a good friend". The slide instructed participants to be sure to prepare enough for the entire 7 min period. After 2 min of silent preparation, another instruction screen appeared (a relief instruction, 15 s duration) that informed participants that they would not have to give the speech. An additional 2 min period of resting baseline followed, which completed the functional run.

The time series were collected every 2 seconds for a total of 215 observations. The analysis in (Lindquist et al., 2007) consisted of using an exponential weighted moving average (EWMA) approach which corrected for an AR error process to find a change point and to determine the duration of the change until a return to baseline had occurred. This methodology does not easily allow the incorporation of multiple change points and requires detrending of the data to be performed prior to the analysis. Using the methodology in this paper, the detrending is added as another set of parameters to estimate within the SMC step providing a combined single step analysis, that is, the detrending within the model. This leads to an extension of Hamilton's MS-AR(r) model which is defined as follows:

$$y_t = \mathbf{m}_t'\beta + \mu_{x_t} + a_t \tag{37}$$

$$a_t = \phi_1 a_{t-1} + \ldots, +\phi_r a_{t-r} + \epsilon_t \qquad \epsilon_t \sim N(0, \sigma^2). \tag{38}$$

Here, $\mathbf{m}_t$ is a $d \times 1$ vector containing the d additional exogenous covariates (detrending basis in this case) at time t associated with the trend mean. β, a $d \times 1$ vector, are the associated trend related coefficients. Note that the Hamilton's MS-AR(r) model specified in Equation 28 can be obtained by fixing $\beta = \mathbf{0}$. In addition, the presented method of this paper allows the uncertainty in the estimation of the change points to be calculated. A 2-state Hamilton MS-AR(r) model with detrending can be used to model the considered time series (Peng, 2008), with the underlying state space being $\Omega_X = \{$"resting", "active"$\}$.

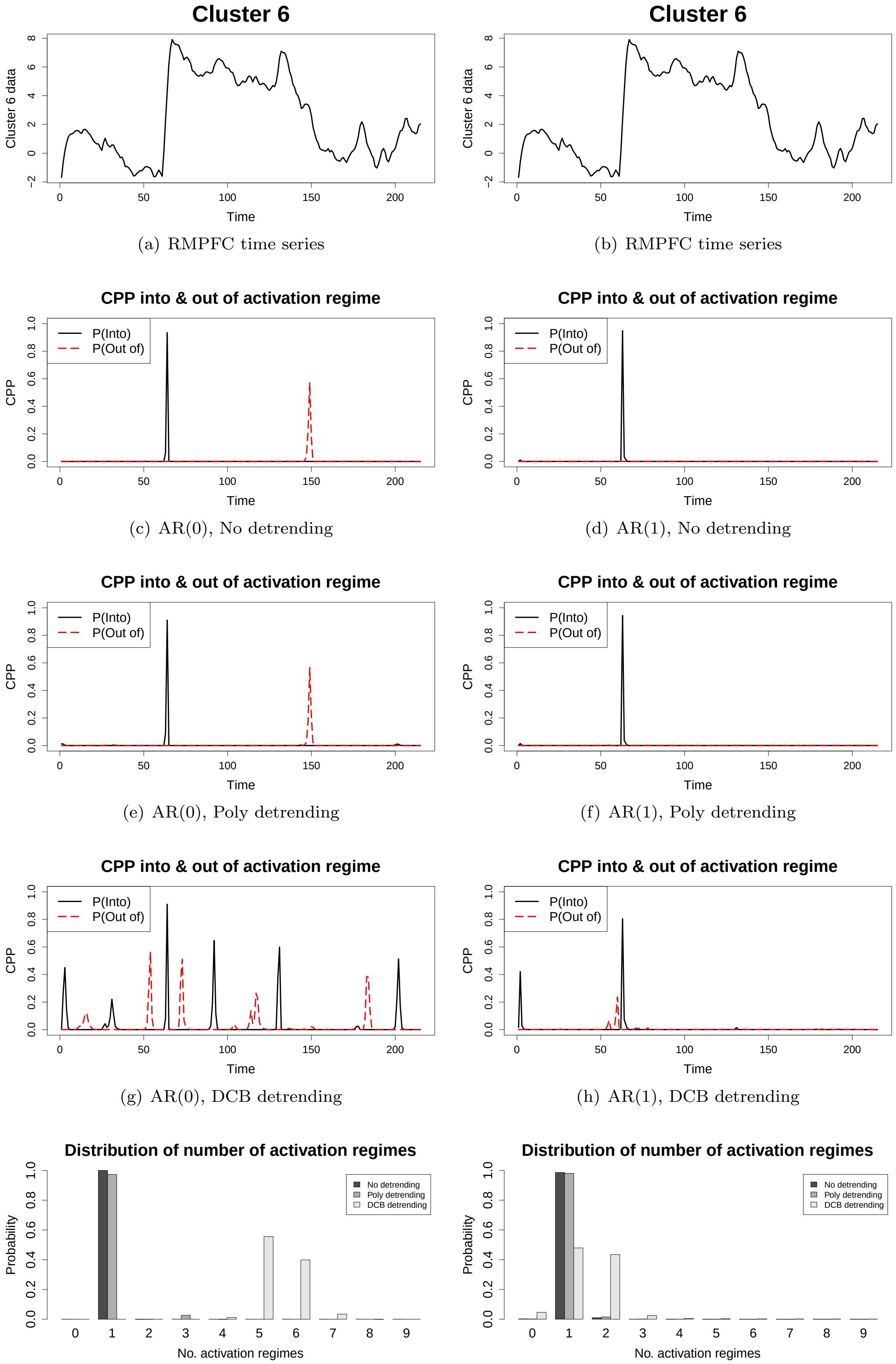
Several models and different detrending options are considered, mainly as an aid to discussion of the importance of taking care of time series properties in any fMRI analysis. Firstly, as a baseline comparison, a model with independent errors (an AR(0) error process) or detrending is used. This will show that this can be particularly unsatisfactory if a change point analysis is being used, which is unsurprising given that change point detection techniques are well known to breakdown in the presence of other forms of non-stationarity such as linear trends. The analysis then proceeds using various combinations of polynomial detrending (Worsley et al., 2002) and discrete cosine basis detrending (Ashburner et al., 1999), along with an AR(1) error model. An AR(1) model for fMRI time series is probably the most commonly used and is default in the Statistical Parametric Mapping (SPM) software (Ashburner et al., 1999).

Two specific regions of the brain are of most interest. In the first, the time series comes from the rostral medial pre-frontal cortex (RMPFC), which is known to be associated with anxiety, while the second is from the visual cortex (VC) and shows activation associated with the task-related instructions (these are denoted (as in Lindquist et al. (2007)) as Cluster 6 and 20 respectively in the results plots and data set). The resulting change point distributions for two regions of the brain can be seen in Figures 2-3 where we deem the region to be activated when there is a sustained changed for at least 5 time points in the activated region, thus $s =$ "active" and $k = 5$. The methodology in this paper finds significant evidence, in terms of the number of change points, that there is a least one change point in both regions of the brain. This accords with the previous EWMA analysis, where both regions were shown to have a change point, with the RMPFC region associated with the anxiety stimulus.

In addition to the actual change point distributions, the HMM analysis allows for different models to be assessed and the effect of the models on the uncertainty regarding change points locations. For the RMFPC region, if an AR(0) with no detrending is used, then two distinct changes, one into an activation region and one out of the activation region are determined. However, if an AR(1) model is assumed, with or without polynomial detrending, the return to baseline is no longer clearly seen, and the series consistent with only one change to activation from baseline during the scan. In this example, little difference is seen with the type of detrending, but considerable differences occur depending on whether independent errors are assumed or not. A little extra variation is found in the change point distribution if a discrete cosine basis is used, but this is likely due to identifiability issues between the cosine basis and the change points present.

On examining the regions of the VC, the choice of detrending is critical. If a suitable detrending is assumed, in this case a discrete cosine basis within the estimation, a clear change point distribution with multiple change points is found. However, if no or a small order polynomial detrending is used, the change point distributions associated with the visual stimuli are masked. It is also noticeable that the assumption of an AR(1) error process increases the inherent variability in the change point distribution.

We also consider an AR(1) error process with $\phi_1 = 0.2$ under all types of detrending. Fixing the AR parameter to this value is a common analysis approach, as featured in the SPM software. The change point results (data not presented) contained features present in both results AR(0) and AR(1) analysis with more peaked and centred change point probability features compared to the presented AR(1) results, due to the accounting for less uncertainty with respect to fixing $\phi_1 = 0.2$.



(a) RMPFC time series

(b) RMPFC time series

(c) AR(0), No detrending

(d) AR(1), No detrending

(e) AR(0), Poly detrending

(f) AR(1), Poly detrending

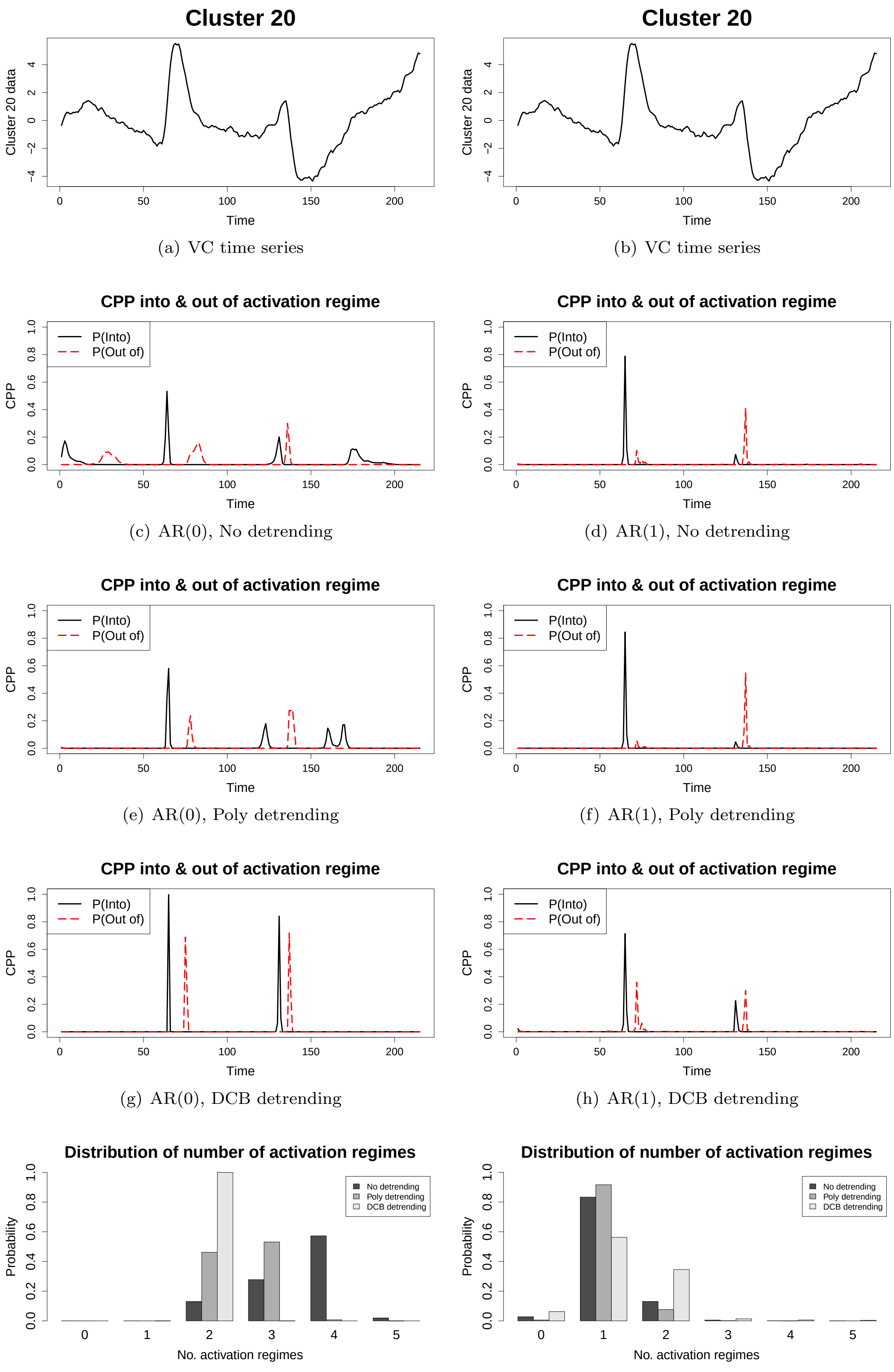
(g) AR(0), DCB detrending

(h) AR(1), DCB detrending

(i) AR(0), Distribution of number of regimes

(j) AR(1), Distribution of number of regimes

Figure 2: Change Point analysis results for the RMPFC (Cluster 6) region of the brain with respect to different order models and detrending.



(a) VC time series

(b) VC time series

(c) AR(0), No detrending

(d) AR(1), No detrending

(e) AR(0), Poly detrending

(f) AR(1), Poly detrending

(g) AR(0), DCB detrending

(h) AR(1), DCB detrending

(i) AR(0), Distribution of number of regimes

(j) AR(1), Distribution of number of regimes

Figure 3: Change Point analysis results for the VC (Cluster 20) region of the brain with respect to different order models and detrending.

5 Discussion

This paper has proposed and demonstrated within a biological context, the use of a new methodology in computing the uncertainty regarding change point estimates in light of model parameter uncertainty for time series. The methodology combines two recent approaches: Sequential Monte Carlo (SMC) samplers and exact change point distributions via Finite Markov Chain Imbedding (FMCI) in a Hidden Markov Model (HMM) framework. A Rao-Blackwellized SMC sampler is used to approximate the model parameter posterior via a weighted cloud of samples, without the need to sample the underlying state sequence. Conditioned on these model parameter samples, we are able to compute exactly the corresponding distributions regarding change point estimate via FMCI without simulating the underlying state sequence.

The proposed methodology introduces sampling error only in the model parameters, and is applicable to a wide class of models used within the literature. There is considerable flexibility within the proposed methodology and a range of different types of change point can be dealt with using it.

Our results have successfully demonstrated good estimation of the posterior distribution for change point characteristics of both simulated and real data without the need for significant application-specific tuning. In addition, the SMC component of our methodology provides a good approximation of the model parameter posterior which did not require sampling the latent state sequence and is not sensitive to change point results. Results for the simulated data demonstrate that parameter uncertainty cannot be safely ignored in change point analysis and that ignoring it can lead to incorrect conclusions.

There are a number of areas in which the proposed methodology could be improved and many of these are currently under investigation. In particular, we have, thus far, assumed the number of underlying states in the underlying Markov chain, H, to be known. However, this is not always the case and a natural extension would be to incorporate the number of states into the collection of parameters to be estimated. Recent work by Scott (2002) and Chopin and Pelgrin (2004); Chopin (2007) have accounted for the uncertainty regarding the number of states using MCMC and SMC techniques respectively. The second of these lends itself particularly to combination with the proposed method.

A limitation of the modelling employed in this paper is that by using a time-homogeneous HMM for the latent state sequence, it implicitly imposes a geometric distribution on the prior holding-time of each state. This can be an unreasonable assumption in several contexts. This difficulty could be resolved via the use of Hidden Semi-Markov models (HSMMs). We refer the reader to Yu (2010) for an overview of HSMMs. HSMMs can be seen as extensions of HMMs except that associated with each state is information regarding the duration spent in that state, for example a probability mass function defined over a possible set of durations.

A wide variety of different models exist for HSMMS, each with different assumptions for the duration distributions and state transitions, for example, whether it is independent to the previous duration spent in the previous state. Variable transition HMMs, where the state transition probabilities are dependent on the state duration, seem a natural extension since they can be collapsed to form an ordinary HMM.

This thus suggests that our existing exact change point distributions via FMCI could be applied.

In addition, some aspects of the SMC algorithm should be considered — achieving the best possible sampling performance will be critical when dealing with problems with large collections of unknown parameters. Areas to be considered include: using non-linear tempering schedules, optimal choice of proposal variances, using different MCMC transition kernels, and mutating samples by blocking correlated sub-components.

From an experimental point of view, one of the most important practical aspects when using change point methodology, such as the one presented here, is that the preprocessing and other scanner related model design issues become highly influential. Linear models with known designs are somewhat robust to misspecification of the trend model or error component model. However, this is not the case with change point procedures. A misspecification of the trend can produce results which look very much like change points, while not accounting for an autocorrelated error process could result in underestimation of the underlying uncertainty in the experimental conclusions. Therefore a careful and thorough time series modelling approach is necessarily when examining experimental data, particularly in applications such as fMRI.

Acknowledgements

The authors are very grateful to Professor Martin Lindquist for allowing us the use of the fMRI data and his discussions on change point methods used in fMRI. JA gratefully acknowledges support from EPSRC grant EP/H016856/1, as well as support from the EPSRC / HEFCE CRiSM grant. AJ gratefully acknowledges support from EPSRC grant EP/I017984/1.

References

Albert, J. and S. Chib (1993). Bayes inference via gibbs sampling of autoregressive time series subject to Markov mean and variance shifts. *Journal of Business & Economic Statistics 11*(1), 1–15.

Albert, P. S. (1991). A two-state Markov mixture model for a time series of epileptic seizure counts. *Biometrics 47*(4), pp. 1371–1381.

Andrieu, C., A. Doucet, and R. Holenstein (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72*(3), 269–342.

Ashburner, J., K. Friston, A. P. Holmes, and J. B. Poline (1999). *Statistical Parametric Mapping* (SPM2 ed.). website(http://www.fil.ion.ucl.ac.uk/spm): Wellcome Department of Cognitive Neurology.

Aston, J. A. and D. E. K. Martin (2007). Distributions associated with general runs and patterns in hidden Markov models. *The Annals of Applied Statistics 1*(2), 585–611.

Aston, J. A. D., J. Y. Peng, and D. E. K. Martin (2009). Implied distributions in multiple change point problems. CRiSM Research Report 08-26, University of Warwick.

Baum, L. E., T. Petrie, G. Soules, and N. Weiss (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics 41*(1), pp. 164–171.

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

Carpenter, J., P. Clifford, and P. Fearnhead (1999). An improved particle filter for non-linear problems. *IEEE Proceedings on Radar, Sonar and Navigation 146*(1), 2–7.

Chen, J. and A. K. Gupta (2000). *Parametric Statistical Change Point Analysis.* Birkhauser.

Chib, S. (1998). Estimation and comparison of multiple change-point models. *Journal of Econometrics 86*, 221–241.

Chopin, N. (2007). Inference and model choice for sequentially ordered hidden Markov models. *Journal of the Royal Statistical Society Series B 69*(2), 269.

Chopin, N. and F. Pelgrin (2004). Bayesian inference and state number determination for hidden Markov models: an application to the information content of the yield curve about inflation. *Journal of Econometrics 123*(2), 327 – 344. Recent advances in Bayesian econometrics.

Davis, R. A., T. C. M. Lee, and G. A. Rodriguez-Yam (2006). Structural break estimation for nonstationary time series models. *Journal of the American Statistical Association 101*, 223–239.

Del Moral, P., A. Doucet, and A. Jasra (2006). Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society Series B 68*(3), 411–436.

Del Moral, P., A. Doucet, and A. Jasra (2011). On Adaptive Resampling Procedures for Sequential Monte Carlo Methods. *Bernoulli To appear.*

Douc, R. and O. Cappé (2005). Comparison of resampling schemes for particle filtering. In *Image and Signal Processing and Analysis, 2005. ISPA 2005. Proceedings of the 4th International Symposium on*, pp. 64–69. IEEE.

Doucet, A. and A. M. Johansen (2011). A tutorial on particle filtering and smoothing: Fifteen years later. In D. Crisan and B. Rozovskiĭ (Eds.), *The Oxford Handbook of Nonlinear Filtering*. Oxford University Press.

Durbin, R., S. Eddy, A. Krogh, and G. Mitchinson (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press.

Eckley, I., P. Fearnhead, and R. Killick (2011). Analysis of changepoint models. In D. Barber, A. Cemgil, and S. Chiappa (Eds.), *Probabilistic Methods for Time-Series Analysis*, pp. 215–238. Cambridge University Press. To appear.

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

Fearnhead, P. and Z. Liu (2007). On-line inference for multiple changepoint problems. *Journal of the Royal Statistical Society Series B 69*, 589–605.

Fu, J. C. and M. V. Koutras (1994). Distribution theory of runs: A Markov chain approach. *Journal of the American Statistical Association 89*(427), 1050–1058.

Fu, J. C. and W. Y. W. Lou (2003). *Distribution Theory of Runs and Patterns and its Applications: A Finite Markov Chain Imbedding Approach*. World Scientific.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter (Eds.) (1996). *Markov Chain Monte Carlo In Practice* (first ed.). Chapman and Hall.

Gordon, N. J., D. J. Salmond, and A. F. M. Smith (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *Radar and Signal Processing, IEEE Proceedings F 140*(2), 107–113.

Hamilton, J. D. (1989, March). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica 57*(2), 357–384.

Huerta, G. and M. West (1999). Priors and component structures in autoregressive time series models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61*(4), 881–899.

Kong, A., J. S. Liu, and W. H. Wong (1994, March). Sequential imputations and Bayesian missing data problems. *Journal of the American Statistical Association 89*(425), 278–288.

Lehmann, E. L. and G. Casella (1998). *Theory of Point Estimation* (Second ed.). Springer.

Lindquist, M. (2008). The statistical analysis of fMRI data. *Statistical Science 23*, 439–464.

Lindquist, M. A., C. Waugh, and T. D. Wager (2007). Modeling state-related fMRI activity using change-point theory. *NeuroImage 35*(3), 1125–1141.

MacDonald, I. L. and W. Zucchini (1997). *Monographs on Statistics and Applied Probability 70: Hidden Markov and Other Models for Discrete-valued Time Series*. Chapman & Hal/CRC.

Neal, R. (2001). Annealed importance sampling. *Statistics and Computing 11*(2), 125–139.

Ogawa, S., T. M. Lee, A. R. Kay, and D. W. Tank (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proc Natl Acad Sci U S A 87*(24), 9868–72.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika 41*, 100–115.

Peng, J.-Y. (2008). *Pattern Statistics in Time Series Analysis*. Ph. D. thesis, Department of Computer Science and Information Engineering College of Electrical Engineering and Computer Science, National Taiwan University.

Peng, J.-Y., J. A. D. Aston, and C.-Y. Liou (2011). Modeling time series and sequences using Markov chain embedded finite automata. *International Journal of Innovative Computing Information and Control 7*, 407–431.

Roberts, G., A. Gelman, and W. Gilks (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability 7*(1), 110–120.

Robinson, L. F., T. D. Wager, and M. A. Lindquist (2010). Change point estimation in multi-subject fMRI studies. *NeuroImage 49*, 1581–1592.

Scott, S. (2002). Bayesian methods for hidden Markov models: Recursive computing in the 21st century. *Journal of the American Statistical Association 97*(457), 337–351.

Stephens, D. A. (1994). Bayesian retrospective multiple-changepoint identification. *Journal of the Royal Statistical Society Series C (Applied Statistics) 43*, 159–578.

Viterbi, A. (1967, April). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *Information Theory, IEEE Transactions on 13*(2), 260 – 269.

Whiteley, N., C. Andrieu, and A. Doucet (2009). Particle MCMC for multiple changepoint models. Research report, University of Bristol.

Worsley, K. J., C. Liao, J. A. D. Aston, V. Petre, G. Duncan, and A. C. Evans (2002). A general statistical analysis for fMRI data. *Neuroimage 15*(1), 1–15.

Yao, Y.-C. (1988). Estimating the number of change-points via Schwarz' criterion. *Statistics and Probabilitiy Letters 6*, 181–189.

Yu, S.-Z. (2010). Hidden semi-Markov models. *Artificial Intelligence 174*(2), 215 – 243. Special Review Issue.