THE UNIVERSITY OF
WARWICK

AUTHOR: **Christopher F. H. Nam** DEGREE: **Ph.D.**

TITLE: **The Uncertainty of Changepoints in Time Series**

DATE OF DEPOSIT: ....

I agree that this thesis shall be available in accordance with the regulations governing the University of Warwick theses.

I agree that the summary of this thesis may be submitted for publication.

I **agree** that the thesis may be photocopied (single copies for study purposes only).

Theses with no restriction on photocopying will also be made available to the British Library for microfilming. The British Library may supply copies to individuals or libraries. subject to a statement from them that the copy is supplied for non-publishing purposes. All copies supplied by the British Library will carry the following statement:

> "Attention is drawn to the fact that the copyright of this thesis rests with its author. This copy of the thesis has been supplied on the condition that anyone who consults it is understood to recognise that its copyright rests with its author and that no quotation from the thesis and no information derived from it may be published without the author's written consent."

AUTHOR'S SIGNATURE: ....

USER'S DECLARATION

1. I undertake not to quote or make use of any information from this thesis without making acknowledgement to the author.
2. I further undertake to allow no-one else to use this thesis while it is in my care.

DATE SIGNATURE ADDRESS

....

....

....

....

....

The Uncertainty of Changepoints in Time Series

by

Christopher F. H. Nam

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Department of Statistics

March 2013

THE UNIVERSITY OF
WARWICK

Contents

Acknowledgments

Several people and services have played some part in the forming of this thesis; each in their own unique ways.

I would like to express some gratitude to the founders and developers of Spotify, YouTube, Grooveshark and BBC iPlayer for providing such services and entertainment which have aided me when working through the days (and nights) in C1.03. A special thanks also go to the writers of the sitcoms *30 Rock* and *Arrested Development* for providing the necessary laughter required to survive a PhD.

My thanks go to the members of staff at the Statistics Department for providing a supportive working environment and many enjoyable conversations over lunch in the Common Room. These will be sorely missed. I would like to thank my fellow PhD students in the department, both past and present, for providing interesting conversations and entertainment, both statistical and non-statistical in context. My special thanks also go to former PhD students: Peter Windridge, Nastasiya Grinberg, Robert Goudie, Bryony Hill and Guy Freeman, for reassuring me through the bad times, entrusting me with their own PhD experiences, and generally making me smile and laugh even if from several miles away. As part of this thesis is also in collaboration with an external affiliation, I would also like to thank the Mathematics and Statistics Department at Lancaster University, in particular Aimée Gott, Rebecca Killick and Matt Nunes for entertaining me sufficiently well during my numerous collaborative visits.

This thesis would not be possible without the mathematical ideas and support of my supervisor and collaborators. I am extremely indebted to John Aston, my supervisor, for his immense guidance and support over the last three and a half

years, and for putting up with my frequent incompetence and rare bouts of competence. I have learnt a great deal both academic and non-academic related, namely a first foray into parenting, under your supervision. I would also like to thank Adam Johansen for practically acting as an unofficial second supervisor to me. My special thanks also go to Idris Eckley from Lancaster University for being a superb collaborator, offering invaluable advice regarding both work and life, and making me get over my dog allergy.

Finally, my thanks go to the Nam clan (Peter, Pamela, Janice, Flora, José) for their continuous love and support over the past seven and half years I have been at the University of Warwick. May you have also enjoyed the Warwickshire and West Midlands area over the past 15 years or so since we first ventured into this area. My how times have changed...

Declarations

I, Christopher Nam, hereby declare that this thesis is based on my own research, except when stated otherwise, in accordance with the regulations of the University of Warwick, and has not been submitted elsewhere.

The results presented in Chapter 2 utilises open source code available from the respective authors. I am grateful to these authors for making such code widely available.

Chapters 3 and 4 contains work featured in the publication: Nam, C. F. H., Aston, J. A. D., and Johansen, A. M. (2012b). Quantifying the uncertainty in change points. *Journal of Time Series Analysis*, 33(5):807–823. This is joint work with my supervisor Dr John Aston, and collaborator Dr Adam Johanson from the University of Warwick. The GNP dataset analysed in Chapter 3 is available from the website[1]. Chapter 4 features fMRI data made available by Professor Martin Lindquist for which I am grateful for.

Chapter 5 contains work that has been submitted for publication. This is joint work with Dr John Aston and Dr Adam Johanson. A draft of the paper can be found in: Nam, C. F. H., Aston, J. A. D., and Johansen, A. M. (2012a). Parallel Sequential Monte Carlo samplers and estimation of the number of states in a Hidden Markov model. *CRiSM Research Report*, 12(23). This chapter features results generated by the computer code of Dr Nicolas Chopin; I am grateful for the availability of such code.

Chapter 6 features work that has been submitted for publication. This paper is joint work with Dr John Aston, and collaborators Dr Idris Eckley and Dr Rebecca

[1] `http://weber.ucsd.edu/~jhamilto/software.htm`

Killick from Lancaster University. A draft of this publication can be found in: Nam, C. F. H., Aston, J. A. D., Eckley, I. A., and Killick, R. (2013). The uncertainty of storm season changes: Quantifying the uncertainty of autocovariance changepoints. *CRiSM Research Report*, 13(5). I would like to thank Shell International Exploration Production for supplying the oceanographic data considered in this chapter. This chapter also utilises some oceanographic results from Killick (2012), which in turn utilises code from Richard Davis and Haeran Cho. I thank Dr Killick for supplying such material.

In all three publications, I have taken a leading role both in terms of preparation of the material and conceptual work. All results presented in this thesis have been produced by me, unless stated otherwise.

List of Abbreviations

AIC	Akaike Information Criteria
AMOC	At Most One Change
AR	Autoregressive
ARMA	Autoregressive Moving Average
AutoPARM	Automatic Piecewise Autoregressive modelling
BIC	Bayesian Information Criteria
CP	Changepoint
CPP	Changepoint Probability
CUSUM	Cumulative Sum
DWT	Discrete Wavelet Transform
ESS	Effective Sample Size
EWS	Evolutionary Wavelet Spectrum
FMCI	Finite Markov Chain Imbedding
fMRI	functional Magnetic Resonance Imaging
GMM	Gaussian Markov Mixture
GNP	Gross National Product
HMM	Hidden Markov Model
HMS-AR(r)	Hamilton's Markov Switching Autoregressive Model of order r

HMS-ARS(r) Hamilton's Markov Switching Autoregressive Switching model of order r

HRF Hemodynamic Response Function

HSMM Hidden Semi-Markov Model

LSW Locally Stationary Wavelet

MA Moving Average

MAP Maximum A Posterior

MC Markov Chain

MCMC Markov Chain Monte Carlo

MLE Maximum Likelihood Estimate

MS-ARMA(r, q) Markov Switching Autoregressive Moving Average model with autoregressive order r and moving average order q

NBER National Bureau of Economic Research

NDWT Non-Decimated Wavelet Transform

PAC Partial Autocorrelation Coefficient

RJ-MCMC Reversible Jump Markov Chain Monte Carlo

RMPFC Rostral Medial Pre-Frontal Cortex

RWMH Random Walk Metropolis Hastings

SMC Sequential Monte Carlo

VC Visual Cortex

Abstract

Analysis concerning time series exhibiting changepoints have predominantly focused on detection and estimation. However, changepoint estimates such as their number and location are subject to uncertainty which is often not captured explicitly, or requires sampling long latent vectors in existing methods. This thesis proposes efficient, flexible methodologies in quantifying the uncertainty of changepoints.

The core proposed methodology of this thesis models time series and changepoints under a Hidden Markov Model framework. This methodology combines existing work on exact changepoint distributions conditional on model parameters with Sequential Monte Carlo samplers to account for parameter uncertainty. The combination of the two provides posterior distributions of changepoint characteristics in light of parameter uncertainty.

This thesis also presents a methodology in approximating the posterior of the number of underlying states in a Hidden Markov Model. Consequently, model selection for Hidden Markov Models is possible. This methodology employs the use of Sequential Monte Carlo samplers, such that no additional computational costs are incurred from the existing use of these samplers.

The final part of this thesis considers time series in the wavelet domain, as opposed to the time domain. The motivation for this transformation is the occurrence of autocovariance changepoints in time series. Time domain modelling approaches are somewhat limited for such types of changes, with approximations often taking place. The wavelet domain relaxes these modelling limitations, such that autocovariance changepoints can be considered more readily. The proposed methodology develops a joint density for multiple processes in the wavelet domain which can then be embedded within a Hidden Markov Model framework. Quantifying the uncertainty of autocovariance changepoints is thus possible.

These methodologies will be motivated by datasets from Econometrics, Neuroimaging and Oceanography.

Chapter 1

Introduction

Change is inevitable - except from a vending machine.

Robert C. Gallagher

Many time series and sequences of observations exhibit structural changes and breaks where a change occurs in the underlying system generating the data. Consequently, the data exhibits changes in statistical properties before and after the occurrence of this structural break. Instances of such structural breaks are becoming more frequent due to technological advances in recent years; time series can now be collected over a longer period and at a greater sampling rate. Analysis thus needs to account for such changes. We refer to instances of structural changes and breaks as changepoints (CPs); instances in time where the statistical properties differ pre and post this instance. This thesis considers aspects of CPs, in particular the uncertainty of them.

CP analysis is important in both theoretical and applied Statistics. For example, in standard time series analysis (see Chatfield (2003) for a good overview), many of the statistical theories and methodologies assume a stationary process where the statistical properties of the time series remain constant over time. Thus in order to consider non-stationary time series, it is necessary to devise methods in which CPs are identified and the non-stationary time series is segmented into smaller stationary time series. Methodologies assuming stationarity can then be applied to these segmented series. Alternatively, the potential presence of CPs can be incorporated into analysis, thus developing new methods which account for non-stationarity.

In an applied context, CPs are often associated with real life events which may consequently lead to a better understanding of the data and aid in decision making. For example, in the Gross National Product data considered in this thesis

(Figure 1.1), CPs correspond to switches in business cycles (between recession and growth periods). In addition these CPs and regimes often correspond to real life events, for example the 1956 recession (grey shaded region) is suspected to be associated with the Suez Crisis. The identification of CPs in time series may also suggest an intervening action on the system considered. For example, Page (1954) developed a methodology to identify whether a machine was faulty or not and needed replacing by assessing the quality of a product from the production line over time. If a fall in production quality is detected (our CP in this case), then the machinery is consequently replaced.

Motivated by the theoretical and applied aspects of CPs, considerable literature is dedicated to detection and estimation aspects of CPs problem. These consider whether a CP has occurred, and if so, how many and where these CPs might occur. In addition, CP methods encompass both offline and online scenarios where the data is made fully and incrementally over time respectively. In comparison, little attention has been focused on the uncertainty of the estimated quantities surrounding CPs.

Whilst detection and estimation of CPs are important aspects, perhaps motivated by the desired objectives when presented with a CP problem, the uncertainty of CP estimates should not be ignored. Considering the uncertainty of CPs may provide a better understanding of the data, highlighting any other potential CP configurations that may have occurred and providing some means of assessing the plausibility of different configurations. This is particularly important when different CP methods provide different results and we want to assess the plausibility of their estimates and their performance. Such a phenomena is successfully demonstrated in Chapter 2 where a variety of different CP methods are applied to the same dataset. Many existing CP approaches do provide some means of uncertainty quantification, but this is often implicit via the use of asymptotics and significance levels in hypothesis testing based methods (for example Chen and Gupta (2000); Davis et al. (2006); Cho and Fryzlewicz (2012)). Those which do quantify the uncertainty regarding CP locations assume the number of CPs to be known a priori, an unreasonable assumption if CP characteristics are generally unknown and of interest (see for example Chib (1998)). Recent Bayesian methods do consider quantifying the uncertainty of CP characteristics more explicitly, although this often requires sampling a long latent sequence which is often difficult to perform and may not be desirable (see for example Chen and Liu (1996); Chib (1998); Fearnhead (2006)).

This thesis considers methods in quantifying the uncertainty of CPs for a time series via the use of Hidden Markov Models (HMMs), a popular framework in the time series and CP community. We initially consider working on the observed time series directly and develop a methodology which provides the posterior distributions for several CP characteristics. This utilises an existing framework to compute exact CP distributions conditioned on model parameters (Aston et al., 2011), and accounts for model parameter uncertainty via the use of Sequential Monte Carlo (SMC) samplers. This combined framework is detailed in Chapter 3 and does not require sampling the underlying state sequence. This leads to a reduction in the Monte Carlo error of the parameters and more importantly, the CP estimates. The resultant methodology thus provides posterior distributions for CP characteristics in light of parameter uncertainty such that a reduction of sampling error is present.

A time-domain HMM framework provides a flexible CP method for changes in mean, variance, and combinations thereof. However, as non-stationarity can also arise from changes in autocovariance structure, it is necessary to consider such changes. Autocovariance CPs have received comparatively little attention compared to changes in mean and variance, with recent methods including Davis et al. (2006); Choi et al. (2008); Cho and Fryzlewicz (2012). However, such methods do not explicitly quantify the uncertainty of their CP estimates, and often provide different results.

A time-domain HMM is able to consider certain types of autocovariance CPs exactly (namely those with changes in autoregressive structure), with an approximation taking place for those which cannot be considered exactly (for example changes in moving average structure). This somewhat limits the type of data and changes that we can consider. We propose considering the observed time series in the wavelet domain which permits a frequency and location decomposition of the time series, and developing a HMM framework in the wavelet domain. By considering the time series in this alternative domain, CPs in second-order structure (autocovariance) may be more readily analysed than in the time domain.

This wavelet-domain approach, outlined in Chapter 6, considers modelling time series under a Locally Stationary Wavelet (LSW) framework, a popular wavelet based framework for modelling time series with evolving second-order structure. This second-order structure is characterised by the Evolutionary Wavelet Spectrum (EWS) at different frequencies and locations, with changes in autocovariance in the observed time series corresponding to changes in spectral structure of the EWS and vice versa. Consequently, focus now turns to assessing the periodogram, an estimate of the EWS, for changes. A HMM framework is established in modelling the periodogram as a multivariate time series with the appropriate emission density. The HMM framework thus allows a multitude of HMM-based CP methods to be applied, with our interest being that of quantifying the uncertainty of CPs. Consequently, the methodology detailed in Chapter 3 can be applied.

By considering the time series in the wavelet domain under the LSW framework, time series may be considered more readily due to their alternative representation. In addition, the proposed wavelet approach removes some sensitivity and concern with respect to model mis-specification compared to a time-domain approximation where an autoregressive component needs to be appropriately specified. By considering the time series in the wavelet domain, this may allow us to consider new types of data exhibiting changes in autocovariance and quantify the uncertainty of them.

The HMM setup considered throughout this thesis assumes that the number of underlying states is known a priori. This is often not the case when presented with time series data. This assumption is common in the statistical analysis and applications of HMMs and is not exclusive to CP analysis. We consider accounting for the uncertainty and determining the number of states of a HMM by extending the use of SMC samplers in their current context (see Chapter 5). The proposed SMC based methodology provides an efficient, flexible procedure in determining the unknown number of states by approximating the model posterior, which reduces the sampling error of estimates due to the absence of state sequence sampling, and requires no additional computational cost.

This thesis is motivated by three real datasets which exhibit different types of CPs. We firstly consider a dataset which is commonly featured in the CP literature; Hamilton's Gross National Product data (GNP, Hamilton (1989)). This data consists of differenced quarterly logarithmic US GNP data between the time periods 1951:II to 1984:IV. CP methods are predominantly used on this dataset in identifying the starts and ends of business cycles, namely when recessions begin and end. Figure 1.1 shows the transformed data that is analysed by various CP methods with recessions periods (grey regions) estimated by the National Bureau of Economic Research (NBER). By quantifying the uncertainty of CPs such as the number and location of recessions, we can assess the plausibility of NBER estimates and those provided by other CP methods. A change in mean is suspected for this time series. This dataset is analysed in Chapter 3 and will also feature as a running example in our literature review (Chapter 2) for demonstrating the performance of CP methods.

In addition, it is common to assume two underlying states are present in generating the GNP data, corresponding to the "contraction" and "expansion" states. Chapter 5 thus assesses whether such an assumption is valid via the HMM model

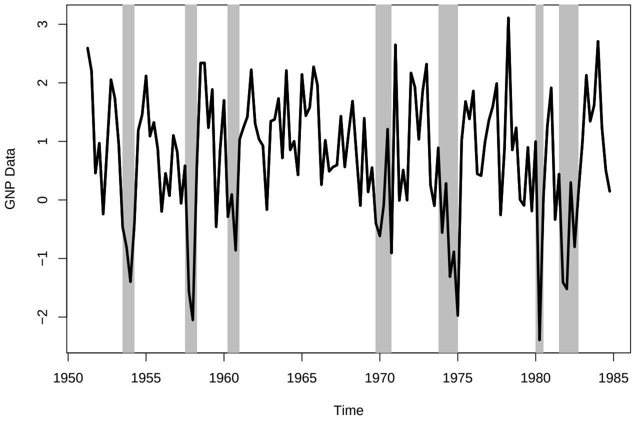
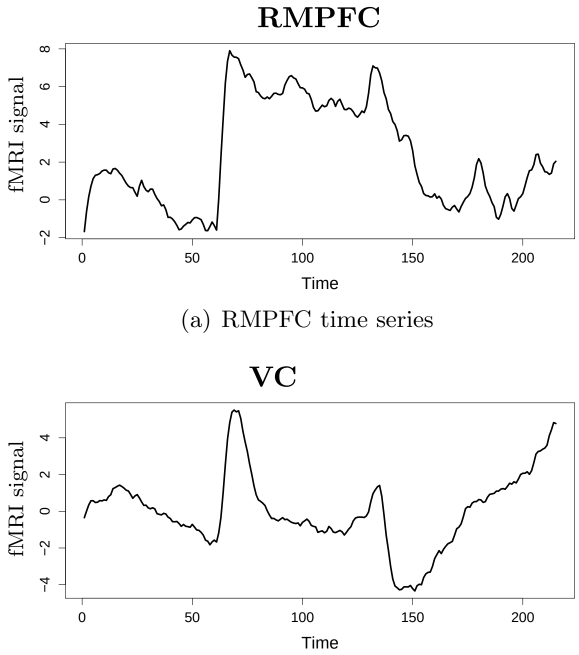


Figure 1.1: Hamilton's GNP: differenced quarterly logarithmic US GNP data from 1951:II to 1984:IV. CP methods are applied to this dataset in determining the starts and ends of business cycles, namely recessions. The grey regions denote the estimated recessions according to the NBER. A change in mean is suspected in this time series.

selection method developed.

The second dataset consists of functional Magnetic Resonance Imaging (fMRI) signals from a psychological experiment. fMRI signals are one way to measure brain activity over time. Two particular regions of the brain are of interest in the dataset (Figure 1.2), namely the Rostral Medial Pre-Frontal Cortex (RMPFC, associated with anxiety and fear emotions) and the Visual Cortex (VC, associated with visual interpretation). Interest lies in whether these regions behave accordingly with respect to the design of the experiment. In addition, statistical analysis for fMRI data typically assumes that the experimental design is known a priori where the onset timing of the stimulus corresponds directly to the onset timing of brain activity. However, this is often not the case, particularly in psychological experiments where the onset of a stimulus may not correspond directly in time to brain behaviour and patients reacting differently to stimulus. CP methods have thus been used to address this issue of unknown experimental design, with the uncertainty of CPs capturing the uncertainty of the onset of the stimulus and different reactions from subjects. A change in mean is associated with this dataset, although a trend is also present due to instabilities associated with fMRI data acquisition; this needs to be accounted for. Analysis of this dataset is considered in Chapter 4 where in addition to quantifying the uncertainty of CPs, detrending and error process assumptions



(a) RMPFC time series

(b) VC time series

Figure 1.2: fMRI signals from two regions of the brain from a psychological experiment. CP methods are used to determine whether the regions behave as expected with respect to the design of the experiment and determining the onset of a stimulus on the brain signal which is often assumed known in fMRI statistical analysis. A change in mean is suspected in both time series, although a trend and error process structure is also present which needs to accounted for.

are incorporated into the proposed methodology. We observe the effect of different detrending and error process assumptions commonly assumed in fMRI analysis on CP results.

The third dataset examines oceanographic data where interest lies in determining storm season changes from historic wave height data. By identifying these changes, ocean engineers may be able to use these results in planning future maintenance and inspection of ocean equipment such as offshore oil rigs. The data analysed is differenced wave heights at a central location in the North Sea from March 1992 – December 1994 (Figure 1.3), where changes in storm season correspond to changes in autocovariance structure of the time series. Differencing has been performed due to trend and seasonality being present in the original wave height time series. Due

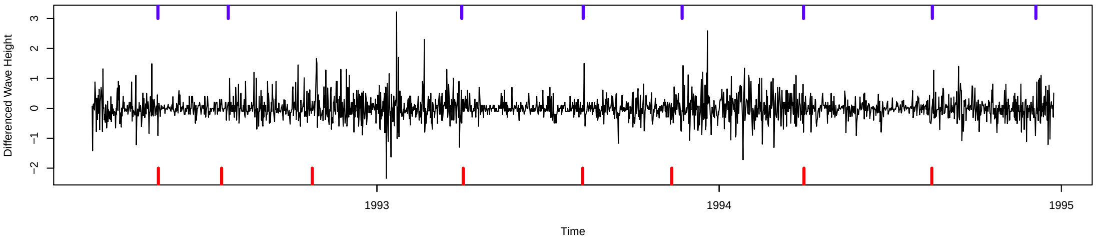


Figure 1.3: Differenced Wave Height Data from a central North Sea location collected at 12 hourly intervals from March 1992 – December 1994. Changes in storm season correspond to changes in autocovariance structure of this time series. The ticks at the top and bottom are estimated storm season changes identified by existing autocovariance CP methods.

to the inherent ambiguity and uncertainty associated with storm seasons such as the number and location of them, quantifying the uncertainty of CPs is of considerable interest. This dataset and the associated methodology of quantifying the uncertainty of autocovariance CPs is considered in Chapter 6.

The thesis features material which has appeared in the following list of publications:

- Nam, C. F. H., Aston, J. A. D., and Johansen, A. M. (2012b). Quantifying the uncertainty in change points. *Journal of Time Series Analysis*, 33(5):807–823
- Nam, C. F. H., Aston, J. A. D., and Johansen, A. M. (2012a). Parallel Sequential Monte Carlo samplers and estimation of the number of states in a Hidden Markov model. *CRiSM Research Report*, 12(23)
- Nam, C. F. H., Aston, J. A. D., Eckley, I. A., and Killick, R. (2013). The uncertainty of storm season changes: Quantifying the uncertainty of autocovariance changepoints. *CRiSM Research Report*, 13(5)

1.1 Structure of Thesis

The structure of the thesis is as follows: Chapter 2 provides a literature review of existing CP methods related to the problems of interest. Chapter 3 proposes a methodology in quantifying the uncertainty of CPs in light of model parameter uncertainty in the time domain via the use of a HMM framework. Chapter 4 further extends this proposed methodology in the context of brain imaging data such that detrending and error process assumptions can be embedded within the framework in a unified manner. Chapter 5 demonstrates how the SMC component of our proposed methodology in Chapter 3, can be further developed to deal with the unknown number of states in a HMM framework. Chapter 6 considers time series in a new domain, namely the wavelet domain. A HMM framework is developed in this alternative domain which allows us to consider the uncertainty of autocovariance CPs, an area which has received little to no attention. Chapter 7 concludes this thesis with a summary and discussion on potential paths for future work.

Chapter 2

Literature Review

> **Take a method and try it. If it fails, admit it frankly, and try another. But by all means, try something.**
>
> *Franklin D. Roosevelt*

2.1 Introduction

Changepoint (CP) analysis dates as far back as the 1950s, where in introductory papers such as Page (1954), methods were heavily motivated by quality control in the manufacturing industry. These problems were fundamentally detection driven; determine whether a change has occurred in the production quality and decide whether to suspend or reset the associated machinery based on the analysis outcome. Naturally, CP problems and analysis have evolved over time to consider a variety of different scenarios. This includes assuming different underlying assumptions on the observations, the presence of multiple CPs, and scenarios where the data increases incrementally over time. In addition, the term "changepoint" appears under a variety of synonyms in the literature due to its varied applications in Biology and Econometrics for example. This includes segmentation (for example in Braun and Müller (1998)), structural breaks (for example in Davis et al. (2006)) and detecting "disorder" (for example in Vostrikova (1981)).

There are a variety of ways in which changes in an observed time series can occur, for example changes in distribution, mean and variance. The majority of CP literature is dedicated to parametric changes, that is, the same model and distributional form is assumed across the data, but different parameters are associated between times of changes. Alternatively, changes in distribution assume that the data does not come from the same distribution, with different distributions being assumed between CPs.

Due to the extensive nature of CPs and the associated literature, we restrict our attention to relevant methods associated with the applications and problems of interest in this thesis. The problems and datasets encountered in this thesis concern estimation of CP characteristics retrospectively when all data is made available prior to analysis (an offline scenario). Consequently methods such as Ross (2012), which are defined within an online scenario where the data increases incrementally over time and analysis, will not be explored in detail. In addition, typically only a single univariate time series is reported with no additional datasets, such as additional exogenous covariates, being provided. Such covariates can be used in a change in regression (trend) context where $y_t = \mathbf{x}_t^T \beta_1$ for $1 \leq t < \tau_1$, and $y_t = \mathbf{x}_t^T \beta_2$ for $\tau_1 \leq t \leq n$, where $\mathbf{x}_t = (x_{t1}, \ldots, x_{tp})$ are the additional exogenous covariates, $y_{1:n}$ is the observed time series, and $\beta_j = (\beta_{j1}, \ldots, \beta_{jp})$ are changing regression parameters. Multivariate CP methods and those concerning changes in regression such as Zeileis et al. (2002) will receive little attention. The problems we shall consider also assume a common parametric distribution and thus methods concerning distributional changes will receive relatively little attention in this thesis.

For comprehensive overviews of CP methods, we refer the reader to Chen and Gupta (2000), Eckley et al. (2011). The website `changepoint.info` (Killick et al., 2012b), a recent initiative amongst the CP community, also provides a useful resource in fostering the research and applications of changepoint analysis with regards to publications and software implementations of both established and upcoming CP methods.

The structure of this chapter is the following. Section 2.2 introduces commonly used terminology and notation within the CP literature and within this thesis. We then proceed to Sections 2.3 to 2.13 which reviews a variety of CP methods. We conclude this chapter with Section 2.14 where we discuss the relative merits and downfalls of the reviewed CP methods with regards to quantifying the uncertainty of CPs.

To motivate why quantifying the uncertainty of CPs is an important aspect, we apply the reviewed CP methods to the aforementioned Hamilton's GNP dataset outlined in Chapter 1 (see Figure 1.1, page 5), where a software implementation of the method is available and is appropriate for the dataset. Hamilton's GNP time series is Gaussian distributed with a change in mean being suspected (Hamilton, 1989). CPs detected by the various methods are denoted by red vertical lines in plots of the data. Code is available from the respective author's website, `changepoint.info` (Killick et al., 2012b) and references therein.

2.2 Terminology and Notation

We now proceed in establishing the terminology and notation commonly used within the CP literature and in this thesis. Let $y_{1:n} = (y_1, \ldots, y_n)$ be an observed univariate time series of length n and $Y_{1:n}$ denote the corresponding sequence of random variables. Many statistical time series analysis assume $y_{1:n}$ is weakly stationary where the mean, variance and covariance of observations remain constant over time, and that the observations belong to the same statistical family. That is $Y_t \sim F(\theta_1)$, for all t, for some common distribution function F with associated parameters θ_1.

However, it is common for time series to be non-stationary such that the statistical properties of $y_{1:n}$ change over time. This is particularly common for time series collected over a long period of time, and can include changes in mean, variance, covariance and distribution. Such changes caused by structural breaks are known as changepoints (CPs) where the statistical properties of the data differ before and after the specified instance.

More formally, $\tau_1 \in \{2, \ldots, n\}$ is defined to be a CP if $y_{1:\tau_1-1}$ and $y_{\tau_1:n}$ possess different statistical properties. For parametric changes, this results in $y_{1:\tau_1-1} \sim F(\theta_1)$ and $y_{\tau_1:n} \sim F(\theta_2)$ with $\theta_1 \neq \theta_2$. Such parametric changes encompasses changes in mean, variance and covariance within the same distribution, F. This definition can also be easily extended to changes in distribution.

However, multiple changes can also occur within time series, particularly those collected over long periods of time. It is therefore necessary to extend the single CP definition into a multiple setting.

Definition 1. $\tau_{1:M}$*, CP configuration for* M *CPs.*
$\tau_{1:M} = (\tau_1, \ldots, \tau_M)$ *is defined to be a CP configuration for* M *CPs where* τ_i *denotes the location of the* i*th CP if*

1. $\tau_i \in \{2, \ldots, n\}$ *for* $i = 1, \ldots, M$ *with* $\tau_0 = 1$ *and* $\tau_{M+1} = n + 1$.

2. $\tau_i < \tau_j$ *if and only if* $i < j$*, for* $i, j \in \{0, 1, \ldots, M + 1\}$.

3. *The configuration partitions the data into* $M + 1$ *disjoint segments as follows:*

$$\begin{aligned} y_{1:n} &= y_{1:\tau_1-1} \cup y_{\tau_1:\tau_2-1} \cup \ldots \cup y_{\tau_{M-1}:\tau_M-1} \cup y_{\tau_M:n} \\ &= \bigcup_{i=1}^{M+1} y_{\tau_{i-1}:\tau_i-1} \end{aligned}$$

such that consecutive segments, $y_{\tau_{i-1}:\tau_i-1}$ *and* $y_{\tau_i:\tau_{i+1}-1}$ *for* $i = 1, \ldots, M$*, are statistically different.*

The additional conditions are necessary to make the multiple CP configuration valid; condition 2 enforces that future CPs in the sequence cannot occur before previous CPs, and condition 3 is necessary to partition the data into statistically different segments. Note however that non-consecutive segments of data need not be statistically different under this definition.

In the most general CP problem, we aim to estimate the unknown number of CPs present, M, and the respective locations of these M changes, $\tau_{1:M}$. In addition, the parameters associated with each of the $M+1$ segments, $\theta = (\theta_1, \ldots, \theta_{M+1})$, are generally unknown with $y_{\tau_{i-1}:\tau_i - 1} \sim F(\theta_i), i = 1, \ldots, M+1$. This needs to be estimated or accounted for in some manner.

In many of the methods considered, the likelihood,
$l(\theta, \tau_{1:M}|y_{1:n}) = p(y_{1:n}|\theta, \tau_{1:M})$ is a key concept in their approaches. How the likelihood is computed is very much dependent on assumptions made on the data and the model. For example, in some methods, it is common to assume independence amongst the segments conditional on the CP configuration $\tau_{1:M}$, in computing the likelihood. If such an assumption is enforced, the likelihood is found to be the product of the segment likelihoods.

More formally under this segment independence assumption, we denote the segment likelihood for segment $i \in \{1, \ldots, M+1\}$ as $l(\theta_i|y_{\tau_{i-1}:\tau_i-1}) = p(y_{\tau_{i-1}:\tau_i-1}|\theta_i)$. If segment independence is assumed conditional on the CP configuration, $\tau_{1:M}$, then the likelihood can be computed as:

$$l(\theta, \tau_{1:M}|y_{1:n}) = p(y_{1:n}|\theta, \tau_{1:M}) = \prod_{i=1}^{M+1} p(y_{\tau_{i-1}:\tau_i-1}|\theta_i) = \prod_{i=1}^{M+1} l(\theta_i|y_{\tau_{i-1}:\tau_i-1})$$

It is common to estimate the unknown θ via a maximum likelihood approach. That is, we estimate θ as that which maximises the likelihood.

$$\hat{\theta} = \arg\max_{\theta} l(\theta, \tau_{1:M}|y_{1:n}) \tag{2.1}$$

In the presence of CP configuration $\tau_{1:M}$, the maximum likelihood estimate (MLE) of θ is computed by considering the maximum likelihood of each segment. That is

$$\hat{\theta}_i = \arg\max_{\theta_i} p(y_{\tau_{i-1}:\tau_i-1}|\theta_i). \tag{2.2}$$

We denote the MLE of θ with respect to the MLE of each segment as $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_{M+1})$. However, not all methods, for example the methods proposed in this thesis, require such an assumption.

We now proceed in reviewing CP methods relevant to the problems of interest. We begin by reviewing two established single CP methods (Section 2.3), and then proceed to review multiple CP methods (Sections 2.4 to 2.13). Whilst single CP methods may appear to be limited in use within single CP scenarios, we shall review a method in which single CP methods can be applied within multiple CP contexts (Section 2.4).

2.3 At Most One Change

The At Most One Change (AMOC) model is as the name suggests, a model which is designed to account for a maximum of one CP occurring within the data. The model is defined as a pair of hypothesis tests as to whether a CP has occurred or not. More specifically, we test the following hypotheses

$$\begin{aligned}&H_0 : \; y_{1:n} \sim F(\theta_1).\\&H_1 : \text{There is an integer } \tau_1 \in \{2, \ldots, n\} \text{ such that}\\&\qquad y_{1:\tau_1 - 1} \sim F(\theta_1), y_{\tau_1 : n} \sim F(\theta_2) \text{ with } \theta_1 \neq \theta_2.\end{aligned}$$

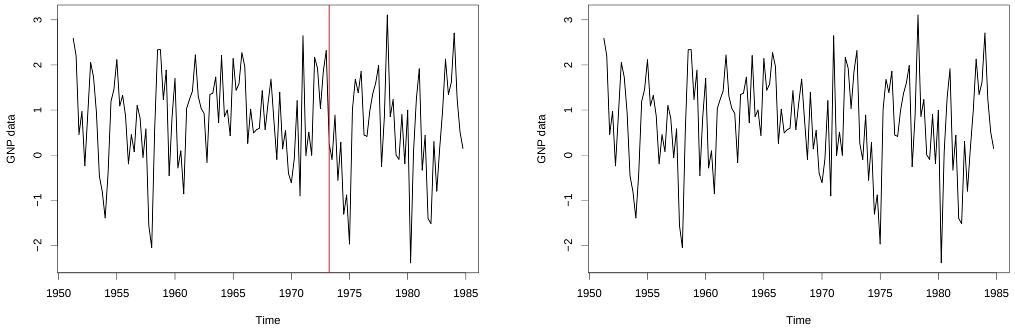
Non-parametric and parametric tests can now be constructed from the AMOC setup and thus used to determine whether a CP has occurred.

The Cumulative Sum (CUSUM, Page (1954)) is a non-parametric approach in testing the AMOC hypotheses. The approach computes a statistic sequentially, and compares this to a baseline statistic defined over the entire time series. For a change of mean scenario, the cumulative mean and sample mean are commonly considered as the sequential and baseline statistic respectively. The intuition of the method is that the sequential statistic is most different to the baseline statistic at the point of change. Thus, if the sequential statistic deviates sufficiently from the baseline statistic with respect to some threshold based on a significance level, H_0 is rejected and a CP is concluded to have occurred. In addition, the location at which the deviation is largest provides an estimate of the CP location.

There are a variety of ways in which the CUSUM statistic can be defined dependent on the type of change and whether observations are independent or not. A classical definition for a suspected change in mean (Kirch, 2006) is:

$$S_m := \frac{1}{\sqrt{m}} |m(\bar{y}_m - \bar{y}_n)| \qquad \text{where } \bar{y}_m = \sum_{i=1}^{m} \frac{y_i}{m}$$

$$T := \max_{m=1,\ldots,n} S_m \qquad \hat{\tau}_1 : = \arg \max_{m=1,\ldots,n} S_m.$$



(a) CUSUM

(b) Likelihood Ratio Test

Figure 2.1: The Cumulative Sum (CUSUM) and Likelihood Ratio Test (LRT) approaches under an At Most One Change setup, applied to Hamilton's GNP data. One CP and no CPs are identified under the two approaches respectively.

$\bar{y}_m$ denotes the cumulative mean, the sequential statistic in this case. S_m denotes the difference between the sequential statistic and the baseline statistic (the sample mean of the time series). If $T > \alpha$, where α is some predetermined significance threshold which encapsulates how certain one is that a CP has occurred, then a CP is concluded to have occurred. The respective location estimate of this CP, $\hat{\tau}_1$, is thus where this maximum deviation has occurred. An additional mathematical property of the CUSUM approach is that the limiting distribution of the CUSUM process T, forms a Brownian bridge under the null hypothesis (Csörgő and Horváth, 1997). In addition, if a confidence interval for $\hat{\tau}_1$ is desired, then through the use of bootstrapping and a suitable threshold, this can be computed (Hinkley, 1971).

Whilst we have presented the CUSUM statistic with respect to a change in mean, corresponding versions exist for changes of variance and other parameters (Inclan and Tiao, 1994; Lee and Lee, 2004). An implementation of the CUSUM procedure exists in the `R` package `changepoint` (Killick and Eckley, 2011). Its application on the GNP data, assuming a change in mean and 95% significance level, is displayed in Figure 2.1(a). A single CP is detected around 1973.

The CUSUM method is a simple intuitive method which is still actively used in the Engineering and quality control community. This may partly be due to the fact that it places no assumption on the distribution of the observations and thus makes it flexible in a variety of scenarios. However, CP uncertainty is only captured via the pre-determined significance level α, which may not be explicit enough for our particular needs.

An alternative approach to the AMOC setup is parametrically via maximum likelihood ratio (see Chen and Gupta (2000) for an extensive overview with respect to different parametric distributions). In order to calculate this ratio, it is necessary to compute the maximum likelihood (or some approximation of it) under both the null and alternative hypothesis. For H_0, this is relatively straightforward as $l(\hat{\theta}_1|y_{1:n})$ where $\hat{\theta}_1$ is the maximum likelihood estimate of θ. For H_1, the profile likelihood, as a function of τ_1, is considered which is defined as:

$$Prl(\tau_1) = l(\hat{\theta}_1|y_{1:\tau_1-1}) \times l(\hat{\theta}_2|y_{\tau_1:n}). \tag{2.3}$$

Let $\hat{\tau}_1 = \arg\max_{\tau_1} Prl(\tau_1)$, the value which maximises the profile likelihood above. These two likelihoods can then be used to define the log likelihood ratio test statistic:

$$T = 2\left(\log Prl(\hat{\tau}_1) - \log l(\hat{\theta}_1|y_{1:n}))\right). \tag{2.4}$$

Similar to the CUSUM approach, the null hypothesis is rejected and it is deduced that a CP has occurred if $T > \alpha$, where α is some significance level specifying how certain one wants to be that a CP has occurred. In addition $\hat{\tau}_1$ is also the estimate of the CP location. The typical threshold levels associated with likelihood ratio test statistics are not applicable in CP problems, and thus new asymptotic distributions for T need to be derived. Chen and Gupta (2000) derive asymptotic distributions for a variety of observational distributions, and these are used to determine the appropriate threshold to consider at various significance levels. For more obscure distributions, simulations are suggested to determine a suitable threshold level.

A likelihood ratio test statistic approach for Gaussian distributed data is available in the package `changepoint` (Killick and Eckley, 2011). Its application to the GNP data assuming a 95% significance level is demonstrated in Figure 2.1(b). Under this approach, no CPs are identified in the observed time series.

Similar to a CUSUM approach, a likelihood ratio test statistic is simple and intuitive to understand as it is formulated from a hypothesis testing framework. In addition, the parametric nature of the ratio also means it can be applied to CP problems concerning changes in known distribution. However, the parametric assumption required on the data is strong, and the resultant test statistic is heavily dependent on the assumptions placed on the data. In terms of the uncertainty of CPs, this is implicit via the specified significance level and is only with respect to the number of CPs, and not its location.

2.4 Binary Segmentation

Binary Segmentation (Vostrikova, 1981) is possibly one of the most established and utilised multiple CP methods in the CP literature. Part of its appeal is that it is intuitive as it iteratively applies a single CP method until no further CPs are detected in each segment of data. It can thus be used in a multiple CP setting. The general idea of the algorithm is that we iteratively segment data using a single CP method until no further CPs are suspected in each of the resultant subsequences of data.

Generic code for Binary Segmentation with respect to single CP methods involving a test statistic is presented in Algorithm 1. The algorithm is computationally efficient with computational cost $\mathcal{O}(n \log n)$, although it is an approximate method since it does not consider every possible CP configuration. Vostrikova (1981) and Venkatraman (1992) discuss the accuracy and consistency of estimates obtained by the algorithm, with two conditions being necessary in obtaining consistent estimates of the CP locations:

(a) The rescaled CP instance, $t_i = \frac{\tau_i}{n}, i = 1, \ldots, \widehat{M}$, are not dependent on the length of the time series n.

(b) The relative instants are sufficiently separated by some positive constant α, That is $t_i - t_{i-1} \geq \alpha$ for some $\alpha \in (0, \frac{1}{2}]$. This intuitively means that CPs cannot occur too close together and long segments generally occur.

An application of the Binary Segmentation algorithm on Hamilton's GNP data is presented in Figure 2.2 with the implementation available in the `changepoint` package. For this particular application, we have considered the CUSUM approach as the single CP method. We observe that numerous CPs are detected, the majority of which lead to very short segment lengths. The poor performance of this method is suspected to be due to the Binary Segmentation conditions being violated, and a low threshold being used in the CUSUM method (the default significance level of 95% in `changepoint`).

The merits of Binary Segmentation include the relative simplicity and intuitiveness of the procedure with a large amount of single CP literature being applicable within a multiple CP context. However, its greatest attraction lies in the fact that it is computationally efficient due to subsequences being tested for CPs in parallel of each other. However, this computational efficiency results in the algorithm being an approximate method and thus estimates are subject to some error. This error becomes more pronounced when the two conditions required for consistent

Algorithm 1 Code for the Binary Segmentation for any single CP method involving a test statistic.

```
Let $T(\cdot)$ be a single CP test statistic.
Let $\alpha$ be the threshold based on some significance level in determining whether a CP has occurred.
Let $\hat{\tau}(\cdot)$ be the estimator of the CP location for a given sequence of observations.
Initialise: $\mathcal{C} = \emptyset$= set of identified CPs, $\mathcal{S} = \{[1, n]\}$ = set of intervals.
Iterate:
while $\mathcal{S} \neq \emptyset$ do
    Select an element from $\mathcal{S}$. Denote this as $[t_1, t_2]$. Compute $T(y_{t_1:t_2})$.
    if $T(y_{t_1:t_2}) < \alpha$ then
        Remove $[t_1, t_2]$ from $\mathcal{S}$. {CP is not identified in the interval considered. Remove interval from considered set.}
    else
        $r = \hat{\tau}(y_{t_1:t_2}) + t_1 - 1$, and add $r$ to $\mathcal{C}$. {CP has been identified. Add $r$, CP location in the original time series, to the set of CPs. Remove $[t_1, t_2]$ from $\mathcal{S}$. }
        if $r \neq t_1$ then
            Add $[t_1, r - 1]$ to $\mathcal{S}$
        end if
        Add $[r, t_2]$ to $\mathcal{S}$ {Replace the interval with two subsequent intervals, segmenting around the identified CP.}
    end if
end while
$\mathcal{C}$ contains the estimated CP locations with $\widehat{M} = |\mathcal{C}|$ being the estimate of the number of CPs.
```

estimates are not satisfied, for example in the presence of short segments occurring. The effect of these two conditions potentially not being satisfied is demonstrated in the GNP application. Extensions of the Binary Segmentation algorithm, such as Circular Binary Segmentation (Olshen et al., 2004), have been developed to allow shorter segments to occur to light of their Genomic application. The main drawback of Binary Segmentation with respect to our CP uncertainty interest is that the uncertainty is only captured implicitly via asymptotic arguments required in obtaining consistent estimates. More specifically, Vostrikova (1981) show that if the two assumptions of Binary Segmentation are satisfied, then $P(|\hat{t_i} - t_i| > \epsilon) \leq \delta$ as $n \to \infty$, is guaranteed. That is, the probability of the estimate of the relative CP instance $\hat{t_i}$, deviating from the true relative CP location t_i by each constant $\epsilon > 0$, then there exists a $\delta > 0$ which provides an upper bound to this probability.

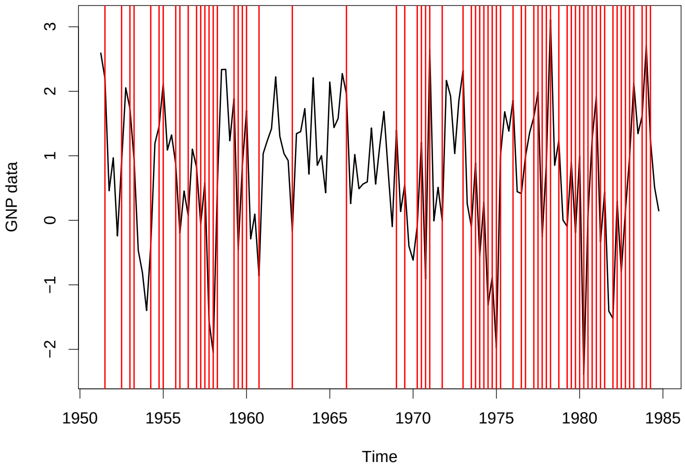


Figure 2.2: CP estimates for the GNP data using Binary Segmentation with CUSUM. Numerous CPs are identified which bare little relation to the estimates provided by the NBER. This poor performance is likely to be due to the necessary conditions of Binary Segmentation being violated and a low threshold (the default setting of 95% significance in `changepoint`) being utilised in the CUSUM method.

2.5 Penalised Likelihood approaches

Multiple CP problems can also be perceived as model selection problems such that each candidate model is associated with a different number of CPs being assumed. In light of this alternative perspective, a variety of model selection approaches and theory can be applied. In addition, a penalising approach can be utilised within dynamic programming based algorithms such as those reviewed in Sections 2.6 and 2.7, to obtain CP estimates in an efficient manner.

A penalised log-likelihood approach is a popular frequentist model selection approach which considers the fit of the model to the data but penalises for more complex models. Such an approach is applicable within a CP context (Yao, 1988; Chen and Gupta, 2000). The intuition for such a method is that the introduction of CPs leads to better fitting models but a penalty is associated with the CPs. This penalisation term is required as it is always possible to obtain a better fitting model by introducing additional CPs, over-segmenting the data such that each observation is considered as its only segment. This is analogous to introducing additional parameters in a linear regression context.

The general form of a penalised log-likelihood is as follows:

$$PL(\hat{\tau}_{1:M}) = -2\log Prl(\hat{\tau}_{1:M}) + p_M\phi(n),$$

where assuming M CPs are present, $Prl(\hat{\tau}_{1:M})$ denotes the maximum profile likelihood with respect to CP configuration $\hat{\tau}_{1:M}$. p_M is the number of parameters associated with assuming M CPs and $\phi(n)$ is the penalty function associated with the length of data. For example, $\phi(n) = 2$ and $\log(n)$ are equivalent to the penalty terms utilised in the Akaike and Bayesian Information Criteria (AIC, Akaike (1974), and BIC, Schwarz (1978)).

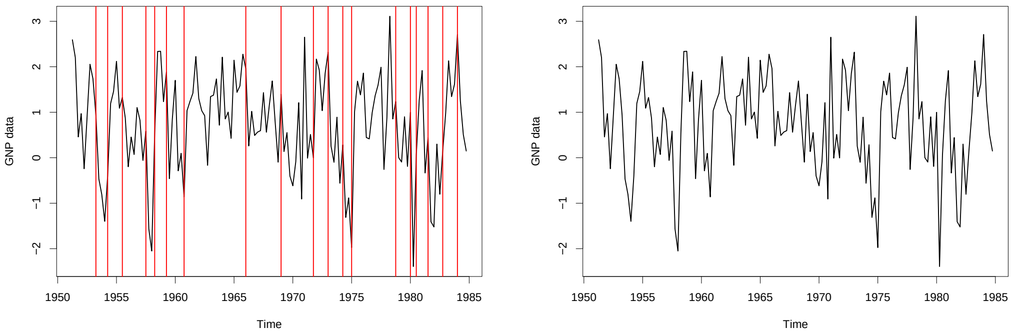
In obtaining an estimate of the number and location of CPs, the associated model and number of CPs which minimises the penalised log-likelihood is selected. The CP location estimates are those associated in achieving this minimisation. That is,

$$\hat{M} = \arg\min_{M=1,2,\ldots} PL(\hat{\tau}_{1:M})$$

with $\hat{\tau}_{1:M}$ being the estimate of the CP locations. This approach is similar to the Global Segmentation approach previously reviewed. However, the penalised log-likelihood approach does not consider all possible CP configurations, often considering a considerably smaller number of candidate CP configurations compared to the exhaustive approach of Global Segmentation. In addition, Global Segmentation is not restricted to the use of the log-likelihood as the chosen target criterion to be minimised.

The penalised log-likelihood approach can also be used as a single CP method by considering the minimum between the penalised log-likelihood of no CP and a single CP being present respectively (that is $\min\{-2\log l(\hat{\theta}|y_{1:n}) + p_0\phi(n), -2\log PL(\hat{\tau}_1) + p_1\phi(n)\}$). As such, if there is no prior knowledge in potential candidate CP configurations to consider, Binary Segmentation (Vostrikova, 1981) can be employed in conjunction with the penalised log-likelihood approach for a more flexible multiple CP method.

An implementation of penalised log-likelihood approaches exist in the `changepoint` package (Killick and Eckley, 2011) in conjunction with the Binary Segmentation algorithm. Application on the GNP data is displayed in Figure 2.3, where Akaike (Akaike, 1974) and Bayesian (Schwarz, 1978) penalty terms are considered. Quite different results are achieved under the two penalty terms used; 20 CPs are identified using an AIC penalisation, whilst no CPs are determined under a BIC penalisation. It is thus suspected that AIC and BIC are overestimating and



(a) Akaike Penalty, $\phi(n) = 2$

(b) Bayesian Penalty, $\phi(n) = \log n$

Figure 2.3: Implementation of penalised log likelihood CP approaches on Hamitlon's GNP data, in conjunction with Binary Segmentation. We consider two different penalty terms; Akaike and Bayesian respectively which yield quite different CP results.

underestimating respectively the number of CPs, a result of under and over penalising respectively, and a manual penalty of $\phi(n)$ between 2 and $\log n$ should thus be considered.

A penalised log-likelihood approach benefits from the model selection perspective of the CP problem in that it is simple and intuitive to understand. Established model selection results such as asymptotic overfitting associated with AIC, also provides additional theoretical results to CP problems. These results also aid in determining a suitable penalisation term, $\phi(n)$, such that one does not overestimate the number of CPs in long time series. Yao (1988) show that consistency in the estimate of the number of CPs is guaranteed if BIC (where $\phi(n) = \log n$) is chosen as the penalisation term. The uncertainty of CP characteristics is thus captured via these asymptotic arguments and is not explicit. In addition, these consistency results are only valid for the number of CPs and not their respective locations. As CP results are sensitive to the chosen penalisation term $\phi(n)$ (see Figure 2.3), fine tuning is often required to obtain the expected CP results. These penalisation terms are often abstract and hard to elicitate with respect to the application in hand

2.6 Global Segmentation

In contrast to the approximate nature of Binary Segmentation, Global Segmentation (Braun and Müller, 1998) provides an exact algorithm that identifies multiple CPs within a time series. The algorithm originates from a DNA segmentation context in Genetics. The general idea of the algorithm is that we find the optimal partitioning of the data when it is assumed M CPs are present, with $M = 0, 1, \ldots, M^{\max}$ where $M^{\max}$ is the maximum number of CPs considered. The optimal partitioning for a given number of CPs is achieved by considering every possible CP configuration possible and selecting the configuration which minimises a chosen target criterion. This target criterion aims to capture the fit of the segment configuration with respect to the data. More formally, one can consider the log of the maximum segment likelihood,

$$R(y_{t_1:t_2}) = -\log p(y_{t_1:t_2}|\hat{\theta}). \tag{2.5}$$

Then the resultant fit for a particular CP configuration $\tau_1, \ldots, \tau_M$, assuming M CPs is,

$$\rho^M(\tau_{1:M}) = \sum_{i=1}^{M+1} R(y_{\tau_{i-1}:\tau_i - 1}) \tag{2.6}$$

Note that this is equivalent to the log profile likelihood as defined in Equation 2.3 if segment independence is assumed. In determining the best segmentation assuming M CPs are present, consider

$$\hat{\rho}^M(\hat{\tau}_{1:M}) = \min_{\tau_{1:M}}\{\rho^M(\tau_{1:M})\}. \tag{2.7}$$

$\hat{\tau}_{1:M}$ denotes the optimal partitioning assuming $M = 1, 2, \ldots, M^{\max}$. In determining the optimal number of CPs to assume, one can consider a penalised log-likelihood approach which takes into consideration the fit of the partitioning but penalises for the introduction of additional CPs. That is,

$$PL(M) = 2\hat{\rho}^M(\hat{\tau}_{1:M}) + p_M\phi(n) \tag{2.8}$$

$$\widehat{M} = \arg \min_{M=0,1,\ldots,M^{\max}} PL(M) \tag{2.9}$$

where p_M is the associated number of parameters by assuming M CPs, and $\phi(n)$ is a penalty function with respect to the length of the n as before in Section 2.5. $\widehat{M}$ is the estimate of the number of CPs, with the associated CP configuration minimising $\hat{\rho}^{\widehat{M}}(\hat{\tau}_{1:\widehat{M}})$ being the estimate of the CP locations.

Computing the optimal segmentation assuming M CPs are present is performed via a dynamic programming approach (Auger and Lawrence, 1989). The basic idea of such an approach is that the optimal segmentation for M CPs is deduced by using the optimal segmentation assuming $M - 1$ CPs and where best to

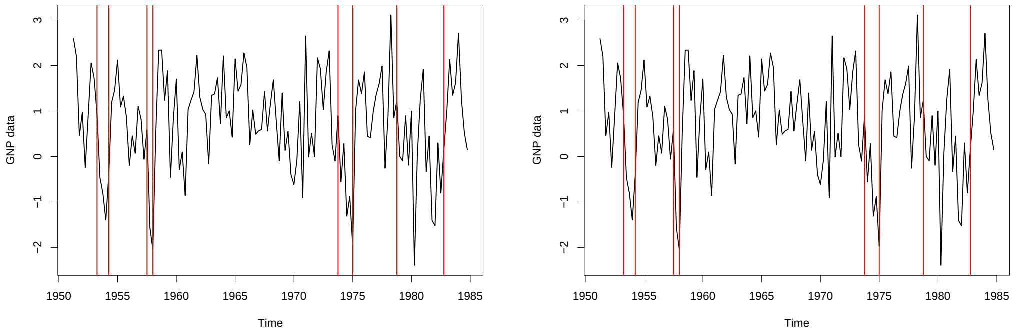
Algorithm 2 Algorithm code for the Global Segmentation.

```
Let $R(\cdot)$ be a target criterion in which we wish to minimise.
Let $1 < M^{\max} \ll n$ be the specified maximum number of CPs considered.
Let $p_M$ be the number of parameters associated with a model assuming $M$ CPs.
Let $\phi(n)$ be a penalty function associated with the length of the data.
Initialise: For all $i, j \in [1, n]$ with $i < j$, compute $q^1_{i,j} = R(y_{i:j})$. {Compute all possible segment likelihoods.}
for $M = 1, \ldots, M^{\max}$ do
    for $j \in \{1, 2, \ldots, n\}$ do
        Compute $q^M_{1,j} = \min_{v=1,\ldots,j}(q^{M-1}_{1,v-1} + q^1_{v,j})$ {Compute optimal target from 1 to j assuming M CPs are present, based on the M − 1 CP configuration and introducing a new CP at v.}
        $\tau^M_1 = \arg\min_v \left(q^{M-1}_{1,v-1} + q^1_{v,n}\right)$
    end for
    for $i = 1, \ldots, M-1$ do
        $\tau^M_i = \arg\min_v \left(q^{M-i-1}_{1,v-1} + q^1_{v,\tau^M_{i-1}}\right)$ {Determine locations of M CPs present by traversing backwards.}
    end for
end for
$\tau^M_{1:M}$ are the CP locations assuming $M$ CPs.
Final Inference: Compute $PL(M) = 2q^M_{1,n} + p_M\phi(n)$
$\widehat{M} = \arg\min_{M=1,\ldots M^{\max}} PL(M)$ is the estimate of number of CPs, and associated $\tau_{1:\widehat{M}}$ minimising this quantity are the estimate of the CP locations.
```

place the new CP. Its implementation within the Global Segmentation algorithm is displayed in Algorithm 2 which describes the generic algorithm.

The `changepoint` package features an implementation of the Global Segmentation algorithm. Figure 2.4(a) displays its application on the GNP data assuming a maximum of 20 CPs ($M^{\max} = 20$), and using a BIC penalty to determine the optimal number of CPs to assume and the associated CP configuration. We observe eight CPs have been identified which appears to concur with how the time series is behaving (CPs are identified when there is a shift between the top and lower half of the data range).

An advantage of Global Segmentation over many algorithms such as Binary Segmentation is that it guarantees the optimal solution is found. This is achieved by considering all possible CP configurations. However this exploration comes at an increased computation cost $\mathcal{O}(n^2)$. This results in the algorithm not being suitable within a long time series context, despite its original motivating application in Genetics. Approximations of the algorithm exist (Braun et al., 2000), although this comprises the accuracy and consistency of the estimates. In addition, the choice in



(a) Global Segmentation (BIC)

(b) PELT (BIC)

Figure 2.4: CP estimates via the Global Segmentation and PELT algorithm. The same CPs are identified under both approaches.

$M^{\max}$ and penalty function , $\phi(n)$ can be quite influential on the results. In additional analysis not shown here, different CP results were obtained when considering an AIC penalty function and lower value of $M^{\max}$. This sensitivity may not be desirable with these parameters set according to the expected CP results one wishes to obtain. The algorithm captures the uncertainty of CP estimates implicitly via the use of asymptotic arguments in providing consistent estimates for CP characteristics and model parameters.

Extensions of the Global Segmentation algorithm exist, namely the Dynamic Programming Algorithm as proposed in Bai and Perron (2003). This extended algorithm allows a minimum distance between two CPs to be specified in addition to the maximum number of CPs considered. By specifying a minimum segment length, this allows additional exogenous information to be incorporated about CP behaviour, for example, the minimum time period between two changes in business cycles in the GNP example. An implementation of this CP algorithm exists in the `R` package `strucchange` (Zeileis et al., 2002) via the function `breakpoints`.

2.7 Pruned Exact Linear Time algorithm

The Pruned Exact Linear Time Algorithm (PELT, Killick et al. (2012a)), combines the computational advantage of Binary Segmentation, but also the exact nature and accuracy of Global Segmentation. Akin to Global Segmentation, the objective is to minimise a chosen target criterion over the possible number and locations of CPs. One of the underlying assumptions of the algorithm in obtaining accurate CP estimates is that the number of CPs increases linearly with the length of the data.

The PELT algorithm considers the data sequentially and the optimal seg-

Algorithm 3 Algorithm code for PELT.

```
Let $R(\cdot)$ be a target criterion in which we wish to minimise.
Let $\phi(n)$ be a penalty function associated with the length of the data.
Let $K$ denote a constant to ensure a time point is kept as a candidate CP.
Initialise: $R^{\min}(0) = -\phi(n)$, $\Delta(0) = \emptyset$ = CP configuration up data point 0, $C_1 = \{0\}$ = optimal set of candidate CP locations.
for $t = 1, \ldots, n$ do
    Compute $R^{\min}(t) = \min_{\tau \in C_t}[R^{\min}(\tau) + R(y_{\tau:t-1}) + \phi(n)]$ and let $\tau^1 = \arg R^{\min}(t)$. {Determine the location of the most recent CP. }
    Set $\Delta(t) = \{\Delta(\tau^1) \cup \tau^1\}$
    Set $C_{t+1} = \{\tau \in C_t \cup t : R^{\min}(\tau) + R(y_{\tau+1:t}) + K \leq R^{\min}(t)\}$ {Pruning step: Will $\tau$ ever be a potential CP?}
end for
$\Delta(n)$ contains the estimates of the CP locations.
```

mentation up to that time point. In particular, the efficient computational cost is achieved by restricting the number of CP configurations considered at each time point. This restriction is enforced by considering the location of the last CP rather than the entire CP configuration up to that time point, and eliminating CP configurations which include time points which could never be a potential CP location (a pruning step). This leads to a linear number of CP configurations being considered at each time point and thus a reduction in the computational cost.

Algorithm 3 describes the generic implementation code for PELT. $R(\cdot)$ and $\phi(n)$ denote the same quantities as in Global Segmentation; the criterion to measure the fit of the data and the data dependent penalty term. $R^{\min}(t)$ denotes the minimum of the target criterion up to time t. $R^{\min}(t)$ is computed recursively, based on the minimum obtained at previous time points. This recursion is based on the Optimal Segmentation algorithm (Jackson et al., 2005). K is a constant which is introduced by the PELT setup as part of the pruning step. This pruning step determines whether the current time point t will ever be a CP in future configurations by significantly improving the target criterion if it is a candidate CP location ($R^{\min}(\tau) + R(y_{\tau+1:t}) + K \leq R^{\min}(t)$ in the Algorithm 3). $\Delta(t)$ is the optimal CP configuration for the data up to time t, $y_{1:t}$. Hence $\Delta(n)$ provides the estimate of the CP locations, with $\widehat{M} = |\Delta(n)|$ being the estimate of the number of CPs

An implementation of PELT exists within the `changepoint` package. Figure 2.4(b) displays results of its implementation on the GNP data with a BIC penalty function in place to control over segmentation. Eight CPs are identified with the same configuration being obtained under Global Segmentation.

PELT offers a good recommendable alternative to the Global Segmentation as it retains the exactness and optimality of Global Segmentation, but at the same computational cost of Binary Segmentation. As such, it can be applied on long time series. However, the algorithm assumes that the number of CPs grows linearly with the length of the time series, and hence segments cannot be too long. In addition, CP uncertainty is captured implicitly via the use of asymptotic arguments.

2.8 AutoPARM

Davis et al. (2006) propose an alternative frequentist model selection approach by obtaining a CP configuration which minimises the Minimum Description Length, another measure of model fit. The proposed method models time series specifically as piecewise autoregressive (AR) processes such that the number, location and orders of the segment AR processes are unknown. Under the parametric assumption that each segment can be modelled as an AR process, the proposed approach is aptly named Automatic Piecewise Autoregressive modelling, AutoPARM.

More explicitly, AutoPARM models the observed process Y_t as a piecewise AR process,

$$Y_t = \mu_{X_t} + \phi_{1,X_t} Y_{t-1} + \ldots + \phi_{p,X_t} Y_{t-p} + \epsilon_t. \qquad \epsilon_t \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2_{X_t}) \tag{2.10}$$

$X_t = \{1, \ldots, M+1\}$ is a latent variable process which denotes which segment and consequently which AR model is being assumed. X_t is constrained to be a non-decreasing process such that it only has two moves at each time; stay in the current segment, or start a new segment. Returning to previously visited segments is therefore not possible.

μ_{X_t} denotes the segment dependent mean associated with the AR segment at time t. ϵ_t is an independent, Gaussian noise process with switching variance $\sigma^2_{X_t}$. $(\phi_{1,X_t}, \ldots, \phi_{p,X_t})$ denotes the p AR coefficients associated with the segment at time t. In addition to the AR coefficients switching between segments, the AR order is also permitted to change. p thus denotes the maximum AR order considered amongst the $M+1$ segments, that is $p = \max\{p_1, \ldots, p_{M+1}\}$ where p_j denotes the AR order associated with segment j. This consequently results in zero AR coefficients, $\phi_{p_j+1} = \ldots = \phi_p = 0$, when $p_j < p$.

In addition to the parameters of each AR segment,
$\theta = \{\mu_j, \sigma^2_j, \phi_{1,j}, \ldots, \phi_{p,j}\}_{j=1}^{M+1}$, being unknown and requiring estimation, both the number of segments $M+1$, and corresponding M breakpoints need to estimated. The methodology determines these quantities by minimising the Minimum Description Length (MDL, Rissanen (1978)). MDL is a term from information theory which

provides an alternative description of a fit of a model with regards to data. More formally, it measures the compression of data with respect to a model with the best fitting model achieving maximum compression of the data and thus achieving a minimum MDL.

Davis et al. (2006) derive the explicit form for the MDL where M breakpoints at locations $\tau_{1:M}$ are present, and AR orders $p_1, \ldots, p_{M+1}$ for each of the resultant segments as:

$$\begin{aligned}\mathrm{MDL}(M, \tau_{1:M}, p_1, \ldots, p_{M+1}) &= \log M + (M+1)\log n + \sum_{j=1}^{M+1} \log p_j \\ &+ \sum_{j=1}^{M+1} \frac{p_j + 2}{2} \log(\tau_j - \tau_{j-1}) + \sum_{j=1}^{M+1} \frac{\tau_j - \tau_{j-1}}{2} \log(2\pi\hat{\sigma}_j^2).\end{aligned} \tag{2.11}$$

This formula of the MDL can be seen as the sum of the code length of fitted values (first four terms) and residuals (last term) under the assumed segmentation and model. The MDL equation is derived by deducing the upper bound on code length on each component, M, p_j and τ_j, from their behaviour (integer valued and bounded for example). The code length corresponding to the residuals under the fitted model is constructed such that larger values of $\hat{\sigma}_j^2$ (the Yule-Walker estimate of σ_j^2) will thus correspond to large residuals (bad fitting models) and a larger code length.

The MDL can be viewed as a target criterion with a penalty term akin to the penalised log-likelihood. The code length for the residuals effectively assesses the fit of the assumed model. The code length regarding the fitted values will however penalise models which are more complex than necessary and require more parameters. The optimal segmentation is that which minimises the MDL as in Equation 2.11.

The number of possible configurations under $(M, \tau_{1:M}, p_1, \ldots, p_{M+1})$ is enormous and thus optimisation of MDL cannot be performed by exhaustive procedures. A Genetic Algorithm (GA, Goldberg and Holland (1988)) is thus implemented to locate the minimum MDL stated in Equation 2.11. GA algorithms are search optimisation algorithms inspired by Darwin's theory of evolution. Algorithms begin with a set of initial possible vector solutions from the search space known as chromosomes. These chromosomes are also assigned weights in relation to how they perform on the objective function (in this case the MDL), with those performing well being assigned higher weights. Parent chromosomes are then randomly selected according to these weights. In exploring the solution space, offspring chromosomes are created by

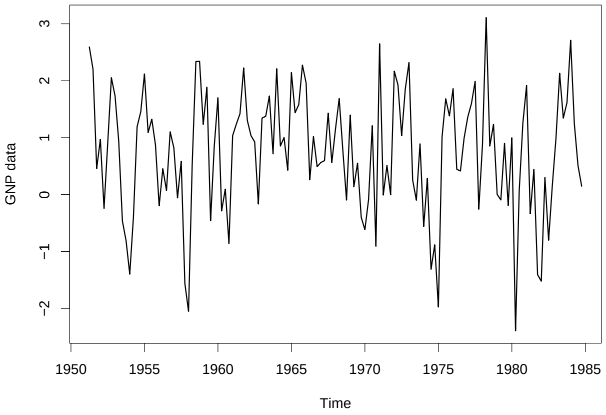


Figure 2.5: Application of AutoPARM on Hamitlon's GNP data where no CPs are determined to have occurred.

mutating the sampled parent chromosomes. These offspring chromosomes form the second generation which are believed to further improve the objective function by taking forward the stronger solutions and the characteristics associated with them. This procedure of creating further generations by mutating existing offspring chromosomes is iterated numerous times and a solution to the optimisation problem is obtained. GA algorithms thus allow a configuration of $(M, \tau_{1:M}, p_1, \ldots, p_{M+1})$ which minimises the MDL to be determined. This provides an estimate of the number and location of CPs which optimally partitions the data.

The setup of AutoPARM as highlighted in Equation 2.10, permits changes in mean, variance and covariance (via changing AR parameters and orders) being identified. An implementation of AutoPARM is available for request from the authors. Its application on the GNP data is displayed in Figure 2.5. No CPs are identified in the data according to the AutoPARM approach.

AutoPARM provides a state-of-the art methodology in identifying changes in mean, variance and covariance for Gaussian time series. The latter type of change has received relatively little attention compared to the former two types of changes. This consequently makes AutoPARM an attractive recent CP method in encompassing a variety of types of changes compared to the approaches considered thus far. The piecewise AR assumption also explicitly permits dependency within $y_{1:n}$ to be considered, an aspect not considered in the methods reviewed thus far. Whilst we have presented AutoPARM with respect to a univariate time series, a multivariate version of AutoPARM is also outlined in Davis et al. (2006).

The parametric assumption of piecewise AR processes as the underlying generating process is strong and may thus not always be appropriate. This parametric assumption is necessary in obtaining consistent CP location estimates with the number of CPs being present also being known. This also results in uncertainty being captured implicitly for the CP location and not at all for the number of CPs present. In addition, parameters are estimated via the typical Yule-Walker equations as in standard time series analysis. Thus, any uncertainty regarding parameters θ is not captured explicitly.

2.9 Bayesian Model Selection methods

Bayesian model selection approaches could also be employed within a CP problem. Bayesian statistics concerns deriving the posterior distribution for the unknown quantities of interest and performing inference on this posterior distribution. Within the CP context, Bayesian approaches are focused on obtaining or approximating the posterior distribution of CP characteristics. Namely, the joint posterior $p(M, \tau_{1:M}|y_{1:n}) = p(M|y_{1:n})p(\tau_{1:M}|y_{1:n}, M)$ is the quantity of interest. Such posterior probabilities can be obtained via applications of Bayes' Theorem and marginalisation. More explicitly,

$$p(\tau_{1:M}, M|y_{1:n}) \propto p(M)p(\tau_{1:M}|M)\ p(y_{1:n}|M, \tau_{1:M}) = p(M)p(\tau_{1:M}|M) \int l(\tau_{1:M}, \theta|y_{1:n})d\theta$$

$$p(M|y_{1:n}) = \sum_{\tau_{1:M}} p(\tau_{1:M}, M|y_{1:n}) \propto p(M) \sum_{\tau_{1:M}} p(\tau_{1:M}|M) \int l(\tau_{1:M}, \theta|y_{1:n})d\theta$$

$$p(\tau_{1:M}|y_{1:n}) \propto p(\tau_{1:M}|M) \int l(\tau_{1:M}, \theta|y_{1:n})d\theta$$

where $p(M)$ and $p(\tau_{1:M}|M)$ denote the prior on the number of CPs and the locations. $p(y_{1:n}|\tau_{1:M})$ denotes the marginal likelihood with respect to CP configuration $\tau_{1:M}$, such that θ has been marginalised out.

The ease in which the posterior is computed is very much dependent on the ease in computing the marginal likelihood $p(y_{1:n}|\tau_{1:M})$, and assumptions placed on the data and the model. For example, if segment independence is assumed, then it is convenient to compute the marginal likelihood as it is the product of segment marginal likelihoods (Eckley et al., 2011). However, in general situations, numerical approximation of $p(y_{1:n}|M)$ is often required to perform the marginalisation. The choice of prior on the number of CPs present and their locations is also an important aspect in calculating the posterior which we shall discuss later on in this section.

An advantage of such Bayesian approaches is that it provides a more explicit quantification of the uncertainty regarding CP characteristics. In addition, the quantities presented above are not conditional on model parameters θ and thus the uncertainty associated with unknown θ has been accounted for. Having obtained the posterior, a variety of inference approaches could thus be applied. This includes Bayes' Factor and Posterior Odds, the ratio between marginal likelihoods and posteriors respectively, which assesses the evidence of one CP configuration over another. For example $\frac{p(y_{1:n}|M=1)}{p(y_{1:n}|M=2)}$ is the Bayes' Factor between one CP being present over two. Larger values of this factor indicate stronger evidence of one CP being present. Bayes' Factor and Posterior Odds can also be used with respect to the posterior of CP configurations. CP estimates can also be obtained by minimising the expected posterior loss function for a suitable loss function. Such Bayesian approaches appear in Smith (1975), Carlin et al. (1992), Stephens (1994) in both single and multiple CP contexts.

An area of ongoing discussion in the Bayesian community is the choice of prior, our initial belief on the unknown quantity of interest. This is known to have an effect on the posterior on which inference is performed. This is no different in a CP context where priors are specified on both the number and location of CPs, $p(M)$ and $p(\tau_{1:M}|M)$. There are variety of ways in which this can be performed, dependent on one's belief. Uninformative priors are often chosen in the Bayesian community if little is known on the unknown quantities. In a CP context this means one does not favour certain CP configurations. As a result, the likelihood has the most influence on the posterior rather than the prior. A naive, misguided prior in achieving this uninformative-ness is to assume the following Uniform distribution on both the number and location of CPs as in Bayesian CP analysis,

$$\begin{aligned}
M &\sim \text{Unif}(\{0, 1, \ldots, M^{\max}\}) \\
p(\tau_{1:M}|M) &= p(\tau_1)\prod_{i=2}^{M} p(\tau_i|\tau_{i-1}) \\
p(\tau_1) &= \frac{1}{n-M} \quad \tau_1 = 2, \ldots, n-M \\
p(\tau_j|\tau_{j-1}) &= \frac{1}{n-\tau_{j-1}-1} \quad \tau_j = \tau_{j-1}+1, \ldots, n-M+j-1, \quad j = 2, \ldots, M.
\end{aligned}$$

Whilst such a prior setup seems to be uninformative via the use of the Uniform distribution, Koop and Potter (2009) show that this is not a case for the location of the CPs with an undesirable clustering effect of CPs towards the end of the data. This effect may not be a true representation of one's uninformative belief and should therefore be avoided if necessary.

In light of this, Koop and Potter (2009) propose the following set of unrestricted uniform priors for the CP location,

$$p(\tau_1) = \frac{1}{\lceil c \cdot n \rceil} \quad \tau_1 = 2, \dots, \lceil c \cdot n \rceil \tag{2.12}$$

$$p(\tau_j | \tau_{j-1}) = \frac{1}{\lceil c \cdot n \rceil} \quad \tau_j = \tau_{j-1} + 1, \dots, \tau_{j-1} + \lceil c \cdot n \rceil, \quad j = 2, \dots, M. \tag{2.13}$$

where c is a tuning parameter controlling the maximum duration for each segment. Larger values correspond to longer segments of data. $\lceil x \rceil$ denotes the ceiling function such that $\lceil x \rceil = \inf\{z \in \mathbb{Z} | x \leq z\}$.

The form of the proposed priors look very similar to that of the "uninformative" uniform priors, although subtle differences occur. Namely CPs can occur beyond the scope of the data. By extending the potential scope of CP instances, this removes the undesirable clustering of CPs towards the end of the data, and thus provides a true uninformative prior for the CP location. In addition, this proposed prior also treats the number of CPs as an unknown with inference now focusing on the number of CPs occurring within the scope of the data. This is despite the number of potential CPs being pre-specified. Nevertheless, the proposed prior provides true uninformative belief and should thus be utilised if an uninformative prior is desired.

An alternative manner to specify a prior on both the number and location of CPs is to consider a prior on the segment length. This prior is introduced with respect to the methodology reviewed in Section 2.13 and we will consider it there in greater detail.

The Bayesian approaches reviewed in this section provide explicit quantification of CP uncertainty in the form of the posterior and is an attractive approach for the problem presented in this thesis. In addition, a Bayesian approach considers the uncertainty associated with the unknown model parameters θ by integrating them out of the joint posteriors obtained. This thus results in CP estimates which are not conditional on specified model parameters. Implementations of these Bayesian methods are scarce and often tailored with a specific problem and application in mind due to the priors and models assumed. Specifying appropriate priors on the number and location of CPs is a difficult task, particularly if it is sensitive on the posterior of interest. This is a potential disadvantage of the Bayesian approaches outlined in this section. If it is thus possible to obtain the posterior of the CP characteristics without having to specify such influential priors on the CP characteristics themselves, this would be a particularly advantageous Bayesian approach. One alternative approach is to specify a prior on the segment durations which is outlined in Section 2.13.

2.10 Reversible Jump Markov Chain Monte Carlo

Reversible jump Markov Chain Monte Carlo (RJ-MCMC) (Green, 1995), is an extension of the Markov Chain Monte Carlo (MCMC) sampling algorithm to sample from target distributions defined on spaces of varying dimensions. A typical application of RJ-MCMC is within Bayesian model selection where the model space varies in dimension with respect to different candidate models, and the potential number of models is unknown. Mixture models with an unknown number of components is a common application of RJ-MCMC (see page 131 of Frühwirth-Schnatter (2005)). Given the setup of the CP problem where the model parameters θ varies in dimension with respect to the number of CPs present and this being unknown, it is evident that RJ-MCMC is a potential method in estimating CP characteristics.

The approach obtains CP estimates by sampling from a joint posterior via the use of MCMC, such that the invariant distribution of the MC is the posterior distribution of interest. The joint posterior is defined with respect to the number of CPs, the CP locations, and the associated model parameters, which is defined as

$$\pi(m, \tau_{1:m}, \theta^{(m)}|y_{1:n}) \propto p(m)p(\tau_{1:m}|m)p(\theta^{(m)}|\tau_{1:m}, m)p(y_{1:n}|\theta^{(m)}, m, \tau_{1:m})$$

where $\theta^{(m)} = (\theta_1, \ldots, \theta_m, \theta_{m+1})$ in this section only. An MCMC sampling algorithm may therefore consider the following moves:

$$\begin{aligned} m_1 &\to m_2 \\ \theta^{(m_1)} &\to \theta^{(m_2)} \\ \tau_{1:m_1} &\to \tau_{1:m_2} \end{aligned}$$

The reversible jump terminology refers to the fact that as the dimension of the posterior varies by assuming a different number of CPs, a mechanism is required such that the sampling MC jumps between these different model spaces. The methodology thus considers both within model moves ($m_1 = m_2$ and thus retain the same dimension), and out of model moves ($m_1 \neq m_2$ and thus varies in dimension) for the sampling MC. In the latter case, a mechanism is required to merge or split the corresponding quantities such as the segment parameters. The sampling MC thus provides samples from the joint distribution, in which the posterior for CP characteristics, $p(M|y_{1:n})$ and $p(\tau_{1:M}|M, y_{1:n})$, can be obtained by marginalisation.

An alternative RJ-MCMC framework is to sample from the CP posterior is via a data augmentation procedure. This procedure introduces a latent process $X_{1:n}$ where X_t can be used to indicate which data segment or generating mechanism Y_t arises from. Interest thus lies in sampling from the posterior $p(x_{1:n}|y_{1:n}) = \int p(x_{1:n}, \theta|y_{1:n})d\theta$ which thus allows one to sample indirectly the respective CP characteristics, for example by determining when there is a change in state in sampled $X_{1:n}$. The number of segments or generating mechanisms present in the data is unknown, and thus the values in which X_t can take, in addition to the potential configuration of $X_{1:n}$, combined with the varying dimension of θ thus results in the RJ-MCMC framework being applicable. The idea of a latent process $X_{1:n}$ being associated with the observed process $y_{1:n}$ is not dissimilar to the ideas presented under the Hidden Markov Model framework approaches reviewed later in this chapter (Section 2.12). An advantage of this alternative RJ-MCMC via a latent process is that priors on the number of CPs present and their locations does not need to specified as they are indirectly determined by the latent process.

RJ-MCMC is a sophisticated Bayesian approach in tackling CP problems. As a Bayesian approach, it allows the quantification of the uncertainty regarding CP characteristics more explicitly by sampling from the respective posterior. This is also in light of parameter uncertainty. Implementations of RJ-MCMC are highly specific to the problem of interest and thus an appropriate open-source implementation for the GNP data does not exist to the best of my knowledge. Certain problems associated with MCMC sampling algorithms are however prevalent. This includes designing efficient sampling moves such that the model spaces are explored sufficiently, and determining whether convergence has been reached. These issues are further exaggerated for a RJ-MCMC framework with instabilities being more common, difficulties in designing good split and merge moves, and a larger number of sampling iterations being required for convergence to be concluded (Fearnhead, 2006).

2.11 Product-Partition models

Barry and Hartigan (1993) and Erdman and Emerson (2008) consider a Bayesian approach to the CP problem by modelling the observed time series as a product-partition model. A product-partition assumes a latent process in addition to the observed time series, which denotes the locations of the CPs and when the parameters switch. More specifically, let $X_{1:n}$ denote the additional latent process which takes values 0 or 1. $X_t = 1$ denotes that a CP occurs at location t for $1 \leq t \leq n$. $p_t = P(X_t = 1)$ denotes that the probability that CP occurs at time t. The use of the latent process is similar to the latent Markov Chain in a HMM framework (Section 2.12) , although the process is treated as a sequence of independent random variables, and the conditional independence amongst the observations given the underlying state sequence is not present. CP inference now centres on postulating the unknown behaviour of $X_{1:n}$, given the observed time series $y_{1:n}$ by sampling from the posterior $p(x_{1:n}|y_{1:n})$.

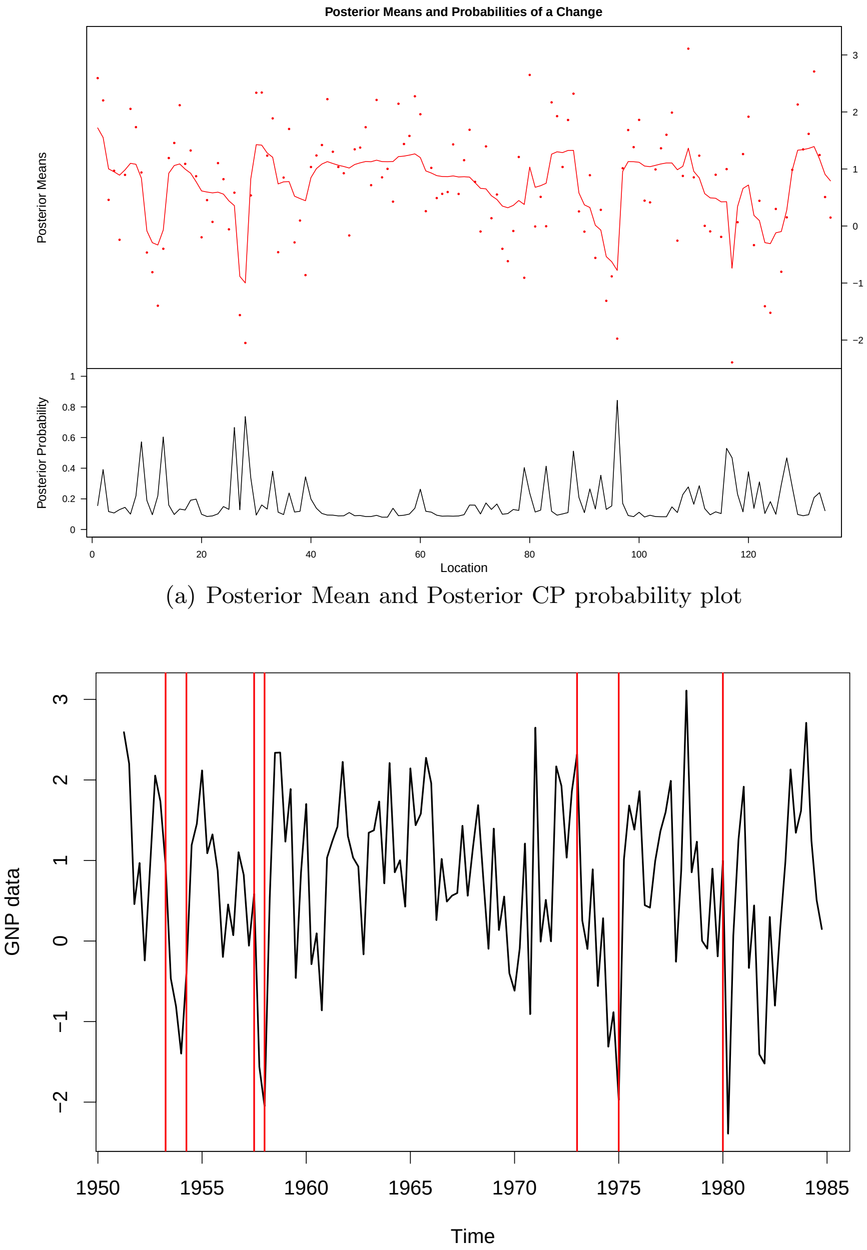
The method assumes Gaussian observations such that the mean differs between each segment, but the variance remains constant during the scope of the data, $y_t \sim N(\mu_i, \sigma^2), i = 1, \dots, M+1$. Independence amongst observations between different segments is also assumed under the framework. An exact inference approach on $X_{1:n}$ is proposed in Barry and Hartigan (1993). However, such an approach is computationally expensive with a computational cost of $\mathcal{O}(n^3)$. In light of this, an approximation is proposed in Erdman and Emerson (2008) which utilises Markov Chain Monte Carlo in sampling from the posterior and performing inference. This reduces the computational cost to $\mathcal{O}(n^2)$. We shall thus outline this approximation method due to its general applicability in CP problems.

CP inference is based on sampling from the posterior of $X_{1:n}$ and the model parameters $\theta = (\{\mu_i\}_{i=1}^{M+1}, \sigma^2)$. That is we sample from $p(x_{1:n}, \theta|y_{1:n})$. This is performed by sampling iteratively from the conditional posterior distributions of $X_{1:n}$ and θ. In sampling $X_{1:n}$, consider sampling X_t, conditional on $y_{1:n}$, θ and all other components of $X_{1:n}$ except t (that is X_j such that $j \neq t$). We refer the reader to Erdman and Emerson (2008) as to how this sampling is specifically performed.

An implementation of the outlined Bayesian Product-Partition method is available in the `R` package `bcp` (Erdman and Emerson, 2007). An application of this method on the GNP data is displayed in Figure 2.6. In particular, we display the posterior means, and the posterior CP probability in Figure 2.6(a), and some initial CP estimates in Figure 2.6(b). These estimates have been obtained by a defining a threshold rule; a CP has occurred when the CP probability exceeds a threshold of 0.5. We thus conclude that seven CPs have occurred at the highlighted locations. These locations correspond to when the mean of the GNP data switches sufficiently. Evidently, these CP estimates are highly sensitive to the threshold used.

The Bayesian Product-Partition method is a sophisticated framework in approximating the posterior distribution of CP characteristics and the associated model parameters. By reporting the posterior distribution, this provides explicit quantification of the uncertainty with regards to CP characteristics. CP estimates can be deduced in the desired manner (for example, thresholding or taking the maximum a posterior estimates). Erdman and Emerson (2008) also extend the framework such that multivariate Gaussian time series can be considered.

However such a method is costly and requires sampling a latent state se-



(a) Posterior Mean and Posterior CP probability plot

(b) Bayesian CP estimates

Figure 2.6: Application of a Bayesian Product Partition CP method on GNP data. Top panel displays the posterior mean and the CPP. Bottom panel presents CP estimates based on a thresholding method on the CPP.

quence under a MCMC sampling regime. As discussed in other CP methods utilising MCMC, it is often difficult to design efficient sampling mechanisms to ensure good mixing and assess convergence. This is more difficult in MCMC algorithms involving sampling latent processes such as this due to their high dimension and correlation.

2.12 Hidden Markov Models based methods

Hidden Markov Models (HMMs) are a popular framework for modelling non-stationary and non-linear time series. Applications include Biology (modelling DNA sequences (Eddy, 2004)), Engineering (speech recognition (Rabiner, 1989)) and Medical (modelling daily epileptic seizure counts of a patient (Albert, 1991)). As they can be used to model non-linearity and non-stationarity within time series, they are also a popular framework for CP analysis. For overviews of HMMs, we refer the reader to MacDonald and Zucchini (1997) and Cappé et al. (2005).

A HMM can be defined as in Cappé et al. (2005): a bivariate discrete time process $\{X_t, Y_t\}_{t\geq 0}$ where $\{X_t\}$ is a latent finite state time-homogeneous Markov chain (MC) with $X_t \in \Omega_X$. The observed process $\{Y_t\}$ is a sequence of independent random variables conditional on $\{X_t\}$ and the conditional distribution of Y_t is completely determined by X_t. Without loss of generality, X_t is assumed to take values in $\Omega_X = \{1, \ldots, H\}, H < \infty$. The underlying states can represent different data generating mechanisms, for example "good" or "bad" days in modelling the number of daily epileptic seizures (Albert, 1991), and thus can be used to capture the non-linearity and non-stationarity of the observed time series Y_t. In specifying a HMM, three components are required:

1. An initial distribution for the underlying MC, $\{X_t\}_{t\geq 0}$, at time 0, that is $P(X_0 = i), \forall i \in \Omega_X$.

2. A transition probability which describes how the underlying MC will evolve over time. For example, $p_{ij} = P(X_t = j|X_{t-1} = i), \forall i, j \in \Omega_X$.

3. An emission probability which describes how the observation's distribution is dependent on the underlying MC. For example, $\gamma_{j,y_t} = f(Y_t = y_t|X_t = j), \forall j \in \Omega_X$, where f is some assumed parametric density.

In the case presented above where the underlying MC is first-order Markov, and the emission probability of Y_t only depends on the underlying state at time t, X_t, we refer to this as the standard HMM. Extensions of the standard HMM exist such that higher order MCs can be considered (see p. 12 of MacDonald and Zucchini (1997)), additional exogenous covariates can be incorporated (see Godfeld and Quandt (1973)), and finite dependency on previous observations and states X_t is permitted for observation Y_t (see p. 357 of Frühwirth-Schnatter (2005)). This thesis will focus on the last extension which are referred to as General Finite State HMMs and are of the form:

$$f(y_t|y_{1:t-1}, x_{1:t}, \theta) = f(y_t|x_{t-r:t}, y_{1:t-1}, \theta) \qquad \text{(Emission)} \tag{2.14}$$

$$p(x_t|x_{1:t-1}, y_{1:t-1}, \theta) = p(x_t|x_{t-1}, \theta) \quad t = 1, \dots, n \qquad \text{(Transition)}. \tag{2.15}$$

where θ denotes the unknown parameters associated with the assumed HMM. The emission density $f(y_t|\cdot)$ describes how the observation distribution depends on a finite number, r, of hidden states and previous observations. The emission density can potentially be any parametric family. Such flexibility in the choice of emission density contributes to the popularity of HMMs as a modelling approach. Associated with the emission density are state dependent parameters which depend on the underlying states of the MC. The transition equation describes how the underlying MC evolves, the simplest setup being that of a first order MC. Extensions to higher order MC behaviour are easily viable via standard embedding arguments (see p. 12 of MacDonald and Zucchini (1997) for example).

In this thesis, the term HMMs is specifically used to refer to the use of a discrete finite state MC, that is $H < \infty$, as in MacDonald and Zucchini (1997), with State Space Models (SSM) referring specifically to Markov Processes defined over an infinite underlying state space Ω_X. Much of the inference and applications of HMMs, including many of the CP methods reviewed later in this section, assume that H, the number of underlying states, is assumed known a priori. This is typically not the case and Chapter 5 will review and propose a method for estimating H.

The model parameters θ, consist of the $H \times H$ transition matrix $\mathbf{P} = \{p_{ij}\}_{i,j \in \Omega_X}$ and the state dependent emission parameters. These parameters will depend on the emission distribution assumed. This can include state dependent emission rates, means and variances for Poisson and Gaussian emission distributions respectively. This leads to Poisson Markov ($Y_t|X_t \sim \text{Poisson}(\lambda_{X_t})$), and Gaussian Markov ($Y_t|X_t \sim \text{N}(\mu_{X_t}, \sigma^2_{X_t})$ models. Not all parameters need to be state dependent however, with some state invariant emission parameters also being applicable. Many inference methods and applications of HMMs are conditional on θ, for example state sequence inference (Viterbi, 1967) and exact CP inference (Aston et al., 2011). However, θ is usually unknown and thus needs to be estimated. The

Expectation-Maximisation (EM) algorithm (Baum et al., 1970) is a popular method which provides a point estimate of θ via maximum likelihood. However, such point estimates usually do not encapsulate the uncertainty that may be associated with the unknown θ. In capturing the potential uncertainty associated with θ, Bayesian approximation of the posterior $p(\theta|y_{1:n})$, is a potential path to consider. We shall explore one potential Bayesian estimation method in Chapter 3.

In many HMM based CP methods, a change in X_t corresponds to a change in the data generating mechanism and thus the statistical properties of the observed time series. That is if $X_{t-1} \neq X_t$, a CP is said to have occurred at the corresponding time. Thus, postulating the potential behaviour of the underlying MC, $X_{1:n}$, with respect to observed $y_{1:n}$ is the key idea in HMM based CP methods. How this state sequence is accounted for is one of the key differences between the various CP methods reviewed.

To aid clarification in this HMM section, the notation of θ and H is suppressed within quantities where necessary, despite many of them being conditioned on them. In addition, these are assumed to be known a priori before analysis with a suitable plug-in estimate being used where necessary. However, in practice, these are unknown and thus need to be estimated. Chapters 3 and 5 consider methods in estimating these quantities and how they can be incorporated within CP analysis.

2.12.1 Deterministic State Sequence Inference

Two popular methods in obtaining a single point estimate of the underlying state sequence $X_{1:n}$, are the Viterbi Algorithm (also known as Global decoding, Viterbi (1967)) and Posterior Decoding (also known as Local Decoding, Juang and Rabiner (1991)). Both of these algorithms are popular in the HMM literature and are not exclusive to CP problems, with applications in speech processing (Rabiner, 1989) and understanding daily epileptic seizures (Albert, 1991) for example.

The Viterbi algorithm (Viterbi, 1967) is a dynamic programming algorithm which computes the most probable state sequence. This is defined as

$$\arg\max_{x_1,\ldots,x_n} P(X_{1:n} = x_{1:n}|Y_{1:n} = y_{1:n}). \tag{2.16}$$

The algorithm for a standard HMM with discrete output (for example a Poisson HMM) is outlined in Algorithm 4 and requires a forward and backward pass through the data. The continuous output case (for example Gaussian HMM) follows analogously. The forward pass computes $\zeta_{t,i}$, the probability of the most probable state sequence ending in state $i \in \Omega_X$ at time t. The backwards pass computes and re-

Algorithm 4 Determining the Viterbi state sequence, the most probable state sequence.

```
Aim: Obtain,
    $\hat{x}_{1:n} = \arg \max_{x_1,\dots,x_n} P(X_{1:n} = x_{1:n}|Y_{1:n} = y_{1:n}).$
Forwards Run: Set
    $\zeta_{1,i} = P(X_1 = i, Y_1 = y_1) = f(Y_1 = y_1|X_1 = i) \sum_{x_0 \in \Omega_X} p_{x_0 i} P(X_0 = x_0).$
for $t = 2, \dots, n$ do
    $\zeta_{t,j} = \max_{x_1,\dots,x_{t-1}} P(X_{1:t-1} = x_{1:t-1}, X_t = j, Y_{1:t} = y_{1:t}) \quad \forall j \in \Omega_X$
    $\qquad = f(Y_t = y_t|X_t = j) \max_{i \in \Omega_X} \{\zeta_{t-1,i} p_{ij}\} = \gamma_{j,y_t} \max_{i \in \Omega_X} \{\zeta_{t-1,i} p_{ij}\}.$
end for
Backwards Run: Set $\hat{x}_n = \arg\max_{i \in \Omega_X} \zeta_{n,i}$.
for $t = n - 1, \dots, 1$ do
    $\hat{x}_t = \arg \max_{i \in \Omega_X} \zeta_{t,i} p_{i,\hat{x}_{t+1}}.$
end for
$\hat{x}_{1:n} = (\hat{x}_1, \dots, \hat{x}_n)$ is the Viterbi state sequence, the most probable state sequence.
```

turns the Viterbi state sequence, $\hat{x}_{1:n}$, by considering the state at time t which leads to the most probable state at the next time $t + 1$. The algorithm is efficient with computational cost $\mathcal{O}(n)$.

Alternatively, the Posterior Decoding algorithm (Juang and Rabiner, 1991) provides an estimate of the underlying state sequence by choosing the states which maximise the marginal smoothed probability for each time t. That is

$$\tilde{x}_t = \arg \max_i P(X_t = i|Y_{1:n} = y_{1:n}) \qquad t = 1, \dots, n.$$

The algorithm is outlined in Algorithm 5 for a standard HMM discrete output, and is computed via the use of the Forward-Backwards equations.

Definition 2. *Forward-Backward equations*

The Forward and Backward probabilities are defined as follows:

$$\alpha_t(i) = P(Y_{1:t} = y_{1:t}, X_t = i) \qquad t = 1, \ldots, n, \forall i \in \Omega_X \tag{2.17}$$

$$\beta_t(i) = P(Y_{t+1:n} = y_{t+1:n} | X_t = i) \qquad t = 1, \ldots, n-1, \forall i \in \Omega_X \tag{2.18}$$

$$\beta_n(i) = 1 \qquad \forall i \in \Omega_X \tag{2.19}$$

These probabilities are computed recursively as demonstrated in Algorithm 5 via the use of the Baum-Welch theorems (Baum et al., 1970). In addition, the Forward-Backwards equations can be used to compute the likelihood for parameter configuration, θ and number of states, H, exactly via,

$$l(\theta, H|y_{1:n}) = P(Y_{1:n} = y_{1:n}|\theta, H) = \sum_{i \in \Omega_X} \alpha_t(i)\beta_t(i) \qquad \forall t = 1, \ldots, n \tag{2.20}$$

Equation 2.20 is important within the HMM literature as it states that the likelihood can be computed without having to sample the unknown underlying state sequence $X_{1:n}$. This is an important property which shall be used throughout this thesis.

Posterior Decoding provides an alternative means of estimating the underlying state sequence. However, a caveat exists as it is possible to obtain an estimate of the underlying state sequence featuring impossible moves under the specified transition probability matrix. This is a result of single states only being considered at time t (hence its alternative name Local Decoding), rather than states and transitions between times as in the Viterbi algorithm.

Having obtained an estimate of the underlying state sequence, $\hat{x}_{1:n}$ and $\tilde{x}_{1:n}$ respectively under the Viterbi and Posterior Decoding Algorithm, CPs can be identified by determining when there is a change in state in the sequence. That is $\hat{x}_{t-1} \neq \hat{x}_t$ for the Viterbi state sequence and analogously for the Posterior Decoding state sequence. Such an approach is simple and intuitive in identifying the number and location of CPs as well as other CP characteristics such as segment lengths.

Implementations of the Viterbi and Posterior Decoding algorithm exist in the `R` package `HiddenMarkov` (Harte, 2012). An application of the two algorithms is demonstrated on the GNP example in Figure 2.7. A 2-state Gaussian Markov Mixture model has been assumed for both algorithms and the maximum likelihood estimates obtained via the EM algorithm have been utilised. Results indicate similar behaviour between both algorithms (identical CP estimates except for one detected in 1973) and the estimates provided by NBER (14 CPs identified corresponding to seven recession periods).

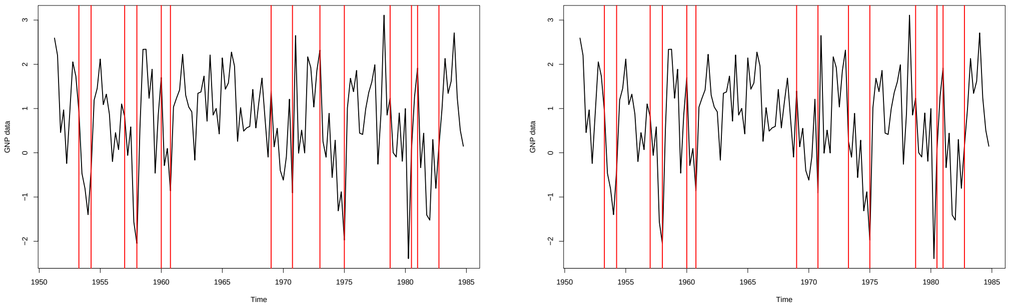
However, the main drawback with the aforementioned algorithms and the

Algorithm 5 Determining the Posterior Decoded State Sequence, the maximum marginal smoothed probabilities.

```
Forward Equations: Compute the forward probabilities, $\alpha_t(j)$.
Set $\alpha_0(i) = P(X_0 = i)$ for $\forall i \in \Omega_X$
for $t = 1, \ldots, n$ do
    $$\alpha_t(j) = \left( \sum_{i \in \Omega_X} \alpha_{t-1}(i) p_{ij} \right) \gamma_{j,y_t} \qquad \forall j \in \Omega_X$$
end for
Backward Equations: Set $\beta_n(i) = 1, \forall i \in \Omega_X$.
for $t = n-1, \ldots, 0$ do
    $$\beta_t(i) = \sum_{j \in \Omega_X} \gamma_{j,y_{t+1}} \beta_{t+1}(j) p_{ij} \qquad \forall i \in \Omega_X$$
end for
Deduce the Posterior Decoded state sequence
    $$\tilde{x}_t = \arg\max_i \frac{\alpha_t(i)\beta_t(i)}{\sum_{j \in \Omega_X} \alpha_t(j)\beta_t(j)} \qquad \forall t = 1, \ldots, n.$$
```

subsequent CP approach is that they provide a single estimate of the underlying state sequence. Within CP inference, these estimates are often used as "deterministically correct" with all CP estimates determined from this single state sequence. It is likely that other state sequences could have led to the observed output and thus different CP configurations may arise from them. Fundamentally, if capturing the uncertainty of CP characteristics is of interest, it would be necessary to postulate all potential state sequences that could have led to $y_{1:n}$. As these algorithms provide only a single state sequence estimate, they do not capture the uncertainty of the underlying state sequence, and thus the uncertainty of the CP estimates.

The Forward-Backward equations presented in Definition 2 are more commonly used to compute the filtering and smoothing probabilities typical in the HMM and SSM literature. Such probabilities denote the probability of the underlying state with respect to partial data up to time t, $P(X_t|y_{1:t})$ (filtering), or conditional on the complete data, $P(X_t|y_{1:n})$ (smoothing). Such probabilities can also be used in forming CP estimates. For example, Hamilton (1989) consider the smoothing probabilities under a particular model, namely Hamilton's Markov Switching Au-



(a) Viterbi state sequence

(b) Posterior Decoding state sequence

Figure 2.7: CP estimation on the GNP dataset via the Viterbi and Posterior Decoding algorithm under a 2-state Gaussian Markov Model framework. CP estimates are almost identical for the two algorithms, with one discrepancy in 1973. These estimates concur with the estimates determined by NBER.

toregressive model of order r, HMS-AR(r). This can be seen as an extension of the Gaussian Markov Mixture model such that dependence on r previous observations is introduced in an autoregressive manner (see Equation 3.34, page 81). Only the mean is state dependent, with variance and AR coefficients and order being state invariant, This model will be discussed and used further in Chapter 3.

In particular, Hamilton (1989) assume a two state HMS-AR(4) model in modelling the US GNP data. where the two underlying states represent "contraction" and "expansion" states of the economy and the autoregressive order of four denotes the annual seasonality from the quarterly data. In determining recession period, an intuitive thresholding argument is used namely

$$y_t \text{ is from a recession regime} \iff P(X_t = \text{“contraction”}|y_{1:n}) > \alpha$$

where Hamilton (1989) consider $\alpha = 0.5$. Under this threshold value and assuming a two state HMM model, this is equivalent to the Posterior Decoding algorithm. The corresponding recession period estimates (grey regions) are presented in Figure 2.8 and generally concur with those provided by NBER and the Viterbi and Posterior Decoding estimates provided in this section. Such a method however is sensitive to the choice of threshold used.

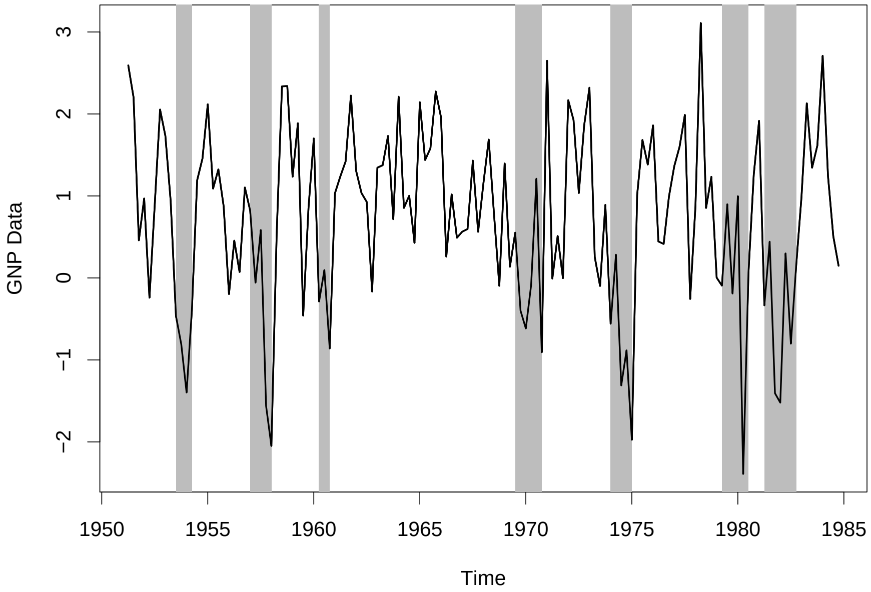


Figure 2.8: Recession estimates (grey regions) provided by Hamilton's Thresholding Method on the smoothed probabilities assuming a 2-state Hamilton's Markov Switching Autoregressive Model of order four (Hamilton, 1989).

2.12.2 Exact CP Distributions

Conditional on a specific model parameter configuration θ and the number of underlying states H, it is possible to compute exact CP distributions under a HMM framework (Aston et al., 2011). The exact nature of this methodology refers to the fact that conditional on θ, results are not subject to sampling or approximation error. The approach forms one of the building blocks of the proposed methodology in Chapter 3. We shall thus review this method in Section 3.2.1, page 60.

This approach provides an efficient and flexible framework in which the uncertainty of several other CP characteristics can also be quantified. This includes the distribution of regime lengths and the probability of a CP falling within a given interval. The main advantage of this approach is that the underlying state sequence is accounted for exactly and does not require sampling which is often a difficult procedure. No approximation or sampling error is thus introduced on estimates.

However, the exact nature of the CP distributions is conditional on θ with a MLE of θ typically being used. As θ is subject to uncertainty itself, it is also important to account for this uncertainty as well. This is particularly important if different configurations of θ give rise to different CP results, despite being equally plausible. We shall return to accounting for parameter uncertainty within this CP approach in Chapter 3.

2.12.3 Constrained HMMs

The HMM framework and methods presented thus far have not placed any restrictions on the behaviour of the underlying MC in that the MC is permitted to visit any of the states freely. Such a HMM is referred to as an unconstrained HMM. Chib (1998) and Luong et al. (2012) consider a constrained HMM such that the underlying MC is restricted to move in a particular way, and construct CP methods around this framework.

Under the constrained HMM framework, the underlying MC cannot return to previously visited states. In a CP context, this results in the underlying states corresponding to the segments between two consecutive CPs. Thus if there are M CPs, then the data is partitioned into $M+1$ segments and the assumed constrained HMM has $H = M+1$ underlying states. As the number of underlying states is assumed known a priori for a HMM whether constrained or unconstrained, this consequently means the number of CPs is known a priori under the constrained HMM framework.

The behaviour of the underlying MC is more formally constrained to move in the following manner. Firstly, $X_0 = X_1 = 1$ and $X_n = H = M+1$. That is, the latent MC and observation process must start in the first segment, and end in the last segment. Secondly, the underlying MC is constructed such that it is unable to return to previously visited segments and thus states. There are consequently only two possible moves for the underlying chain at each time t. Explicitly, if $X_t = i, i = 1, \ldots, M$, then either

(i) Remain in the current state and segment, thus $X_{t+1} = X_t = i$.

(ii) Alternatively, move to the next segment and state in the state space. Thus, $X_{t+1} = i+1 \neq X_t = i$

$\mathbf{P}$, the corresponding transition matrix, is a matrix with non-zero entries on the diagonal and immediate super-diagonal, and zeroes elsewhere. That is $p_{ij} > 0$ if $j = \{i, i+1\}$, else $p_{ij} = 0$. Under this setup, each row of the transition matrix only has one unknown transition probability as $p_{i,i+1} = 1 - p_{i,i}$. Such restriction on the transition matrix needs to be accounted for in parameter estimation methods in order to maintain the constrained HMM framework.

Luong et al. (2012) provide a method in which the posterior CP probability and confidence intervals for CP location estimates can be computed via the use of a constrained HMM framework. These pre-determined location estimates could be provided by CP estimates computed under the Viterbi or Posterior Decoding algorithm discussed earlier, or by alternative means. Via the Forward-Backward Equations, it is shown that the probability of a CP occurring at a specified time, can be computed in addition to usual smoothed probabilities under a constrained HMM framework. That is, for $i = 1, \dots, M$,

$$\begin{aligned} P(i\text{th CP at time } t+1) &= P(X_{t+1} = i+1, X_t = i | y_{1:n}) && (2.21) \\ &= \frac{\alpha_t(i)\beta_{t+1}(i+1)p_{i,i+1}f(y_{t+1}|X_{t+1} = i+1)}{\alpha_1(1)\beta_1(1)} && (2.22) \end{aligned}$$

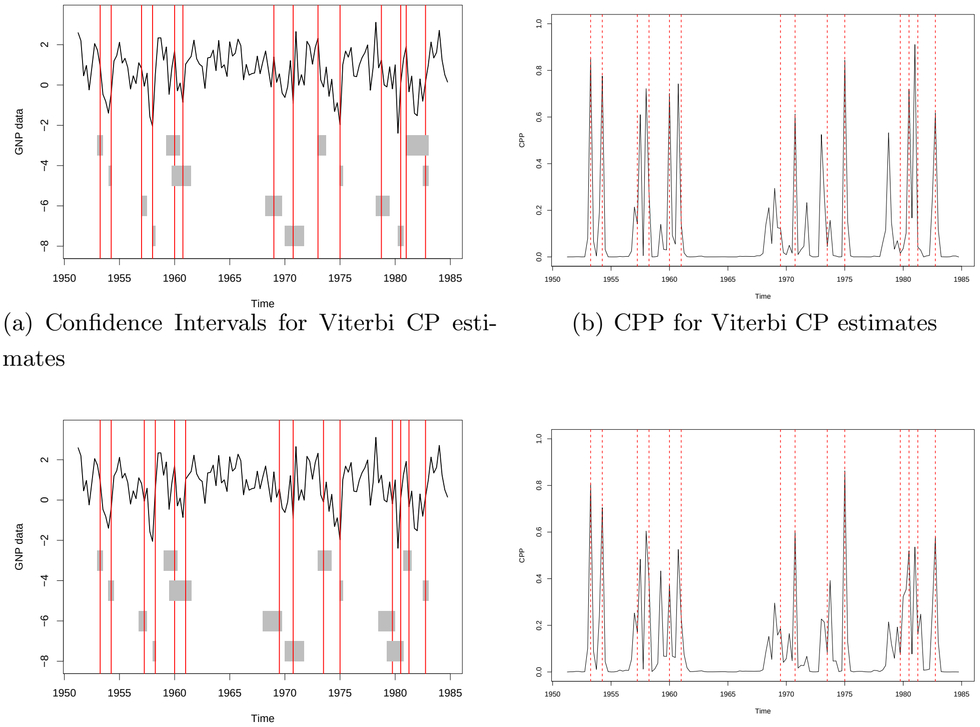
Such probabilities can thus be used to determine the CP probability (CPP, the probability of any CP occurring at a specified time). The α confidence intervals for the ith CP location, (L_i^α, U_i^α) can also be provided by:

$$\begin{aligned} L_i^\alpha &= \inf\left\{ L \in \{1, \dots, n\} | \sum_{t=1}^{L} P(i\text{th CP at time } t+1) \geq \frac{1-\alpha}{2} \right\} \\ U_i^\alpha &= \inf\left\{ U \in \{1, \dots, n\} | \sum_{t=1}^{U} P(i\text{th CP at time } t+1) \geq \frac{\alpha+1}{2} \right\} \end{aligned}$$

Such quantities provide quantification of the uncertainty regarding the CP location.

An implementation of the methodology is provided in the `R` package `postCP` (Nuel and Luong, 2012) and its application on the GNP dataset are displayed in Figure 2.9. We consider the 95% confidence intervals and CPP plot for the Viterbi and NBER CP estimates, assuming a 2-state Gaussian Markov Mixture model . We observe that the confidence intervals are a mixture of narrow and wide (the initial CPs and the middle CPs respectively), highlighting that some of the CP estimates provided are more certain than others and other CP configurations are possible. The CPP plots provide further reasoning as to the shape and behaviour of the confidence intervals, with narrow intervals associated with centred and peaked CPPs, and wide intervals associated with more diffused CPPs around the CP estimates. Such CPP behaviour corresponds to how the GNP data is behaving and whether the CPs are obvious or not. By quantifying the uncertainty of CPs via the CPP plot for example, this provides a better understanding of the data and the CP estimate.

Whilst the uncertainty of CP locations has now been addressed, there are several disadvantages to such an approach, namely that CP location estimates need to be provided preliminary and this is also dependent on the number of CPs being known a priori. Luong et al. (2012) remark that the accuracy of the CP posterior



(a) Confidence Intervals for Viterbi CP estimates

(b) CPP for Viterbi CP estimates

(c) Confidence Intervals for NBER CP estimates

(d) CPP for NBER CP estimates

Figure 2.9: Confidence Intervals (grey bars) and Changepoint Probability (CPP) plots for the Viterbi and NBER estimates on the GNP dataset. These quantities are computed via a constrained HMM framework as proposed in Luong et al. (2012).

probabilities reported are highly dependent on the estimates of the CP locations and number provided, due to its influence in the estimation of θ. This is demonstrated in the GNP implementation (see Figure 2.9, around 1980) where the CPP plots are noticeably different for the two sets of CP estimates initially provided. Such sensitivity is not particularly desirable or sensible if CP characteristics are generally unknown.

Chib (1998) propose a framework in which the uncertainty of CP locations is quantified more explicitly by considering the uncertainty of the underlying state sequence. This is performed by sampling from the posterior of the underlying state sequence, $p(x_{1:n}|y_{1:n})$, and thus sampling the location of CPs when there is a change in state in the underlying state sequence. That is $X_t = i \neq X_{t+1} = i + 1$ for $i = 1, \ldots, M$.

Sampling the underlying state sequence is achieved by sampling from the joint posterior distribution of the model parameters and underlying state sequence,

$p(x_{1:n}, \theta | y_{1:n}, H)$. This is typically not a conventional, standard distribution and thus a MCMC sampling scheme is employed. In particular, they iteratively sample from the following two full conditionals,

- $\theta | y_{1:n}, X_{1:n} = x_{1:n}$
- $X_{1:n} | y_{1:n}, \theta$.

It is thus possible to obtain a posterior of the state sequence by marginalising out the model parameters from the joint posterior, $p(x_{1:n}|y_{1:n}, H) = \int p(x_{1:n}, \theta | y_{1:n}, H) d\theta$. Consequently a posterior of the CP locations can be obtained by determining when there is a change in state in the sampled state sequence from its posterior.

Chib (1998) also provide an ad-hoc solution in determining the number of underlying states and thus the number of CPs. This is achieved by framing the unknown number of CPs problem as a Bayesian model selection problem, similar to that explored in Section 2.9. Each model assumes a different number of states and thus number of CPs. The marginal likelihood can thus be approximated for each model, and Bayesian model selection methods such as Bayes' factor can be employed in determining which model is suitable, and thus how many CPs to assume.

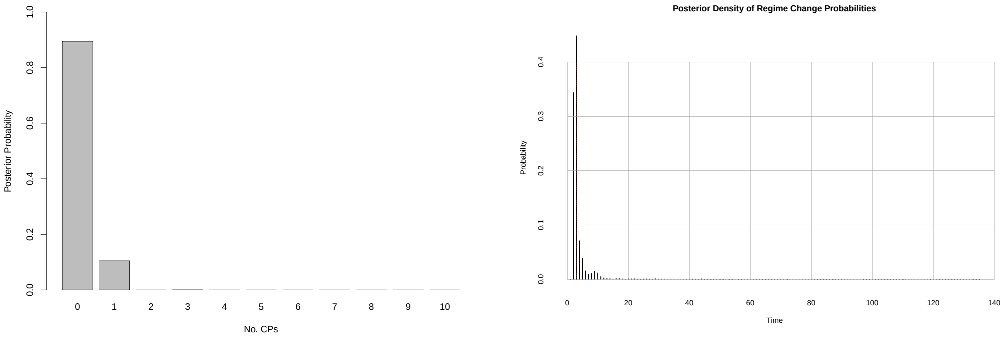
Chib (1998) remark that the marginal likelihood, $p(y_{1:n}|H = h)$ which assesses the likelihood of the data arising from a model assuming $H = h$ states, can be approximated and obtained additionally from the MCMC sampling algorithm for the joint posterior distribution of the underlying state sequence and parameters.

Having obtained the marginal likelihood, the model posterior distribution can also be approximated in combination with a model prior. Chib (1998) use the Bayes' Factor to determine which model, and thus how many CPs, to assume. Bayes' Factor in assessing the relative evidence of one model over another. Thus, suppose one wants to assess whether to assume m_1 or m_2 CPs, and consequently whether to assume $m_1 + 1$ or $m_2 + 1$ underlying states in a constrained HMM framework. Then the Bayes' Factor between these two models is defined as,

$$B_{m_1,m_2} = \frac{p(y_{1:n}|H = m_1 + 1)}{p(y_{1:n}|H = m_2 + 1)}. \tag{2.23}$$

Larger values of B_{m_1,m_2} indicate that the data supports a model assuming m_1 CPs over m_2 CPs.

Figure 2.10 displays the results of Chib's implementation on the GNP example. In particular, we assume the GNP data arises from a Gaussian Markov Mixture model such that the mean and the variance are state dependent. As the number of CPs is unknown a priori, this needs to be estimated firstly. We consider models with



(a) Posterior Distribution of Number of CPs (b) CP probability assuming one CP occurs

Figure 2.10: Posterior Distribution of Number of CPs and location of first CP under the constrained HMM framework of Chib (1998). Zero CPs are most probable but if a single CP is assumed to have occurred, then this is most likely to occurred towards the beginning of the data.

zero to ten CPs and approximate their respective posterior distributions, assuming a Uniform prior over the number of CPs (Figure 2.10(a)). As the posterior highlights, zero CPs are the most probable, with some probability associated with one recession potentially occurring. The use of Bayes' Factor also concludes the same result. Up to 14 potential CPs were also considered in concordance with the 14 detected by NBER; identical results were achieved with nearly all probability mass on zero CPs occurring.

We could thus conclude that no CPs have occurred during the data if we take the maximum a posterior estimate of the number of CPs. However, if we condition that one CP has occurred, this CP appears to occur towards the beginning of the data.

The constrained HMM approach as proposed by Chib (1998) provides a state-of-the art framework in tackling CP problems and providing quantification of CP characteristics. The uncertainty is captured by sampling the underlying state sequence via a MCMC algorithm, and model parameter uncertainty is captured by marginalising out this quantity. However, this is typically a high-dimensional correlated vector and thus care is required in designing good moves such that the sampling MC is mixing well. In addition, the uncertainty of both the number and location of CPs are not considered simultaneously which may be desired.

2.13 Exact Sampling of the Posterior via Recursions

Fearnhead (2005); Fearnhead and Liu (2007) propose a framework in which exact sampling from the CP posterior distribution can be performed in an offline and online context. The exact and efficient sampling relies on the assumption that segments are independent, conditional on the CP locations. Assuming such a conditional independence assumption results in probability recursions which allow exact sampling to be performed. These recursions are similar to the Forward-Backward algorithm in HMMs. In addition, the CP posterior distributions sampled from are not conditional on model parameters, and thus the CP estimates obtained are in light of parameter uncertainty (the CP estimates are not conditional on specific model parameter configurations).

The framework proposed also provides an alternative elicitation approach in specifying the prior over the CP characteristics. This alternative prior considers the distribution of the segment lengths and is derived by modelling the event of a CP as a point process. This prior setup indirectly implies a prior on both the number and locations of CPs. This segment prior will be assumed in reviewing the methodology and we refer the reader to Fearnhead (2006) with regards to the implementation of standard priors directly on the CP characteristics of interest.

In this section only, we review the exact sampling methodology as in Fearnhead and Liu (2007), the online context with the offline scenario following in a similar manner. Under this methodology τ is a CP if it segments the data into $y_{1:\tau}$ and $y_{\tau+1:n}$. Under this definition of a CP, $\tau_0 = 0$ and $\tau_{M+1} = n$. The constraints on the intermediary CPs remain unchanged. Let $g(t)$ denote the probability mass function for a segment of length $t \in [1, n-1]$. Let $G(t) = \sum_{s=1}^{t} g(s)$ denote the corresponding distribution function of the segment length, and let $G^C(t) = 1 - G(t)$. Thus the prior probability of M CPs occurring at locations $\tau_{1:M} = (\tau_1, \ldots, \ldots, \tau_M)$ is:

$$p(\tau_{1:M} = (\tau_1, \ldots, \ldots, \tau_M)) = \left(\prod_{j=2}^{M} g(\tau_j - \tau_{j-1}) \right) G^C(n - \tau_M).$$

This alternative prior setup is equivalent to the usual prior defined over CP locations. Typical segment priors implemented are those from the negative Binomial family such as the Geometric distribution, and result in a Binomial prior on the number of CPs. Specifying a prior over the segment length can often be more intuitive and natural compared to the usual practice of specifying a prior over the potential CP locations. For example, prior information and beliefs with respect to the length of segments may be more accessible, and segment priors do not need to be adapted if the length of the time series changes.

The exact sampling approach samples from the joint posterior of the CP characteristics, $p(M, \tau_{1:M}|y_{1:n})$, by performing a forward and backwards pass on the data. The forward pass is essentially a filtering recursion which computes filtering probabilities for a latent variable denoting the time of the most recent CP. The backwards pass simulates the changepoints of interest by traversing backwards in time. Before proceeding with the main framework, it is necessary to introduce the partial marginal likelihood,

$$P(s, t, q) = \int p(y_{s-1:t}|\theta, \text{model } q)p(\theta|\text{model } q)d\theta, \tag{2.24}$$

where $p(\theta|\text{model } q)$ is the model parameter prior by assuming model q. It is assumed that this partial marginal likelihood can be computed for all $s < t$ and q, either by assuming conjugate priors for θ or numerical integration. The model q is one model from a set of Q possible models for the data from each segment, for example each model could assume a different regression model. Consequently, this methodology is not limited by the types of changes compared to others. The model prior is denoted by $p(q)$.

The latent process introduced is denoted by C_t, which captures the time of the most recent CP prior to time t. Consequently, the variable takes values from $C_t \in \{0, 1, \ldots, t-1\}$ where $C_t = 0$ denotes that no CP as occurred prior to time t. At time t there are only two possible moves; either $C_t = C_{t-1}$ or $C_t = t-1$ which indicates that $t-1$ is not and is a CP respectively. C_t can be thought of as a Markov Chain with the corresponding constrained behaviour. The transition probabilities for this latent MC are based on the distribution of the segment durations as follows:

$$P(C_{t+1} = j|C_t = i) = \begin{cases} \frac{G^C(t-i)}{G^C(t-1-i)} & \text{if } j = i \ (t \text{ is not a CP}), \\ \frac{g(t-i)}{G^C(t-1-i)} & \text{if } j = t \ (t \text{ is a CP}), \\ 0 & \text{otherwise.} \end{cases}$$

The forward pass of the algorithm concerns computing the filtering probability of this MC, that is $P(C_t = i|y_{1:t})$, in a recursive manner. From the standard filtering recursions,

$$\begin{aligned} P(C_{t+1} = j|y_{1:t+1}) &\propto P(y_{t+1}|C_{t+1} = j, y_{1:t})P(C_{t+1} = j|y_{1:t}) \\ &= P(y_{t+1}|C_{t+1} = j, y_{1:t})\sum_{i=0}^{t-1} P(C_{t+1} = j|C_t = i)P(C_t = i|y_{1:t}). \end{aligned} \tag{2.25}$$

Then it can be shown that the recursions are

$$P(C_{t+1} = j|y_{1:t+1}) \propto \begin{cases} \frac{\sum_{q=1}^{Q} P(j,t+1,q)p(q)}{\sum_{q=1}^{Q} P(j,t,q)p(q)} \frac{G^C(t-i)}{G^C(t-1-i)} P(C_t = j|y_{1:t}) & \text{if } j < t, \\ \frac{\sum_{q=1}^{Q} P(j,t+1,q)p(q)}{\sum_{q=1}^{Q} P(j,t,q)p(q)} \sum_{i=0}^{t-1} \frac{g(t-i)}{G^C(t-1-i)} P(C_t = j|y_{1:t}) & \text{if } j = t, \end{cases}$$

where $P(C_1 = 0|y_1) = 1$ is the initialisation setting.

Having obtained and stored these filtering probabilities, $P(C_t = i|y_{1:t})$ for all $t = 1, \dots, n$ and $i = 0, \dots, t-1$, a backwards pass is then performed to sample from the joint posterior distribution of the CP locations. To obtain one sample of a CP configuration from the joint posterior, we begin by simulating the location of the last CP using the probability $P(C_n|y_{1:n})$. Denote this sampled CP location as t. If $t = 0$, terminate the algorithm as this indicates no CPs have occurred. Else if $t > 0$, the next CP is simulated backwards in time from the conditional distribution:

$$\begin{aligned} P(C_t = i|y_{1:n}, C_{t+1} = t) &\propto P(C_t = i|y_{1:t})P(C_{t+1} = t|C_t = i) \qquad \text{for } i = 1, \dots, t-1, \\ &= P(C_t = i|y_{1:t}) \frac{g(t-i)}{G^C(t-1-i)}, \end{aligned}$$

which utilises the fact data after CP at time t is independent of the CP prior to time t. We continue this simulation process until $C_t = 0$. This provides a sample of CP locations and thus the number of CPs from the joint posterior. This sampling recursion is efficient since these probabilities only need to be calculated once throughout the whole sampling algorithm.

Fearnhead (2006); Fearnhead and Liu (2007) also develop a Viterbi algorithm in calculating the maximum a posterior (MAP) CP estimates and the model for each segment. Let $\mathcal{M}_s$ indicate the MAP choice of CP configuration and models prior to time s, given that a CP occurs at time s. Then for $t = 1, \dots, n$, $s = 0, \dots, t-1$ and $q = 1, \dots, Q$

$$P_t(s,q) = P(C_t = s, \text{model } q, \mathcal{M}_s, y_{1:t}) \quad \text{and} \quad P_t^{\text{MAP}} = P(\text{CP at } t, \mathcal{M}_s, y_{1:t}).$$

Then the following equations provide the MAP estimates regarding the CP and models,

$$P_t(s,q) = G^C(t-s-1)P(s,t,q)p(q)P_s^{\text{MAP}} \quad \text{and} \quad P_t^{\text{MAP}} = \max_{s,q} \left\{ \frac{P_t(s,q)g(t-s)}{G^C(t-s-1)} \right\}.$$

An implementation of the offline method is available on the author's website[1]. This

[1] http://www.maths.lancs.ac.uk/~fearnhea/software/ARPS.html

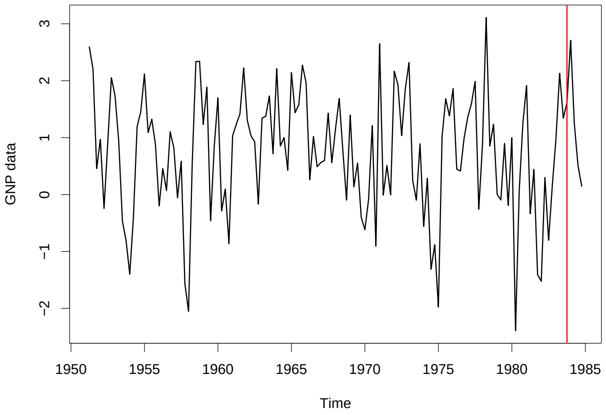


Figure 2.11: Maximum A Posterior Estimates of the CP locations for GNP data using the exact sampling approach of Fearnhead (2005). Piecewise autoregressive models of order up four have been considered.

assumes a piecewise constant autoregressive model for each segment. Figure 2.11 displays the MAP estimates of the CP locations when applied to the GNP data. Autoregressive models of order up to four have been considered due to the belief that there is some annual seasonality present in the data. We observe that a single CP has been identified towards the end of the data. This again provides another CP configuration which is different to NBER's estimates. However, it is believed that few CPs have been identified under this model since the piecewise constant autoregressive model considered, assumes a constant zero mean in each of the autoregressive models. Thus, if the GNP data is suspected to contain changes in mean, it is unlikely that this method will be able to identify the CPs.

This methodology has the advantage over many other Bayesian sampling methods in that it can sample directly and efficiently from the CP posterior of interest. The exact sampling is favourable compared to approximations via MCMC for example, in that it is not necessary to design good mixing algorithms and one need not worry about whether our sampling Markov Chain has reached convergence. The exact algorithm also has a computation cost of $\mathcal{O}(n^2)$ due to support of C_t increasing linearly with t. An approximation is possible such that the summation in Equation 2.25 is truncated due to the majority of previous filtering probabilities $P(C_t = i|y_{1:t})$, being negligible. Such an approximation only introduces negligible approximation error according to empirical results (Fearnhead and Liu, 2007). An additional advantage compared to the quantification of CP uncertainty as proposed in Aston et al. (2011) (see Chapter 3) is that it considers parameter uncertainty. However, this requires computing the segment marginal likelihood $P(s, t, q)$, which may not be possible directly and may thus require some numerical approximation.

2.14 Conclusion

This chapter has presented an overview of a variety of CP methods in the literature. These methods are based on a variety of different assumptions placed on the data including the underlying distribution, the type of change suspected and whether observations are independent or not. In addition, the CP problem can also be perceived in a variety of different perspectives in which statistical literature may be more developed for the alternative perspective considered. This includes hypothesis testing as in the AMOC setup, model selection for penalised log-likelihood approaches, and the use of latent processes in HMM based approaches. CP methods can also be characterised as to whether they are frequentist or Bayesian and thus how explicit they are with regards to CP uncertainty.

Several of the reviewed methods have been applied to the running example of Hamilton's GNP data and successfully demonstrated that quite different CP results can be obtained. This motivates the need to assess the plausibility of the CP estimates provided and the performance of the various CP approaches available. Quantifying the uncertainty of CPs thus provides a means of doing so.

The majority of the CP approaches reviewed in this chapter do provide some quantification of CP uncertainty. For frequentist approaches however, this is often implicit via the use of significance levels (AMOC approaches) or via asymptotic arguments (for example penalised log-likelihood). In addition, the CP uncertainty may also be partially captured. For example, penalised log-likelihood approach via Bayesian Information Criterion can only provide a consistent estimate of the number of CPs and not their respective locations, and AutoPARM provides consistent CP location estimates if the number of CPs is known. Bayesian CP methods are often more explicit with regards to CP uncertainty via the derivation of the CP posterior. However, many partially capture the CP uncertainty to some degree (for example in Chib (1998), the posterior of the CP locations is conditional on the number of CPs), whilst others require MCMC sampling and numerical approximation to obtain quantities such as the marginal likelihood. This can be difficult and computationally costly to obtain, particularly those involving sampling long latent vectors due to their high dimensional and induced correlation. Fearnhead (2006) appear to provide a promising approach in fully capturing CP uncertainty for both the number and location of CPs, for a variety of potential changes. However, the trade-off in doing so is that it is computationally intensive and highly specific to the problem of interest.

Another important aspect to consider is how the unknown model parameters θ, are accounted for. Frequentist approach such as Global Segmentation (Braun and Müller, 1998) and AutoPARM (Davis et al., 2006), estimate θ via maximum likelihood and condition on these values in their respective methods. Any uncertainty associated with θ is captured implicitly via the use of consistency arguments and not considered within the CP results. This approach does not seem entirely desirable if CP results are sensitive to the θ that they are conditioned on. Bayesian approaches are considerably more explicit with regards to the uncertainty of θ and incorporating this into the CP results. This is performed by integrating out θ from the joint posterior involving θ and the CP quantities. A Bayesian CP approach thus provides a more promising path in tackling CP problems as we can account for the uncertainty of θ in some manner, and remove its sensitivity on the CP results of interest.

The Hidden Markov Model framework is an attractive CP framework as it allows a wide range of changes and parametric emission distributions to be considered. In addition it provides an intuitive framework in that the latent Markov Chain represents how the underlying system may be behaving, and allows dependent observations to be modelled. In particular, the approach proposed by Aston et al. (2011) (see Section 3.2.1, page 60) is a promising HMM approach in that it efficiently computes conditional exact CP distributions. A noticeable advantage of this approach is that the underlying state sequence is accounted for exactly and is not sampled compared to other methods involving latent processes (for example Chib (1998), Green (1995), Fearnhead (2006)). This is a particular benefit as it reduces the computational cost with Figure 4b of Aston et al. (2011) showing that a large number of samples are required before the difference between an exact and simulation based estimation procedure becomes negligible. However this exact CP approach is conditional on θ and thus does not consider the uncertainty associated with θ.

A large proportion of this thesis is thus focused on how we can account for parameter uncertainty and how this can be incorporated within the conditional exact CP approach proposed in Aston et al. (2011). Chapter 3 reviews this proposal with subsequent chapters showing further extensions of this framework with respect to model selection and changes in autocovariance structure. The latter further develops CP methods concerning changes in autocovariance in which there are relatively few methods in comparison to changes in mean and variance. In addition to accounting for parameter uncertainty, the core framework proposed retains the exact nature of the underlying state sequence and the CP distribution computed from it, and sampling error is only introduced in sampling θ. This thus provides a flexible, efficient Bayesian CP approach in comparison to other Bayesian methods.

Chapter 3

Exact Changepoint Distributions and Sequential Monte Carlo Samplers

Jack Donaghy: First of all, never bad mouth synergy.

"Retreat to Move Forward", Episode 3.09, 30 Rock, Tami Sagher

3.1 Introduction

Detecting and estimating the number and location of changepoints (CPs) in time series is becoming increasingly important as both a theoretical research problem and a necessary part of data analysis. Chapter 2 has highlighted that a variety of different CP methods exist, each assuming different assumptions and often providing different CP estimates regarding the number and location of CPs for example. In addition, many of these methods fail to capture fully or explicitly the uncertainty associated with CPs, with those which do capture the uncertainty explicitly requiring simulation of large vectors of dependent latent variables. It is important to account for the uncertainty of CPs in a bid to assess the confidence of CP estimates and provide a better understanding of the data analysed.

This chapter proposes a methodology which fully quantifies the uncertainty of CPs for an observed time series, without estimating or simulating latent state sequences. The absence of such simulation is desirable in some settings where a reduction in computational cost is important for example, and is thus one significant motivation of the technique proposed in this chapter.

The proposed methodology is based upon three areas of existing work in the literature. We model our observed time series and consider CPs in a Hidden Markov Model (HMM) framework. HMMs and the general use of dependent latent state variables are widely used in CP estimation (Chib, 1998; Fearnhead, 2006; Fearnhead and Liu, 2007). In these approaches, each state of the underlying chain represents a segment of data between CPs and thus a CP 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 CP has occurred), or move to the next state in the sequence, corresponding to a new segment and thus a CP has occurred. Returning to previously visited states is thus not possible. Interest now lies predominantly in determining the latent state sequence (usually through simulation, by MCMC for example), in order to determine the relevant CP characteristics. In the case of Chib (1998), this consequently means the number of CPs is assumed known which may appear restrictive since these are also often unknown and of interest.

We consider an alternative use of HMMs 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 CPs to be unknown a priori and inferred from the data. We assume that the number of different underlying states is known a priori, a common assumption made in the HMM literature. This latter point seems less restrictive in a CP context than assuming the number of CPs which are usually of great interest. However, Chapter 5 proposes a method for estimating the number of underlying states if necessary. By modelling the observations under a HMM framework, we are able to compute exactly the likelihood via the Forward equations (Rabiner, 1989), which does not require the underlying state sequence to be estimated or sampled.

We also consider a generalised definition of CPs corresponding to a *sustained* change in the underlying state sequence. This means that we are alternatively looking for runs of particular states in the underlying state sequence which corresponds to a CP into a particular regime. 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 CP distributions to be computed, conditional upon model parameters. In practice, it is common for these parameters to be treated as known and fixed, with MLEs typically being used. In most applications where parameters are estimated from the data itself, it is desirable to account for parameter uncertainty in CP estimates. As the above approach provides posterior CP distributions conditional on a parameter, it seems natural to extend this Bayesian approach to account for parameter uncertainty.

Recent Bayesian CP approaches have dealt with model parameter uncertainty by integrating the parameters out in some fashion in order to ultimately sample from the joint CP posterior. This is usually achieved by also sampling the aforementioned latent state sequence (Chib, 1998; Fearnhead, 2006). However, this introduces additional sampling error into the CP 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 posterior distribution of the model parameters via Sequential Monte Carlo. This does not require simulating the latent state sequence as we exploit the exact computation of the likelihood under a HMM framework. This approach introduces sampling error only in the model parameters and retains, conditionally, the exact CP distributions: we will show that this amounts to a Rao-Blackwellised form of the estimator, a variance reduced estimator.

Quantifying the uncertainty in CP 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 CPs can lead to a better understanding of the data and the system generating the data. Whenever estimates are provided for the location of CPs, we should be interested in determining how confident we are about these estimates and whether other CP configurations are plausible. In many situations, it may be desirable to average over models rather than choosing a most probable explanation. In addition, different CP approaches can often lead to different estimates when applied to the same time series, as demonstrated successfully in Chapter 2. This motivates the need to assess the performance and plausibility of these different approaches and their estimates. Quantifying the uncertainty provides a means of so doing.

As a motivating example, we return to the US GNP data presented in Chapter 1, Figure 1.1 (page 5) and analysed throughout Chapter 2. By quantifying the uncertainty of the recessions, our CPs in this instance, we can express the confidence of NBER's recession estimates and if any other recession configurations are possible.

The exact CP distributions computed via FMCI methodology (Aston et al., 2011) 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 also consider parameter uncertainty because the use of different model parameters can give quite different CP 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 CPs, we are able to account for several types of CP behaviour under a variety of model parameter scenarios and thus fully quantify the uncertainty regarding CPs. This is demonstrated in both the simulated data and Econometric GNP data we shall analyse.

The remainder of this chapter has the following structure: Section 3.2 details the statistical background of the methodology which is proposed in Section 3.3. This methodology is applied to both simulated and Econometric GNP data in Section 3.4. Section 3.5 concludes the chapter with some discussion of our findings and potential paths for future work.

3.2 Background

Let $y_{1:n} = (y_1, \ldots, y_n)$ be an observed time series which is potentially non-stationary. This non-stationarity could be due to a changing mean, variance or covariance present in the observations. Let $Y_{1:n} = (Y_1, \ldots, Y_n)$ denote a general sequence of random variables. One particular framework for modelling such a time series is via Hidden Markov Models (HMMs), as discussed in Section 2.12 (page 35), where $\{X_t\}_{t \geq 0}$ denotes our unobserved underlying MC. The methods presented in this chapter and thesis are applicable to general finite state HMMs such that finite dependency on previous states of X_t and previous observations can be incorporated.

Let θ be our unknown model parameters associated with the HMM that need to be estimated. These are dependent on the emission density assumed, but typically consist of the transition probability matrix $\mathbf{P}$ and parameters associated with the emission density, of which some must be dependent on the underlying MC X_t.

We stress that the HMM framework defined in Equation 2.14 (page 36), and indeed throughout the entire HMM literature, is conditional on the number of underlying states H being known a priori. This is typically not the case and is pre-specified prior to statistical analysis such as parameter estimation. Throughout this chapter and Chapter 4, we assume that H is known priori to analysis. Chapter 5 shall address how one may want to estimate the number of underlying states.

A common definition within the HMM framework and the use of latent vectors in modelling time series, is that a CP has occurred at time t whenever there is a change in state in the underlying MC or latent process, that is $X_{t-1} \neq X_t$. This definition is currently adopted in existing work such as Hamilton (1989); Chib (1998); Durbin et al. (1998); Fearnhead (2006). However, in some applications, a sustained change is required before a change to a new regime is said to have occurred. Examples include Economics where a recession is said to have occured when there are at least two consecutive negative growth (contraction) states, 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, before being deemed to have occurred. Motivated by such instances and applications, we define a sustained CP as follows.

Definition 3. *A changepoint to a regime occurs at time t when a change in state persists for at least k_{CP} time periods. That is*

$$X_{t-1} \neq X_t = \ldots = X_{t+j} \tag{3.1}$$

where $j \geq k_{\mathrm{CP}} - 1$.

For example, in the Economic example concerning recession analysis, $k_{\mathrm{CP}} = 2$ and interest lies in the sustained changes to the "contraction" state. This definition can be interpreted as a generalised version of the "change in state" definition defined on a suitably defined space and it is both easier to interpret and computationally convenient to make use of this explicit form. The standard CP definition can be recovered by setting $k_{\mathrm{CP}} = 1$.

A graphical representation for this CP definition on the standard HMM is presented in Figure 3.1. This graphical representation highlights two important features. Firstly, similar to the other HMM based CP methods reviewed in Section 2.12, CP analysis is based on inference of the underlying state sequence of X_t. Secondly, rather than analysing for changes in state in the underlying MC, attention turns to analysing for runs in state of a minimum length k_{CP} in the underlying MC. This latter point motivates one of the main building blocks of the proposed methodologies in this chapter.

Interest often lies in determining the time of a CP and the number of CPs occurring within a time series. Let $M^{(k_{\mathrm{CP}})}$ and $\tau^{(k_{\mathrm{CP}})} = (\tau_1^{(k_{\mathrm{CP}})}, \ldots, \tau_{M^{(k_{\mathrm{CP}})}}^{(k_{\mathrm{CP}})})$ be variables denoting the number and times of CPs respectively. Given a vector $\tau^{(k_{\mathrm{CP}})}$, we use $t \in \tau^{(k_{\mathrm{CP}})}$ to indicate that one of the elements of $\tau^{(k_{\mathrm{CP}})}$ is equal to t: if $t \in \tau^{(k_{\mathrm{CP}})}$, then $\exists j \in \{1, \ldots, M^{(k_{\mathrm{CP}})}\}$ such that $\tau_j^{(k_{\mathrm{CP}})} = t$. This chapter will propose a methodology to quantify the uncertainty in estimates of these CP characteristics

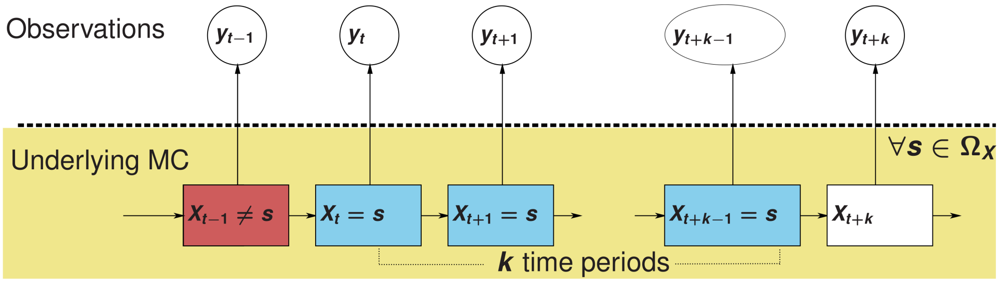


Figure 3.1: Graphical representation of the sustained CP definition utilised. $k \equiv k_{\text{CP}}$, the required sustained time period in the underlying state sequence for a CP into a new regime to have occurred. In this example, a CP into the regime corresponding to state $s \in \Omega_X$ is said to have occurred at time t if $X_{t-1} \neq X_t = \ldots = X_{t+j} = s$ for $j \geq k_{\text{CP}} - 1$.

by estimating:

$$P(M^{(k_{\text{CP}})} = m|y_{1:n}) \qquad m = 0, 1, 2 \ldots, \tag{3.2}$$

$$\text{and } P(\tau^{(k_{\text{CP}})} \ni t|y_{1:n}) \qquad t = 2, \ldots, n \tag{3.3}$$

where $P(\tau^{(k_{\text{CP}})} = t) \equiv P(\tau^{(k_{\text{CP}})} \ni t|y_{1:n}) \equiv \sum_m P(M^{(k_{\text{CP}})} = m|y_{1:n}) \sum_{i=1}^{m} P(\tau^{(k_{\text{CP}})} = t|y_{1:n}, M^{(k_{\text{CP}})} = m)$, that is , the probability distribution of the number of CPs, and the marginal posterior probability that a CP occurs at a particular time (the CP probability, CPP). The CPP is commonly denoted using the equality symbol in the CP literature. That is $P(\tau^{(k_{\text{CP}})} = t) \equiv P(\tau^{(k_{\text{CP}})} \ni t|y_{1:n})$, with the latter being used since we shall be decomposing the event of a CP occurring into the event of the uth CP occurring.

3.2.1 Exact CP Distributions via Finite Markov Chain Imbedding

Under the generalised CP definition and conditioned on a particular model parameter setting θ, it is possible to compute exact CP distributions for a variety of CP characteristics (Aston et al., 2011). That is, it is possible to compute $P(\tau^{(k_{\text{CP}})} \ni t|y_{1:n}, \theta)$ and $P(M^{(k_{\text{CP}})} = m|y_{1:n}, \theta)$ exactly, such that they are not subject to sampling or approximation error.

The generalised CP definition presented motivates why we are analysing for runs of a minimum length in the underlying chain X_t. A run of length k_{CP} in state $s \in \Omega_X$ in the underlying state sequence is k_{CP} consecutive occurrences of s in X_t. That is $X_t = s = X_{t+1} = \ldots = X_{t+k_{\text{CP}}-1}$ and if $X_{t-1} \neq s$ then the run of desired length has occurred at time $t + k_{\mathrm{CP}} - 1$. Thus in order to consider whether a CP has occurred at time t, we can reformulate this problem as determining whether a run of length k_{CP} has occurred at time $t + k_{\mathrm{CP}} - 1$ in the underlying chain.

One popular approach for inferring the behaviour in the underlying state sequence for HMMs given an observation process, is via the Viterbi and Posterior Decoding algorithms (Viterbi (1967) and Juang and Rabiner (1991) respectively). However, as discussed in Section 2.12.1 (page 37), these provide a single estimate of the underlying state sequence and all CP estimates are obtained deterministically from this single estimate. However, other state sequences may also be possible under the observed data. These algorithms thus fail to capture the uncertainty regarding other potential state sequences occurring and consequently, the uncertainty associated with the run and CP statistics derived from it is not captured.

In order to fully capture the uncertainty of CPs under a HMM framework, it is necessary to consider all possible state sequences. This can be achieved by computing posterior, time-inhomogeneous transition probabilities with respect to the observed time series $P(X_t|X_{t-1}, y_{1:n})$ for $t = 1, \dots, n$. These can be obtained from the smoothed probabilities, the probability of the chain being in particular states conditioned on the entire time series for example $P(X_{t-r:t-1}, X_t = s|y_{1:n})$, as follows:

$$P(X_t = s|X_{t-r:t-1}, y_{1:n}) = \frac{P(X_{t-r:t-1}, X_t = s|y_{1:n})}{\sum_{s \in \Omega_X} P(X_{t-r:t-1}, X_t = s|y_{1:n})} \tag{3.4}$$

These posterior transition probabilities form a sequence of time dependent posterior transition probabilities matrices $\{\tilde{\mathbf{P}}_1, \dots, \tilde{\mathbf{P}}_n\}$ and permits us to consider the general evolution of the underlying MC with respect to the observed time series. This thus allows us to quantify the uncertainty of the underlying state sequence, the uncertainty of runs in the underlying state sequence and ultimately, the uncertainty of the CP themselves.

In doing so, we firstly decompose the event of a CP occurring at time t. Let $\tau_u^{(k_{\mathrm{CP}})}$ denote the time of the uth CP with $u \geq 1$. The CP probability (CPP) can thus be decomposed as follows via the total law of probability:

$$P(\tau^{(k_{\mathrm{CP}})} \ni t|y_{1:n}, \theta) = \sum_{m} P(M^{(k_{\mathrm{CP}})} = m|y_{1:n}, \theta) \sum_{u=1}^{m} P(\tau_u^{(k_{\mathrm{CP}})} = t|M^{(k_{\mathrm{CP}})} = m, y_{1:n}, \theta) \tag{3.5}$$

$$= \sum_{u=1,2,\dots} P(\tau_u^{(k_{\mathrm{CP}})} = t, M^{(k_{\mathrm{CP}})} \geq u|y_{1:n}, \theta). \tag{3.6}$$

The event of the uth CP occurring at time t can be re-expressed as a quantity involving runs, specifically: whether the uth run of minimum length k_{CP} has occurred at time $t + k_{\mathrm{CP}} - 1$. Let $W_s(k_{\mathrm{CP}}, u)$ denote the waiting time for the uth occurrence of a run of minimum length k_{CP} in state $s \in \Omega_X$. Thus $W_s(k_{\mathrm{CP}}, u) = t + k_{\mathrm{CP}} - 1$ denotes that the uth run of interest has successfully occurred at time $t + k_{\mathrm{CP}} - 1$. Similarly, $W(k_{\mathrm{CP}}, u)$ denotes the waiting time for the uth occurrence of a run in any state in Ω_X of at least length k_{CP}. Specific regimes are associated with specific runs of particular states $s \in \Omega_X$. Consequently, $W_s(k_{\mathrm{CP}}, u)$, is the main focus of analysis if interest lies in CPs into these specific regimes. For example, if recession regimes are of interest where two consecutive "contraction" states are required for a CP into a recession regime, then $W_1(k_{\mathrm{CP}} = 2, u)$ is of interest where $s = 1 =$ "contraction". For general CP inference regarding CPs into any regime, $W(k_{\mathrm{CP}}, u)$ is the main focus.

By re-expressing the uth CP event as the waiting time for the uth occurrence of a run, it is thus possible to compute the corresponding probabilities:

$$P(\tau_u^{(k_{\mathrm{CP}})} = t | y_{1:n}, \theta) = P(W(k_{\mathrm{CP}}, u) = t + k_{\mathrm{CP}} - 1 | y_{1:n}, \theta). \tag{3.7}$$

It is exactly the waiting time probability on the right of the above equation that can be computed exactly. More specifically, it is possible to compute exact distributions regarding the waiting times of run and pattern statistics, namely $P(W(k_{\mathrm{CP}}, u) \leq t | \theta, y_{1:n})$. This is achieved by an efficient framework called Finite Markov Chain Imbedding (FMCI, Fu and Koutras (1994), Fu and Lou (2003)). This framework is not exclusive to the use of HMMs, originating from a multistate trials context (Fu and Koutras, 1994), and applied in a Markov Chain scenario for generalised patterns (see Aston and Martin (2005) for example). Motivated by the sustained CP definition with respect to runs of states, we focus on reviewing FMCI with respect to runs as opposed to patterns (a defined configuration of symbols).

As the name suggests, FMCI imbeds the random variables of interest into finite auxiliary MCs such that the run and pattern statistic of interest, in this case the waiting time statistic, can be computed via MC results. More specifically, FMCI introduces several auxiliary MCs $\{Z_t^{(1)}, Z_t^{(2)}, Z_t^{(3)}, \ldots\}$ which are defined over the common state space $\Omega_Z^{(k_{\mathrm{CP}})} = \Omega_X \times \{-1, 0, 1, \ldots, k_{\mathrm{CP}}\}$. $\Omega_Z^{(k_{\mathrm{CP}})}$ can be considered as an expanded version of Ω_X which consists of tuples (X_t, l). The first component of the tuple denotes the behaviour of the underlying MC as before, and the new variable $l = -1, 0, 1, 2, \ldots, k_{\mathrm{CP}}$ indicates the progress of any runs that are of interest. The uth auxiliary MC $\{Z_t^{(u)}\}$, corresponds to tracking the occurrence of the uth run of length k_{CP}, conditional of $(u-1)$ runs having already occurred.

The states of the auxiliary MCs can be categorised into three groups dependent on the pattern progress value l: continuation $(l=-1)$, run in progress $(l=0,1,\ldots,k_{\mathrm{CP}}-1)$ and absorption $(l=k_{\mathrm{CP}})$. For the uth auxiliary MC $\{Z_t^{(u)}\}$, which tracks the occurrence of the uth run, the absorption states denotes that the uth run of desired length has successfully occurred, the run in progress states indicate the progress of any potential initiated runs, and the continuation states denote that the $(u-1)$th run is still in progress (its length exceeds the required length k_{CP}) and needs to end before the occurrence of the new uth run can be officially tracked. The continuation states are also known as waiting states and there is a one-to-one correspondence with the absorption states.

The auxiliary MCs are constructed such that at time 0, $\{Z_t^{(1)}\}$ is initialised in the initialisation states where $l=0$. At each time step thereafter, an operation is performed such that any probability associated in the absorption states of each chain $\{Z_t^{(u)}\}_{u=1}^{\infty}$, is mapped to the corresponding continuation state in the $(u+1)$th chain $\{Z_t^{(u)}\}_{u=2}^{\infty}$. Consequently, the uth auxiliary chain for $u=2,3,\ldots$ is initialised with non-zero probability in the continuation states when the previous chain in the sequence has reached the corresponding absorption state.

The transition probabilities of these auxiliary MCs $\{Z_t^{(u)}\}_{u=1}^{\infty}$ between the states are obtained deterministically from the original MC. Let $\mathbf{Q}$ denote the transition matrix for the auxiliary MCs $\{Z_t^{(u)}\}_{u=1}^{\infty}$ which is populated by entries from the transition probability matrix $\mathbf{P}=(p_{ij})_{i,j\in\Omega_X}$ for the original MC X_t. The transition matrix $\mathbf{Q}$ is used as in standard MC theory to describe how the uth auxiliary MC $Z_t^{(u)}$, evolves over time.

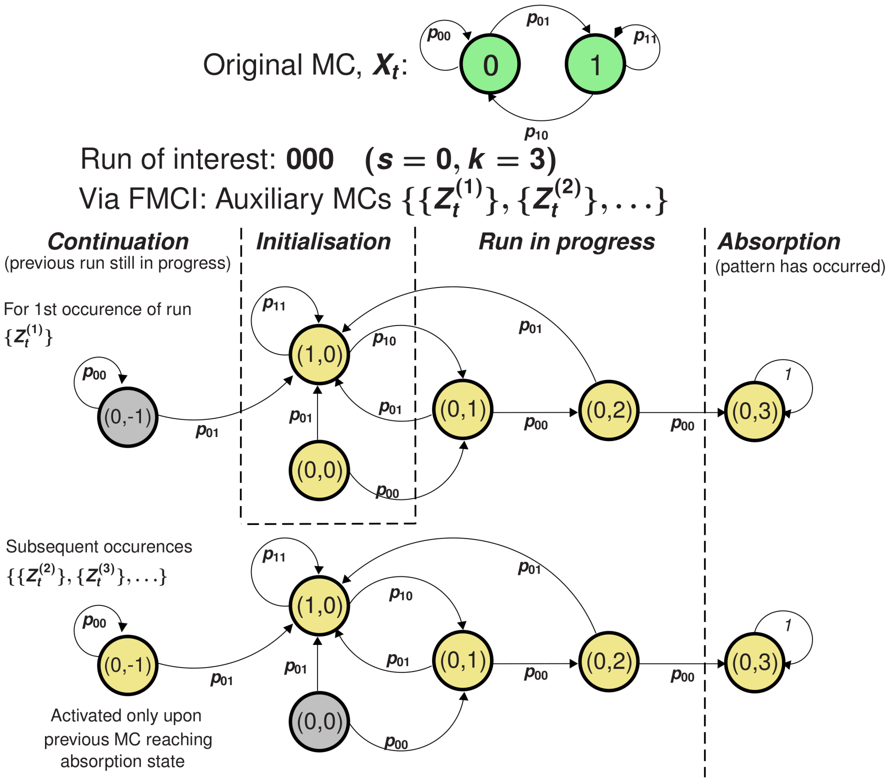
To fix ideas and terminology regarding FMCI, Figure 3.2 presents a toy example with respect to a standard time homogeneous MC. We consider a two state MC, $\Omega_X=\{0,1\}$ with the run of interest being 000 and thus $s=0$, $k_{\mathrm{CP}}=3$. The transition probabilities of the auxiliary MCs, $\{Z_t^{(u)}\}_{u=1}^{\infty}$, are those from the original MC $\{X_t\}$ with some modifications in places (for example the transition probabilities for the absorption state, $(0,3)$). The first chain $\{Z_t^{(1)}\}$, tracks the movement of the first occurrence of the run. The chain is typically initialised in the states corresponding to no pattern being in progress $(l=0)$, more specifically in states $(0,0)$ and $(1,0)$. Such initialisation can be based on the initialisation of the original chain $\{X_t\}$. If a 0 is observed at time 1 $(X_1=0)$, the auxiliary MC moves to state $(0,1)=Z_1^{(1)}$, due to the initiation of a potential run and being one step closer in observing the run of interested. For each $X_t=0$, the pattern progress variable l increases by one each time since we are one step closer in potentially seeing the run of interest. However, if a 1 is observed at any time ($X_t = 1$), this terminates any initiated runs in progress, and the chain returns to $Z_t^{(1)} = (1, 0)$, the state corresponding to no pattern in progress. Upon reaching $(0, 3)$, the absorption state in this example, the run has successfully occurred and thus the first occurrence of the run of interest has occurred. The auxiliary MC remains in this state for all subsequent time points regardless of whether a 0 or 1 is observed.

Successfully reaching the absorption state activates the next chain in the sequence $\{Z_t^{(2)}\}$, which tracks the movement of the second occurrence of the run, conditioned on the first occurrence having successfully happened. Upon $Z_t^{(1)} = (0, 3)$ reaching the absorption state for some time t, the new chain is immediately initialised with the non-zero probability associated with this absorption state in the corresponding continuation state $Z_t^{(2)} = (0, -1)$ to denote that the previous occurrence of the run is still in progress. If for future times $t' > t$, $X_{t'} = 0$ is observed, then $Z_{t'}^{(2)} = (0, -1)$, to denote that the previous run is still in progress. However if $X_{t'} = 1$, then this officially terminates the previous run, a new run can be officially tracked and $Z_{t'}^{(2)} = (1, 0)$. $Z_{t'}^{(2)}$ then proceeds as before, with new auxiliary chains being fully initialised with non-zero probability when they reach absorption states.

In the context of HMMs where an observed time series is available and with respect to CP problems, a few modifications are made to the FMCI framework. Most importantly, the time-homogeneous transition probability matrix associated with the auxiliary MCs $\mathbf{Q}$, is replaced with a sequence of posterior, time-inhomogeneous transition probabilities $\{\tilde{\mathbf{Q}}_t\}_{t=1}^n$. These transition probabilities are based on the sequence of posterior, time-inhomogeneous transition probabilities $\{\tilde{\mathbf{P}}_t\}_{t=1}^n$ defined with respect the underlying MC $\{X_t\}$, and are determined from the posterior transition probabilities as calculated from Equation 3.4. The use of these posterior, time inhomogeneous transition probabilities thus allows all potential underlying state sequences to be postulated with respect to the observed time series.

Under the initialisation configuration presented where $\{Z_t^{(1)}\}$ is initialised in the states where $l = 0$, it is possible for a CP to occur at time 1 (for example, if $X_0 = 1, X_{1:3} = 0$ in the example presented in Figure 3.2). As a CP occurring at time 1 often makes little sense, it is thus possible to initialise in the equivalent continuation states ($l = -1$) to resolve this issue.

Computing waiting time distributions is achieved by Markov Chain theory. Before doing so, it is necessary to define the technical concepts discussed above in linking the sequence of auxiliary MCs together such that multiple occurrences of runs can be modelled. Let $\Psi_t, t = 0, 1, \ldots, n$ be a $M^{\max} \times |\Omega_Z^{(k_{\mathrm{CP}})}|$ matrix where



Original MC, X_t:
0
1
p_{00}
p_{01}
p_{11}
p_{10}
Run of interest: **000** $(s = 0, k = 3)$
Via FMCI: Auxiliary MCs $\{\{Z_t^{(1)}\}, \{Z_t^{(2)}\}, \ldots\}$
Continuation
(previous run still in progress)
Initialisation
Run in progress
Absorption
(pattern has occurred)
For 1st occurence of run $\{Z_t^{(1)}\}$
(0,-1)
(1,0)
(0,0)
(0,1)
(0,2)
(0,3)
p_{00}
p_{01}
p_{11}
p_{10}
1
Subsequent occurences $\{\{Z_t^{(2)}\}, \{Z_t^{(3)}\}, \ldots\}$
Activated only upon previous MC reaching absorption state

Figure 3.2: Graphical Example of Finite Markov Chain Imbedding (FMCI), an efficient mechanism to compute exact run and pattern distributions of Markov Chains. As our CP definition is defined in terms of sustained changes in states, we can thus compute exact distribution regarding CPs.

$M^{\max} = \lfloor \frac{n}{k_{\text{CP}}} \rfloor$ denotes the maximum number of runs, and consequently the maximum number of CPs, that can occur during the scope of the data. The uth row of Ψ_t is denoted by $\psi_t^{(u)}$ which will store state probabilities for the uth auxiliary MC $\{Z_t^{(u)}\}$, at time t. The initial matrix Ψ_0, thus has $\psi_0^{(1)}$ with non-zero probabilities in the initialisation or continuation states, and zeroes elsewhere in the row vector and the initialisation matrix Ψ_0. This latter remark is due to the fact that no further runs and subsequent chains can be in progress at $t = 0$. In order to connect absorptions states to their corresponding continuation in the next auxiliary MC in the sequence, this is achieved by the following mechanism. Denote the collection of states representing the collection absorption and continuation states as $A = \{Z_t^{(u)} = (X_t, l) \in \Omega_Z^{(k_{\text{CP}})} | l = k_{\text{CP}}\}$ and $C = \{Z_t^{(u)} = (X_t, l) \in \Omega_Z^{(k_{\text{CP}})} | l = -1\}$

respectively. Then let Υ be $|\Omega_Z^{(k_{\mathrm{CP}})}| \times |\Omega_Z^{(k_{\mathrm{CP}})}|$ matrix is defined as follows:

$$\Upsilon(z_1, z_2) = \begin{cases} 1, & \text{if } z_1 \in A \text{ and } z_2 \text{ is the corresponding continuation state in } C; \\ 0, & \text{otherwise.} \end{cases} \tag{3.8}$$

Finally, let $\{\tilde{\mathbf{Q}}_t\}_{t=1}^n$ denote the sequence of time in-homogeneous, posterior transition probabilities defined over the auxiliary MCs, and $U(A)$ be a $|\Omega_Z^{(k_{\mathrm{CP}})}|$ length column vector with ones in the locations of the absorption states and zeroes elsewhere. Then the waiting time for the uth occurrence of a run, $W(k_{\mathrm{CP}}, u)$ can be computed as follows. For $t = 1, \ldots, n$,

$$\Psi_t = \Psi_{t-1}\tilde{\mathbf{Q}}_t \tag{3.9}$$

$$\psi_t^{(u)} \leftarrow \psi_t^{(u)} + \psi_{t-1}^{(u-1)}(\tilde{\mathbf{Q}}_t - \mathbf{I})\Upsilon, \qquad u = 2, \ldots, M^{\max} \tag{3.10}$$

$$P(W(k_{\mathrm{CP}}, u) \le t | y_{1:n}, \theta) = P(Z_t^{(u)} \in A | y_{1:n}, \theta) = \psi_t^{(u)} U(A) \tag{3.11}$$

where $\mathbf{I}$ is a $|\Omega_Z^{(k_{\mathrm{CP}})}| \times |\Omega_Z^{(k_{\mathrm{CP}})}|$ identity matrix. The intuition of this computation is as follows: Equation 3.9 computes the general evolution of all $M^{\max}$ auxiliary MCs simultaneously. Equation 3.11 denotes the probability of the uth chain being in any of the absorption states and thus the probability the runs of interest having occurred by time t, conditional on the $(u-1)$th run having already occurred. Equation 3.10 is the necessary modification which links the auxiliary MCs together and such that a chain is assigned non-zero probability (activated) when the previous chain in the sequence has reached an absorption state. This is ultimately a row updating operation which transfers the probability of the $(u-1)$th auxiliary MC being in absorption state into the corresponding continuation states of the uth auxiliary MC. By expressing the above equations in terms of matrices and vector, this leads to an efficient mechanism to compute waiting time distributions.

Having computed exactly the waiting time distributions for runs via the FMCI framework presented above, it is thus possible to compute exact distributions for a variety of CP characteristics. For example, the probability of the uth CP at time t is provided by

$$\begin{aligned} P(\tau_u^{(k_{\mathrm{CP}})} = t | y_{1:n}, \theta) &= P(W(k_{\mathrm{CP}}, u) = t + k_{\mathrm{CP}} - 1 | y_{1:n}, \theta) \\ &= P(W(k_{\mathrm{CP}}, u) \le t + k_{\mathrm{CP}} - 1 | y_{1:n}, \theta) - \\ &\quad P(W(k_{\mathrm{CP}}, u) \le t + k_{\mathrm{CP}} - 2 | y_{1:n}, \theta). \end{aligned}$$

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

$$P(M^{(k_{\mathrm{CP}})} = m|y_{1:n}, \theta) = P(W(k_{\mathrm{CP}}, m) \leq n|y_{1:n}, \theta) - P(W(k_{\mathrm{CP}}, m+1) \leq n|y_{1:n}, \theta).$$

Exact distributions for other CP characteristics such as the probability of a CP within a given time interval and the distribution of regime lengths, can also be computed via the FMCI framework. This thus provides a flexible methodology in capturing the uncertainty of CP problems. Aston et al. (2011) discuss that the computational complexity for this conditional exact CP method is $\mathcal{O}(n)$, when H and $M^{\max}$ are fixed between different runs. This computational complexity is expected to increase as H and $M^{\max}$ increase.

These exact CP distributions are conditioned on the model parameters θ. However, it is typical for θ to be unknown, with the Expectation-Maximisation algorithm (Baum et al., 1970) being a typical frequentist approach in obtaining a point estimate of θ under the HMM framework. θ is also subject to error and uncertainty which needs to captured. Consequently, in order to fully consider uncertainty of CPs, it is also necessary to consider the uncertainty of the parameters. In capturing the uncertainty fully and explicitly, we turn to Bayesian methods in accounting for θ which explicitly considers the uncertainty compared to frequentist approaches. In particular, we account for the model parameters via the use of Sequential Monte Carlo samplers.

3.2.2 Sequential Monte Carlo methods

In dealing with parameter uncertainty, we adopt a Bayesian approach by integrating out the model parameters to obtain marginal posterior distributions of the CP quantities alone. However, it is not possible to perform this integration analytically for the models of interest and thus numerical approximation is necessary.

Sequential Monte Carlo (SMC) methods, also known as particle filters (Kitagawa, 1996), permit such numerical approximation and are more specifically a class of simulation algorithms for sampling from a sequence of related distribution $\{\pi_b\}_{b=1}^{B}$ via importance sampling and resampling techniques. Common applications of SMC methods in Statistics, Scientific Computing and Engineering include sequential Bayesian inference on the posterior where the data increases incrementally (that is online inference such that $\pi_b \propto p(\theta)l(\theta|y_{1:b}), b = 1, \ldots, B = n$), the self avoiding random walk model in modelling the growth of a polymer, and online filtering in radar tracking problems. We refer the reader to Liu (2001) and Doucet and Johansen (2011) for recent surveys on the SMC literature.

Importance sampling is a fundamental concept in Monte Carlo sampling such that if one wants to sample from a single, complex target distribution π_B, we sample instead from a similar, tractable distribution q_B (importance distribution), and reweight the samples accordingly such that they are samples from π_B. Namely, if $\theta \sim q_B$, then the corresponding importance weight is,

$$w_B(\theta) \propto \frac{\pi_B(\theta)}{q_B(\theta)} \tag{3.12}$$

where support $[q_B(\cdot))] \geq$ support $[\pi_B(\cdot)]$. The target distribution π_B, is therefore approximated by a weighted cloud of N samples, $\pi_B \approx \{\theta^i, W_B^i\}_{i=1}^N$, where W_B^i are the normalised importance weights. In addition, the normalising constant for π_B can also be approximated.

In the SMC context where one wants to sample from multiple distributions $\{\pi_b\}_{b=1}^B$, we firstly obtain samples from π_1 via importance sampling. For $b = 2, \ldots, B$, we use the existing samples of π_{b-1} to obtain samples of π_b. This is achieved by perturbing existing samples in some fashion (namely via a Markov kernel) and re-weighting accordingly. This thus approximates the sequence of distributions $\{\pi_b\}_{b=1}^B$ by weighted clouds of N samples, $\{\theta_b^i, W_b^i\}_{i=1}^N$ for $b = 1, \ldots, B$, and such an algorithm is known as Sequential Importance Sampling (SIS) in the literature.

A resampling mechanism is often introduced within SMC algorithms, such as SIS, consequently leading to the Sequential Importance Resampling algorithm. This avoids the weight degeneracy problem. Such a problem occurs due to the variance of the importance weights increasing with b as a result of several samples with small weights close to zero, and only a few samples with large weights being present. As a consequence, this does not provide accurate approximations or samples from the distribution of interest π_b. A resampling step is thus introduced to resolve this issue such that we discard samples with small weights, and replicate those with higher weights. This consequently allows greater focus on more probable areas of the distribution and preserves the expectation of the approximation of the integral to any bounded function. More specifically, if $\{W^i, \theta^i\}_{i=1}^N$ is a collection of weighted samples, then resampling consists of selecting a collection of samples $\{\tilde{\theta}^i\}_{i=1}^N$ 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. Resampling selection is determined by the importance weights of the samples and there are a variety of methods in which resampling can performed; Douc and Cappé (2005) provide an overview and comparison of different resampling schemes available. The simplest approach is termed multinomial resampling as we draw N samples with replacement from the existing collection of samples with multinomial probabilities $(W^1, \dots, W^N)$. However, this approach unnecessarily increases the Monte Carlo variance and other techniques such as residual resampling are preferable. All resampled samples are then set to have equal importance weights (that is $W^i = \frac{1}{N}$).

Whilst resampling is beneficial in the long run, resampling unnecessarily at every iteration is not desired since it introduces unnecessary Monte Carlo variance. Thus resampling should not be performed at every iteration b. A dynamic resampling scheme is therefore implemented within the SMC community such that one only resamples when the variance of weights exceeds a pre-specified threshold. This can be implemented by considering the Effective Sample Size (ESS),

$$ESS = \frac{1}{\sum_{i=1}^{N}(W^i)^2}. \tag{3.13}$$

This is obtained via a Taylor expansion of the variance of associated estimates (Kong et al., 1994) and acts as a proxy for the variance of importance weights. Intuitively, the ESS provides an approximation of the number of independent samples required from the distribution π_b, that would provide an estimate of comparable variance. Resampling is performed when the ESS falls below a pre-specified threshold, $ESS < T$, with $T = N/2$ commonly being used in the literature. Resampling at such stopping times rather than deterministic time is valid and has recently been demonstrated that convergence results can be extended to this case (Del Moral et al., 2012).

Sequential Monte Carlo samplers

The standard application of SMC algorithms such as those presented above require that the sequence of distributions $\{\pi_b\}_{b=1}^{B}$, are defined upon a sequence of increasing state spaces. For example in sequential Bayesian inference, the state space increases systematically with respect to each new observation. Sequential Monte Carlo samplers (SMC samplers, Del Moral et al. (2006)) are a particular class of SMC algorithms such that $\{\pi_b\}_{b=1}^{B}$ can be defined over any sequence of spaces. One particular use of the SMC samplers framework is to ultimately sample from a complex target distribution, π_B, such that we sample initially from a tractable distribution π_1 which shares the same state space as the target distribution, and define a sequence of intermediary distributions in which we move through to sample from the target distribution of interest. For example, in Bayesian inference where one may be interested in sampling from a complex parameter posterior $\pi_B = p(\theta|y_{1:n})$, it is possible to define the sequence of distributions as follows:

$$\pi_b \propto p(\theta) l(\theta|y_{1:n})^{\gamma_b} \qquad b = 1, \ldots, B \tag{3.14}$$

where $p(\theta)$ is the model parameter prior, $l(\theta|y_{1:n})$ is the likelihood, and $\{\gamma_b\}_{b=1}^{B}$ is a non-decreasing tempering schedule such that $\gamma_1 = 0, \gamma_B = 1$. Such a sequence ultimately allows us to sample from the parameter posterior, $\pi_B \propto p(\theta)l(\theta|y_{1:n})$, by sampling from the prior initially and introducing the effect of the likelihood gradually. It is exactly this sequence of distributions which shall be the focus of this chapter. Sampling via SMC samplers has computational complexity $\mathcal{O}(N)$ where N is the number of samples.

The general idea of SMC samplers is graphically represented in Figure 3.3 in sampling from the parameter posterior. Each distribution in the sequence $\{\pi_b\}_{b=1}^{B}$ is approximated by the weighted cloud of samples. Samples are represented graphically by circles and their associated weights by their radii in the figure. We sample initially from the first distribution in the sequence, $\pi_1 = p(\theta)$ the parameter prior, either directly or via importance sampling and compute the associated importance weights. If π_2 is similar enough to π_1, then the intuition is that we can approximate π_2 by moving the existing samples approximating π_1 by mutating them via local moves into regions of higher probability density and re-weighting accordingly. There is a great deal of flexibility in the mutation step, with Markov kernels such as Metropolis-Hastings being a possibility. This idea of approximating π_b via mutation of existing samples of π_{b-1} and re-weighting persists throughout the SMC samplers algorithm. In addition, to avoid weight degeneracy a dynamic resampling scheme is employed, which encourages samples with higher weights in higher probability areas to survive. Such an algorithm consequently allows one to sample and approximate the defined sequence of distributions, and ultimately the posterior of interest, $\pi_B \propto p(\theta)l(\theta|y_{1:n})$.

The SMC samplers framework also provides approximations of the normalising constants for the distributions $\{\pi_b\}_{b=1}^{B}$. This is an important feature that will become more relevant in the model selection methodology proposed in Chapter 5.

We refer the reader to Del Moral et al. (2006) for specific details regarding the asymptotics of the SMC samplers algorithm. For example, it is shown that as $N \to \infty$ (that is as the number of samples in the approximation increases), the approximations asymptotically converges to the distribution of interest.

In ensuring that mutated samples are correctly re-weighted to approximate the next distribution in the sequence, it is necessary to discuss this re-weighting procedure in greater detail. A collection of Markov kernels $\{L_b\}$ is firstly intro-

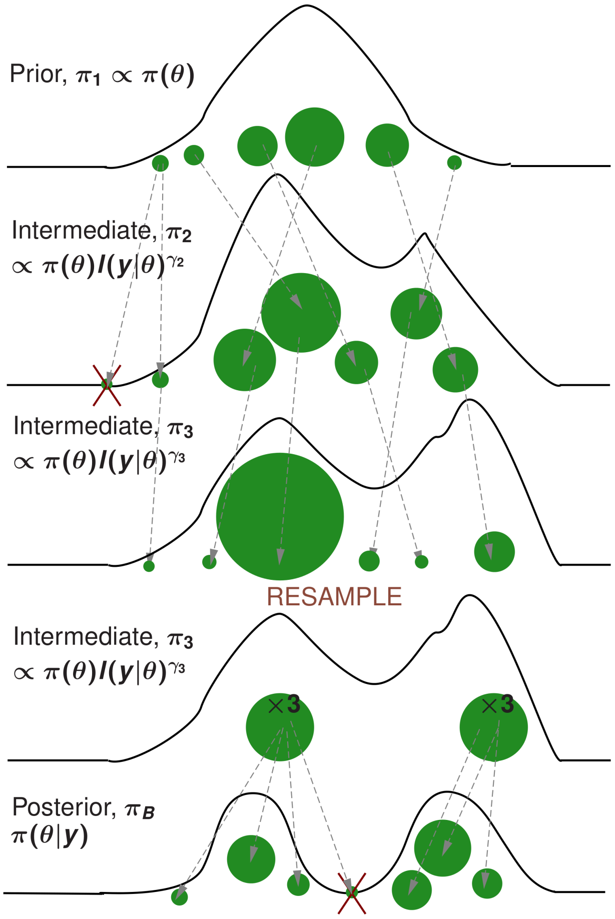

Prior, $\pi_1 \propto \pi(\theta)$
Intermediate, π_2
$\propto \pi(\theta)l(y|\theta)^{\gamma_2}$
Intermediate, π_3
$\propto \pi(\theta)l(y|\theta)^{\gamma_3}$
RESAMPLE
Intermediate, π_3
$\propto \pi(\theta)l(y|\theta)^{\gamma_3}$
×3
×3
Posterior, π_B
$\pi(\theta|y)$

Figure 3.3: Graphical representation of Sequential Monte Carlo samplers, an algorithm to sample from a sequence of connected distributions defined over any arbitrary sample space. Each distribution in the sequence is approximated by weighted clouds of samples. Samples are represented by circles, and their corresponding weights by their radii in the graphic above. The sampling of each distribution in the sequence is achieved by mutating and resampling existing samples from the previous distribution in the sequence. For the application of interest, the ultimate aim is to sample from $\pi_B = p(\theta|y_{1:n})$, the parameter posterior. This is achieved by initially sampling from the prior $\pi_1 = p(\theta)$, and sampling from a sequence of intermediary distributions by introducing the effect of the likelihood $l(\theta|y_{1:n}) \equiv l(y|\theta)$ gradually via the use of a non-decreasing tempering schedule $\{\gamma_b\}_{b=1}^{B}$.

duced with the distributions of interest $\{\pi_b(u_b)\}$ being formally augmented with the aforementioned collection of Markov kernels to produce auxiliary distributions $\{\tilde{\pi}_b\}$ with $\tilde{\pi}_b = \pi_b(u_b)\prod_{j=1}^{b-1} L_j(u_{j+1}, u_j)$.

Given a weighted sample $\{W_{b-1}^i, \theta_{b-1}^i\}$ which is correctly weighted to approximate $\pi_{b-1}(\theta_{b-1})$, the SMC sampler with proposal kernel $K_b(\theta_{b-1}^i, \theta_b^i)$ is used which leads to the sample $\{W_{b-1}^i, (\theta_{b-1}^i, \theta_b^i)\}$. Such a sample is properly weighted to 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 modify the weights of the sample such that it is correctly weighted to the target distribution $\pi_b(\theta_b)L_{b-1}(\theta_b, \theta_{b-1})$. This is achieved by multiplying the weights by the appropriate incremental weights $\widetilde{w}_b(\theta_{b-1}^i, \theta_b^i)$ such that $W_b^i \propto W_{b-1}^i \cdot \widetilde{w}_b(\theta_{b-1}^i, \theta_b^i)$. These incremental weights are:

$$\widetilde{w}_b(\theta_{b-1}^i, \theta_b^i) = \frac{\pi_b(\theta_b^i)L_{b-1}(\theta_b^i, \theta_{b-1}^i)}{\pi_{b-1}(\theta_{b-1}^i)K_b(\theta_{b-1}^i, \theta_b^i)}, \tag{3.15}$$

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

$$L_{b-1}^{\text{opt}}(\theta_b, \theta_{b-1}) = \frac{\pi_{b-1}(\theta_{b-1})K_b(\theta_{b-1}, \theta_b)}{\int \pi_{b-1}(\theta'_{b-1})K_b(\theta'_{b-1}, \theta_b)d\theta'_{b-1}}, \tag{3.16}$$

if resampling is performed at every iteration b of the SMC samplers algorithm. However, the integral in the denominator is generally intractable and it is therefore necessary to use approximations. These approximations only increase the variance of the estimator but do not introduce any further approximation. If K_b is chosen to be a π_b MCMC invariant kernel, then a widely-used approximation of this optimal quantity can be obtained such that if $\pi_{b-1} \approx \pi_b$ (that is consecutive distributions in the sequence are similar), then we can replace π_{b-1} with π_b in the optimal backward kernel. This thus provides the approximated optimal backward kernel:

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

where K_b is a π_b-invariant Markov kernel. The incremental weight expressed in Equation 3.15 is thus:

$$\widetilde{w}_b(\theta_{b-1}^i, \theta_b^i) = \frac{\pi_b(\theta_b^i)}{\pi_{b-1}(\theta_{b-1}^i)K_b(\theta_{b-1}^i, \theta_b^i)} \times \frac{\pi_b(\theta_{b-1}^i)K_b(\theta_{b-1}^i, \theta_b^i)}{\pi_b(\theta_b^i)} = \frac{\pi_b(\theta_{b-1}^i)}{\pi_{b-1}(\theta_{b-1}^i)}. \tag{3.17}$$

Note that such a incremental weight is independent of the present sample θ_b^i at iteration b, and dependent only on the sample at the previous iteration θ_{b-1}^{i}. Consequently, the importance weights of the sample at iteration b are independent of the sample itself. Due to the independence between the weights and the sample at iteration b, resampling can thus be performed prior to the mutation step. This prior resampling before the mutation ultimately leads to greater diversity of the resulting sample compared to post resampling.

Algorithm 6 presents a generic SMC sampler algorithm in sampling from a sequence of distributions $\{\pi_b\}_{b=1}^{B}$. The SMC samplers application of sampling from a complex target distribution π_B through a sequence of distributions is similar to Annealed Importance Sampling (Neal, 2001), although a resampling mechanism is present in the SMC samplers framework.

Other Monte Carlo sampling strategies are also possible. Markov Chain Monte Carlo (MCMC, Gilks et al. (1996)) is a popular approach to sample from the complex distribution π_B, where an ergodic Markov chain is constructed with transition kernels K such that the stationary distribution of the sampling MC is π_B, the target distribution of interest. In the general context of SMC methods where interest lies in a sequence of distributions, this is impossible to perform efficiently via MCMC. Consequently, MCMC cannot be used in a sequential Bayesian estimation context with respect to incremental data. In general, if π_B is ultimately of interest such as the example presented in this section, it is typically difficult to design transition kernels such that the sampling chain is mixing well (exploring the state space well), and it is often difficult to determine whether the chain has reached convergence and thus sampling from the desired distribution π_B is achieved. Ensuring that the chain is mixing well is particularly important when π_B is multimodal, and it is thus necessary to ensure that the chain can move between these modes if necessary. In comparison, SMC samplers has the advantage of considering several samples simultaneously which explore the state space in a local fashion. Designing good MCMC algorithms with acceptable performance is also often specific to the application and problem, whereas SMC works well even under generic settings.

Data augmentation is a common strategy used within MCMC algorithms, particularly when considering HMMs and mixture models. Such a strategy introduces a latent vector sequence which postulates which component or state the observation may have arisen from. This latent sequence is sampled along with the parameters and via marginalisation, the parameter posterior can be obtained. However, due to the inherent correlation within the latent sequence itself and the parameters, it is often harder to obtain a fast, good mixing MCMC algorithm.

Particle MCMC (Andrieu et al., 2010) is a recently proposed sampling algo-

Algorithm 6 Generic SMC Sampler algorithm to sample from the sequence of distributions $\{\pi_b\}_{b=1}^{B}$. (Del Moral et al., 2006)

```
Step 1: Initialisation Set $b = 1$
for $i = 1, \ldots, N$ do
    Draw $\theta_1^i \sim q_1$ ($q_1$ is a tractable importance distribution for $\pi_1$).
    Compute the corresponding importance weight $\{w_1(\theta_1^i)\} \propto \pi_1(\theta_1^i)/q_1(\theta_1^i)$.
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 $i = 1, \ldots, N$ do
    Draw $\theta_b^i \sim K_b(\theta_{b-1}^i, \cdot)$ where $K_b$ is a $\pi_b$ invariant Markov kernel.
    Compute the incremental weights:
        $$\widetilde{w}_b\left(\theta_{b-1}^i, \theta_b^i\right) = \frac{\pi_b(\theta_{b-1}^i)}{\pi_{b-1}(\theta_{b-1}^i)}.$$
end for
Compute the new normalised importance weights:
    $$W_b^i = W_{b-1}^i \widetilde{w}_b(\theta_{b-1}^i, \theta_b^i) \Bigg/ \sum_{j=1}^{N} W_{b-1}^j \widetilde{w}_b(\theta_{b-1}^j, \theta_b^j). \tag{3.18}$$
if $b < B$ then
    Go to step 2.
end if
Output:
    $$\{\theta_b^i, W_b^i\}_{i=1}^{N} \approx \pi_b \qquad \text{with } \sum_{i=1}^{N} W_b^i = 1 \tag{3.19}$$
a weighted cloud of $N$ samples approximating the distribution $\pi_b$, for $b = 1, \ldots, B$.
```

rithm which considers the use of SMC proposal kernels within the MCMC framework. This thus provides high-dimensional proposals. Whiteley et al. (2009) investigate the use of Particle MCMC algorithms within a CP context which appears promising. In more general settings than that considered here in which it is not possible to integrate-out the underlying state sequence, this seems a sensible strategy.

Both the conditional CP distribution method via FMCI in a HMM framework and SMC samplers are powerful tools in their respective areas. One advantage that both components share is that the latent state sequence of the underlying MC does not need to be sampled. It is therefore worth considering whether it is possible to combine the two components such that both parameter and CP uncertainty can be considered without the need to sample the underlying state sequence. The next section presents a methodology in doing so.

3.3 Methodology

In CP problems, the main characteristics of interest are often the posterior probability of a CP occurring at a certain time, $P(\tau^{(k_{\text{CP}})} \ni t|y_{1:n})$, and the posterior distribution of the number of CPs, $P(M^{(k_{\text{CP}})} = m|y_{1:n})$. The aim of this chapter is to estimate these quantities which are in light of parameter uncertainty. In particular, they can be seen as integrating out the model parameters θ from the joint posterior, and manipulating as follows:

$$P(\tau^{(k_{\text{CP}})} \ni t|y_{1:n}) = \int P(\tau^{(k_{\text{CP}})} \ni t, \theta|y_{1:n})d\theta = \int P(\tau^{(k_{\text{CP}})} \ni t|\theta, y_{1:n})p(\theta|y_{1:n})d\theta, \tag{3.20}$$

in the case of the probability of a CP at a specific time t (the CPP). A similar expression regarding the distribution of number of CPs can be obtained. We focus on the posterior CPP throughout this section; the distribution of number of CPs can be dealt with analogously.

Equation 3.20 highlights that we can replace the joint posterior probability of a CP and model parameters by the integral of the product of two familiar quantities: $P(\tau^{(k_{\text{CP}})} \ni t|\theta, y_{1:n})$, the CP probability conditioned on θ, and $p(\theta|y_{1:n})$, the posterior of the model parameters. We have shown in Section 3.2.1 that it is possible to compute exactly $P(\tau^{(k_{\text{CP}})} \ni t|\theta, y_{1:n})$ via the use of FMCI in an HMM setting. However, it is generally not possible to evaluate the right hand side of Equation 3.20 and so numerical and simulation based approaches need to be considered.

Viewing the integral of Equation 3.20 as an expectation with respect to

$p(\theta|y_{1:n})$, that is

$$P(\tau^{(k_{\text{CP}})} \ni t|y_{1:n}) = \mathbb{E}_{p(\theta|y_{1:n})}[P(\tau^{(k_{\text{CP}})} \ni t|\theta, y_{1:n})], \tag{3.21}$$

then this reduces the estimation of the distribution of interest to a standard Monte Carlo approximation of the expectation with respect to drawing samples from $p(\theta|y_{1:n})$, and standard SMC convergence results can be applied.

Equation 3.21 can be viewed as a Rao-Blackwellised version of the estimator one would obtain by simulating both the state sequence of the underlying MC 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, Theorem 7.8 of Lehmann and Casella (1998)).

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, then by Monte Carlo results, we can approximate Equation 3.20 and 3.21 by

$$P(\tau^{(k_{\text{CP}})} \ni t|y_{1:n}) \approx \widehat{P^N}(\tau^{(k_{\text{CP}})} \ni t|y_{1:n}) = \sum_{i=1}^{N} W^i P(\tau^{(k_{\text{CP}})} \ni t|\theta^i, y_{1:n}). \tag{3.22}$$

The proposed methodology is consequently formed of three stages:

1. Approximate the model parameter posterior $p(\theta|y_{1:n})$ by a cloud of N weighted samples $\{\theta^i, W^i\}_{i=1}^N$ via the aforementioned SMC samplers in Section 3.2.2.

2. For each sample $\{\theta^i\}_{i=1}^N$, compute the conditional exact CP distribution $P(\tau^{(k_{\text{CP}})} \ni t|\theta^i, y_{1:n})$, via the FMCI and HMM framework discussed in Section 3.2.1.

3. To obtain the general CP distribution of interest in light of model parameter uncertainty, take the weighted average of the conditional exact CP distributions from step 2 with respect to weights $\{W^i\}_{i=1}^N$.

A more in-depth procedure of the proposed methodology is displayed in Algorithm 7.

An alternative Monte Carlo approach to the evaluation of Equation 3.20 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 under the proposed approach in

Algorithm 7 SMC algorithm for quantifying the uncertainty in CPs.

```
Define the following sequence of distributions

    $\pi_b \propto p(\theta) l(\theta|y_{1:n})^{\gamma_b} \qquad b = 1, \dots, b$

where $\{\gamma_b\}_{b=1}^{B}$ is a non-decreasing tempering schedule with $\gamma_1 = 0$ and $\gamma_B = 1$.
Approximating $p(\theta|y_{1:n})$
Initialisation: Set $b = 1$
for $i = 1, \dots, N$ do
    Sample $\theta_1^i \sim q_1$ where $q_1(\theta)$ is a tractable importance distribution of $\pi_1(\theta) = p(\theta)$.
end for
Compute for each $i$

    $W_1^i = \frac{w_1(\theta_1^i)}{\sum_{i=1}^N w_1(\theta_1^i)} \qquad \text{where } w_1(\theta_1) = \frac{p(\theta_1)}{q_1(\theta_1)}.$    (3.23)

if  $ESS < T$ then Resample.
for $b = 2, \dots, B$ do
    Reweighting:
    For each $i$ compute

        $W_b^i = \frac{W_{b-1}^i \widetilde{w}_b(\theta_{b-1}^i)}{\sum_{i=1}^N W_{b-1}^i \widetilde{w}_b(\theta_{b-1}^i)}$    (3.24)

        $\text{where } \widetilde{w}_b(\theta_{b-1}^i) = \frac{\pi_b(\theta_{b-1}^i)}{\pi_{b-1}(\theta_{b-1}^i)} = \frac{l(\theta_{b-1}^i|y_{1:n})^{\gamma_b}}{l(\theta_{b-1}^i|y_{1:n})^{\gamma_{b-1}}}.$    (3.25)

    Selection:
    if  $ESS < T$ then Resample.
    Mutation:
    for each $i = 1, \dots, N$ do
        Sample $\theta_b^i \sim K_b(\theta_{b-1}^i, \cdot)$ where $K_b$ is a $\pi_b$ invariant Markov kernel.
    end for
end for
Intermediary Output:

    $\pi_b \approx \{\theta_b^i, W_b^i\}_{i=1}^N, \qquad b = 1, \dots, B$

Obtaining the change point estimates of interest using FMCI
Using,
    $p(\theta|y_{1:n}) \approx \{\theta_B^i, W_B^i\}_{i=1}^N \equiv \{\theta^i, W^i\}_{i=1}^N,$

compute the CP quantities of interest in light of parameter uncertainty:

    $\widehat{P^N}(\tau^{(k_{\text{CP}})} \ni t|y_{1:n}) = \sum_{i=1}^N W^i P(\tau^{(k_{\text{CP}})} \ni t|y_{1:n}, \theta^i)$    (3.26)

    $\widehat{P^N}(M^{(k_{\text{CP}})} = m|y_{1:n}) = \sum_{i=1}^N W^i P(M^{(k_{\text{CP}})} = m|y_{1:n}, \theta^i)$    (3.27)

where $P(\tau^{(k_{\text{CP}})} \ni t|y_{1:n}, \theta^i)$ and $P(M^{(k_{\text{CP}})} = m|y_{1:n}, \theta^i)$ can be computed exactly via FMCI.
```

order to compute the CP quantities of interest, as we can account for it exactly under the FMCI and HMM framework reviewed in Section 3.2.1. For generic MCMC strategies, it is typical to sample the underlying state sequence along with the model parameters and then marginalising to obtain samples from the target distribution of interest. That is we sample from $p(\theta, x_{1:n}|y_{1:n})$, and then marginalise to obtain $p(theta|y_{1:n})$. However, it often difficult to design good MCMC moves to ensure that the chain mixes well due to the high dimensionality and inherent correlation of the state sequence. Our methodology has one advantage that we do not need to sample this underlying state sequence and that we only introduce Monte Carlo error only on the model parameters. This thus retains the exactness of the CP distributions when conditioned on model parameters. In addition, parameter estimation can be performed directly by using the sample approximation of the marginal posterior distribution of the parameters. This estimation does not require knowledge of the underlying state sequence and CP characteristics.

Other MCMC sampling strategies which do not require sampling $x_{1:n}$ are available, such as the Metropolis-Hastings sampler (see Scott (2002) and references therein for further details), and may thus be an alternative to the SMC samplers utilised in this thesis. However, these algorithms tend to perform poorly when θ is of high dimension. We thus advocate the use of SMC samplers over MCMC strategies along with the other potential benefits discussed earlier.

3.3.1 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, defining the sequence of distributions $\{\pi_b\}_{b=1}^B$ as

$$\pi_b(\theta) \propto l(\theta|y_{1:n})^{\gamma_b} p(\theta), \tag{3.28}$$

where $p(\theta)$ denotes the prior on the model parameters and $l(\theta|y_{1:n})$ is the likelihood. As the likelihood does not require sampling the underlying state sequence to evaluate it for a HMM framework, each distribution in the sequence including the parameter posterior, consequently does not need require sampling this quantity. There is great flexibility in the choice of non-decreasing tempering schedule, $\{\gamma_b\}_{b=1}^B$ such that $\gamma_1 = 0$ and $\gamma_B = 1$, ranging from a simple linear sequence, with $\gamma_b = \frac{b-1}{B-1}$ for $b = 1, \dots, B$, to more sophisticated tempering schedules. We approximate each distribution, π_b with the weighted empirical measure associated with a cloud of N samples, with the weighted sample denoted by $\{\theta_b^i, W_b^i\}_{i=1}^N$. As the weighted cloud of samples approximating the posterior $\pi_B = p(\theta|y_{1:n})$ is ultimately of interest, we simplify the notation by dropping the subscript as follows, $\{\theta^i, W^i\}_{i=1}^N \equiv \{\theta^i_B, W^i_B\}_{i=1}^N$

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 3.4.

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 dependence structure between the transition and emission parameter sets and hence assume prior independence amongst the emission parameters and the transition probabilities. Consequently,

$$p(\theta) = p(\eta)p(\mathbf{P}). \tag{3.29}$$

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

$$p(\mathbf{P}) = \prod_{h=1}^{H} p(p_h) \tag{3.30}$$

$$p(p_h) \sim \text{Dirichlet}_H(\alpha_h), \qquad h = 1, \ldots H \tag{3.31}$$

where $p_h = (p_{h1}, \ldots, p_{hH})$ 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 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 sequence remains in the same state for a significant number of time periods. We thus choose our hyperparameters to reflect this. We also note that since $\mathbf{P}$ is a stochastic matrix, there are only $H(H-1)$ unknown transition probabilities that need to be estimated.

There is also considerable flexibility when implementing the prior specification 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.

By assuming standard distributions for the prior of each component of θ, this means that we can sample from the parameter prior directly. Consequently, 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, importance sampling could be implemented for non-standard distributions: 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 $W_1^i \propto \frac{p(\theta_1^i)}{q_1(\theta_1^i)}$. 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 $\pi_1 = p(\theta)$.

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

Having obtained an approximation of distribution π_{b-1} in terms of a weighted cloud of samples $\{\theta_{b-1}^i, W_{b-1}^i\}_{i=1}^N$, it is now necessary to mutate and re-weight samples such that it approximates π_b. This can be achieved by reweighting, possibly resampling and then mutating existing samples with a π_b-invariant Markov kernel, $K_b(\theta_{b-1}^i, \cdot)$. There is a great deal 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 any 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 Hastings (RWMH) strategy on a logit scale. Mutation of the logit scale 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 3.4.

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

$$w_b(\theta_b^i) = W_{b-1}^i \widetilde{w}_b(\theta_{b-1}^i) \qquad W_b^i = \frac{w_b(\theta_b^i)}{\sum_{i=1}^N w_b(\theta_b^i)}, \tag{3.32}$$

where $\widetilde{w}_b(\theta_{b-1}^i) = \frac{l(\theta_{b-1}^i|y_{1:n})^{\gamma_b}}{l(\theta_{b-1}^i|y_{1:n})^{\gamma_{b-1}}}$ by substituting π_{b-1} and π_b into Equation 3.17. Note that the incremental weights do not depend on the new mutated particle θ_b^i, allowing resampling to be performed before sampling $\{\theta_b^i\}$ in the mutation step. Indeed, it is more intuitive to consider reweighting the existing sample approximation to target π_b, to resample, and then to mutate the sample approximation of π_b according to a π_b-invariant Markov kernel.

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}.

3.4 Results and Applications

The following section applies the proposed methodology of Section 3.3 to simulated and Econometric datasets. The Econometric dataset analysed is more specifically the aforementioned Hamilton's US GNP (Hamilton, 1989) where interest lies in determining the starts and ends recessions.

Hamilton's US GNP data can be modelled by Hamilton's Markov Switching Autoregressive model of order r, HMS-AR(r) (Hamilton, 1989). The model for the observation at time t, y_t, is defined as:

$$y_t = \mu_{x_t} + a_t \tag{3.33}$$

$$a_t = \phi_1 a_{t-1} + \ldots + \phi_r a_{t-r} + \epsilon_t \qquad \epsilon_t \sim N(0, \sigma^2), \tag{3.34}$$

where the underlying mean μ_{x_t} is dependent on the underlying hidden state x_t, and y_t is dependent on the previous r observations in an autoregressive manner via the 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

$$\begin{aligned} f(y_t|x_{1:t}, y_{1:t-1}, \theta) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2} \left[a_t - \left(\sum_{j=1}^r \phi_j a_{t-j} \right) \right]^2 \right\} \qquad (3.35) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2} \left[(y_t - \mu_{x_t}) - \left(\sum_{j=1}^r \phi_j (y_{t-j} - \mu_{x_{t-j}}) \right) \right]^2 \right\}. \end{aligned}$$

Note that Y_t is dependent on $r+1$ underlying states of the Markov chain $X_{t-r:t}$, in addition to the previous r observations $y_{t-r:t-1}$. The model parameters to be estimated under such a model are the transition probabilities, state dependent means, global precision and AR coefficients, $\theta = (\mathbf{P}, \eta) = (\mathbf{P}, \mu_1, \ldots, \mu_{|\Omega_X|}, \lambda = 1/\sigma^2, \phi_1, \ldots, \phi_r)$. We consider a common Bayesian practice of working with the precision as opposed to the variance.

In addition to modelling business cycles in Econometrics, HMS-AR(r) models are also applied in Biology to model functional Magnetic Resonance Imaging data (fMRI, see (Peng, 2008)). We consider such an application to brain signals in Chapter 4. A two state HMS-AR(r) is often assumed in modelling Hamilton's GNP data (see Hamilton (1989); Aston et al. (2011)) with the two underlying states corresponding to a "contraction" and "expansion" state. Motivated by the potential behaviour that can arise from such a model, we consider analysing simulated data from a two state HMS-AR(r) model in Section 3.4.1, before analysing the aforementioned GNP data in Section 3.4.2.

3.4.1 Simulated Data

We consider simulated data from a 2-state Hamilton MS-AR model of order 1, HMS-AR(1). The model parameter vector is explicitly $\theta = (p_{11}, p_{22}, \mu_1, \mu_2, \lambda, \phi_1)$. We now proceed in discussing a potential implementation for such a model.

Implementation for a 2-state HMS-AR(1)

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. Consequently, we can employ the following prior distributions for the emission parameters:

$$\begin{aligned} \mu_1 \sim & \mathrm{N}(0, \sigma^2_{\mu_1} = 50) & \mu_2 \sim & \mathrm{N}(-1, \sigma^2_{\mu_2} = 50) \\ \lambda \sim & \mathrm{Gamma}(\mathrm{shape} = 5, \mathrm{scale} = 2) & \phi_1 \sim & \mathrm{Unif}(-1, 1) \end{aligned} \tag{3.36}$$

Other priors could also be implemented, dependent on one's belief about the parameters. The chosen prior distributions respect our belief and the domain of the parameters. To obtain interpretable results and aid with state identifiability, we introduce the constraint $\mu_1 < \mu_2$, which can be viewed as specifying a joint prior distribution proportional to $\mathrm{N}(\mu_1; 0, \sigma^2_{\mu_1})\mathrm{N}(\mu_2; -1, \sigma^2_{\mu_2})\mathbf{1}_{(\mu_1,\infty)}(\mu_2)$ where $\mathbf{1}_A(x)$ denotes the indicator function on set A evaluated at x. To maintain stationarity within regimes, we constrain the roots of the AR polynomial to lie within the unit circle;

that is $|\phi_1| < 1$ in this example. In additional, as no information is provided regarding the AR parameter, we 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. In the case of the AR order being greater than one, the corresponding Partial Autocorrelation Coefficients (PACs) can be considered in place of the AR coefficients to maintain AR stationarity.

As mentioned previously in Section 3.3, we assume an asymmetric Dirichlet prior for the transition probabilities such that transition matrices encouraging persistent behaviour in states are favoured a priori. Using the benchmark that the majority of mass should occur in the $(0.5, 1)$ interval similar to that of Albert and Chib (1993), we employed the following priors in this particular case.

$$p_{11} \sim \text{Beta}(3,1) \qquad p_{22} \sim \text{Beta}(3,1) \tag{3.37}$$

We mutate current samples, θ via a RWMH proposal applied sequentially to component(s) of θ, conditioned on the most recent values of the other other components (akin to a Gibbs samplers). More specifically the mutation strategy is:

i. Mutate p_{11}, p_{22} simultaneously via RWMH on a logit scale, with some specified correlation structure. That is, proposals for the transition probabilities, $p_{11}^\star, p_{22}^\star$ are obtained via:

$$\begin{bmatrix} l_{11}^\star = \log\left(\frac{p_{11}^\star}{1-p_{11}^\star}\right) \\ l_{22}^\star = \log\left(\frac{p_{22}^\star}{1-p_{22}^\star}\right) \end{bmatrix} \sim \text{N}\left(\begin{bmatrix} l_{11} = \log\left(\frac{p_{11}}{1-p_{11}}\right) \\ l_{22} = \log\left(\frac{p_{22}}{1-p_{22}}\right) \end{bmatrix}, \Sigma = \begin{bmatrix} \sigma_p^2 & \rho_p \\ \rho_p & \sigma_p^2 \end{bmatrix} \right), \tag{3.38}$$

where σ_p^2 is the proposal variance for the transition probabilities, and ρ_p is a specified covariance between l_{11} and l_{22}. This proposal is accepted with the following acceptance probability

$$\min\left\{1, \frac{p(p_{11}^\star)p(p_{22}^\star)l(\theta^\star|y_{1:n})^{\gamma_b}|\prod_{i,j\in\Omega_X} p_{ij}^\star|}{p(p_{11})p(p_{22})l(\theta|y_{1:n})^{\gamma_b}|\prod_{i,j\in\Omega_X} p_{ij}|}\right\}, \tag{3.39}$$

where $\theta^\star$ denotes θ with the proposal $p_{11}^\star$ and $p_{22}^\star$, in place.

ii. Mutate μ_1, μ_2 independently via RWMH on the standard scale. That is, proposals, $\mu_i^\star$ are randomly sampled from

$$\mu_i^\star \sim \mathrm{N}(\mu_i, \sigma_\mu^2) \qquad i = 1, 2, \tag{3.40}$$

where σ_μ^2 is the specified proposal variance for the means. The corresponding acceptance probability from this proposal is consequently,

$$\min\left\{1, \frac{p(\mu_i^\star) l(\theta^\star | y_{1:n})^{\gamma_b}}{p(\mu_i) l(\theta | y_{1:n})^{\gamma_b}}\right\} \tag{3.41}$$

where $\theta^\star$ is the proposal model parameter containing the proposal mean $\mu_i^\star$.

iii. Mutate λ via RWMH on a log scale. Proposals, $\lambda^\star$ are thus sampled via

$$\log(\lambda^\star) \sim \mathrm{N}(\log(\lambda), \sigma_\lambda^2), \tag{3.42}$$

where σ_λ^2 is the specified proposal variance for the precision. This is accepted with probability

$$\min\left\{1, \frac{p(\lambda^\star) l(\theta^\star | y_{1:n})^{\gamma_b} |\lambda^\star|}{p(\lambda) l(\theta | y_{1:n})^{\gamma_b} |\lambda|}\right\} \tag{3.43}$$

where $\theta^\star$ is the proposal parameter sample as a result of the proposal precision $\lambda^\star$.

iv. Mutate ϕ_1 by transforming onto the interval $(0, 1)$ and then performing RWMH on a logit scale. That is, proposals $\phi_1^\star$ are obtained as follows,

$$l^\star = \log\left(\frac{\phi_1^\star + 1}{1 - \phi_1^\star}\right) \sim \mathrm{N}\left(l = \log\left(\frac{\phi_1 + 1}{1 - \phi_1}\right), \sigma_{\phi_1}^2\right), \tag{3.44}$$

where $\sigma_{\phi_1}^2$ is the proposal variance for the AR parameter. The corresponding acceptance probability is

$$\min\left\{1, \frac{p(\phi_1^\star) l(\theta^\star | y_{1:n})^{\gamma_b} |(\phi_1^\star + 1)(1 - \phi_1^\star)|}{p(\phi_1) l(\theta | y_{1:n})^{\gamma_b} |(\phi_1 + 1)(1 - \phi_1)|}\right\} \tag{3.45}$$

where $\theta^\star$ contains the proposal AR parameter $\phi_1^\star$.

The SMC framework presented above with the proposal kernels K_b on the sample θ_b^i corresponds to the composition of a sequence of Metropolis-Hastings kernels (and the associated backward kernel). We note that the RWMH mutations are performed on different scales due to the differing domains and constraints of the parameters. To ensure good mixing, we mutated the transition probabilities simultaneously as we believe that there 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 simulated data results, 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 carried out whenever the ESS fell below $T = N/2$.

There is evidently a trade-off between the accuracy of approximations to their target distributions, and computational costs with large values of N and B leading to better approximations. The current values were motivated by pilot studies: we found that essentially indistinguishable estimates are produced when using $N = 10,000$ samples.

Results

The following results consider a variety of data where the AR parameter, ϕ_1, varies in value. We fix however, the underlying state sequence (consequently the CPs) 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 simulate sequences of 200 observations under a variety of

AR parameter values ranging from 0.5 to 0.9 resulting in the defined CPs becoming increasing less obvious in their location and number.

Figure 3.4 displays the various simulated time series and the respective underlying state sequence together with the CPP plot (left column) and the distribution of the number of CPs (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 CP results consider changes into and out of regime 1 which is that with smaller mean for at least 2 time periods ($k_{\text{CP}} = 2$ and $s = 1$). The CPP plots display the probability of switching into and out of this regime (black solid and red dotted line respectively). 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 CP characteristics of interest. CPPs are centred around the true locations of the starts and ends of the regime of interest and the general features of the observed time series. The shape and peaks of the CPPs provide a good indication of potential estimates of the CP location. 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 CP characteristics becomes more diffuse. This is a result of the data being less informative with respect to the defined CPs as ϕ_1 increases and the behaviour associated with each regime is less distinct (see for example the data concerning $\phi_1 = 0.9$ at around time 60). 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 of no CPs is not negligible for $\phi_1 = 0.75$ and for $\phi_1 = 0.90$ which captures the uncertainty regarding the general CP configuration. These results illustrate the necessity of accounting for CP uncertainty in CP estimates.

Table 3.1 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. That is, $\bar{\theta} = \sum_{i=1}^{N} W^i \theta^i$. In addition, we provide Monte Carlo estimates of the posterior standard deviation, $\sqrt{\sum_{i=1}^{N} W^i (\theta^i - \bar{\theta})^2}$. We observe that the posterior values are reasonably close to the true values used to generate the time series.

As ϕ_1 increases and consequently the data becomes less informative with respect to the defined CPs, the estimates are less accurate with greater deviation from the true values. Nevertheless, we observe that the model parameter posterior has been reasonably well approximated.

To highlight why capturing the uncertainty of both parameter and CP characteristics is important, we also consider the exact CP distributions obtained by conditioning on these posterior means. From the corresponding plots in Figure 3.4, quite different results can be achieved; some of the uncertainty concerning the possible additional CPs has not been captured (see, for example the two CPP plots when $\phi_1 = 0.75$). The slight nuances around time 150 have not been captured in the CPP plot under the exact approach compared to the proposed SMC based approach. This "ironing-out" of the CPP is due to the absence of the parameter uncertainty. This is reflected in the distribution of the number of switches to the regime of interest where almost all mass is placed on two switches having occurred. Further possible CP configurations and estimates are thus not captured under the exact CP approach. This apparently improved confidence could be dangerously misleading in real applications.

The importance of accounting for parameter uncertainty in CP problems is successfully illustrated further in the $\phi_1 = 0.75, 0.9$ scenarios due to the differences in CPP plots and CP distributions between the exact approach conditional on the posterior mean and SMC approach. For $\phi_1 = 0.9$, 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. The true behaviour of the underlying system is therefore not correctly identified in this instance. Thus obtaining results by conditioning on model parameters may provide misleading CP conclusions and accounting for model parameter uncertainty is able to provide an general overview with regards to different types of possible CP behaviours that may be occurring. The proposed approach concurs with Bayesian inference in that all inference is based upon the full posterior distribution where nuisance parameters (the model parameters) have been marginalised out.

3.4.2 Hamilton's GNP data

We now return to our Econometric application. Hamilton's GNP data (Hamilton, 1989) consists of differenced quarterly logarithmic US GNP between 1951:II to 1982:IV. The data is found to be adequately modelled by a HMS-AR(r) model where y_t represents the logged and differenced GNP data. Of particular interest in

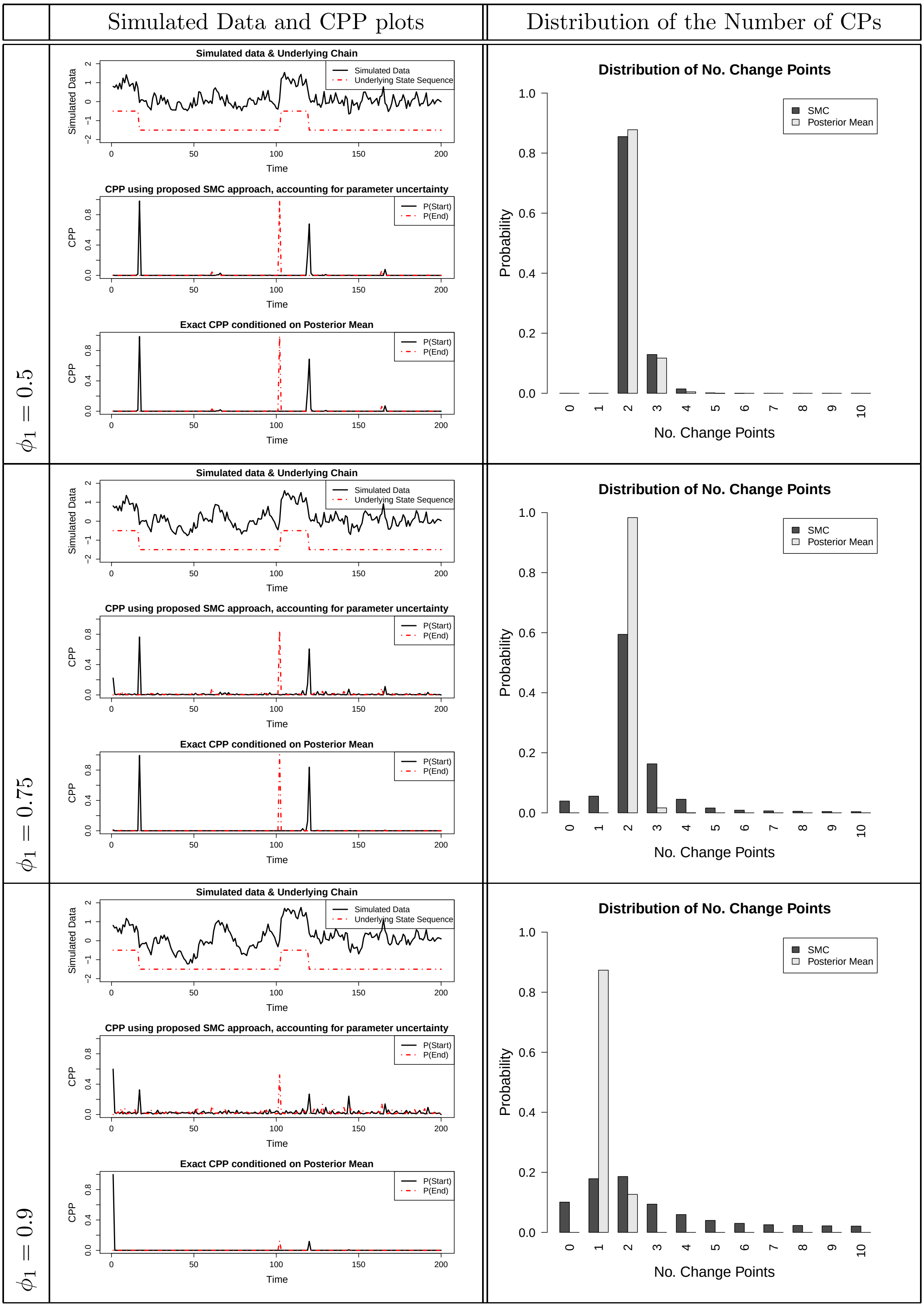


Figure 3.4: Results on Simulated Data from a HMS-AR(1) model. We consider a variety of data and display the CPP plots and distribution of number of CP under our proposed SMC based methodology. Comparisons with the exact CP distributions conditioned on the posterior mean is also presented.

	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 3.1: Estimated posterior means and posterior standard deviations (in parentheses) of parameters for the three simulated time series from a HMS-AR(1) model.

this dataset is to identify the starts and ends of business cycles, namely recessions. CP methods have thus been proposed as a means of determining these recession characteristics.

Official estimates of recession characteristics are provided by NBER. These estimates are provided by considering other economical measures such as unemployment rates. There is evidently uncertainty and ambiguity associated with these recession estimates which needs to be captured. This thus makes the dataset ideal in applying our proposed methodology. Several existing CP methods have also been applied to the dataset. Hamilton (1989) determine the starts and ends of recession by a thresholding method on smoothed probabilities of the underlying state at each time assuming a HMS-AR(4) model. Albert and Chib (1993) consider an auxiliary HMM in the spirit of Chib (1998), to sample from the joint posterior of the underlying state sequence and parameters via Gibbs sampling. More recently, Aston et al. (2011) compute the exact CP distributions conditional on MLE. Some sensitivity analysis of CP results with respect to the conditioned parameters has been performed in Aston et al. (2011) although as our simulated data results have highlighted, it is important to capture the parameter uncertainty more explicitly in CP results. In addition to quantifying the uncertainty of NBER's recession estimates, we are also able to assess the estimates provided by these CP methods under our proposed methodology.

GNP data and other measures of economical performance are often modelled as arising from two potential states; "contraction" and "expansion". In addition, a dependence structure is typically present in Econometric datasets. Consequently, a two state HMS-AR(4) is found to be adequate in modelling the GNP data of interest, where the underlying state space is $\Omega_X = \{$"contraction", "expansion"$\}$ and the autoregressive order is four (Hamilton, 1989). A widely held definition of a recession to be in progress is two consecutive "contraction" periods. As a result, we deem a recession to have occurred at time t when there is a run of minimum length of 2 ($=k_{\text{CP}}$) in the "contraction" ($=s$) state of the underlying Markov chain.

Figure 3.5 displays the CP results generated under our proposed methodology. Similar SMC settings as those in the simulated data section have been utilised: $N = 500$ samples, $B = 100$ distributions, $\mu_1, \mu_2 \sim \text{N}(0, 10)$, $\lambda \sim \text{Gamma}(1, 1)$, $p_{11}, p_{22} \sim \text{Beta}(10, 1)$. An arbitrary strong prior for the transition probabilities has been utilised, namely to reflect a stronger persistent nature in the underlying MC; switches between "contraction" and "expansion" states and their corresponding regimes do not occur too frequently. As a consequence, we set $\rho_p = 0$, that is, there is no correlation structure between the proposals of the transition probabilities. The tighter mean priors also reflects the range of the data considered. In particular, Figure 3.5(a) presents a plot of the US GNP data analysed (first panel) and the CPP plot under an exact MLE and proposed SMC approach (second and third panel, see Hamilton (1989) for MLE). The CPP plot in particular displays the probability of a recession starting (black line) and ending (red dotted line) respectively. The grey regions denote the recession periods estimated by NBER. Figure 3.5(b) displays the distribution of the number of recessions under an exact MLE and proposed SMC approach.

Accounting for parameter uncertainty under the proposed SMC approach generates promising CP results. The CPPs are still peaked and centred around NBER's estimates and the mode of the distribution of the number of recessions is seven. There is therefore evidence that NBER's estimates are plausible, although the uncertainty quantified by the proposed approach also highlights that other recession configurations are also plausible.

In comparison with an exact MLE approach, we observe that both the CPP plot and distribution of number of recessions are less peaked and less pronounced under the SMC approach. For example, in the distribution of the number of recessions, less probability is assigned to six–eight recessions occurring and assigned to other configurations including zero recessions. The CPP shape has also changed quite noticeably for the fourth to seventh recession. Such less pronounced behaviour and different CPP profiles is not surprising since we are accounting for additional uncertainty and therefore potentially highlighting different CP (recession) configurations under different parameter settings.

Having obtained posterior distributions for the CP characteristics of interest, it is now possible to obtain CP estimates which may be more useful in decision making. There are variety of ways in which estimates can be obtained with any Bayesian loss function being applicable. We take the mode of the corresponding distribution as the estimate of the number of recessions, $\hat{M}$ (the MAP estimates). The estimate of the CP locations can be obtained by locating the time point at which at least half of the probability for the uth CP lies. More specifically,

$$\hat{M} = \arg \max_{m=0,\dots,M^{\max}} \left\{ P(M^{(k_{\mathrm{CP}})} = m | y_{1:n}) \right\}, \tag{3.46}$$

$$\hat{\tau}_u = \inf \left\{ t \in \{u+1, \dots, n\} \middle| P(\tau_u^{(k_{\mathrm{CP}})} \leq t | y_{1:n}) \geq \frac{P(\tau_u^{(k_{\mathrm{CP}})} \leq n | y_{1:n})}{2} \right\} \qquad u = 1, \dots, \hat{M}, \tag{3.47}$$

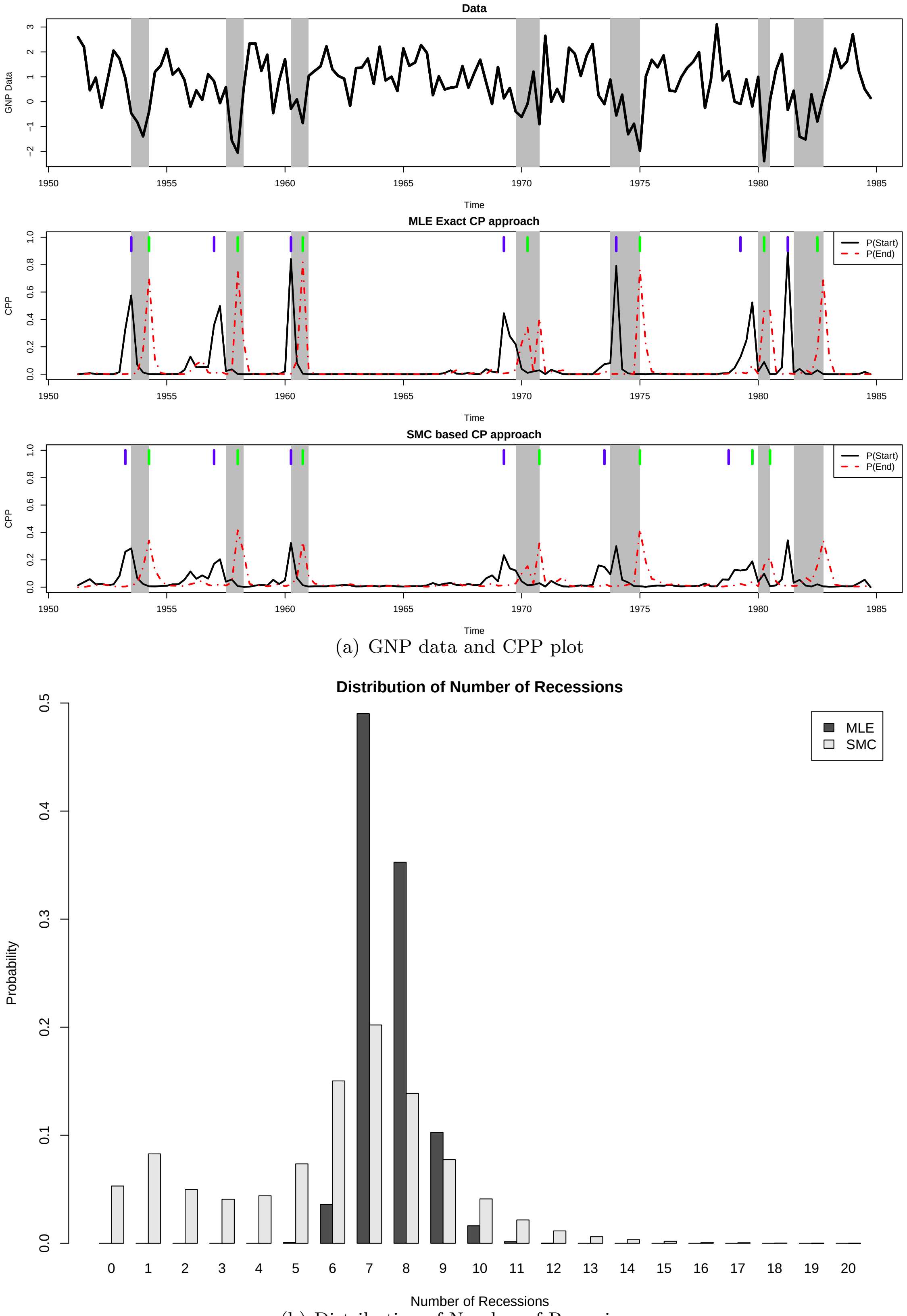
where $P(\tau_u^{(k_{\mathrm{CP}})} \leq t | y_{1:n}) = \sum_{q=1}^{t} P(\tau_u^{(k_{\mathrm{CP}})} = q | y_{1:n})$, the distribution of the time of the uth CP. Our recession estimates under the proposed approach are represented by the blue and green ticks at the top of the CPP plots.

Under such an approach, the estimate of the number of recessions concurs with NBER's estimate (that is seven recession is estimated), and estimates of the start and end of recessions fall near those provided by NBER. The discrepancies occurring for the final two recessions is a result of the uncertainty and highlighting another possible recession configuration. The final NBER recession is also estimated under the proposed approach if an eighth recession is assumed to have occurred under our estimation procedure.

Sensitivity Analysis

There may be some interest as to how sensitive our CP results are with respect to the SMC implementation, for example the number of samples considered in approximating distributions and the hyperparameters assumed for the prior distributions. Figures 3.6 and 3.7 present CP results (CPP plots and distribution of number of CPs) under different SMC scenarios, namely considering 1000 samples, and more disperse priors: $p_{11}, p_{22} \sim \mathrm{Beta}(5,1)$, $\mu_1, \mu_2 \sim \mathrm{N}(0,100)$ and $\lambda \sim \mathrm{Gamma}(1,2)$. We consider each of these scenarios individually, keeping all other conditions as in the main GNP analysis presented previously.

We observe that the CPP plots remain largely unchanged under the new scenarios, with a few subtle changes present. For example, the CPP profile for the recession around 1980 retains the same shape generally although subtle nuances are present in all scenarios. Larger discrepancies in CP results are exhibited in the



(a) GNP data and CPP plot

(b) Distribution of Number of Recessions

Figure 3.5: CP results for US GNP data under the proposed methodology; the data, CPP plots and the distribution of number of recessions. We also compare results under an exact CP approach when conditioned on the MLE of θ. The blue and green ticks represent the estimate of the start and end of recessions, assuming seven recession (the MAP estimate).

distribution of number of recessions (see Figure 3.7).

The top panel displays the use of 1000 samples in the SMC approximation; this produces similar results to the original GNP analysis presented and thus suggests that 500 samples is sufficient in our analysis if the same hyperparameter settings is used. Consequently, there is no real incentive in considering more samples in the approximation. More noticeable differences are exhibited under different hyperparameters settings with new modes being present. The second, third and fourth panels display the recession distributions if a more diffuse prior is associated for the transition probabilities, mean and precision. We observe that the general shape of the distributions remains largely intact compared to the original conditions considered, with the distributions placing substantial probability on seven to nine recessions occurring and being centred in this region. However, there is also a noticeable change in the mode of the distribution with nine (transition probability), zero (mean) and eight (precision) being now being the most probable under the conditions considered. In the case of the mean scenario where the mode switches to zero recessions occurring with substantial probability associated with it, this is suspected to have occurred as the corresponding prior distribution is extremely disperse. The sensitivity of results to hyperparameters reinforces that as in standard Bayesian analysis, care needs to be taken in choosing prior hyperparameters.

3.5 Conclusion and Discussion

This chapter has proposed a new methodology in which the uncertainty of CP estimates has been quantified in light of parameter uncertainty. The methodology combines two recent approaches in the field of Statistics; Sequential Monte Carlo samplers and exact CP distributions via Finite Markov Chain Imbedding in a Hidden Markov Model framework. A Rao-Blackwellised 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. Conditional on these model parameter samples, exact CP distributions can be computed via FMCI without additional sampling. Consequently, sampling error is introduced only in the model parameters and less variance is associated with the CP estimate. The methodology is applicable and flexible such that a wide class of HMM models and different type of changes can be considered.

Our results have successfully demonstrated good estimation of the posterior distribution for CP characteristics for both simulated and Econometric data. This is without the need for significant application specific tuning. In addition, good ap-

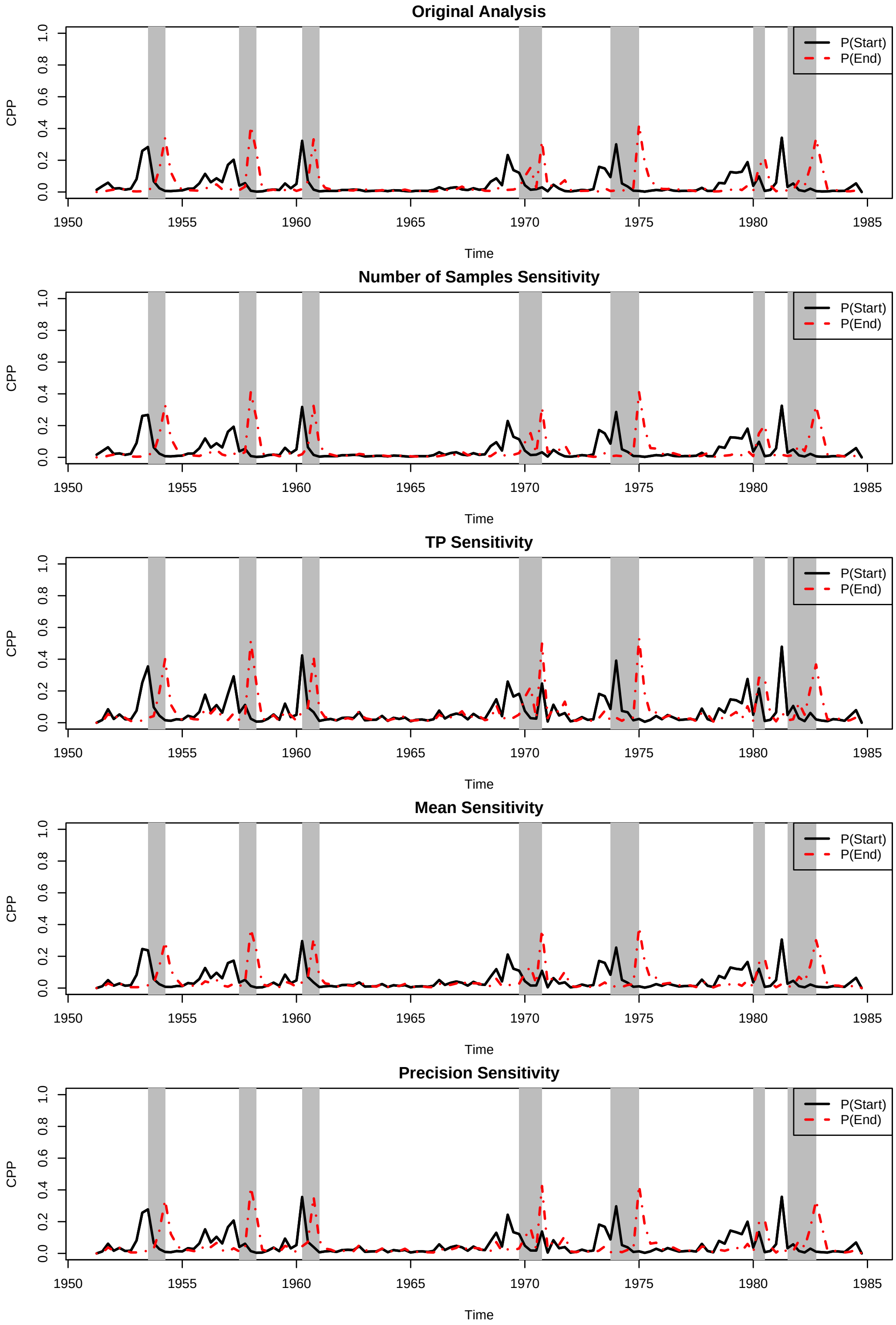


Figure 3.6: CPP plots using a different number of samples and hyperparameter values in the SMC component of our proposed framework. CPP under original SMC settings (first panel), 1000 samples used in distribution approximations (second panel), diffuse transition probability prior (third panel), diffuse mean prior (fourth panel), and diffuse precision prior (fifth panel).

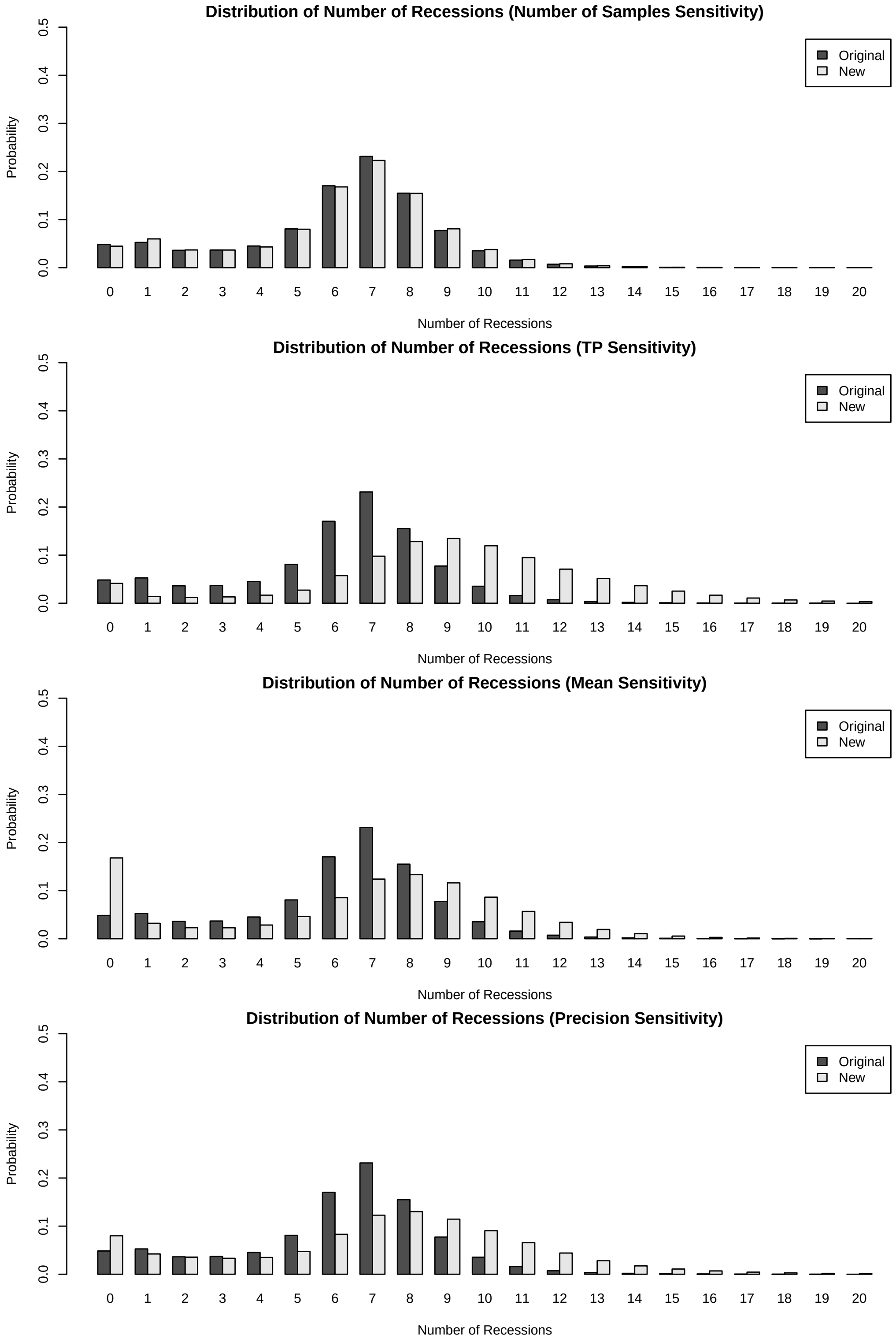


Figure 3.7: Distribution of number of recessions using a different number of samples and hyperparameter values in the SMC component of our proposed framework. Distributions using 1000 samples in distribution approximations (first panel), diffuse transition probability prior (second panel), diffuse mean prior (third panel), and diffuse precision prior (fourth panel). Distributions are displayed alongside the distribution under the original SMC settings utilised in the previous section.

proximation of the model parameter posterior is provided by the SMC framework, without the need to sample the underlying state sequence or CP characteristics. Both simulated and Econometric results demonstrate that parameter uncertainty cannot be safely ignored in CP analysis with other CP configurations being highlighted if parameter uncertainty is accounted for and ignoring it can lead to misleading conclusions. For the GNP data analysed, we deem the recession estimates provided by NBER to be plausible although other configurations are also possible.

There are a number of areas in which the proposed methodology could be improved and extended. Firstly, the number of underlying states, H, in our HMM is current assumed known a priori to analysis. This is typically not the case and thus needs to be accounted for appropriately. Recent work by Robert et al. (2000), Scott (2002) and Chopin (2007) propose Bayesian methods to account for the unknown H by approximating the posterior distribution of it via MCMC and SMC methods. We show in Chapter 5 however, that the presented SMC samplers methodology can also be used to approximate the posterior of the number of underlying states in a simple yet effective manner. This is at no additional computational cost if we also approximate the parameter posterior as performed in this chapter.

In addition, some aspects of the SMC component of this framework could be investigated further in achieving the best possible sampling performance. This will be more critical when dealing 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. Nevertheless, the SMC implementation presented in this chapter provides promising results even under generic settings.

The current definition of a CP presented considers changes into a regime explicitly, but not changes out of a regime. This poses a slight issue if brief but unsustained changes in state occur, such that transitions to a new regime do not occur. For example, in the case of $\Omega_X = \{1, 2\}$, $k_{\text{CP}} = 2$ and runs in all states are of interest, consider the scenario of $X_{t-1} = 1, X_t = X_{t+1} = 2, X_{t+2} = 1, X_{t+3} = X_{t+4} = 2$. In this case, a CP into regime 2 successfully occurs at time t. There is a brief change to state 1 at time $t+2$ but not sustained enough for a CP into regime 1 to have occurred. The underlying chain then returns to state 2 for two time periods such that under the current definition of a CP, a CP into regime 2 is said to have occurred at time $t+3$ also. However, this poses an issue in that a switch to regime 1 never occurred successfully, and thus the change at time $t+3$ should not be deemed to be a proper CP. It may therefore be worth accounting for this properly if switches between regimes as opposed to states are of interest.

This can be rectified by defining more explicitly a CP out of a regime as follows:

Definition 4. *We say a changepoint-out-of the regime corresponding to state $s \in \Omega_X$ is said to have occurred at time t' when X_t has not been in state s for k'_{CP} time periods. That is*

$$X_{t'-j} \neq s \qquad \forall\ 0 \leq j \leq k'_{\text{CP}} - 1. \tag{3.48}$$

Such a definition can be easily incorporated into the FMCI framework with the necessary modifications made in how the auxiliary MC transitions between states. This definition has the advantage that our CP results are not sensitive to brief unsustained switches in state such as outliers that may occur whilst in a regime, and in Chapter 6 where frequent but unsustained changes in state (almost periodic in nature) occur in the time series analysed. The standard change in state for a termination of a regime can also be recovered by setting $k'_{\text{CP}} = 1$.

A limitation of the modelling employed in this chapter, and in general the use of HMMs, is that by assuming a time-homogeneous HMM for the underlying Markov chain, this implicitly imposes a Geometric distribution on the duration of regime lengths. This may be an unreasonable assumption if segment and regime lengths do not follow such a distribution. This assumption could be relaxed via the use of Hidden Semi-Markov models (HSMMs, see Murphy (2002) and Yu (2010) for introductory overviews). HSMMs have a variety of applications including CP analysis (Dong and He, 2007) and can be seen as an extension of HMMs such that associated with each underlying state is information regarding the distribution spent in the corresponding state. For example, a probability mass function defined over a probable set duration times based on prior information, could be associated with the underlying state. For example, existing information and data for durations of recession and non-recession periods could be embedded via the HSMMs framework.

A wide variety of HSMMs exist, each with different assumptions for the duration distributions and state transitions, for example whether these quantities are independent of 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 onto a HMM construction. One way of observing such a framework is an underlying Markov chain with time inhomogeneous transition probabilities (Sin and Kim, 1995). This therefore naturally implies the presented FMCI methodology could be employed and lead to a CP approach in which additional information can be utilised.

Chapter 4

Quantifying the Uncertainty of Brain Activity

> **A Fox entered the house of an actor and, rummaging through all his properties, came upon a Mask, an admirable imitation of a human head. He placed his paws on it and said, "What a beautiful head! Yet it is of no value, as it entirely lacks brains."**
>
> *Aesop (from Aesop's Fables, The Fox and the Mask)*

4.1 Introduction

Functional Magnetic Resonance Imaging (fMRI) is a non-invasive imaging technique used to study and understand brain activity. Such a technique has become one of the most popular methods in the Neuroimaging community in recent years. Data collected from a fMRI scan, a series of 3D brain images collected over time, are extensively used in medical research, for example in investigating the effect of drugs on specific regions of the brain (for example Minas et al. (2012)), or determining the connectivity between different regions of the brain (see Friston (2009) for example). In answering such questions, a variety of statistical methods are employed in accounting for the spatial and temporal nature of the data, and several subjects typically partaking in a single fMRI experiment.

Statistical analysis used to model fMRI data often assumes however that the exact experimental design is known a priori in that the exact timing of the stimulus on the response is known (Worsley et al., 2002). However, this is often not the case, especially for psychological experiments where the exact onset of the stimulus on the response is unknown and subjects often react differently to stimulus. The unknown nature of activation thus needs to be accounted for in analysis. Changepoint (CP) methods have thus been proposed in determining when regions of the brain are activated and thus the onset timings of the stimulus on the response. The CPs (activations) effectively act as a latent experimental design for each time series in such a scenario. Lindquist et al. (2007) propose a CP approach under the At Most One Change assumption using a control type CP method, similar to the cumulative sum statistic (CUSUM) as reviewed in Section 2.3. However, multiple activations in a single fMRI scan can occur and the number of activations and their timings are inherently subject to uncertainty which are not accounted for explicitly by this method. There is thus interest in quantifying the uncertainty of these brain activations.

This chapter applies the method proposed in Chapter 3 to fMRI data from an anxiety induced experiment initially presented in Lindquist et al. (2007) and previously displayed in Figure 1.2 (page 6). By considering the proposed Hidden Markov Model (HMM) based CP method, this allows multiple activations to be considered and the uncertainty of the activations to be quantified. In addition, error process assumptions and detrending typically performed in fMRI statistical analysis can be included within the CP approach which thus provides a unified approach. We also demonstrate how different assumptions on the error process and detrending performed influence our analysis and results.

The structure of this chapter is as follows. Section 4.2 provides a brief background to fMRI in order to appreciate this chapter, in particular the paradigm of a known experimental design. Section 4.3 details the anxiety induced experiment from which our fMRI data is collected from. Section 4.4 presents the results of the dataset considered under the proposed methodology. Section 4.5 concludes this chapter which details potential paths of further research.

4.2 A Brief Introduction to Neuroimaging and fMRI

Neuroimaging is a discipline within the field of medicine and neuroscience which provides various non-invasive imaging techniques in studying the structure and behaviour of the human brain. The disciple combines research from a variety of areas of Science including Physics, Biology, Engineering and Statistics. Such imaging techniques have proven to be intensively useful in answering clinical and medical research questions in a safe, efficient manner.

Neuroimaging techniques can be categorised into two distinct groups; structural and functional imaging techniques. Structural imaging techniques, such as Magnetic Resonance Imaging (MRI), aims to provide a single 3D image of the brain in order to examine the structure and anatomy of the brain. Functional neuroimaging in contrast, studies the activity of the brain both spatially and temporally. A variety of techniques are available, each with their own merits and disadvantages arising from the logistical and statistical challenges. We focus our attention to a particular functional neuroimaging technique.

Functional neuroimaging techniques can be further categorised into those which measure brain activity directly and indirectly via a biological by-product of brain activity. Electroencephalography (EEG) and Magnetoencephalography (MEG) fall into the former category and account for brain activity by measuring the electrical and magnetic activity respectively from the surface of the skull. Such methods provide excellent temporal resolution (data can be sampled at a high frequency up to milliseconds) but poor spatial resolution (difficulty in identifying the region of the brain where the signal has arisen from). In contrast, functional Magnetic Resonance Imaging (fMRI) measures signals associated with the changes in blood flow due to neuronal activity via the concentration of oxygenated blood, and thus fall in the latter category. This method provides greater spatial resolution compared to EEG and MEG, but poorer temporal resolution due to slower rates associated with brain hemodynamics (changes in blood flow which take a few seconds) and how quickly the data can be collected under physical constraints of the scanner. Research regarding fMRI studies has exploded considerably in the last two decades and is the type of data of considered in this chapter. We refer the reader to Lindquist (2008) and Poldrack et al. (2011) for good overviews of fMRI and the statistical methods associated with such data.

4.2.1 Data Acquisition

As remarked earlier, fMRI data arises due to the change in concentration of oxygenated blood (oxyhemoglobin) which is a consequence of neuronal activity. Thus, when neurons in the brain become active, they require more oxygen which is supplied by the blood flow. As a result, the concentration of oxyhemoglobin in the activated region of the brain decreases. This is known as the hemodynamic response function (HRF), which describes the behaviour of change in concentration of oxyhemoglobin over time due to activation. From the blood oxygenation level dependent effect (BOLD, Ogawa et al. (1990)), oxyhemoglobin and deoxyhemoglobin possess different magnetic properties, namely diamagnetic and paramagnetic respectively. By placing a participant in a magnetic field (an MRI scanner) and examining the small changes in magnetic field arising due to the decrease (increase) of oxyhemoglobin (deoxyhemoglobin) associated with neuronal activity, it is possible to measure brain activity at different locations and time via this biological by-product. The neuronal activity is achieved by asking the subject to perform a specific task (a stimulus) such as responding to certain images or tapping their fingers. This is the essence of fMRI data acquisition. We refer the reader to Lindquist (2008) and references therein for more technical descriptions of the data acquisition procedure.

After image reconstruction of an fMRI scan, a series of 3D brain images collected over time is collected per individual. Each image consists of voxels, uniformly spaced volume elements which partition the brain into equally spaced sized cubes; this is analogous to pixels in a 2D image. Each voxel corresponds to a specific part of the brain, with the fMRI signal (the recorded change in magnetic field) over time at a particular voxel being a measure of the brain activity associated with the voxel. Regions of the brain, clusters, are a collection of voxels constituting that region. It is typical for a brain volume to consist of approximately 100,000 voxels under standard conditions of an fMRI scan (a $64 \times 64 \times 30$ image). The signal from each voxel over n time periods can be considered as a time series. Thus, one potential statistical perspective of fMRI data is as a multivariate time series dataset consisting of 100,000 time series. This multivariate time series exhibits correlation both within individual time series (temporal) and across time series (spatially).

The low temporal resolution associated with fMRI arises from two factors. Firstly, it takes approximately two seconds to obtain a single full brain volume under the standard conditions of a fMRI scan. We can thus only detect changes in magnetic field and the corresponding brain activity up to this degree of accuracy. Secondly, whilst neuronal activity typically lasts for only a few milliseconds, the HRF can last for up to twenty seconds under the canonical HRF typically assumed (Grinband et al., 2008). This affects how closely we can identify activation timings, and ultimately the design of experiments considered with the stimulus being sufficiently separated in time. The data considered in this chapter concerns a block design stimulus where stimulus is applied over a sustained period. This latter point should therefore not pose too much of a problem.

Both spatial and temporal correlation exists within the fMRI data with neighbouring voxels behaving similarly to one another (spatial correlation) and measurements collected at nearby time points being possibly correlated (temporal correlation). This correlation structure, and general noise associated with the signal, thus needs to be accounted for in analysis.

4.2.2 Preprocessing

Before any statistical analysis can be performed on a time series of 3D brain images, preprocessing is performed to remove undesired artefacts that may be present from the scanning session (both associated with the subject and the scanner itself) and to validate model assumptions to ensure that the data is suitable for statistical analysis. The various stages of preprocessing are as follows:

Slice Timing Correction Each 3D image of the brain is formed by taking several 2D images of the brain at different slices (parallel planes) of the brain. Each slice is typically taken sequentially at different times (each separated by a few milliseconds), and thus the corresponding measurements between each slice are taken at slight lags. The slice timing correction step thus corrects for the temporal shift that occurs between each of the sampled slices, and consequently each slice can be assumed to be taken simultaneously from the same time point. Such correction is typically performed by interpolation or the Fourier shift theorem.

Motion Correction It is highly likely that subjects will move their heads whilst in the MRI scanner during the experiment. Even small amounts of movement can cause a large amount of error which causes signals from a specific voxel to be contaminated by signals from neighbouring signals. It is thus necessary to correct for this in order to match the measured fMRI signals to the corresponding voxels, and remove contamination of signals from neighbouring voxels. Such correction is performed in two steps: linear transformations of the series of brain images (namely translations, rotations and scaling operations) such that it matches a target image, and then interpolating the image to create new motion corrected voxel values.

In the case of the head movement being too severe such that no amount of correction will make it viable for analysis, the subject and its corresponding scan are removed from further analysis.

Coregistration fMRI data is typically of low spatial resolution compared to MRI data and thus provides little information regarding the anatomical structures of the brain. For presentational purposes and estimating localisation, it is therefore common to map fMRI images onto a high resolution structure MRI image of the brain. This is achieved by coregistration which aligns structural and functional images via rigid body or affine transformations.

Normalisation It is common in fMRI experiments for multiple subjects to be scanned. However, it is highly likely that these subjects have different shaped and sized brains; it is thus common to map these onto a standardised template of the brain (for example the Talairach or Montreal Neurological Institute (MNI) brain (Talairach and Tournoux, 1988)) so that we can consider the activity from the same regions of the brain between individuals. In addition, this spatial normalisation provides a consistent manner in which fMRI data is reported in which fair comparisons between individuals and studies can be made. This spatial normalisation is performed by non-linear transformations to match local features of each subject's brain to the template brain.

Spatial Smoothing It is common to perform spatial smoothing to fMRI data prior to analysis which removes undesired high frequency behaviour such as noise artefacts. Spatial smoothing may improve inter-subject registration to the chosen brain template, blur any residual anatomical differences that may have resulted from spatial normalisation, to ensure that the assumptions of random field theory (for spatial analysis corrections) are valid, and to denoise images such that the signal-to-noise ratio within a region is increased. Smoothing is typical performed by convoluting the image with a 3D Gaussian kernel, with the choice of suitable bandwidth being an area of ongoing discussion.

The preprocessing steps outlined above have an obvious effect on the spatial-temporal correlation structure in the raw fMRI data. The effect of each preprocessing step thus needs to be understood, along with its consequences on the correlation structure. In addition, it is also necessary to understand the interactions between the different steps, the order in which they may be performed and its impact on the resulting data. There is thus considerable potential research in these preprocessing steps and whether they can be included directly within statistical analysis frameworks (see Lindquist (2008)).

For almost all fMRI scans, slice timing correction, motion correction and coregistration are performed in order to satisfy the assumptions necessary for statistical analysis. In addition, the dataset considered in this chapter is of a multi-subject nature, and thus the data has been normalised between individuals. Spatial smoothing has not been performed on the data provided, although as we shall consider specific regions of interest (a collection of voxels), we perform this by averaging over the voxel time series forming the regions.

4.2.3 Statistical Analysis

Having preprocessed the fMRI data into a suitable format, we can now proceed in performing statistical analysis and answering questions regarding how certain stimuli lead to changes in neuronal activity in the brain. This includes locating the regions of the brain which are associated with certain stimuli, determining the potential connectivity of regions of the brain with respect to a stimuli, and making predictions about psychological or disease states of the brain.

However, many challenges are encountered with fMRI datasets. Firstly, despite the intuitive perspective that fMRI data is a collection of 100,000 time series, the complex nature of the correlation structure and the size of the data both for individual and group analysis, makes it difficult to construct a complete statistical model which can account for this type of behaviour fully. A number of simplifications are thus required in order to balance computational feasibility with model efficiency. The most common simplification is to consider a massive univariate approach where each voxel is considered individually and independently from others initially, with some sort of spatial correction implemented towards the end. This review will focus on the massive univariate statistical approaches although approaches accounting for spatial correlation more explicitly are available (see Lindquist (2008) and references therein).

The measured fMRI signal at a specific voxel for one individual can be decomposed into three components: the BOLD signal, drift and noise. Consequently, a commonly assumed model for the fMRI signal is

$$\begin{aligned}\text{fMRI signal} &= \text{BOLD signal} && +\text{Drift} && +\text{Noise},\\ \mathbf{y} &= \boldsymbol{\Delta}\mu && +\mathbf{G}\beta && +\mathbf{a},\\ y_t &= \delta_t'\mu && +\mathbf{g}_t'\beta && +a_t, \end{aligned} \tag{4.1}$$

where $\mathbf{y} = (y_1, \ldots, y_n)'$, is a n length column vector containing the fMRI time series. The BOLD signal comprises of $\boldsymbol{\Delta} = (\delta_1, \ldots, \delta_n)'$, a $n \times k$ design matrix which typically denotes whether a stimulus is on-or-off at time t, and μ, is a k length column vector which denotes the underlying BOLD signal associated with the corresponding stimulus. The drift component consists of $\mathbf{G} = (\mathbf{g}_1, \ldots, \mathbf{g}_n)'$, a $n \times d$ matrix containing the drift covariates and β is a d length column vector containing the corresponding drift coefficients. $\mathbf{a}$ is a n length column vector corresponding to the noise associated with the signal.

The fMRI signal typically drifts slowly over time due to scanner instabilities. Such drift is associated with low frequency behaviour, and thus detrending using a high-pass filter is performed, that is removing the low frequency behaviour associated with drift and retaining the high frequency behaviour of the BOLD signal. Two common approaches in modelling drift are as a dth order polynomial with respect to (rescaled) time (polynomial detrending), or as a series of d low frequency cosine functions (discrete cosine transform basis detrending). More specifically,

$$\mathbf{g}_t'\beta = \sum_{b=1}^{d} \beta_b t^b \qquad \text{polynomial,}$$

$$\mathbf{g}_t'\beta = \sum_{b=1}^{d} \beta_b \sqrt{\frac{2}{n}} \cos\left(\frac{b\pi t}{n}\right) \qquad \text{discrete cosine transform basis.}$$

In order to model the temporal correlation present in the time series y_t, a model is associated with the noise variable a_t. In contrast to standard time series analysis however, this noise correlation is specified prior to analysis rather than being estimated from the data. This prior specification is due to the large number of time series being considered, and thus estimation is not computationally feasible. This is despite each time series being considered independently. An autoregressive model of order r is found to be adequate in capturing the potential autocorrelation present in the data.

The BOLD response signal is the main underlying signal within an fMRI signal as it corresponds to the neuronal activity in which we wish to infer. The BOLD response is typically modelled in terms of the stimuli via a linear time invariant (LTI) system, where the stimulus acts as the input, and the BOLD response is the output response function. The LTI system permits the following relationship between the stimulus and BOLD response; scaling, superposition and time-invariance. Scaling implies that scaling of the stimuli (the input) by some factor c, causes a scaling by the same amount in the BOLD response. Consequently, this means that the amplitude of the measured signal provides a measure of the amplitude of neuronal activity, and difference between measured signals corresponds to difference between neuronal activity. Superimposition implies that the response of two different stimuli corresponds to the sum of their individual responses. Time-invariance implies that a shift in time for the stimulus corresponds to a shift in the response BOLD signal by the same amount. These three properties thus allow us to differentiate between responses in various regions of the brain to multiple closely spaced stimuli.

The assumptions made regarding the BOLD response are crucial to the analysis when assuming Model 4.1. It is typical to assume that the stimulus function timings are known and thus the exact form of the experimental design matrix corresponding to the stimulus $\mathbf{\Delta}$, is known. In addition, if the HRF behaviour is assumed known a priori (typically assumed to be canonical), then Model 4.1 reverts to a multiple regression with known input signal components $\mathbf{\Delta}$, and unknown amplitudes, μ. Such assumptions lead to the popular general linear model approach in modelling fMRI signals (Worsley et al., 2002). The model essentially "models time series as a linear combination of different signal components and tests whether activity in a brain region is systematically related to any of these known input function" (Lindquist, 2008). That is, we fit Model 4.1 to $\mathbf{y}$, modelling it as a linear model using $\mathbf{\Delta}$ as the design matrix, and estimating μ. In addition, we test whether μ is significant, corresponding to association between the stimuli and BOLD response, and thus whether neural activation has occurred. Significance results are represented by a statistical map, a brain image which highlights the voxels in which μ is found to be significant and thus activation has occurred in that voxel.

This can be further extended into mixed-effects analysis for multi-subject fMRI analysis. Akin to mixed-effects models in standard statistical analysis, this allows for two levels of variation; a global level which effects all subjects of the experiment in a similar manner, and a local level which is specific to that individual. Such models can thus be used in population level inference, determining whether the activation of brain regions is generally associated with the stimulus by testing the significance of the global effects μ.

In accounting for spatial correlation, this is typically accounted post voxel-wise analysis via the use of random field theory on test statistics. As a result, this corrects the test statistics computed for an individual voxel and determines the statistical significance for the entire set of voxels. Methods which do account for spatial correlation more explicitly in modelling individual voxels have recently been developed, for example via the use of Markov random fields (see Lindquist (2008) and references therein).

The statistical methods concerning fMRI data presented above assume that the timings of the stimulus function are known exactly, such that the structure of the design matrix $\mathbf{\Delta}$ is known exactly. This is a strong assumption and may not always be the case in experiments, for example in psychological experiments where the exact onset timing of the stimulus on the BOLD response is unknown. In addition, there is no reason to assume that applying a stimulus causes an immediate effect on the BOLD response, with a delay between the two being highly plausible for any experiment. It is thus necessary to account for the unknown structure of $\mathbf{\Delta}$. One way in which this can be estimated is via CP methods, as proposed in Lindquist et al. (2007).

4.3 An Anxiety Inducing Experiment

The data presented in Lindquist et al. (2007) and analysed in this chapter concerns an anxiety inducing experiment. The experimental protocol, as outlined in Lindquist et al. (2007), is described below with a graphical representation of the experiment displayed in Figure 4.1.

> The design was an off-on-off design, with an anxiety-provoking speech preparation task 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 two minute (resting baseline). At the end of this period, participants viewed an instruction slide for 15 seconds 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 seven minute period. After two minutes of silent preparation, another instruction screen appeared (a relief instruction, 15 seconds duration) that informed participants that they would not have to give the speech. An additional two minute period of resting baseline followed, which completed the functional run.

The fMRI dataset consists of 215 images where an image is collected every two seconds. The study features 24 valid fMRI scans from 24 subjects, where scans involving excessive head motion and other prominent scanner instabilities being removed from statistical analysis.

Lindquist et al. (2007) propose a Hierarchical Exponential Weighted Moving Average based CP method (HEWMA) in determining whether regions of the brain become activated over the course of the scanning period, and estimates of any potential activation times. A massive univariate based approach is considered in analysing the entire brain. The Exponential Weight Moving Average approach (EWMA), is a control type CP method similar to the CUSUM statistic, in that the EWMA statistic is computed sequentially and compared to a baseline value. That

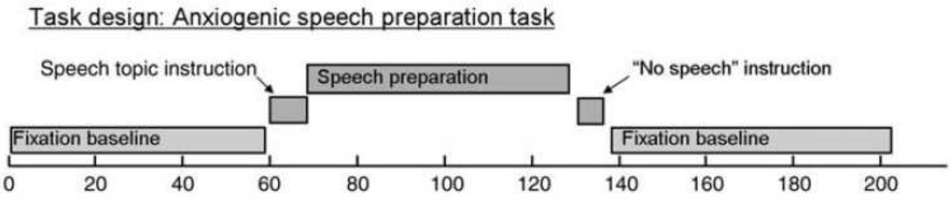


Figure 4.1: Graphical representation of the design of the anxiety inducing experiment.

is

$$z_t = \lambda y_t + (1 - \lambda) z_{t-1} \qquad t = 1, \ldots, n, \tag{4.2}$$

where $\lambda \in (0, 1)$ is a user specified smoothing parameter. A CP into the activation regime is deemed to have occurred if the EWMA statistic, z_t, deviates sufficiently (that is exceeds a threshold) from the baseline value. The duration of the activation period is also estimated by determining when the EWMA statistic returns to baseline behaviour. The hierarchical aspect is introduced as they test for activations on the 24 subjects simultaneously to obtain a general activation time over all subjects. In addition, the proposed HEWMA method corrects for the autoregressive error assumptions used within statistical fMRI analysis.

This methodology does not however easily allow for the incorporation of multiple activations (CPs) and requires detrending of the data prior to CP analysis (that is, as a preprocessing step effectively). In addition, the uncertainty of activations is only captured implicitly under the HEWMA approach via the chosen significance level of the threshold. As there is ambiguity with respect to the timing of these activations and the number of them, there is thus interest in quantifying the uncertainty of activation regimes. The methodology proposed in Chapter 3 provides one approach in doing so.

4.4 Results

The fMRI signal model presented in Model 4.1 can be rewritten in terms of a modified version of the Hamilton's Markov Switching Autoregressive model of order r (HMS-AR(r)) as presented in Equation 3.34 (page 81) with an additional trend component. More specifically, if Δ consists of zeroes and ones denoting whether a stimulus is on or off with only one stimulus being activated at most at time t. Consequently μ can be seen as the stimulus dependent BOLD response where we select the corresponding entry dependent on the stimulus configuration. If an underlying latent Markov Chain, $X_t \in \Omega_X$ is used to equivalently denote the stimulus configuration δ_t at time t in Δ, then this results in the state dependent BOLD response μ_{X_t}. Model 4.1 can thus be re-expressed as,

$$y_t = \mu_{X_t} + \mathbf{g}_t'\beta + a_t, \tag{4.3}$$

$$a_t = \phi_1 a_{t-1} + \ldots, +\phi_r a_{t-r} + \epsilon_t \qquad \epsilon_t \sim N(0, \sigma^2). \tag{4.4}$$

As before, $\mathbf{g}_t'$ denotes the d known drift covariates (either polynomial or discrete cosine) at time t, with corresponding d unknown drift coefficients β. Note that setting $\beta = \mathbf{0}$ results in the standard HMS-AR(r) model and assuming that no drift is present. As the anxiety induced experiment considers an on-off-on design, we consequently assume a 2-state modified HMS-AR(r) model where the underlying state space is $\Omega_X = \{$"resting", "active"$\}$.

Peng (2008) investigated the uncertainty of brain activation associated with this dataset, assuming the HMS-AR(r) presented above and conditional on model parameters. This chapter quantifies the brain activation with respect to model parameter uncertainty where the unknown parameters, $\theta = (p_{11}, p_{22}, \mu_1, \mu_2, 1/\sigma^2, \beta_1, \ldots, \beta_d, \phi_1, \ldots, \phi_r)$, are estimated via the Sequential Monte Carlo samplers (SMC) methodology presented in Chapter 3. As Chapter 3 has demonstrated, accounting for model parameter is equally important in CP analysis.

We focus on two specific regions of the brain; the rostral medial pre-frontal cortex (RMPFC), which is known to be associated with anxiety, and the visual cortex (VC) which is suspected to show activation behaviour associated with the presentation of the task-related instructions. The time series from these regions have been obtained by averaging over the time series from the voxels forming these clusters. In addition, we are interested in the general activation behaviour of the experiment across all subjects, and thus consider the averaged time series from each subject's cluster signal. These are the time series that shall be considered under our proposed methodology.

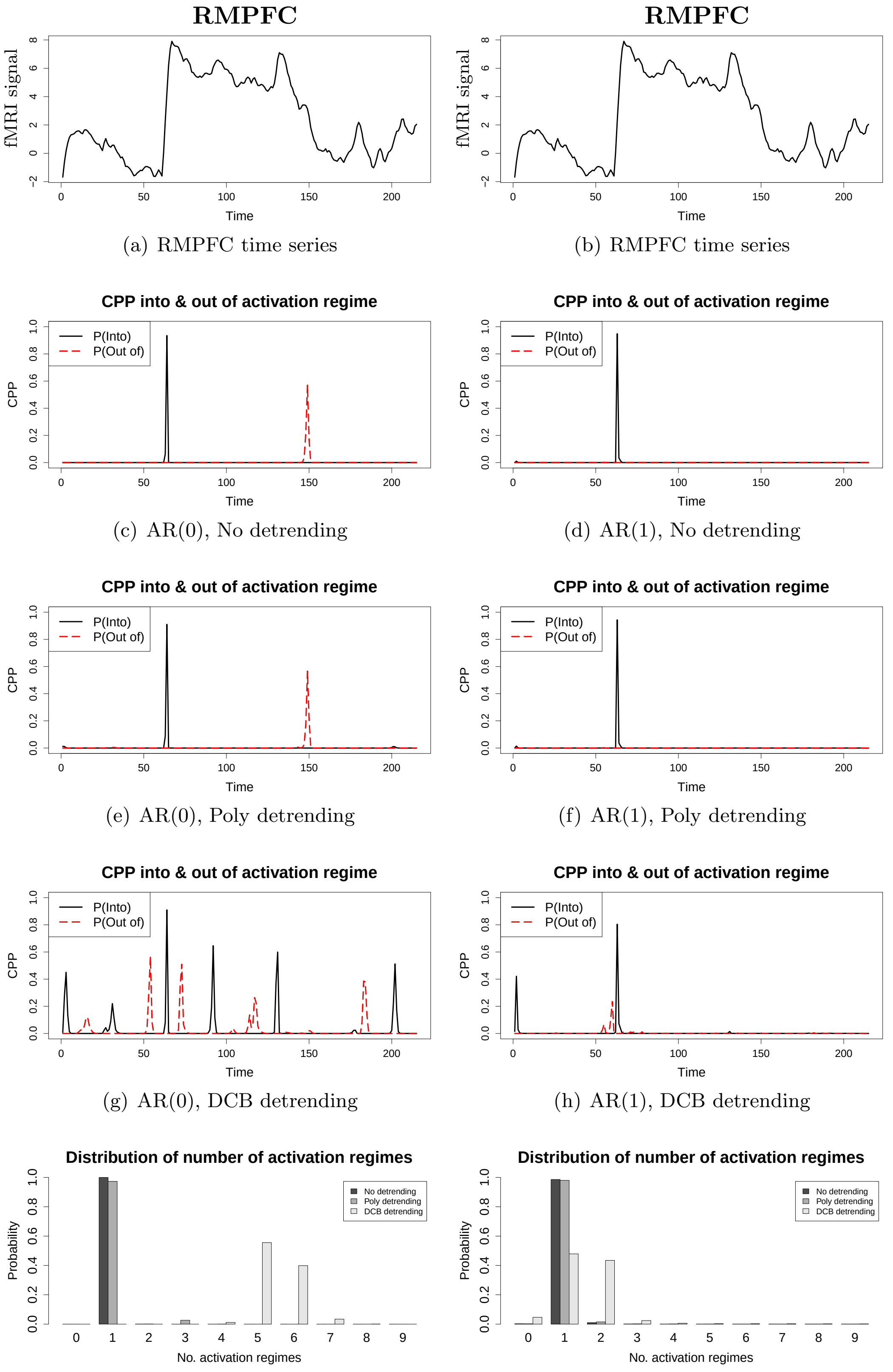
We consider several different models as a result of assuming different AR orders and performing different types of detrending. This is one of the benefits of considering the HMM based framework in that it allows model assumptions to be varied with ease. Firstly, as a baseline comparison, a model with independent errors (an AR(0) error process) and no detrending is performed. This is shown to provide unsatisfactory CP results, which is unsurprising given that changepoint detection techniques are well known to breakdown in the presence of other forms of non-stationarity such as linear trends and drift. The analysis then proceeds using various combinations of polynomial detrending of order three (Worsley et al., 2002) and discrete cosine basis detrending of order twelve (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 the default in the Statistical Parametric Mapping (SPM) software (Ashburner et al., 1999) available in the Neuroimaging community.

We deem a region to be activated when there is a sustained change into the "active" state for at least five time points in the region, thus $s =$ "active" and $k_{\text{CP}} = 5$. This is equivalent to an activation of at least 10 seconds in real time, which accounts for the biological behaviour of the HRF. Other values of k_{CP} were also considered and provided similar results (results not presented here). Similar SMC settings as those consider in Chapter 3 were employed: $N = 500$ samples, $B = 100$ distributions, linear tempering schedule, $p_{11}, p_{22} \sim \text{Beta}(3, 1)$, $\mu_1, \mu_2 \sim \text{N}(0, 50)$, $\frac{1}{\sigma^2} \sim \text{Gamma}(1, 1)$.

The resulting CP distributions for the two regions of the brain are presented in Figures 4.2 and 4.3, where we display CPP plots (the probability of an activation regime starting and ending) and the distribution of the number of activation regimes. The CP results under the proposed methodology finds significant evidence that there is at least one CP, and thus activation, in both regions of the brain. This accords with the HEWMA analysis, where both regions were shown to have a CP, with the RMPFC region associated with the anxiety stimulus, and VC with the visual instruction timings. In addition, this concurs with the design of the experiment.

However, quite different CP results are obtained under the different error and detrending assumptions implemented. For the RMFPC region, if an AR(0) with no detrending is used, then two distinct changes, one into the activation region and one out of the activation region are determined. This corresponds to the speech preparation period when subjects are suspected to experience some anxiety. 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 is consistent with only one change to activation from baseline during the scan. 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 CP distribution if a discrete cosine basis is used, but this is likely due to identifiability issues between the cosine basis and the CPs present.

On examining the regions of the VC, the choice of detrending is critical. If



(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 4.2: Changepoint analysis results for the RMPFC region of the brain with respect to different order models and detrending. This region is associated with emotions such as fear and anxiety.

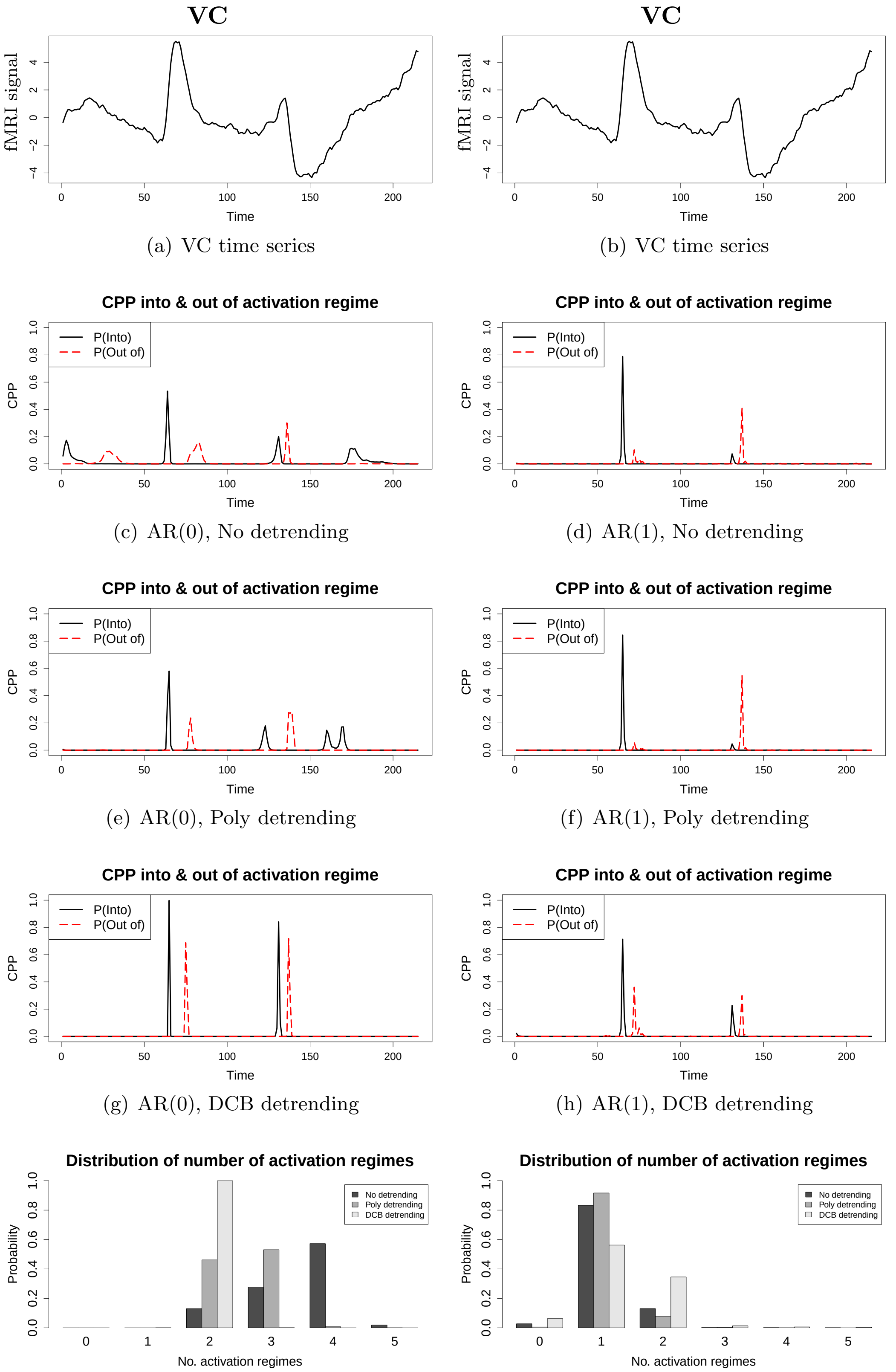
a suitable detrending is assumed, in this case a discrete cosine basis trending, a clear CP distribution with multiple CPs is found, corresponding directly to the two visual instructions presented to subjects. However, if no or a small order polynomial detrending is used, the CP distributions associated 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 CP distribution.

It is typical to assume and fix the AR coefficient associated with the noise model in analysis. Consequently, we also consider fixing the AR error process coefficient to $\phi_1 = 0.2$ as featured in the SPM software (Ashburner et al., 1999). The CP results (not presented) contain features present in both results AR(0) and AR(1) analysis with more peaked and centred CP probability features compared to the presented AR(1) results. This is not surprising since less uncertainty is present by fixing the value of the AR parameter.

4.5 Conclusion and Discussion

This chapter has applied the HMM-based changepoint method presented in Chapter 3 on fMRI data in quantifying the uncertainty of brain activity. This is an important aspect of statistical analysis regarding fMRI data as the experimental design is often assumed known with the exact timings of the stimulus on the BOLD response being known a priori. This is a strong assumption and is typically not the case, especially for psychological experiments such as those considered in this chapter. CP methods have thus been proposed in estimating the timings regarding the experimental design (Lindquist et al., 2007). In addition, these timings are subject to uncertainty with subjects reacting differently to the stimuli which thus needs to be accounted for. The proposed HMM-based CP approach thus provides one way of both estimating and accounting for the uncertainty regarding the unknown timings. The results under the proposed methodology concur with the activation results of other CP methods and the general design of the experiment.

The proposed methodology also provides a unified framework in which different assumptions regarding scanner instabilities can be made, namely the assumptions of the drift model and error process. Such assumptions are typically considered as a preprocessing step and to be known and fixed in other methods such as Lindquist (2008). Typical statistical analysis assuming a known experimental design are found to be robust to such assumptions (Worsley et al., 2002). However, these assumptions are found to be highly influential on our CP results. A misspecification of the drift model produces CP results which identify the expected CPs according



(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 4.3: Changepoint analysis results for the VC region of the brain with respect to different order models and detrending. This region is associated with visualisation.

to the design of the experiment, although variation does exist between the different models assumed. Not accounting for potential autocorrelated error provides more noticeable discrepancies with underestimation of the uncertainty associated with the CPs and thus the underlying experimental design. Care is therefore required when analysing time series arising from experimental design such as fMRI experiments.

There are various areas of potential further research associated with the anxiety induced fMRI data and other fMRI datasets with unknown experimental design. As the CP results are sensitive to the error process structure assumed, further investigation on such an influential factor is worth pursuing. Autoregressive orders which capture the autocorrelation induced by physiological artefacts such as heart beats and respiration should thus be investigated.

We have thus far considered two specific regions of the brain and performed independent analysis on them. However, analysis regarding the entire brain is often of interest and thus worth further investigation. In addition, it may be advantageous to exploit the spatial correlation present in the brain in producing more efficient methodologies. For example, rather than performing the SMC samplers for each cluster of interest, it may be possible to perform the SMC samplers algorithm at a fewer number of clusters and to "borrow" model parameters samples generated under the algorithm for different clusters. This "borrowing" scheme is determined via the spatial correlation and could be implemented via the use of Markov Random Fields (Chellappa and Jain, 1991), similar to Woolrich et al. (2005). Such a framework would thus allow us to capture both the temporal and spatial correlation structure which is inherently present in fMRI data.

This chapter has considered analysis on a single fMRI signal at a specific region of the brain which has been obtained by averaging over the fMRI signals from each subject of the experiment at the corresponding region. However, it seems wasteful to reduce the data in such a way and although we have provided a global estimate of the potential underlying experimental design, subjects often react differently to stimuli and it is thus worth investigating the experimental design corresponding to individual subjects. There are two potential paths in which this can be considered.

Firstly, similar to the analysis presented in this chapter, we apply our HMM-based CP methodology to each subject's fMRI signal, independently of each other. We consequently have a set of J CP results for each of the J subjects involved in the experiment. To infer the global experimental design, these J CP results could then be combined and summarised in some way such that the variation between individual's experimental designs could also be captured.

The second approach is to assume some sort of hierarchical structure is present, a common approach as in other multi-subject experiments (see Lindquist (2008)). Here, it seems intuitive to assume a two level structure; a global level which represents how all subjects will generally react to the stimuli, and an individual specific level which models how individuals uniquely react to the stimulus. A potential model would thus be

$$y_{jt} = \delta_t'\mu + \mathbf{v}_{jt}'\eta_j + \mathbf{g}_t'\beta + a_{jt} \qquad j = 1, \ldots, J, \tag{4.5}$$

where $\mathbf{v}_{jt}'\eta_j$ denotes the specific BOLD response for individual j. Such a model may thus also allow us to capture the inter variability between subjects and their experimental designs. Chapter 6 proposes a methodology which considers changepoints in a multivariate time series setting which may be feasible in this multi-subject context.

Chapter 5

Model Selection for Hidden Markov Models

> **Statisticians, like artists, have the bad habit of falling in love with their models.**
>
> *George E. P. Box*

5.1 Introduction

The Hidden Markov Model (HMM) based changepoint methods considered in this thesis have assumed that the number of underlying states, H, is known a priori to analysis. This assumption is made for the majority of the modelling and inference methods regarding HMMs such as estimating the underlying state sequence (Viterbi, 1967) and parameter estimation (Baum et al., 1970). However, this is often not the case when presented with time series data, where the number of underlying states is unknown.

Assuming a particular number of underlying states without performing any statistical justification can sometimes be advantageous if the states correspond directly to a particular phenomena. For example in the Econometric GNP analysis (Hamilton, 1989) considered throughout this thesis, two states are often assumed a priori, "Contraction" and "Expansion", due to the interest in recessions which are defined as two consecutive "contraction" states in the underlying state sequence (Shiskin, 1974). Without such an assumption, this definition of a recession and the conclusions we can draw from the resulting analysis may be lost.

However, it may be necessary to assess whether such an assumption on the number of underlying states is adequate, and typically, we are presented with time series data for which we are uncertain about the appropriate number of states to assume. This chapter concerns model selection for HMMs when H is unknown. Throughout this chapter, we use "model" and the "number of states in a HMM" interchangeably to denote the same statistical object.

Several methods for determining the number of states of a HMM currently exist; Section 5.3 provides a review of some of these methods. Bayesian methods appear to dominate the model selection problem of interest, and quantify more explicitly the model uncertainty by providing approximations of the model posterior distribution. These approximations are often obtained by jointly sampling the parameter and underlying state sequence, and marginalising as necessary to obtain the desired distribution. However, sampling the underlying state sequence can be particularly difficult due to its high dimension and correlation, and reduces statistical efficiency if the state sequence is not of interest. Alternative sampling techniques may thus be more suitable if they can avoid having to sample the state sequence.

This chapter proposes approximating the model posterior via the use of parallel Sequential Monte Carlo (SMC) samplers, where each SMC sampler approximates the marginal likelihood and parameter posterior conditioned on the number of states as previously considered in Chapter 3. These approximations are combined to approximate the model posterior of interest. One major advantage of the proposed methodology is that the underlying state sequence is not sampled and thus less complex sampling designs can be considered. We demonstrate in this chapter that the simple yet effective SMC sampler approach works well even with simple, generic sampling strategies which do not require application specific tuning. In addition, if we are already required to approximate numerous model parameter posteriors conditioned on several different number of states (as would be the case for sensitivity analysis, for example), the framework requires no additional computational effort and leads to parameter estimates with smaller standard errors than competing methods.

The structure of this chapter is as follows: Section 5.2 fixes terminology and notation used within the HMM literature and within this chapter. Section 5.3 provides a brief review of existing model selection methods concerning HMMs. Section 5.4 outlines the proposed method. Section 5.5 applies the proposed methodology to both simulated data and the Econometric GNP example considered throughout this thesis. Section 5.6 concludes the chapter.

5.2 Background

Let us consider the HMM setup and notation as defined in Sections 2.12 and 3.2 (page 35 and 58 respectively). We are still interested in general finite state HMMs, however we now make it explicit that these models are conditioned on H underlying states being present. Without loss of generality, we assume $\Omega_X = \{1, \dots, H\}, H < \infty$ with H being known a priori before inference is performed. Consequently, the emission and transition equations are rewritten as follows.

$$\begin{aligned} y_t|y_{1:t-1}, x_{1:t} &\sim f(y_t|x_{t-r:t}, y_{1:t-1}, \theta, H) && \text{(Emission)} \\ p(x_t|x_{1:t-1}, y_{1:t-1}, \theta, H) &= p(x_t|x_{t-1}, \theta, H) \quad t = 1, \dots, n && \text{(Transition)}. \quad (5.1) \end{aligned}$$

The use of HMMs allows us to compute exactly the likelihood $l(\theta|y_{1:n}H)$, via the use of the Forward-Backward algorithm (Baum et al., 1970), such that the underlying state sequence is accounted for exactly and does not need to be sampled.

In dealing with unknown θ, the model parameters of the assumed HMM, a maximum likelihood point estimate can be obtained via the Expectation-Maximisation algorithm (Baum et al., 1970) or a Bayesian approach can be employed which considers the model parameter posterior conditioned on there being H states, $p(\theta|y_{1:n}, H)$. This is typically a complex distribution which cannot be sampled from directly, with numerical approximations such as Monte Carlo methods being utilised. Approaches include MCMC (see for example Scott (2002) and Chib (1998)) and Sequential Monte Carlo algorithms such as Sequential Monte Carlo samplers (SMC, Del Moral et al. (2006)) as demonstrated in Chapter 3. We redirect the reader to Section 3.2.2 (page 67) for the relative merits of these two approaches. In addition to sampling from a sequence of connected distributions $\{\pi_b\}_{b=1}^{B}$ via SMC samplers, the sequence of normalising constants, $\{Z_b\}_{b=1}^{B}$ associated with these distributions can also be approximated in a natural way.

5.3 Literature Review

Standard model selection approaches via maximum likelihood and Akaike's and Bayesian Information Criteria are not suitable for HMMs as it is always possible to optimise these criteria via the introduction of additional states. In addition, information criteria methods have not been theoretically justified in the context of HMMs (Titterington, 1984). We begin by reviewing Mackay (2002), a frequentist information theoretic approach, before proceeding to Bayesian methods (Scott, 2002; Robert et al., 2000; Chopin, 2007) which dominate the literature. However, such

Bayesian approaches often require sampling the underlying state sequence which may not be necessary for the problems of interest, and difficult to perform due to their high dimensionality and inherent correlation. A more efficient Bayesian methodology which can avoid sampling the additional nuisance state sequence is thus of interest.

5.3.1 An Information Theoretic Approach

In light of the limitations of model selection via maximum likelihood and information criteria, Mackay (2002) proposes an information theoretic approach which yields a consistent estimate of the number of underlying states in addition to the model parameters. This is achieved by minimising the penalised distance function. For a k-dimensional process consisting of $(Y_t, \ldots, Y_{t+k-1})$ for $t = 1, \ldots, n-k+1$, the penalised distance for the distribution of the process $\{Y_t\}_{t=1}^{n-k+1}$ is defined as,

$$D(\bar{F}_n^k, F^k) = d_1(\bar{F}_n^k, F^k) - c_n \sum_{i=1}^{H} \log P(X_t = i), \tag{5.2}$$

where $\bar{F}_n^k$ is the k-dimensional empirical distribution function,

$$\bar{F}_n^k(y_1, \ldots, y_k) = \frac{\sum_{t=1}^{n-k+1} \mathbf{1}(Y_t \leq y_1, \ldots, Y_{t+k-1} \leq y_k)}{n-k+1}, \tag{5.3}$$

$P(X_t = i)$ is the stationary distribution of $\{X_t\}$, and c_n is a sequence of positive constant such that $c_n \to 0$. k is chosen based on identifiability conditions as discussed in Mackay (2002), and is also chosen to be as small as possible to minimise the computation burden of the methodology. Mackay (2002) utilises $k = 2H^{\max}$ where $H^{\max}$ is an upper bound of the number of underlying states. d_1 is assumed to be the Kolmogorov-Smirnov distance, where for distribution functions F_1 and F_2,

$$d_1(F_1, F_2) = \sup_y |F_1(y) - F_2(y)|.$$

The intuition of Equation 5.2, is that as $H \to \infty$, that is there are more underlying states, then there are more invariant probabilities $P(X_t = i)$ which are closer to zero, and consequently $\sum_{i=1}^{H} \log P(X_t = i)$ tends to minus infinity. As a result, the distance between the two distributions is penalised more for the introduction of these unnecessary states.

By minimising the penalised distance function presented in Equation 5.2, Mackay (2002) shows that consistent estimates of the number of underlying states and the model parameters can be obtained simultaneously, under mild conditions regarding the HMM (see (Mackay, 2002, Section 2) for further details).

This frequentist approach appears to work well, although Mackay (2002) highlights that global minimisation of Equation 5.2 can never be guaranteed which is critical to the parameter estimates. More importantly, the choice of penalisation constants c_n is highly influential on the estimate of H, with tuning on these abstract parameters being required. The mild conditions required for the consistent results are generally applicable for simple HMMs (those in which the emission density is only dependent on the underlying MC and not previous observations), although may not be satisfied for the general finite state HMMs also of interest in this thesis. In general, the uncertainty regarding the number of states is implicit, relying on the use of asymptotic arguments in obtaining the consistency results which may not be appropriate for time series of short length.

5.3.2 Parallel Markov Chain Monte Carlo

Scott (2002) proposes the use of a parallel Gibbs sampler in approximating the model posterior, stemming from the fact that the parameter posterior, conditional on the number of states, can be approximated via a Gibbs sampler. The parallel nature of the methodology arises from the fact that the approximation of each conditional parameter posterior can be performed independently of each other, and thus in parallel.

The methodology assumes that the number of underlying states is from a finite set, that is $H \in \{1, \ldots, H^{\max}\}$. Scott (2002) remarks that the use of the upper bound $H^{\max}$, is a mild restriction. $H^{\max}$ can be set to n, the length of the time series considered, such that each observation has its own state. However, in combination with an uninformative uniform prior over $1, \ldots, H^{\max}$, this does not lead to a parsimonious statistical model, and leads to estimation instabilities. Consequently, $H^{\max} \ll n$ is a recommended choice.

Let $\underline{\theta} = (\theta_1, \ldots, \theta_{H^{\max}})$ where θ_H denotes the model parameter θ associated with the HMM assuming H states. Consequently, $p(y_{1:n}|\underline{\theta}, H) \equiv p(y_{1:n}|\theta_H, H)$. It is assumed that $\theta_1, \ldots, \theta_{H^{\max}}$ are conditionally independent given H, and consequently,

$$p(\underline{\theta}, H) = p(H) \prod_{H=1}^{H^{\max}} p(\theta_H), \tag{5.4}$$

where $p(H)$ is a model prior. This property also transfers to the posterior distribution in that $\theta_1, \ldots, \theta_{H^{\max}}$ are conditionally independent given $y_{1:n}$ and H. That is,

$$p(\underline{\theta}|y_{1:n}) = \prod_{H=1}^{H^{\max}} p(\theta_H|y_{1:n}, H). \tag{5.5}$$

Each conditional parameter posterior $p(\theta_H|y_{1:n}, H)$ can be sampled from via an individual Gibbs sampler, and due to the conditional independence, these samplers can be performed in parallel. These conditional samples can then be collated to form a sample of $\underline{\theta}$. Sampling from each posterior $p(\theta_H|y_{1:n}, H)$ requires sampling the underlying state sequence $x_{1:n}$, in addition to the model parameter θ_H. We refer the reader to Scott (2002) for further details regarding the Gibbs sampler procedure.

A Monte Carlo approximation of $p(H|y_{1:n})$ is thus obtained as follows:

$$\begin{align}
p(H|y_{1:n}) &= \int p(H|y_{1:n}, \underline{\theta})p(\underline{\theta}|y_{1:n})d\theta \tag{5.6}\\
&= \mathbb{E}_{p(\underline{\theta}|y_{1:n})}[p(H|y_{1:n}, \underline{\theta})] \tag{5.7}\\
&\approx \frac{1}{N}\sum_{i=1}^{N} p(H|\underline{\theta}^i, y_{1:n}) = \frac{1}{N}\sum_{i=1}^{N} \frac{p(y_{1:n}|H, \theta_H^i)p(H)}{\sum_{h=1}^{H^{\max}} p(y_{1:n}|H=h, \theta_h^i)p(H=h)} \tag{5.8}
\end{align}$$

where $\{\underline{\theta}^i = (\theta_1^i, \ldots, \theta_{H^{\max}}^i)\}_{i=1}^N$, are N samples from $p(\underline{\theta}|y_{1:n})$ obtained via the parallel Gibbs sampler framework outlined above.

Scott (2002) discusses that computing $p(H|y_{1:n})$ via MCMC is an improvement over the BIC which provides an asymptotic approximation to $\log p(H|y_{1:n})$, and assumes a uniform prior over the models in this approximation. As the Monte Carlo approximation provides a more explicit approximation compared to asymptotic approximations which may not be satisfied in short time series, this is one advantage of Bayesian methods. Explicit approximation thus avoids over penalisation associated with BIC in small sample cases and greater control over the model prior. However, as with any other MCMC algorithm, this methodology also requires careful sampling designs such that we can ensure that the sampling MC is mixing well in the sample space and certainty that convergence has been reached for the sampling MC.

5.3.3 Reversible Jump Markov Chain Monte Carlo

A reversible jump Markov chain Monte Carlo (RJ-MCMC, Green (1995)) framework seems natural for such a model selection problem where the parameter space varies in dimension. RJ-MCMC are extensively used in model selection for mixture distributions (see Frühwirth-Schnatter (2005) and references therein), and in turn, they can also be applied to HMMs (Robert et al., 2000) where the sample space varies in dimension with respect to the number of underlying states assumed. This is an example of Variable-Dimension Monte Carlo as discussed in Scott (2002).

A fundamental concept of the RJ-MCMC is that the sampling MC needs to be able to move between the spaces of varying dimension, in addition to the moves within spaces of the same dimension size. This mechanism for moving between spaces is achieved by one-to-one transitions, for example by split and combine moves, where a state is broken into two states, and two states are merged to form a single state respectively. Consequently, this forms one of the steps of the MCMC sampling algorithm for this methodology.

Robert et al. (2000), present the RJ-MCMC based methodology for Gaussian Markov Mixture models such that only the variances are state dependent and the means are zero. However, they remark that the methodology is applicable for other distributions. The objective of the methodology is thus to sample from the conditional joint posterior density

$$p(\xi, H, \mathbf{P}, x_{1:n}, \sigma, |y_{1:n}) = p(\xi|y_{1:n})p(H|y_{1:n})p(\mathbf{P}|H, \delta)p(\sigma|H, \xi),$$

where ξ is a hyperparameter for standard deviations σ_h, such that $\sigma_h \sim \text{Unif}(0, \xi)$. Similar to Scott (2002), H the number of underlying states, is assumed to be from the finite set $\{1, \ldots, H^{\max}\}$. $\mathbf{P}$ denotes the $H \times H$ transition matrix, where δ denotes the associated hyperparameter assuming a Dirichlet prior distribution. A single iteration of the MCMC algorithm is as follows:

(a) Update the transition probability matrix $\mathbf{P}$.

(b) Update the standard deviations $\sigma = (\sigma_1, \ldots, \sigma_H)$.

(c) Update the underlying state sequence $x_{1:n}$.

(d) Update the hyperparameter ξ.

(e) Either split an existing state into two, or merge two states into one.

(f) Consider the birth or death of an empty state, a state in which no observations have been allocated to it.

We refer the reader to Robert et al. (2000) for specific details of the algorithm. (a)-(d) are performed via a Gibbs move, whilst (e) and (f) are complex Metropolis-Hastings steps which allow for the number of underlying states to increase or decrease by one. The split-combine move works by splitting a randomly selected single state with probability b_h, and combining a randomly selected adjacent pair of states with probability $d_h = 1 - b_h$. Under the conditions presented, $d_1 = b_{H^{\max}} = 0$ naturally, and Robert et al. (2000) implement $b_h = d_h = 0.5$ for $h = 2, 3, \ldots, H^{\max}-1$. As part of the split-combine move, all parameters and the state sequence, are modified accordingly in a systematic manner (and random for the case of splitting case). An acceptance probability is then computed to determine whether the move is accepted or rejected or not, and in turn whether it is beneficial in modelling the observed time series.

For the birth-death move, a birth and death action are selected at random with probabilities b_k and d_k respectively. As the death action is only performed on empty states, this action involves deleting the corresponding parameters associated with the deleted state and re-normalising the transition matrix. The underlying state sequence remains unchanged. The birth action involves creating a new empty state which is not associated with any of the existing states in comparison to a split move, with the associated parameters being drawn from the prior distributions. An acceptance probability is also computed for the birth-death move in determining whether such a move accepted or not.

An approximation of the model posterior of interest, the model posterior $p(H|y_{1:n})$, can then be obtained by marginalisation. However, RJ-MCMC is often computationally intensive and care is required in designing moves such that the sampling MC mixes well both within model spaces (same number of states, different parameters) and amongst model spaces (different number of states). These disadvantages mainly arise from the moves between model spaces which often lead to lower acceptance rates, and the need to sample the underlying latent state sequence, a high dimensional latent vector exhibiting correlation. In addition, RJ-MCMC is typically more unstable compared to standard MCMC algorithms, and it is typically more difficult to assess whether convergence has been reached (see Fearnhead (2006) for example).

5.3.4 Sequential Hidden Markov Model

Chopin (2007) proposes a model selection methodology which reformulates the HMM framework such that SMC can be used in approximating the model posterior by re-expressing the problem as a filtering problem. The reformulation of the HMM framework replaces the underlying Markov chain with an augmented hidden Markov chain $\tilde{X}_t = (h_t, x_t)$, where x_t represents the current state of the underlying MC as before, and h_t is a new variable which denotes the number of unique states that have appeared up to time t. The augmented Markov chain has the following construction, for $i, j \in \Omega_X$:

$$\begin{aligned} \tilde{X}_1 &= (h_1, x_1) = (1, 1) \\ p(x_{t+1} = j | x_t = i, h_t = h) &= \begin{cases} p_{ij} & \text{if } i, j \leq h \leq H; \\ \sum_{j=i+1}^{H} p_{ij} & \text{if } i \leq j = h + 1 \leq H; \\ 0 & \text{otherwise.} \end{cases} \\ h_{t+1} &= \max(h_t, x_{t+1}) \end{aligned} \tag{5.9}$$

The state dependent emission density remains unchanged, being dependent only on x_t at time t as in the standard HMM. Equation 5.9 can be seen as sequentially relabelling the states with respect to the order in which they appear: at time t with $\tilde{X}_t = (h_t, x_t) = (h, x)$, the chain can either return to a previously visited state l for $l \leq h$ with probability p_{xl}, or alternatively jump to a new state with probability $\sum_{l=h+1}^{H} p_{xl}$. In the latter case, $h_{t+1} = h + 1$ to denote that a new unique state has been visited; otherwise in the former case $h_{t+1} = h$ to represent that no new state has been visited. The transition matrix associated with this new HMM formulation is denoted by $\tilde{\mathbf{P}}$. Chopin (2007) show that this new reformulation is equivalent to the original HMM as presented in Equation 5.1 and alleviates the problem of state identifiability as the states are labelled as observed in the data sequence. It is this sequential reformulation of the HMM framework that gives rise to the name of the approach, Sequential HMM. Under this method, the transition probabilities and emission parameters to be estimated are the same as those in Equation 5.1 (the original HMM), but model selection inference is performed via the augmented HMM highlighted in Equation 5.9 by inferring on the $\tilde{X}_t = (H_t, X_t)$. This latter point is one of the key differences between the SHMM method and those reviewed thus far.

Having reformulated the HMM framework as follows, SMC can then be employed in estimating filtering probabilities such as $p(\tilde{x}_t | y_{1:n}, H)$, and the model posterior $p(H | y_{1:n})$. An outline of the SMC based algorithm is as follows:

Step 1 **Initialisation:** For $b = 1$, draw N independent samples of (H, θ) from the prior $p(H, \theta)$. This is performed by assuming the following prior structure,

$$p(H, \theta) = p(H) p(\theta | H) \tag{5.10}$$

$$p(H) \sim \text{Unif}(\{1, \ldots, H^{\max}\}) \tag{5.11}$$

$$p(\theta | H) = \prod_{h=1}^{H} p(\eta_h) \prod_{h=1}^{H} \text{Dirichlet}((p_{h1}, \ldots, p_{hH}) | \alpha_h), \tag{5.12}$$

where η_h are the state-dependent emission parameters, and α_h are hyperparameters associated with the transition probabilities $(p_{h1}, \dots, p_{hH})$.

Let $\{H^i, \theta^i\}_{i=1}^N$ denote the N samples drawn independently from the prior. Set the associated importance weights to $W_1^i = \frac{1}{N}$ for all samples $i = 1, \dots, N$.

Step 2 **Reweight:** Set $b = b+1$, compute the new importance weights for iteration b.

$$W_b^i = \frac{W_{b-1}^i p(y_b|y_{1:b-1}, \theta^i)}{\sum_{j=1}^N W_{b-1}^j p(y_b|y_{1:b-1}, \theta^j)} \qquad \text{for all } i = 1, \dots, N. \tag{5.13}$$

Step 3 **Resample-Move:** If $ESS = \frac{1}{\sum_{i=1}^N W_b^i} < T = \frac{N}{2}$, then resample particles $\{\theta^i\}_{i=1}^N$ according to their weights $\{W_b^i\}$. Let $\{\hat{\theta}^i\}_{i=1}^N$ denote the resampled particles, and reweight with weights $W_b^i = \frac{1}{N}$.

Mutate resampled particles with respect to some Markov kernel $K_b(\cdot, \cdot)$ with invariant distribution $p(\theta|y_{1:b})$. That is

$$\theta^i \sim K_b(\hat{\theta}^i, \cdot) \qquad i = 1, \dots, N. \tag{5.14}$$

A suitable choice of $K_b(\cdot, \cdot)$ is the Gibbs sampler for the Sequential HMM. This involves sampling iteratively from the full conditionals regarding the latent state sequence $\tilde{X}_{1:n}$, the transition matrix $\tilde{\mathbf{P}}$, and the emission parameters $(\eta_1, \dots, \eta_H)$.

Step 4 **Positive Discrimination:** Compute

$$p(H|y_{1:b}) \approx \hat{p}_{H,b} = \sum_{i:H^i=H} W_b^i. \tag{5.15}$$

For each $H \in \{1, \dots, H^{\max}\}$ such that $\hat{p}_{H,b} < \rho$, where $\rho \in (0,1)$ say $\rho = 0.1$, resample ρN particles from the sub-population of particles corresponding to model H. Reweight these resampled particles with importance weights $\frac{\hat{p}_{H,b}}{\rho}$. To retain N samples throughout the duration of the algorithm, resample $N - \kappa\rho N$ particles from the remaining samples present in the system, where κ denotes the number of models subjected to the positive discrimination mechanism outlined above. Reweight these completion resample particles with importance weights $\frac{1}{N}$.

Step 5 If $b < n$, then go to step 2. Else, terminate the algorithm.

The resample-move and positive discrimination steps are implemented in order to avoid the weight degeneracy issue common in all SMC algorithms. In particular, the positive discrimination mechanism avoids particles associated with larger values of H becoming extinct before there is enough data to be associated to H states.

Equation 5.13 provides the weights of the samples such that the weighted cloud of samples $\{\theta^i, W_b^i\}_{i=1}^N$ approximates the distribution $\tilde{\pi}_b = p(\theta|y_{1:b}, H)$. $p(y_b|y_{1:b-1}, \theta, H)$ denotes in particular, the likelihood of the observation y_b, for parameter value θ and given $y_{1:b-1}$. This is computed iteratively using the forward recursion of HMMs as outlined in the Appendix of Chopin (2007). Note that this sequence of distributions is different to that proposed in Chapter 3 and in this chapter, which is with respect to all observations being present and a tempering schedule on the likelihood. That is $\pi_b \propto l(\theta|y_{1:n}, H)^{\gamma_b} p(\theta|H)$, where $\{\gamma_b\}_{b=1}^B$ is some non-decreasing tempering scheme on the likelihood. Whilst the final distribution under both tempering schedules will both be the same parameter posterior $p(\theta|y_{1:n}, H)$, the data tempering scheme of Chopin (2007) naturally facilitates online estimation and applications.

Chopin (2007) remark that the SMC based algorithm compares favourably to MCMC based algorithms in terms of computational cost. However, they stress that one of the main advantages of such SMC based algorithms is that there is greater robustness compared to MCMC algorithms (one can be confident about the results if several different runs of the algorithm lead towards the same results and conclusions), and there is less concern about whether the algorithm has converged under the mixture setting of interest. However, this SMC based methodology also requires sampling the underlying state sequence which can often be difficult to perform since it is typically of high dimension and highly correlated.

5.4 Methodology

Similar to the approaches of Robert et al. (2000); Scott (2002); Chopin and Pelgrin (2004); Chopin (2007), we take a Bayesian model selection approach in determining H, the number of underlying states in a HMM. That is, we approximate $p(H|y_{1:n})$, the posterior over the number of underlying states for a given realisation of data $y_{1:n}$ (the model posterior). In addition, similar to these approaches, we assume a finite number of states, $H \in \{1, \ldots, H^{\max}\}$, where $H^{\max} \ll n$ in order to obtain stable estimates. Some methods, for example the Infinite HMM proposed in Beal et al. (2002), place no restriction on $H^{\max}$ via the use of a Dirichlet process based methodology. However, this also requires sampling the underlying state sequence via Gibbs samplers and requires approximating the likelihood via particle filters, neither of which is necessary under the proposed approach.

Via Bayes' Theorem, the model posterior of interest can be re-expressed as follows,

$$\begin{aligned} p(H|y_{1:n}) &\propto p(y_{1:n}|H)p(H) && (5.16) \\ &= \frac{p(y_{1:n}|H)p(H)}{\sum_{h=1}^{H^{\max}} p(y_{1:n}|H=h)p(H=h)} && (5.17) \end{aligned}$$

where $p(y_{1:n}|H)$ denotes the marginal likelihood for model H, and $p(H)$ denotes the model prior. We are thus able to approximate the model posterior if we are able to approximate the marginal likelihood associated with each model.

As Chapter 3 has demonstrated, SMC samplers can be used to approximate the conditional parameter posterior, $\pi_B \propto p(\theta|y_{1:n}, H)$, and its normalising constant Z_B. In contrast to Chapter 3, we now make explicit that these quantities are conditional on the number of underlying states assumed. Recall, that we can define the sequence of distributions $\{\pi_b\}_{b=1}^{B}$ as follows:

$$\pi_b(\theta) \propto l(\theta|y_{1:n}, H)^{\gamma_b} p(\theta|H), \qquad b = 1, \ldots, B \tag{5.18}$$

where conditioned on a specific model H, $p(\theta|H)$ is the prior of the model parameters and γ_b is a non-decreasing temperature schedule. We thus sample initially from $\pi_1(\theta) = p(\theta|H)$ either directly or via importance sampling, and introduce the effect of the likelihood $l(\theta|y_{1:n}, H)$, gradually. We in turn sample and approximate the target distribution, the parameter posterior $p(\theta|y_{1:n}, H)$. This does not require sampling the underlying state sequence as the likelihood and prior do not require this sampling. This latter point leads to Rao-Blackwellised estimates, a reduction in Monte Carlo variance.

Note that this setup is different to that proposed in Chopin and Pelgrin (2004) and Chopin (2007), where distributions are defined as $\tilde{\pi}_b = p(\theta, \tilde{x}_{1:b}|y_{1:b})$ with respect to incoming observations, and $\tilde{x}_b$ denotes the augmented hidden Markov Chain. A different tempering schedule is consequently employed due to the increasing data sequence over time.

Z_B, the normalising constant for the parameter posterior $p(\theta|y_{1:n}, H) = \frac{l(\theta|y_{1:n},H)p(\theta|H)}{Z_B}$, is more specifically of the following form,

$$Z_B = \int l(\theta|y_{1:n}, H)p(\theta|H)d\theta = \int p(y_{1:n}, \theta|H)d\theta = p(y_{1:n}|H). \tag{5.19}$$

That is, the normalising constant for the parameter posterior conditioned on model H, is the conditional marginal likelihood of interest required in Equation 5.17. Thus, given that we can approximate the marginal likelihood, we can thus approximate the model posterior as follows:

Algorithm outline:

1. For $h = 1, \ldots, H^{\max}$,
   (a) Approximate $p(y_{1:n}|H = h)$ and $p(\theta|y_{1:n}, H = h)$, the marginal likelihood (see Section 5.4.1) and parameter posterior (see Section 3.3.1, page 78) conditioned on h states, via SMC samplers.
2. Approximate $p(H = h|y_{1:n})$, the model posterior, via the approximation of $p(y_{1:n}|H = h)$ and model prior $p(H)$.

5.4.1 Approximating $p(y_{1:n}|H)$

In addition to sampling from a sequence of distributions $\pi_b, b = 1, \ldots, B$, SMC samplers can be used to approximate their respective normalising constants, Z_b. As presented in Section 3.2.2, SMC samplers work on the principle of providing weighted particle approximations of distributions through importance sampling and resampling techniques. For a comprehensive exposition of SMC samplers, we refer the reader to Del Moral et al. (2006).

The first part of Algorithm 7 (page 77) presents the formulation of SMC samplers within the HMMs framework. We now make it explicit however that the quantities obtained via SMC samplers are conditional on H underlying states being assumed. The main output of the SMC samplers algorithm is a series of weighted sample approximations of π_b, namely $\{\theta_b^i, W_b^i|H\}_{i=1}^N$, where N is the number of samples used in the SMC approximation. The approximation of the ratio between consecutive normalising constants can then be found as:

$$\frac{Z_b}{Z_{b-1}} \approx \widehat{\frac{Z_b}{Z_{b-1}}} = \sum_{i=1}^{N} W_{b-1}^i \widetilde{w}_b(\theta_{b-1}^i) := \overline{W}_b. \tag{5.20}$$

This ratio corresponds to the normalising constant for un-normalised weights at iteration b (that is the denominator in Equation 3.24 in Algorithm 7). Z_B, can thus be approximated as:

$$\widehat{Z_B} = \widehat{Z_1} \prod_{b=2}^{B} \overline{W}_b \tag{5.21}$$

which, remarkably, is an unbiased estimator of the true normalising constant (Del Moral, 2004). Equation 5.21 can also be more condensely expressed by only considering the ratio of normalising constants at resampling times. We refer the reader to Del Moral et al. (2006) for more details regarding this reduction in calculation, which may be more beneficial with respect to computational storage. Note that the normalising constant, Z_b, corresponds to the the following quantity

$$\pi_b(\theta) = \frac{\varphi_b(\theta)}{Z_b}, \tag{5.22}$$

where φ_b is the unnormalised density. We can thus approximate the marginal likelihood by simply recording the normalising constants for the weights, $\overline{W}_b$, at each iteration of our SMC algorithm.

As discussed in Chapter 3, there is a great deal of flexibility with the SMC implementation and some design decisions are necessarily dependent upon the model considered. We have found that a reasonably straightforward strategy works well for the class of HMMs which we consider without the need for application specific tuning. An example implementation, similar to that discussed in Nam et al. (2012b) and in Section 3.3.1 and 3.4.1 (page 78 and 82 respectively) is followed. This implementation consists of the use of a linear tempering schedule, asymmetric priors for the transition probability vectors, relatively flat priors for the emission parameters and Random Walk Metropolis Hastings proposal kernels. Details of specific implementation choices are given for representative examples in the following section.

5.5 Results

This section applies the proposed methodology to a variety of simulated and real data. All results have been obtained using the approach of Section 5.4 with the following settings. $N = 500$ samples and $B = 100$ iterations have been used to approximate the sequence of distributions. Additional sensitivity analysis has been performed with respect to larger values of N and B which we found further reduced the Monte Carlo variability of estimates, as would be expected, but for practical purposes samples of size 500 were sufficient to obtain good results. α_h is a H-long hyperparameter vector full of ones, except in the h-th position where a ten is present. This has been set arbitrary based on our belief and encourages the aforementioned persistent behaviour in the underlying MC typically associated with HMMs; other hyperparameters are of course available for other beliefs. The linear tempering schedule and proposal variances used have not been optimised to ensure optimal acceptance rates. Promising results are obtained with these simple default settings.

A uniform prior has been assumed over the model space in approximating the model posterior. We consider selecting the maximum a posterior (MAP) model, that is $\arg\max_{h=1,\ldots,H^{\max}} p(H = h|y_{1:n})$, which indicates the strongest evidence for the model favoured by the observed data.

5.5.1 Simulated Data

We consider simulated data generated by two different models; Gaussian Markov Mixture (GMM) and Hamilton's Markov Switching Autoregressive model of order r (HMS-AR(r), Hamilton (1989)). The first model has been chosen due to its relative simplicity and connection to Gaussian mixture distributions such that the Gaussian distributions have state dependent means and variances dependent on the underlying Markov Chain. In addition, the computer code for the SHMM method proposed by Chopin (2007) is available for such a model and thus comparisons can be made. The latter, as presented in Equation 3.34 on page 81, can be used to model Econometric GNP data (Hamilton, 1989) and brain imaging signals (Peng et al., 2011), as explored in Chapters 3 and 4 respectively. HMS-AR models induce dependency on previous observations via an autoregressive nature.

For various scenarios under these two models, Figures 5.1 and 5.3 present an example realisation of the data from the same seed (left column) and the model selection results from 50 simulations using different seeds (right column). Changes in state occur at times 151, 301 and 451. We consider a maximum of five states, $H^{\max} = 5$, as we believe that no more than five states are required to model the business cycle data example we will consider later, and the simulations are designed to reflect this.

The following priors have been implemented for the state dependent mean and precision (inverse of variance) parameters: $\mu_h \overset{\text{iid}}{\sim} \text{N}(0, 100)$, $\frac{1}{\sigma_h^2} \overset{\text{iid}}{\sim} \text{Gamma}(\text{shape} = 1, \text{scale} = 1), h = 1, \ldots, H$. For the HMS-AR model, we consider the partial autocorrelation coefficients (PAC, ψ_1) in place of AR parameter, ϕ_1, with the following prior, $\psi_1 \sim \text{Unif}(-1, 1)$. As discussed in Section 3.4.1 (page 82), the use of PAC allows us to maintain stationarity amongst the AR coefficients more efficiently (AR polynomial roots lying within the unit circle). Baseline proposal variances of 10 have been used for each parameters' mutation step which decrease linearly as a function of sampler iteration. For example, the proposal variance $\frac{\sigma_\mu^2}{b}$ is used for μ_h mutations.

Gaussian Markov Mixture

Figure 5.1 displays a variety of results generated by a GMM model under the proposed parallel SMC methodology. In addition, we compare our model selection results to the SHMM approach as proposed in Chopin (2007) where computer code for a GMM is freely available[1]. Recall that one of the key differences between the SHMM and proposed parallel SMC approach is that model selection inference is performed on the augmented MC via inference on $\tilde{X}_t = (H_t, X_t)$. The same transition probabilities and emission parameters are however being estimated. The following settings have been utilised for the SHMM implementation; $N = 5000$ samples have been used to approximate the sequence of distributions defined as $\tilde{\pi}_b = p(\theta, \tilde{x}_{1:b}|y_{1:b})$ with respect to the augmented MC, $H^{\max} = 5$ as the maximum number of states possible and one SMC replicate per dataset. The same prior settings under the proposed parallel SMC samplers have been implemented. Other default settings in the SHMM code such as model averaging being performed have been utilised. The model posterior approximations from this approach are displayed alongside the parallel SMC posterior approximations.

Figures 5.1(a) and 5.1(b) concern a simple two state scenario with changing mean and variance simultaneously. From the data realisation, it is evident that two or more states are appropriate in modelling such a time series. This is reflected in the model selection results with a two state model being significantly the most probable under the model posterior from all simulations, and always correctly selected under MAP. However, uncertainty in the number of appropriate states is reflected with some small probability assigned to a three state model amongst the simulations. These results indicate that the proposed methodology works well on a simple, well defined toy example. Results concur with the SHMM framework; a two state model is most probable for all simulations, and less model uncertainty is exhibited.

Figure 5.1(c) and 5.1(d) displays results from a similar three state model, where different means correspond to the different states with subtle changes in mean present, for example around the 151 time point. Such subtle changes are of interest because the GNP data also contains subtle changes in mean. The correct number of states is significantly the most probable under all simulations, and always correctly identified under MAP selection. In contrast under the SHMM approach, more variability is present amongst the simulations. A three state model is largely the most probable, although some approximations display a four or two state model also being the most probable. Such variability is reflected in the MAP selection with

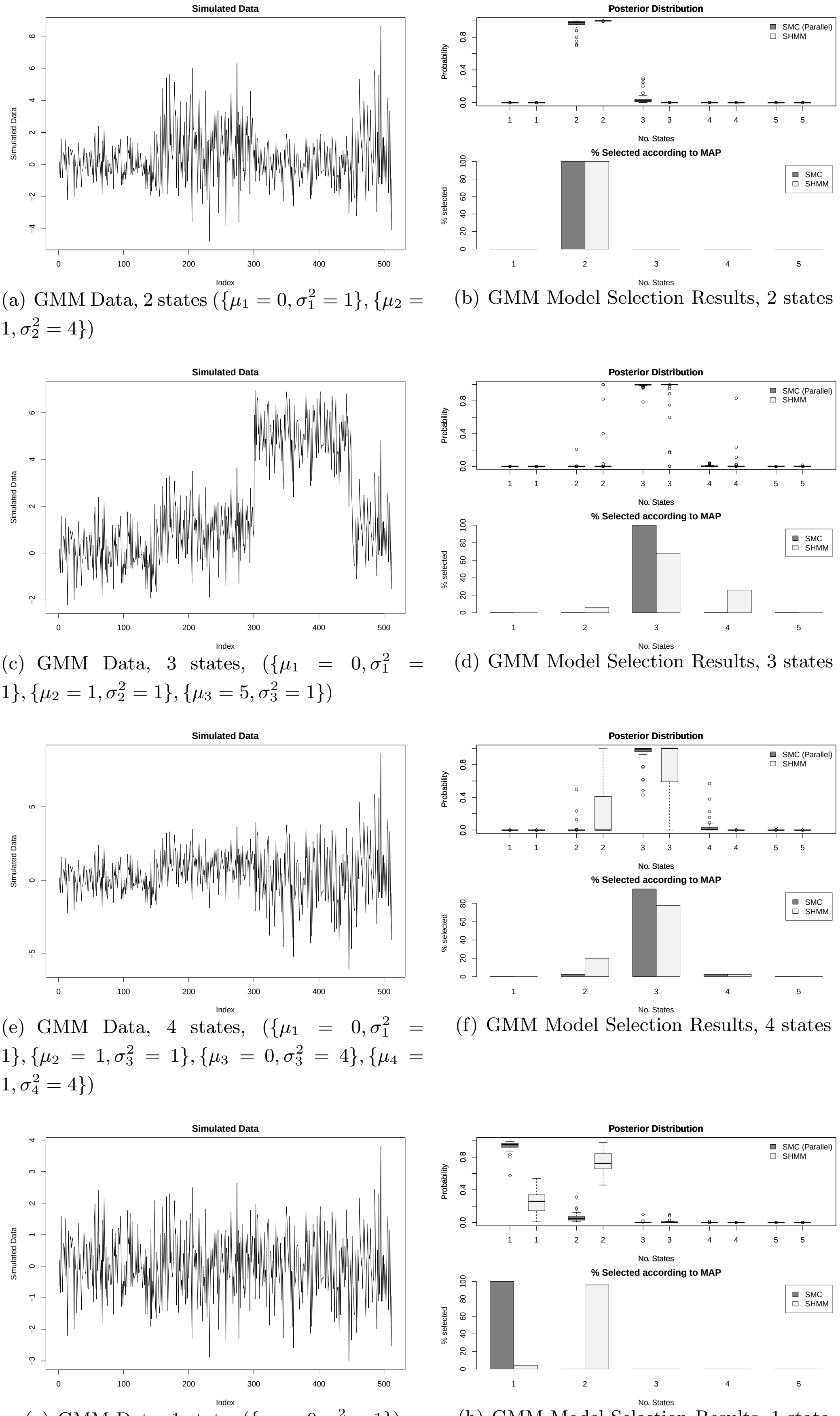
[1] `http://www.blackwellpublishing.com/rss/Volumes/Bv69p2.htm`

two and four states being selected in addition to the majority of three state models.

Figure 5.1(e) and 5.1(f) displays results from a challenging scenario of changes in both subtle mean and variance, independently of each other, with four states being present. The SMC methodology is unable to correctly identify the number of states, with three states being the most probable and most selected model from the majority of the simulations. However, given the example data realisation, it is not particularly surprising that such a result has been obtained; it is not entirely evident that the realisation is generated from a four or even three state model. This underfitting in the number of states is suspected to have occurred due to the segment of data from 451 to 512 being too short and associated with the previous segment of data from the previous state. In light of this challenging scenario, greater variability is present in the model posterior with significant probability assigned to two and four state models, in addition to the majority of probability assigned to a three state model. The SHMM also performs similarly, with significant probability being assigned to two state and three state models, and negligible probability assigned to four state models.

Figure 5.1(g) and 5.1(h) presents results from a one state GMM model, a stationary Gaussian process. The interest in this particular scenario is whether our methodology is able to avoid overfitting even though a true HMM is not present. The model selection results highlight that overfitting is successfully avoided with a one state model being most probable under the model posterior for all simulations and always the most selected under MAP. The SHMM method, in contrast, attaches substantial probability to a two state model over a one state model, and is nearly always selected under MAP. The successful avoidance of overfitting under the proposed SMC methodology compared to the SHMM methodology is another advantage of the presented methodology.

We also consider comparing the samples approximating the true emission parameters under the two methods. We consider the presented data scenarios of Figures 5.1(a) and 5.1(c) where the proposed SMC and SHMM method both concur with respect to the number of underlying states identified via MAP. In order to allow fair comparisons between the two approaches and the truth, the same labelling procedure has been utilised; namely by ordering the means ($\mu_1 < \mu_2$ for example). Table 5.1 displays the averaged posterior means and standard error for each emission parameter over the 50 simulations. We observe that the SMC methodology is more accurate in estimating the true value, and the standard error is smaller compared to the estimates provided by SHMM. This is as expected since the SHMM methodology requires sampling the underlying state sequence which ultimately induces additional



(a) GMM Data, 2 states ($\{\mu_1 = 0, \sigma_1^2 = 1\}, \{\mu_2 = 1, \sigma_2^2 = 4\}$)

(b) GMM Model Selection Results, 2 states

(c) GMM Data, 3 states, ($\{\mu_1 = 0, \sigma_1^2 = 1\}, \{\mu_2 = 1, \sigma_2^2 = 1\}, \{\mu_3 = 5, \sigma_3^2 = 1\}$)

(d) GMM Model Selection Results, 3 states

(e) GMM Data, 4 states, ($\{\mu_1 = 0, \sigma_1^2 = 1\}, \{\mu_2 = 1, \sigma_3^2 = 1\}, \{\mu_3 = 0, \sigma_3^2 = 4\}, \{\mu_4 = 1, \sigma_4^2 = 4\}$)

(f) GMM Model Selection Results, 4 states

(g) GMM Data, 1 state, ($\{\mu_1 = 0, \sigma_1^2 = 1\}$)

(h) GMM Model Selection Results, 1 state

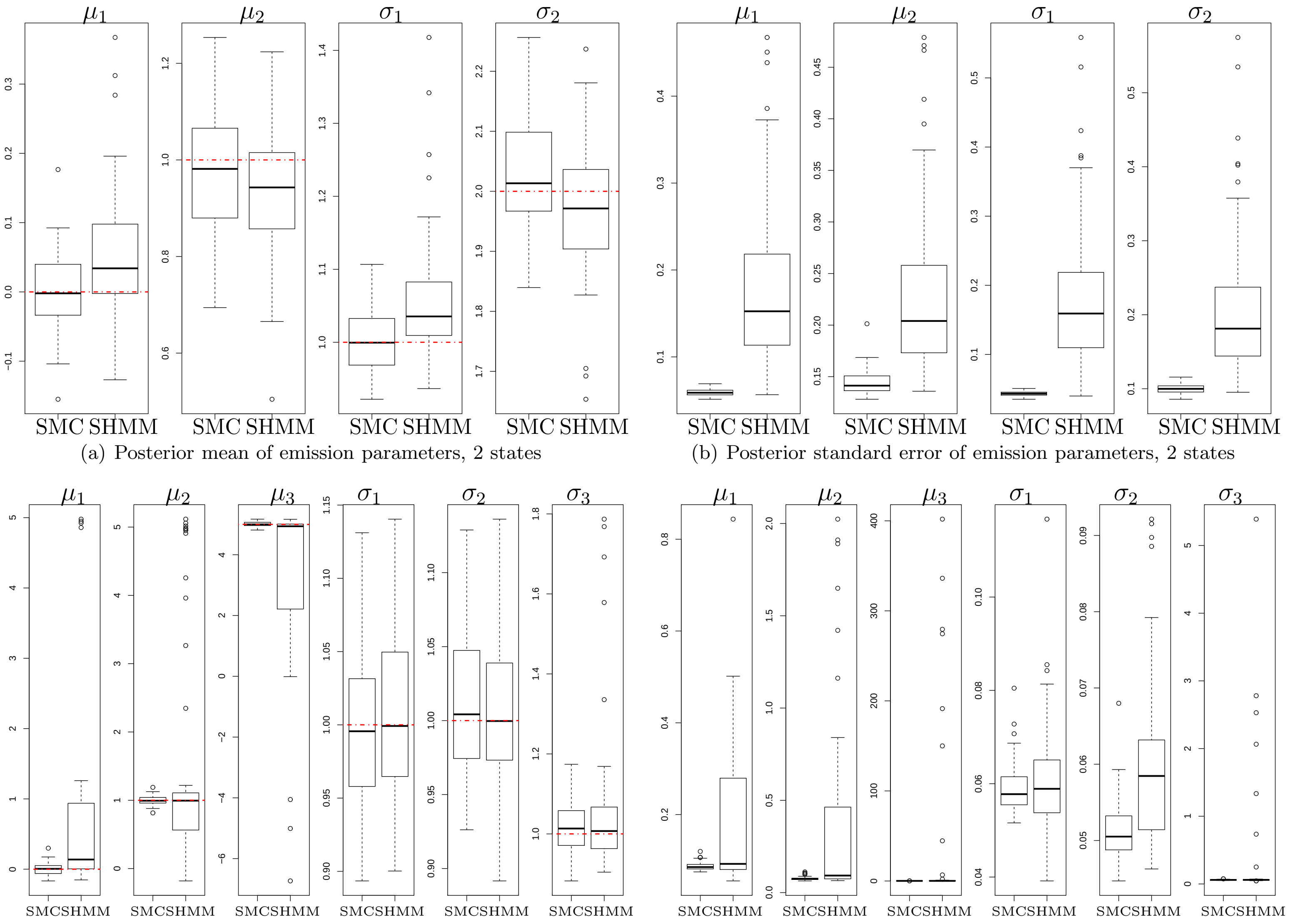
Figure 5.1: Model Selection Results for variety of Gaussian Markov Mixture Data. Left column shows examples of data realisations, right column shows the model selection results; boxplots of the model posterior approximations under the parallel SMC and SHMM approaches, and percentage selected according to MAP. Results are from 50 realisations.

	μ_1	μ_2	μ_3	σ_1	σ_2	σ_3
Truth	0	1	–	1	2	–
SMC	0.00 (0.06)	0.99 (0.14)	–	1.00 (0.04)	2.03 (0.10)	–
SHMM	0.05 (0.18)	0.94 (0.23)	–	1.06 (0.19)	1.97 (0.21)	–
Truth	0	1	5	1	1	1
SMC	0.00 (0.09)	1.00 (0.08)	5.00 (0.08)	1.00 (0.06)	1.01 (0.05)	1.01 (0.06)
SHMM	0.70 (0.19)	1.50 (0.38)	3.51 (33.86)	1.01 (0.06)	1.01 (0.06)	1.07 (0.35)

Table 5.1: Averaged posterior means and standard error for each emission parameter over the 50 simulations for the two data scenarios considered. We compare the proposed parallel SMC and SHMM method. Averaged standard errors are denoted in the parentheses. The same labelling procedure of the states has been utilised to allow valid comparisons between the two methods and the true parameter values. Results indicate that the SMC approach outperforms the SHMM method with greater accuracy in approximating the true values and smaller standard errors.

sampling error into the standard error of the estimates. This does not occur under the parallel SMC approach.

Figure 5.2 displays box plots of the posterior means (5.2(a) and 5.2(c)) and standard error (5.2(b) and 5.2(d)) of the emission parameter estimates for all 50 simulations. The posterior mean box plots indicate further that the proposed parallel SMC approach is generally more accurate and centered around the true emission parameter values (horizontal red dotted lines) across all simulations. The SHMM estimates are generally less precise with greater variability in the values present. Similarly, the standard error box plots indicate that the standard error is less under the proposed SMC methodology compared to the SHMM method, presumably due to the lower dimension of the sampling space resulting from the absence of the state sequence.



(a) Posterior mean of emission parameters, 2 states

(b) Posterior standard error of emission parameters, 2 states

(c) Posterior mean of emission parameters , 3 states

(d) Posterior standard error of emission parameters, 3 states

Figure 5.2: Boxplots of the posterior means and standard error for each emission parameter over 50 simulations. Red dotted values denote the value of the emission parameter used to generate the simulated data in the posterior mean box plots. We compare the results under the two approaches, parallel SMC and SHMM. We observe that the proposed SMC approach fares better with posterior means centered more accurately around the true values, and the standard error for the samples being smaller.

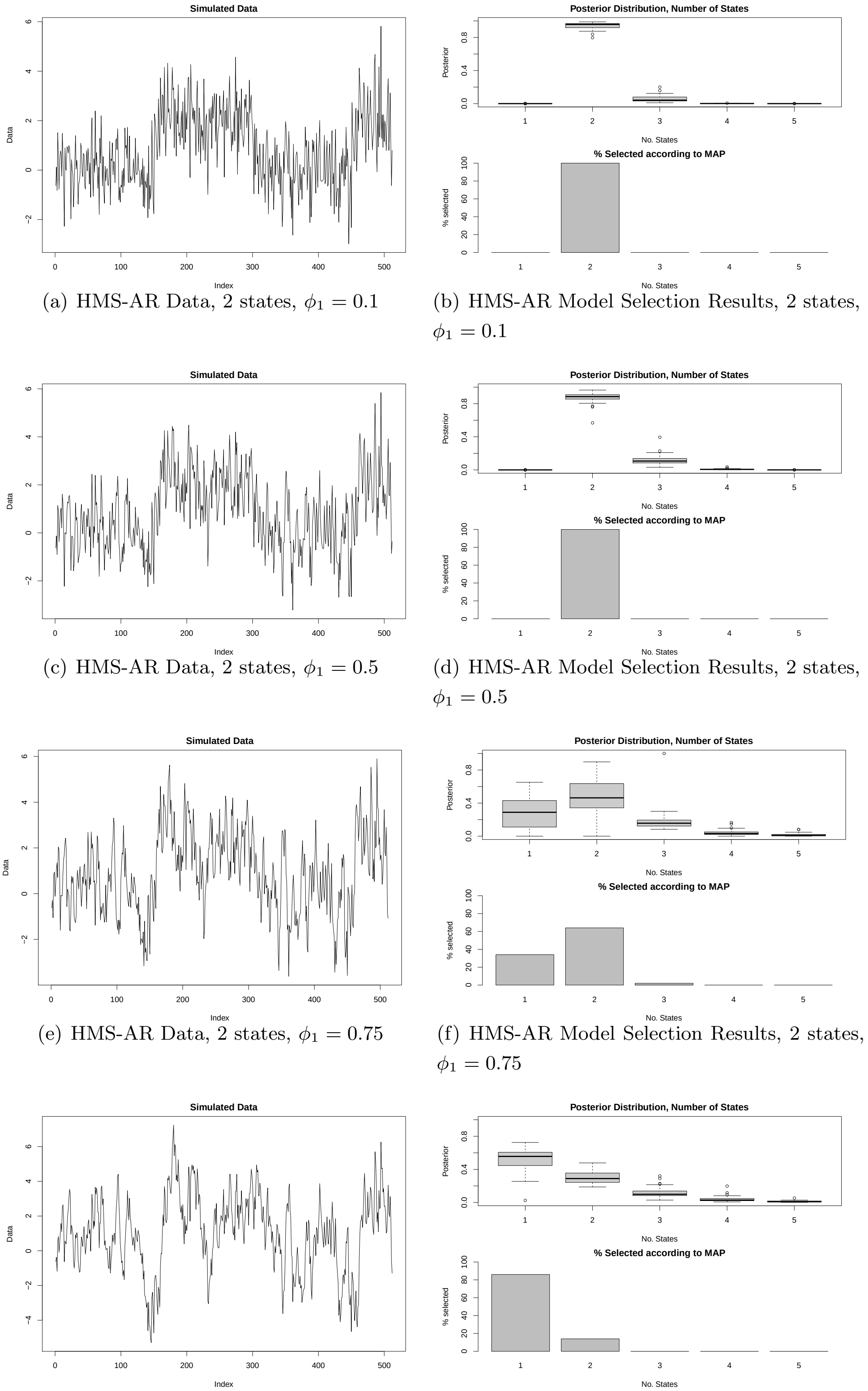
These additional results indicate that more accurate estimates are obtained under the proposed SMC approach, compared to the existing SHMM method, in addition to identifying the correct model more frequently. This is a result of the Rao-Blackwellised estimator provided by the SMC samplers framework and despite more samples being used under the SHMM approach. As fewer samples are required to achieve good, accurate estimates, the proposed parallel SMC method would appear to be more computationally efficient.

In addition, while not directly comparable, the runtime for the SMC samplers approach for one time series was approximately 15 minutes to consider the five possible model orders using $N = 500$ samples (implemented in R (R Development Core Team, 2011)), while it takes approximately 90 minutes for the SHMM approach with the default $N = 5000$ particles (implemented in MATLAB (MATLAB, 2012)).

Hamilton's Markov Switching Autoregressive Model of order r

Figure 5.3 shows results from a HMS-AR model with autoregressive order one; we assume that this autoregressive order is known prior to analysis although the proposed methodology could easily be extended to consider model selection with respect to higher AR orders. The following results were obtained using data generated using a two state model, with varying autoregressive parameter, ϕ_1, and the same means and variance used for each scenario ($\mu_1 = 0, \mu_2 = 2, \sigma^2 = 1$). Interest lies in how sensitive the model selection results are with respect to ϕ_1.

For small values of ϕ_1 (for example $\phi_1 = 0.1, 0.5$) indicating small dependency on previous observations, our proposed methodology works well with the correct number of true states being highly probable and always the most selected according to MAP. Relatively little variability exists in the approximation of the model posterior. However, as ϕ_1 begins to increase and tend towards the unit root, for example $\phi_1 = 0.9$, we observe that more uncertainty is introduced into the model selection results, with greater variability in the model posterior approximations and alternative models being selected according to MAP. However, as the data realisation in Figure 5.3(g) suggests, the original two state model is hard to identify by eye and thus our methodology simply reflects the associated model uncertainty. These results indicate that the proposed model selection method works for sophisticated models such as HMS-AR models, although the magnitude of the autoregressive nature evidently affect results.



(a) HMS-AR Data, 2 states, $\phi_1 = 0.1$

(b) HMS-AR Model Selection Results, 2 states, $\phi_1 = 0.1$

(c) HMS-AR Data, 2 states, $\phi_1 = 0.5$

(d) HMS-AR Model Selection Results, 2 states, $\phi_1 = 0.5$

(e) HMS-AR Data, 2 states, $\phi_1 = 0.75$

(f) HMS-AR Model Selection Results, 2 states, $\phi_1 = 0.75$

(g) HMS-AR Data, 2 states, $\phi_1 = 0.9$

(h) HMS-AR Model Selection Results, 2 states, $\phi_1 = 0.9$

Figure 5.3: Model Selection Results for variety of HMS-AR(1) data with $\mu_1 = 0, \mu_2 = 2, \sigma^2 = 1$ and varying ϕ_1. Left column shows examples of data realisations, right column shows the parallel SMC model selection results from 50 realisations; approximations of the model posterior, and percentage selected according to MAP.

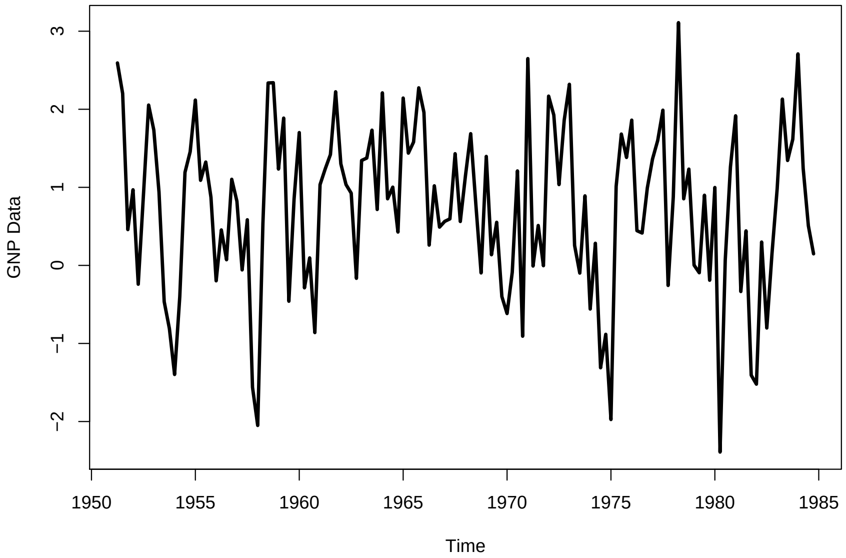
5.5.2 Hamilton's GNP data

We now return to the GNP dataset which has featured throughout this thesis. Recall that Hamilton's GNP data (Hamilton, 1989) consists of differenced quarterly logarithmic US GNP between 1951:II to 1984:IV. Following Hamilton (1989) and Aston et al. (2011), a two state HMS-AR(4) model was assumed in Chapter 3 to model y_t, the aforementioned transformed data, in order to analyse and identify the starts and ends of recessions. The two underlying states correspond to "Contraction" and "Expansion" states with respect to the typical definition of a recession; two consecutive quarters of contraction.

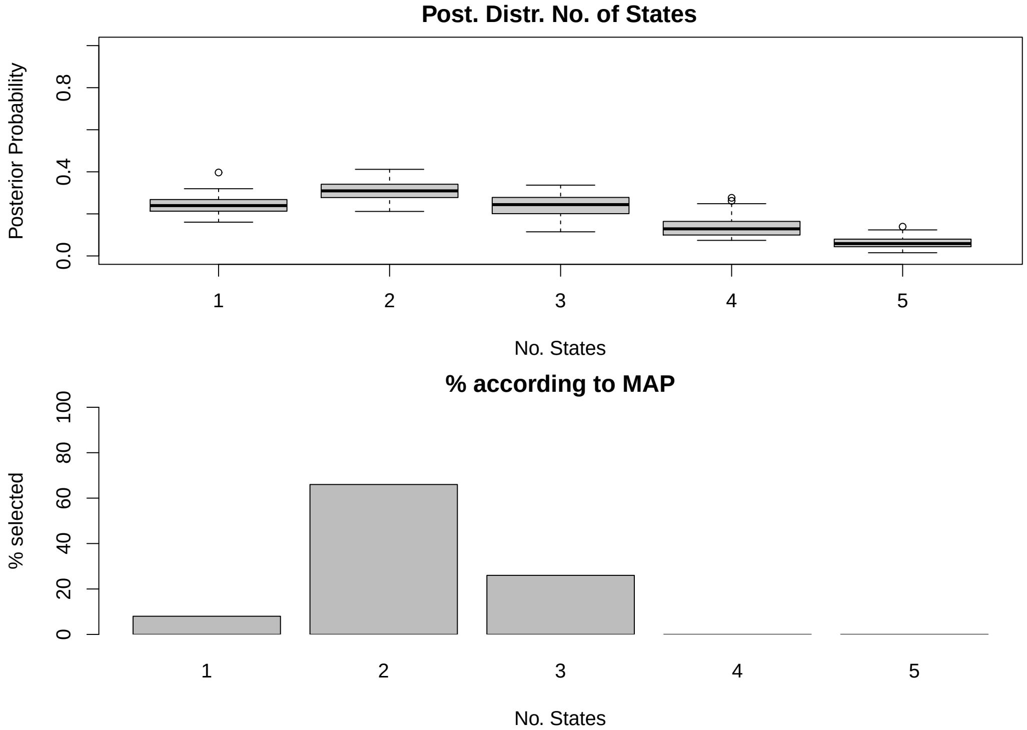
Whilst such a model works well in practice for recession inference, we investigate whether a two state HMS-AR(4) model is indeed appropriate. We assume the autoregressive order of four, is known a priori relating to some dependence on past observations within the year. We assume a maximum of five possible underlying states in the HMM framework ($H^{\max} = 5$) as we believe that the data arises from at most five possible states for the particular duration considered.

The following priors have been utilised: for the means, $\mu_h \overset{\text{iid}}{\sim} \text{N}(0, 10), h = 1, \ldots, H$, precision (inverse variance) $\frac{1}{\sigma^2} \sim \text{Gamma}(\text{shape} = 1, \text{scale} = 1)$, PAC coefficients $\psi_j \overset{\text{iid}}{\sim} \text{Unif}(-1, 1), j = 1, \ldots, 4$. A uniform prior has been used over the number of states H. As in the simulated data analysis, the baseline proposal variance is 10 which diminishes linearly with each iteration of the sampler.

Figure 5.4 displays the corresponding dataset and model selection results from fifty different SMC replicates. The model selection results, Figure 5.4(b), demonstrate that there is uncertainty in the appropriate number of underlying states with non-negligible probability assigned to each model considered and variability amongst the SMC replication results. Some of the alternative models seem plausible, for example a one-state model given the plot of the data and the additional underlying states modelling the subtle nuances and features in the data. However, a two state model is the most probable the majority of the time according to MAP. In addition, the distribution appears to tail off as we consider more states, thus indicating that the value of $H^{\max}$ used is appropriate. In conclusion, the two state HMS-AR(4) model assumed by Hamilton (1989) does seem adequate in modelling the data although this is not immediately evident and uncertainty is associated with the number of underlying states.



(a) Hamilton's GNP data: differenced quarterly logarithmic US GNP between 1951:II to 1984:IV.



(b) Model Selection Results for GNP data

Figure 5.4: Model selection results for Hamilton's GNP data under the proposed model selection methodology. 5.4(a) displays the analysed transformed GNP data. 5.4(b) displays the model posterior approximations from 50 SMC replications, and percentage selected under MAP.

Sensitivity Analysis

This section briefly performs sensitivity analysis with respect to different SMC conditions for our GNP model selection results. Figure 5.5 displays these sensitivity analysis results (boxplots of the posterior distributions from different seeds), under new conditions involving a different number of samples, and different hyperparameters associated with the prior. The latter set of conditions are of particular interest as Bayesian model selection results are known to be sensitive to choice of prior and hyperparameters in the literature (see Hoeting et al. (1999) for example). All other SMC settings as presented in Figure 5.4(b), remain unchanged (for example the number of distributions considered, the use of a linear tempering schedule).

The first panel on the left displays the model posterior with the use of 1000 samples as opposed to 500 samples in our approximations of the distributions. We observe that there is little change in the posterior compared to the original analysis as presented in Figure 5.4(b). Thus our model selection results remain fairly robust to further number of samples being utilised. We stress that this is conditional on other settings remaining unchanged, and for this particular dataset.

More noticeable changes in the model posterior arise when considering hyperparameters associated with more diffuse priors on the model parameters. Such changes include greater variability in the estimates of the model posterior probability, and ultimately the model conclusions drawn from the resultant distribution. The second panel considers a more diffuse prior associated with the transition probabilities, namely α_h, the H lengthed vector of mostly ones, contains a five in the h-th co-ordinate (previously set as ten in Section 5.5). Under such a setting, more probability is assigned to models with a larger number of underlying states, with the mode of the posterior shifting from two to three states. Such a change in posterior is suspected to be due to the underlying MC being less persistent in the same state under this prior choice, and thus the underlying MC is able to move between states more frequently. This latter remark consequently allows more subtle features of the GNP data to be modelled by the additional states.

The third panel considers a diffuse prior for the state-dependent means, namely $\mu_h \overset{\text{iid}}{\sim} \text{N}(0, 100), h = 1, \ldots, H$. We observe that the posterior becomes positively skewed with a parsimonious one-state model being the mode. This behaviour is suspected to have arisen as the prior is very diffuse such that it extends beyond the scope of the data, and due to the potential diversity of the initial sample, one state will capture the global behaviour of the time series, with additional underlying states failing to capture the finer nuances present unless in the correct region. Due to the variance associated with this prior, low probability is associated with this region.

The final and fourth panel considers sensitivity with respect to global precision by considering the prior $\lambda \sim \text{Gamma}(1, 2)$. We observe that the model distribution is slightly less peaked towards a two-state model, with some of this probability being assigned to a three-state model. However, the mode of the posterior still remains at two and the probabilities associated with the other models remaining the same as in original analysis. A potential explanation for more evidence towards a three-state model is that larger precision values are sampled under this prior configuration and there are more samples with smaller global variances. Consequently, more states are required to model that data appropriately, each with different means associated with them.

Our sensitivity results thus demonstrate that prior specification does influence our HMM model selection results which is no different to other Bayesian model selection methods (see Hoeting et al. (1999) for a good overview). Care must therefore be taken in prior specification as the model posterior, and the inference we perform on it, are very sensitive to such specification.

5.6 Conclusion and Discussion

This chapter has proposed a methodology in which the number of underlying states in a HMM framework, H, can be determined by the use of parallel Sequential Monte Carlo samplers. Conditioned on the number of states, the conditional marginal likelihood can be approximated in addition to the parameter posterior via SMC samplers. By conditioning on a different number of states and thus model, we can obtain several conditional marginal likelihoods. These conditional marginal likelihoods can then be combined with an appropriate prior to approximate the model posterior, $p(H|y_{1:n})$, of interest. The use of SMC samplers within a HMM framework results in an computationally efficient and flexible framework such that the underlying state sequence does not need to be sampled unnecessarily compared to other methods which reduces Monte Carlo error of parameter estimates, and complex design algorithms are not required. In comparison to MCMC based methodologies, our simple yet effective SMC based algorithm does not need to be assessed as to whether it has reached convergence, or any application specific tuning.

The proposed methodology has been demonstrated on a variety of simulated data and GNP data and shows good results, even in challenging scenarios where subtle changes in emission parameters are present. Results on the GNP data have

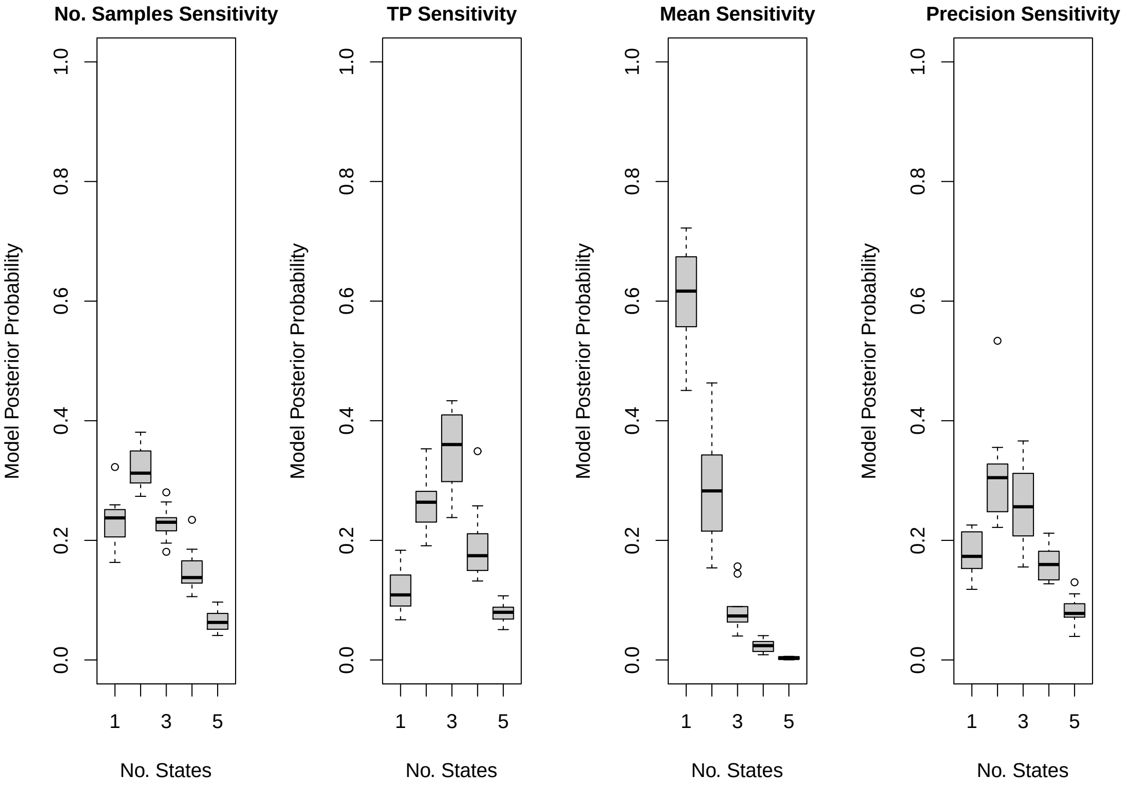


Figure 5.5: Boxplots of model posterior for the GNP data under different SMC settings. We consider the use 1000 samples in the approximation of distributions (first panel on the left), a transition probability vector prior of $p_h \overset{\text{iid}}{\sim} \text{Dirichlet}(\alpha_h), h = 1, \ldots, H$ where α_h is a H length vector of ones except for the hth element where a five is present (second panel), a mean prior of $\mu_h \overset{\text{iid}}{\sim} \text{N}(0, 100), h = 1, \ldots, H$ (third panel), and a precision prior of $\lambda \sim \text{Gamma}(1, 2)$ (fourth panel). Results indicate as with other Bayesian model selection methods, our posterior is sensitive to the choice of hyperparameter.

further confirmed that a two state HMS-AR model assumed in previous studies and analysis is appropriate, although the uncertainty associated with the number of underlying states has now been captured. In the settings considered, the method performs at least as well as other state of the art approaches in the literature such as the SHMM approach proposed in Chopin (2007).

From a modelling perspective, the model selection results presented in this thesis have assumed a uniform prior over the collection of models considered but there would be no difficulty associated with the use of more complex priors. Perhaps more important in the context of model selection is the specification of appropriate priors over model parameters, which can have a significant influence on model selection results, as demonstrated by the sensitivity analysis results presented in Figure 5.5. Such sensitivity on the prior is common to all Bayesian model selection methodologies and stresses the fact that some sensitivity analysis should always be conducted with respect to prior specification on model selection results. In determining suitable hyperparameters and reducing some sensitivity of the prior on the model posterior, it would be worth investigating an empirical Bayes approach or introducing a prior associated with the hyperparameters themselves (a hierarchical structure). Empirical Bayes', in which hyperparameters are estimated from the data itself, is a feasible, intuitive approach although the methodology no longer remains fully Bayesian. By implementing priors on the hyperparameter, this retains the overall Bayesian philosophy and should not provide any real difficulty with respect to implementation, although this does increase the computational cost.

From the perspective of computational efficiency and statistical estimation, it is desirable to identify a value of $H^{\max}$ which is sufficiently large to allow for good modelling of the data but not so large that the computational cost of evaluating all possible models becomes unmanageable (noting that the cost of dealing with any given model is an increasing function of the complexity of that model) and stable estimates with relatively small standard errors (achieved by guaranteeing that a reasonable number of observations are associated with the state and its parameters). Such aspects are associated with the length of data and the order of the HMM and thus need to be considered in determining $H^{\max}$.

We consider two areas of further research regarding the methodology presented in this chapter. Having proposed a method in which we are able to determine the number of unknown states, the next natural step is to embed this information within existing HMM methodologies, for example the changepoint methods presented in this thesis. The most straightforward approach is to determine an estimate of the number of states from the model posterior (for example, the maximum a posterior estimate $\widehat{H} = \arg\max_{H=1,\ldots,H^{\max}} p(H|y_{1:n}))$, and simply condition our existing HMM methodologies on this estimate. For example, in the case of the changepoint probability, we compute $P(\tau^{(k_{\text{CP}})} \ni t|y_{1:n}) = P(\tau^{(k_{\text{CP}})} \ni t|y_{1:n}, \widehat{H})$.

The second, more advantageous approach is to account for the model posterior in our applications and perform model averaging effectively. In the case of the changepoint probability, this would lead to the following formulation,

$$P(\tau^{(k_{\text{CP}})} \ni t|y_{1:n}) = \sum_{h=1}^{H^{\max}} P(\tau^{(k_{\text{CP}})} \ni t|y_{1:n}, H = h)P(H = h|y_{1:n}). \tag{5.23}$$

This latter approach is particularly attractive as our changepoint estimates are now able to account for a degree of model uncertainty, in addition to model parameter uncertainty. As our methodology can also be potentially used in the uncertainty of autoregressive orders, this could also be similarly accounted for in the changepoint estimates, for example in Chapter 4. Due to the parallelised nature of the SMC based methodology proposed in this chapter, this model averaging procedure could also be performed in an efficient manner.

This chapter has focused on a retrospective, offline context where all data is available prior to analysis. The second path of further research is to consider model selection in an online scenario where data is made available incrementally. Under such a scenario, the sequence of distributions would be defined as,

$$\pi'_b(\theta) \propto l(\theta|y_{1:b}, H)p(\theta|H) \qquad b = 1, \ldots, n \tag{5.24}$$

where $l(\theta|y_{1:b}, H)$ is the partial likelihood with respect to the incremental observations available up to time b. This is in the spirit of Sequential HMM (Chopin, 2007). The normalising constants for this new sequence of distribution are the partial marginal likelihoods $Z'_b = p(y_{1:b}|H)$. In turn, these can be used in approximating the partial model posterior $p(H|y_{1:b})$. By defining the sequence of distributions as above under the SMC samplers framework, this would provide an online approach which does not require sampling the underlying state sequence (leading to a reduction in sampling variance), and which retains many of the implementation procedures and benefits presented in this chapter.

Chapter 6

Quantifying the Uncertainty of Autocovariance Changepoints

> **Liz Lemon: But that cookie jar says "mom" on it.**
> **Jack Donaghy: Er, I don't think so. I've always viewed it as an upside down "wow".**
>
> *"The Collection", Episode 2.03, 30 Rock, Matt Hubbard*

6.1 Introduction

This thesis has thus far focused on changepoint (CP) methods regarding changes in mean and variance predominantly. This is a result of work in CP detection and estimation predominantly dedicated to changes in mean, trend (regression), variance, and combinations there of. However, non-stationarity can also arise from changing autocovariance structure and potentially exhibited in financial time series (Cho and Fryzlewicz, 2012) and oceanography (Killick, 2012) for example. However, there is comparatively little CP literature dedicated to such changes. In addition, those methods which do exist for such changes often provide different estimates and many fail to capture explicitly the uncertainty associated with these estimates. As argued and demonstrated in Chapter 3, there is a need to assess the plausibility of estimates provided by different autocovariance CP methods and this can be performed by quantifying the uncertainty associated with the estimates.

Certain changes in autocovariance can be modelled adequately by the general finite state Hidden Markov Model framework presented throughout this thesis, and thus the methodologies presented in Chapters 3 and 5 are still applicable. More specifically, it is possible to consider CPs arising from piecewise autoregressive (AR) processes by modelling the time series as a Markov Switching AR model where the AR coefficients are state dependent in addition to means and variances.

However, changes in autocovariance can also arise from piecewise moving average (MA) or piecewise generalised autoregressive conditionally heteroscedastic (GARCH) time series. However, the corresponding Markov Switching MA and Markov Switching GARCH models have posterior Markov chains which require the entire history of the underlying state sequence for analysis. Consequently, the methodologies presented in Chapters 3 and 5 cannot be applied since they assume finite dependence on previous states. We stress that to the best of our knowledge, exact CP methods for from piecewise MA and GARCH models do not exist with approximations being necessary (see for example Berkes et al. (2004); Tahmasbi and Rezaei (2008)).

One approach is to approximate the time series as a piecewise AR process regardless of how it may be generated. Under the HMM framework, the use of a Markov Switching AR model would induce the finite dependency necessary for analysis. Approximating by an AR process, forms the basis of the Automatic Piecewise Autoregressive Modelling procedure (AutoPARM, Davis et al. (2006)), which as detailed in Section 2.8 (page 25), models observed time series as piecewise AR processes with varying orders and AR coefficients to capture the changing autocovariance. Changepoints are identified via optimisation of the Minimum Description Length criteria (Rissanen, 1978) which provides the best segmentation configuration with respect to the CP locations, and the corresponding AR models for each segment. However the parametric assumption of piecewise AR processes is a strong assumption and may not always be appropriate, for example in the case of piecewise MA processes although it has been shown empirically to work well in some MA cases. Uncertainty is implicitly captured via asymptotic arguments in obtaining consistent estimates of the CP locations, conditional on the number of CPs being known and assuming a true piecewise AR structure, and thus not reported explicitly.

An alternative approach is to consider the time series in an alternative domain such as the frequency domain, and consider CPs in the associated periodograms of the time series. Periodograms are estimates of spectra which describe the autocovariance structure of a time series at different frequency bands. Representations in the frequency domain can be achieved in a variety of different manners, each with respect to different sets of basis functions, and consequently leading to different transformations. These basis functions include sinusoidal functions (Fourier transform, Condon (1937)), Smooth Localised complex Exponential functions (SLEX transform, Ombao et al. (2002)), and wavelets (wavelet transform, Daubechies (1990)).

In addition, the SLEX and wavelet basis functions possess a time-localisation property which permits spectra with time varying behaviour to be considered. Such basis functions are thus ideal in considering CPs and non-stationarity in time series. In comparison, the sinusoidal functions of the Fourier transform are global, and are thus inadequate in capturing non-stationarity in time series. The short time Fourier transform (Allen, 1977), is one possible solution in utilising the Fourier transform analysis in a non-stationary context. This transform considers analysing the time series in small time windows specified by the user, thus allowing for time varying spectral structure. However, this transform requires specifying a bandwidth and is still inadequate in modelling discontinuities.

Ombao et al. (2001) propose the Auto-SLEX method which is an automatic statistical procedure which simultaneously segments the time series (thus providing the optimal CP configuration), and provides the periodogram estimate of a time varying spectrum via the use of SLEX basis functions. Under a SLEX transformation, this provides a library of different orthonormal basis representations (corresponding to different CP configurations), each with different partitions at several frequency bands. A cost function for a particular CP configuration is computed which is the sum of the cost functions for each segment as a result of the assumed configuration. The optimal CP configuration is that which minimises the cost function. However, Auto-SLEX only considers partition configurations where segments have dyadic length (that is an integer power of two) which constrains the locations of CP estimates. As there is no reason to assume or guarantee that segments have dyadic length and that CPs occur at the constrained set of location estimates, such an approach does not seem adequate for the datasets of interest.

Choi et al. (2008) propose a sequential CP detection method for changes in autocorrelation structure which is rooted in considering the time series in the frequency domain. By performing a short time Fourier transform or wavelet transform, periodograms from consecutive windows of data are compared against each other with a similarity statistic being computed. It is this similarity statistic which is used to determine whether a CP has occurred by assessing over time whether this process drops below a specified threshold.

Alternatively, Cho and Fryzlewicz (2012) consider modelling time series under the Locally Stationary Wavelet (LSW) framework, where the building blocks of the time series are the localised wavelets at different frequencies and locations. Under the LSW framework, the Evolutionary Wavelet Spectrum (EWS) describes the autocovariance structure of a time series at different scales (frequency bands) and locations. Autocovariance CPs in the time series thus correspond to changes in the scale processes of the EWS and vice versa. Cho and Fryzlewicz (2012) analyse each scale process independently for CPs via a non-parametric test statistic (an extension of the circular binary segmentation algorithm), and then combine CP results from each scale to obtain a single set of results for the observed time series. This post-processing step is necessary as a CP in the time series can appear at several different scales of the periodogram, and thus it is necessary to correct for the possible over detection of a CP. The non-parametric test statistic places less restriction on the time series considered although several tuning parameters are required under this approach and so care is required. Uncertainty for CP estimates is captured via the use of asymptotic arguments in obtaining consistent estimates.

This chapter proposes a methodology to quantify the uncertainty of autocovariance CPs. Building upon the existing wavelet-based approach of Cho and Fryzlewicz (2012), we model the time series as a LSW process and perform our analysis using the wavelet periodogram. We derive a joint density for scale processes of the raw wavelet periodogram which can be embedded into a Hidden Markov Model (HMM) framework. By modelling the periodogram as a HMM, this allows a variety of existing CP methods to potentially be applied (for example changes in state in the Viterbi sequence (Viterbi, 1967)), with our focus being that of quantifying the uncertainty of CPs as proposed in Nam et al. (2012b) and in Chapter 3.

By considering time series in the frequency domain, and more specifically at different locations and frequencies under the wavelet transform, this may allow time series exhibiting changes in autocovariance to be more readily considered. This includes piecewise MA processes, which as remarked by Nason et al. (2000), have a piecewise constant EWS.

We motivate the proposed methodology with an oceanographic application, as presented in Figure 1.3 (page 7). In oceanography, historic wave height data is often used to determine storm season changes. Identifying such changes in storm seasons provides a better understanding of the data for oceanographers which may help them in planning future maintenance work of equipment such as offshore oil rigs. Changes in autocovariance structure are associated with these storm season changes, and thus autocovariance CP methods are employed in determining these changes. However, there is evidently ambiguity associated with these changes, such as their number and location, which traditional CP methods often fail to capture. By quantifying the uncertainty associated with such changes, we can thus address the ambiguity associated with storm season changes.

The structure of this chapter is as follows: Section 6.2 provides the motivation for the proposed methodology. Section 6.3 provides background into wavelet analysis which is involved in the proposed methodology. Section 6.4 details the proposed wavelet based HMM framework and modelling approach. Section 6.5 applies the proposed framework to a variety of simulated data and the oceanographic dataset as presented in Figure 1.3 (page 7). Section 6.6 concludes the chapter.

6.2 Motivation

Let $y_1, \ldots, y_n$ denote a potential non-stationary time series, observed at equally spaced discrete time points. We assume in this chapter that the non-stationarity arises due to a varying second-order structure such that for any lag $v \geq 0$, there exists a τ such that

$$\mathrm{Cov}(Y_1, Y_v) = \ldots = \mathrm{Cov}(Y_{\tau-1}, Y_{\tau-v}) \neq \mathrm{Cov}(Y_\tau, Y_{\tau-v+1}) = \ldots = \mathrm{Cov}(Y_{n-v+1}, Y_n),$$

and that the mean remains constant. In situations where the mean is not constant, pre-processing of the data can be performed. We refer to τ as a CP. Changes in second-order structure can be constructed easily; for example by a piecewise autoregressive moving average (ARMA) process.

As demonstrated in Chapter 3, one approach in modelling time series exhibiting non-stationarity such as changes in mean and variance is via Hidden Markov Models (HMMs), and are extensively used in CP analysis (for example Chib (1998), Aston et al. (2011)). We retain the same notation and framework as in Section 2.12 (page 35) and throughout this thesis.

It is possible to model certain types of autocovariance changes under the HMM framework. For example, one can consider a generalised Markov Switching Autoregressive Moving Average model of order r and q, MS-ARMA(r, q), which we define as follows.

$$Y_t = \sum_{r'=1}^{r} \delta_{X_t,r'} Y_{t-r'} + \epsilon_t + \sum_{q'=1}^{q} \kappa_{X_t,q'} \epsilon_{t-q'} \qquad \epsilon_t \sim \mathrm{N}(0, \sigma^2_{X_t}). \tag{6.1}$$

Here, $\delta_{X_t,r'}, r' = 1, \ldots, r$ are state-dependent AR coefficients, $\kappa_{X_t,q'}, q' = 1, \ldots, q$ are state-dependent MA coefficients, μ_{X_t} denotes a state-dependent mean and $\sigma^2_{X_t}$ is a state-dependent innovation variance.

When $q = 0$, this reduces to a Markov Switching Autoregressive model which retains finite dependency on the underlying state sequence under analysis (the state dependent emission depends only on a finite number of previous underlying states and observations). This permits standard algorithms associated with HMMs such as computation of the likelihood via the Forward Backward equations (Baum et al., 1970) and the methodologies presented in Chapter 3 and 5 to be applied. However, assuming an autoregressive structure is a strong assumption and somewhat limits the type of behaviour that can be modelled under such a model. For example, changes in autocovariance structure which are a result of changing moving average behaviour will not be captured fully under a Markov Switching Autoregressive model.

One misguided approach in modelling data exhibiting such behaviour would be to include a Moving Average component ($q \neq 0$). However, upon introducing this Moving Average structure, the emission density becomes dependent on the entire history of underlying state sequence $X_{1:t}$ and previous observation $Y_{1:t-1}$, and thus the model loses its Markovian structure. For example, consider the case of a MS-ARMA(1,1) model. The model can be expressed as follows:

$$\begin{aligned}
& & Y_t &= \delta_{X_t,1} Y_{t-1} + \kappa_{X_t,1} \epsilon_{t-1} + \epsilon_t \\
&\text{Then,} & \epsilon_t &= Y_t - \delta_{X_t,1} Y_{t-1} - \kappa_{X_t,1} \epsilon_{t-1} \\
&\text{Via recursions,} & Y_t &= \delta_{X_t,1} Y_{t-1} + \epsilon_t + \kappa_{X_t,1} (Y_{t-1} - \delta_{X_{t-1},1} Y_{t-2} \\
& & &\quad - \kappa_{X_{t-1},1} \left[Y_{t-2} - \delta_{X_{t-2},1} Y_{t-3} - \kappa_{X_{t-2},1} (\ldots) \right])
\end{aligned}$$

Thus the state-dependent emission density of Y_t depends on $Y_{1:t-1}$ and $X_{1:t}$. As the entire history of the underlying state sequence needs to be recorded, the model loses its Markovian structure and standard inference methods such as computing the likelihood via filtering cannot be performed. This loss of the Markovian structure is also applicable for Switching GARCH models, as described in Frühwirth-Schnatter (p.383, 2005). Approximations are thus required in order to perform inference regarding such models.

One potential approach is to model the time series as a Markov Switching AR process, regardless of how it is potentially generated, with the AR order approximating the dependence structure. This is a common approach even in a non-HMM framework, for example, as observed in the AutoPARM approach (Davis et al., 2006). Alternatively, it may also be possible to consider the time series in an alternative domain, for example as observed in Ombao et al. (2001) and Cho and Fryzlewicz (2012). This chapter will investigate the potential of transforming the problem to an alternative domain, namely the wavelet domain, which permits a time and frequency decomposition of the time series.

6.3 Wavelets

Wavelets are compactly supported oscillating functions which are used in a variety of scientific areas including signal processing (Rioul and Vetterli, 1991), Statistics (Abramovich et al., 2000) and data compression (Salomon, 2004). Statistical applications include time series analysis, density estimation and non-parametric regression (see Abramovich et al. (2000) for a good introductory paper). Applications predominantly focus on the Discrete Wavelet Transform (DWT), due to the discrete and finite nature in which data is collected. The DWT is analogous to the Fast Fourier Transform (FFT) in that instead of sinusoidal functions at different frequencies forming an orthonormal basis, it is translations and dilations of a specified wavelet function which forms the orthonormal basis. Wavelet analysis therefore permits a time series or function to be equivalently represented at different scales (frequency bands) and locations.

A graphical representation of the DWT is demonstrated in Figure 6.1, with respect to the simplest possible mother wavelet $\psi(x)$, the Haar wavelet. An orthonormal basis is formed from dyadic translations and dilations of the mother wavelet which are denoted by $\{\psi_{j,k}(x)\}_{j,k\in\mathbb{Z}}$. j and k are known as the scale and location in the literature and correspond to the dilation and translation factors. These translated and dilated versions of the mother wavelet, $\{\psi_{j,k}(x)\}_{j,k\in\mathbb{Z}}$, are referred to as daughter wavelets (the red curves in Figure 6.1 where only the non-zero behaviour has been displayed). In turn, the function of interest $f(\cdot)$, can be written as a linear combination of daughter wavelets $\psi_{j,k}$, where $d_{j,k}$ represents the contribution of the corresponding daughter wavelet. This thus provides a scale and location decomposition of the function where the latter property arises from the localised behaviour of the wavelets. In contrast, the FFT provides only a scale decomposition and not location wise due to the global nature of the sinusoidal basis functions considered.

As the data considered in this thesis is of a temporal nature, we shall focus our review on wavelet methods and analysis concerning time series data. Section 6.3.1 reviews the Discrete Wavelet Transform (DWT), an algorithm which provides the wavelet decomposition of a time series. However certain disadvantages exist with the DWT, especially with respect to CP analysis. Thus the Non-Decimated Wavelet Transform (NDWT), an extension of the DWT, is reviewed in Section 6.3.2. Finally, Section 6.3.3 reviews the Locally Stationary Wavelet (LSW) process framework which permits time series with varying second-order structure (variance and covariance) to be considered. We refer interested readers to Vidakovic (1999), Percival and Walden (2007) and Nason (2008) for comprehensive overviews of wavelets

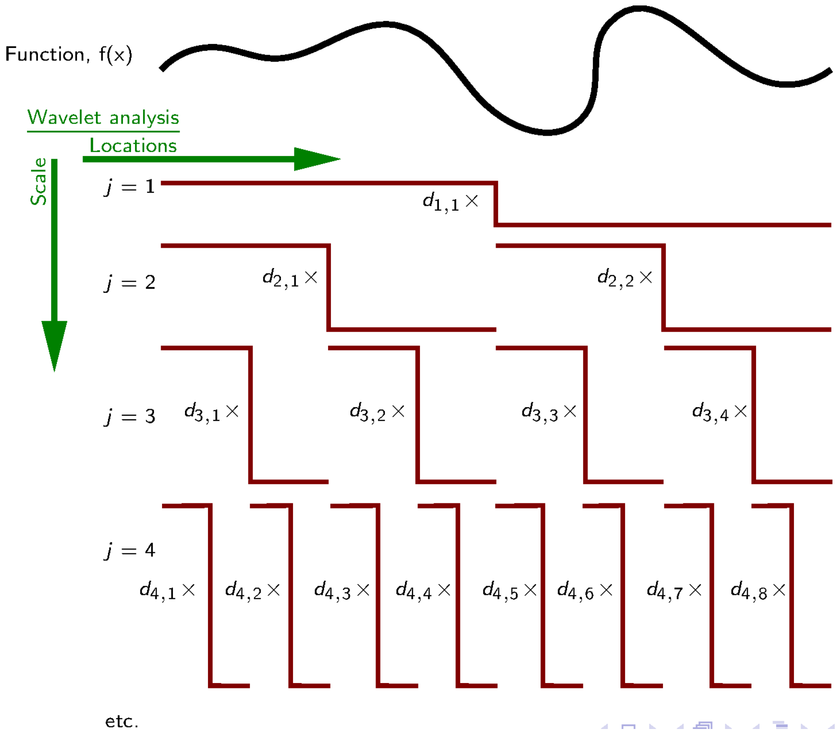


Figure 6.1: Graphical representation of the Discrete Wavelet Transform using the Haar mother wavelet. The function of interest f is equivalently represented as a linear combination of daughter wavelet $\{\psi_{j,k}(x)\}_{j,k\in\mathbb{Z}}$, where the daughter wavelets are translations and dilations of the Haar mother wavelet $\psi(x)$. This representation allows decomposition of the function at different scales j (frequency bands) and locations k.

in Statistics and time series analysis.

Wavelet analysis is associated with functions that are square integrable, that is $f(\cdot) \in \mathbb{L}_2(\mathbb{R})$. A wavelet is more formally defined as follows.

Definition 5. *$\psi(\cdot) \in \mathbb{L}_2(\mathbb{R})$ is defined to be a wavelet function (mother wavelet) if it satisfies the following conditions.*

1. *The admissibility condition*

$$C_\psi = \int_{\mathbb{R}} \frac{|\Psi(\omega)|^2}{|\omega|} d\omega < \infty$$

where $\Psi(\omega)$ is the Fourier transformation of $\psi(x)$. That is $\Psi(\omega) = \langle \psi(x), \exp(i\omega x)\rangle = \int_{\mathbb{R}} \psi(x) \exp(-i\omega x) dx$.

2. *$\Psi(0) = \int_{\mathbb{R}} \psi(x) dx = 0$.*

3. *The dyadic translations and dilations of the wavelet $\psi(\cdot)$, form an orthonormal basis of $\mathbb{L}_2(\mathbb{R})$. These translated and dilated versions are of the form:*

$$\psi_{j,k}(x) = 2^{-\frac{j}{2}}(2^{-j}x - k) \qquad j, k \in \mathbb{Z}$$

j and k are noted as scales and location parameters.

Condition 1 results in $\psi(\cdot)$ having exponential decay over $\mathbb{L}_2(\mathbb{R})$ which permits localised behaviour to be captured. Condition 2 ensures that $\psi(\cdot)$ possess an oscillating behaviour such that the area of the function is equal to 0 and therefore this oscillating behaviour is controlled. Condition 3 states that shifts and stretches (translations and dilations) of $\psi(\cdot)$ form an orthonormal basis of $\mathbb{L}_2(\mathbb{R})$ in which functions of interest lie. The scale parameter j corresponds to the frequency band that will be captured by $\psi_{j,k}$, and at the respective location k.

A variety of wavelets exist, each with varying degrees of smoothness and localised support. The wavelets that are commonly considered in statistical applications are Daubechies' Compactly Supported wavelets; a family of wavelets which have compact finite support. However other wavelets also exist, for example Shannon's Wavelets, the Mexican Hat wavelet and Meyer's Wavelets (Vidakovic, 1999, pp. 60–80) which have exponential decay. The smoothness of a wavelet is classified by the number of vanishing moments it possesses. This is defined as follows:

Definition 6. *The mother wavelet function $\psi(\cdot)$ is said to have $v \in \mathbb{Z}_+$ vanishing moments if*

$$\int_{\mathbb{R}} x^m \psi(x) dx = 0$$

holds for $m = 0, 1, \ldots, v$.

As v increases, the mother wavelet becomes smoother and has a larger support. A variety of wavelets with various vanishing moments from Daubechies' Compactly Supported wavelet family (which are consequently further divided into Daubechies' Extremal Phase and Least Asymmetric wavelets) are presented in Figure 6.2. A consequence of the vanishing moment property is that in representing a polynomial function with degree $(v - 1)$ or less, the wavelet representation will consist of zeroes. This leads to sparse wavelet representations and it is this sparsity which is part of the attraction of wavelet analysis in a variety of applications. The sparseness of this representation is determined by the number of vanishing moment the chosen mother wavelet possesses.

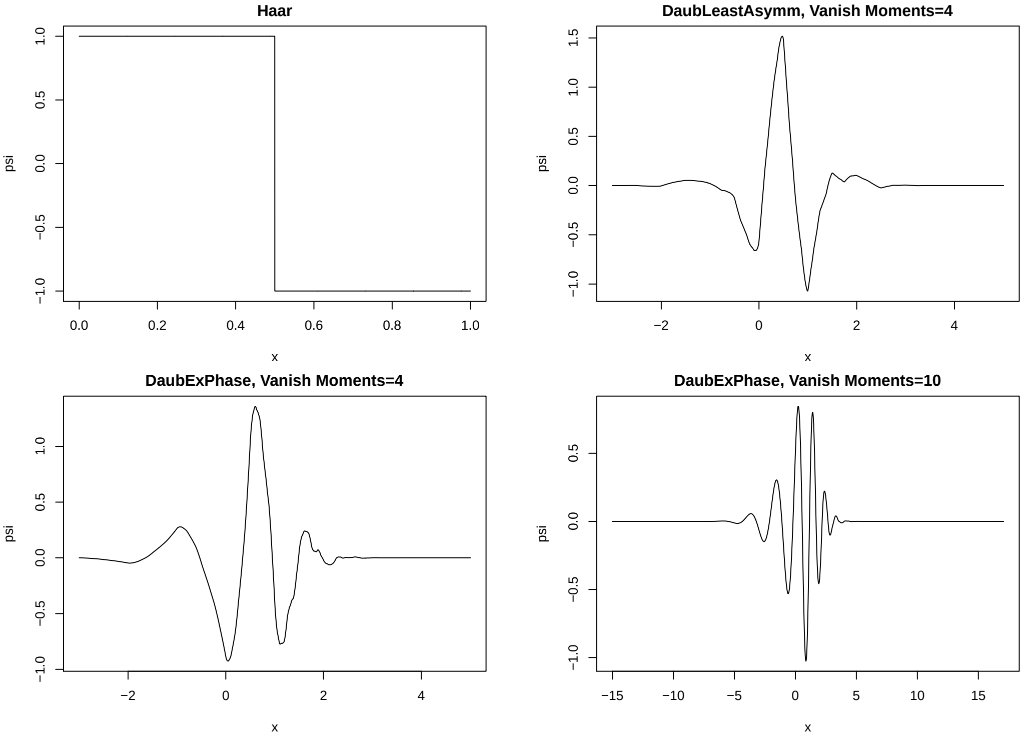


Figure 6.2: Examples of mother wavelet functions with various vanishing moments. These are: Haar wavelet with $v = 1$ (top left), Daubechies' Least Asymmetric wavelet with $v = 4$ (top right), Daubechies' Extremal Phase wavelet with $v = 4$ (bottom left), and Daubechies' Extremal Phase wavelet with $v = 10$ (bottom right).

The main backbone of wavelet analysis is Multiresolution Analysis (MRA, see Section 3.3, p. 51 Vidakovic (1999) for further details). This permits $f(\cdot) \in \mathbb{L}_2(\mathbb{R})$ (a square integrable function) to be approximated at different resolutions and equivalently via the orthonormal basis, $\{\psi_{j,k}(x)\}_{j,k\in\mathbb{Z}}$; translations and dilations of the mother wavelet, $\psi(x)$. In obtaining the approximations of $f(\cdot)$ at different resolutions (scales), this involves translation and dilations of the father wavelet $\phi(x)$. The scaling equation, $\phi_j(x) = \sum_{k\in\mathbb{Z}} h_k 2^{-\frac{j}{2}}\phi(2^{-j}x - k)$, describes how the functions are related at different resolutions (scale j). The coefficients $\{h_k\}_{k\in\mathbb{Z}}$ are known as the low-pass (averaging) filter.

The mother and daughter wavelets, $\psi(x)$ and $\psi_{j,k}(x)$ can also be expressed in terms of the father wavelet $\phi(x)$, namely $\psi_{j,k}(x) = \sum_{k\in\mathbb{Z}} g_k 2^{-\frac{(j-1)}{2}}\phi(2^{-\frac{(j-1)}{2}}x - k)$. The coefficients $\{g_k\}_{k\in\mathbb{Z}}$ are known as the high-pass filter respectively. The quadrature mirror relation states the relationship between the high-pass filter $\{g_k\}_{k\in\mathbb{Z}}$, and the low-pass filter $\{h_k\}_{k\in\mathbb{Z}}$, as follows.

$$g_k = (-1)^k h_{1-k} \qquad k \in \mathbb{Z}. \tag{6.2}$$

This thesis will focus on the use of the Haar mother wavelet in analysis which has many advantages including its simplicity and intuitiveness. The proposed methodology in this chapter can however be extended to the use of other mother wavelet in Daubechies' Compactly Supported wavelet family. More specifically, the Haar wavelet has the corresponding father scaling wavelet

$$\phi(x) = \begin{cases} 1, & 0 \leq x < 1; \\ 0, & \text{otherwise.} \end{cases} . \tag{6.3}$$

From the scaling equation, this gives rise to the low-pass filter coefficients, $h_0 = h_1 = \frac{1}{\sqrt{2}}$ and $h_k = 0$ for all other values of $k \in \mathbb{Z}$. From the quadrature mirror filter relationship (Equation 6.2), the corresponding high-pass filter coefficients are $g_0 = \frac{1}{\sqrt{2}}$, $g_1 = -\frac{1}{\sqrt{2}}$, and $g_k = 0$ for all remaining values of $k \in \mathbb{Z}$. The mother wavelet function is thus derived as

$$\begin{aligned} \psi(x) &= \sum_{k\in\mathbb{Z}} g_k \phi_0(x) = \sum_{k\in\mathbb{Z}} g_k 2^{\frac{1}{2}} \phi(2x-k) = \phi(2x) - \phi(2x-1) && (6.4) \\ &= \begin{cases} 1, & 0 \leq x < \frac{1}{2}; \\ -1, & \frac{1}{2} \leq x < 1; \\ 0, & \text{otherwise.} \end{cases} . && (6.5) \end{aligned}$$

6.3.1 Discrete Wavelet Transform (DWT)

The Discrete Wavelet Transform (DWT) is an efficient, fast procedure which transforms a dyadic length time series observed at equally spaced points, $\mathbf{y} = y_{0:n-1} = (y_0, \ldots, y_{n-1}), n = 2^J$ for $J \in \mathbb{N}^+$, into an equivalent wavelet representation defined by the scaling and detail coefficients, $c_{j,k}$ and $d_{j,k}$ for $j = 1, \ldots, J$, $k = 0, \ldots, 2^{J-j} - 1$. These coefficients denote the contribution of the father and daughter wavelet at respective scales in the equivalent representation. More specifically, the wavelet representation consists of,

$$\mathbf{y}^\star = \left(\{c_{J,k}\}_{k=0}^{1}, \{d_{J,k}\}_{k=0}^{1}, \{d_{J-1,k}\}_{k=0}^{3}, \ldots, \{d_{1,k}\}_{k=0}^{2^{J-1}-1} \right) = \left(\mathbf{c}_J, \mathbf{d}_J, \ldots, \mathbf{d}_1 \right).$$

That is, the detail coefficients from all scales $j = 1, \ldots, J$, and the scaling coefficients at the coarsest scale $j = J$. Each coefficient vector, $\mathbf{c}_j$ and $\mathbf{d}_j$, contains 2^{J-j} elements for $j = 1, \ldots, J$. The total number of elements in the complete wavelet representation $\mathbf{y}^\star$, is thus n. This wavelet representation is an equivalent representation of the original time series $\mathbf{y}$, such that the $||\mathbf{y}^\star||_2 = ||\mathbf{y}||_2$ where $||\mathbf{z}||_2$ is the $\mathbb{L}_2$-norm of vector $\mathbf{z}$. This is more formally known as Parseval's relation in the literature and states that the energy of the original signal is retained under the new wavelet representation.

This wavelet decomposition can be efficiently computed by performing Mallat's Pyramid algorithm (Mallat, 1989). The general principal is that one sets $\mathbf{c}_0 = \mathbf{y}$, and computes $\mathbf{c}_j$ and $\mathbf{d}_j$ from $\mathbf{c}_{j-1}$, for $j = 1, \ldots, J$ via the use of the low and high-pass filters $\{h_k\}_{k\in\mathbb{Z}}, \{g_k\}_{k\in\mathbb{Z}}$. That is, coarser scale scaling and detail coefficients are computed from the scaling coefficients from the previous finer scale. More formally, the algorithm recursions are of the form:

$$\begin{aligned} c_{j+1,k} &= \sum_{l\in\mathbb{Z}} h_l c_{j,l+2k} = \sum_{l\in\mathbb{Z}} h_{l-2k} c_{j,l}. \\ d_{j+1,k} &= \sum_{l\in\mathbb{Z}} g_{l-2k} c_{j,l}. \end{aligned} \tag{6.6}$$

The algorithm can also be expressed in terms of filter and decimation operators. Let $\mathcal{H} = \{h_k\}_{k\in\mathbb{Z}}$ and $\mathcal{G} = \{g_k\}_{k\in\mathbb{Z}}$ denote the low and high-pass filter operators. These have the following effects on sequences.

Definition 7. *For a doubly infinite sequence* $(\ldots, z_{-1}, z_0, z_1, \ldots)$*, the operator* $\mathcal{H}$ *has the following effect on the sequence*

$$(\mathcal{H}z)_k = \sum_{n\in\mathbb{Z}} h_{n-k} z_n. \tag{6.7}$$

Similarly, the operator $\mathcal{G}$ *has the following effect,*

$$(\mathcal{G}z)_k = \sum_{n\in\mathbb{Z}} g_{n-k} z_n. \tag{6.8}$$

The binary decimation operator $\mathcal{D}_0$, has the following effect on a sequence.

Definition 8. *The binary decimation operator* $\mathcal{D}_0$ *is defined such that it chooses every even element of the sequence. That is*

$$(\mathcal{D}_0 z)_j = z_{2j}. \tag{6.9}$$

These operators are defined with respect to doubly infinite sequences. It is noted however that in the DWT, we have a finite dyadic length sequence. In applying these operators to finite length sequences, the periodic and symmetric boundary conditions are commonly implemented within the wavelet community (see pages 55-57 Nason (2008)). Such conditions effectively repeat and reflect the sequence around the origin respectively; that is $(\ldots, z_{n-2}, z_{n-1}, z_0, z_1, z_2, \ldots)$ and $(\ldots, z_1, z_0, z_0, z_1, \ldots)$.

Equations 6.6 can thus be written in terms of these filter and decimation operators,

$$\mathbf{c}_j = \mathcal{D}_0 \mathcal{H} \mathbf{c}_{j-1} \qquad \mathbf{d}_j = \mathcal{D}_0 \mathcal{G} \mathbf{c}_{j-1} \quad j = 1, \ldots, J. \tag{6.10}$$

Letting $\mathbf{c}_0 = \mathbf{y} = y_{0:n-1}$, then Equation 6.10 can also be expressed with regards to the original time series,

$$\mathbf{c}_j = (\mathcal{D}_0 \mathcal{H})^j \mathbf{c}_0 \qquad \mathbf{d}_j = (\mathcal{D}_0 \mathcal{G})(\mathcal{D}_0 \mathcal{H})^{j-1} \mathbf{c}_0 \quad j = 1, \ldots, J. \tag{6.11}$$

Figure 6.3 demonstrates Mallat's Pyramid algorithm in practice on a simple short time series using a Haar wavelet. Recall that corresponding Haar filter coefficients are $h_0 = h_1 = \frac{1}{\sqrt{2}}$, $g_0 = \frac{1}{\sqrt{2}}$, $g_1 = -\frac{1}{\sqrt{2}}$ and zeroes elsewhere in the corresponding filters. It is noted that Parseval's relation is satisfied with $||\mathbf{y}^\star||_2 = ||\mathbf{y}||_2$. In addition, the wavelet decomposition is sparse compared to the original time series with several zeroes being present in the decomposition. This sparser representation is part of the attraction of the wavelets, particularly in the data compression community. Due to the non-overlapping nature of the filters on the observations at each scale (a result of the orthogonal transform), the DWT can also remove some of the unknown dependent structure in the time series (see p. 341 of Percival and Walden (2007) for further details). This is also another benefit of wavelet analysis.

As the DWT is effectively a change in basis representation into the wavelet domain, the DWT can consequently be expressed in terms of matrix and vector notation where the matrix represents the change in basis transformation. That is

$$\mathbf{y}^\star = \mathbf{K}\mathbf{y} \tag{6.12}$$

where $\mathbf{y}^\star, \mathbf{y}$ are n length column vectors and $\mathbf{K}$ is an $n \times n$ matrix. The entries of $\mathbf{K}$ are the effective high and low pass-filter coefficients with respect to being applied to the data $\mathbf{y}$ directly. As the DWT is an orthogonal transformation, $\mathbf{K}$ is consequently an orthogonal matrix. Whilst this matrix representation provides another perspective of the DWT, it is seldomly considered in performing the DWT due to its higher computational cost compared to Mallat's Pyramid algorithm; the computational cost of Mallat's Pyramid algorithm and the matrix operation are respectively $\mathcal{O}(n)$ and $\mathcal{O}(n^2)$.

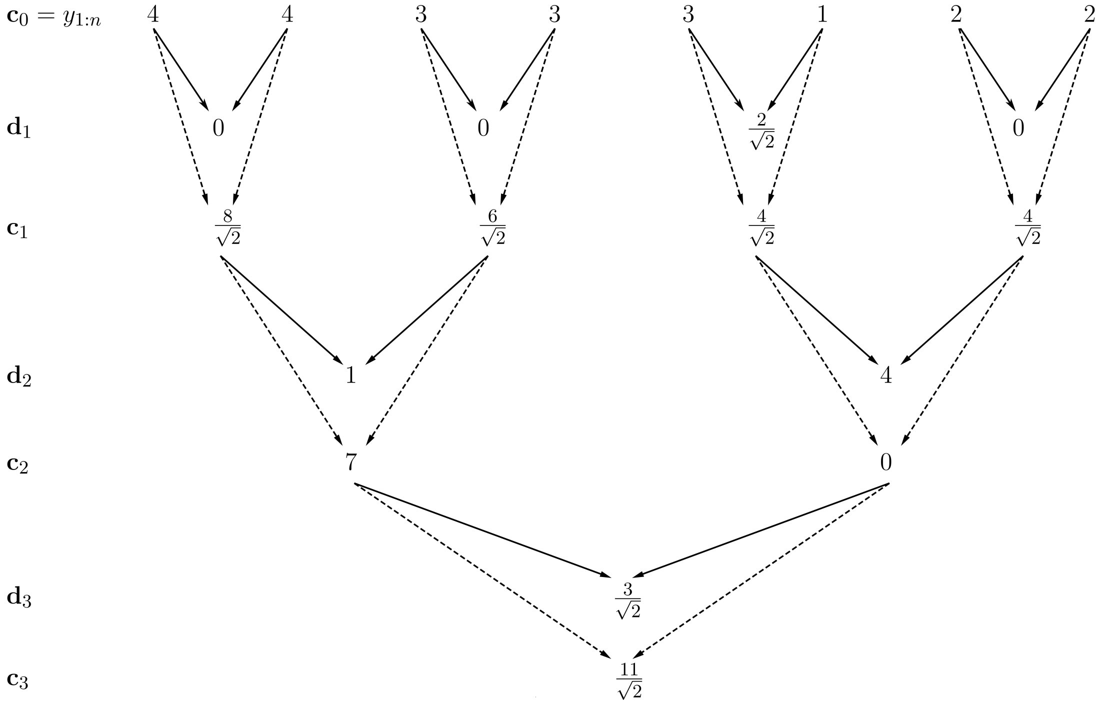


Figure 6.3: Example of the DWT in practice on a toy time series. This figure demonstrates how Mallat's Pyramid algorithm is used to efficiently compute the wavelet decomposition associated with the DWT. A Haar wavelet has been used with the following filter coefficients: $h_0 = h_1 = \frac{1}{\sqrt{2}}$, $g_0 = \frac{1}{\sqrt{2}}$, $g_1 = -\frac{1}{\sqrt{2}}$ and zeroes elsewhere in the corresponding filters. The observed time series $\mathbf{y} = (4, 4, 3, 3, 3, 1, 2, 2)$ has the following wavelet decomposition $\mathbf{y}^{\star} = (\frac{11}{\sqrt{2}}, \frac{3}{\sqrt{2}}, 1, 4, 0, 0, \frac{2}{\sqrt{2}}, 0)$.

However, the matrix representation illustrates well that an inverse DWT does exist such that one can recover the original time series $\mathbf{y}$, if provided with the wavelet decomposition $\mathbf{y}^{\star}$, and the mother wavelet used in obtaining this wavelet decomposition. In terms of matrix notation, this results in

$$\mathbf{y} = \mathbf{K}^{-1}\mathbf{y}^{\star} = \mathbf{K}^{T}\mathbf{y}^{\star}. \tag{6.13}$$

With regards to Mallat's Pyramid algorithm perspective, the inverse transform expresses $\mathbf{c}_{j-1}$ in terms of $\mathbf{c}_j$ and $\mathbf{d}_j$, $j = J, \ldots, 1$. That is, finer scale scaling coefficients are calculated from coarser scale scaling and detail coefficients. In terms of the filter and decimation and operator notation, it is necessary to pad out intermediary sequences due to the number of coefficients halving from finer to coarser scales. This is achieved by defining the inverse of the binary decimation operator, $\mathcal{D}_0^{-1}$, which inserts zeroes between each element in the vector that it is applied to. Consequently the inverse DWT is provided by

$$\mathbf{c}_{j-1} = \mathcal{H}\mathcal{D}_0^{-1}\mathbf{c}_j + \mathcal{G}\mathcal{D}_0^{-1}\mathbf{d}_j \qquad j = J, \dots 1. \tag{6.14}$$

For individual coefficients, this results in the expression,

$$c_{j-1,l} = \sum_{k\in\mathbb{Z}} h_{l-2k}c_{j,k} + \sum_{k\in\mathbb{Z}} g_{l-2k}d_{j,k}. \tag{6.15}$$

The wavelet representation obtained by the DWT is specific to the wavelet basis that one transforms onto. Consequently, modifications of the DWT exist which lead to alternative orthonormal bases being considered. For example, in the DWT presented in this section, a decimation operator is performed at each step which takes forward only the even elements of a vector, and discards the odd elements. However, it is also possible to perform the reverse; retain the odd elements and discard the even. This thus leads to a different wavelet representation with respect to the new basis. This can be further extended such that a mixture of odd and even decimation takes place, with the sequence of even-odd decimation operations performed being recorded. This again leads to another wavelet basis and is termed ϵ-decimated DWT in the literature. We refer the reader to the aforementioned reference texts with more details regarding modifications of the DWT. However, decimation is an important part of the DWT in order for the transform to remain orthogonal.

One disadvantage of the DWT is that it is not translation equivariant in that a shift in the time series does not correspond to a shift by the same amount in the wavelet decomposition. This is demonstrated successfully in Figure 6.4, where the DWT has been performed on both a block test function and a shifted version of this function (to the left by 75 observations). The plots, which are typical in the wavelet community, displays only the detail coefficients of the wavelet decomposition $\mathbf{y}^\star$ at their respective scales and locations (the observations that they are computed from). Such sensitivity of the wavelet decomposition to the orientation of the data is not desired. In addition, the location of CPs in the test function can become lost in the wavelet decomposition obtained, dependent on the orientation of the data. For example, the second jump in both versions of the time series analysed appears in the unshifted DWT analysis (Figure 6.4(b)) but does not appear in the shifted DWT representation (Figure 6.4(d)) when considering the finest scale coefficients (resolution level nine). Such sensitivity to the orientation of the data and the potential obscurance of CPs quantities in the wavelet decomposition is not desirable for CP problems of interest. These issues are addressed in the non-decimated Discrete Wavelet Transform.

6.3.2 Non-Decimated Wavelet Transform (NDWT)

The non-decimated wavelet transform (NDWT, also known as the Stationary Wavelet Transform, Nason and Silverman (1995)) can be thought of "filling in the gaps" that is resultant from the DWT and interpreted in several different manners. Firstly, as the name suggests, the NDWT does not perform the decimation at each step. Consequently, the same number of scaling and detail coefficients are retained at each scale with the number of coefficients at each scale being equal to the length of the time series analysed. This retention of coefficients thus fills in the gaps. Alternatively, the NDWT can be thought of as performing the DWT on every possible shift configuration of $\mathbf{y}$ and ordering the coefficients obtained in a systematic manner (for example, time ordered with respect to the moving window of observations the filter is performed on). By considering these overlapping filter windows of observations, compared to the non-overlapping orthogonal filter windows in the DWT, this fills in the gaps that is lost under the DWT. This retention of additional coefficients provides an over complete, redundant representation of the data and consequently means the NDWT is not an orthogonal transform. There is therefore no unique inverse NDWT in which the original time series can be recovered when given the wavelet representation from a NDWT.

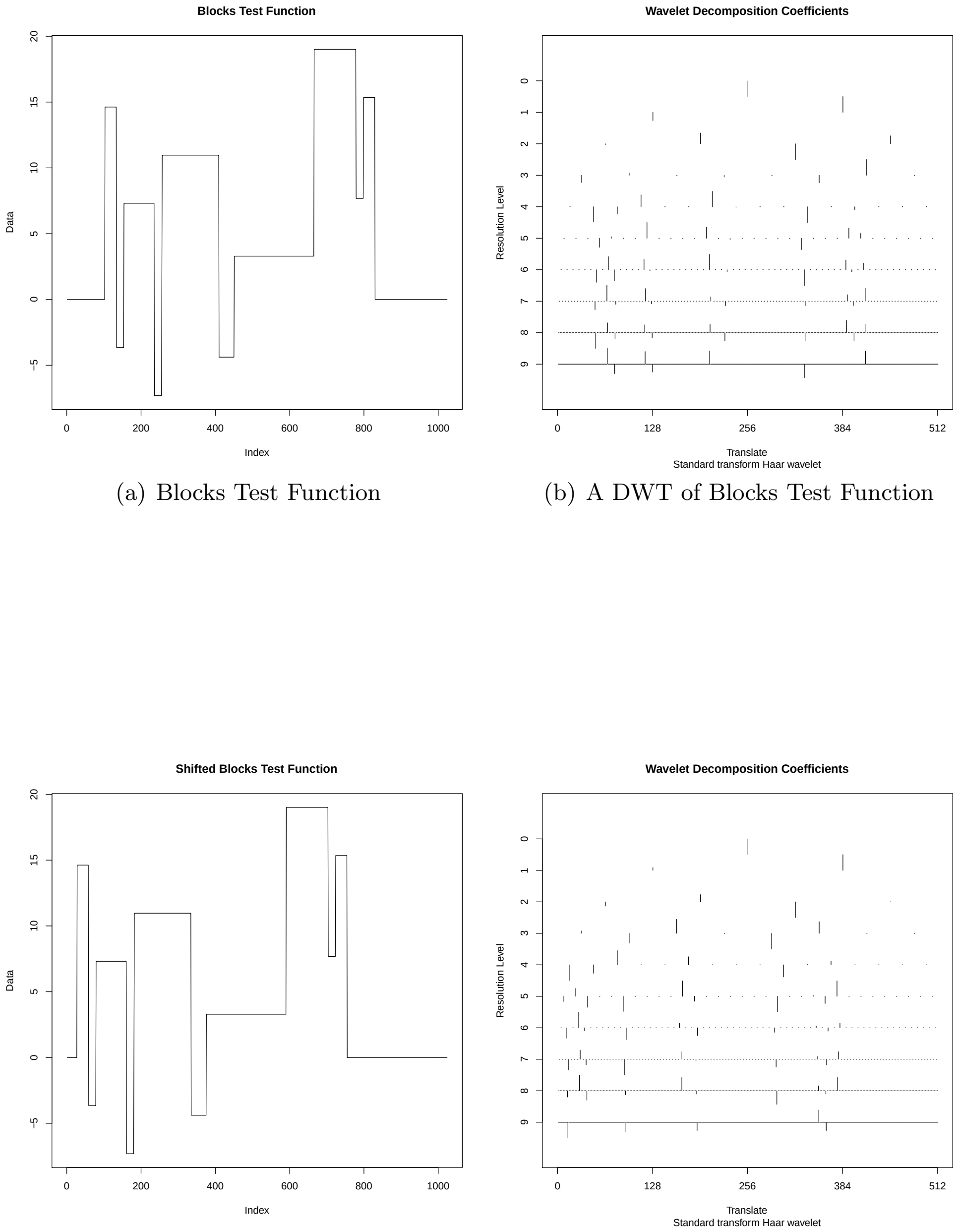
In defining the NDWT, the high and low-pass filter operations defined in Definition 7 need to be modified such that we retain the same number of coefficients at each scale. This is achieved by defining the new set of filter operators.

Definition 9. *The non-decimated wavelet transform uses low and high-pass filters which are defined recursively as*

$$\begin{aligned} \mathcal{H}^{[0]} &= \mathcal{H} = \{h_k\}_{k\in\mathbb{Z}} & \qquad \mathcal{G}^{[0]} &= \mathcal{G} = \{g_k\}_{k\in\mathbb{Z}} \\ \mathcal{H}^{[r]} &= \mathcal{D}_0^{-1}\mathcal{H}^{[r-1]} & \qquad \mathcal{G}^{[r]} &= \mathcal{D}_0^{-1}\mathcal{G}^{[r-1]}. \end{aligned}$$

The effective filter is therefore the original filter with numerous zeroes between each element. The NDWT can thus be defined in terms of these new filters.

Definition 10. *Let $\mathbf{c}'_j$ and $\mathbf{d}'_j$ be the over-complete scaling and detail coefficients at scale j respectively from a NDWT. Then the coarser scaling and detail coefficients*



(a) Blocks Test Function

(b) A DWT of Blocks Test Function

(c) Shifted Blocks Test Function

(d) A DWT of Shifted Blocks Test Function

Figure 6.4: A Discrete Wavelet Transform on the blocks test function and a shifted version of the blocks test function. This demonstrates that the DWT is not translation invariant as a shift in the data does not correspond to a shift in the wavelet representation. Such sensitivity to the orientation of the data may not be desirable for CP analysis.

at the next coarser scale are defined recursively by

$$\mathbf{c}'_j = \mathcal{H}^{[j-1]}\mathbf{c}'_{j-1} \qquad \mathbf{d}'_j = \mathcal{G}^{[j-1]}\mathbf{c}'_{j-1} \qquad j = 1, \dots J. \tag{6.16}$$

where $\mathbf{c}'_0 = \mathbf{c}_0 = \mathbf{y} = y_{0:(n-1)}$.

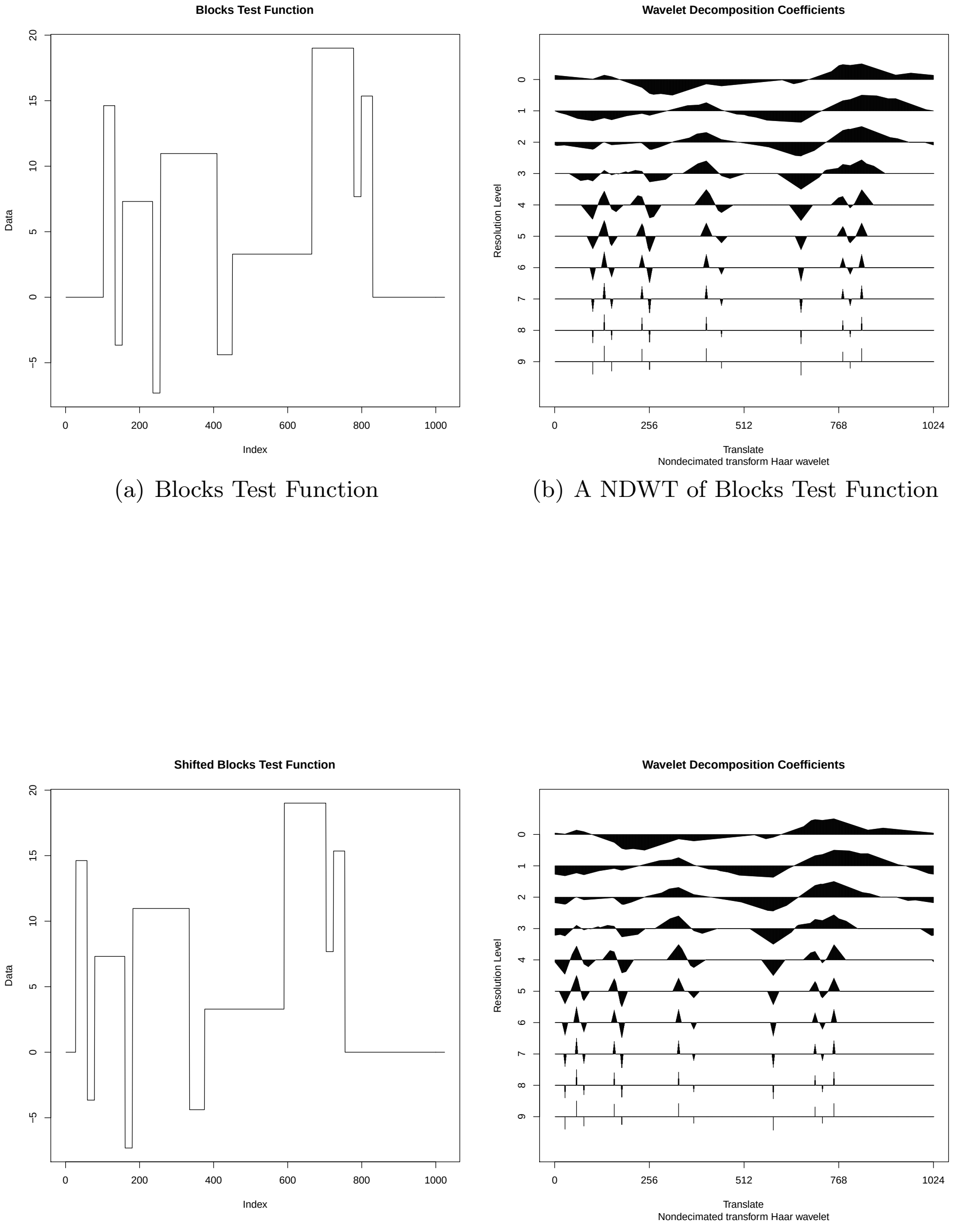
The general principle of the DWT where coarser scale coefficients are computed from finer scale coefficients still exists within the NDWT algorithm, although no decimation occurs. This results in the overcomplete wavelet decomposition $\mathbf{y}' = (\mathbf{c}'_J, \mathbf{d}'_J, \dots, \mathbf{d}'_1)$ where both $\mathbf{c}'_j$ and $\mathbf{d}'_j$ contain $n = 2^J$ elements for all $j = 1, \dots, J$. $\mathbf{y}'$ thus has $n(J+1)$ elements. Due to the retention of coefficients and more calculations being involved, the NDWT evidently has a higher computational cost of $\mathcal{O}(n \log_2 n)$ compared to the DWT, although this is still considered to be fast. An equivalent matrix representation of transformation also exists. Namely,

$$\mathbf{y}' = \mathbf{K}'\mathbf{y}, \tag{6.17}$$

where $\mathbf{y}'$ is a $n(J+1)$ lengthed column vector, $\mathbf{y}$ is a column vector of length n, and $\mathbf{K}'$ is a $n(J+1) \times n$ matrix. The entries of $\mathbf{K}'$ are the effective high and low-pass filter coefficients when applied to the observations. However, $\mathbf{K}'$ is not an orthogonal matrix and an inverse does not exist. Consequently, this further illustrates that inverse NDWT is not possible straightaway.

An example of the NDWT is demonstrated on the aforementioned blocks test function in Figure 6.5. Observe that the wavelet decomposition plots (right column) retains the same number of coefficients present at each scale, thus providing the overcomplete representation. We note that a shift in the test function data now results in a shift in the wavelet representation. Hence the NDWT is translation invariant. We also note that the location of the jumps in the function are retained and much clearer in the NDWT output, regardless of the orientation of the data. This property should therefore be useful with regards to CP analysis compared to a DWT decomposition.

The NDWT seems more relevant and useful in the context of CP analysis in contrast to the DWT, as it provides a much more complete picture of the data and the location of any CPs present. However the transform is not orthogonal and the coefficients are no longer independent due to the overlapping nature of the wavelet filters considered. However, this dependence structure amongst the coefficients is known and it is determined from the mother wavelet used in analysis. This suggests it can therefore be incorporated into statistical analysis. The NDWT is powerful and utilised in the construction of Locally Stationary Wavelet processes,



(a) Blocks Test Function

(b) A NDWT of Blocks Test Function

(c) Shifted Blocks Test Function

(d) A NDWT of Shifted Blocks Test Function

Figure 6.5: A Non-Decimated Wavelet Transform on the blocks test function and a shifted version of the blocks test function. This demonstrates that the NDWT is translation invariant as a shift in the data corresponds to a shift in the wavelet representation. This removes sensitivity of results to the orientation of the data. In addition, coefficients are retained such that the same number of coefficients is present in each scale of the representation.

a framework which allows time series with time-varying autocovariance structures to be considered. Such a framework may therefore be useful if we want to consider changes in autocovariance structure more actively.

6.3.3 Locally Stationary Wavelet processes

One common approach in modelling and representing stationary time series is in the frequency domain via Fourier analysis. As alluded to earlier, the basis functions under this representation are sinusoidal functions at different frequencies, defined globally over the entire scope of the time series. However, such a representation is not appropriate or adequate for time series exhibiting non-stationarity due to its global nature not capturing these localised features.

The Locally Stationary Wavelet (LSW) framework is a popular wavelet based modelling framework for non-stationary time series with a time varying second-order, covariance structure (Nason et al., 2000). The motivation for such a framework is that whilst time series may not be stationary over the entire scope of the data (globally), it may be stationary in smaller time windows (locally). This localised stationarity is achieved via the use of wavelets and the localised behaviour property they possess. Associated with the LSW process is the Evolutionary Wavelet Spectrum (EWS) which provides a decomposition of the autocovariance structure at different scales (frequencies) and locations. Recent applications of the LSW framework include forecasting (Fryzlewicz et al., 2003), classification (Fryzlewicz and Ombao, 2009) and CP identification (Cho and Fryzlewicz, 2012).

The main building blocks of the LSW framework are more specifically discrete non-decimated wavelets. Due to the local nature of wavelets, this makes them apt for capturing the local stationarity in the time series compared to other potential basis functions (for example, sinusoidal functions of Fourier analysis). In defining the LSW process, we first need to define non-decimated wavelet vectors. Let $\{h_k\}_{k\in\mathbb{Z}}$ and $\{g_k\}_{k\in\mathbb{Z}}$ be the aforementioned low and high-pass filter. The associated discrete wavelet vector at scale $j \geq 1$ is represented by

$$\psi_j = (\overset{\mathbf{0}}{\ldots}, \psi_{j,0}, \psi_{j,1}, \ldots, \psi_{j,(N_j-1)}, \overset{\mathbf{0}}{\ldots}), \qquad j = 1, \ldots.$$

where $\overset{\mathbf{0}}{\ldots}$ denotes an infinite long zero vector. These vectors are compactly supported with $N_j < \infty$ non-zero entries. These vectors are computed recursively via

$$\begin{aligned}
\psi_{1,l} &= \sum_k g_{l-2k}\delta_{0k} = g_l && l = 0, 1, \ldots, N_1 - 1 \\
\psi_{j+1,l} &= \sum_k h_{l-2k}\psi_{j,k} && l = 0, 1, \ldots, N_{j+1} - 1 \quad j + 1 = 2, 3, \ldots
\end{aligned}$$

where $N_j = (2^j - 1)(N_h - 1) + 1$. δ_{0k} is the Kronecker delta and N_h is the finite number of non-zero elements in $\{h_k\}_{k\in\mathbb{Z}}$.

For example, in the case of Haar wavelets at scale 1 and 2,

$$\begin{aligned}
\psi_1 &= (\overset{\mathbf{0}}{\ldots}, g_0, g_1, \overset{\mathbf{0}}{\ldots}) = (\overset{\mathbf{0}}{\ldots}, \frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}, \overset{\mathbf{0}}{\ldots}) \\
\psi_2 &= (\overset{\mathbf{0}}{\ldots}, h_0g_0, h_1g_0, h_0g_1, h_1g_1, \overset{\mathbf{0}}{\ldots}) = (\overset{\mathbf{0}}{\ldots}, \frac{1}{2}, \frac{1}{2}, -\frac{1}{2}, -\frac{1}{2}, \overset{\mathbf{0}}{\ldots}).
\end{aligned}$$

These are the effective filter coefficients when applied directly to the observations and also populate the transform matrices $\mathbf{K}$ and $\mathbf{K}'$.

ψ_j denotes the non-decimated wavelet vector at scale j, where $\psi_{j,k}$ denotes the kth non-zero entry of ψ_j. $\psi_{j,k}(t) = \psi_{j,(k-t)}$ denotes the $(k-t)$th non-zero element in ψ_j. This can also be interpreted as the kth non-zero element in a shifted version of ψ_j by amount t.

Under the DWT, $\{\psi_j\}_{j=1}^{\infty}$ is an orthonormal set of shifted vectors if we shift them by multiples of dyadic amounts 2^j. This results in an orthonormal transform. However, the NDWT lifts this restriction such that the wavelet vectors can be shifted by any desired amount, not necessarily dyadic. As a result, the discrete non-decimated wavelet vectors $\{\psi_j\}_{j=1}^{\infty}$ are no longer orthonormal, but an overcomplete collection of shifted vectors.

Following Fryzlewicz and Nason (2006), we define a LSW process as follows.

Definition 11. *$\{Y_t\}_{t=1}^n$ for $n =, 2, \ldots, 2^J, J \in \mathbb{N}_+$ is said to be a Locally Stationary Wavelet (LSW) process if the following mean-square representation exists,*

$$Y_t = \sum_{j=1}^{J}\sum_{k\in\mathbb{Z}} \psi_{j,k}(t) U_j\left(\frac{k}{n}\right)\xi_{j,k} \tag{6.18}$$

where $j \in \mathbb{N}$ and $k \in \mathbb{Z}$ denote the scale and location parameters respectively. $\psi_j = \{\psi_{j,k}\}_{k\in\mathbb{Z}}$ is a discrete, real-valued, compactly supported, non-decimated wavelet vector with support lengths $\mathcal{L}_j = \mathcal{O}(2^j)$ at each scale. $\xi_{j,k}$ is a zero-mean, orthonormal, identically distributed incremental error process (that is $\mathbb{E}[\xi_{j,k}] = 0, \mathbb{E}[\xi_{j,k}\xi_{l,m}] = \delta_{jl}\delta_{km}$).

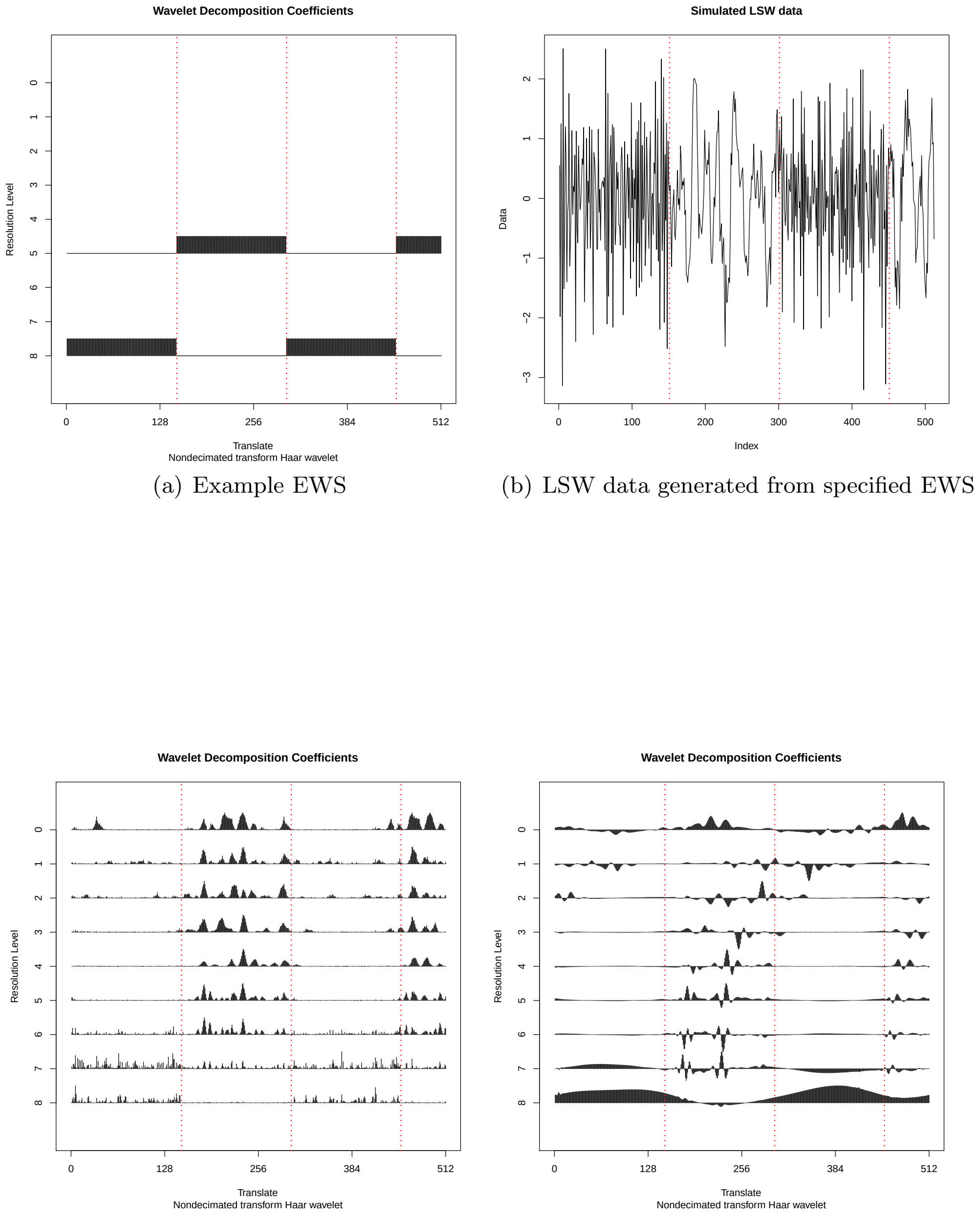
For each $j \geq 1$, $U_j(z) : [0,1] \rightarrow \mathbb{R}$ is a real valued, piecewise constant function with a finite (but unknown) number of jumps. Let $\mathcal{N}_j$ denote the total magnitude of jumps in $U_j^2(z)$, the variability of function $U_j^2(z)$ is controlled so that:

- $\sum_{j=1}^{\infty} U_j(z) < \infty$ *uniformly in* z.
- $\sum_{j=1}^{J} 2^j \mathcal{N}_j = \mathcal{O}(\log n)$ *where* $J = \lfloor \log_2 n \rfloor$.

A consequence of this definition is that the LSW process assumes Y_t has mean zero for all t due to the non-zero mean error process. Analogous to classical Fourier time series analysis, the Evolutionary Wavelet Spectrum, $\{U_j^2(\frac{k}{n})\}_{j=1}^{J}$ characterises the autocovariance structure of Y_t at different scales (frequency bands) and locations. This characterisation is unique up to choice of mother wavelet, $\psi(\cdot)$.

As $\{U_j^2(\frac{k}{n})\}_{j=1}^{J}$ describes the second-order structure of the time series and the time series is assumed to be locally stationary, $U_j^2(\frac{k}{n})$ is constrained to evolve gradually and slowly for each j in order to maintain this local stationarity. Under the definition of a LSW process presented, this is in a piecewise constant manner maintained by the final two conditions. This permits processes exhibiting piecewise second-order structures to be modelled. In general, Nason et al. (2000) only require $U_j^2(\frac{k}{n})$ to be a Lipschitz function for all j instead of the final two conditions presented in the definition above. This consequently means LSW processes can have second-order structures which are not piecewise, thus giving rise to different types of data. However, this thesis will focus on time series exhibiting piecewise covariance structure.

If one specifies a spectrum $\{U_j^2(\frac{k}{n})\}_{j=1}^{J}$, a generating mother wavelet $\psi(\cdot)$ and a parametric distribution for the error process (for example Gaussian), it is possible to simulate data according to the LSW framework. For example Figure 6.6(a) displays a user specified EWS with a piecewise constant power structure (power denotes the contribution to the autocovariance at that particular scale and location). An LSW process instance has been simulated according to this specified EWS as displayed Figure 6.6(b). We observe that power at finer scales of the EWS (higher resolution level) corresponds to higher frequency behaviour in the time series and vice versa for power placed at coarser scales (lower resolution level). Where there is no power at a location at all scales in the EWS, this results in observations being equal to zero in the time series and no variation being present. The important aspect to note however is that a change in the EWS corresponds to a change in autocovariance structure of the simulated time series Y_t. Thus, instead of performing autocovariance CP analysis on the time series, we consider performing CP analysis on the power structure of the EWS, or rather an estimate of the EWS.



(a) Example EWS

(b) LSW data generated from specified EWS

(c) Raw Wavelet Periodogram

(d) Smooth, Corrected, Wavelet Periodogram

Figure 6.6: Example of an Evolutionary Wavelet Spectrum (EWS, 6.6(a)), an LSW process simulated according to the specified EWS (6.6(b)), and the raw and smoothed corrected wavelet periodogram estimates of the EWS (6.6(c) and 6.6(d)). A change in autocovariance structure in the simulated time series corresponds to a change in the EWS and its periodogram estimate.

We are typically not presented with the EWS from which the observed time series has been generated. Instead, we are more commonly presented an observed time series and required to estimate the EWS if the LSW framework is appropriate. An estimate of the EWS can be obtained by considering the square of the empirical detail coefficients from a NDWT on $\{Y_t\}_{t=1}^n$. That is

$$U_j^2\left(\frac{k}{n}\right) \approx I_{j,k} = D_{j,k}^2 = \left(\sum_{t=1}^{n} \psi_{j,k}(t) Y_t\right)^2 . \tag{6.19}$$

This estimate is referred to as the raw wavelet periodogram. For a sequence of random variables $Y_{1:n}$, we denote the corresponding unknown detail coefficients from a NDWT as $\tilde{\mathbf{D}}_{1:n} = (\tilde{\mathbf{D}}_1, \ldots, \tilde{\mathbf{D}}_n), \tilde{\mathbf{D}}_k = \{D_{jk}\}_{j=1}^J$, and the corresponding unknown raw wavelet periodogram as $\mathbf{I}_{1:n} = (\mathbf{I}_1, \ldots, \mathbf{I}_n), \mathbf{I}_k = \{I_{jk}\}_{j=1}^J$. We use their lower case counterparts to denote observed, empirical values of them.

$\mathbf{I}_{1:n}$ and $\tilde{\mathbf{D}}_{1:n}$ can be thought of as a multivariate time series consisting of $J = \lfloor \log_2 n \rfloor$ components at each location with each component denoting a different scale. Due to the use of NDWT and its overlapping wavelets both within and across scales, a dependence structure is present within both of these multivariate time series.

The raw wavelet periodogram for the presented simulated LSW time series is displayed in Figure 6.6(c). We note that the CPs in the observed time series correspond directly to changes in power in scale processes of the periodogram (the same on-off power behaviour is present at resolution level eight and five of the periodogram). Thus in considering changes in autocovariance structure, it is possible to consider changes in the estimated power structure in the scale processes of the periodogram, $\mathbf{I}_{1:n}$. It is this primary idea that forms the main motivation of the methodology proposed in this chapter.

It is worth noting that the raw wavelet periodogram is a biased estimate of the EWS. More specifically,

$$\mathbf{I}_{1:n} = \mathbf{A}\mathbf{U}_{1:n}^2$$

where $\mathbf{U}_{1:n}^2$ is the $J \times n$ matrix representation of the EWS, and $\mathbf{A}$ is a $J \times J$ matrix based on the inner product between the autocorrelation wavelets. An effect of this biased-ness is that a leakage effect occurs in that power at finer scales will diffuse into the coarser scales. Consequently, CPs defined at finer scales may protrude into the coarser scales and in general, the raw wavelet periodogram may not provide an accurate estimate of how the power is truly distributed across scales (see resolution level seven of Figure 6.6(c) for example which exhibits some slight on-off behaviour from resolution level eight). In obtaining an unbiased estimate of the EWS, the corrected periodogram is considered,

$$\tilde{\mathbf{I}}_{1:n} = \mathbf{A}^{-1}\mathbf{I}_{1:n}.$$

Similar to the periodogram estimate of the frequency spectrum in classical Fourier analysis of time series, the raw wavelet periodogram is also not a consistent estimator. Thus in obtaining a consistent unbiased estimate, smoothing (a method of denoising) is typically performed on the raw wavelet periodogram prior to correction by $\mathbf{A}^{-1}$. We refer the reader to Nason et al. (2000) regarding the specifics of the smoothed, corrected periodogram estimate.

The smoothed, corrected version of the raw wavelet periodogram presented previously is displayed in Figure 6.6(d). We observe that the smoothed corrected periodogram estimate is a much more faithful representation of the true spectrum and how the power is truly distributed across the scales. For example, see resolution level eight and five where the on-off power behaviour is much more explicit, and all other resolution levels feature some fluctuation in power, including non-negative power. The smoothed corrected version of the periodogram is therefore often used as an estimate of the EWS with regards to the true spectral structure.

However, with regards to the CP problems of interest in this thesis, we are not necessarily interested in which scale of the EWS the CP may occur, but whether a CP occurs across the scales of the EWS at a location. In turn, accurate estimation of EWS with respect to how the autocovariance structure is decomposed over scales is of little interest compared to whether a change in power across scales occurs at a certain location. As a result, the raw wavelet periodogram can be considered in CP analysis as it still permits accurate CP detection with respect to the time location, but does not indicate the true spectral power structure. This latter point is usually not of interest when considering CPs in the observed time series. The proposed methodology of this chapter thus focuses on analysis of the raw wavelet periodogram as opposed to the smoothed, corrected periodogram. This has many advantages for our analysis including the fact that an explicit distribution for the raw wavelet periodogram can be computed which is more difficult for the smoothed, corrected periodogram. The raw wavelet periodogram is also analysed in the CP approach proposed by Cho and Fryzlewicz (2012).

The LSW framework is a powerful tool in modelling locally stationary time series with time varying autocovariance structure, and can provide an alternative wavelet representation for time domain models such as piecewise MA processes. A natural question to pose therefore is whether it is possible to consider a HMM framework within the LSW framework. Such a proposed hybrid framework may allow us to consider changes in autocovariance structure more actively, whilst utilising a large multitude of existing HMM based CP methods. This includes the HMM based CP method proposed in Chapter 3 in quantifying the uncertainty of CPs. This LSW-HMM framework may thus allow us to consider the uncertainty of autocovariance CPs, an area which has received little to no attention.

6.4 Methodology

As previously described, our goal is to quantify the uncertainty of autocovariance CPs for a time series by considering its spectral structure. Quantities of interest include the CP probability $P(\tau \ni t|y_{1:n})$ (CPP, the probability of a CP at time t), and the distribution of number of CPs within the observed time series $P(M = m|y_{1:n})$. Other CP characteristics such as joint or conditional CP distributions are also available using the proposed methodology.

As detailed in Section 6.3.3, the raw wavelet periodogram characterises how the autocovariance structure of a time series evolves over time if a LSW process is assumed. Consequently, we perform analysis on the periodogram to quantify the uncertainty of autocovariance CPs. This is achieved by modelling the periodogram via a HMM framework, and quantifying the CP uncertainty via the existing HMM approach proposed in (Nam et al., 2012b) and Chapter 3. In proposing the new methodology, several challenges need to be addressed.

Firstly, the multivariate joint density of $\mathbf{I}_k$ is unknown and needs to be derived. This density captures the dependence structure introduced by the use of the NDWT in estimating the periodogram. The derivation of this joint density and its embedding in a HMM modelling framework is detailed in Section 6.4.1. As the model parameters, θ, associated with the HMM framework are unknown, these need to be estimated and we turn to Sequential Monte Carlo samplers (SMC, Del Moral et al. (2006)) in considering the posterior of the parameters as in Chapters 3 and 5. These model parameters can be shown to be directly associated with the EWS. An example SMC implementation is provided in Section 6.4.2. Section 6.4.3 details some aspects concerning the computation of the distribution of CP characteristics. Section 6.4.4 provides an outline of the overall proposed approach.

There are many advantages to considering the observed time series under the LSW framework. In particular, time series exhibiting piecewise second-order structure can be more readily analysed under this framework compared to a time-domain approach. For example, for a piecewise moving average processes, the associated EWS has a piecewise constant structure at each scale; a sparser representation where the discontinuities can be analysed with fewer issues potentially arising from changes in mean methods. This sparser representation is not possible in the time-domain.

By combining the use of wavelets in conjunction with an HMM framework, we can systematically induce a dependence structure in the HMM framework by selecting a suitable number of scale process of the periodogram to analyse, compared to choosing an arbitrary dependence structure in a time-domain approximation.

We assume in this chapter that the error process in the LSW model is Gaussian, that is $\xi_{j,k} \overset{\text{iid}}{\sim} \text{N}(0,1)$. This leads to Y_t being Gaussian itself and is commonly referred to as a Gaussian LSW process. Recall from Section 6.3.3 that our EWS is piecewise constant. That is,

$$U_j^2\left(\frac{k}{n}\right) = \sum_{s=1}^{H^*} u_{j,s}^2 \mathbf{1}_{\mathcal{U}_s}(k) \qquad j = 1, \ldots, J, \tag{6.20}$$

where $u_{j,s}^2$ are some unknown constants, and $\mathcal{U}_s, s = 1, \ldots, H^*$ is an unknown disjoint partitioning of $1, \ldots, n$ over all scales j simultaneously. Each $\mathcal{U}_s$ has a particular EWS power structure associated with it, such that consecutive $\mathcal{U}_s$ have changes in power in at least one scale. H^* denotes the unknown number of partitions there are in the EWS, and ultimately correspond to the segments in the data and in turn the number of CPs.

We now propose the LSW-HMM modelling framework in quantifying the uncertainty of autocovariance CPs under the assumptions outlined above.

6.4.1 LSW-HMM modelling framework

Recall that the raw wavelet periodogram, an estimate of the EWS, is provided by the square of the empirical wavelet coefficients under a NDWT,

$$U_j^2\left(\frac{k}{n}\right) \approx I_{j,k} = D_{j,k}^2 = \left(\sum_{t=1}^{n} \psi_{j,k}(t) Y_t\right)^2. \tag{6.21}$$

We consider modelling the raw wavelet periodogram at a single location k over the different scales j. We adopt the convention that $j = 1$ is the finest scale, and $j = 2, \ldots, J$ as the subsequent coarser scales (where $J = \lfloor \log_2 n \rfloor$). Within-scale dependence induced by the NDWT can be accounted for by the HMM framework.

We refer to the collection of J periodogram coefficients at a particular time point as $\mathbf{I}_k = \{I_{j,k}\}_{j=1,\ldots,J}$ (random variable) and $\tilde{\mathbf{d}}_k^2 = \{d_{j,k}^2\}_{j=1,\ldots,J}$ (observed, empirical) from here onwards.

Note that under the definition of the transform above, the wavelet coefficients at location k are a function of observations in the future. For example, the above equation can be rewritten as

$$\begin{aligned} D_{1,k} &= \frac{1}{\sqrt{2}}(Y_{k+1} - Y_k) \qquad 1 \leq k \leq (n-1) \\ D_{1,n} &= \frac{1}{\sqrt{2}}(Y_1 - Y_n) \end{aligned}$$

for scale one of the Haar wavelet. However, it is also possible to re-write the above equation and transform in terms of past observations by relabelling the time label. Namely $D_{1,k} \leftarrow D_{1,k+1}$ for scale one of the Haar wavelet. A similar relabelling procedure exists for other scales and other wavelets.

We next turn to deriving the joint density of $\mathbf{I}_k$.

Distribution of $\mathbf{I}_k$

Recall that since we have assumed an LSW model and Gaussian innovations, Y_t is Gaussian with mean zero. By performing a wavelet transform, the wavelet coefficients $D_{j,k}$ are Gaussian distributed themselves with mean zero. The use of NDWT however induces a dependence structure between the coefficients $D_{j,k}$. We consider in particular, $\tilde{\mathbf{D}}_k = \{D_{j,k}\}_{j=1,\ldots,J}$, the coefficients across J scales considered at a given location, k. Thus,

$$\tilde{\mathbf{D}}_k \sim \mathrm{MVN}(\mathbf{0}, \Sigma_k^D) \qquad k = 1, \ldots, n,$$

where Σ_k^D specifies the covariance structure between the wavelet coefficients at location k across the J scales considered. The subsection below discusses how Σ_k^D can be computed from the spectrum $U_j^2(\frac{k}{n})$.

As $\mathbf{I}_k = \tilde{\mathbf{D}}_k^2 = (D_{1,k}^2, \ldots, D_{J,k}^2)$, the following result can be established.

Proposition 1. *The density of* $\mathbf{I}_k$ *is,*

$$\begin{aligned} g(\tilde{\mathbf{d}}_k^2 | \Sigma_k^D) &= g(d_{1,k}^2, \ldots, d_{J,k}^2 | \Sigma_k^D) \\ &= \frac{1}{2^J \prod_{j=1}^J |d_{j,k}|} \sum_{a_1,\ldots,a_J=\{+,-\}} f\left(a_1|d_{1,k}|, \ldots, a_J|d_{J,k}| \middle| \mathbf{0}, \Sigma_k^D\right), \end{aligned} \tag{6.22}$$

where $f(\cdot|\mathbf{0}, \Sigma_k^D)$ *is the joint density corresponding to* $\mathrm{MVN}(\mathbf{0}, \Sigma_k^D)$*.*

Proof. This is based on a change of variables argument detailed further in Section 6.A.1 (page 200). □

We can thus use the joint density of wavelet coefficients, $\tilde{\mathbf{D}}_k$, to deduce the joint density for the squared wavelet coefficients $\mathbf{I}_k = \tilde{\mathbf{D}}_k^2$. A similar joint density can be computed if we consider each scale process of the periodogram across all time locations, that is $\mathbf{I}^j = \{I_{j,k}\}_{k=1}^n$, although the order of computation increases exponentially.

Computing Σ_k^D

We next turn to the problem of accounting for the dependence between the wavelet coefficients, induced by a NDWT. This dependence structure feeds into the joint densities of $\tilde{\mathbf{D}}_k$ and $\mathbf{I}_k$. Recall that the EWS characterises the autocovariance structure of the observation process for any orthonormal incremental process as follows (Nason et al., 2000):

$$\mathrm{Cov}(Y_t, Y_{t-v}) = \sum_l \sum_m U_l^2 \left(\frac{m}{n}\right) \psi_{l,m}(t)\psi_{l,m}(t-v).$$

It is possible to compute this autocovariance quantity without knowing the entire EWS due to the compact support of wavelets, that is the product $\psi_{l,m}(t)\psi_{l,m}(t-v)$ will only be non-zero for some values of v.

As the following proposition demonstrates, the autocovariance structure of the observations also feeds into the covariance structure of the wavelet coefficients.

Proposition 2. *For a LSW process, the covariance structure between the pair of wavelet coefficients, $D_{j,k}$ and $D_{j',k'}$, of a NDWT is of the following form:*

$$\mathrm{Cov}(D_{j,k}, D_{j',k'}) = \sum_t \sum_v \psi_{j,k}(t)\psi_{j',k'}(t-v)\mathrm{Cov}(Y_t, Y_{t-v}). \tag{6.23}$$

Proof. See Section 6.A.2 (page 202). □

We can thus deduce the covariance structure for the wavelet coefficients $\tilde{\mathbf{D}}_k$, Σ_k^D, from the EWS. Similar to the autocovariance structure of the observation series, only a finite number of covariances in the summation are needed to evaluate Σ_k^D due to the compact support property associated with wavelets. Consequently, the entire EWS does not need to be known to calculate the covariance between the wavelet coefficients of $\tilde{\mathbf{D}}_k$.

More specifically, one can show that to compute Σ_k^D, the covariance structure of the wavelet coefficients at location k, the power from locations $k-2(\mathcal{L}_j-1),\ldots,k$ for scale $j=1,\ldots,J$ needs to be recorded where $\mathcal{L}_j$ denotes the number of non-zero filter elements in the wavelet at scale j (see Section 6.A.3, page 203).

Example 6.4.1. Computing Σ_k^D when provided with the EWS, $U_j^2(\frac{k}{n})$.
We assume $n=4$ and the Haar mother wavelet as the generating wavelet in this example. Thus $J=2$ and recall that the wavelet vectors for the two scales are $\psi_1=(\overset{\mathbf{0}}{\ldots},\frac{1}{\sqrt{2}},-\frac{1}{\sqrt{2}},\overset{\mathbf{0}}{\ldots})$ and $\psi_2=(\overset{\mathbf{0}}{\ldots},\frac{1}{2},\frac{1}{2},-\frac{1}{2},-\frac{1}{2},\overset{\mathbf{0}}{\ldots})$. Let the pre-specified EWS have the following matrix form.

$$\mathbf{U}_{1:4}^2=\{u_{j,k}^2\}_{j=1,2,\,k=1,2,3,4}=\begin{bmatrix}1&1&2&2\\2&2&2&2\end{bmatrix}$$

We begin by computing the variances and covariances between observations (that is $\mathrm{Cov}(Y_t,Y_{t-v})$) which feed in directly to the covariance of the coefficients, $D_{j,k}$. Due to the finite support of wavelets, it is only necessary to consider $0\geq v\geq 3$. Hence, using Equation 6.23,

$$\begin{aligned}\mathrm{Var}(Y_1)&=\overbrace{u_{1,1}^2\psi_{1,1}^2+u_{1,2}^2\psi_{1,2}^2}^{\text{Scale 1}}+\overbrace{u_{2,1}^2\psi_{2,1}^2+u_{2,2}^2\psi_{2,2}^2+u_{2,3}^2\psi_{2,3}^2+u_{2,4}^2\psi_{2,4}^2}^{\text{Scale 1}}\\&=1\cdot\left(\frac{1}{\sqrt{2}}\right)^2+1\cdot\left(-\frac{1}{\sqrt{2}}\right)^2+2\cdot\left(\frac{1}{2}\right)^2+2\cdot\left(\frac{1}{2}\right)^2+2\cdot\left(-\frac{1}{2}\right)^2+2\cdot\left(-\frac{1}{2}\right)^2\\&=1+2=3.\end{aligned}$$

Similarly,

$$\mathrm{Var}(Y_2)=3.5\qquad\mathrm{Var}(Y_3)=4\qquad\mathrm{Var}(Y_4)=3.5,$$

recalling that the $\mathbf{U}_{1:4}^2$ loops round to the beginning when $k\to n=4$. Similarly the other covariances are computed as follows,

$$\begin{aligned}\mathrm{Cov}(Y_1,Y_2)&=\overbrace{u_{1,1}^2\psi_{1,1}\psi_{1,2}}^{\text{Scale 1}}+\overbrace{u_{2,1}^2\psi_{2,1}\psi_{2,2}+u_{2,2}^2\psi_{2,2}\psi_{2,3}+u_{2,3}^2\psi_{2,3}\psi_{2,4}}^{\text{Scale 2}}\\&=1\cdot\frac{1}{\sqrt{2}}\cdot-\frac{1}{\sqrt{2}}+2\cdot\frac{1}{2}\cdot\frac{1}{2}+2\cdot\frac{1}{2}\cdot-\frac{1}{2}+2\cdot-\frac{1}{2}\cdot-\frac{1}{2}\\&=-0.5+0.5-0.5+0.5=0.\end{aligned}$$

$$\mathrm{Cov}(Y_2,Y_3)=0\qquad\mathrm{Cov}(Y_3,Y_4)=-0.5$$

$$\begin{aligned}
\mathrm{Cov}(Y_1, Y_3) &= \overbrace{u_{2,1}^2 \psi_{2,1}\psi_{2,3} + u_{2,2}^2 \psi_{2,2}\psi_{2,4}}^{\text{Scale 2}} \\
&= 2 \cdot \frac{1}{2} \cdot -\frac{1}{2} + 2 \cdot \frac{1}{2} \cdot -\frac{1}{2} \\
&= -0.5 - 0.5 = -1 \\
\mathrm{Cov}(Y_2, Y_4) &= -1 \\
\mathrm{Cov}(Y_1, Y_4) &= u_{2,1}^2 \cdot \psi_{2,1} \cdot \psi_{2,4} = -0.5.
\end{aligned}$$

These variances and covariances regarding the observations consequently feed into the computation of the variance and covariances between the wavelet coefficients as demonstrated in Proposition 2. For example,

$$\begin{aligned}
\mathrm{Var}(D_{1,1}) &= \psi_{1,1}^2 \mathrm{Var}(Y_1) + \psi_{1,2}^2 \mathrm{Var}(Y_2) + 2\psi_{1,1}\psi_{1,2}\mathrm{Cov}(Y_1, Y_2) \\
&= \frac{1}{2} \cdot 3 + \frac{1}{2} \cdot 3.5 + 2 \cdot -\frac{1}{2} \cdot 0 = \frac{13}{4} \\
\mathrm{Var}(D_{2,1}) &= \psi_{2,1}^2 \mathrm{Var}(Y_1) + \psi_{2,2}^2 \mathrm{Var}(Y_2) + \psi_{2,3}^2 \mathrm{Var}(Y_3) + \psi_{2,4}^2 \mathrm{Var}(Y_4) \\
&+ 2\left[\psi_{2,1}\psi_{2,2}\mathrm{Cov}(Y_1, Y_2) + \psi_{2,2}\psi_{2,3}\mathrm{Cov}(Y_2, Y_3) + \psi_{2,3}\psi_{2,4}\mathrm{Cov}(Y_3, Y_4)\right. \\
&\left. + \psi_{2,1}\psi_{2,3}\mathrm{Cov}(Y_1, Y_3) + \psi_{2,2}\psi_{2,4}\mathrm{Cov}(Y_2, Y_4) + \psi_{2,1}\psi_{2,4}\mathrm{Cov}(Y_1, Y_4)\right] \\
&= \frac{1}{4} \cdot 3 + \frac{1}{4} \cdot \frac{7}{2} + \frac{1}{4} \cdot 4 + \frac{1}{4} \cdot \frac{7}{2} \\
&\qquad + 2\left[\frac{1}{4} \cdot 0 + \frac{-1}{4} \cdot 0 + \frac{1}{4} \cdot \frac{-1}{2} + \frac{-1}{4} \cdot -1 + \frac{-1}{4} \cdot -1 + \frac{-1}{4} \cdot -\frac{1}{2}\right] \\
&= \frac{7}{2} + \frac{1}{2} = 4 \\
\mathrm{Cov}(D_{1,1}, D_{2,1}) &= \psi_{1,1}\psi_{2,1}\mathrm{Var}(Y_1) + \psi_{1,1}\psi_{2,2}\mathrm{Cov}(Y_1, Y_2) + \psi_{1,1}\psi_{2,3}\mathrm{Cov}(Y_1, Y_3) \\
&+ \psi_{1,1}\psi_{2,4}\mathrm{Cov}(Y_1, Y_4) + \psi_{1,2}\psi_{2,2}\mathrm{Var}(Y_2) + \psi_{1,2}\psi_{2,3}\mathrm{Cov}(Y_2, Y_3) + \psi_{1,2}\psi_{2,4}\mathrm{Cov}(Y_2, Y_4 \\
&= \frac{1}{2\sqrt{2}} \cdot 3 + \frac{1}{2\sqrt{2}} \cdot 0 + \frac{-1}{2\sqrt{2}} \cdot -1 \\
&+ \frac{-1}{2\sqrt{2}} \cdot \frac{-1}{2} + \frac{-1}{2\sqrt{2}} \cdot \frac{7}{2} + \frac{1}{2\sqrt{2}} \cdot 0 + \frac{1}{2\sqrt{2}} \cdot -1 \\
&= \frac{1}{4\sqrt{2}}.
\end{aligned}$$

Thus for $k = 1$, the corresponding covariance matrix between the coefficients is

$$\Sigma_1^D = \begin{bmatrix} \frac{13}{4} & \frac{1}{4\sqrt{2}} \\ \frac{1}{4\sqrt{2}} & 4 \end{bmatrix}.$$

The HMM framework

Having derived a joint density for the wavelet periodogram, we now turn our attention to the question of how this can be incorporated appropriately within a HMM framework. The J multivariate scale processes from a raw wavelet periodogram can be modelled simultaneously via a single HMM framework with a multivariate emission density. That is, at location k, we consider $\mathbf{I}_k = \{I_{j,k}\}_{j=1,\dots,J}$, and model it as being dependent on a single underlying, unobserved Markov chain (MC), X_k, which takes values from $\Omega_X = \{1, \dots, H\}$ with $H = |\Omega_X| < \infty$,

$$p(x_k|x_{1:k-1}, \theta) = p(x_k|x_{k-1}, \theta) \qquad k = 1, \dots, n \quad \text{(Transition)}$$

$$\mathbf{I}_k|\{X_{1:k-1}, \mathbf{I}_{1:k-1} = \tilde{\mathbf{d}}^2_{1:k-1}\} \sim g(\mathbf{I}_k = \tilde{\mathbf{d}}^2_k|x_{k-2(\mathcal{L}_J-1):k}, \theta) \qquad k = 1, \dots, n \quad \text{(Emission)}$$

The HMM framework assumes that the emission density of $\mathbf{I}_k$ is determined by the latent process X_k, such that the process follows the Markov property and the $\mathbf{I}_{1:n}$ are conditionally independent given $X_{1:n}$. This latter remark allows us to account for some of the within-scale dependence induced by a NDWT via the underlying MC. H denotes the number of underlying states the latent MC, X_k, can take and corresponds to different data generating mechanisms, for example "stormy" and "non-stormy" seasons in the motivating oceanographic application. Under our setup, this corresponds to the number of unique power configurations over the disjoint partitioning $\mathcal{U}_1, \dots, \mathcal{U}_{H^*}$. That is $H \leq H^*$ is the number of states that generate the H^* partitions, with some partitions possibly being generated by the same state. We assume in our analysis that H is known a priori, as we want to give a specific interpretation to the states in the application, that of "stormy" or "non-stormy" seasons. However as discussed in Chapter 5 and Zhou et al. (2012); Nam et al. (2012a), the number of states can be deduced via the existing use of SMC samplers and we examine this assumption in Section 6.5.2 with regards to our oceanographic application. We assume that the underlying unobserved MC, X_k, is first order Markov, although extensions to a higher order Markov Chain are permitted via the use of embedding arguments.

The state-dependent emission density, $g(\mathbf{I}_k|X_{k-2(\mathcal{L}_J-1):k})$, is that proposed in Equation 6.22, with the covariance structure Σ^D_k being dependent on $X_{k-2(\mathcal{L}_J-1):k}$. Rather than estimating entries of Σ^D_k directly, we instead estimate the powers, $u^2_{j,s}$ as in Equation 6.20, that feed directly into and populate Σ^D_k. More specifically, we estimate state-dependent powers $u_{j,s}^2$ in

$$U_{j,X_k}^2\left(\frac{k}{n}\right) = \sum_{s=1}^{H} u_{j,s}^2 \mathbf{1}_{[X_k=s]} \qquad j = 1, \ldots, J. \tag{6.24}$$

This state-dependent power structure is equivalent to the piecewise constant EWS as in Equation 6.20. As X_k is permitted to move freely between all states of Ω_X, we are able to reduce the summation limit in Equation 6.20 to H from H^*. Returning to previous power configurations in the EWS is therefore possible, with a change in state corresponding to a change in power in at least one scale. Σ_k^D is dependent on the underlying states of X_k from times $k - 2(\mathcal{L}_J - 1), \ldots, k$ (see Section 6.A.3) and thus the order of the HMM is $2\mathcal{L}_J - 1$. We highlight again that in the case when future observations or observations are considered and thus future states of the underlying MC, it is always possible to relabel the time order such that it depends only past quantities.

Here, θ denotes the model parameters that need to be estimated which consists of the transition matrix $\mathbf{P}$ and the aforementioned state-dependent power $U^2 = \{U_{\cdot,1}^2, \ldots, U_{\cdot,H}^2\}$, where $U_{\cdot,s}^2 = \{u_{j,s}^2\}_{j=1}^J$ for all $s \in \Omega_X$, is associated with the emission density. We can thus partition the model parameters into transition and emission parameters, $\theta = (\mathbf{P}, U^2)$. As θ is unknown, we turn to SMC samplers (Del Moral et al., 2006) for their estimation.

6.4.2 SMC samplers implementation

This section outlines an example SMC implementation in approximating the parameter posterior, $p(\theta|\tilde{\mathbf{d}}_{1:n}^2, H)$ via a weighted cloud of N particles, $\{\theta^i, W^i|H\}_{i=1}^N$, since $\theta = (\mathbf{P}, U^2)$ is unknown. As highlighted in Section 3.2.2 (page 67), SMC samplers provide an algorithm to sample from a sequence of connected distributions via importance sampling and resampling techniques (Del Moral et al., 2006). Analogous to the sequence of distributions defined in Chapter 3 and 5, we can define the following sequence of distributions,

$$\pi_b(\theta) \propto l(\theta|\tilde{\mathbf{d}}_{1:n}^2, H)^{\gamma_b} p(\theta|H) \qquad b = 1, \ldots, B, \tag{6.25}$$

where $l(\theta|\tilde{\mathbf{d}}_{1:n}^2, H)$ denotes the likelihood with respect to the periodogram, and $p(\theta|H)$ as the prior of the model parameters. $\{\gamma_b\}_{b=1}^B$ denotes a non-decreasing tempering schedule such that $\gamma_1 = 0$ and $\gamma_B = 1$. As with the other sequence of distributions sampled via SMC samplers in this thesis, this sequence of distributions similarly does not need the latent state sequence to be sampled. This is because under the HMM framework, the likelihood can be computed exactly via the Forward-Backward Equations (Baum et al., 1970) which does not require sampling the latent state sequence. This has many advantages, including a reduction in Monte Carlo sampling error.

Section 6.A.5 (page 204) provides a detailed outline of the example SMC implementation used within our framework. We proceed in a similar fashion as that presented in Section 3.4.1 and Chapter 5 in that we consider the transition probability row vectors, $p_s, s = 1, \ldots, H$ forming the transition matrix $\mathbf{P}$, independently from the inverse state dependent powers $\frac{1}{u_{j,s}^2}, j = 1, \ldots, J, s = 1, \ldots, H$. The re-parametrisation of the state dependent powers to its inverse is analogous to the re-parametrisation of variance to precision (inverse variance) in typical time-domain models (see Section 3.4.1 for example). In practice, the series we consider will all contain at least a small portion of variation, and as such issues regarding zero or infinite power for particular frequencies will not arise.

We initialise by sampling from a Dirichlet and Gamma prior distribution respectively for transition probability vectors and inverse state dependent powers, and mutate according to a Random Walk Metropolis Hastings Markov kernel on the appropriate domain for each component, namely on the logit scale for the transition probability vectors since they are non-negative and must sum to one, and on the log-scale for the non-negative inverse powers. There is a great deal of flexibility within the SMC samplers framework with regards to the type of mutation and sampling schemes from the prior. The example implementation presented is in no way the only implementation or optimal with respect to optimising mixing and acceptance rates. However, this design provides results which appear sensible without a great deal of manual tuning.

6.4.3 Exact CP distributions

Having formulated an appropriate HMM framework to model the periodogram $\tilde{\mathbf{d}}_{1:n}^2$, and accounting for unknown θ via SMC samplers, it is now possible to compute the CP distributions of interest. As detailed in Section 3.2.1 (page 60), it is possible to compute exact CP distributions, such as $P(\tau^{(k_{\mathrm{CP}})} \ni t | \tilde{\mathbf{d}}_{1:n}^2, \theta, H)$, conditional on θ via Finite Markov Chain Imbedding (FMCI) in a HMM framework (see Aston et al. (2011) and references therein). In particular, the use of the k_{CP} and k'_{CP} variables under the generalised CP definition denoting the sustained nature of regimes, also correspond to the sustained nature of the EWS such that it evolves gradually to maintain local stationarity. This sustained nature also has an intuitive interpretation in the oceanographic application with a storm season deemed to be in progresses when there are several consecutive “stormy” measurements.

6.4.4 Outline of Approach

An outline of the final algorithm is as follows:

1. Perform a NDWT to time series $y_{1:n}, n = 2^J, J \in \mathbb{N}$ to obtain the wavelet periodogram. Let $\tilde{\mathbf{d}}^2_{1:n}$ denote the periodogram, a J multivariate time series.

2. Assuming H underlying states, model $\tilde{\mathbf{d}}^2_{1:n}$ by a HMM framework with the corresponding joint emission density (Equation 6.22). This joint density also accounts for the dependence structure between scale processes.

3. Account for the uncertainty of the unknown HMM model parameters, θ, via Sequential Monte Carlo samplers. This results in approximating the posterior, $p(\theta|\tilde{\mathbf{d}}^2_{1:n}, H)$, by a weighted cloud of N particles $\{\theta^i, W^i|H\}_{i=1}^N$.

4. To obtain the CP probability of interest, approximate as follows. Let k_{CP} denotes the sustained condition under a generalised CP definition (see Chapter 3), and $P(\tau^{(k_{\text{CP}})} \ni t|\mathbf{d}^2_{1:n}, \theta^i, H)$ to be the exact CP distribution conditional on θ^i. Then the CP probability is,

$$P(\tau \ni t|y_{1:n}) \equiv P(\tau^{(k_{\text{CP}})} \ni t|\tilde{\mathbf{d}}^2_{1:n}, H) \approx \sum_{i=1}^{N} W^i P(\tau^{(k_{\text{CP}})} \ni t|\tilde{\mathbf{d}}^2_{1:n}, \theta^i, H). \tag{6.26}$$

   That is, the weighted average of conditional exact CP distributions with respect to different model parameter configurations.
   $P(M = m|y_{1:n}) \equiv P(M^{(k_{\text{CP}})} = m|\tilde{\mathbf{d}}^2_{1:n}, H)$ follows analogously.

Computationally, it is not possible to consider all J scales of the periodogram as the order of the HMM increases exponentially and the intended Markovian structure becomes lost (see Section 6.A.4, page 204 for further details). Consequently, we approximate the periodogram by considering $J^* \leq J$ finer scales of the periodogram, a common approach in time series analysis (see for example Cho and Fryzlewicz (2012)). This restricts our attention to changes in autocovariance structure associated at higher frequencies which seems more appropriate in the oceanographic data of interest. This should therefore not hinder our proposed methodology with regards to the motivating oceanographic application.

We assume that the choice of analysing wavelet used for the transform is known a priori, and is the same as the generating wavelet. However, this is often unknown and we note that wavelet choice is an area of ongoing interest with the effect between differing generating and analysing wavelets for EWS estimation investigated in Gott and Eckley (2013).

The assumption of dyadic lengthed data, $n = 2^J, J \in \mathbb{N}$, is typically unrealistic in real statistical applications. We stress that this assumption is made throughout the wavelet literature and is only required when performing the NDWT in order to estimate the EWS. It is not required for the proposed HMM-based modelling framework. To address scenarios where this assumption is not satisfied, we propose performing two common approaches in the wavelet literature; truncate the data so that it becomes dyadic in length, or "pad out" the beginning and/or end of the data with white noise or a constant so it becomes dyadic in length (Ogden, 1997, p. 116). In the latter case, it suffices to only analyse the part of the periodogram which corresponds to the original data for CP analysis under the proposed LSW-HMM framework. Future work may want to explore the use of maximal overlap discrete wavelet transform (MODWT) (see Whitcher et al. (2000) and Choi et al. (2008)), where the dyadic length assumption is not required. However, such a transform remains undeveloped with respect to the established LSW framework. Alternatively, if only $J^* \leq J$ scales are of interest, then it is possible to relax the dyadic length restriction and consider time series of length $n = C \cdot 2^{J^*}$ where $C \in \mathbb{N}_+$, due to the J^* scale wavelets having smaller support than th J scale wavelets. Nevertheless, the datasets considered in this chapter are of dyadic length.

6.5 Results and Applications

We next consider the performance of our proposed methodology on both simulated and oceanographic data.

We first consider simulated white noise and MA processes with piecewise second-order structures. White noise processes are considered and compared to a time-domain HMM approach because this type of process can be modelled exactly in the time-domain with no approximation being necessary. Hence our proposed wavelet method should compliment it. The potential benefit of the proposed wavelet approach is then demonstrated on piecewise MA processes in which an exact time-domain HMM is not possible without some sort of approximation taking place.

We also return to the oceanographic application concerned with determining changes in storm season from wave height data. In addition to quantifying the uncertainty of storm season changes, we demonstrate concurrence with estimates provided by other autocovariance CP methods and those provided by expert oceanographers.

The `R` package `wavethresh` (Nason, 2012) has been used to obtain the raw wavelet periodogram in analysis.

6.5.1 Simulated Data

We consider simulated processes of length 512 and with defined CPs (red dotted lines at times 151, 301 and 451). We initially compare our proposed method to a time-domain Gaussian Markov Mixture model on the time series itself, regardless of how the data is actually generated and statistical features present. In the case of the piecewise MA data, such a model mis-specification is a possible approximation. We assume state dependent means and variances under this time domain model, that is $Y_t|X_t \sim \mathrm{N}(\mu_{X_t}, \sigma^2_{X_t})$.

In generating our results, the following SMC samplers settings have been used; $N = 500$ samples to approximate the defined sequence of $B = 100$ distributions. The hyperparameter for the s-th transition probability vector, α_s, is a H-long vector of ones with 10 in the s-th position which encourages the underlying MC to remain in the same state. The shape and scale hyperparameters for the inverse power parameters priors are $\alpha_\lambda = 1$ and $\beta_\lambda = 1$ respectively. These hyperparameters have been arbitrarily set. A linear tempering schedule, that is $\gamma_b = \frac{b-1}{B-1}, b = 1, \ldots, B$, and a baseline line proposal variance of ten which decreases linearly with respect to the iteration of the sampler, are utilised.

The simulated data considered arises from two possible generating mechanisms in the time-domain, and we thus assume $H = 2$ in our HMM framework, and $k_{\mathrm{CP}} = 20, k'_{\mathrm{CP}} = 10$ for the required sustained change in state under our CP definition. $J^* = 3$ scale processes of the periodogram under a Haar LSW framework are considered, a computationally efficient setting under the conditions presented.

In the case of the time-domain Gaussian Markov Mixture, the following priors are considered in the SMC implementation: $\mu_s \overset{\mathrm{iid}}{\sim} \mathrm{N}(0, 10), \frac{1}{\sigma_s^2} \overset{\mathrm{iid}}{\sim} \mathrm{Gamma}(\mathrm{shape} = 1, \mathrm{scale} = 1), s = 1, 2$.

Gaussian White Noise Processes with Switches in Variance

The following experiment concerns independent Gaussian data which exhibits a change in variance at defined time points. It is well known that the corresponding true EWS is $U_j^2(\frac{k}{n}) = \frac{\sigma^2_k}{2^j}, j = 1, \ldots, J$. A change in variance thus causes a change in power across all scales simultaneously. The corresponding EWS for such data is presented in Figure 6.7(a). This type of data can be modelled exactly in the time domain via a Gaussian Markov Mixture model. A realisation of the data and corresponding CP analysis are displayed in Figure 6.8. The top panel is a plot of the simulated data analysed. The second and third panel display the CPP plot under the wavelet and time-domain approaches respectively. The fourth panel presents the distribution of the number of CPs from both approaches.

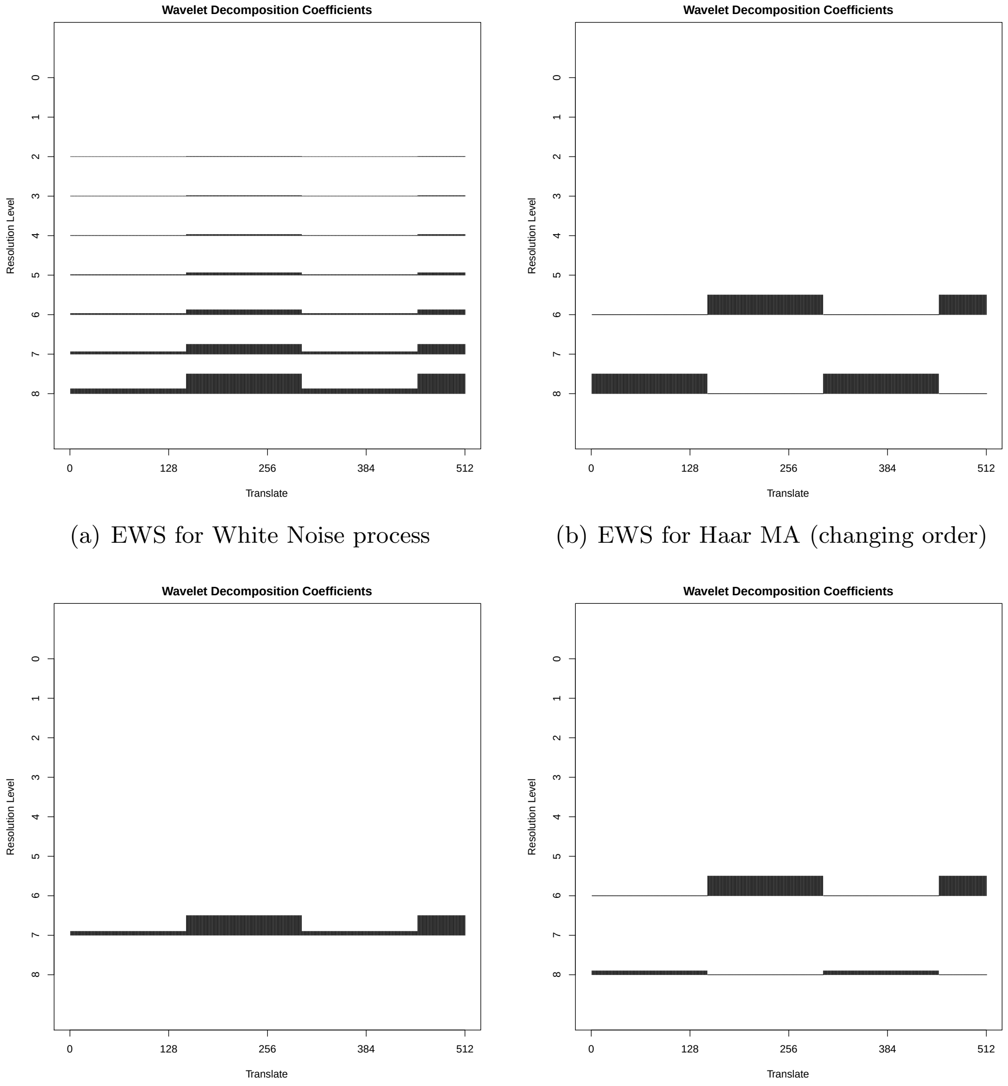
We observe that our proposed methodology has peaked and centred CPP around the defined CP locations and provides similar results to the time-domain approach. This type of CPP behaviour provides a potential indication of the CP location estimates. In some instances, the wavelet approach outperforms the time-domain approach, for example the CP associated with time point 301 is more certain. We note that there is some significant CPP assigned to the first few time points under the wavelet approach. This arises due to a label identifiability issue common with HMMs (see Scott (2002), states are identifiable up to the permutation of them). As such, an additional CP is often detected at the start of the data and this is reflected in the CP distribution. Disregarding this artefact, we observe that three CPs occurring is almost certain under the wavelet approach. This is in accordance with the time-domain approach and truth.

The results demonstrate that there is potential in providing an alternative method when dealing with this type of data as the wavelet based method identifies CPs near the defined locations. However some differences and discrepancies do exist between the proposed wavelet approach, the truth and time-domain approach. In particular, the CPP under the proposed approach is slightly offset from the truth. However, these estimates are still in line with what we might observe in the time series realisation and compares favourably to the time-domain approach.

Piecewise MA processes - Piecewise Haar MA processes

The following scenario considers piecewise MA processes with changing MA order, variance and both simultaneously. We consider in particular piecewise Haar MA processes where the coefficients of the MA process are the Haar wavelet coefficients with a piecewise constant power structure in the EWS being present. Such processes are the types of data that our proposed methodology should perform well on and for which time-domain HMM methods require some approximation. In this case, we approximate the observed time series by modelling it as Gaussian Markov Mixture model, ignoring any of the autocorrelation present in the time series. This incorrect modelling approach is also equally applicable when dealing with real data where the "true" model is unknown. We later account for the autocorrelation present in the series by introducing some AR structure (page 188)

Stationary Haar MA processes have constant power structure in a single



(a) EWS for White Noise process

(b) EWS for Haar MA (changing order)

(c) EWS for Haar MA (changing variance)

(d) EWS for Haar MA (changing order and variance)

Figure 6.7: Corresponding Evolutionary Wavelet Spectrum for the Simulated White Noise (6.7(a)) and Haar Moving Average processes (6.7(b) – 6.7(d)) considered in Section 6.5.1.

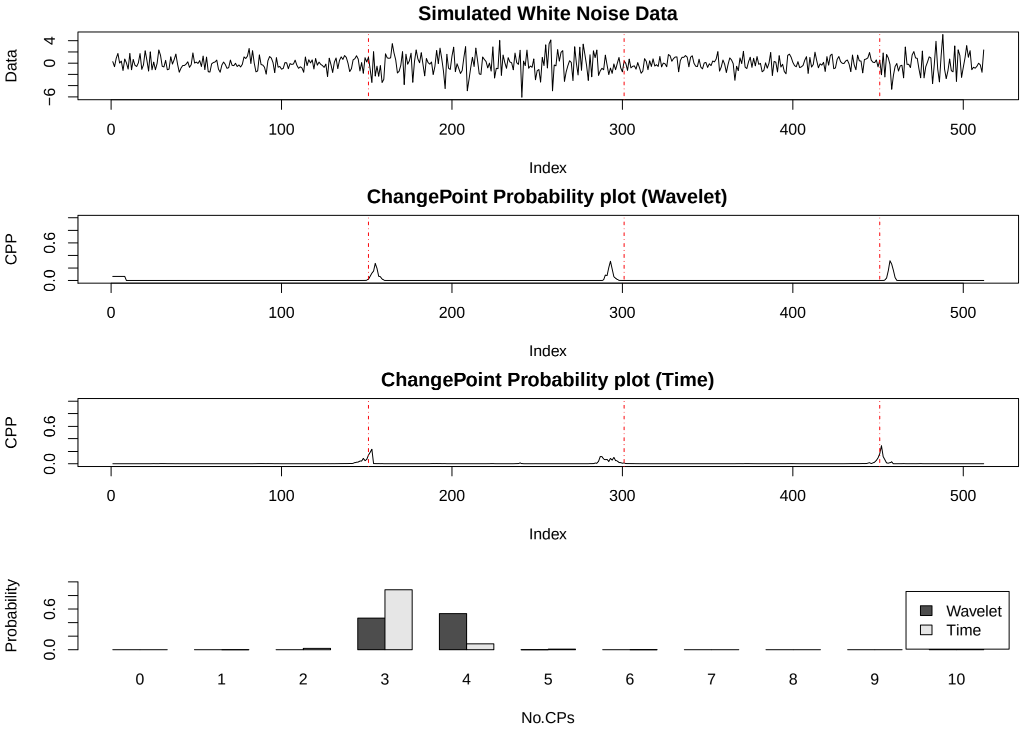


Figure 6.8: Changepoint results (CPP plot and distribution of number of CPs) for simulated Gaussian white noise data with a change in variance (1 and 4, see Figure 6.7(a) for corresponding EWS). 1st panel presents the simulated data analysed. 2nd and 3rd panel displays the CPP plots under the wavelet and time-domain approaches respectively. 4th panel presents the distribution of number of CPs. The proposed methodology compliments the time-domain approach and concurs with the truth.

scale j' of the EWS, namely $U_j^2(\frac{k}{n}) = \mathbf{1}_{[j=j']}\sigma^2, \quad j' \in \{1, \dots, J\}$, and a Haar generating wavelet, where σ^2 is the time-domain innovation variance of the process. The equivalent time-domain representation of this model is a MA($2^{j'} - 1$) process with innovation variance σ^2 and MA coefficients determined by the Haar wavelet at scale j'. Piecewise Haar MA processes can thus be constructed by considering piecewise constant EWS. Changes in power across scales correspond to changes in MA order and changes in power within-scales correspond to changes in variance of Y_t. Figures 6.7(b) – 6.7(d) display the corresponding EWS we shall be considering for simulated Haar MA processes. Nason et al. (2000) remark that any MA process can be written as a linear combination of Haar MA processes, hence highlighting a potentially more favourable representation in the wavelet domain.

Figure 6.9 considers a change in order from MA(1) $\leftrightarrow$ MA(7) and constant variance $\sigma^2 = 1$. The associated EWS is presented in Figure 6.7(b). These results show the real potential of the proposed method in that it outperforms the time domain approach. Under the proposed wavelet approach, the CPP are centred and peaked around the defined CP locations, with additional CP potentially being present corresponding to the subtle nuances arising in the data. The potential presence of additional CPs is also reflected in the distribution of the number of CPs with probability assigned to these number of CPs. In contrast, the time-domain method is unable to identify these CPs completely due to the highly correlated nature and change of autocovariance present in the the data. This thus demonstrates that there is an advantage in considering the CP problem in the wavelet-domain over the time-domain, in light of incorrect model specification.

Figure 6.10 displays CP results for a Haar MA process with constant order, changing innovation variance (MA(3), $\sigma^2 = 1, 5$). This is achieved by changing power within a single scale of the EWS as demonstrated in Figure 6.7(c). Results indicate that both the wavelet and time domain approach perform reasonably well with the CPP peaked and centred around the defined CP, and assigning a significant amount of probability to the true number of CPs after the necessary correction has taken place. The wavelet domain also appears to be less sensitive to false CPs potentially occurring, for example the CP detected in the time domain approach at around 250. It is surprising how well the time domain approach performs despite the presence of autocorrelation in the time series. However, its acceptable performance is likely to be due to the underlying MC capturing some of the autocorrelation present in the time series, and the change in variance being a dominant feature of the data which can be successfully modelled by the Gaussian Markov Mixture model.

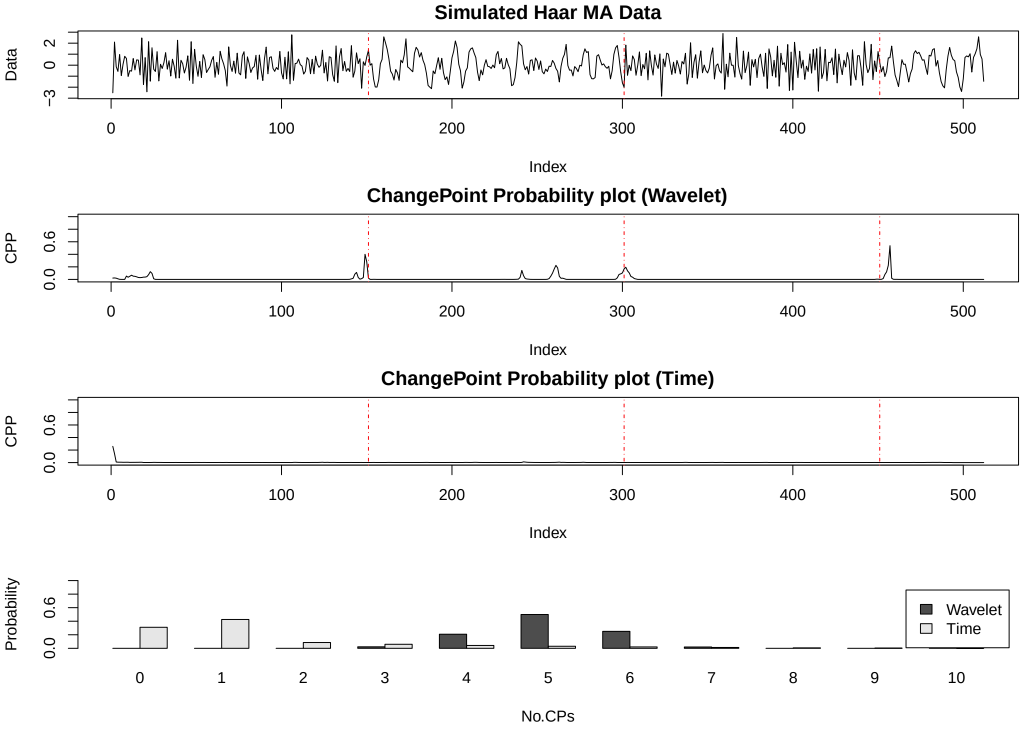


Figure 6.9: Changepoint results for piecewise Haar MA data with a change in order, constant variance (change in power across scales → MA(1) ↔ MA(7), $\sigma^2 = 1$, see Figure 6.7(b) for corresponding EWS). 1st panel presents the simulated data analysed. 2nd and 3rd panel displays the CPP plots under the wavelet and time-domain approaches respectively. 4th panel presents the distribution of number of CPs. The wavelet-domain approach is successfully able to identify the defined CP locations, in addition to other CPs. This is reflected in the distribution of the number of CPs. The time-domain fails to identify the CP characteristics however due to high autocorrelation present in the data and the change within it.

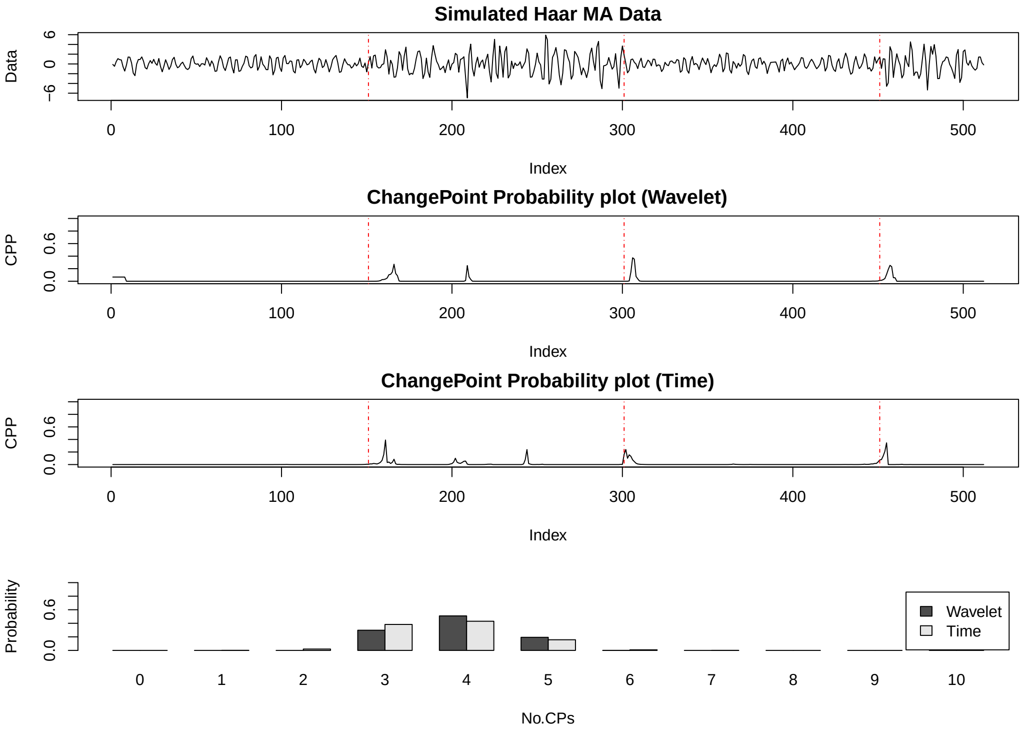


Figure 6.10: CP results for piecewise Haar MA data with a change in innovation variance, constant order (change in power within a scale → MA(3), $\sigma^2 = 1, 5$, see Figure 6.7(c) for corresponding EWS). 1st panel presents the simulated data analysed. 2nd and 3rd panel displays the CPP plots under the wavelet and time domain approaches respectively. 4th panel presents the distribution of number of CPs. Results largely concur with the truth and time approach, with some discrepancies present (offset peaked CPP around defined CP location). However, this is still in line with the behaviour of the data.

Figure 6.11 considers the case of changing power between scales. This results in a piecewise MA process with varying order and innovation variance (MA(1), $\sigma^2 = 1 \leftrightarrow$ MA(7), $\sigma^2 = 5$). The associated EWS is presented in Figure 6.7(d). We observe that both the wavelet and time domain approach perform well with CPP peaked and centred around the defined CP locations, and true number of CPs being the most probable after discarding the artefact at the beginning of the time series. However, the wavelet approach is generally more certain for all potential CP locations than the time domain approach. Again, the good performance of the time domain approach is suspected to be because the change in variance dominants the change in covariance, and this is successfully captured by the Gaussian Markov Mixture framework.

Further piecewise MA simulations were performed with respect to different power configurations at different scales (results not shown here). Under these scenarios, the proposed methodology shows similar performance to those presented in this section by outperforming or compared favourably to the approximating time-domain approach.

Markov Switching Autoregressive Switching Approximation It is clear from Figures 6.9 - 6.11 that the simulated Haar MA processes considered exhibit autocorrelation in the time series. Consequently, a Gaussian Markov Mixture model may not be an appropriate model as it captures little to no autocorrelation structure potentially present. In an attempt to capture some of this autocorrelation structure, we propose an alternative time domain modelling approach for the Haar MA time series. Namely, we consider an extension of Hamilton's Markov Switching Autoregressive model of order r, HMS-AR(r), as defined earlier (Equation 3.34, page 81). This extension is of the following form:

$$a_t = Y_t - \mu_{X_t} \tag{6.27}$$

$$a_t = \sum_{p=1}^{r} \phi_{p,X_t} a_{t-p} + \epsilon_t \qquad \epsilon_t \sim N(0, \sigma^2_{X_t}), \tag{6.28}$$

where the mean, innovation variance and AR coefficients are state dependent with respect to the underlying chain. The state dependent AR coefficients thus allow us to consider changes in autocovariance structure which is not possible under the original HMS-AR(r) model. We refer to this HMS-AR(r) model with switching AR coefficients as Hamilton's Markov Switching Autoregressive Switching model with order r, HMS-ARS. Note that the HMS-ARS(r) model allows us to model piecewise

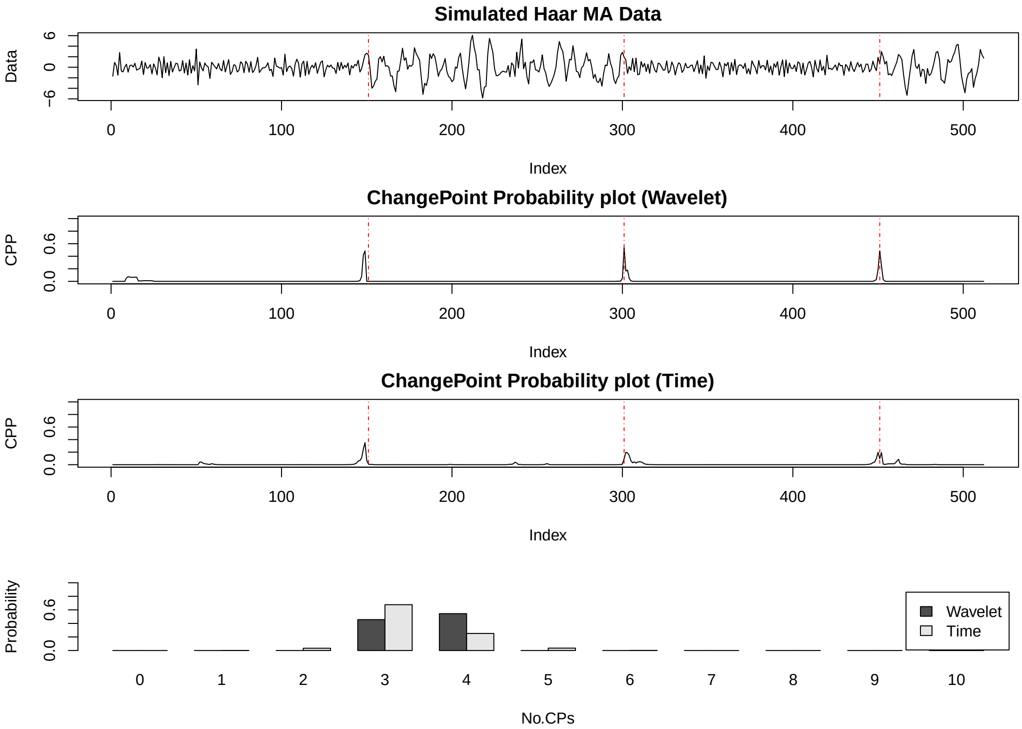


Figure 6.11: CP results for piecewise Haar MA data with a change in order and variance (change in power across scales → MA(1), $\sigma^2 = 1 \leftrightarrow$ MA(7), $\sigma^2 = 5$, see Figure 6.7(d) for corresponding EWS). 1st panel presents the simulated data analysed. 2nd and 3rd panel displays the CPP plots under the wavelet and time domain approaches respectively. 4th panel presents the distribution of number of CPs. Both the wavelet and time domain perform well in identifying the location and number of the defined CPs. However, the wavelet approach appears to fare better with greater certainty in the potential CP location estimates.

AR processes with switching mean, innovation variance and AR structure. Note that although the AR order is fixed under the HMS-ARS(r), changes in AR order can be achieved by setting the corresponding AR coefficients to zero. An alternative modelling approach would also be the Markov Switching AR model defined in Equation 6.1 (page 149) where $q = 0$.

Figures 6.12 – 6.14 present the CP results assuming a HMS-ARS(4) model in the time domain for the same Haar MA processes considered in Figures 6.9 – 6.11. An autoregressive order of four has been chosen arbitrary and sufficiently large enough to account for some of the autocorrelation structure present in the time series. The same SMC settings have been used in obtaining the results ($N = 500$ particles, $B = 100$ distributions, linear tempering schedule et cetera).

We observe that the HMS-ARS approximation method provides promising results for all three data realisations, with CPP centred and peaked around the defined CP locations, and the correct number of CPs being the most probable from the distribution of number of CPs. This alternative time domain approximation clearly outperforms assuming a Gaussian Markov Mixture, and is clearly a more suitable model as a time-domain approximation.

The HMS-ARS time domain approximation also performs as well as the LSW-HMM approach, with greater certainty on the current number of CP locations present and their respective locations in some instances. This thus poses the question as to why one would want to consider the proposed wavelet based approach for analysis. We argue that whilst the HMS-ARS approach performs on a par with the LSW-HMM approach, the latter may be more robust to model mis-specification as we do not need to worry as to whether any autocorrelation is present in the time series since this is modelled automatically under the proposed approach. In addition, under the HMS-ARS approach, a suitable AR order needs to be determined in modelling any potential autocorrelation present. This is systematically accounted for under the proposed LSW-HMM framework by modelling the periodogram directly, and a dependency order in the underlying MC is systematically deduced from the choice of analysing wavelet and the number of periodogram scales considered. Finally, the HMS-ARS framework may not correctly identify changes in AR or MA order as it relies on the associated coefficients being set to zero. The proposed wavelet approach is able to account for these changes in order more explicitly by considering the change in power configuration across scales of the periodogram and EWS.

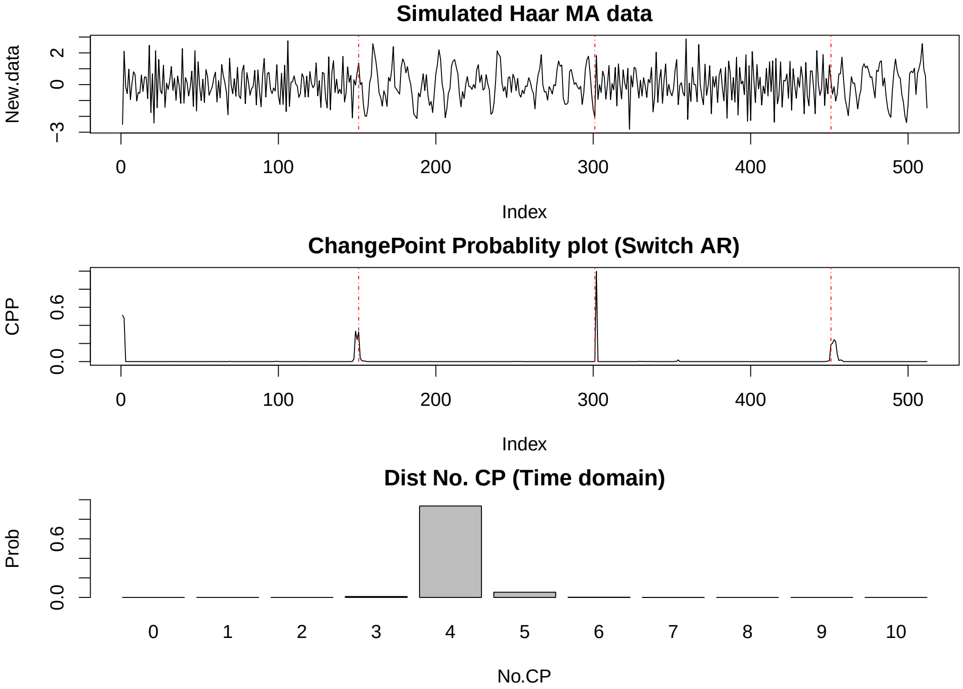


Figure 6.12: CP results for piecewise Haar MA data with a change in order (change in power across scales → MA(1), $\sigma^2 = 1 \leftrightarrow$ MA(7), $\sigma^2 = 1$, see Figure 6.7(b) for corresponding EWS). Analysis assumes a HMS-ARS(4) in the time domain. 1st panel presents the simulated data analysed. 2nd displays the CPP plots under the HMS-ARS(4) time domain approach. 3rd panel presents the distribution of number of CPs.

6.5.2 Oceanographic Application

We now return to consider the oceanographic data example introduced in Section 6.1. Clearly there is ambiguity as to when storm seasons start and the number that have occurred. Hence there is particular interest in quantifying the uncertainty of storm seasons. We therefore apply our proposed methodology to the data from a location in the North Sea.

The analysed data is plotted in the top panel of Figure 6.15 along with CP estimates from existing change in autocovariance methods namely, Cho and Fryzlewicz (2012) (CF, blue top ticks) and Davis et al. (2006) (AutoPARM, red bottom ticks). The data consists of differenced wave heights measured at 12 hour intervals from March 1992 - December 1994 in a central North Sea location.

The following inputs have been used to achieve the presented CP results in Figure 6.15: $J^* = 2$ corresponding to higher frequency time series behaviour (where

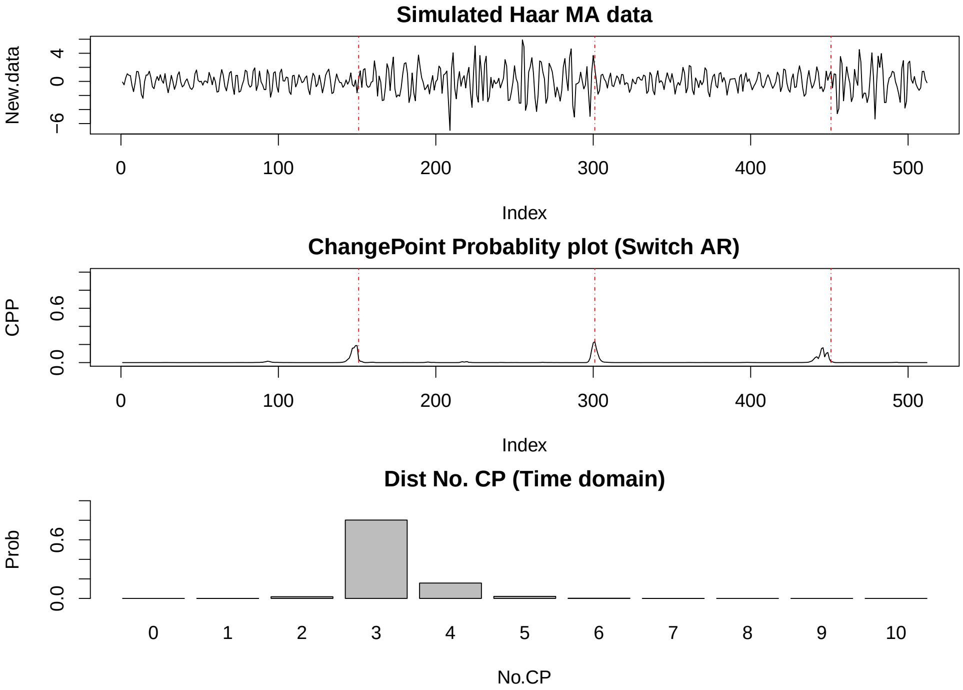


Figure 6.13: CP results for piecewise Haar MA data with a change in innovation variance, constant order (change in power within a scale → MA(3), $\sigma^2 = 1, 5$, see Figure 6.7(c) for corresponding EWS). Analysis assumes a HMS-ARS(4) in the time domain. 1st panel presents the simulated data analysed. 2nd displays the CPP plots under the HMS-ARS(4) time domain approach. 3rd panel presents the distribution of number of CPs.

changes are expected), and $H = 2$ states have been assumed reflecting the belief that there are "stormy" and "non-stormy" seasons. Assuming two states also aids ocean engineers in interpreting the model; we validate this assumption later on. The same SMC samplers settings utilised in the simulated data analysis have been used ($N = 500$ particles, $B = 100$ distributions, linear tempering schedule). Under a sustained CP definition, $k_{\mathrm{CP}} = 40$ and $k'_{\mathrm{CP}} = 30$, have been used to reflect the general sustained nature of seasons (seasons last for at least a few weeks).

Ocean engineers have indicated that it is typical to see two changes in storm season each year occurring in the Spring (March-April) and Autumn (September-October). The results displayed in Figure 6.15 concur with this statement; five and six storm season changes are most likely according to the number of CPs distribution, and with the CPP being centred and peaked around these times. The uncertainty encapsulated by the number of CP distribution demonstrates that there are potentially more or fewer storm seasons than five or six, although these are less

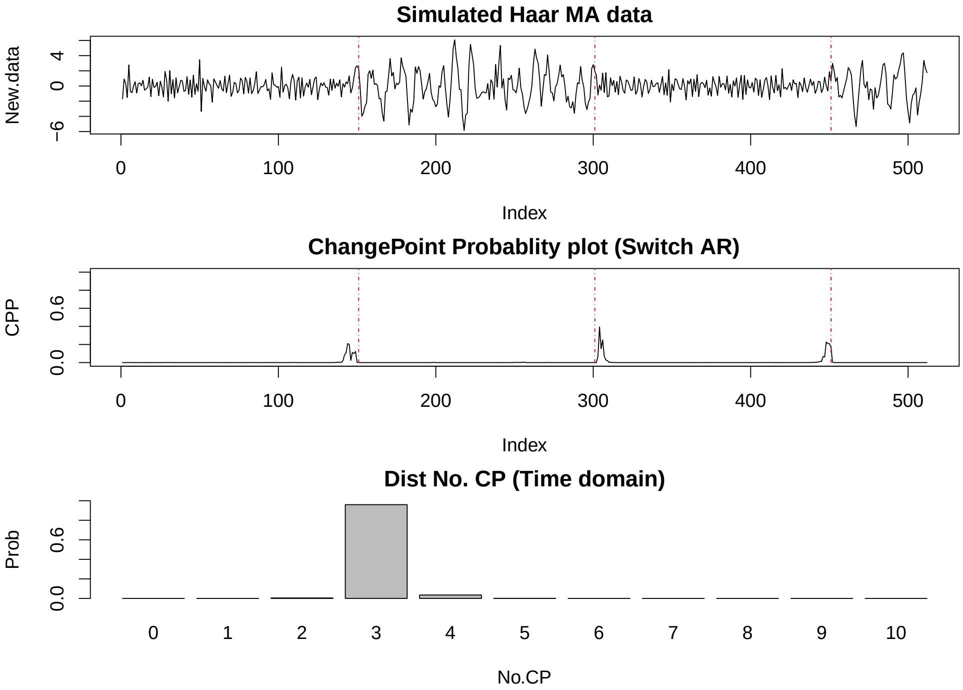


Figure 6.14: CP results for piecewise Haar MA data with a change in order and variance (change in power across scales → MA(1), $\sigma^2 = 1 \leftrightarrow$ MA(7), $\sigma^2 = 5$, see Figure 6.7(d) for corresponding EWS). Analysis assumes a HMS-ARS(4) in the time domain. 1st panel presents the simulated data analysed. 2nd displays the CPP plots under the HMS-ARS(4) time domain approach. 3rd panel presents the distribution of number of CPs.

certain, along with the corresponding locations.

Results also concur with CP estimates from the other two methods, with our method highlighting another possible configuration. A few discrepancies exist, for example the CP estimated in the middle of 1993 and 1994 according to CF and AutoPARM. These potential changes in state do not seem sufficiently sustained for a change in season to have occurred and thus our methodology has not identified them. Lowering the associated values of k_{CP} and k'_{CP}, does begin to identify these as CP instances, in addition to others. However, we justify our values of k_{CP} and k'_{CP} by the sustained nature of storm seasons.

Changes identified in the middle of 1992 and end of 1994 by CF and AutoPARM are suspected to be due to an insufficient number of states to account for these more subtle changes. This suggests that HMM model selection methods may be worth implementing in order to assess whether the two state assumption of “stormy” and “non-stormy” seasons is adequate in modelling the observed wave height data.

Model Selection on North Sea data

Figure 6.16 displays the model selection results obtained via the parallel SMC based HMM model selection approach proposed in Chapter 5. In particular, the top panel displays boxplots of posteriors from fifty different SMC runs, and the bottom panel presents the percentage of a model being selected via maximum a posterior. These results are achieved using the SMC inputs as described in Section 6.5.2 and we consider a maximum of three underlying states ($H^{\max} = 3$) due to the belief that at most three states are required to model the data.

We observe that in nearly all SMC replications, a two state HMM model is assigned almost all probability, and the remaining small amount of probability is assigned to a three state model. The additional third state may thus be capturing the more subtle features associated with the changes identified in the middle of 1992 and end of 1994, although there is relatively little evidence for this third state being needed. In addition, no probability is assigned to a one state model which suggests that a HMM framework is appropriate in modelling the data. Consequently under these posterior approximations, a two state model is selected almost always under MAP.

In the SMC instance where posterior probabilities appear as outliers from the other approximations (for example when the posterior is split almost equally between a two and three state model), this is suspected to be due to sampling error and suggests running the SMC sampler with more samples present (currently $N = 500$). More SMC replications would also provide further confidence in these model selection results. However, these model selection results provide initial strong evidence that a two state HMM is appropriate in modelling the wave height data.

6.6 Discussion

This chapter has proposed a methodology for quantifying the uncertainty of autocovariance CPs in time series, an area which has received little to no attention in the CP community. This is achieved by assuming a Locally Stationary Wavelet framework and considering the estimate of the Evolutionary Wavelet Spectrum which fully characterises the potentially varying second-order structure of a time series. By appropriately modelling this estimate as a multivariate time series under a Hid-

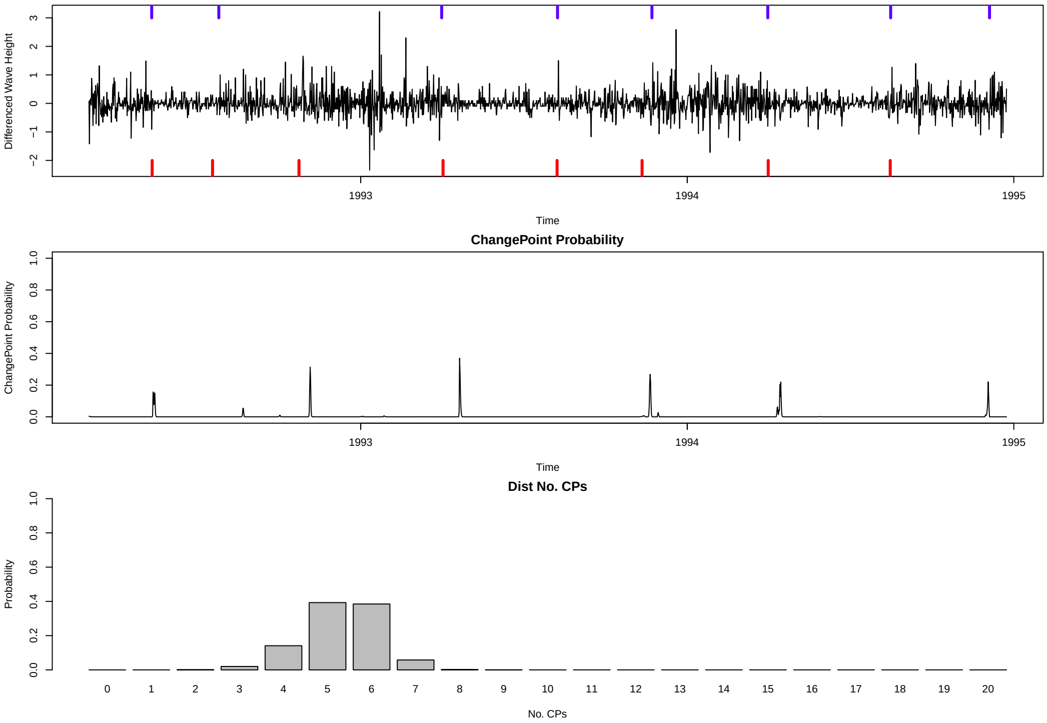


Figure 6.15: Changepoint results for North Sea data. Top panel displays the analysed data and the CP estimates from existing approaches (CF= blue top ticks, AutoPARM= red bottom ticks). Middle and bottom panel display the CPP plot and distribution of number of CPs respectively under the proposed methodology. This corresponds to the start of storm seasons and the number of them. Analysis considers the two finest scales of the wavelet periodogram ($J^* = 2$), and assumes two underlying states ($H = 2$) reflecting "stormy" and "non-stormy" seasons.

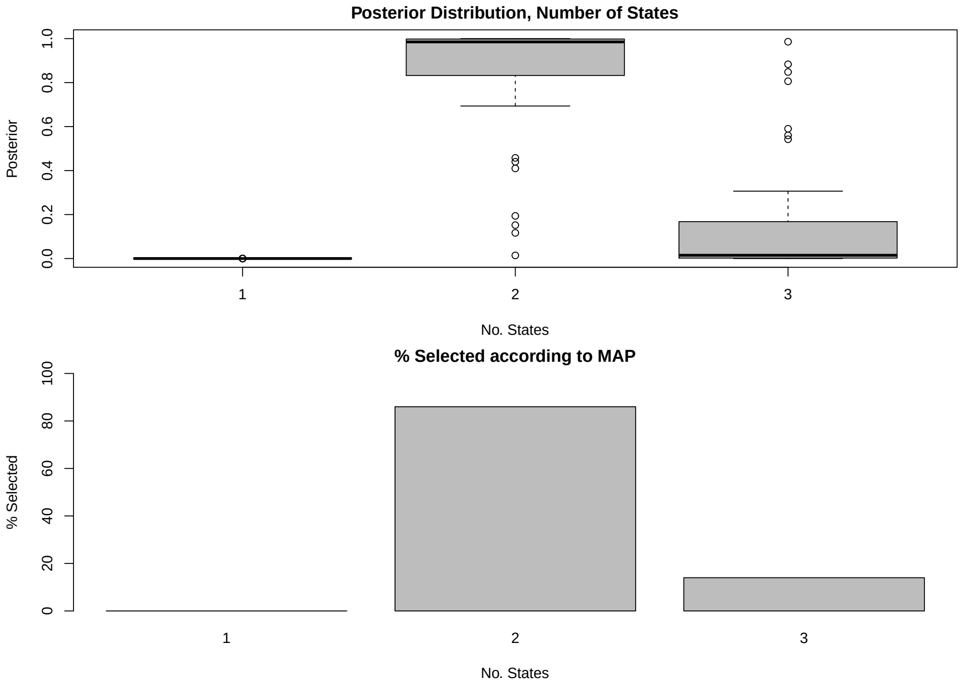


Figure 6.16: Model selection on the North Sea data presented in Figure 6.15 using the parallel SMC model selection approach proposed in Chapter 5. Top panel displays boxplots of the model posteriors from 50 SMC runs, bottom panel displays the percentage selected according to maximum a posterior.

den Markov Model framework and deriving the corresponding multivariate emission density, we can quantify the uncertainty of CPs by the methodology proposed in Chapter 3.

Results on a variety of simulated data indicate that the methodology works well in quantifying the uncertainty of CP characteristics. Application to white noise data show that the proposed methodology compliments the equivalent exact time domain approach. The real advantage of our proposed methodology potentially lies in considering piecewise MA processes which are not readily analysed using the HMM framework in the time-domain without some approximation taking place. For such data, the wavelet approach outperforms a time-domain approximation method which ignores the autocorrelation structure present, and performs on a par with an approximation which accounts for at least some of the autocorrelation via autoregressive coefficients. We believe our proposed wavelet approach is more robust to model mis-specification with less concern as to whether a autocorrelation structure is present, the order of autoregressive nature required and changes that may occur within it.

This methodology has also been applied to an oceanographic dataset, namely wave height data, where interest lies in determining changes in storm season. Such changes correspond to changes in second-order structure where there is also interest in the uncertainty of these changes due to the inherent ambiguity of storm seasons. Our method has showed accordance with various existing CP methods including expert ocean engineers. Our methodology allows us to assess the plausibility and performance of CP estimates and provide further information in planning future operations.

A few discrepancies do exist between the various methods, a potential result of the sustained CP definition implemented and number of states assumed in our HMM. However, the settings used to achieve the results seem valid given the oceanographic application and are more intuitive in controlling CP results compared to abstract tuning parameters and penalisation terms present in other CP methods. In addition, supplementary model selection results indicate that the number of underlying states assumed in generating our CP results is valid.

Extensions of this work include investigating piecewise MA processes further with subtle changes in the piecewise constant EWS structure being associated with changes in MA coefficients and order. Other types of EWS structures, not necessarily piecewise constant in structure could also be further investigated. This may thus extend the types of data and changes in the resultant time series we can consider, and ultimately whether new types of data could be readily analysed which may not be permissible in the time domain.

We note that an offset appears to occur in the CPP from the defined CP locations under the proposed approach, for example Figure 6.8. In assessing whether this lag is specific to the data analysed or due to the proposed methodology, Figure 6.17 displays a summary of the CPP from 50 different sets of analysis for white noise processes. The main black line denotes the median CPP, the grey region is the interquartile range, and the red vertical line denotes the defined CP location. The top and bottom panel display the summary CPP plots for the proposed wavelet and time domain approach respectively. The summary CPP under the wavelet approach indicates that the offset is systematic rather than data specific. This is suspected to be due to `wavethresh` performing some slight re-ordering of the coefficients when computing the periodogram, and indeed the new reordering procedure may cause the detected CP to be off. This offset also appears in Haar MA simulations and a similar offset in Cho and Fryzlewicz (2012) which also utilises `wavethresh` to compute the periodogram. Future work may thus want to explore this offset further and whether a correction mechanism could be developed to correct for it as necessary.

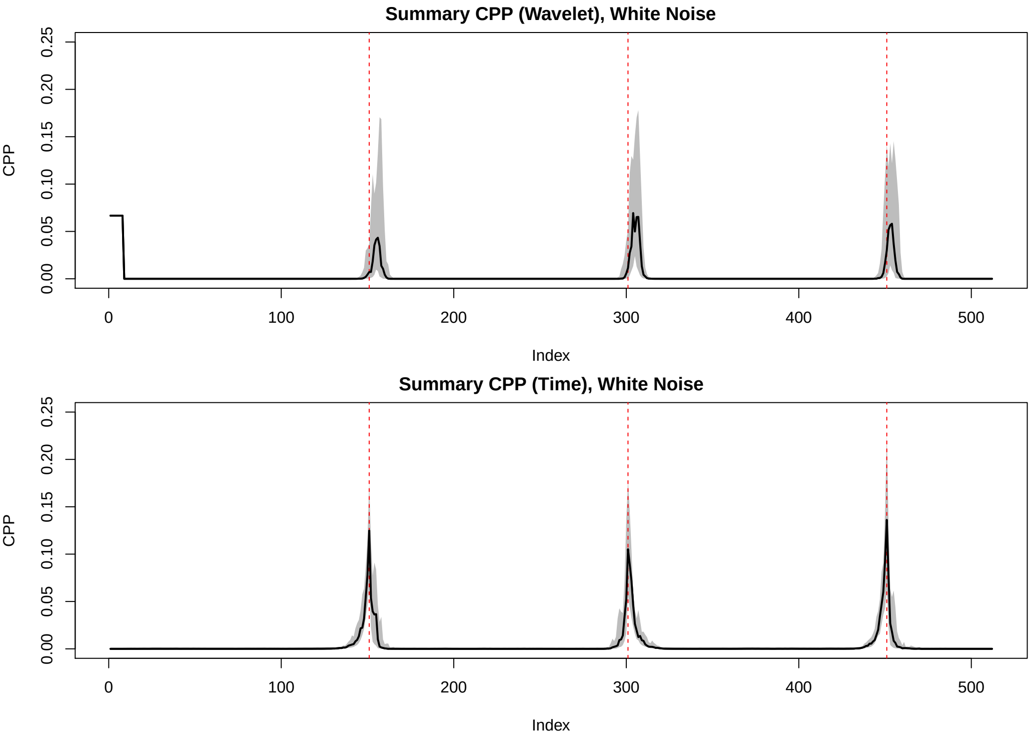


Figure 6.17: Summary of CPP from 50 simulated white noise processes under the proposed wavelet approach (top panel) and time domain approach (bottom panel). The main black line denotes the median CPP from the 50 sets of analysis, and the grey region is the interquartile range of the CPP.

Nevertheless, the summary CPP still provides a good approximation of the CP behaviour in an alternative wavelet manner. This plot also highlights that the CP detected at the beginning of the data is systematic (there is no variation in its location or probability), thus suggesting further that it is possible to correct for it.

In contrast, the time domain plot does not exhibit an offset in CPP, with the CPP being peaked and centred around the defined location. This is not particularly surprising since we are modelling the observed time series directly (no transformation) and exactly (no approximation taking place). This highlights if anything, further confidence in that the CP time domain methodology presented in Chapter 3.

The LSW-HMM framework presented assumes that all dependence between wavelet coefficients is captured for by conditioning on the underlying MC. This is a strong assumption and may want to be relaxed if one does not believe it is adequate. This could be performed by devising an autoregressive model such that dependency on a finite number of previous wavelet coefficients is incorporated into the emission density. This additional dependency could be easily incorporated into the framework (similar to the Markov Switching AR models considered in this thesis) but does add additional complexity. Note that the autoregressive order and coefficients are dependent on the EWS and the mother wavelet used for analysis.

The current LSW framework assumes that the observed time series is mean zero and constant with prior detrending occurring before analysis is performed. However, as exhibited throughout this thesis, non-stationarity can also arise from changes in mean. As demonstrated in Chapter 4, it is important to account for changes in mean and trend within a unified framework in analysis.

A potential path for future research is thus to modify the current LSW framework such that the orthonormal incremental noise process $\xi_{j,k}$, is no longer mean zero and is dependent on k, μ_k say. This thus allows non-zero mean processes to be considered, and also LSW processes which permit changes in mean. More specifically, the modified framework could potentially take the form:

$$\begin{aligned} Y_t &= \sum_{j=1}^{\infty}\sum_{k\in\mathbb{Z}} \psi_{j,k}(t)\epsilon_{j,k} && \epsilon_{j,k} \overset{\text{iid}}{\sim} \mathrm{N}\left(\mu_k, U_j^2\left(\frac{k}{n}\right)\right) \\ &= \sum_{j=1}^{\infty}\sum_{k\in\mathbb{Z}} \psi_{j,k}(t)\left[\mu_k + U_j\left(\frac{k}{n}\right)\xi_{j,k}\right] && \xi_{j,k} \overset{\text{iid}}{\sim} \mathrm{N}(0,1) \\ &= \underbrace{\sum_{j=1}^{\infty}\sum_{k\in\mathbb{Z}} \psi_{j,k}(t)U_j\left(\frac{k}{n}\right)\xi_{j,k}}_{\text{LSW as presented in Equation 6.18}} + \underbrace{\sum_{j=1}^{\infty}\sum_{k\in\mathbb{Z}} \psi_{j,k}(t)\mu_k}_{\text{Mean component}}. \end{aligned} \tag{6.29}$$

As the father scaling wavelet coefficients capture the behaviour of the mean of a time series in standard wavelet analysis, it is suspected that statistical analysis will now focus on the behaviour of $\tilde{\mathbf{c}}_{1:n}$, where $\tilde{\mathbf{c}}_k = \{c_{j,k}\}_{j=1}^{J}$ are the father scaling wavelet coefficients from a non-decimated wavelet transform.

Figure 6.18 presents results from a preliminary study involving simulated piecewise MA processes where changes in mean, variance and autocovariance are exhibited in the time series. Figure 6.18(a) displays an example realisation of such a process with changes in mean and MA order occurring at times 201, 513 and 713, and a change in variance occurring at 513. The left column of Figure 6.18(b) presents information from standard LSW analysis, namely the periodogram and an estimate of the EWS from 250 data realisation, both concerning the mother detail wavelet coefficients $\tilde{\mathbf{d}}_{1:n}$, whereas the right column presents information concerning the father coefficients $\tilde{\mathbf{c}}_{1:n}$ from a single realisation and averaged across 250 realisations. It is clear that a change in autocovariance and variance structure still corresponds to changes in the power configuration of the EWS estimate (both within and across scales due to changes in variance and autocovariance), but also that the changes in mean transpires in the $\tilde{\mathbf{c}}_{1:n}$ coefficients at the corresponding locations in a similar fashion with changes in "power" occurring in scales of the "periodogram". This suggests that a methodology similar to that proposed in this chapter is a potential path of further research. In addition, by providing separate channels in which different types of changes are reported from could provide a better understanding of the data.

By considering the modified version of the LSW framework as detailed in Equation 6.29 and considering analysis of father scaling coefficients $\tilde{\mathbf{c}}_{1:n}$, this could thus potentially provide a powerful unified framework such that changes in mean and autocovariance can be considered simultaneously in the wavelet domain.

6.A Appendix

6.A.1 Joint density of $\mathbf{I}_k = (I_{1,k}, I_{2,k}, \ldots, I_{J^*,k})$

The following section considers the generalised version of computing the density of a transformed random vector. $\mathbf{X}$ and $\mathbf{Y}$ denote standard random vectors here with no connections to the HMM or wavelet framework. This material is taken from Grimmett and Stirzaker (2001). As $\mathbf{Y} = (Y_1 = X_1^2, Y_2 = X_2^2) = T(\mathbf{X} = (X_1, X_2))$ is a many-to-one mapping, direct application of a standard change of variable via the Jacobian argument (Grimmett and Stirzaker, 2001, p. 109) is not permissible. In the one-dimensional case, the following proposition is proposed.

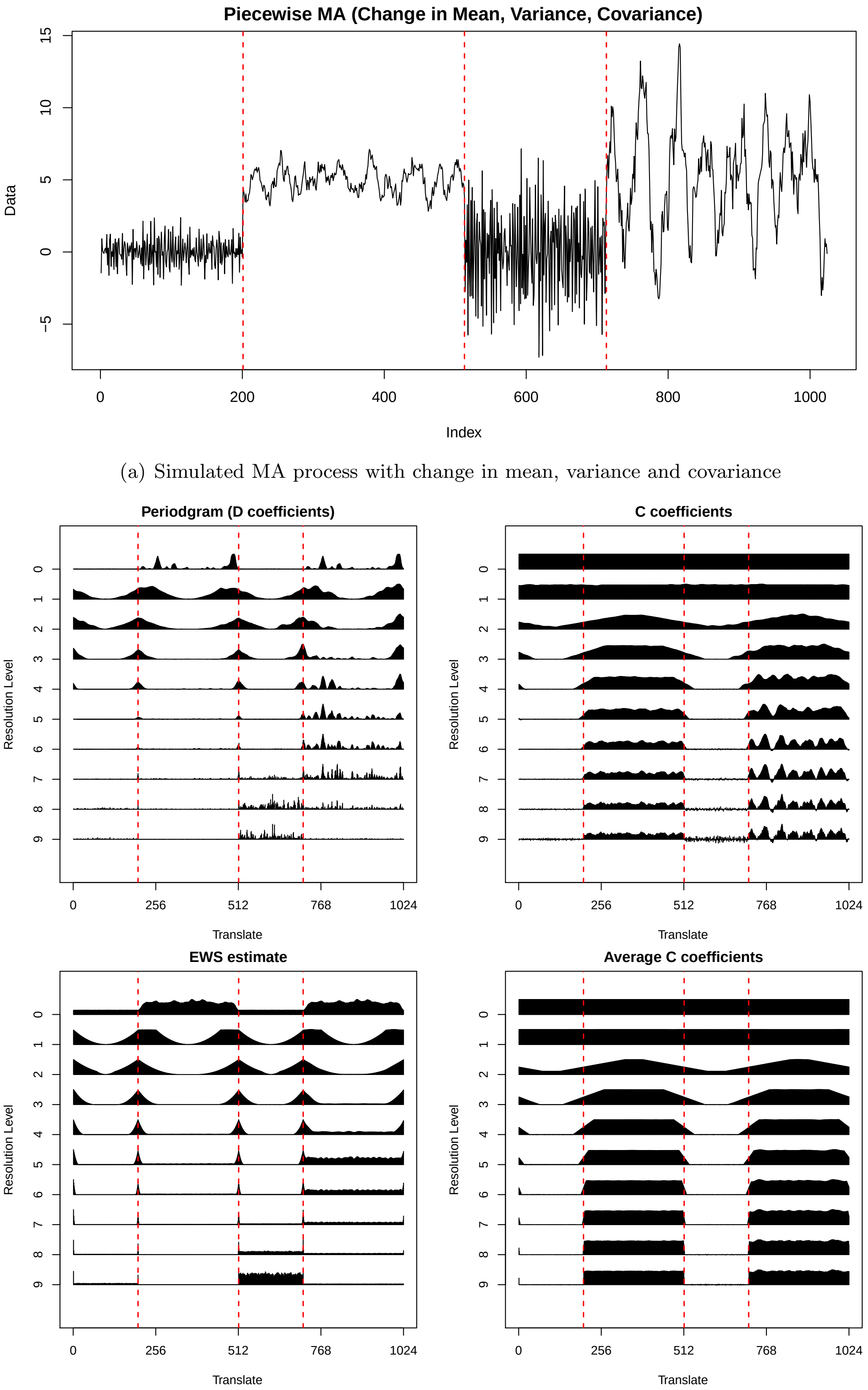
Proposition A-1. *Let $I_1, I_2, \ldots, I_n$ be intervals which partition $\mathbb{R}^2$, and suppose that $Y = g(x)$ where g is strictly monotone and continuously differentiable on every I_i. For each i, the function $g : I_i \to \mathbb{R}$ is invertible on $g(I_i)$ with the inverse function h_i. Then*

$$f_Y(y) = \sum_{i=1}^{n} f_X(h_i(y)) \Big| h_i'(y) \Big|,$$

with the convention that the ith summand is 0 if h_i is not defined at y, and $h_i'(\cdot)$ is the first derivative of $h_i(\cdot)$.

Proof. See (Grimmett and Stirzaker, 2001, page 112). □

Therefore,



(a) Simulated MA process with change in mean, variance and covariance

(b) Periodogram ($\tilde{\mathbf{d}}^2_{1:n}$, top left panel) and father coefficients ($\tilde{\mathbf{c}}_{1:n}$, top right panel) associated with Simulated MA data. Bottom left and right respectively display an estimate of the EWS (averaged 250 periodograms from 250 different simulated time series), and the averaged father coefficients.

Figure 6.18: Wavelet analysis for a simulated MA process with change in mean, variance and covariance.

Proposition A-2. *For* $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n) = (X_1^2, X_2^2, \dots, X_n^2)$

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{Y}}(y_1, \dots, y_n) = \frac{1}{2^n \prod_{i=1}^n |x_i|} \sum_{a_1,\dots,a_n \in \{+,-\}} f_{\mathbf{X}}\left(a_1|x_1|, \dots, a_n|x_n|\right).$$

Proof. Applications of Propositions A-1 and Corollary 4 (Grimmett and Stirzaker, 2001, p. 109). □

6.A.2 Computing Σ_k^D, the covariance structure of $\tilde{\mathbf{D}}_k$

This section outlines how the covariance structure of $\tilde{\mathbf{D}}_k = (D_{1,k} \dots, D_{J^*,k})$, can be computed from the Evolutionary Wavelet Spectrum $U_j^2(\frac{k}{n})$.

Proposition A-3. *The autocovariance structure for the observation process, Y_t, can be characterised by the Evolutionary Wavelet Spectrum as follows:*

$$\mathrm{Cov}(Y_t, Y_{t-v}) = \sum_l \sum_m U_l^2 \left(\frac{m}{n}\right) \psi_{l,m-t} \psi_{l,m-t+v}.$$

Proof. See proof of Proposition 1 in Nason et al. (2000). □

Proof of Proposition 2 (page 173). As LSW processes are assumed to have mean zero, $\mathbb{E}[Y_t] = 0$, then it follows that the wavelet coefficients are mean zero themselves since they can be seen as a linear combination of Gaussian observations. Thus $\mathbb{E}[D_{j,k}] = \mathbb{E}[D_{j',k'}] = 0$. Then

$$\begin{aligned}
\mathrm{Cov}(D_{j,k}, D_{j',k'}) &= \mathbb{E}[D_{j,k} D_{j',k'}] - \mathbb{E}[D_{j,k}]\mathbb{E}[D_{j',k'}] = \mathbb{E}[D_{j,k} D_{j',k'}] \\
&= \mathbb{E}\left[\left(\sum_t Y_t \psi_{j,k-t}\right)\left(\sum_s Y_s \psi_{j',k'-s}\right)\right] \\
&= \mathbb{E}\left[\sum_t \left(\sum_l \sum_m U_l\left(\frac{m}{n}\right) \psi_{l,m-t} \xi_{l,m}\right) \psi_{j,k-t}\right. \\
&\qquad \left. \times \sum_s \left(\sum_p \sum_q U_p\left(\frac{q}{n}\right) \psi_{p,q-s} \xi_{p,q}\right) \psi_{j',k'-s}\right] \\
&= \sum_{t,l,m,s,p,q} U_l\left(\frac{m}{n}\right) \psi_{l,m-t} \psi_{j,k-t} U_p\left(\frac{q}{n}\right) \psi_{p,q-s} \psi_{j',k'-s} \mathbb{E}[\xi_{l,m} \xi_{p,q}]
\end{aligned}$$

By definition,

$$\mathbb{E}[\xi_{l,m}\xi_{p,q}] = \begin{cases} 1, & \text{iff } l = p,\ m = q; \\ 0, & \text{otherwise.} \end{cases}$$

Thus,

$$\begin{aligned}\mathrm{Cov}(D_{j,k}, D_{j',k'}) &= \sum_{t,l,s,m} U_l^2\left(\frac{m}{n}\right)\psi_{l,m-t}\psi_{l,m-s}\psi_{j,k-t}\psi_{j',k'-s}\\ &= \sum_t \psi_{j,k-t}\sum_s \psi_{j',k'-s}\sum_l\sum_m U_l^2\left(\frac{m}{n}\right)\psi_{l,m-t}\psi_{l,m-s}.\end{aligned}$$

Let $s = t - v$, then

$$\begin{aligned}\mathrm{Cov}(D_{j,k}, D_{j',k'}) &= \sum_t \psi_{j,k-t}\sum_{t-v} \psi_{j',k'-t+v}\sum_l\sum_m U_l^2\left(\frac{m}{n}\right)\psi_{l,m-t}\psi_{l,m-t+v}\\ &= \sum_t \psi_{j,k-t}\sum_v \psi_{j',k'-t+v}\sum_l\sum_m U_l^2\left(\frac{m}{n}\right)\psi_{l,m-t}\psi_{l,m-t+v}\\ &= \sum_t\sum_v \psi_{j,k-t}\psi_{j',k'-t+v}\mathrm{Cov}(Y_t, Y_{t-v}).\end{aligned}$$

Thus,

$$\mathrm{Cov}(D_{j,k}, D_{j',k'}) = \sum_t\sum_v \psi_{j,k-t}\psi_{j',k-t+v}\mathrm{Cov}(Y_t, Y_{t-v}). \tag{6.30}$$

□

6.A.3 Determining how much of the EWS one needs to know to compute Σ_k^D

In determining how much of the EWS needs to be known when computing the covariance structure at location k, we consider the following lines of logic. Let $\mathcal{L}_j$ denote the support for the wavelet at scale j (number of non-zero filter coefficients in ψ_j). The number of non-zero product filtering coefficients, $\psi_{l,m-t}\psi_{l,m-t+v}$, is greatest when no lag is present ($v = 0$) and thus we consider the variance of the wavelet coefficients and observations process, $\mathrm{Var}(D_{j,k})$ and $\mathrm{Var}(Y_t)$ respectively. In addition, the number of non-zero product terms will be greatest for the coarsest scale considered, J^*, with corresponding support $\mathcal{L}_{J^*}$

$\mathrm{Var}(D_{j,k})$ will be dependent on observations $Y_k, \ldots, Y_{k-(\mathcal{L}_j-1)}$ for any scale $j = 1, \ldots, J^*$. Thus for the coarsest scale $\mathrm{Var}(D_{J^*,k})$ will be dependent on observations $Y_k, \ldots, Y_{k-(\mathcal{L}_{J^*}-1)}$. The variance for the most distant observation $Y_{k-(\mathcal{L}_{J^*}-1)}$ is dependent on the power from the following locations: $k-(\mathcal{L}_j-1)-(\mathcal{L}_j-1), \ldots, k-(\mathcal{L}_j-1)$, for scale j. The coarsest scale requires the most power feeding into it: $U_{J^*}^2\left(\frac{k-2(\mathcal{L}_{J^*}-1)}{n}\right), \ldots, U_{J^*}^2\left(\frac{k-(\mathcal{L}_{J^*}-1)}{n}\right)$. For the most recent observation Y_k at the coarsest scale, the following power needs to be known $U_{J^*}^2\left(\frac{k-(\mathcal{L}_{J^*}-1)}{n}\right), \ldots, U_{J^*}^2\left(\frac{k}{n}\right)$.

Thus to compute Σ_k^D, the covariance structure of the wavelet coefficients at location k, we must record the power from the locations $k - 2(\mathcal{L}_j - 1), \ldots, k$ for scale $j = 1, \ldots, J^*$.

6.A.4 Order of HMM with respect to analysing wavelet and J^*

We briefly comment on the behaviour of the order of the HMM as we consider more scales and different choices in analysing wavelet. Recall that the order of the HMM is associated with the analysing wavelet considered and J^*, the number of scales considered. More specifically, the HMM order is $2\mathcal{L}_{J^*} - 1$.

For the case of the Haar wavelet, where $\mathcal{L}_j = 2, 4, 8, 16$ for $j = 1, 2, 3, 4$, the corresponding order of the induced HMM is $3, 7, 15, 31$ for $J^* = 1, 2, 3, 4$. Similarly, Daubechies Extremal Phase wavelets with two vanishing moment has the following supports $\mathcal{L}_j = 4, 10, 22, 46$ for $j = 1, 2, 3, 4$. The induced order of the HMM is thus $7, 19, 43, 91$ for $J^* = 1, 2, 3, 4$ scale processes respectively. Thus by considering coarser scales and smoother analysing wavelets, the order of the induced HMM grows exponentially which causes computational problems eventually. The use of a Haar wavelet and only considering a few finer scale processes is thus advocated.

6.A.5 SMC samplers example implementation

This section describes more explicitly the SMC samplers implementation described in Section 6.4.2 (page 177). Defining $l(\theta|\tilde{\mathbf{d}}^2_{1:n}, H)$ as the likelihood, and $p(\theta|H)$ as the prior of the model parameters, we can define the following sequence of distributions,

$$\pi_b(\theta) \propto l(\theta|\tilde{\mathbf{d}}^2_{1:n}, H)^{\gamma_b} p(\theta|H) \qquad b = 1, \ldots, B, \tag{6.31}$$

where $\{\gamma_b\}_{b=1}^B$ is a non-decreasing tempering schedule such that $\gamma_1 = 0$ and $\gamma_B = 1$. We could therefore sample from the sequence of distribution $\{\pi_b\}_{b=1}^B$ as follows:

Initialisation, Sampling from $\pi_1 = p(\theta|H)$: Assume independence between the transition probability matrix, $\mathbf{P}$ and the state dependent power, U^2.

$$p(\theta|H) = p(\mathbf{P}|H)p(U^2|H). \tag{6.32}$$

Transition Probability matrix, P: Sample each of the H transition probability rows $p_s = (p_{s1}, \ldots, p_{sH})$, $s = 1, \ldots, H$ independently from a Dirichlet prior distribution. As HMMs are typically associated with persistent behaviour in the same underlying state, asymmetric priors encouraging persistent behaviour are generally implemented. That is,

$$p_s \overset{\text{iid}}{\sim} \text{Dir}(\alpha_s) \qquad s = 1, \dots, H$$
$$p(\mathbf{P}|H) = \prod_{s=1}^{H} p(p_s|H),$$

where α_s is the associated hyperparameter encouraging persistency.

State Dependent Power, U^2: Sample each of the state dependent inverse power for each scale independently from a Gamma distribution. That is,

$$\lambda_{j,s} = \frac{1}{u_{j,s}^2} \overset{\text{iid}}{\sim} \text{Gamma}(\alpha_\lambda, \beta_\lambda) \qquad j = 1, \dots, J^*, s = 1, \dots, H$$
$$p(\Lambda = \frac{1}{U^2}|H) = \prod_{j=1}^{J^*} \prod_{s=1}^{H} p(\frac{1}{u_{j,s}^2}|H),$$

where α_λ and β_λ are associated shape and scale hyperparameters.

Mutation and Reweighting, approximating π_b from π_{b-1}: We consider Random Walk Metropolis Hastings proposal kernels on different domains given the constraints of the parameters; $\mathbf{P}$ is a stochastic matrix, $u_{j,s}^2$ are non-negative. We consider mutating and updating components of θ separately, using the most recent value of the components (akin to Gibbs sampling). In particular, we consider the following mutation strategies to move from θ_{b-1}^i to θ_b^i, for particle i at iteration b.

Transition Probability matrix, P: Consider each of the H transition probability rows p_s separately, and mutate on the logit scale. That is, we propose moving from p_s to p_s^P via:

$$\begin{aligned}
&\text{Define the current logits:} && l_s = \left(l_{s1} = \log \frac{p_{s1}}{p_{sH}}, \dots, l_{sH} = \log \frac{p_{sH}}{p_{sH}} = 0 \right), \\
&\text{Proposal logits:} && l_s^P = l_s + \epsilon_l \qquad \epsilon_l \sim \text{MVN}(0, \Sigma_l), \quad \text{with } l_{sH}^P = 0, \\
&\text{Proposal probability vectors:} && p_s^P = \left(\frac{\exp l_{s1}^P}{\sum_{n=1}^{H} \exp l_{sn}^P}, \dots, \frac{\exp l_{sH}^P}{\sum_{n=1}^{H} \exp l_{sn}^P} \right),
\end{aligned}$$

where Σ_l is a suitable $H \times H$ proposal covariance matrix.

State Dependent Power, U^2: Consider each of the state dependent inverse powers for each scale independently, and mutate on the log scale. That is we propose moving from $\lambda_{j,s}$ to $\lambda_{j,s}^P$ via:

$$\lambda_{j,s}^P = \exp(\log \lambda_{j,s} + \epsilon_\lambda) \qquad \epsilon_\lambda \sim \mathrm{N}(0, \sigma_\lambda^2), j = 1, \ldots, J^*, s = 1, \ldots, H,$$

where σ_λ^2 is a suitable proposal variance.

Reweighting: From Equation 3.17 (page 72), one can show that under general conditions of SMC samplers, the re-weighting formula for particle i to approximate π_b is:

$$W_b^i = \frac{W_{b-1}^i \tilde{w}_b(\theta_{b-1}^i, \theta_b^i)}{\sum_{i=1}^N W_{b-1}^i \tilde{w}_b(\theta_{b-1}^i, \theta_b^i)}$$
$$\text{with } \tilde{w}_b(\theta_{b-1}^i, \theta_b^i) = \frac{\pi_b(\theta_{b-1}^i)}{\pi_{b-1}(\theta_{b-1}^i)} = \frac{l(\theta_{b-1}^i | \tilde{\mathbf{d}}_{1:n}^2, H)^{\gamma_b}}{l(\theta_{b-1}^i | \tilde{\mathbf{d}}_{1:n}^2, H)^{\gamma_{b-1}}}.$$

Final Output: We have a weighted cloud of N particles approximating the parameter posterior:

$$p(\theta | \tilde{\mathbf{d}}_{1:n}^2, H) \approx \{\theta_B^i, W_B^i | H\}_{i=1}^N \equiv \{\theta^i, W^i | H\}_{i=1}^N. \qquad (6.33)$$

Chapter 7

Discussion and Future Work

Farmer Hoggett: That'll do, pig. That'll do.

Babe (1995), George Miller and Chris Noonan

7.1 Summary of thesis

This thesis has presented aspects and methodologies in quantifying the uncertainty of changepoints (CPs). This is often overlooked compared to detection and estimation of CPs, but is nevertheless important in expressing the plausibility of CP estimates and assessing the performance of different CP methods.

The core methodology in quantifying the uncertainty of CPs (see Chapter 3) draws on the use of Hidden Markov Models (HMMs), Finite Markov Chain Imbedding (FMCI) and Sequential Monte Carlo samplers (SMC). The combination of the three leads to a flexible efficient methodology which does not require sampling the latent sequence and in which inference on a variety of CP characteristics is possible. Incorporation of additional trend and error process structures can also be embedded into this framework with ease (see Chapter 4).

This methodology has been extended into the wavelet framework via the use of the Locally Stationary Wavelet framework (see Chapter 6) . By transforming the time series into the wavelet domain, a joint density between wavelet processes of a periodogram is derived. Consequently, this can be embedded and modelled by a wavelet-based HMM framework which allows for the quantification of autocovariance CP uncertainty in a more robust manner compared to time domain approaches. In general, the quantification of uncertainty regarding autocovariance CPs has received little attention in the literature.

A methodology in determining the unknown number of underlying states of a HMM has also been proposed (see Chapter 5). This utilises the existing SMC framework in a simple efficient manner, such that no additional computations are required in approximating the model posterior. This proposed methodology has been shown to either outperform or perform on a par compared to an existing state of the art method.

These methodologies have been applied to datasets form Econometrics, Neuroimaging and Oceanography. Results indicate promise and potential in the proposed methodologies, concurring with CP estimates supplied by experts and other methods, and providing further evidence of model assumptions used in existing studies. However, the main gain of these results and proposed methodologies is that we have captured the CP uncertainty explicitly, highlighting that other CP configurations may also have occurred.

7.2 Future Work

There are a variety of potential paths for future research, many of which have already been outlined in the concluding sections of their respective chapters. In addition, the following may provide fruitful future research.

7.2.1 Changepoints for multivariate time series

The methods presented in this thesis concern univariate time series. However, multivariate time series are also common, for example when data is measured at several locations over time, or the multi-subject brain imaging dataset considered in Chapter 4. In addition to a temporal structure being present, dependence between the individual time series can also exist which may correspond to the spatial structure between sites for example. However, multivariate CP methods remain relatively undeveloped compared to univariate CP methods. Existing works in CPs for multivariate time series include Kiefer (1959), Srivastava and Worsley (1986), Davis et al. (2006) and Cheon and Kim (2010). Like the univariate time series methods considered in this thesis, these multivariate approaches often do not quantify the uncertainty of CPs explicitly.

However, the methodology presented in Chapter 6 considers a multivariate time series by analysing multiple scale processes of the periodogram simultaneously. This is despite a univariate observed time series being considered originally. Consequently, we have been able to consider quantifying the uncertainty of CPs in the original univariate time-domain series by analysing a multivariate time series in the wavelet domain. There is thus no real reason as to why time-domain multivariate time series cannot be considered under the multivariate HMM framework presented in Chapter 6 and modified as necessary, for time-domain usage.

Section 3.4 (page 121) of MacDonald and Zucchini (1997) outline the main framework for multivariate HMMs (multiple observation time series, a single underlying Markov chain) and the significant results. However, rather surprisingly, the applications and research of multivariate HMMs remains rather sparse (see Zucchini and Guttorp (1991) for one early application). It would therefore be beneficial to spearhead the use of multivariate HMMs, both in modelling multivariate time series and CP analysis.

An initial step for further research is to extend the Gaussian Markov Mixture Model to a Multivariate Normal Markov Mixture model. This multivariate model is intended to have mean vector and covariance matrix which are state dependent in the most general case. That is, for a multivariate time series of J components, $\mathbf{y}_t = (y_{1,t}, \ldots, y_{J,t})$, the corresponding emission distribution would be:

$$\mathbf{y}_t | (X_t = x_t) \overset{\text{iid}}{\sim} \text{MVN}(\mu_{x_t}, \Sigma_{x_t}) \qquad t = 1, \ldots, n \tag{7.1}$$

where μ_{x_t} is a J length column vector, and Σ_{x_t} is a $J \times J$ symmetric, positive definite covariance matrix. Under such a setup, at least some of the entries of the mean vector and the covariance matrix would change in response to a change in underlying state. In addition to capturing changes in mean and variance by the corresponding entries, changes between the covariance structure between time series can also be captured by state-dependent off diagonal entries of Σ_{x_t}, another potential type of change to consider. In estimating these quantities, SMC samplers can still be employed with the Cholesky decomposition of Σ_{x_t} being considered in maintaining its positive definite constraint. Further research may then extend to a Markov Switching Vector Autoregressive model, a Vector Autoregressive model which contains (state dependent) autoregressive parameters. This would thus allow further temporal structure to be incorporated.

For the J multi-subject functional Magnetic Resonance Imaging dataset considered in Chapter 4, we discussed the possibility of capturing a global effect which all subjects encounter, and a subject specific effect. Under the initial multivariate HMM framework presented in Equation 7.1, this could be captured by constructing the state-dependent mean vector as follows,

$$\mu_{x_t} = \mathbf{g}_{x_t} + \eta_{x_t} \qquad t = 1, \ldots, n \tag{7.2}$$

$$\mu_{j,x_t} = g_{x_t} + \eta_{j,x_t} \qquad j = 1, \ldots, J, \tag{7.3}$$

where $\mathbf{g}_{x_t}$ is a J length vector with entries g_{x_t} to denote the global effect and $\eta_{x_t} = (\eta_{1,x_t}, \ldots, \eta_{J,x_t})^T$ which contains subject individual effects. $\mu_{x_t} = (\mu_{1,x_t}, \ldots, \mu_{J,x_t})^T$ consequently captures both global and individual effects. All vectors are state dependent with respect to different stimuli conditions, and it may be reasonable to assume that Σ_{x_t} is diagonal due to subjects and scans typically being independent of one another.

7.2.2 Changepoint uncertainty in light of missing data

This thesis has considered time series data which is complete and does not exhibit missingness. However, missing data is a common occurrence in time series; for example consider missing measurements in the wave height data considered in Chapter 6 due to a temporary defect in the sensor. Rubin (1976) provide a good introductory paper in missing data. This is an area which to the best of my knowledge has received little attention (see Vidal et al. (2008) for some research in the computer vision community). Nunes et al. (2012) have recently indicated their own initial statistical research in this area.

The use of HMMs, and indeed Hidden Semi Markov Models, are one approach in dealing with missing observations (see for example Bahl et al. (1983); Yu and Kobayashi (2003)) with the Expectation-Maximisation algorithm (Dempster et al., 1977) for parameter estimation used in both the HMM and missing data community. Consequently, it seems natural to continue using the HMM framework in both modelling CPs and dealing with missingness. Under the definition of how a CP is typically defined under the HMM framework, the problem may be more straightforward as we are less interested in the missing observation itself, but rather the corresponding underlying state. It is envisaged that accounting for the potential partial state sequence in the missing data regions is the core idea, with a (sustained) change in state corresponding to the CP.

However, the uncertainty of CPs is even more apparent in the case of missing data and thus needs to be captured. Research will focus on $p(x_{t^\star}|y_{1:n})$ and consequently $P(\tau \ni t^\star|y_{1:n})$, where $t^\star$ is the location of a missing observation. Similar to the core methodology presented in this thesis, the uncertainty of the CPs corresponds to the uncertainty regarding the partial state sequence. As a result the FMCI mechanism and Markov Chain theory surrounding the evolution of the underlying Markov Chain will no doubt be valuable resources in tackling this problem.

In addition, Knight et al. (2012) have investigated the effect of missing observations and data collected at irregular times on the estimation of the Evolutionary Wavelet Spectrum under the Locally Stationary Wavelet framework. This could also potentially be useful in considering time series with autocovariance CPs where missingness is present.

7.2.3 Forecasting Changepoints

This thesis has considered CP analysis in a retrospective manner where all estimation and detection is performed with hindsight. However, CPs can also occur in the future and thus forecasting CPs may also be of interest. Developing forecasting CP methods may indicate when to be aware that a system will potentially change in the future and if one needs to prepare for these possible changes.

Forecasting using HMMs have currently been explored by Pesaran et al. (2006); Meligkotsidou and Dellaportas (2011), with the former accounting for the presence of CPs both within the data and potentially beyond the scope of the data. However, these place focus on the observation series and derive predictive densities of the form $f(y_{n+1:n+q}|y_{1:n})$, for a q-step ahead forecast where $q \in \mathbb{N}_+$. Within the context of CPs and under the HMM framework however, focus will shift onto the future of the underlying state sequence, and thus predictive densities with respect to underlying MC, for example $p(x_{n+1:n+q}|y_{1:n})$, are of greater focus. This shift in focus is similar to that of CPs in missing data with inference of how the MC behaves in the future in order to obtain predictive CP densities $P(\tau^{(k_{\text{CP}})} \ni t + q|y_{1:n})$. For the standard change in state definition of a CP, this leads to standard Markov Chin theory with the underlying MC being considered as a standard MC since no observations are available to indicate how it may perform in the future. For the generalised sustained definition of CP, waiting time distributions via FMCI (Aston and Martin, 2005, 2007) for a standard MC are the equivalent framework to consider.

As part of the forecasting frameworks and predictive CP distributions developed, it would also be useful to incorporate the retrospective CP results obtained thus far. For example, if past CPs have occurred around the same time of year, then it may be possible to incorporate this information into our analysis such that a CP is favoured in this same time of year in the future. The use of time inhomogeneous transition probabilities may aid with such incorporation which the FMCI mechanism can handle with ease. In addition, information pertaining to the duration of segments between CPs may be incorporated in some manner, with the use of Hidden Semi Markov Models providing a natural solution since this provides an explicit manner in which one can specify the distribution of state durations.

Bibliography

Abramovich, F., Bailey, T. C., and Sapatinas, T. (2000). Wavelet analysis and its statistical applications. *Journal of the Royal Statistical Society: Series D (The Statistican)*, 49(1):1–29.

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723.

Albert, J. and Chib, S. (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):1371–1381.

Allen, J. (1977). Short term spectral analysis, synthesis, and modification by discrete Fourier transform. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 25(3):235–238.

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

Ashburner, J., Friston, K., Holmes, A. P., and Poline, J. B. (1999). *Statistical Parametric Mapping*. Wellcome Department of Cognitive Neurology, website(http://www.fil.ion.ucl.ac.uk/spm), spm2 edition.

Aston, J. A. D. and Martin, D. E. K. (2005). Waiting time distributions of competiting patterns in higher-order Markovian sequences. *Journal of Applied Probability*, 42(4):977–988.

Aston, J. A. D. and Martin, D. E. K. (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., Peng, J. Y., and Martin, D. E. K. (2011). Implied distributions in multiple change point problems. *Statistics and Computing*, 22(4):981–993.

Auger, I. and Lawrence, C. (1989). Algorithms for the optimal identification of segment neighborhoods. *Bulletin of mathematical biology*, 51(1):39–54.

Bahl, L. R., Jelinek, F., and Mercer, R. (1983). A maximum likelihood approach to continuous speech recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-5(2):179–190.

Bai, J. and Perron, P. (2003). Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, 18(1):1–22.

Barry, D. and Hartigan, J. A. (1993). A Bayesian analysis for change point problems. *Journal of the American Statistical Association*, 88(421):309–319.

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

Beal, M., Ghahramani, Z., and Rasmussen, C. (2002). The infinite hidden Markov model. *Advances in neural information processing systems*, 14:577–584.

Berkes, I., Gombay, E., Horváth, L., and Kokoszka, P. (2004). Sequential change-point detection in GARCH (p, q) models. *Econometric Theory*, 20(6):1140–1167.

Braun, J. V., Braun, R. K., and Müller, H.-G. (2000). Multiple changepoint fitting via quasilikelihood, with application to DNA sequence segmentation. *Biometrika*, 87(2):301–314.

Braun, J. V. and Müller, H.-G. (1998). Statistical methods for DNA sequence segmentation. *Statistical Science*, 13(2):142–162.

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

Carlin, B. P., Gelfand, A. E., and Smith, A. F. M. (1992). Hierarchical Bayesian analysis of changepoint problems. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 41(2):389–405.

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

Chatfield, C. (2003). *The analysis of time series: an introduction*, volume 59. Chapman & Hall/CRC.

Chellappa, R. and Jain, A., editors (1991). *Markov Random Fields: Theory and Application*. Academic Press.

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

Chen, R. and Liu, J. (1996). Predictive updating methods with application to Bayesian classification. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(2):397–415.

Cheon, S. and Kim, J. (2010). Multiple change-point detection of multivariate mean vectors with the Bayesian approach. *Computational Statistics & Data Analysis*, 54(2):406–415.

Chib, S. (1998). Estimation and comparison of multiple change-point models. *Journal of Econometrics*, 86(2):221–241.

Cho, H. and Fryzlewicz, P. (2012). Multiscale and multilevel technique for consistent segmentation of nonstationary time series. *Statistica Sinica*, 22(1):207–229.

Choi, H., Ombao, H., and Ray, B. (2008). Sequential change-point detection methods for nonstationary time series. *Technometrics*, 50(1):40–52.

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 Pelgrin, F. (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.

Condon, E. U. (1937). Immersion of the Fourier transform in a continuous group of functional transformations. *Proceedings of the National Academy of Sciences of the United States of America*, 23(3):158.

Csörgő, M. and Horváth, L. (1997). *Limit Theorems in Change-Point Analysis*. Wiley Series in Probability and Statistics.

Daubechies, I. (1990). The wavelet transform, time-frequency localization and signal analysis. *Information Theory, IEEE Transactions on*, 36(5):961–1005.

Davis, R. A., Lee, T. C. M., and Rodriguez-Yam, G. A. (2006). Structural break estimation for nonstationary time series models. *Journal of the American Statistical Association*, 101(473):223–239.

Del Moral, P. (2004). *Feynman-Kac formulae: genealogical and interacting particle systems with applications*. Probability and Its Applications. Springer, New York.

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

Del Moral, P., Doucet, A., and Jasra, A. (2012). On adaptive resampling strategies for sequential Monte Carlo methods. *Bernoulli*, 18(1):252–278.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

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.

Douc, R. and Cappé, O. (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*, pages 64–69. IEEE.

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

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

Eckley, I., Fearnhead, P., and Killick, R. (2011). Analysis of changepoint models. In Barber, D., Cemgil, A., and Chiappa, S., editors, *Bayesian Time Series Models*, pages 215–238. Cambridge University Press.

Eddy, S. R. (2004). What is a hidden Markov model? *Nature Biotechnology*, 22(10):1315 – 1316.

Erdman, C. and Emerson, J. W. (2007). bcp: An R package for performing a Bayesian analysis of change point problems. *Journal of Statistical Software*, 23(3):1–13.

Erdman, C. and Emerson, J. W. (2008). A fast Bayesian change point analysis for the segmentation of microarray data. *Bioinformatics*, 24(19):2143–2148.

Fearnhead, P. (2005). Exact Bayesian curve fitting and signal segmentation. *Signal Processing, IEEE Transactions on*, 53(6):2160–2166.

Fearnhead, P. (2006). Exact and efficient Bayesian inference for multiple changepoint problems. *Statistical Computing*, 16(2):203–213.

Fearnhead, P. and Liu, Z. (2007). On-line inference for multiple changepoint problems. *Journal of the Royal Statistical Society Series B*, 69(4):589–605.

Friston, K. (2009). Causal modelling and brain connectivity in functional magnetic resonance imaging. *PLoS biology*, 7(2):e1000033.

Frühwirth-Schnatter, S. (2005). *Finite Mixture and Markov Switching Models.* Springer Series in Statistics.

Fryzlewicz, P. and Nason, G. (2006). Haar–fisz estimation of evolutionary wavelet spectra. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(4):611–634.

Fryzlewicz, P. and Ombao, H. (2009). Consistent classification of nonstationary time series using stochastic wavelet representations. *Journal of the American Statistical Association*, 104(485):299–312.

Fryzlewicz, P., Van Bellegem, S., and Von Sachs, R. (2003). Forecasting non-stationary time series by wavelet process modelling. *Annals of the Institute of Statistical Mathematics*, 55(4):737–764.

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

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

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

Godfeld, S. and Quandt, R. (1973). The estimation of structural shifts by switching regressions. *Annals of Economic and Social Measurement*, 2(4):473–483.

Goldberg, D. E. and Holland, J. H. (1988). Genetic algorithms and machine learning. *Machine Learning*, 3(2):95–99.

Gott, A. N. and Eckley, I. A. (2013). A note on the effect of wavelet choice on the estimation of the evolutionary wavelet spectrum. *Communications in Statistics - Simulation and Computation*, 42(2):393–406.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–32.

Grimmett, G. and Stirzaker, D. (2001). *Probability and Random Processes.* Oxford University Press, 3rd edition.

Grinband, J., Wager, T. D., Lindquist, M., Ferrera, V. P., and Hirsch, J. (2008). Detection of time-varying signals in event-related fMRI designs. *NeuroImage*, 43(3):509–520.

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

Harte, D. (2012). *HiddenMarkov: Hidden Markov Models.* Statistics Research Associates, Wellington. R package version 1.7-0.

Hinkley, D. V. (1971). Inference about the change-point from cumulative sum tests. *Biometrika*, 58(3):509–523.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical science*, 14(4):382–401.

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

Inclan, C. and Tiao, G. C. (1994). Use of cumulative sums of squares for retrospective detection of changes of variance. *Journal of the American Statistical Association*, 89(427):913–923.

Jackson, B., Scargle, J., Barnes, D., Arabhi, S., Alt, A., Gioumousis, P., Gwin, E., Sangtrakulcharoen, P., Tan, L., and Tsai, T. (2005). An algorithm for optimal partitioning of data on an interval. *Signal Processing Letters, IEEE*, 12(2):105–108.

Juang, B. and Rabiner, L. (1991). Hidden Markov models for speech recognition. *Technometrics*, 33(3):251–272.

Kiefer, J. (1959). K-sample analogues of the Kolmogorov-Smirnov and Cramér-V. Mises tests. *The Annals of Mathematical Statistics*, 30:420–447.

Killick, R. (2012). *Novel methods for changepoint problems.* PhD thesis, Lancaster University.

Killick, R. and Eckley, I. A. (2011). *changepoint: An R package for changepoint analysis.* R package version 0.5.

Killick, R., Fearnhead, P., and Eckley, I. A. (2012a). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598.

Killick, R., Nam, C. F. H., Aston, J. A. D., and Eckley, I. A. (2012b). changepoint.info: The changepoint repository.

Kirch, C. (2006). *Resampling Methods for the Change Analysis of Dependent Data.* PhD thesis, Universität zu Köln,.

Kitagawa, G. (1996). Monte carlo filter and smoother for non-gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics*, 5(1):pp. 1–25.

Knight, M. I., Nunes, M. A., and Nason, G. P. (2012). Spectral estimation for locally stationary time series with missing observations. *Statistics and Computing*, 22(4):877–895.

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

Koop, G. and Potter, S. M. (2009). Prior Elicitation in Multiple Change-Point Models. *International Economic Review*, 50(3):751–772.

Lee, S. and Lee, T. (2004). CUSUM test for parameter change based on the maximum likelihood estimator. *Sequential Analysis: Design Methods and Applications*, 23(2):239–256.

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

Lindquist, M. (2008). The statistical analysis of fMRI data. *Statistical Science*, 23(4):439–464.

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

Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing.* Springer.

Luong, T., Rozenholc, Y., and Nuel, G. (2012). Fast estimation of posterior probabilities in change-point models through a constrained hidden Markov model. *arXiv preprint arXiv:1203.4394.*

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

Mackay, R. (2002). Estimating the order of a hidden Markov model. *Canadian Journal of Statistics*, 30(4):573–589.

Mallat, S. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 11(7):674–693.

MATLAB (2012). *version 7.14.0 (R2012a).* The MathWorks Inc., Natick, Massachusetts.

Meligkotsidou, L. and Dellaportas, P. (2011). Forecasting with non-homogeneous hidden Markov models. *Statistics and Computing*, 21(3):439–449.

Minas, G., Rigat, F., Nichols, T. E., Aston, J. A. D., and Stallard, N. (2012). A hybrid procedure for detecting global treatment effects in multivariate clinical trials: theory and applications to fMRI studies. *Statistics in Medicine*, 31(3):253–268.

Murphy, K. (2002). Hidden semi-Markov models (HSMMs). Unpublished notes.

Nam, C. F. H., Aston, J. A. D., Eckley, I. A., and Killick, R. (2013). The uncertainty of storm season changes: Quantifying the uncertainty of autocovariance changepoints. *CRiSM Research Report*, 13(5).

Nam, C. F. H., Aston, J. A. D., and Johansen, A. M. (2012a). Parallel Sequential Monte Carlo samplers and estimation of the number of states in a Hidden Markov model. *CRiSM Research Report*, 12(23).

Nam, C. F. H., Aston, J. A. D., and Johansen, A. M. (2012b). Quantifying the uncertainty in change points. *Journal of Time Series Analysis*, 33(5):807–823.

Nason, G. (2012). *wavethresh: Wavelets statistics and transforms.* R package version 4.6.1.

Nason, G. and Silverman, B. (1995). The Stationary Wavelet Transform and some Statistical Applications. *Lecture Notes In Statistics*, pages 281–281.

Nason, G., Von Sachs, R., and Kroisandt, G. (2000). Wavelet processes and adaptive estimation of the evolutionary wavelet spectrum. *Journal of the Royal Statistical Society: Series B*, 62(2):271–292.

Nason, G. P. (2008). *Wavelet Methods in Statistics with R.* Springer Verlang.

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

Nuel, G. and Luong, T. M. (2012). *postCP: A package to estimate posterior probabilities in change-point models using constrained HMM.* R package version 1.05.

Nunes, M. A., Killick, R., and Knight, M. I. (2012). Personal Communication, Joint Statistical Meeting 2012.

Ogawa, S., Lee, T., Kay, A., and Tank, D. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences*, 87(24):9868–9872.

Ogden, R. T. (1997). *Essential wavelets for statistical applications and data analysis.* Birkhäuser.

Olshen, A., Venkatraman, E., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5(4):557–572.

Ombao, H., Raz, J., Von Sachs, R., and Guo, W. (2002). The SLEX model of a non-stationary random process. *Annals of the Institute of Statistical Mathematics*, 54(1):171–200.

Ombao, H. C., Raz, J. A., von Sachs, R., and Malow, B. A. (2001). Automatic statistical analysis of bivariate nonstationary time series. *Journal of the American Statistical Association*, 96(454):543–560.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, 41(1/2):100–115.

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

Peng, J.-Y., Aston, J. A. D., and Liou, C.-Y. (2011). Modeling time series and sequences using Markov chain embedded finite automata. *International Journal of Innovative Computing Information and Control*, 7(1):407–431.

Percival, D. B. and Walden, A. T. (2007). *Wavelet Methods for Time Series Analysis.* Cambridge Series in Statistical and Probabilistic Mathematics.

Pesaran, M. H., Pettenuzzo, D., and Timmermann, A. (2006). Forecasting time series subject to multiple structural breaks. *The Review of Economic Studies*, 73(4):1057–1084.

Poldrack, R., Mumford, J., and Nichols, T. (2011). *Handbook of functional MRI data analysis.* Cambridge University Press.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257 –286.

Rioul, O. and Vetterli, M. (1991). Wavelets and signal processing. *Signal Processing Magazine, IEEE*, 8(4):14–38.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14(5):465–471.

Robert, C. P., Rydén, T., and Titterington, D. M. (2000). Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(1):57–75.

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

Ross, G. J. (2012). *Nonparametric Sequential Change Detection in R: The cpm package.*

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Salomon, D. (2004). *Data compression: the complete reference.* Springer-Verlag New York Incorporated.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

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.

Shiskin, J. (1974). The changing business cycle. *New York Times*, 12:22.

Sin, B. and Kim, J. H. (1995). Nonstationary hidden Markov model. *Signal Processing*, 46(1):31 – 46.

Smith, A. F. M. (1975). A Bayesian approach to inference about a change-point in a sequence of random variables. *Biometrika*, 62(2):407–416.

Srivastava, M. S. and Worsley, K. J. (1986). Likelihood Ratio Tests for a Change in the Multivariate Normal Mean. *Journal of the American Statistical Association*, 81(393):199–204.

Stephens, D. A. (1994). Bayesian retrospective multiple-changepoint identification. *Journal of the Royal Statistical Society Series C (Applied Statistics)*, 43(1):159–578.

Tahmasbi, R. and Rezaei, S. (2008). Change point detection in GARCH models for voice activity detection. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(5):1038–1046.

Talairach, J. and Tournoux, P. (1988). *Co-planar stereotaxic atlas of the human brain*, volume 147. Thieme New York:.

Titterington, D. M. (1984). Comments on “Application of the conditional population-mixture model to image segmentation”. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-6(5):656 –658.

Venkatraman, E. S. (1992). Consistency results in multiple change-point situations. Technical report, Department of Statistics Standford University.

Vidakovic, B. (1999). *Statistical Modeling by Wavelets.* Wiley Series in Probability and Statistics.

Vidal, R., Tron, R., and Hartley, R. (2008). Multiframe motion segmentation with missing data using PowerFactorization and GPCA. *International Journal of Computer Vision*, 79(1):85–105.

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

Vostrikova, L. J. (1981). Detecting "disorder" in multidimensional random processes. *Soviet Mathetmatics Doklady*, 24:55–59.

Whitcher, B., Guttorp, P., and Percival, D. (2000). Multiscale detection and location of multiple variance changes in the presence of long memory. *Journal of Statistical Computation and Simulation*, 68(1):65–87.

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

Woolrich, M., Behrens, T., Beckmann, C., and Smith, S. (2005). Mixture models with adaptive spatial regularization for segmentation with an application to fMRI data. *Medical Imaging, IEEE Transactions on*, 24(1):1–11.

Worsley, K. J., Liao, C., Aston, J. A. D., Petre, V., Duncan, G., and Evans, A. C. (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(3):181–189.

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

Yu, S.-Z. and Kobayashi, H. (2003). A hidden semi-Markov model with missing data and multiple observation sequences for mobility tracking. *Signal Processing*, 83(2):235–250.

Zeileis, A., Leisch, F., Hornik, K., and Kleiber, C. (2002). strucchange: An R package for testing for structural change in linear regression models. *Journal of Statistical Software*, 7(2):1–38.

Zhou, Y., Johansen, A. M., and Aston, J. A. D. (2012). Bayesian model comparison via path-sampling sequential Monte Carlo. In *Statistical Signal Processing Workshop (SSP), 2012 IEEE*, pages 245–248. IEEE.

Zucchini, W. and Guttorp, P. (1991). A hidden Markov model for space-time precipitation. *Water Resources Research*, 27(8):1917–1923.