MONASH University

Bayesian Change-Point Modeling with Segmented ARMA Model

Farhana Sadia

M.S in Applied Statistics

A thesis submitted for the degree of Doctor of Philosophy

at the School of Mathematics

Monash University, Australia

February, 2020

Copyright notice

Abstract

An intensely investigated problem in time series analysis is identifying change-points (that is, segment boundary points) and modeling shifts in the dynamical properties of each segment. Time series segmentation plays a vital role in many applications. An important and interesting focus of current research in this field is to segment multiple sequences in parallel instead of a single sequence, which can make the inference of change-point locations more accurate, precise and sensitive. This task has recently received much attention.

The autocorrelation structure of parallel time series may display more complex patterns than are observed for individual time series and some dependency may also exist between parallel time series. Therefore, the task of developing methods to account for these dependencies in order to avoid false change-point detection is important in this context.

In this research, my focus was to use a Bayesian approach to segment time series data and develop methods to segment multiple parallel time series. The analysis was mainly carried out using a Bayesian change-point segmented ARMA model. In the first phase of the research, I introduced and validated this model. This novel methodology presented a promising direction to estimate the locations of change-points by segmenting a time series using an ARMA model to account for the autocorrelations in time series in a better way. ARMA models express a time series as a linear function of its past values and consider the dependence between residual terms by incorporating a moving average component. This methodology used a prior on the locations of change-points as well as on different parameters of the ARMA model and determined the posterior probability distributions of these change-points. Parameters were inferred using the Generalized Gibbs sampler Markov chain Monte Carlo technique. A second methodological innovation was a simple one-dimensional approach to segmenting parallel time series. I proposed an event detection approach for segmenting spatio-temporal data with background noise which adapted the segmented ARMA model. I pre-processed the data by finding the maximum over locations for each time point, thus producing a single summary time series amenable to analysis with

the segmented ARMA model instead of segmenting two-dimensional data. This dimension reduction method involves information loss and provided false positive change-points in high background noise. A third, and most important, methodological innovation was to build a more general and flexible method for the segmentation of multiple parallel sequences. I developed three alternative models to simultaneously segment multiple series, each of which were alternative generalizations of the Bayesian change-point segmented ARMA model. Incorporating multiple parallel series in segmentation greatly helped to clearly identify all change-point locations for a data set where the segmentation of a single time series detected only some change-points.

Finally, my focus turned to zero-inflated data, raised because of the presence of excessive numbers of zeros in an interesting data set pertaining to the composition of sediment cores extracted from a floodplain lake. These time series directed my research interest to investigate the literature of models regarding zero-inflated data. The insight of this review provides an interesting direction for future research, to generalize the Bayesian change-point segmented ARMA model for handling zero-inflated time series data.

Declaration

This thesis contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

Farhana Sadia
08 February 2020

Thesis including published works declaration

I hereby declare that this thesis contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

This thesis includes one original paper published in a peer reviewed journal, two submitted articles and one article to be submitted. The core theme of the thesis is segmenting time series data with a Bayesian change-point model and developing methods to segment multiple parallel time series. The ideas, development and writing up of all the papers in the thesis were the principal responsibility of myself, the student, working within the School of Mathematics at Monash University under the supervision of Associate Prof. Jonathan M. Keith, Dr. Robert Bryson-Richardson and Prof. Kate Smith-Miles.

In the case of Chapters 3, 4, 5 and 6, my contribution to the work involved the following:

Thesis Chapter	***Publication Title***	***Status***	***Nature and % of student contribution***	***Co-author name(s) Nature and % of Co-author's contribution****	***Co-author(s), Monash student Y/N****
3	*Bayesian change-point modelling with segmented ARMA model*	*Published*	*80% Concept, collecting data, proposing and implementing method, validation, writing code and manuscript*	*Jonathan M. Keith, supervised, proposing method, writing code, input into manuscript and proofread, 15%*	*No*
				Sarah Boyd, writing code, 5%	*No*
4	*Event detection in spatio-temporal data using a one-dimensional summary statistic*	*Submitted*	*75% Concept, proposing method, analysing data and results, writing code and manuscript*	*Jonathan M. Keith, supervised, input into manuscript and proofread, 15%*	*No*
				Sevvandi Kandanaarachchi, data collection, analysing data and results, input and proofread, 5%	*No**
				Kate Smith-miles, supervised, data collection, proofread manuscript, 5%	*No*

5	*Segmenting multiple sequence in parallel: three paradigms*	*Submitted*	*80% Concept, proposing method, analysing data and results, writing code and manuscript*	*Jonathan M. Keith, supervised, input into manuscript and proofread, 13%*	*No*
				Anna Lintern, data collection, and proofread manuscript, 5%	*No**
				David McCarthy, data collection, 2%	*No*
6	*A study of models for zero-inflated time series data*	*To be submitted*	*90% Concept, model, writing manuscript*	*Jonathan M. Keith, supervised, concept, input into manuscript and proofread, 10%*	*No*

**Former Monash PhD student but involvement with this project occurred after the award of their degree*

I have not renumbered sections of submitted or published papers in order to generate a consistent presentation within the thesis.

Student signature: **Date:**

The undersigned hereby certify that the above declaration correctly reflects the nature and extent of the student's and co-authors' contributions to this work. In instances where I am not the responsible author, I have consulted with the responsible author to agree on the respective contributions of the authors.

Main Supervisor signature: **Date:**

Acknowledgements

First and foremost, I am thankful to almighty Allah that by his grace and bounty, I am able to finish my PhD research work and write my thesis.

I would like to express my sincere gratitude to my supervisor Associate Prof. Jonathan M. Keith for granting me the opportunity to pursue this wonderful research project under his excellent supervision. I am grateful to him for his exceptional support, insightful suggestions, motivating presence, friendly attitude, careful guidance and inspiration throughout this challenging period of PhD that allowed me to explore the various aspects of statistical sciences and implanted confidence for conducting research. I sincerely appreciate his valuable contribution for supervising me with his immense knowledge.

I would thank my associate supervisors Prof. Kate Smith-Miles and Dr. Robert Bryson-Richardson for their help and suggestions in my research project. I would also like to express my appreciation to Dr. Sarah Boyd, Dr. Sevvandi Kandanaarachchi, Dr. Anna Lintern and Dr. David McCarthy for their help and collaborations. I also found myself motivated by the insightful comments and feedbacks from my PhD panel members: Associate Prof. Tianhai Tian, Associate Prof. Tim Garoni and Associate Prof. Kais Hamza during my milestone seminars.

I wish to thank my friend and colleague Amani Alahmadi for supporting me and sharing different discussions with me. I would also acknowledge the support from our administration staff, especially, John Chan for his continuous support throughout my PhD. I am truly grateful to Monash University, Faculty of Science, School of Mathematics and my supervisor Associate Prof. Jonathan M. Keith for providing me the scholarship throughout my PhD candidature. I would also like to express my heartiest gratitude to the University of Dhaka

and Institute of Statistical research and Training (ISRT) for granting me study leave to carry out my PhD research work.

Finally, a heartfelt thanks to my beloved husband, Dr. Maizbha Uddin Ahmed whose love, support and encouragement allowed me to finish this journey. I would like to thank my daughter Mersiha Ibnat Ahmed who came to my life during my PhD journey and became an inspiration. Last but not least, I would like to thank my parents for their unconditional blessings throughout my life.

The List of publications

1. Chapter 3: **Sadia F**, Boyd S, Keith JM. (2018). Bayesian change-point modeling with segmented ARMA model. *PloS one* 13(12):e0208927. doi:10.1371/journal.pone.0208927.

2. Chapter 4: **Sadia F**, Kandanaarachchi S, Smith-Miles K, Keith JM. (2020) Event detection in spatio-temporal data using a one-dimensional summary statistic. Submitted.

3. Chapter 5: **Sadia F**, Lintern A, McCarthy D, Keith JM. (2020). Segmenting multiple sequences in parallel: three paradigms. Submitted.

4. Chapter 6: **Sadia F**, Keith JM. (2020). A Study of Models for Zero-inflated Time Series Data. In Preparation.

Notation

Symbol	Description
T	The total length of the signal.
$t = 1, 2 \cdots, T$	A time point in the signal.
x_t	Signal at time t.
$\mathbf{X} = (x_1, \cdots, x_T)$	A vector of the values of the signal.
ϕ	Probability of starting a new segment.
K	Total number of segments.
$k = 1, \cdots, K$	Index for the segments.
$\mathbf{s} = (s_1, \cdots, s_K)$	Starting positions of the segments.
N	The number of groups.
$\mathbf{g} = (g_1, \cdots, g_K)$	A vector containing the assignment of each segment to a group.
$\boldsymbol{\pi} = (\pi_1, \cdots, \pi_N)$	The probabilities of assigning segments to groups.
a	Order of the AR model.
m	Order of the MA model.
ψ	Parameter of AR model.
θ	Parameter of MA model.
$\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_T)$	A vector of error terms.
σ^2	variance of the error term.
c_k	Mean signal level (or mean of the ARMA model) for the segment k.
AR(a)	Autoregressive model with order a which is: $\mathrm{x}_t = c + \sum_{i=1}^{a} \psi_i (x_{t-i} - c) + \epsilon_t$
MA(m)	Moving average model with order m which is: $\mathrm{x}_t = c + \sum_{i=1}^{m} \theta_i \epsilon_{t-i} + \epsilon_t$
ARMA(a, m)	Autoregressive moving average model with order a and m which is: $\mathrm{x}_t = c + \epsilon_t + \sum_{i=1}^{a} \psi_i (x_{t-i} - c) + \sum_{i=1}^{m} \theta_i \epsilon_{t-i}$

Contents

1 Introduction **1**

2 Background and Literature Review **13**

2.1 Introduction 13

2.2 Some Background to Segmentation 13

2.3 Sampling 25

2.4 Segmenting parallel sequences 40

2.5 Models and methods used in this thesis 46

3 Bayesian change-point modeling with segmented ARMA model **62**

3.1 Introduction 64

3.2 Methodology 66

3.3 Validation of the methodology 72

3.4 Illustrative examples 74

3.5 Discussion 79

4 Event detection in spatio-temporal data using a one-dimensional summary statistic **87**

4.1 Abstract 89

4.2 Introduction 89

4.3 Methods 91

4.4 Results 95

4.5 Discussion 111

5 Segmenting multiple sequences in parallel: three paradigms **118**

5.1 Abstract 120

5.2 Introduction 120

5.3 Methodology . . . 123
5.4 Illustrative Examples . . . 130
5.5 Conclusion . . . 140

6 A Study of Models for Zero-inflated Time Series Data 145
6.1 Abstract: . . . 147
6.2 Introduction . . . 147
6.3 Models for zero-inflated time series data . . . 149
6.4 Summary and Future directions . . . 165

7 Discussion, Conclusion and Future Directions 170

A Appendix Chapter 3 177
A.1 Details of Posterior Distribution . . . 177
A.2 Generalized Gibbs Sampling . . . 182
A.3 Supplementary Material A . . . 186
A.4 Supplementary Material B . . . 188

B Appendix Chapter 4 190
B.1 Summary of Bayesian change-point modeling with segmented ARMA model 190
B.2 Algorithm of the proposed method using PCA as a summary statistic . . . 193

C Appendix Chapter 5 194
C.1 Posterior Distributions . . . 194
C.2 Supplementary material . . . 203

Chapter 1

Introduction

A time series is a sequence of observations made for a well-defined variable at successive time-points. In many applications, the time points are equally spaced, although this is not a necessary requirement. For example, the total retail sales made by a specific company each month of the year would form a time series because sales revenue is a well-defined variable, consistently estimated at approximately equally spaced time intervals [1]. Time series are generally modelled as random processes that may be stationary or non-stationary. A non-stationary time series is a time series that exhibits temporal heterogeneity, thus statistical properties such as localised mean, variance and autocorrelations at various lags are not constant over time. For example, some time series are characterized by sudden changes of local statistical properties at certain time instants known as change-points, and constant local statistical properties in between change-points. Identification of change-points and modelling the segments of the time-series between change-points is an intensely investigated problem in time series analysis. Most of the literature on this topic (reviewed in Chapter 2) focuses on abrupt changes to the first few moments of the series (mean and variance). The aim is to detect locations of change-points and the amount of change in those moments over time.

The conclusions drawn from time-series data can be inaccurate if the lack of homogeneity is not taken into account [2]. Failing to detect change-points can also have a harmful impact on predictive performance. Recognition of these facts has led to an interest in change-point detection. As a result, it has become a beneficial tool in a diverse set of application fields including bioinformatics [3, 4], EEG analysis [5–7], finance [8, 9], industrial monitoring [10], signal processing [11], econometrics [12, 13] and disease demographics [14].

In its simplest form, the change-point problem considers whether one or more changes

occur in a time series and if so, the most likely times of any such changes. A more complex and difficult problem is to identify, locate and estimate multiple change-points in a time series. Both frequentist and Bayesian approaches have generated significant amounts of literature on estimation for multiple change-point models. This literature is surveyed in [10, 15–20], where numerous models and methods are suggested, discussed and a comprehensive list of references is reviewed. As an approach to change-point detection, the Bayesian approach is particularly appealing because it automatically achieves a compromise between model intricacy (quantified by the number of change points) and model fit. A significant advantage of Bayesian change-point detection approaches over frequentist approaches is that they quantify uncertainty about the number and location of change points [21]. Bayesian approaches also produce a probability distribution to quantify uncertainty in the parameters, rather than using a point estimate. In addition, for some frequentist approaches to detecting multiple change-points, the computational expense increases drastically with the number of change-points present in the time series. A combinative cost in the optimization task is induced by some forms of frequentist approach with maximum likelihood, and this may make calculations prohibitive whenever the number of change-points is more than two [22]. In contrast, several authors have used dynamic programming [8, 23] to render Bayesian approaches computationally feasible [24].

A variety of Bayesian change-point approaches have been developed and widely used in time-series segmentation. The Bayesian approach to multiple change-points dates back to the seminal paper of Chernoff and Zacks [25]. Here I highlight two aspects of Bayesian change-point models: one is the types of model used to model homogenous behaviour within segments and the other is the types of method used to simulate sampling from a posterior distribution. In Bayesian approaches, there are many different types of model that have been used to describe the behaviour of a sequence within a segment (between two successive change-points). There are so many candidate models that is infeasible to enumerate all of them here. As a result, I mention only those models of most direct relevance to this thesis, specifically those which used an autoregressive model (AR) for each segment [26–34]. These Bayesian approaches model a nonstationary time series by segmenting the series into blocks of different AR processes. However, the presence of autocorrelation in time series may sometimes lead to an excessively large estimate of the number of change-points.

This thesis introduces an autoregressive moving average (ARMA) model. ARMA models take into account the dependence between residual terms by adding a moving average

component, as discussed in Chapter 2. Using a segmented ARMA model in a Bayesian approach is advantageous in that it permits the fitting of more flexible Bayesian change-point models than is possible with an AR model. Pei-Gee [33] argues that ARMA models may be useful for analysing texture and contextual information in image segmentation. However, Pei-Gee used a two-dimensional AR model instead of an ARMA model to simplify the computation in his proposed time series segmentation approach.

The complexity of posterior distributions for parameters of change-point models usually necessitates specialised simulation methods. McCulloch and Tsay [35] expanded Chernoff-Zack's model of normal mean shift Gaussian autoregressive models with possible changes in level and error variance and used the Gibbs sampler to approximate the posterior distribution of time-varying parameters. Succeeding improvements of the Bayesian approach make use of reversible jump Markov chain Monte Carlo (MCMC) developed by Green [36], Metropolis-Hastings (MH) algorithm [37] or Gibbs sampling used in conjunction with Metropolis-Hastings steps, as in Albert and Chib [2], Chib [38], Liu and Lawrence [39], Wang and Zivot [40], Chib, Nardari and Shephard [41]. All these methods involved simulation-based inference via MCMC algorithms. The method of time series segmentation used in this thesis uses a highly efficient sampling technique (Generalized Gibbs Sampler, or GGS) to generate samples from a posterior distribution. This technique (developed by Keith *et al.* [42]) is similar to a conventional Gibbs sampler but provides an alternative to the reversible jump sampler. The dimension in this algorithm does not need to be fixed and it provides flexibility to sample from varying dimensional spaces.

Detection of multiple changes in multiple parallel time series has recently received much attention in time series segmentation. Segmentation of multiple sequences is an important and interesting but also a challenging problem because of its complexity. Three approaches are considered in the literature to segment multiple time series in parallel. The first is the simultaneous segmentation of all series where changes are in common among series [43–45], the second approach is the joint segmentation of all the series where each series has its own specific number and location of changes [46–49] and the third approach considers all series independently and compares the location of change-points between the series [44]. Segmenting multiple sequences instead of a single sequence potentially makes the inference of change-point locations more accurate, more precise and more sensitive. Also, some dependency may exist between multiple series and it is necessary to consider these dependencies in order to avoid false change-point detection [49]. Segmentation of multiple

sequences in parallel can potentially enhance robustness and power by pooling information across sequences [43]. These motivating reasons are relevant in an increasing number of applications in which analysis of multiple series at a time enhances the ability to infer an intricate underlying phenomenon.

Keeping the above points in mind, the main objective of this thesis is to:

'segment time series data with a Bayesian change-point segmented ARMA model and develop methods to segment multiple parallel time series'.

To achieve this objective, I used a Bayesian change-point segmented ARMA model throughout this research. This method considers the problem of modeling a time series by segmenting the series into blocks of autoregressive moving average (ARMA) processes. The model, methods and software were developed in collaboration with Assoc. Prof. Jonathan M. Keith and Dr. Sarah Boyd. In Chapter 2, I discuss: i) basic terminologies related to time series segmentation; ii) a Bayesian change-point model [42] used for segmenting binary sequences, which is closely related to the Bayesian change-point segmented ARMA model; iii) a wider review of models used in time series segmentation, a review of MCMC samplers used in time series segmentation and a review of the models used to segment multiple sequence in parallel; and iv) background to some other methods used in this thesis.

The focal point of Chapter 3 is to introduce and validate a Bayesian change-point segmented ARMA model and asses the performance of this model by segmenting one dimensional time series data. The Bayesian change-point segmented ARMA model has two novel features: i) a separate ARMA model for each segment and ii) a highly efficient technique (GGS) to generate samples from a posterior distribution. I demonstrate that this novel method provides a promising way to estimate the locations of change-points. To the best of my knowledge, the use of ARMA models in change-point analysis is an approach that is surprisingly absent from the existing literature. A key advantage of ARMA models over AR models is that the former can better account for the autocorrelations in a time series [1]. ARMA models not only express a time series as a linear function of its past values but also consider the dependence between residual terms by including a moving average component. In the model presented here, each segment in the time series is assumed generated by an ARMA model with different means but the same variance. I validated this method using Cook's method [50] and tested its performance by applying it to simulated data and to a well known real world dataset used in the literature for change-point detection. I applied

AR(1), MA(1) and ARMA(1,1) models to that real data and compared these three models using DICV values. Results obtained using simulated data demonstrated that fitting a segmented ARMA model potentially detects a larger number of change-points than the AR model and for the real world data, the segmented ARMA model identifies a greater number of change-points than have been detected by comparable methods in the existing literature illustrating the higher sensitivity of the proposed model. This work was published in the journal PLoS ONE [51] and the methodology was presented as a poster at the BioInfoSummer 2015 conference held in Sydney.

As this thesis is primarily aimed at developing improved methods for parallel segmentation, I explored methods and statistical resources currently available segmenting parallel sequences. This motivates the work of Chapter 4, where I propose a simple one-dimensional approach to segment parallel time series. This is an event detection approach for spatio-temporal data in presence of background noise using a one-dimensional summary statistic. Here I adapted the Bayesian change-point segmented ARMA model presented in Chapter 3 to detect such events. As the Bayesian change-point segmented ARMA model is designed for one-dimensional data, I pre-processed parallel spatially correlated time-series data to produce a single summary sequence. The goal of this chapter was to identify an event of interest in parallel sequences partially obscured by background noise, without explicitly segmenting the two-dimensional data. Instead of segmenting the two-dimensional image, I developed a complementary approach to segment parallel time series collected at roughly evenly spaced spatial locations along a line or curve. Before applying the Bayesian change-point segmented ARMA model, the data were pre-processed using the maximum over locations as a summary statistic for each time point. The performance of the method was evaluated via synthetic data as well as on real world data and the results compared with a simple event extraction method. Results of synthetic data found the model to be effective in identifying events in backgrounds with a variety of noise levels. Results obtained using real world data identified the boundary of an event with high probability in data with a low level of background noise. The method was less effective in identifying events in the presence of a high level of background noise, for which it found some false positive change-points in addition to true events. In the same chapter, I also investigated an alternative summary statistic, obtained using principal component analysis (PCA). As a summary statistic, the maximum produced better results than PCA in our examples. However, dimension reduction method using any summary statistic includes information loss,

and is therefore suitable only for preliminary exploratory analysis. Further development of this model is required to improve the sensitivity of this method in the presence of high noise variance. This work is currently submitted to a journal.

In Chapter 5, I propose a more general and flexible method for the segmentation of multiple sequences in parallel. This method is also a generalization of the Bayesian change-point segmented ARMA model presented in Chapter 3. In this chapter, I present three alternative models for simultaneous segmentation of multiple series in parallel. The first generalization assumes change-points occur in corresponding locations in all series, the second adds the additional assumption that all time-series have the same probabilities of assigning segments to a number of segment classes and the third assumes adds the further assumption that corresponding segments in each time-series belong to a common segment class. The performances of these three generalizations were assessed using simulated data and real world data. These alternative models were also compared using three information criteria (approximated AIC, BIC and DICV values) in real life examples. This work is currently submitted to a renowned journal.

The real world data used in Chapter 5 motivates the work in Chapter 6. The data set is taken from a sediment core and includes concentrations of thirty-seven metal elements detected at each depth (in mm) in the cores. The time series of these elements represent specific sediment characteristics such as magnetic susceptibility, organic matter, sediment particle size and elemental composition of the sediments [52, 53]. Most of the elements were not detected at all depths in the core, resulting in time-series that include zeros over extended time-periods. For this reason, I segmented only ten time-series in parallel in Chapter 5, specifically those that did not contain any zeros. The excessive number of zeroes in the remaining sequences directed my research interest to models for zero-inflated data. A variable is described as zero-inflated semicontinuous when it has a continuous distribution except for a probability mass at 0. In Chapter 6, I review models for zero-inflated data. This chapter will ultimately form the basis for a review article (in preparation) and provides an interesting direction for further research.

As this thesis is intended to be written in fulfillment of the requirement for 'Thesis by Publications', Chapter 3, Chapter 4, Chapter 5 and Chapter 6 are presented as journal articles formatted for their respective journals. The methods sections of Chapter 3, Chapter 4 and Chapter 5 partially overlap since they all use the same Bayesian change-point segmented ARMA model as the basic model. The bibliographies of these chapters are

incorporated at the end of each chapter.

In synopsis, the main objective of this thesis is to segment time series data with a Bayesian change-point segmented ARMA model and develop methods to segment multiple parallel sequences. To achieve this, I have investigated the following aspects:

1. Review literature to gain knowledge about the models used for time series segmentation, sampling techniques used in time series segmentation and about the models used to segment parallel sequence [Chapter 2].

2. Present a Bayesian change-point modelling approach where the data in segments are modeled by an autoregressive moving average (ARMA) model [Chapter 3].

3. Validate the segmented ARMA model and assess the performance of Bayesian change-point segmented ARMA model via simulated and real world data [Chapter 3].

4. Segment parallel spatially correlated time-series data using a simple approach based on a summary statistic [Chapter 4].

5. Segment multiple sequences in parallel using a generalization of the segmented ARMA model assuming the same change-points for all time series [Chapter 5].

6. Segment multiple sequences in parallel using a generalization of the segmented ARMA model assuming the change-points and probabilities of assigning segments to classes are identical for all time series [Chapter 5].

7. Segment multiple sequences in parallel using a generalization of the segmented ARMA model by considering parallel sequences have common segment classes in addition to the assumptions of the above two generalizations [Chapter 5].

8. Review models for zero-inflated time-series data [Chapter 6].

9. Discuss key findings, issues regarding methodological challenges, and future works [Chapter 7].

References

1. Hamilton, J. D. *Time series analysis* (Princeton university press Princeton, NJ, 1994).
2. Albert, J. H. & Chib, S. Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts. *Journal of Business & Economic Statistics* **11,** 1–15 (1993).
3. Bleakley, K. & Vert, J.-P. The group fused lasso for multiple change-point detection. *arXiv preprint arXiv:1106.4199* (2011).
4. Muggeo, V. M. & Adelfio, G. Efficient change point detection for genomic sequences of continuous measurements. *Bioinformatics* **27,** 161–166 (2010).
5. Barlow, J., Creutzfeldt, O., Michael, D., Houchin, J. & Epelbaum, H. Automatic adaptive segmentation of clinical EEGs. *Electroencephalography and Clinical Neurophysiology* **51,** 512–525 (1981).
6. Bodenstein, G. & Praetorius, H. M. Feature extraction from the electroencephalogram by adaptive segmentation. *Proceedings of the IEEE* **65,** 642–652 (1977).
7. Kaplan, A. Y. & Shishkin, S. L. in *Non-parametric statistical diagnosis* 333–388 (Springer, 2000).
8. Bai, J. & Perron, P. Computation and analysis of multiple structural change models. *Journal of applied econometrics* **18,** 1–22 (2003).
9. Talih, M. & Hengartner, N. Structural learning with time-varying components: tracking the cross-section of financial time series. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67,** 321–341 (2005).
10. Basseville, M., Nikiforov, I. V., *et al. Detection of abrupt changes: theory and application* (Prentice Hall Englewood Cliffs, 1993).
11. Fearnhead, P. Exact Bayesian curve fitting and signal segmentation. *IEEE Transactions on Signal Processing* **53,** 2160–2166 (2005).
12. Chen, J. & Gupta, A. K. Testing and locating variance changepoints with application to stock prices. *Journal of the American Statistical association* **92,** 739–747 (1997).
13. Potter, S. & Koop, G. Forecasting and Estimating Multiple Change-Point Models with an Unknown Number of Change Points. *FRB of New York Staff Report* (2004).

14. Denison, D. G. & Holmes, C. C. Bayesian partitioning for estimating disease risk. *Biometrics* **57,** 143–149 (2001).

15. Algama, M. & Keith, J. M. Investigating genomic structure using changept: A Bayesian segmentation model. *Computational and structural biotechnology journal* **10,** 107–115 (2014).

16. Jensen, U. & Lütkebohmert, C. Change-point models. *Encyclopedia of Statistics in Quality and Reliability* **1** (2008).

17. Reeves, J., Chen, J., Wang, X. L., Lund, R. & Lu, Q. Q. A review and comparison of changepoint detection techniques for climate data. *Journal of Applied Meteorology and Climatology* **46,** 900–915 (2007).

18. Aminikhanghahi, S. & Cook, D. J. A survey of methods for time series change point detection. *Knowledge and information systems* **51,** 339–367 (2017).

19. Rodionov, S. A brief overview of the regime shift detection methods. *Large-scale disturbances (regime shifts) and recovery in aquatic ecosystems: challenges for management toward sustainability,* 17–24 (2005).

20. Truong, C., Oudre, L. & Vayatis, N. Selective review of offline change point detection methods. *Signal Processing,* 107299 (2019).

21. Ruggieri, E. A Bayesian approach to detecting change points in climatic records. *International Journal of Climatology* **33,** 520–528 (2013).

22. Harlé, F., Chatelain, F., Gouy-Pailler, C. & Achard, S. Bayesian Model for Multiple Change-points Detection in Multivariate Time Series. *arXiv preprint arXiv:1407.3206* (2014).

23. Lung-Yut-Fong, A., Lévy-Leduc, C. & Cappé, O. Homogeneity and change-point detection tests for multivariate data using rank statistics. *arXiv preprint arXiv:1107.1971* (2011).

24. Paquet, U. Empirical Bayesian change point detection. *Graphical Models* **1995,** 1–20 (2007).

25. Chernoff, H. & Zacks, S. Estimating the current mean of a normal distribution which is subjected to changes in time. *The Annals of Mathematical Statistics,* 999–1018 (1964).

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

27. Lai, T. L. & Xing, H. A simple Bayesian approach to multiple change-points. *Statistica Sinica,* 539–569 (2011).

28. Lai, T. L., Liu, H. & Xing, H. Autoregressive models with piecewise constant volatility and regression parameters. *Statistica Sinica,* 279–301 (2005).

29. Joseph, L., Vandal, A. C. & Wolfson, D. B. Estimation in the multipath change point problem for correlated data. *Canadian Journal of Statistics* **24,** 37–53 (1996).

30. Wood, S., Rosen, O. & Kohn, R. Bayesian mixtures of autoregressive models. *Journal of Computational and Graphical Statistics* **20,** 174–195 (2011).

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

32. Ruanaidh, J. J. O. & Fitzgerald, W. J. *Numerical Bayesian methods applied to signal processing* (Springer Science & Business Media, 2012).

33. Ho, P.-G. *Image segmentation* (BoD–Books on Demand, 2011).

34. Punska, O., Doucet, C. A. A., Fitzgerald, W., Andrieu, C. & Doucet, A. Bayesian segmentation of piecewise constant autoregressive processes using MCMC methods (1999).

35. McCulloch, R. E. & Tsay, R. S. Bayesian inference and prediction for mean and variance shifts in autoregressive time series. *Journal of the American Statistical Association* **88,** 968–978 (1993).

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

37. Billio, M., Monfort, A. & Robert, C. P. Bayesian estimation of switching ARMA models. *Journal of econometrics* **93,** 229–255 (1999).

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

39. Liu, J. S. & Lawrence, C. E. Bayesian inference on biopolymer models. *Bioinformatics (Oxford, England)* **15,** 38–52 (1999).

40. Wang, J. & Zivot, E. A Bayesian time series model of multiple structural changes in level, trend, and variance. *Journal of Business & Economic Statistics* **18,** 374–386 (2000).

41. Chib, S., Nardari, F. & Shephard, N. Markov chain Monte Carlo methods for stochastic volatility models. *Journal of Econometrics* **108,** 281–316 (2002).

42. Keith, J. M., Kroese, D. P. & Bryant, D. A generalized Markov sampler. *Methodology and Computing in Applied Probability* **6,** 29–53 (2004).

43. Zhang, N. R., Siegmund, D. O., Ji, H. & Li, J. Z. Detecting simultaneous changepoints in multiple sequences. *Biometrika* **97,** 631–645 (2010).

44. Cleynen, A. & Robin, S. Comparing change-point location in independent series. *Statistics and Computing* **26,** 263–276 (2016).

45. Barigozzi, M., Cho, H. & Fryzlewicz, P. Simultaneous multiple change-point and factor analysis for high-dimensional time series. *Journal of Econometrics* **206,** 187–225 (2018).

46. Dobigeon, N., Tourneret, J.-Y. & Davy, M. Joint segmentation of piecewise constant autoregressive processes by using a hierarchical model and a Bayesian sampling approach. *Signal Processing, IEEE Transactions on* **55,** 1251–1263 (2007).

47. Dobigeon, N., Tourneret, J.-Y. & Scargle, J. D. Joint segmentation of multivariate astronomical time series: Bayesian sampling with a hierarchical model. *IEEE Transactions on Signal Processing* **55,** 414–423 (2007).

48. Picard, F., Lebarbier, É., Budinská, E. & Robin, S. Joint segmentation of multivariate Gaussian processes using mixed linear models. *Computational Statistics & Data Analysis* **55,** 1160–1170 (2011).

49. Collilieux, X., Lebarbier, E. & Robin, S. A factor model approach for the joint segmentation with between-series correlation. *Scandinavian Journal of Statistics* **46,** 686–705 (2019).

50. Cook, S. R., Gelman, A. & Rubin, D. B. Validation of software for Bayesian models using posterior quantiles. *Journal of Computational and Graphical Statistics* **15,** 675–692 (2006).

51. Sadia, F., Boyd, S. & Keith, J. M. Bayesian change-point modeling with segmented ARMA model. *PloS one* **13,** e0208927 (2018).

52. Lintern, A. *et al.* Sediment cores as archives of historical changes in floodplain lake hydrology. *Science of the Total Environment* **544,** 1008–1019 (2016).

53. Wolfe, B. B. *et al.* Reconstruction of multi-century flood histories from oxbow lake sediments, Peace-Athabasca Delta, Canada. *Hydrological Processes: An International Journal* **20,** 4131–4153 (2006).

Chapter 2

Background and Literature Review

2.1 Introduction

This chapter is organized into four sections. First, I discuss some background definitions related to segmentation: I define time series segmentation and change-point detection, review literature about previously used models in time series segmentation and give an overview of a Bayesian change-point model. The second section reviews the Markov chain Monte Carlo (MCMC) method, different types of MCMC sampler, the steps to sample from the posterior distribution of a Bayesian change-point model and includes a discussion of literature about different types of sampler used for time series segmentation. In the third section, I discuss literature about previously used methods to segment parallel sequences. Lastly in the fourth section, I review some models and methods that are used in the upcoming chapters of this thesis.

2.2 Some Background to Segmentation

2.2.1 Time Series Segmentation

A time series is a sequence of measurements made at consecutive (and usually equally spaced) points in time. Some time series can be partitioned into discrete segments, each with characteristic properties. That is, time series segmentation decomposes time series into homogeneous segments consisting of similar observations, different from those of neighbouring segments [1]. Segmentation is useful in such disparate fields as bioinformatics, industrial monitoring, audiovisual data and financial data [2]. This process of partitioning a sequence

into individual segments with an intention of revealing changes in the properties of the sequence is known as time series segmentation. Examples of application of time-series segmentation include speaker segmentation, in which an audio recording signal is divided into segments to identify who is speaking at different times [3, 4]; segmenting time series of stock market trajectories to quantify the impact of important world events [5]; segmentation of handwriting to identify the several words or letters of which it is composed [6], and many more. The goals of segmentation may include to: i) identify time intervals of stability and homogeneity in the behavior of the process; ii) delineate the time instants of change; iii) identify characteristics of each segment; and iv) use these facts and information to ascertain shapes and patterns in a nonstationary time series [7]. Overall, the main intent of time-series segmentation is to pinpoint the segment boundary points in the time-series, and to determine the dynamical properties corresponding to each segment.

2.2.2 Change-point detection in time series

Change-point detection refers to the problem of detecting changes and determining the times of changes in stochastic processes [8]. Change-point detection methods can be a way to segment time series data, as the purpose of such methods is to segment a succession of observations by selecting a series of change-point locations such that the observations are, in some sense, homogeneous within segments and heterogenous between segments. Mathematically, if we have an ordered sequence of observations $x_{1:T} = (x_1, x_2, \cdots, x_T)$, a change-point is said to be present at time point s where $s \in \{1, 2, \cdots, T-1\}$ if the statistical properties of $x_1, \cdots, x_s$ and $x_{s+1} \cdots x_T$ differ in some way. This definition refers to a single change-point; if we extend the concept to multiple change-points we now have a number of change-points k with positions, $s_{1:k} = (s_1, \cdots, s_k)$. Here, every change-point position is an integer between 1 and $T-1$ inclusive. We assume $s_0 = 0$ and $s_{k+1} = T$, and the change-points are ordered so that $s_i < s_j$ if, and only if, $i < j$. The k change-points will divide the data into $k+1$ segments where the i-th segment holds data $x_{(s_{i-1}+1):s_i}$. Each segment has a corresponding set of parameters. Change-point detection methods represent the data by finding the number of segments and by estimating the values of parameters associated with each segment [9].

Figure 2.1 shows common four types of change-point problems. In the first row, segments differ in mean. In the second row, segments differ in variance. In the third row, segments differ in their internal patterns of autocorrelation. The segments shown in the fourth row

differ in the entire form of their probability distribution [10]. In this thesis, I consider those problems where changes are present in the mean and variance.

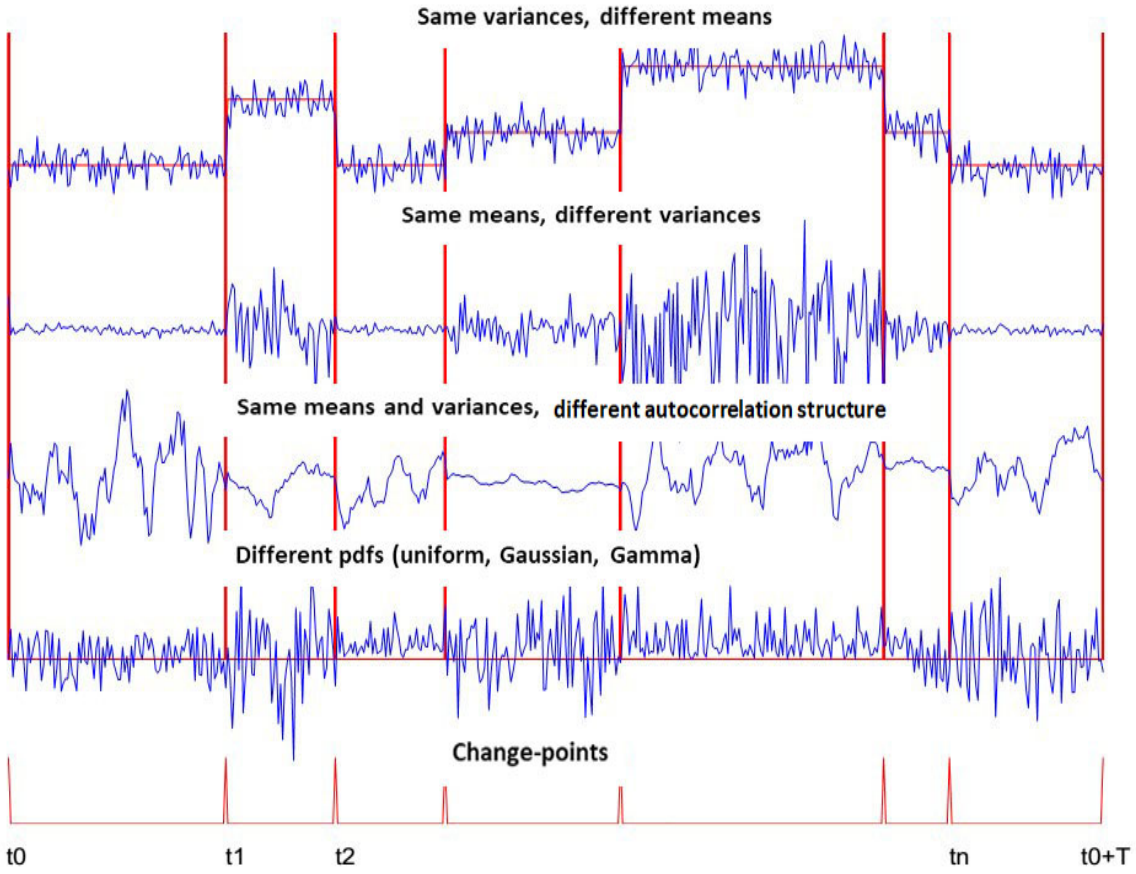


Figure 2.1: Types of common change-point problems. First row: changes in the mean, second row: changes in the variance, third row: changes in the autocorrelation and fourth row: changes in the shape of the probability distribution.

2.2.3 Change-point detection model

There are many models that have been used to detect change-points in time-series and it is beyond the scope of this thesis to enumerate all of them. Miscellaneous methods of non-, semi- and fully parametric change-point modeling have flourished [11–17].

Models for time series segmentation

Change-point detection in time series segmentation has been the subject of intensive research in the past half-century and the literature on this subject is extensive (see reviews, [11, 14–16]). Here I review some change-point detection approaches focusing on Bayesian approaches. Chernoff and Zacks [18] introduced the first Bayesian approach to change-point detection in 1964, in which they inferred the mean of a normal distribution for each of two

segments in a Bayesian framework. They considered $X_1, X_2, \cdots, X_n$ independent and normally distributed random variables with means $\mu_1, \mu_2, \cdots, \mu_n$ and variance 1. Every mean at each time point μ_i was the same as the mean at previous time point μ_{i-1}, except at a single change-point. The focus of their approach was to estimate the final mean μ_n.

They made three assumptions: i) the time point when a change occurs follows an arbitrary specified *a priori* probability distribution, ii) the difference between the means is normally distributed with mean 0 and variance σ^2 and iii) the final mean μ_n is a normally distributed random variable with means 0 and variance τ^2. They derived a Bayes estimator of μ_n for a quadratic loss function by letting $\tau^2 \to \infty$ and a minimum variance linear unbiased estimator (MVLU) of μ_n based on the previous three assumptions. They showed the Bayes estimator was more efficient than the MVLU estimator if the location of the change-point is neither 0 nor $n-1$. The Bayes estimator is in general very difficult to apply as it involves many computations. Consequently, they considered a simplification of the general Bayes estimator formula by assuming the *a priori* variance of the changes, $\sigma^2 \to 0$. However, the simplified estimator was not efficient when the magnitude of changes was large. Then they studied the case where the *a priori* distribution of change-point times is such that there is at most one change-point and derived a simple Bayes estimator called AMOC (at most one change) under this assumption. For sequences with two change-points, they developed an '*ad hoc*' estimator in order to avoid the shortcomings of the AMOC estimator. The *ad hoc* estimator used the AMOC Bayes estimator in combination with a sequence of tests intended to identify the last time point of change. These estimators were then compared using Monte Carlo computations [18]. This method ignores the dynamic and compound nature of the problem. That is, in some real applications, data becomes available sequentially, with each new observation requiring a new estimate for the mean of the last observation. If the times between observations are short, the AMOC and *ad hoc* estimators may be too computationally expensive to re-estimate for each new observation.

Smith [19] also considered the problem of finding a single change-point in a finite series of independent observations. He considered three cases: i) the initial parameter distribution and the changed parameter distribution both are known; ii) only the initial distribution is known; and iii) both distributions are unknown [20]. He introduced Bayesian approaches for each of these cases, in which the probability distributions of the observed signals are binomial or normal.

To model the relationship between a response and an explanatory variable, Fearnhead

[21] proposed a novel algorithm for exact Bayesian inference in regression models. This method models the functional relationship between the response and explanatory variables as a sequence of independent linear regressions on disjoint segments. This method uses an efficient dynamic programming algorithm to exactly estimate the posterior distribution over the number and location of change-points in one-dimensional time series. An unknown number of segments and an unknown model order for the linear regressions within each segment was allowed. He first estimated the joint posterior distribution of the number and positions of change-points using dynamic programming and then sampled change-points from this posterior distribution using a perfect simulation algorithm. This algorithm makes use of the independence between segments and the Markovian nature of the time-series within segments. The algorithm includes a recursion for the probability of the data from time t onwards, conditional on a change-point immediately before time t, in terms of the equivalent probabilities for all times after t. When these probabilities have been calculated for all time-points, it is possible to directly simulate from the posterior distribution of the time of the first change-point, and then the conditional distribution of the time of the second change-point, given the first, and so on. An advantage of the perfect simulation algorithm is that it takes independent samples from the posterior distribution and avoids the problems of diagnosing convergence encountered by other MCMC methods. However, the method is not suitable for comparison of models involving different numbers of change-points.

Punskaya *et al.* [4] developed a Bayesian method for estimating the parameters of an assumed functional relationship between response and explanatory variables given noisy measurements. In their model, certain parameters are piecewise constant functions of time, thus dividing the time series into segments within which standard models (such as autoregressive or Lotka-Volterra models) can be applied, with the number and locations of change-points being additional parameters to be estimated. Mathematically, for any generic sequence κ_t, they defined $\boldsymbol{\kappa}_{i:j} \triangleq (\kappa_i, \kappa_{i+1}, \cdots \kappa_j)^T$ and let $\mathbf{y}_{0:T-1}$ be a vector of T real observations. The elements of $\mathbf{y}_{0:T-1}$ were assumed generated by one of the models $\mathbf{M}_{k,\mathbf{p}_k}$ where the signal is in the form of a linear regression model with piecewise constant parameters and k ($k \in 0, \cdots, k_{max}$) change-points. That is,

$$\mathbf{M}_{k,\mathbf{p}_k} : \mathbf{y}_{\tau_i:\tau_{i+1}-1} = \mathbf{G}_i^{(p_i)} \boldsymbol{\beta}_i^{(p_i)} + \mathbf{n}_{\tau_i:\tau_{i+1}-1}, \quad i = 0, \cdots, k. \tag{2.1}$$

where, $\boldsymbol{\beta}_i^{(p_i)}$ is a vector of p_i model parameters for the ith segment $i = (0, \cdots, k)$ and $\mathbf{n}_{\tau_i:\tau_{i+1}-1}$ is a vector of i.i.d. Gaussian noise samples of variance σ_i^2 associated with the

ith model. The change-points of the model $\mathbf{M}_{k,\mathbf{p}_k}$ are arranged in the vector $\boldsymbol{\tau}_k \triangleq \boldsymbol{\tau}_{1:k}$ and they adopted $\tau_0 = 0$ and $\tau_{k+1} = T - 1$ for notational convenience. They also denoted $\boldsymbol{\sigma}_k^2 \triangleq \boldsymbol{\sigma}_{0:k}^2$ and $\mathbf{p_k} \triangleq \mathbf{p_{0:k}}$, where $p_i = 0, \cdots, p_{max}$. The matrix $\mathbf{G}_i^{(p_i)}$ is the matrix of the basis functions for the ith segment $i = (0, \cdots, k)$. For the piecewise polynomial model, $\mathbf{G}_i^{(p_i)}$ is of the following form:

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

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

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

The orders of the different linear regression models $\mathbf{p}_{0:k}$ were assumed equal and unknown, that is, $p_i = p_j = p_0$ for any $(i, j) \in \{0, \cdots, k\}$. The number of change-points k and the associated parameters $\boldsymbol{\Psi}_{k,\mathbf{p}_k} \triangleq (\boldsymbol{\tau}_k, \mathbf{p}_k, \{\boldsymbol{\beta}_i^{(p_i)}\}_{i=0}^k, \boldsymbol{\sigma}_k^2)$ were also unknown. Given $\mathbf{y}_{\tau_0:T-1}$, their aim was to estimate k and $\boldsymbol{\Psi}_{k,\mathbf{p}_k}$. They used joint prior distributions for the number of the change-points, their locations and the unknown orders of the linear regression models within each segment. They also developed hierarchical prior distributions in which the hyperparameters were presumed random with a vague prior distribution. Reversible jump Markov chain Monte Carlo (MCMC) methods were used to estimate the resulting posterior probability distributions, as these did not admit closed-form analytical expressions. The proposed approach is flexible and can be used to compute the predictive distribution from the MCMC samples. This algorithm can be extended for a more general noise distribution or for multivariate signals.

Another piecewise autoregressive (AR) process was suggested by Davis *et al.* [22]. They allowed for an unknown number and locations of segments and also unknown orders of the respective AR processes. They provided an algorithm for finding the "best"-fitting model from the class of piecewise AR processes, that is, the "best" combination of the number

of breakpoints, the lengths of the segments, and the orders of the piecewise AR processes. To achieve this, they proposed an automatic piecewise autoregressive modeling procedure, referred to as Auto-PARM. The minimum description length principle was used to quantify the fit of the model to the data. The "best" combination was defined as the optimizer of an objective function and the authors used a genetic algorithm to solve this optimization problem. After specifying the number of change-points, their positions and the order of the respective AR process, they determined the maximum likelihood estimates of the AR parameters for each segment. They also considered the segmentation of multivariate time series data.

To segment possibly nonstationary time series data, a class of time-domain models (models used to measure variation of amplitude of signal with time known as time-domain model) was developed by Wood *et al.* [23]. Models in this class are structured as a mixture of time series, each with fixed and unknown parameters, and time-varying mixture probabilities. They estimated the number of mixture components using the data. In particular, the authors studied mixtures of autoregressive models with unknown but finite lags and an unknown number of components. The data set was divided into small non-overlapping segments such that all observations within one segment were always assigned to the same component. The model parameters, including the number of mixture components, were estimated using Markov chain Monte Carlo methods. Many current methods in time series analysis whose parameters change over time are based on segmentation of the time series [4, 22, 24]. Another approach which permits the parameters to change is to model their evolution [25, 26]. The method suggested by Wood [23] differs from these methods in that it does not determine which parameters change and which do not and it also doesn't specify the model for parameter evolution to allow for structural changes. This is a significant advantage of Wood's method, because in any model with more than a few parameters, it is likely that not all parameters will evolve in the same way or that not all parameters change abruptly at the same time. Moreover, sometimes it is hard to model the evolution of some parameters. In such cases, this method formulation allows some of the parameters to be the same over time by making them common across all components.

The observations in change-point detection problems can not in general be assumed to be independent. The autocovariance structure of a time series may exhibit complicated patterns of dependence and this needs to be considered in change-point estimation. However, considering the dependence structure may invalidate the classical inference approach

of the independent case. Chakar *et al.* [27] proposed a robust approach for detecting multiple change-points in the mean of a Gaussian AR(1) process by taking into account the dependency structure. The proposed method involves two phases: i) creating a robust estimator of the autocorrelation parameter to whiten the original series (that is, a transformed time series which behaves like statistical white noise) and ii) estimating change-points in the mean of the now approximately independent random variables. They considered the segmentation of an AR(1) process with homogeneous auto-correlation coefficient ρ^*:

$$y_i = \mu_k^* + \eta_i, \;\; t_{n,k}^* + 1 \leq i \leq t_{n,k+1}^*, \;\; 0 \leq k \leq m^*, \;\; 1 \leq i \leq n, \tag{2.2}$$

where, $(\eta_i)_{i\in\mathbb{Z}}$ is a zero-mean stationary AR(1) Gaussian process and $\eta_i = \rho^*\eta_{i-1} + \epsilon_i$. Here, $|\rho^*| < 1$, the ϵ_i's are i.i.d. zero-mean Gaussian random variables with variance σ^{*^2} and y_0 is a Gaussian random variable with mean μ_0^* and variance $\sigma^{*^2}/(1-\rho^{*^2})$. The purpose of their methodology was to estimate both the change-point locations $\mathbf{t_n^*} = (t_{n,k}^*)_{1\leq k\leq m^*}$ and the means $\boldsymbol{\mu}^* = (\mu_k^*)_{0\leq k\leq m^*}$, considering the presence of autocorrelation ρ^*. Before performing segmentation, they proposed to estimate ρ^*, but the estimation is more complex in the presence of change-points. Consequently, they considered the data observed at the change-point positions as outliers and proposed an estimate of ρ^* which was robust to the existence of such outliers. For the segmentations, they developed a criterion equivalent to the classic least-squares and applied it to a decorrelated version of the series, calculated using the estimated ρ^*. They showed that the resulting change-point estimators have the same asymptotic properties as the classical estimators in the independent framework, indicating that the performances of the estimators were not influenced by the dependence assumption. Finally, the authors identified the number of segments by proposing a model selection criteria. This method is computationally effective and displays better performance on finite sample size data than established approaches that do not consider the dependence structure of the time series.

Chan and Tong [28] suggested a new class of non-linear models to consider the problem of regime switching behaviour in the conditional mean. They named these Smooth Transition Autoregressive (STAR) models. STAR models are an expansion of Threshold Autoregressive (TAR) models, which were proposed by Tong and Lim [29–31]. The TAR model is a piecewise linear model consisting of two or more regimes of linear submodels. The TAR model uses an indicator variable representing a switch from one regime to another and taking a value zero or one, conditioning upon the values of a transition variable

and threshold parameter. If the transition variable is a lagged endogenous variable, a special class of the TAR model is produced, called the Self Exiting Threshold Autoregressive (SETAR) model [30] and specified by the following equation:

$$y_t = \mathbf{X^T}\boldsymbol{\psi}^{(j)} + \sigma^{(j)}\epsilon_t, \quad r_{j-1} < z_t < r_j. \tag{2.3}$$

where, $\mathbf{X^T} = (1, y_{t-1}, y_{t-2}, \cdots, y_{t-p})$ is a column vector of variables, j is the indicator variable, $y_t = \psi_0 + \psi_1 y_{t-1} + \cdots + \psi_a y_{t-a} + \epsilon_t$, $\boldsymbol{\psi}$ is the vector of autoregressive parameters, $\epsilon_t \sim^{iid} WN(0,1)$, z_t is a transition variable and $-\infty = r_0 < r_1 < \cdots < r_k = +\infty$ are $k+1$ non-trivial thresholds dividing the domain of z_t into k different regimes.

Chan and Tong [28] replaced the indicator variable of the TAR model with a smooth transition function in their STAR model, reasoning that abrupt jumps from one regime to another may not be an adequate representation of the underlying mechanism generating observed data. A smooth transition between regimes may be of use when the transition from one regime to another arises as a result of several actions that occur over time. The STAR model can be presented as follows:

$$y_t = \mathbf{X^T} + F(z_t, \zeta, c)\mathbf{X^T} + \sigma^{(j)}\epsilon_t. \tag{2.4}$$

where, $F(z_t, \zeta, c)$ is the transition function bounded between 0 and 1, ζ is the smoothing parameter and c is the location parameter. In a Bayesian framework, Péguin-Feissolle [32] presented an estimation and prediction procedure for a general non-linear model and then described it using an LSTAR model with simulated data. A comparison of non-linear models with regime changes in the conditional mean, namely Markov switching autoregressive (MSAR) models, TAR and STAR models, was presented by Potter [33]. The review included estimation techniques from a classical and Bayesian perspective as well as a review of some of the parametric tests for non-linearity.

Traditional time series modelling assumes a constant conditional variance. Engle (1982) [34] introduced the Autoregressive Conditional Heteroscedastic (ARCH) model which allows the conditional variance to change over time. This revolutionary new class of models could be applied to processes with non-constant variances, conditional on the past, but required constant unconditional variances. The ARCH model was generalised by Bollerslev (1986) [35], who introduced past conditional variances into the model's conditional variance equation. The resulting model is known as the Generalised Autoregressive Conditional

Heteroscedastic (GARCH) model and has been widely applied in several fields, though most commonly in economics and finance. GARCH models have been used for Value at Risk estimation [36, 37], and to estimate volatility for financial markets and indices [38–40]. GARCH models were also used to model speech signals by Cohen (2004) [41]. Estimation of GARCH model parameters has also been performed in Bayesian framework. Bauwens and Lubrano [42] applied a Griddy-Gibbs sampler for inference of the parameters of GARCH models with student errors. They showed that their method is feasible and competitive for importance sampling and the Metropolis-Hastings algorithm.

Hidden Markov models (HMMs) are well-established as an efficient approach in time series segmentation. A HMM is a statistical model in which the observed time series is fractionated into segments and the time series within each segment is modeled by a Markov process with hidden states. In each segment, hidden states determine the transition probabilities and the transitions between these hidden states occur at segment ends [11].

Albert and Chib [43] considered HMMs for regime switching problems. Regime shifts were conceptualised as the outcome of an unobserved two-state indicator variable and modeled by a Markov process with unknown transition probabilities. They developed the model in a Bayesian framework in which the unobserved states, one for each time point, were considered as missing data and these unobserved states were analyzed using Gibbs sampling. This method is useful, as the conditional posterior distribution of the parameters, given the states, and the conditional posterior distribution of the states, given the parameters, have an amenable structure for Monte Carlo sampling. This method is straightforward and generates the marginal posterior distributions for all parameters of interest. The authors also obtained the posterior distributions of the states, future observations, and the residuals, averaged over the parameter space. Bayesian HMM approaches are computationally intense and are generally impracticable for segmenting large sequences, without simplifying heuristics.

Kehagias [44] also used HMMs for the segmentation of hydrological and enviromental time series data. In this model, the unobservable state process is Markovian and can take a finite number of values $1, 2 \cdots, k$. At every time step, the state process can either remain the same or increase by one and the observable process generates a sample from a normal distribution with mean value depending on the current state. They assumed that the time series $\mathbf{x} = (x_1, x_2, \cdots, x_T)$ is a realization of the observable process and represent each time interval during which the hidden state doesn't change as a separate segment. Under this

correspondence, the segmentation problem is simplified to estimating the underlying state sequence $z_1, z_2, \cdots, z_T$.

Successive generalizations and expansions of Bayesian methods of change-point detection problems include [45–48] and many others.

Bayesian change-point detection model

Now I describe one Bayesian change-point modeling approach [49] in detail that is closely related to the Bayesian change-point segmented ARMA model which is used throughout my thesis to achieve my objectives. This Bayesian change-point model generates binary sequences.

Let T be the length of the binary sequence and let ϕ be the probability of starting a new segment, assuming change-point times were generated by a sequence of Bernoulli trials at each time point to decide whether to start a new segment. The probability of generating a segmentation with k change-points at positions $s = (s_1, \cdots, s_k)$ is given by

$$p(k, s|\phi) = \phi^{k-f}(1-\phi)^{T-1-k}. \tag{2.5}$$

where f is the number of fixed change-points. Fixed change-points are the boundaries between the segments of sequences that can not be moved or removed. Suppose that T and f are given. A decision is made whether to start a new segment at each position in the sequence except the first and those immediately following fixed change-points. Set $s_0 = 1$ and $s_{k+1} = T + 1$.

Segments are classified into groups that share similar properties. Each segment is assigned to one of ϑ groups. Denote the group to which segment j is assigned by $g_j \in \{0, 1, \cdots, \vartheta - 1\}$ and let π_t be the probability of assigning a segment to group t. Set $g = (g_0, \cdots, g_k)$. Then the probability of a specific assignment of the segments such that x_0 segments are assigned to group '0', . . . , $x_{\vartheta-1}$ segments are assigned to group 'ϑ-1' is :

$$p(\mathbf{g}|k, \pi) = \pi_0^{x_0} \cdots \pi_{\vartheta-1}^{x_{\vartheta-1}} = \prod_{j=0}^{k} \pi_{g_j}. \tag{2.6}$$

Let θ_j be the probability of generating a '1' at each position of segment j. For each segment j, the parameter θ_j is drawn from a beta distribution with as yet unspecified parameters $\alpha_0^{(t)}, \alpha_1^{(t)}$, where $t = g_j$, that is,

$$p(\theta_j|\alpha_0^{(t)},\alpha_1^{(t)}) = B(\theta_j|\alpha_0^{(t)},\alpha_1^{(t)}) = \frac{\Gamma(\alpha_0^{(t)}+\alpha_1^{(t)})}{\Gamma(\alpha_0^{(t)})\Gamma(\alpha_1^{(t)})}\theta_j^{\alpha_1^{(t)}-1}(1-\theta_j)^{\alpha_0^{(t)}-1}. \tag{2.7}$$

Set $\theta = (\theta_0,\cdots,\theta_k), \alpha^{(t)} = (\alpha_0^{(t)},\cdots,\alpha_k^{(t)})$ and $\alpha = (\alpha^{(0)},\cdots,\alpha^{(\vartheta-1)})$. Finally, the binary sequence within each segment is generated by independent Bernoulli trials at each position in the segment. The probability that segment j contains a specific sequence X_j including m_j zeros and n_j ones is :

$$p(X_j|T_j,\theta_j) = \theta_j^{n_j}(1-\theta_j^{m_j}). \tag{2.8}$$

where $T_j = s_{j+1} - s_j$ is the length of segment j. Also note, $m_j + n_j = T_j$. The final binary sequence S is obtained by concatenating $X_0,\cdots,X_k$. Thus, the joint distribution of k, s, g, θ and X is given by:

$$p(k,s,g,\theta,S|\phi,\pi,\alpha) = p(k,s|\phi)p(g,k|\pi)\prod_{j=0}^{k} B(\theta_j,\alpha^{(g_j)})p(X_j|T_j,\theta_j). \tag{2.9}$$

Figure 2.2 illustrates the conditional dependencies of the parameters.

One must assign prior probabilities to the parameters to complete the Bayesian model. For π and ϕ, a uniform prior $p(\pi) = p(\phi) = 1$ is assigned on the interval [0,1]. For $\alpha^{(t)}$, uniform priors are assigned on mean $\mu^{(t)}$ and standard deviation $\sigma^{(t)}$ of the beta distribution, given by $\mu_i^{(t)} = \frac{\alpha_i^{(t)}}{\alpha_0^{(t)}+\alpha_1^{(t)}}$ and $\sigma^{(t)} = \sqrt{\frac{\mu_0^{(t)}\mu_1^{(t)}}{\alpha_0^{(t)}+\alpha_1^{(t)}+1}}$.

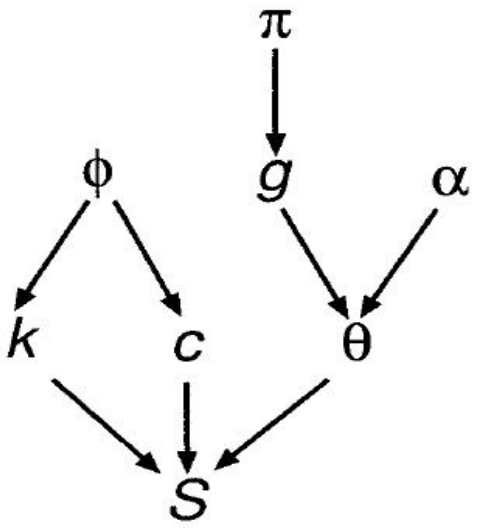


Figure 2.2: The parameters of the model with their dependencies

One can obtain the following posterior distribution by integrating over ϕ and θ and summing over g with the help of Bayes rule.

$$p(k, s, \pi, \mu, \sigma) \propto \Gamma(T-k)\Gamma(k-f+1)\prod_{j=0}^{k} f(m_j, n_j|\pi, \alpha). \tag{2.10}$$

where, $\mu = (\mu^{(0)}, \mu^{(1)}, \cdots, \mu^{(\vartheta-1)})$ and $\sigma = (\sigma^{(0)}, \sigma^{(1)}, \cdots, \sigma^{(\vartheta-1)})$, α is a function of μ and σ and

$$f(m, n|\pi, \alpha) = \sum_{t}\left[\pi_t \frac{\Gamma(\alpha_0^{(t)}+\alpha_1^{(t)})}{\Gamma(\alpha_0^{(t)})\Gamma(\alpha_1^{(t)})} \times \frac{\Gamma(m+\alpha_0^{(t)})\Gamma(n+\alpha_1^{(t)})}{\Gamma(m+\alpha_0^{(t)}+n+\alpha_1^{(t)})}\right]. \tag{2.11}$$

2.3 Sampling

2.3.1 Markov Chain Monte Carlo Method (MCMC)

Markov Chain Monte–Carlo (MCMC) is a popular method for acquiring information about distributions, especially posterior distributions in Bayesian inference. Bayesian analysis often requires the integration of intricate and high-dimensional functions including the estimation of i) the normalising constant of proportionality in Bayes' theorem ii) the marginal distributions and iii) inferences in the form of posterior expectations. Explicit calculation of such intricate integrals are often intractable or computationally intense even with powerful computational resources. MCMC methods provide an alternative way of performing such computation by sampling from the posterior distribution and estimating quantities of interest using those simulated samples [50]. MCMC combines two concepts: Monte Carlo integration and Markov chains.

Monte Carlo is the method of estimating properties of a distribution by analyzing random samples from the distribution. Let $\int_a^b h(\theta)d\theta$ be an integral to be computed. Then $h(\theta)$ can be decomposed into a product of two functions, $f(\theta)$ and $p(\theta)$, using the Monte Carlo method, where $f(\theta)$ is a function of θ and $p(\theta)$ is a probability density function defined over the interval (a, b). Then the original integral $\int_a^b h(\theta)d\theta$ can be written as an expectation of $f(\theta)$ over the density $p(\theta)$ as follows:

$$\int_a^b h(\theta)d\theta = \int_a^b f(\theta)p(\theta)d\theta = E_{p(\theta)}[f(\theta)]. \tag{2.12}$$

The Monte Carlo method draws a sufficiently large number of random samples from the density $p(\theta)$, so that the original integration can be estimated as follows:

$$\int_a^b h(\theta)d\theta = E_{p(\theta)}[f(\theta)] \simeq \frac{1}{n}\sum_a^b f(\theta_i). \tag{2.13}$$

The Markov chain element of MCMC is the concept that random samples are generated using a sequential process. Each random sample is used as a stepping stone to generate the next random sample. A defining property of Markov chains is that, conditional on the chain's present state, future states are independent of the past states (this is the "Markov" property). A Markov chain is defined to be a sequence of random variables $\theta_1, \cdots, \theta_n$ such that θ_{n+1} is conditionally independent of $\theta_1, \cdots, \theta_{n-1}$, given θ_n, that is [51],

$$p(\theta_{n+1}|\theta_1, \cdots, \theta_n) = p(\theta_{n+1}|\theta_n), \tag{2.14}$$

where p may represent a probability or a probability density. The range of possible values for the random variable θ is known as the state space of the Markov chain [51]. A Markov chain has stationary transition probabilities if the conditional distribution of θ_{n+1} given θ_n does not depend on n. Let $f(\theta)$ be a target probability distribution of a quantity of interest on a target state space S. If $f(\theta)$ can not be sampled directly, then the MCMC method involves forming a Markov chain in the target state space S such that its stationary distribution is the same as the target posterior distribution. In a countable state space, the distribution f is stationary with respect to a transition matrix $\mathbf{P} = p_{ij}$, if $f\mathbf{P} = f$. The transition matrix $\mathbf{P}$ consists of transition probabilities:

$$p_{ij} = p(\theta_{n+1} = j|\theta_n = i), \quad i = 1, 2 \cdots, n \text{ and } j = 1, 2 \cdots, n. \tag{2.15}$$

The concept of a stationary distribution can be generalised in a straightforward manner for an uncountable state space, using probability densities and transition functions. The Markov chain effectively converges to its stationary distribution after running the chain for a sufficient time and the samples drawn from the chain can be regarded as samples from the target posterior distribution [50]. Then Monte Carlo integration is used to approximate the posterior quantities of interest.

In Bayesian analysis, there are many MCMC methods available, including the Metropolis-Hastings algorithm, Gibbs sampler, reversible jump sampler and Generalized Gibbs sampler (GGS). I used GGS as the sampling technique throughout this thesis.

Metropolis-Hastings algorithm

The Metropolis algorithm is one of the simplest MCMC algorithms, named for the American physicist and computer scientist Nicholas C. Metropolis. This algorithm can be used to obtain random samples from an arbitrarily complex target distribution of any dimension, where the normalizing constant may not be known. Assume the aim is to draw samples from a distribution with probability (or probability density) $p(\theta|y)$. The basic steps of the Metropolis algorithm are as follows [52]:

1. Choose an initial value θ^0. This need not be from a high density region of the target.

2. Sample a proposal θ^* from a distribution $q(\theta^{t-1}, \theta)$ using the current θ^{t-1} . Here, $q(\theta^{t-1}, \theta)$ is the probability (or probability density) of θ given a previous value of θ^{t-1} and is known as the proposal distribution. A condition of the Metropolis algorithm is that the proposal distribution must be symmetric, that is, $q(\theta^{t-1}, \theta) = q(\theta, \theta^{t-1})$

3. Given θ^*, compute an acceptance probability (α) at the proposal θ^* and current θ^{t-1} points,

$$\alpha = \min\left\{\frac{p(\theta^*|y)}{p(\theta^{t-1}|y)}, 1\right\}. \tag{2.16}$$

4. Accept proposal θ^* as θ^t with probability α. If θ^* is not accepted, $\theta^t = \theta^{t-1}$.

5. Repeat steps 2-4 (K times).

The above steps generate a Markov chain $(\theta^0, \theta^1, \cdots, \theta^K)$ since the transition probabilities from θ^t to θ^{t+1} depend only on θ^t and not $(\theta^0, \cdots, \theta^{t-1})$. The chain moves to its stationary distribution after a burn-in period (say, m steps) and the samples $\theta^{m+1}, \cdots, \theta^K$ are taken to be samples from $p(\theta|y)$.

The Metropolis algorithm is a special case of the Metropolis-Hastings algorithm, which was proposed by Hastings (1970). The main difference is that the Metropolis-Hastings algorithm does not require a symmetric distribution (in Step 2 above). Here, an asymmetric proposal distribution, $q(\theta^*, \theta^t) \neq q(\theta^t, \theta^*)$ is used. Then the acceptance probability becomes:

$$\alpha = \min\left\{\frac{p(\theta^*|y)q(\theta^*, \theta^{t-1})}{p(\theta^{t-1}|y)q(\theta^{t-1}, \theta^*)}, 1\right\}. \tag{2.17}$$

Other steps remain the same.

Gibbs Sampling

The Gibbs sampler is a special case of the Metropolis-Hastings Algorithm, named by Geman and Geman (1984) after the American physicist Josiah W. Gibbs. The Gibbs sampler actually existed before this, and was known as Glauber dynamics in the context of statistical physics [53].

In the Gibbs sampler, the proposal distribution cycles through the posterior conditional distributions and the proposals are accepted 100% of the time. Gibbs sampling is advantageous when it is easy to sample from the conditional posterior distributions for each parameter in the model, conditional on all other parameters. Therefore, the main idea of Gibbs sampling is that it decomposes the joint posterior distribution into full conditional distributions for each parameter in the model and these conditionals are sampled sequentially and iteratively [54].

Assume a parameter vector $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_m)$ of K samples need to be obtain from the joint probability distribution $p(\theta_1, \cdots, \theta_m)$ where m is the number of parameters. Let the i-th sample be $\boldsymbol{\theta}^{(i)} = (\theta_1^{(i)}, \cdots, \theta_m^{(i)})$. The sampling steps are as follows:

1. Start with $i = 0$ and choose an arbitrary initial value of $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \cdots, \theta_m^{(0)})$.

2. For the $(i+1)$-th sample, the parameter vector will be defined as $\boldsymbol{\theta}^{(i+1)} = (\theta_1^{(i+1)}, \cdots, \theta_m^{(i+1)})$. To generate the $(i+1)$-th sample, each component parameter $\theta_j^{(i+1)}$ is sampled in turn from the distribution specified by $p(\theta_j^{(i+1)} | \theta_1^{(i+1)}, \cdots, \theta_{j-1}^{(i+1)}, \theta_{j+1}^{(i)}, \cdots, \theta_m^{(i)})$.

3. Repeat step 2 until $i = K$.

The above steps generate K samples after a burn-in period. These samples can be considered as samples from the posterior joint distribution. Monte Carlo integration can be performed using these draws to obtain quantities of interest [54].

Reversible Jump Sampler

The reversible jump sampler introduced by Green [46] allows simulation of sampling from posterior distributions on spaces of varying dimensions. This simulation is feasible even if the number of parameters in the model is unknown. Suppose that the competing fixed-dimensional models are represented by the set $\mathbb{M} = \{\mathbb{M}_1, \mathbb{M}_2, \cdots\}$. The posterior distribution under model $\mathbb{M}_k$ is $p(\theta_k | y, k) \propto p(y | \theta_k, k) p(\theta_k | k)$, where $p(y | \theta_k, k)$ is the likelihod model and $p(\theta_k | k)$ is the prior distribution of the parameters of model $\mathbb{M}_k$, respectively.

The reversible jump sampler generalizes the Metropolis-Hastings algorithm by allowing transitions between models defined by (k, θ_k) to $(k', \theta_{k'})$ with different dimensions k and k'. The resulting Markov chain includes transitions between such distinct models and may be treated as samples from the joint distribution $p(\theta_k, k)$. If the current state of the Markov chain is (k, θ_k), then the steps of the reversible jump sampler are as follows [55, 56]:

1. Propose a move to a state $(k', \theta_{k'})$ in model $\mathbb{M}_{k'}$ with proposal probability $J(k \to k')$ The model $\mathbb{M}_{k'}$ has higher dimension than the model $\mathbb{M}_k$, so that $n_{k'} > n_k$ where $n_{k'}$ and n_k are the dimension of parameter θ'_k under model $\mathbb{M}_{k'}$ and θ_k under model $\mathbb{M}_k$ respectively .

2. Generate u of length $d_{k\to k'} = n_{k'} - n_k$ with a proposal probability (or probability density) $q(u|\theta_k, k, k')$.

3. Set $(\theta_{k'}, u') = g_{k,k'}(\theta_k, u)$, where $g_{k,k'}$ is a bijection between (θ_k, u) and $(\theta_{k'}, u')$, where u and u' work in dimension-matching condition, $n_k + d_{k\to k'} = n_{k'} + d_{k'\to k}$; $d_{k'\to k}$ is the length of vector u'.

4. The acceptance probability of the new model, $(\theta_{k'}, k')$ is

$$\min\left\{1, \frac{p(y|\theta_{k'}, k')p(\theta_{k'})p(k')}{p(y|\theta_k, k)p(\theta_k)p(k)} \frac{J(k' \to k)q(u'|\theta_k, k', k)}{J(k \to k')q(u|\theta_k, k, k')} \left|\frac{\partial g_{k,k'}(\theta_k, u)}{\partial(\theta_k, u)}\right|\right\}. \tag{2.18}$$

Repeating steps 1-4 generates a sample $\theta_{kl}, l = 1, \cdots, L$.

Generalized Gibbs Sampling (GGS)

This section summarises the Generalized Gibbs Sampler (GGS) presented in [57]. GGS is a generalization of the Gibbs sampler. Moreover, it can consider as a generalisation of all the well known MCMC samplers.

The conventional Gibbs sampler is used to sample from a distribution f over a space $\mathcal{X}$ with fixed dimension d. Each iteration of the Gibbs sampler involves d coordinate updates where new values for each of the d coordinates are drawn from the one-dimensional conditional distributions of p with the other coordinates fixed. On the other-hand, the GGS can be used when points in $\mathcal{X}$ do not have fixed dimension, and may not even have a representation in terms of coordinates. The move types of GGS are analogous to the coordinate updates of the conventional Gibbs sampler.

A set $\mathcal{I}$ is defined, which will be referred to, in the following, as the index set. Let $\mathcal{U} \subset \mathcal{I} \times \mathcal{X}$, such that the projections of $\mathcal{U}$ onto $\mathcal{X}$ and $\mathcal{I}$ are surjective. For each $x \in \mathcal{X}$, let $\mathcal{Q}(x)$ be the set $\{(k,z) \in \mathcal{U} : z = x\}$. $\mathcal{Q}(x)$ is a catalogue of the types of transitions available from x. For every $x \in \mathcal{X}$, a transition matrix or density Q_x is defined on $\mathcal{Q}(x)$. The probability (or probability density) of the stationary distribution of the Markov chain induced by Q_x is denoted by q_x. Define Q the global transition matrix or density on $\mathcal{U}$:

$$Q((i,x),(j,y)) = \begin{cases} Q_x((i,x),(j,y)) & \text{for}(j,y) \in \mathcal{Q}(x) \\ 0 & \text{otherwise} \end{cases}$$

For each $(i,x) \in \mathcal{U}$, let $\mathcal{R}(i,x)$ be the set of possible transitions of type i available at x. These sets are required to form a partition of $\mathcal{U}$:

$$(j,y) \in \mathcal{R}(i,x) \Leftrightarrow (i,x) \in \mathcal{R}(j,y)$$

$$\left.\begin{array}{l} (j,y) \in \mathcal{R}(i,x) \\ (k,z) \in \mathcal{R}(j,y) \end{array}\right\} \Rightarrow (k,z) \in \mathcal{R}(i,x)$$

Here, $(i,x) \in \mathcal{R}(i,x)$. On $\mathcal{R}(i,x)$, a transition matrix or density $R(i,x)$ with respect to η_r on $r = \mathcal{R}(i,x)$ is defined as follows :

$$R_{(i,x)}((i,x),(j,y)) = \frac{f(y)q_y(j,y)}{\int_r f(z)q_z(k,z)d\eta_r(k,z)}$$

where the integral is replaced by a sum if r is countable (and similarly for subsequent integrals in this section). A global transition matrix or density R is defined on $\mathcal{U}$:

$$R((i,x),(j,y)) = \begin{cases} R_{(i,x)}((i,x),(j,y)) & \text{for}(j,y) \in \mathcal{R}(i,x) \\ 0 & \text{otherwise} \end{cases}$$

This formula can be generalized by replacing $R((i,x),(j,y))$ with $R((i,x),(j,y))S((i,x),(j,y))$ in the definition of the density R, given that $S((i,x),(j,y)) = S((j,y),(i,x))$ and adjusting S so that for all (i,x), $\int_r R((i,x),(j,y))S((i,x),(j,y))d\eta_r(j,y) = 1$. This can be useful to divide the $\mathcal{R}(i,x)$ set into two (or more) subsets. Consider a Markov chain $\{U_1, U_2, \cdots\}$ on $\mathcal{U}$ with a transition matrix or density $P = QR$.

$$P((i,x),(j,y)) = \int_{r \setminus \mathcal{Q}(x) \cap \mathcal{R}(j,y))} Q((i,x),(k,z))R((k,z),(j,y))d\eta_r(k,z)$$

Let μ be the probability (or density) $\mu(i,x) = f(x)q_x(i,x)$. Now,

$$\mu(i,x)R((i,x),(j,y)) = \begin{cases} \frac{f(x)q_x(i,x)f(y)q_y(j,y)}{\int_{r\setminus(i,x)} f(z)q_z(k,z)d\eta_r(k,z)} & \text{if}(j,y) \in \mathcal{R}(i,x) \\ 0 & \text{otherwise} \end{cases}$$

$$\mu(j,y)R((j,y),(i,x)) = \begin{cases} \frac{f(x)q_x(i,x)f(y)q_y(j,y)}{\int_{r\setminus(j,y)} f(z)q_z(k,z)d\eta_r(k,z)} & \text{if}(i,x) \in \mathcal{R}(i,x) \\ 0 & \text{otherwise} \end{cases}$$

Here $(j,y) \in \mathcal{R}(i,x), (k,z) \in \mathcal{R}(j,y) \Rightarrow (k,z) \in \mathcal{R}(i,x)$ is known, so

$$\mu(i,x)R((i,x),(j,y)) = \mu(j,y)R((j,y),(i,x))$$

(note that if $S((i,x),(j,y)) = S((j,y),(i,x))$ and

$$\begin{aligned} \int_{r\setminus\mathcal{U}} \mu(i,x)Q((i,x),(j,y))d\eta_r(i,x) &= \int_{r\setminus\mathcal{Q}(y)} f(y)q_y(i)Q((i,y),(j,y))d\eta_r(i,y) \\ &= f(y)q_y(j) = \mu(j,y) \end{aligned}$$

Consequently, μ is stationary with respect to Q and to R and with respect to P. If P is ergodic, μ is the limiting distribution of the process P. So, the steps of the generalized Gibbs sampler are as follows :

1. **Q-step** : Given $U_n = (i,x)$, generate $V \in \mathcal{Q}(x)$ by drawing from the distribution with probability or density $Q((i,x),.)$.

2. **R-step** : Given $V = (j,y)$, generate $W \in \mathcal{R}(j,y)$ by drawing from the distribution with probability or density $R((j,y),.)$.

3. Let $U_{n+1} = W$.

This algorithm creates a Markov chain, so that the limited distribution is $\mu(i,x)$ provided that the density P is ergodic.

2.3.2 Sampling steps of Bayesian change-point detection model

To estimate values of the parameters k, s, π and α of the Bayesian change-point detection model discussed in Section 2.2.3, a sample from the posterior distribution in Equation 2.10 is drawn using the Generalized Gibbs Sampler (GGS). The GGS sampler cycles through a sequence of steps by updating parts of the current element of a Markov chain and holding

other parts constant, in a manner resembling the conventional Gibbs sampler. Unlike the conventional Gibbs sampler, the GGS can sample from spaces in which the dimension varies from point to point. The GGS algorithm was applied to the change-point model in Section 2.2.3, using the move-types defined below.

Move Types

- (I, i) : decide whether to insert a new change-point in segment i, and at what position.
- (D, i) : decide whether to delete change-point i (if it is not a fixed change-point).
- (S, i) : slide change-point i to a new position between $s_i - 1$ and $s_i + 1$ (if it is not a fixed change-point).
- (π_{t_1}, π_{t_2}) : simultaneously update π_{t_1} and π_{t_2} for $(t_1, t_2) \in \{0, \cdots, \vartheta - 1\}^2$.
- (t_1, t_2) : simultaneously update $\pi_{t_1}, \pi_{t_2}, \alpha^{(t_1)}$ and $\alpha^{(t_2)}$ for $(t_1, t_2) \in \{0, \cdots, \vartheta - 1\}^2$.
- π_t: update π_t, scaling all other π values by a constant factor.
- σ^t: update $\sqrt{\frac{1}{z^t+1}}$ while holding μ^t fixed.
- μ^t: update μ^t while holding σ^t fixed.

There are: $k + 1$ I-moves; k D-moves; k S-moves; 3ϑ moves for each group to update π_t, σ^t, μ^t; $\vartheta(\vartheta - 1)/2$ moves to update (π_{t_1}, π_{t_2}) and $\vartheta(\vartheta - 1)/2$ moves to update (t_1, t_2). Then the total number of moves for a sequence with k change-points is:

$$N(k) = 3k + 1 + \vartheta(\vartheta - 1) + 3\vartheta.$$

where ϑ is the number of groups. There is a possibility no change is made in each of the first 3 moves, in which case the current segmentation is repeated. The sampler cycles through the available moves in a systematic manner, illustrated in Figure 2.3.

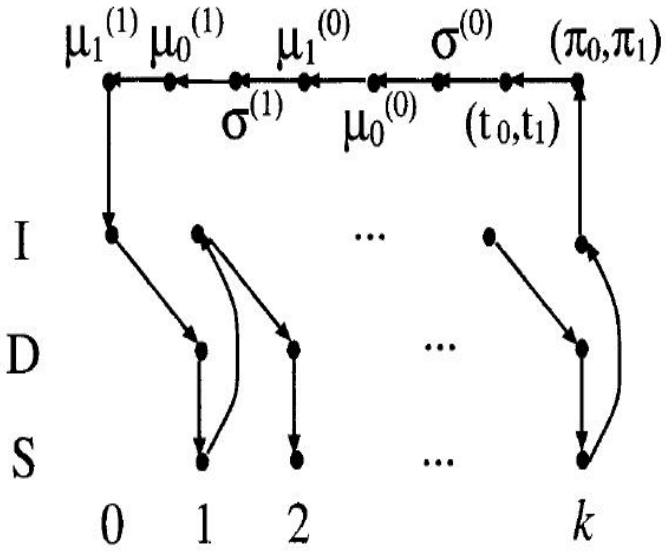


Figure 2.3: The order in which the updates are carried out.

Insertion Step:$(I; i)$

For each segment $i = 0, \cdots, k$, determine the conditional posterior distribution over the set consisting of the current segmentation and the segmentation obtained by inserting a new change-point between s_i and s_{i+1}, while holding π, α and the positions of the other change-points constant. The distribution of the current segmentation is proportional to (using Equation 2.10):

$$\Gamma(T-k)\Gamma(k-f+1)f(m_i, n_i|\pi, \alpha).$$

where m_i and n_i are respectively the numbers of zeros and ones in segment i. The conditional posterior probability of a segmentation with a new change-point at position z is proportional to (using Equation 2.10):

$$\Gamma(T-k-1)\Gamma(k-f+2)f(m'_i, n'_i|\pi, \alpha)f(m'_{i+1}, n'_{i+1}|\pi, \alpha).$$

where m'_i and n'_i are respectively the numbers of zeros and ones between s_i and $z-1$ and m'_{i+1} and n'_{i+1} are respectively the numbers of zeros and ones between z and $s_{i+1}-1$. Select the current segmentation (that is, the current segmentation is repeated with no change) with probability proportional to:

$$w_- = \frac{\Gamma(T-k)\Gamma(k-f+1)f(m_i,n_i|\pi,\alpha)}{N(k)}. \tag{2.19}$$

Select a new segmentation with a new change-point at z for each $z \in \{s_i+1,\cdots,s_{i+1}-1\}$ with probability proportional to:

$$w_z = \frac{\Gamma(T-k-1)\Gamma(k-f+2)f(m'_i,n'_i|\pi,\alpha)f(m'_{i+1},n'_{i+1}|\pi,\alpha)}{N(k+1)}. \tag{2.20}$$

Improving the efficiency of insertion step $(I;i)$

The last stage of the sampling procedure given in the previous section can be improved with the help of the following method. According to the generalized Gibbs sampler discussed in Section 2.3.1:

$$\mathcal{R}((I,i),(k,s_0)) = ((I,i),(k,s_0)) \bigcup_{s_z\in\{s_i+1,\cdots,s_{i+1}-1\}} ((D,i+1),(k+1,s_z)).$$

The probability of inserting a change-point at position x is

$$R_{((I,i),(k,s_0))}(((I,i),(k,s_0)),((D,i+1),(k+1,s_x))) = \frac{w_x}{w_- + \sum_z w_z}.$$

In fact, the generalized form of the algorithm is used with the matrix S (see Section 2.3.1)

$$S((I,i)(k,x),(D,i+1)(k+1,y)) = [w_- + \sum_z w_z]\min\left[\frac{1}{w_-},\frac{1}{\sum_z w_z}\right].$$

and choose $S(((I,i),(k,x)),((I,i),(k,x)))$ so that the sum over all the possibilities is one. The fact that S is symmetrical, so $S(((D,i+1),(k+1,y)),((I,i),(k,x))) = S(((I,i),(k,x)),((D,i+1),(k+1,y)))$ will be seen in the deletion move. This matrix S is used to separate the set $\mathcal{R}((I,i)(k,x))$ into two subsets: $\{((I,i)(k,x))\}$ on the one hand and on the other hand, $\bigcup_{s_z\in\{s_i+1,\cdots,s_{i+1}-1\}}((D,i+1),(k+1,s_z))$. Hence, the probability of inserting a new change-point at some point x, which is the probability of a transition to any element other than the current one, is given by:

$$\frac{w_x}{w_- + \sum_z w_z}[w_- + \sum_z w_z]\min\left[\frac{1}{w_-},\frac{1}{\sum_z w_z}\right] = w_x\min\left[\frac{1}{w_-},\frac{1}{\sum_z w_z}\right].$$

Then the probability of making an insertion can be replaced by probability:

$$\min \left[1, \frac{\sum_{z=s_i+1}^{s_{i+1}-1} w_z}{w_-} \right].$$

If the decision is made to insert a new change-point, its position $z \in \{s_i+1, \cdots, s_{i+1}-1\}$ is selected with probability proportional to w_z. This modification enhances the probability of accepting an insertion, and thus improves the efficiency of the algorithm. If a change-point is inserted, the move-type is updated to $(D, i+1)$, otherwise it remains (I, i). In either case, the move-type is then updated as in Figure 2.3.

Deletion Step: $(D; i)$

For each non-fixed change-point $i = 1, \cdots, k$, determine the conditional posterior distribution over the set consisting of every segmentation with a change-point at z between $s_{i-1}+1$ and $s_{i+1}-1$ and the segmentation obtained by deleting a non-fixed change-point s_i, while holding π, α and the positions of the other change-points constant.

The conditional posterior distribution of a segmentation obtained by deleting the change-point s_i is proportional to (using Equation 2.10):

$$\Gamma(T-k+1)\Gamma(k-f)f(m_i, n_i|\pi, \alpha).$$

where m_i and n_i are respectively the numbers of zeros and ones between s_{i-1} and $s_{i+1}-1$. The conditional posterior distribution of a segmentation obtained by sliding s_i to a (possibly) new change-point at z between $s_{i-1}+1$ and $s_{i+1}-1$ is proportional to (using Equation 2.10):

$$\Gamma(T-k)\Gamma(k-f+1)f(m_i', n_i'|\pi, \alpha)f(m_{i+1}', n_{i+1}'|\pi, \alpha).$$

where m_i' and n_i' are respectively the numbers of zeros and ones between s_{i-1} and $z-1$ and m_{i+1}' and n_{i+1}' are respectively the numbers of zeros and ones between z and $s_{i+1}-1$.

A straight forward GGS update would be to delete the change-point with probability proportional to:

$$w_- = \frac{\Gamma(T-k+1)\Gamma(k-f)f(m_i, n_i|\pi, \alpha)}{N(k-1)} \tag{2.21}$$

or slide the change-point to position z with probability proportional to:

$$w_z = \frac{\Gamma(T-k)\Gamma(k-f+1)f(m'_i, n'_i|\pi, \alpha)f(m'_{i+1}, n'_{i+1}|\pi, \alpha)}{N(k)}. \tag{2.22}$$

Improving the efficiency of deletion step $(D; i)$

As with the insertion step, the probability of deleting change-point s_i can be modified to:

$$\min \left[1, \frac{w_-}{\sum_{z=s_{i-1}+1}^{s_{i+1}-1} w_z} \right].$$

If the decision is made not to delete the change-point, its position remains unchanged. If a change-point is inserted, the move-type is updated to $(I, i-1)$, otherwise, it remains (D, i). For a fixed change-point, the (D, i) move is replaced by the trivial move of repeating the same segmentation. In either case, the move-type is then updated as in Figure 2.3.

Slide Step: $(S; i)$

For each non-fixed change-point $i = 1, \cdots, k$, determine the conditional posterior distribution over the set of segmentations obtained by sliding s_i to a (possibly) new change-point between s_{i-1} and s_{i+1}, while holding $\pi; \alpha$ and the positions of the other change-points constant.

The conditional posterior probability of a segmentation obtained by sliding change-point i to z is proportional to (using Equation 2.10):

$$\Gamma(T-k)\Gamma(k-f+1)f(m'_i, n'_i|\pi, \alpha)f(m'_{i+1}, n+1'_i|\pi, \alpha).$$

where m'_i and n'_i are respectively the numbers of zeros and ones in the segment with endpoints s_{i-1} and $z-1$ and m'_{i+1} and n'_{i+1} are respectively the numbers of zeros and ones in the segment with endpoints z and $s_{i+1}-1$. The conditional posterior probability of the current segmentation is proportional to (using Equation 2.10):

$$\Gamma(T-k)\Gamma(k-f+1)f(m_i,n_i|\pi,\alpha)f(m_{i+1},n_{i+1}|\pi,\alpha)$$

where m_i and n_i are respectively the numbers of zeros and ones between s_{i-1} and s_i-1 and m_{i+1} and n_{i+1} are respectively the numbers of zeros and ones between s_i and $s_{i+1}-1$.

A straight forward GGS update would be to re-select the current segmentation with probability proportional to:

$$w_- = \frac{\Gamma(T-k)\Gamma(k-f+1)f(m_i,n_i|\pi,\alpha)f(m_{i+1},n_{i+1}|\pi,\alpha)}{N(k)} \tag{2.23}$$

or slide the change-point to position z with probability proportional to:

$$w_z = \frac{\Gamma(T-k)\Gamma(k-f+1)f(m'_i,n'_i|\pi,\alpha)f(m'_{i+1},n+1'_i|\pi,\alpha)}{N(k)}. \tag{2.24}$$

Improving the efficiency of sliding step $(S;i)$

Here, the matrix S described in GGS (Section 2.3.1) is not used. So, the probability of sliding the change-point s_i to x is

$$\left[\frac{w_x}{w_- + \sum_{z=s_{i-1}+1}^{s_{i+1}-1} w_z}\right].$$

For a fixed change-point, the current segmentation is repeated. The move-type is then updated as in Figure 2.3.

Steps π, α, μ

Updates of π_t, σ^t and μ^t include sampling the conditional posterior distributions over several one-dimensional subspaces of the target space, holding k and s fixed. These updates are conventional Gibbs updates. The conditional distributions are straight forward to derive, but attention must be given to multiply by the appropriate Jacobian when a change of variables is involved. Further details of these updates may be found in [49].

Monte Carlo Integration

For each character position in the binary sequence, the posterior probability that position is within a given group is calculated by Monte Carlo integration. Further details may be found in [49].

2.3.3 Sampling from a posterior distribution in time series segmentation

In Bayesian change-point analysis, a variety of sampling techniques have been used to simulate samples from a posterior distribution. Carlin *et al.* [20] faced difficulties in computing marginal posterior distributions for the location of the change-point and of the model parameters in their hierarchical Bayesian change-point model. They suggested a straightforward Markov chain Monte Carlo sampling method, which employs the Gibbs sampler to obtain marginal posterior distributions. The Gibbs sampler was described in Section 2.3.1. The Gibbs sampler, in general, deals with missing data in a straightforward manner. Missing data values can be treated as additional model parameters whose conditional distributions can be sampled.

Green [46] used the reversible jump sampler to estimate the number of change-points for a non-homogenous Poisson process. In Green's approach, segmentations of a time series with different numbers of change-points corresponded to subspaces with differing dimensionality. At each step of Green's algorithm, a choice is made whether to: (1) delete a change-point, thus joining two consecutive segments (death move); (2) add a new change-point, thus dividing a segment into two (birth move); or (3) alter the location of a change-point. A move is accepted at each step with a probability similar to that used by the Metropolis-Hastings algorithm. One practical issue for this methodology is to identify

moves which are likely to be useful. Moreover, in many cases the reversible jump algorithm converges slowly.

Chib [45] also used the reversible jump MCMC for time series segmentation. He generalized reversible jump MCMC to analyse a change-point model in terms of latent discrete state variables. A latent variable is associated with each position in the time series, and is used to indicate which of several alternative discrete-time, discrete-state Markov processes describes the dependence between adjacent time-points.

Billio *et al.* [58] introduced a Markov-Chain Monte Carlo (MCMC) algorithm to sample the regime-indicator variables using a proposal density which is an accurate approximation of the target density. The approximation error in the proposal density was corrected by applying the Metropolis Hastings (MH) algorithm. When the entire series of the regime indicator variable was drawn from the proposal density, the approximation error was corrected for by globally accepting or rejecting the newly drawn regime indicator variables with a carefully defined probability. The authors used a Gibbs sampler for drawing the regime-indicator variables. In this sampler, a single indicator variable is drawn one at a time, conditional on the remaining regime-indicator variables.

Lavielle and Lebarbier [59] proposed a Bayesian methodology in order to investigate a change-point process instead of estimating the sequence of change-point instants $(\tau_k, k \geqslant 0)$. They defined the change-point process as follows:

$$r_t = \begin{cases} 1 & \text{if there exists } k \text{ such that } t = \tau_k, \\ 0 & \text{otherwise.} \end{cases}$$

They assumed $y = (y_t, t \geqslant 1)$ is a real process such that, for any $t \geqslant 1, y_t = s(t) + \epsilon_t$. Here, ϵ_t is a sequence of random variables with mean 0 and s is a piecewise constant function between two successive change-points instants (τ_k). A sequence m_k also exists such that, for any $k \geqslant 1$, $s(t) = m_k$ for all $\tau_{k-1} + 1 \leqslant t \leqslant \tau_k$ with the condition $\tau_0 = 0$. The unknown function s was recovered by estimating the sequences r_t and m_k. To obtain a good estimate of the configuration of change-points, they investigated the joint posterior distribution of the mean sequence $m = m_k$ and the change-point process $r = r_t$ instead of the posterior distribution of τ. Here, the dimension of the series r_t is fixed. This article used Metropolis-Hasting to sample a series of zeros and ones, of fixed length $n - 1$, where the length of the observed signal is n.

To estimate the distribution of the mean of the newly defined series r_t, the authors

used a hybrid MCMC algorithm, combining elements of the Metropolis-Hasting algorithm with elements of the Gibbs sampler. This hybrid algorithm produced ergodic Markov chains $(s^{(i)}, r^{(i)})$ which converged to the joint posterior distribution $p(s, r|y; \theta)$, where θ is the set of hyper-parameters of the model. As $p(s, r|y; \theta)$ cannot be described completely, they considered estimating the posterior expectation of the mean $(\mathbb{E}(s|y; \theta))$. However, the posterior expectation of the mean was not suitable for finding a good configuration of change-points as it produced a smooth version of the signal instead of a step function.

They considered another approach which conditionally provided a good estimation of the configuration of change-points. In this approach, they used the MAP estimate of r and the fact that the conditional distribution $p(s|r, y; \theta)$ is a Gaussian distribution with known parameters. They also used a modification of an MCMC algorithm to estimate the hyper-parameters of the model (θ) instead of setting the hyper-parameters to a particular value. To update the set of hyperparameters at each iteration of MCMC, they used a stochastic approximation to expectation maximization (SAEM) introduced by Delyon *et al.* [60]. This resulted in an iterative algorithm which needs an initial configuration of change-points $r^{(0)}$ and an initial guess $\theta^{(0)}$. The algorithm is composed of two steps at each iteration i: a simulation step and an estimation step. In the simulation step, a new configuration $r^{(i)}$ is generated with M iterations of the MCMC algorithm, using the current values of the hyper-parameters $\theta^{(i-1)}$ and the current configuration $r^{(i-1)}$. Then $\theta^{(i)}$ is updated at the estimation step using the new configuration $r^{(i)}$ and a stochastic approximation. After that, $\theta^{(i)}$ is computed by maximizing the complete likelihood $f(r, y; \theta)$.

One important advantage of this method is the capability to automatically execute different tasks (estimate the posterior distribution of the unknown sequence of change-points instants, estimate the hyper-parameters of the model and compute the change-points sequences of highest probabilities). This method is much faster than the reversible jump algorithm. Another advantage was that the hyper-parameters of the model were estimated, rather than arbitrarily selected. This method can be extended to detect changes in the spectrum of a signal.

2.4 Segmenting parallel sequences

Many methods have been discovered and introduced in the literature in the last and recent decades for the segmentation of a single sequence. However, numerous applications require

the segmentation of multiple sequences simultaneously, in order to better understand a complex underlying phenomenon. Here I review some methods from the literature that have been used to segment parallel sequences.

Chamroukhi *et al.* [61] developed methods for human activity recognition using joint segmentations of multidimensional time series of acceleration data. This acceleration data was measured in a three-dimensional space using body-worn accelerometers. To perform automatic temporal segmentation, they developed a statistical latent process model by assuming the observed acceleration series was determined by a series of hidden (unobserved) activities. The method depends on a multiple regression model, including a hidden discrete logistic process. The logistic process determines switching from one activity to another over time.

They studied the model in an unsupervised context by maximizing the observed-data log-likelihood through a dedicated expectation-maximization (EM) algorithm. The performance of the proposed model of joint segmentation was compared to alternative approaches, including well-known supervised static classifiers and the standard hidden Markov model (HMM). Two significant advantages of this method are that it directly utilizes the raw acceleration data and works in an unsupervised way. This approach used regression models with polynomial bases to segment human activity data and the method can be extended by using non-linear models for improving the representation of each activity signal. To describe any kind of complex activities, this method can also be extended by integrating it into a Bayesian non-parametric model.

Another joint segmentation of correlated time-series data was considered by Collilieux *et al.* [62]. They segmented spatio-temporal data and dealt with multiple series, considering the dependency between them to avoid false change-point detection. They assumed each series to be affected by changes at series-specific breakpoints and also assumed the observed sequences at each location were correlated. They proposed a model for correlated Gaussian series based on a segmentation model combined with a factor model. To remove the between series dependence, they used a variant of the EM algorithm which combines EM and dynamic programming (DP). The DP algorithm is applicable only when the log-likelihood is additive with respect to the segments [63]. However, the log-likelihood is not additive if dependency exists, so they transformed the data to remove the between-series dependency and then applied the DP algorithm to the transformed data. The authors also proposed a heuristic model selection procedure combining two BIC criteria: the classical

BIC to determine the number of factors and a modified BIC criterion to determine the number of segments in the context of segmentation.

The advantage of this segmentation approach is that it considers a wide range of possible dependency structures between series, rather than assuming the simplest form of correlation between all series. The inferred number of segments obtained using this method is close to the true number and the change-points are accurately positioned if the noise variance is small, regardless of the strength of dependency between the series. However, when the noise variance is large, the change-point detection becomes less accurate. In that case the method tends to underestimate the number of segments, although it is still able to accurately infer change-point locations.

Cleynen and Robin [64] were interested in the segmentation of independent series in order to compare the locations of change-points between several series that have been observed independently. In this work, independent series are the series that have been observed during the same period of time, but not necessarily simultaneously. For example, in a study of finding changes in the stream flow of a river over a year, independent series can arise when the purpose is to examine whether those changes occur at the same time every year. Independent series can also arise for data sets with a non-temporal one-dimensional structure. For instance, the authors considered genomic locations instead of time in their examples. Two exact approaches in a Bayesian framework were developed by the authors to compare the positions of change-points. The first approach was used to compare the location of change-points between two series and the second compared more than two series. In their first approach, they determined the posterior credibility interval of the shifts in location of change-points in two independent series of the same length. In the second approach, they estimated the posterior probability for a given change-point to have the same location in all series. The computations in both approaches were performed in an exact manner in quadratic time. Both approaches used a Bayesian segmentation model introduced by Rigaill *et al.* [65] with conjugate priors to make exact inference on the change-point model. They used a Gaussian heteroscedastic distribution, a Gaussian homoscedastic distribution with known variance, a Poisson, and a negative binomial homoscedastic with known dispersion as alternative data models in their Bayesian segmentation model.

Harlé *et al.* [2] proposed a Bayesian model for joint segmentation in multivariate time series data. They presented a vector $J \in \{0,1\}^N$ to model the presence or absence of change-points at different time points where j_i is an entry defined as an indicator variable

such as

$$j_i = \begin{cases} 1 & \text{if } x_i \text{ is a change-point,} \\ 0 & \text{otherwise,} \end{cases}$$

for all $1 \leq i \leq n$ and $j_1 = j_N = 1$. Their model combines a robust non-parametric statistical test acting on individual time segments and a Bayesian framework. It makes weak assumptions regarding the locations of change-points across multivariate time series and regarding the distributions of the signals. The authors derived a composite marginal likelihood in order to approximate the full likelihood and created an inference function using the p-values of certain statistical tests in that composite marginal likelihood. These p-values were considered as a random variable p_i which were calculated from the data. And the inference function is as follows:

$$L_*(X|J) = \prod_{i=2}^{N-1} f(p_i|J). \tag{2.25}$$

$L_*(X|J)$ is not a proper likelihood and is called a composite marginal likelihood as it was composed of a product of the marginal likelihood of the $(p_i)_{1\leq i\leq N}$ using their univariate distributions. Then they used the non-parametric Wilcoxon rank sum test in order to be free from any Gaussian assumption and to be robust to outliers. They choose a prior distribution for the change-point indicator vector and derived the corresponding posterior distribution. They identified the Maximum A Posteriori (MAP) parameter values using a Gibbs sampler strategy. An advantage of this method is that it provides information about the underlying dependency structure between time series without making any assumption on the number of change-points. A limitation of this method is the relatively long computation time.

Zhang and David [66] also considered the problem of detecting simultaneous change-points in multiple sequences to identify DNA copy number variants, which are gains and losses of segments in chromosomes in multiple samples. They looked at the problem of detecting local signals that occur at the same location in multiple one-dimensional noisy sequences, giving special attention to relatively weak signals that occur in only a fraction of the sequences. They proposed simple scan and segmentation algorithms which depend on summing chi-square statistics across samples. The resulting statistic is equivalent to the generalized likelihood ratio for models in which the errors in each sample are independent. In these algorithms, they investigated the particular problem of detecting a shared abrupt

jump in mean, assuming the noise within each profile to be independent and identically distributed Gaussian variables. They used this mean shift model to detect DNA copy number variants. They also used a multisample segmentation algorithm to analyse a cohort of tumor samples that holds complex nested and overlapping copy number aberrations. Their algorithm provided a sparse and intuitive summary across these tumor samples. The segmentation algorithm is slower than the multi-sample scan, because each time a change-point is found, the entire interval must be re-scanned in the next step of the recursion.

A joint segmentation algorithm to address the problem of segmenting correlated signals recorded from several sensors was proposed by Dobigeon *et al.* [67]. This algorithm was based on a hierarchical Bayesian model and was developed for piecewise constant autoregressive (AR) processes where the orders of these processes are fixed on each segment. They studied J sensors delivering J signals, each with sample size n. They denoted individual signals as $\mathbf{x}_j = (x_{j,1}, \cdots, x_{j,n})$ and modeled each of the J signals as a piecewise constant AR process as follows:

$$x_{j,t} = \sum_{i=1}^{a} \psi_{j,k,i} x_{j,t-i} + \epsilon_{j,t}, \quad j = 1, 2, \cdots, J \text{ and } k = 1, 2, \cdots, K_j \tag{2.26}$$

where, $x_{j,t}$ is the sample of signal j at time t, k is the segment index, K_j is the number of segments, $\boldsymbol{\psi}_{j,k} = (\psi_{j,k,1}, \cdots, \psi_{j,k,a})^T$ is the set of AR parameters, a is the order of the AR model and $\boldsymbol{\epsilon}_j = (\epsilon_{j,1}, \cdots, \epsilon_{j,n})$ are i.i.d. zero mean Gaussian noise samples. The segment k in the signal j has boundaries $[i_{j,k-1} + 1, i_{j,k}]$ where $i_{j,k}$ is the time index immediately after a change-point is found, with condition $i_{j,0} = 0$ and $i_{j,K_j} = n$. In matrix form, the piecewise constant AR process becomes:

$$\mathbf{x}^T_{j,i_{j,k-1}+1:i_{j,k}} = \mathbf{X}_{j,k}\boldsymbol{\psi}_{j,k} + \boldsymbol{\epsilon}^T_{j,i_{j,k-1}+1:i_{j,k}}, \tag{2.27}$$

where, $\mathbf{X}_{j,k}$ denotes a matrix of J observed signals with size $(i_{j,k} - i_{j,k-1}) \times a$ and

$$\mathbf{X}_{j,k} = \begin{bmatrix} x_{j,i_{j,k-1}} & x_{j,i_{j,k-1}-1} & \cdots & x_{j,i_{j,k-1}-a+1} \\ x_{j,i_{j,k-1}+1} & x_{j,i_{j,k-1}} & \cdots & x_{j,i_{j,k-1}-a+2} \\ \vdots & \vdots & \vdots & \vdots \\ x_{j,i_{j,k}-1} & x_{j,i_{j,k}-2} & \cdots & x_{j,i_{j,k}-a} \end{bmatrix}$$

The authors suggested a Bayesian approach for estimating the change-point locations $i_{j,k}$ from the J observed time series $\mathbf{x}_j$. They chose a suitable prior to take account of

the correlations between change-point locations of the observed signals, and sampled the resulting posterior distribution using a Gibbs sampling strategy. They also described an extension of this model, assuming unknown model order of the piecewise constant AR processes. The assumptions this model makes regarding the observed signals are weak to handle a large class of real signals (for example, seismic or biomedical signals).

Dobigeon *et al.* [68] also considered the problem of identifying and characterizing structure in two or more related astronomical time series. This article proposed a Bayesian time-series segmentation algorithm, permitting the joint segmentation of multiple signals coming from different sensors. This model is hierarchical and involves a piecewise constant Poisson rate model with prior distributions for the unknown parameters of change-point locations and Poisson parameters. They used a Gibbs sampling strategy for joint estimation of the unknown parameters and hyperparameters. One significant feature of this method is that its treatment of possible relationships between observed times series can be used to investigate interrelationships between time series arising in a wide range of applications, especially in astronomical data. Another significant feature is that information regarding uncertainties in the parameter estimates emerges from the sampling strategy. However, this approach can be very expensive in terms of computational time.

Some other approaches to segment parallel sequence include [69–74].

Autocorrelation may occur in time series data and some dependency may also exist between multiple time series. It is important to consider these dependencies to avoid false change-point detection. Consequently, it is necessary to consider time series models that can take into account these dependencies. In all aforementioned studies, the methods used for segmenting parallel time series data did not allow for possible correlations between noise terms at consecutive time points. This thesis aims to fill this gap by proposing an autoregressive moving average (ARMA) model in each segment. AR models have been used in the context of segmenting multiple sequences: Dobigeon *et al.* [67] used piecewise constant segmented autoregressive (AR) models for joint segmentation of multiple time series. But ARMA models can consider the dependency between the residual terms of the model by including a moving average component. This provides an advantage of ARMA models over AR models, that may be important in some applications. The segmentation model used in this thesis also provides flexibility to sample from varying dimensional spaces in which the number of change-points is unknown, using a highly efficient Generalized Gibbs Sampler. These reasons make a segmented ARMA model in a Bayesian setting more flexible

than segmented AR models for the segmentation of multiple sequences in parallel. A better way to handle the dependencies between multiple time-series is to segment a multivariate ARMA model. We intend to explore such approaches in future work.

2.5 Models and methods used in this thesis

Here I present some models and methods used in upcoming chapters.

2.5.1 ARMA model

One difference between the Bayesian change-point model presented in Chapter 3 and the model discussed in Section 2.2.3 is that the latter model is designed for binary sequences whereas the segmented ARMA model presented in Chapter 3 is designed for time series data. These distinct data types necessitate different models within each segment. In the model described in Section 2.2.3, the binary sequence within each segment is generated by independent Bernoulli trials at each position in that segment whereas the Bayesian change-point segmented ARMA model segments the sequence by assuming an ARMA process within each segment. A concise description of ARMA models follows.

An Autoregressive–moving-average (ARMA) model gives a parsimonious representation of a (weakly) stationary stochastic time series process. This model provides a general framework for studying stationary processes in terms of two polynomials, one for the autoregression (AR) component and the second for the moving average (MA) component [75]. In 1951, the general ARMA model was illustrated by Peter Whittle in his thesis, "Hypothesis testing in time series analysis", and was popularised in the 1970 book by George E. P. Box and Gwilym Jenkins [76]. The ARMA model is used to model and forecast the future values in a time series. The AR component involves regression of the variable on its past values and the MA component involves modeling of the error term, which depends linearly on its previous values. A real-valued stochastic process $\{X_t\}$ is an ARMA(a,m) process if

1. $\{X_t\}$ is stationary or weakly stationary.

2. The process or the deviation of the mean from the process itself $(X_t - E(X_t))$ fulfills the linear difference equation written in "regression form" as

$$X_t - \psi_1 X_{t-1} - \cdots - \psi_a X_{t-a} = \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_m \epsilon_{t-m} \tag{2.28}$$

where, $\psi_1, \cdots, \psi_a$ are the AR parameters, $\theta_1, \cdots, \theta_m$ are the MA parameters and $\epsilon_1, \epsilon_{t-1}, \cdots$ are the white noise error terms with $\epsilon_t \sim N(0, \sigma^2)$.

The polynomials

$$\psi(z) = 1 - \psi_1 z - \cdots - \psi_a z^a \tag{2.29}$$

and

$$\theta(z) = 1 + \theta_1 z + \cdots + \theta_m z^m \tag{2.30}$$

have no common factors.

The ARMA model described in Equation 2.28 can not be simplified if the polynomials of Equation 2.29 and Equation 2.30 have no common factors. However, there are excess parameters if the polynomials do have common factors and this unnecessarily complicates analysis of the model. We can write Equation 2.28 concisely using the backshift operator B as

$$\psi(B)X_t = \theta(B)\epsilon_t. \tag{2.31}$$

where $\psi(B)$ and $\theta(B)$ are known as the regressive operator (polynomial in B) and the moving average operator (polynomial in B), respectively. When $\theta(B) = 1$, an ARMA(a,m) model reduces to an AR(a) and for $\psi(B) = 1$, ARMA(a,m) reduces to MA(m). These processes are denoted AR(a) and MA(m), indicating that the autoregressive model and the moving average model are members of the ARMA models family [77].

Properties of ARMA Process

Stationarity

A stochastic process is said to be strictly stationary if the joint distribution of that process does not change over time. That is, the joint distribution of $X_{t_1}, X_{t_2}, \cdots, X_{t_n}$ is the same as the joint distribution of $X_{t_1+h}, X_{t_2+h}, \cdots, X_{t_n+h}$ for any integer $t_1, t_2, \cdots, t_n$ and h [78, 79]. The process is weakly stationary if the mean is a fixed constant for all time points and the autocovariances of the process depend only on the time difference [78].

An ARMA(p, q) process given by Equation 2.28 is stationary if the characteristic equation $\psi(B) = 0$ has all its roots outside the unit circle [78]. We can write from Equation 2.30

$$X_t = \phi(B)\epsilon_t. \tag{2.32}$$

where, $\phi(B) = \frac{\theta(B)}{\psi(B)} = \frac{1+\theta_1 B+\theta_2 B^2+\cdots+\theta_m B^m}{1-\psi_1 B-\psi_2 B^2-\cdots-\psi_a B^a}$. The process X_t is stationary if

$$\sum_{j=1}^{\infty} |\phi_j| < \infty. \tag{2.33}$$

This occurs if the series $\phi(Z)$ converges for every Z with the condition $|Z| \leq 1$ and the series converges if the complex zeros of $\phi(Z)$ lie outside the unit circle [80].

Stationarity is an important property of time series. For stationary series, a finite realisation of the process can be taken as representative of the process, thus permitting one to make statistical conclusions about the whole process [81].

Causality

Causality of a stationary time series indicates that the time series is dependent only on past and present noise, not on future values of the noise. This property is important for a process as only the past or present shocks can influence the current state [81].

An ARMA(a, m) process X_t is causal if there is a $\phi(B) = \phi_0 + \phi_1 B + \phi_2 B^2 + \cdots$ with $\sum_{j=0}^{\infty} |\phi_j| < \infty$ [75, 80] and

$$X_t = \sum_{j=0}^{\infty} \phi_j \epsilon_{t-j} \text{ for all } t. \tag{2.34}$$

The weights ϕ_j can be estimated from the relation $\phi(B)\psi(B) = \theta(B)$ to satisfy

$$\phi_j = \psi_1\phi_{j-1} + \psi_2\phi_{j-2} + \cdots + \psi_a\phi_{j-a} - \theta_j, \quad j > 0. \tag{2.35}$$

with the condition $\phi_0 = 1$, $\phi_j = 0$ for $j < 0$ and $\theta_j = 0$ for $j > m$.

Invertibility

The process X_t is said to be *invertible* if the characteristic equation $\theta(B) = 0$ has all its roots outside the unit circle [78]. An ARMA (a, m) process is invertible if there exists constants $|\pi_j|$ and if there is a $\pi(B) = \pi_0 + \pi_1 B + \pi_2 B^2 + \cdots$ with $\sum_{j=0}^{\infty} |\pi_j| < \infty$ [75, 80] and

$$\epsilon_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}. \tag{2.36}$$

The above equation implies that the errors can also be expressed as weighted sums of present and previous observations. The weights π_j can be estimated from the relation $\theta(B)\pi(B) = \phi(B)$ to satisfy [78]

$$\pi_j = \theta_1 \pi_{j-1} + \theta_2 \pi_{j-2} + \cdots + \theta_m \pi_{j-m} + \psi_j, \quad j > 0. \tag{2.37}$$

with the condition $\pi_0 = -1$, $\pi_j = 0$ for $j < 0$ and $\psi_j = 0$ for $j > a$.

2.5.2 Model Comparison

Akaike information criterion (AIC)

Statistician Hirotugu Akaike introduced the Akaike information criterion (AIC) in the early 1970s to compare the quality of different statistical models[82]. The AIC is used in Chapter 5. It is defined as:

$$AIC = 2d - 2\ln f(x|\hat{\theta}). \tag{2.38}$$

where d is the number of estimated parameters in the model, x is the observed data, $\hat{\theta}$ is the maximum likelihood estimate (MLE) for the model parameters θ and $f(x|\hat{\theta})$ is the likelihood of the parameter θ. The first term in AIC provides a penalty function that increases with the number of estimated parameters in the model and the second term provides a measure of goodness of fit of the model that is assessed by the likelihood function [83].

Keith *et al.* (2008) suggested an approximation of AIC for the Bayesian segmentation and classification model given as [49]:

$$\widehat{AIC} = 2\overline{K} - 2\overline{\ln f(x|\hat{\theta})}. \tag{2.39}$$

where, $\overline{K}$ is the average number of segments over the set of segmentations sampled by

MCMC and $\overline{\ln f(x|\hat{\theta})}$ is the average of log-likelihood over the set of MCMC samples. The difference between the approximated AIC and the regular AIC is that the approximated AIC uses averages for model fit and model complexity whereas the regular AIC uses a fixed number of parameters d and an optimized log-likelihood $\ln f(x|\hat{\theta})$. The advantage of this approximated criterion is that it can be estimated using the already sampled segmentations [49].

Bayesian information criterion (BIC)

Bayesian information criterion (BIC) (also used in Chapter 5) was first developed by Gideon E. Schwarz (1978) in the paradigm of the maximum likelihood methodology and originated from a Bayesian approach to model selection [84]. The BIC is defined as:

$$BIC = d\ln n - 2\ln f(x|\hat{\theta}). \tag{2.40}$$

where, n is the sample size. To solve the overfitting problem in models, BIC introduces a penalty term ($d\ln n$). This is similar to the penalty term used in AIC but it penalizes the number of parameters more heavily. Hence, BIC favours models with fewer parameters [85].

This thesis uses an approximation to BIC proposed for the Bayesian change-point model in Oldmeadow *et al.* (2010) given as [86]:

$$\widehat{BIC} = -2\overline{\ln f(x|\hat{\theta})} + \overline{K}\ln T. \tag{2.41}$$

where, T is the total length of the signal.

Deviance information criterion (DICV)

A model selection tool developed in generalized linear models is known as the DIC (Spiegelhalter *et al.*, 2002 [87]) where model assessment is based on the model deviance, $D(\theta) = -2\ln f(x|\theta)$, $f(x|\theta)$ is the likelihood function. DICV (used in Chapter 3 and Chapter 5) is a Bayesian analogue of the Akaike Information Criterion (AIC) and similar to AIC, but the complexity penalty depends on the number of effective model parameters, and not the actual number of free parameters as in AIC. Spiegelhalter *et al.* (2002) defined the complexity measure p_D, known as the effective number of parameters, to be the difference between the mean posterior deviance and the deviance evaluated at the posterior estimates of the

parameters given as [85]:

$$P_D = \overline{D(\theta)} - D(\hat{\theta}). \tag{2.42}$$

where $\overline{D(\theta)}$ is the mean posterior deviance ($E_\theta[D(\theta)]$) and $\hat{\theta}$ is an estimate of θ, often the posterior mean, median or mode. The DIC is then defined as:

$$DIC = \overline{D(\theta)} + p_D = D(\hat{\theta}) + 2p_D. \tag{2.43}$$

The deviance used in the DIC is computed at posterior means (or alternatively posterior medians or posterior modes). The penalty function for model complexity in the AIC is estimated by the nominal number of parameters in the model whereas DIC uses an estimate of the effective number of parameters in the model. The nominal number of parameters allows zero covariance among parameters whereas the effective number of parameters assumes non-zero covariance among parameters. Moreover, the nominal number of parameters is complicated to determine, especially in hierarchical models, and the effective number of parameters is typically estimated from the data. However, the effective number of parameters p_D cannot be intrinsically specified in missing data models such as mixture models [85]. The parameters θ may not be identifiable in these models and use of the posterior mean can give a poor estimate of p_D. In these cases, the posterior median or mode is the more pertinent choice. To identify the posterior mode is a problematic task in the Bayesian segmentation model as the set over which we are sampling is large [85]. Moreover, p_D can be negative for three reasons: i) if the posterior distribution is very dissimilar from the normal distribution, so that $f(\theta)$ does not give a good estimate of θ; ii) if the sampling distribution is non-log concave; and iii) if strong prior-data conflict is present (Gelman *et al.* [88]). Gelman *et al.* [88] and Sturtz *et al.* [89] proposed an alternative estimator of the effective dimension size instead of p_D used in Raftery *et al.* [90]. It is denoted as p_v and defined as:

$$p_v = Var(D(\theta))/2.$$

giving

$$DICV = \overline{D(\theta)} + p_v. \tag{2.44}$$

Raftery *et al.* [90] suggested that the effective number of parameters is substantially

overestimated by this version for normal random effect models.

2.5.3 Principal Component Analysis (PCA)

Principal component analysis (used as a summary statistic in Chapter 4 for dimensionality reduction) is a statistical approach that uses orthogonal transformation in order to convert a large number of possibly correlated variables into a small number of linearly uncorrelated variables. The resulting linearly uncorrelated variables are called principal components. It is a useful technique to analyze the interrelationships among a large number of variables and to reduce the number of variables needed to describe a data set with a minimum loss of information. Principal component analysis was invented by Karl Pearson in 1901, and is often used as a tool in predictive modeling and exploratory data analysis [91]. It is a common technique for finding patterns in data of high dimension [92, 93]. The technique is widely used in fields such as face recognition, image compression and computer graphics. PCA has also been used in time series analysis. Jolliffe wrote a chapter about PCA for time series and non-independent data in his book [94]. Some other authors used PCA for time series data including [95–98].

Let $X = [x_i]$ be any $p \times 1$ random vector. The first principal component (Y_1) is a linear combination of the variables $X_1, X_2, \cdots, X_p$

$$Y_1 = a_{11}X_1 + a_{12}X_2 + \cdots + a_{ap}X_p. \tag{2.45}$$

or, in matrix notation

$$\boldsymbol{Y_1 = a_T X.} \tag{2.46}$$

The first principal component is computed such that it achieves the greatest possible variance for such a linear combination. Note that, the sum of squares of the weights is constrained to be 1.

$$a_{11}^2 + a_{12}^2 + \cdots + a_{1p}^2 = 1.$$

The second principal component is then computed in the same way using the residuals obtained by subtracting the first principal component from the data. The second principal component is uncorrelated with (i.e., orthogonal to) the first principal component and accounts for the next highest variance. This process is iterated until a total of p principal components have been computed, where p is the original number of variables. The total

variance of all of the principal components will equal the total variance among all of the variables at this point. In matrix notation, the transformation of the original variables to the principal components is written as

$$\mathbf{Y}_1 = \mathbf{AX}. \tag{2.47}$$

The principal components are found by calculating the eigenvectors and eigenvalues of the data covariance matrix. One can calculate the variance-covariance matrix of the principal components as follows:

$$\mathbf{S_Y} = \mathbf{AS_X A^T}. \tag{2.48}$$

The rows of matrix $\mathbf{A}$ are the eigenvectors. The values within a particular row of matrix A are known as weights a_{ij}, and are also known as loadings. Large loadings imply that a particular variable has a strong relationship with a particular principal component. The elements in the diagonal of matrix $\mathbf{S_Y}$ are the eigenvalues, which are the variances of each principal component. These variances reduce monotonically from the first principal component to the last. Eigenvalues are commonly plotted on a scree plot, which facilitates visualising the decreasing rate at which variance is explained by additional principal components. The off-diagonal elements of matrix $\mathbf{S_Y}$ are zero, indicating zero covariance, as the principal components are independent.

The positions of each observation in this new coordinate system of principal components are called scores. These are computed as linear combinations of the original variables and the weights a_{ij}. For example, the score for the rth sample on the kth principal component is obtained as

$$Y_{rk} = a_{1k}x_{r1} + a_{2k}x_{r2} + \cdots + a_{pk}x_{rp}.$$

The correlations of the original variables with the principal components are used to interpret the principal components. The correlation of variable X_i and principal component Y_j is

$$r_{ij} = \sqrt{a_{ij}^2 Var(Y_j)/s_{ii}}.$$

The goal of principal components analysis is reduction of dimensionality. Many criteria have been suggested for ascertaining how many PCs should be examined and how many should be discarded. Four criteria are best evaluated by examining the scree plot [92]:

- Disregard principal components at the point at which the next PC shows little increase in the total explained variation.
- Incorporate all PCs up to a predetermined total percent explained variation, such as 90%.
- Disregard components whose explained variation is less than one when a correlation matrix is used, or less than the average variation explained when a covariance matrix is used, with the idea being that such a PC offers less than one variable's worth of information.
- Disregard the last PCs with variances all roughly equal.

In this thesis, I disregarded principal components according to the first criteria as the retained principal components can explain most of the variation from the data, and thus the discarded ones provide little information about the total explained variation.

References

1. Lovrić, M., Milanović, M. & Stamenković, M. Algoritmic methods for segmentation of time series: An overview. *Journal of Contemporary Economic and Business Issues* **1,** 31–53 (2014).
2. Harlé, F., Chatelain, F., Gouy-Pailler, C. & Achard, S. Bayesian Model for Multiple Change-points Detection in Multivariate Time Series. *arXiv preprint arXiv:1407.3206* (2014).
3. Kotti, M., Moschou, V. & Kotropoulos, C. Speaker segmentation and clustering. *Signal processing* **88,** 1091–1124 (2008).
4. Punskaya, E., Andrieu, C., Doucet, A. & Fitzgerald, W. J. Bayesian curve fitting using MCMC with applications to signal segmentation. *Signal Processing, IEEE Transactions on* **50,** 747–758 (2002).
5. Andreou, E. & Ghysels, E. Detecting multiple breaks in financial market volatility dynamics. *Journal of Applied Econometrics* **17,** 579–600 (2002).
6. Zheng, Y., Li, H. & Doermann, D. *The segmentation and identification of handwriting in noisy document images* in *International Workshop on Document Analysis Systems* (2002), 95–105.

7. Badagián, A. L., Kaiser, R. & Peña, D. in *Empirical Economic and Financial Research* 45–59 (Springer, 2015).

8. Khodadadi, A. & Asgharian, M. Change-point problem and regression: an annotated bibliography. *COBRA Preprint Series,* 44 (2008).

9. Killick, R. & Eckley, I. changepoint: An R package for changepoint analysis. *Journal of statistical software* **58,** 1–19 (2014).

10. Mohammad-Djafari, A. & Féron, O. Bayesian approach to change points detection in time series. *International Journal of Imaging Systems and Technology* **16,** 215–221 (2006).

11. Algama, M. & Keith, J. M. Investigating genomic structure using changept: A Bayesian segmentation model. *Computational and structural biotechnology journal* **10,** 107–115 (2014).

12. Aminikhanghahi, S. & Cook, D. J. A survey of methods for time series change point detection. *Knowledge and information systems* **51,** 339–367 (2017).

13. Basseville, M., Nikiforov, I. V., *et al. Detection of abrupt changes: theory and application* (Prentice Hall Englewood Cliffs, 1993).

14. Jensen, U. & Lütkebohmert, C. Change-point models. *Encyclopedia of Statistics in Quality and Reliability* **1** (2008).

15. Reeves, J., Chen, J., Wang, X. L., Lund, R. & Lu, Q. Q. A review and comparison of changepoint detection techniques for climate data. *Journal of Applied Meteorology and Climatology* **46,** 900–915 (2007).

16. Rodionov, S. A brief overview of the regime shift detection methods. *Large-scale disturbances (regime shifts) and recovery in aquatic ecosystems: challenges for management toward sustainability,* 17–24 (2005).

17. Truong, C., Oudre, L. & Vayatis, N. Selective review of offline change point detection methods. *Signal Processing,* 107299 (2019).

18. Chernoff, H. & Zacks, S. Estimating the current mean of a normal distribution which is subjected to changes in time. *The Annals of Mathematical Statistics,* 999–1018 (1964).

19. Smith, A. A Bayesian approach to inference about a change-point in a sequence of random variables. *Biometrika* **62,** 407–416 (1975).

20. Carlin, B. P., Gelfand, A. E. & Smith, A. F. Hierarchical Bayesian analysis of change-point problems. *Applied statistics,* 389–405 (1992).

21. Fearnhead, P. Exact Bayesian curve fitting and signal segmentation. *IEEE Transactions on Signal Processing* **53,** 2160–2166 (2005).

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

23. Wood, S., Rosen, O. & Kohn, R. Bayesian mixtures of autoregressive models. *Journal of Computational and Graphical Statistics* **20,** 174–195 (2011).

24. Kitagawa, G. & Akaike, H. A procedure for the modeling of non-stationary time series. *Annals of the Institute of Statistical Mathematics* **30,** 351–363 (1978).

25. Gerlach, R., Carter, C. & Kohn, R. Efficient Bayesian inference for dynamic mixture models. *Journal of the American Statistical Association* **95,** 819–828 (2000).

26. West, M., Prado, R. & Krystal, A. D. Evaluation and comparison of EEG traces: Latent structure in nonstationary time series. *Journal of the American Statistical Association* **94,** 375–387 (1999).

27. Chakar, S., Lebarbier, E., Lévy-Leduc, C., Robin, S., *et al.* A robust approach for estimating change-points in the mean of an AR(1) process. *Bernoulli* **23,** 1408–1447 (2017).

28. Chan, K. S. & Tong, H. On estimating thresholds in autoregressive models. *Journal of time series analysis* **7,** 179–190 (1986).

29. Tong, H. *On a threshold model in Pattern Recognition and Signal Processing* **29,** 575–586 (Sijthoff & Noordhoff, 1978).

30. Tong, H. & Lim, K. S. Threshold autoregression, limit cycles and cyclical data. *Journal of the Royal Statistical Society. Series B (Methodological),* 245–292 (1980).

31. Tong, H. *Threshold models in non-linear time series analysis* (Springer Science & Business Media, 2012).

32. Peguin-Feissolle, A. Bayesian estimation and forecasting in non-linear models application to an LSTAR model. *Economics Letters* **46,** 187–194 (1994).

33. Potter, S. Nonlinear time series modelling: An introduction. *Journal of Economic Surveys* **13,** 505–528 (1999).

34. Engle, R. F. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica: Journal of the Econometric Society,* 987–1007 (1982).

35. Bollerslev, T. Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics* **31,** 307–327 (1986).

36. Angelidis, T., Benos, A. & Degiannakis, S. The use of GARCH models in VaR estimation. *Statistical methodology* **1,** 105–128 (2004).

37. So, M. K. & Philip, L. Empirical analysis of GARCH models in value at risk estimation. *Journal of International Financial Markets, Institutions and Money* **16,** 180–197 (2006).

38. Awartani, B. M. & Corradi, V. Predicting the volatility of the S&P-500 stock index via GARCH models: the role of asymmetries. *International Journal of Forecasting* **21,** 167–183 (2005).

39. Gokcan, S. Forecasting volatility of emerging stock markets: linear versus non-linear GARCH models. *Journal of forecasting* **19,** 499–504 (2000).

40. Chong, C. W., Ahmad, M. I. & Abdullah, M. Y. Performance of GARCH models in forecasting stock market volatility. *Journal of forecasting* **18,** 333–343 (1999).

41. Cohen, I. Modeling speech signals in the time–frequency domain using GARCH. *Signal Processing* **84,** 2453–2459 (2004).

42. Bauwens, L., Hafner, C. M. & Laurent, S. *Handbook of volatility models and their applications* (John Wiley & Sons, 2012).

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

44. Kehagias, A. A hidden Markov model segmentation procedure for hydrological and environmental time series. *Stochastic Environmental Research and Risk Assessment* **18,** 117–130 (2004).

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

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

47. Keith, J. M. Segmenting eukaryotic genomes with the generalized Gibbs sampler. *Journal of Computational Biology* **13,** 1369–1383 (2006).

48. Sofronov, G. Y., Evans, G. E., Keith, J. M. & Kroese, D. P. Identifying change-points in biological sequences via sequential importance sampling. *Environmental Modeling & Assessment* **14,** 577–584 (2009).

49. Keith, J. M., Adams, P., Stephen, S. & Mattick, J. S. Delineating slowly and rapidly evolving fractions of the Drosophila genome. *Journal of Computational Biology* **15,** 407–430 (2008).

50. Brooks, S. Markov chain Monte Carlo method and its application. *Journal of the royal statistical society: series D (the Statistician)* **47,** 69–100 (1998).

51. Ruanaidh, J. J. O. & Fitzgerald, W. J. *Numerical Bayesian methods applied to signal processing* (Springer Science & Business Media, 2012).

52. Metropolis, N. & Ulam, S. The monte carlo method. *Journal of the American statistical association* **44,** 335–341 (1949).

53. Mossel, E. & Sly, A. Gibbs rapidly samples colorings of G (n, d/n). *Probability theory and related fields* **148,** 37–69 (2010).

54. Carlo, C. M. Markov chain monte carlo and gibbs sampling. *Lecture notes for EEB* **581** (2004).

55. Fan, Y. & Sisson, S. A. Reversible jump MCMC. *Handbook of Markov Chain Monte Carlo,* 67–92 (2011).

56. Lopes, H. F. A note on Reversible Jump Markov Chain Monte Carlo. *Graduate School of Business, The University of Chicago* (2006).

57. Keith, J. M., Kroese, D. P. & Bryant, D. A generalized Markov sampler. *Methodology and Computing in Applied Probability* **6,** 29–53 (2004).

58. Billio, M., Monfort, A. & Robert, C. P. Bayesian estimation of switching ARMA models. *Journal of econometrics* **93,** 229–255 (1999).

59. Lavielle, M. & Lebarbier, E. An application of MCMC methods for the multiple change-points problem. *Signal processing* **81,** 39–53 (2001).

60. Delyon, B., Lavielle, M., Moulines, E., *et al.* Convergence of a stochastic approximation version of the EM algorithm. *The Annals of Statistics* **27,** 94–128 (1999).

61. Chamroukhi, F., Mohammed, S., Trabelsi, D., Oukhellou, L. & Amirat, Y. Joint segmentation of multivariate time series with hidden process regression for human activity recognition. *Neurocomputing* **120,** 633–644 (2013).

62. Collilieux, X., Lebarbier, E. & Robin, S. A factor model approach for the joint segmentation with between-series correlation. *Scandinavian Journal of Statistics* **46,** 686–705 (2019).

63. Bai, J. & Perron, P. Computation and analysis of multiple structural change models. *Journal of applied econometrics* **18,** 1–22 (2003).

64. Cleynen, A. & Robin, S. Comparing change-point location in independent series. *Statistics and Computing* **26,** 263–276 (2016).

65. Rigaill, G., Lebarbier, É. & Robin, S. Exact posterior distributions and model selection criteria for multiple change-point detection problems. *Statistics and computing* **22,** 917–929 (2012).

66. Zhang, N. R., Siegmund, D. O., Ji, H. & Li, J. Z. Detecting simultaneous changepoints in multiple sequences. *Biometrika* **97,** 631–645 (2010).

67. Dobigeon, N., Tourneret, J.-Y. & Davy, M. Joint segmentation of piecewise constant autoregressive processes by using a hierarchical model and a Bayesian sampling approach. *Signal Processing, IEEE Transactions on* **55,** 1251–1263 (2007).

68. Dobigeon, N., Tourneret, J.-Y. & Scargle, J. D. Joint segmentation of multivariate astronomical time series: Bayesian sampling with a hierarchical model. *IEEE Transactions on Signal Processing* **55,** 414–423 (2007).

69. Barigozzi, M., Cho, H. & Fryzlewicz, P. Simultaneous multiple change-point and factor analysis for high-dimensional time series. *Journal of Econometrics* **206,** 187–225 (2018).

70. Ehsanzadeh, E., Ouarda, T. B. & Saley, H. M. A simultaneous analysis of gradual and abrupt changes in Canadian low streamflows. *Hydrological Processes* **25,** 727–739 (2011).

71. Hoai, M., Lan, Z.-Z. & De la Torre, F. *Joint segmentation and classification of human actions in video* in *In: Proceedings of CVPR, IEEE.* (2011), 3265–3272.

72. Picard, F., Lebarbier, É., Budinská, E. & Robin, S. Joint segmentation of multivariate Gaussian processes using mixed linear models. *Computational Statistics & Data Analysis* **55,** 1160–1170 (2011).

73. Picard, F. *et al.* Joint segmentation, calling, and normalization of multiple CGH profiles. *Biostatistics* **12,** 413–428 (2011).

74. Vert, J.-P. & Bleakley, K. *Fast detection of multiple change-points shared by many signals using group LARS* in *Advances in neural information processing systems* (2010), 2343–2351.

75. Brockwell, P. J. & Davis, R. A. *Introduction to time series and forecasting* (springer, 2016).

76. Hannan, E. J. *Multiple time series* (John Wiley & Sons, 2009).

77. Hamilton, J. D. *Time series analysis* (Princeton university press Princeton, NJ, 1994).

78. Box, G. E., Jenkins, G. M., Reinsel, G. C. & Ljung, G. M. *Time series analysis: forecasting and control* (John Wiley & Sons, 2015).

79. Enders, W. *Applied econometric time series* (John Wiley & Sons, 2008).

80. Folarin, S. B. & Iyiola, O. CAUSALITY AND INVERTIBILITY OF AUTOREGRESSIVE INTEGRATED MOVING AVERAGE (ARIMA) MODEL. *Asian Journal of Current Engineering and Maths* **2,** 260–266 (2013).

81. Palma, W. *Time series analysis* (John Wiley & Sons, 2016).

82. Akaike, H. in *Selected Papers of Hirotugu Akaike* 215–222 (Springer, 1974).

83. Anderson, D., Burnham, K. & White, G. Comparison of Akaike information criterion and consistent Akaike information criterion for model selection and statistical inference from capture-recapture studies. *Journal of Applied Statistics* **25,** 263–282 (1998).

84. Schwarz, G. *et al.* Estimating the dimension of a model. *The annals of statistics* **6,** 461–464 (1978).

85. Oldmeadow, C. & Keith, J. M. Model selection in Bayesian segmentation of multiple DNA alignments. *Bioinformatics* **27,** 604–610 (2011).

86. Oldmeadow, C., Mengersen, K., Mattick, J. S. & Keith, J. M. Multiple evolutionary rate classes in animal genome evolution. *Molecular biology and evolution* **27,** 942–953 (2009).

87. Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & Van Der Linde, A. Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology)* **64,** 583–639 (2002).

88. Gelman, A. *et al. Bayesian data analysis* (Chapman and Hall/CRC, 2004).

89. Sturtz, S., Ligges, U. & Gelman, A. E. R2WinBUGS: a package for running WinBUGS from R (2005).

90. Raftery, A. E., Newton, M. A., Satagopan, J. M. & Krivitsky, P. N. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. *Bayesian Statistics* **8,** 1–45 (2007).

91. Pearson, K. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **2,** 559–572 (1901).

92. Johnson, R. A., Wichern, D. W., *et al. Applied multivariate statistical analysis* **8** (Prentice hall Upper Saddle River, NJ, 2002).

93. Sanguansat, P. *Principal component analysis* (BoD–Books on Demand, 2012).

94. Jolliffe, I. Principal component analysis for time series and other non-independent data. *Principal Component Analysis,* 299–337 (2002).

95. Page, R. M., Lischeid, G., Epting, J. & Huggenberger, P. Principal component analysis of time series for identifying indicator variables for riverine groundwater extraction management. *Journal of hydrology* **432,** 137–144 (2012).

96. Abonyi, J., Feil, B., Nemeth, S. & Avra, P. Principal component analysis based time series segmentation: A new sensor fusion algorithm. *preprint* (2004).

97. Long, D. W., Brown, M. & Harris, C. *Principal components in time-series modelling* in *1999 European Control Conference (ECC),IEEE* (1999), 1705–1710.

98. Lansangan, J. R. G. & Barrios, E. B. Principal components analysis of nonstationary time series data. *Statistics and Computing* **19,** 173 (2009).

Chapter 3

Bayesian change-point modeling with segmented ARMA model

Chapter Objectives

The overall objective of this thesis is to extend and develop a Bayesian change-point segmented ARMA model for time series data and develop methods to segment multiple parallel time series. This chapter addresses the first part of my objective by introducing and validating a Bayesian change-point segmented ARMA model which considers the problem of modeling a time series by segmenting the series into blocks of autoregressive moving average (ARMA) processes. The performance of this model was assessed by applying it to a simulated and a real-world data. The results of this approach showed high detection accuracy for both simulated and real data. I compared the results of real data with the findings of other methods which are available in the existing literature. This model showed high sensitivity and detected a larger number of change-points than had been identified by those comparable methods.

Authorship

Farhana Sadia[1], Sarah Boyd[2], Jonathan M. Keith[1]

1 School of Mathematics, Monash University, Clayton, VIC 3800, Australia

2 Faculty of Information Technology, Monash University, Clayton, VIC 3800, Australia

Reference

Sadia F, Boyd S, Keith JM. (2018). Bayesian change-point modeling with segmented ARMA model. *PloS one* 13(12):e0208927. doi:10.1371/journal.pone.0208927.

PLOS ONE

RESEARCH ARTICLE

Bayesian change-point modeling with segmented ARMA model

Farhana Sadia[1], **Sarah Boyd**[2], **Jonathan M. Keith**[1]*

1 School of Mathematical Sciences, Monash University, Clayton, VIC, Australia, **2** Faculty of Information Technology, Monash University, Clayton, VIC, Australia

* jonathan.keith@monash.edu

OPEN ACCESS

Citation: Sadia F, Boyd S, Keith JM (2018) Bayesian change-point modeling with segmented ARMA model. PLoS ONE 13(12): e0208927. https://doi.org/10.1371/journal.pone.0208927

Editor: Tiago P. Peixoto, University of Bath, UNITED KINGDOM

Received: June 28, 2018

Accepted: November 26, 2018

Published: December 31, 2018

Data Availability Statement: All relevant data are available at: https://rss.onlinelibrary.wiley.com/hub/journal/14679868/series-b-datasets/pre_2016a. Those interested would be able to access the data in the same manner as the authors. The authors had no special access privileges.

Funding: This work was funded by the Australian Research Council (http://www.arc.gov.au/) grant DP1095849. The authors are grateful to the Australian Research Council (ARC) and the ARC Centre of Excellence for Mathematical and Statistical Frontiers for their support of this project (DP1095849, CE140100049). The funders had no

Abstract

Time series segmentation aims to identify segment boundary points in a time series, and to determine the dynamical properties corresponding to each segment. To segment time series data, this article presents a Bayesian change-point model in which the data within segments follows an autoregressive moving average (ARMA) model. A prior distribution is defined for the number of change-points, their positions, segment means and error terms. To quantify uncertainty about the location of change-points, the resulting posterior probability distributions are sampled using the Generalized Gibbs sampler Markov chain Monte Carlo technique. This methodology is illustrated by applying it to simulated data and to real data known as the well-log time series data. This well-log data records the measurements of nuclear magnetic response of underground rocks during the drilling of a well. Our approach has high sensitivity, and detects a larger number of change-points than have been identified by comparable methods in the existing literature.

Introduction

A time series is a succession of measurements made over a time interval. Some time series can be divided into a sequence of individual segments, each with its own unique characteristic properties. Identifying segment boundaries and inferring dynamical properties of different segments is referred to as time series segmentation. Change-point detection methods can be used to segment time series data since the goal of such methods is to select a sequence of change-point locations such that the observations are, in some sense, homogeneous within segments and different between segments [1].

Statistical analysis of change-point problems has been the subject of intensive research in the past half-century and there has been a large amount of literature on this subject (for reviews, see [2–5]). Literature on Bayesian change-point modeling is as extensive as for the classical change-point model. The Bayesian change-point model was pioneered by Chernoff and Zacks [6], who estimated the mean of a normal distribution for each segment in a Bayesian framework. Smith [7] proposed a Bayesian change-point model for finite series with normal and binomial models. To detect change-points in multivariate time series, Harlé *et al.* [8] used a Bayesian approach where change-points are modeled using Bernoulli variables for the

role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

change-point indicator vector. Successive generalizations and extensions of Bayesian methods for change-point problems include [9–13] and many others.

Different authors propose diverse models for each segment. For example, Punskaya *et al.* [14] suggested a Bayesian method for fitting piecewise linear regression models such as auto-regressive models. To explain regime switching behaviour in the conditional mean, Chan and Tong [15] proposed a new class of non-linear models, called the Smooth Transition Autoregressive (STAR) models. This model is an extension of the Threshold Autoregressive (TAR) model, introduced by Tong and Lim [16–18]. The TAR model is a piecewise linear model consisting of two or more linear sub-models. An indicator variable is used in the TAR model which represents a switch from one regime to another and takes a value zero or one, depending upon the values of a transition variable and a threshold parameter. This indicator variable implies abrupt jumps from one regime to the next. Chan and Tong [15] suggested the replacement of the indicator function with a smooth transition function in their STAR model, since sudden jumps from one regime to another may not be the best representation of the underlying mechanism generating observed data. Davis *et al.* [19] estimated structural breaks in a nonstationary time series by segmenting the series into blocks of distinct autoregressive (AR) processes. They assumed the number of breakpoints, their locations and the orders of the respective AR models are unknown. To segment non-stationary time series data, Wood *et al.* [20] developed a Bayesian mixture of autoregressive models where components of the model are time series and mixture probabilities. The time series have constant but unknown parameters and those mixture probabilities depend on time. They assumed unknown lag of the AR processes as well as an unknown number of components in the mixture model. To estimate the number and locations of multiple change-points in the mean of a Gaussian AR(1) process, Chakar *et al.* [21] proposed a new approach where the unknown autocorrelation coefficient and the variance of an "innovation" term remain unchanged from one segment to the other. They firstly estimated the autocorrelation coefficient and then decorrelate the series. After that they applied a dynamic programming algorithm to the decorrelated series.

As noted above, some existing models for time series segmentation have used a segmented AR model [19, 22]. This paper proposes to segment time series data using a segmented ARMA model, an approach that is surprisingly absent from the existing literature. Results for the well-log time series data discussed below show that fitting an ARMA model in each segment potentially identifies a greater number of change-points than the AR model in both real and simulated data and thus provides higher sensitivity. Moreover, we find that this is true for both large and small values of the variance σ^2 of the innovation term. Detailed explanations and comparison with AR model are provided in S3 Appendix.

The Bayesian segmented ARMA change-point model presented here resembles the approach of Keith *et al.* [23]. The major modelling innovation here is the use of an ARMA model in all segments. The posterior marginal distribution of this model are difficult to analyse directly since they have nonstandard form. However, simulated sampling can be performed via a Markov chain Monte Carlo (MCMC) algorithm. Several MCMC algorithms are available in the literature including the Metropolis-Hastings algorithm [24, 25], Gibbs Sampler [26], the Reversible Jump MCMC algorithm [9]; Multiple-Try Metropolis algorithm [27] and Delayed Rejection Metropolis-Hastings algorithm [28]. The Bayesian segmented ARMA change-point model here uses a highly efficient sampling technique known as the Generalized Gibbs Sampler (GGS) for generating samples from a posterior distribution [29]. The dimension in this algorithm does not need to be fixed and it provides flexibility to sample from varying dimensional spaces. The Generalized Gibbs Sampler (GGS) has been applied to some very high dimensional problems (see [30, 31]). It has resulted in highly efficient sampling for these

problems. However, no systematic comparison of the advantages and disadvantages of the GGS versus the reversible jump sampler has been performed.

Methodology

Problem statement

We consider the problem of modeling a time series by segmenting the series into blocks of autoregressive moving average (ARMA) processes. Let $t = 1, \cdots, T$ be time points in the signal or time series, where T represents the total length of the signal. Let x_t represent a real valued signal at time point, t. Let, $\mathbf{X} = (x_t)_{t=1}^{T}$ represent the time series vector or the signal that we want to segment. The ARMA model is:

$$x_t = c + \epsilon_t + \sum_{i=1}^{a} \psi_i (x_{t-i} - c) + \sum_{i=1}^{m} \theta_i \epsilon_{t-i}.$$

where $\psi_1, \cdots, \psi_a$ and $\theta_1, \cdots, \theta_m$ are the parameters of the AR and MA sub-models, respectively; a and m denote the order of the AR and MA sub-models; c is the mean of the ARMA model; ϵ is white noise and x_t is the time series. The number of change-points and their locations are assumed unknown. Each segment in the time-series is assumed to be generated by an ARMA model with different mean, and the goal is to infer the most probable segment locations and model parameters that describe them.

Likelihood model

We start by writing down the likelihood function of a model in which the sequence within each segment is generated by an ARMA process. This determines the probability of generating the observed sequence for any given parameter values. For each position in the signal except the first, the probability of starting a new segment at that position is denoted by ϕ. Thus a time series with K segments that have starting positions $\mathbf{s} = (1 = s_1 < \cdots < s_K \leq T)$ is generated with probability:

$$p(K, \mathbf{s}|\phi) = \phi^{K-1}(1 - \phi)^{T-K}. \quad (1)$$

Here, $s_1 = 1$, indicating that the first segment always starts at the beginning of the signal. Let the right hand points of the segments be $\mathbf{d} = (d_1, \cdots, d_K)$ where $d_K = T$ so that the last segment always finishes at the end of the signal. Let X_k be the signal of the segment between positions s_k and d_k inclusive. Each segment is then assigned to one of N groups with probabilities $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_N)$ where π_n is the probability of assigning any segment to group n. We denote the group to which segment k is assigned by $g_k \in \{1, \cdots, N\}$ where $\mathbf{g} = (g_1, \cdots, g_K)$. Then the probability of a specific assignment of the K segments to the N groups is

$$p(\mathbf{g}|K, \boldsymbol{\pi}) = \prod_{k=1}^{K} \boldsymbol{\pi}_{g_k}. \quad (2)$$

Let b_n be the number of segments with $g_k = n$. The probability of a specific assignment of the K segments to the N groups can be alternatively defined as:

$$p(\mathbf{g}|K, \boldsymbol{\pi}) = \prod_{n=1}^{N} \boldsymbol{\pi}_n^{b_n}. \quad (3)$$

Each segment is then modeled by an ARMA model. We write the ARMA model in each segment as:

$$x_t = c_k + \epsilon_t + \sum_{i=1}^{a} \psi_i (x_{t-i} - c_k) + \sum_{i=1}^{m} \theta_i \epsilon_{t-i}. \tag{4}$$

Here, c_k is the mean signal level for segment k and $c_k \sim \mathcal{N}(\mu_{g_k}, \tau^2_{g_k})$, where μ_{g_k} and $\tau^2_{g_k}$ are the mean and variance of the distribution of these means for the group g_k. Also, $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_T)$ is the vector of error terms and $\epsilon_t = x_t - \lambda_t$, where $\lambda_t = c_k + \sum_{i=1}^{a} \psi_i (x_{t-i} - c_k) + \sum_{i=1}^{m} \theta_i \epsilon_{t-i}$. We suppose that $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$, where σ^2 is the variance of error terms and this applies for $t \in (s_k, \cdots, d_k)$. In our model only these segment means differ between segments; other parameters of the ARMA model, $\boldsymbol{\psi}$ and $\boldsymbol{\theta}$, are the same for all segments. This assumption is appropriate for the applications used in this paper but other data sets may need to allow all of these parameters to differ between segments. In this paper, when we define the order of the AR (a) and MA (m) submodels we only intend for a and m to take values 0 or 1. Note that when $t - s_k$ is less than a or m, the expression for λ_t includes terms from the previous segment. However, for the left end of the signal when we can not look back a or m order steps, λ_t includes only the c_k term. An alternative is to choose the initial values for each segment by initializing them from the stationary distribution for the ARMA process in that segment.

The probability density of the observed signal is a product over all segments, that is, the probability of the signal $\mathbf{X}$ conditioning on parameters K, $\mathbf{s}$, $\boldsymbol{\theta}$, $\boldsymbol{\psi}$, $\mathbf{c}$ and σ^2 is expressed as a product of normal distributions with mean λ_t and variance σ^2 as follows:

$$\begin{aligned} p(\mathbf{X}|K, \mathbf{s}, \boldsymbol{\theta}, \boldsymbol{\psi}, \mathbf{c}, \sigma^2) &= \prod_{t=1}^{T} p(x_t|K, s, \theta, \psi, c, \sigma^2, x_{<t}) \\ &= \prod_{t=1}^{T} \mathcal{N}(x_t|\lambda_t, \sigma^2). \end{aligned} \tag{5}$$

Here, $x_{<t}$ indicates the signal value at time points $1, 2, \cdots, t-1$. The joint distribution of $\mathbf{X}$, K, $\mathbf{s}$, $\mathbf{g}$ and $\mathbf{c}$ conditional on the other parameters is given by:

$$\begin{aligned} p(\mathbf{X}, K, \mathbf{s}, \mathbf{g}, \mathbf{c}|\phi, \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\psi}, \sigma^2, \boldsymbol{\mu}, \boldsymbol{\tau}) \quad = \\ p(\mathbf{X}|K, \mathbf{s}, \boldsymbol{\theta}, \boldsymbol{\psi}, \mathbf{c}, \sigma^2) \times \\ p(K, \mathbf{s}|\phi) \times \\ p(\mathbf{c}|\mathbf{g}, \boldsymbol{\mu}, \boldsymbol{\tau}) \times \\ p(\mathbf{g}|K, \boldsymbol{\pi}) \end{aligned} \tag{6}$$

Here, $p(\mathbf{c}|\mathbf{g}, \boldsymbol{\mu}, \boldsymbol{\tau})$ is the probability of the ARMA mean for all segments given by:

$$p(\mathbf{c}|\mathbf{g}, \boldsymbol{\mu}, \boldsymbol{\tau}) = \prod_{k=1}^{K} \mathcal{N}(c_k|\mu_{g_k}, \tau^2_{g_k}). \tag{7}$$

Fig 1 shows the parameters of this model and their conditional dependencies. A parameter at the head of an arrow is conditionally dependent on the parameter at the tail.

Prior distribution

Since the segmented ARMA model is presented here in a hierarchical Bayesian framework, we have to assign prior distributions for the unspecified parameters. A beta prior is assigned for ϕ

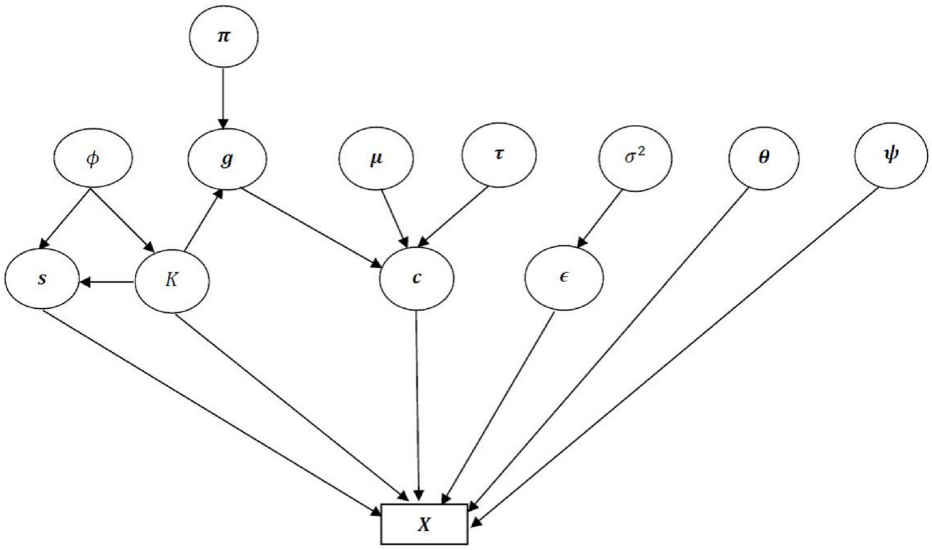


Fig 1. The conditional dependencies of the parameter.

https://doi.org/10.1371/journal.pone.0208927.g001

with parameters $a = 1.0$ and $b = 1.0$. Our computational algorithm is not sensitive to the choice of prior for ϕ, as we show in S4 Appendix. We assign a Dirichlet distribution to $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_N)$ with parameters $(\alpha_1, \cdots, \alpha_N) = (1.0, \cdots, 1.0)$ and $\sum_{n=1}^{N} \pi_N = 1$. Inferences are performed here using a weakly informative normal prior with mean 0.0 and variance 1.0 for the mean μ_n of the distribution of ARMA means in segment class N. We also assign an inverse gamma prior distribution with parameters $\alpha = 3.0$ and $\beta = 3.0$ for the variance (τ_n^2) of the distribution of $\mathbf{c}$ and the variance of the error terms (σ^2) of the ARMA model. The order of the AR model and the MA model will be considered fixed. Since we have no strong prior beliefs about the parameters of the ARMA model, we choose a uniform prior distribution for these on the interval (-1,1) and assume they are independent of each other. The forms of these hyper-priors were chosen to reflect the degree of prior belief about their respective parameters. Note that in general the signal x_t can be shifted and scaled so that the above prior is appropriate.

Posterior distribution

Using Bayes' theorem the posterior distribution of parameters is:

$$\begin{aligned} &p(K, \mathbf{s}, \mathbf{g}, \mathbf{c}, \phi, \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\psi}, \sigma^2, \boldsymbol{\mu}, \boldsymbol{\tau} | \mathbf{X}) \\ &\quad \propto p(\mathbf{X}, K, \mathbf{s}, \mathbf{g}, \mathbf{c} | \phi, \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\psi}, \sigma^2, \boldsymbol{\mu}, \boldsymbol{\tau}) p(\phi) p(\boldsymbol{\pi}) p(\boldsymbol{\theta}) p(\boldsymbol{\psi}) p(\sigma^2) p(\boldsymbol{\mu}) p(\boldsymbol{\tau}) \end{aligned} \tag{8}$$

See S1 Appendix for details of the calculation of the conditional posterior distribution of each parameter.

Sampling

The posterior distribution obtained in S1 Appendix is sampled using a Markov chain Monte Carlo technique known as the Generalized Gibbs Sampler, or GGS [29]. The GGS algorithm cycles through a sequence of steps in which parts of a sampled element are updated, while other parts are held constant. These different types of update are analogous to the coordinate updates of the conventional Gibbs sampler and are known as "move-types". This technique resembles a

conventional Gibbs sampler but can be applied in a transdimensional setting, where it provides an alternative to the reversible jump sampler. This technique allows the number of change-points to vary: it cycles through segments inserting and deleting change-points, and shifting change-point positions. See S2 Appendix for details of the GGS algorithm. The three main stages of this algorithm in Bayesian change-point segmented ARMA model are: (also see Fig 2)

- Iterate through the segments doing insertion and deletion updates, segment group assignments (g_k) and segment mean updates (c_k).
- Iterate through the groups updating group parameters (group mean μ_g and group variance τ_g^2).
- Update all the other parameters ($\boldsymbol{\pi}$, σ^2, ϕ, $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$).

Move types. Fig 2 illustrates the following defined move-types:

- (I, k): Decide whether to insert a new change-point in segment k, and at what position.
- (D, k): Decide whether to remove change-point k or move it to a new position (for each change-point except the first).
- c_k: Update mean signal level c_k in segment $k = 1, 2, 3, \cdots, K$.
- g_k: Update segment group assignments g_k in segment $k = 1, 2, 3, \cdots, K$.
- μ_g: Update group mean μ_g for group g.
- τ_g^2: Update group variance τ_g^2 for group g.
- $(\boldsymbol{\theta}, \boldsymbol{\psi})$: Update $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$.
- $(\boldsymbol{\pi}, \sigma^2, \phi)$: Update all other parameters, $\boldsymbol{\pi}$, σ^2 and ϕ.

There are K I-moves, $K - 1$ D-moves, K moves for updating segment group assignments, K moves for updating mean signal level, N moves for updating group mean, N moves for updating group variance, a moves to update parameters of the AR model, m moves to update parameters of the MA model and finally three moves for updating other parameters $\boldsymbol{\pi}$, σ^2 and ϕ. The total number of moves for a sequence with K segments is:

$$T(K) = 4K - 1 + 2N + a + m + 3. \tag{9}$$

where N is the number of groups, a is the order of the AR model and m is the order of the MA model.

Insertion: Step (I, k). Each move-type of a GGS sampler involves drawing from a distribution over a subset of the target space. The insertion and deletion move-types mentioned in Fig 2 both involve selecting an element from a set containing only two elements. In this respect, these move-types resemble a Metropolis-Hastings update, in which a new element is proposed and either accepted or rejected with some probability. In fact the insertion and deletion move-types involve the same subsets: they differ only in which element of the set is regarded as the source or current element, and which is regarded as the proposed element. The two elements in these subsets each have non-zero probability of being selected, proportional to the target distribution multiplied by the probability of selecting the move type when that element is the current element (see [29] for details).

In an (I, k) insertion step (where $k \in \{1, \cdots, K\}$) the subset contains the current segmentation and a new segmentation with a change-point inserted somewhere in segment k. A new segment end-point position z is proposed between s_k and $d_k - 1$, inclusive. The location of z is selected from a uniform distribution over the set $\{s_k, \cdots, d_k - 1\}$. Then for the left segment

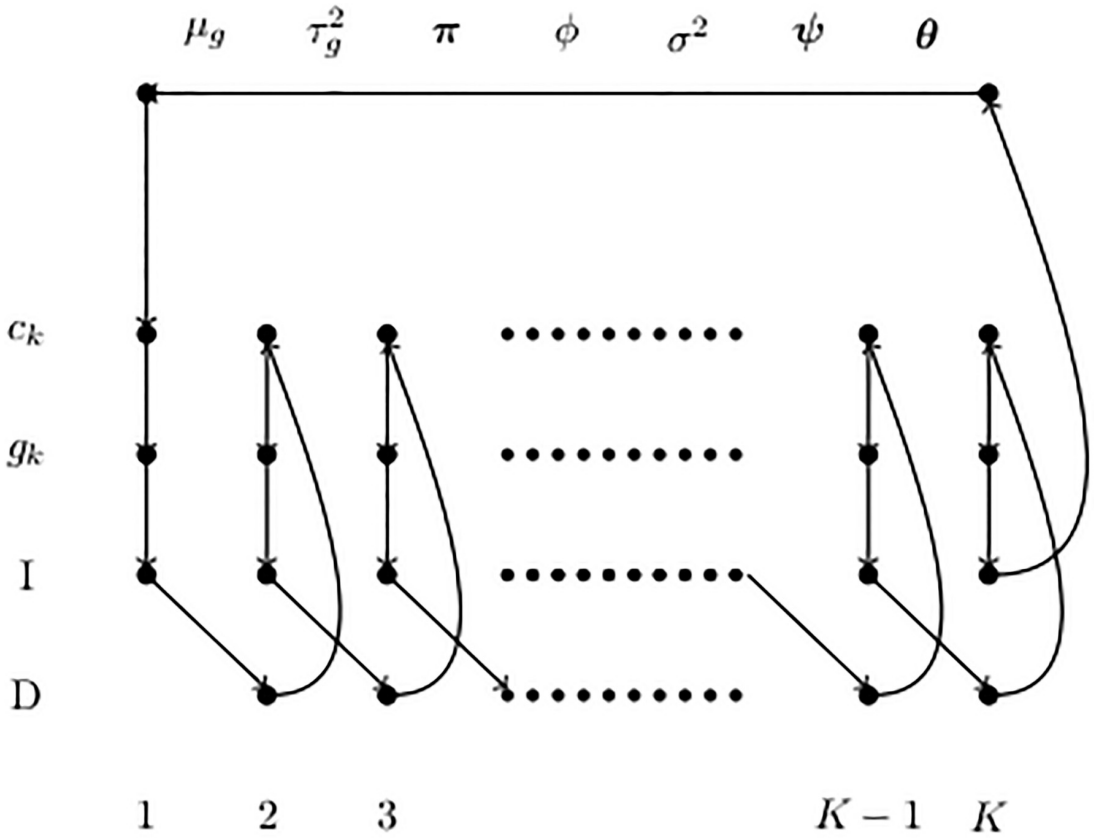


Fig 2. Order of move-types for the sampler. Note that *I* updates run from 1 to *K* whereas *D* updates run from 2 to *K*.

https://doi.org/10.1371/journal.pone.0208927.g002

(from s_k to z) new values of g'_k and c'_k are proposed and for the right segment (from $z+1$ to d_k) new values of g''_{k+1} and c''_{k+1} are proposed. Here, g'_k and g''_{k+1} are selected from a discrete distribution with parameters $\pi_1, \ldots, \pi_N$ and then c'_k and c''_{k+1} are selected from normal distributions with parameters $(\mu_{g'_k}, \tau^2_{g'_k})$ and $(\mu_{g''_{k+1}}, \tau^2_{g''_{k+1}})$ respectively. Choosing g'_k, g''_{k+1}, c'_k and c''_{k+1} in this manner results in cancellations such that the terms $p(\mathbf{c}|\mathbf{g}, \boldsymbol{\mu}, \boldsymbol{\tau})$ and $p(\mathbf{g}|K, \boldsymbol{\pi})$ in Eq 6 disappear when calculating the acceptance ratio.

After further cancellations, the new change-point at position $z+1$ is rejected with a probability proportional to

$$P_1 = (1-\phi)\prod_{t=s_k}^{d_k} p(\epsilon_t|0,\sigma^2)\,\frac{1}{(d_k-s_k)}\,\frac{1}{T(K)} \tag{10}$$

where

$$p(\epsilon_t|0,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{\epsilon_t^2}{2\sigma^2}\right]$$

and

$$\epsilon_t = x_t - \lambda_t = x_t - \left(c_k + \sum_{i=1}^{a}\psi_i(x_{t-i} - c_k) + \sum_{i=1}^{m}\theta_i\epsilon_{t-i}\right).$$

In this expression, $1/(d_k - s_k)$ is the probability of proposing the location of the new change-point and $1/T(K)$ is a correction factor used by the GGS to account for the number of move types available for the current segmentation with K segments.

Alternatively, the new change-point is accepted with probability proportional to:

$$P_0 = \phi \prod_{t=s_k}^{z} p(\epsilon'_t|0,\sigma^2) \times \prod_{t=z+1}^{d_k} p(\epsilon''_t|0,\sigma^2) \frac{1}{T(K+1)} \tag{11}$$

where

$$\begin{aligned} p(\epsilon'_t|0,\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{\epsilon'^2_t}{2\sigma^2}\right] \\ p(\epsilon''_t|0,\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{\epsilon''^2_t}{2\sigma^2}\right] \\ \epsilon'_t = x_t - \lambda'_t &= x_t - \left(c'_k + \sum_{i=1}^{a} \psi_i (x_{t-i} - c'_k) + \sum_{i=1}^{m} \theta_i \epsilon'_{t-i}\right) \end{aligned} \tag{12}$$

and

$$\epsilon''_t = x_t - \lambda''_t = x_t - \left(c''_k + \sum_{i=1}^{a} \psi_i (x_{t-i} - c''_k) + \sum_{i=1}^{m} \theta_i \epsilon''_{t-i}\right) \tag{13}$$

Here, Eq 12 applies for the left segment from s_k to z and Eq 13 applies for the right segment from $z + 1$ to d_k.

Thus the new change-point at $z + 1$ is accepted with probability $\frac{P_0}{P_0+P_1}$ or rejected with probability $\frac{P_1}{P_0+P_1}$. An alternative (which we have not implemented) is to use the Metropolis-Hastings acceptance probability $\min\{1, P_0/P_1\}$.

If a change-point is inserted, the move-type is updated to $(D, k+1)$, otherwise it remains (I, k). In either case, the move-type is then further updated as in Fig 2.

Deletion: Step (D, k). For each segment $k = 2, \cdots, K$, a (D, k) deletion step also involves selecting an element from a set containing only two elements: the current segmentation and a new segmentation with the change-point at the left end of segment k removed, so that segments $k - 1$ and k merge to form a new segment. Values of g_k and c_k are then chosen for the new merged segment: g_k is selected from a discrete distribution with parameters $\pi_1, \ldots, \pi_N$ and then c_k is selected from a normal distribution with parameters μ_{g_k} and $\tau^2_{g_k}$ respectively. Other parameters and the positions of other change-points are held constant.

The probability of accepting the deletion is proportional to:

$$P_1 = (1-\phi) \prod_{t=s_{k-1}}^{d_k} p(\epsilon_t|0,\sigma^2) \frac{1}{(d_k - s_{k-1})} \frac{1}{T(K-1)}$$

where

$$p(\epsilon_t|0,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{\epsilon^2_t}{2\sigma^2}\right]$$

and

$$\epsilon_t = x_t - \lambda_t = x_t - \left(c_k + \sum_{i=1}^{a} \psi_i (x_{t-i} - c_k) + \sum_{i=1}^{m} \theta_i \epsilon_{t-i}\right)$$

Here K is the number of segments in the current segmentation, that is, without making the deletion.

The probability of rejecting the deletion is proportional to:

$$P_0 = \phi \prod_{t=s_{k-1}}^{z} p(\epsilon_t'|0,\sigma^2) \times \prod_{t=z+1}^{d_k} p(\epsilon_t''|0,\sigma^2)\frac{1}{T(K)}.$$

Thus the deletion is accepted with probability $\frac{P_1}{P_0+P_1}$ or rejected with probability $\frac{P_0}{P_0+P_1}$.
The conditional posterior distribution for $\mathbf{c}$ and $\boldsymbol{\epsilon}$ is proportional to:

$$\prod_{t=s_k}^{d_k} p(x_t|\lambda_t,\sigma^2) \times p(c_k|\mu_{g_k},\tau_{g_k}^2).$$

The $\mathbf{c}$'s and the corresponding $\epsilon_{s_k},\cdots,\epsilon_{s_{k+1}}$ can be updated one segment at a time with this conditional posterior distribution. Now, the conditional posterior distribution for $\boldsymbol{\theta}$ and ϵ is proportional to:

$$\prod_{t=1}^{T} p(x_t|\lambda_t,\sigma^2) \times p(\theta).$$

To update $\boldsymbol{\psi}$, similar procedures were used. The sampler iterates through the groups, updating group parameters ($\boldsymbol{\mu}$ and $\boldsymbol{\tau}$) with their respective conditional posterior distributions.

The sampler uses conventional Gibbs updates for other parameters c_k, g_k, θ_i, ϕ_i, σ^2, $\boldsymbol{\pi}$, ϕ, μ_g and τ_g. We use the Slice sampler [32] where needed to draw from non-standard distributions.

Validation of the methodology

We have used a simulation-based method for testing the correctness of software for fitting Bayesian models using posterior simulation [33]. This validation technique is based on posterior quantiles. Consider the general Bayesian joint distribution $p(y|\Theta)p(\Theta)$, where $p(y|\Theta)$ presents the sampling distribution of the data, $p(\Theta)$ presents the proper prior distribution of the parameter vector Θ, and inferences are based on the posterior distribution, $p(\Theta|y)$. The validation method samples a parameter vector $\Theta^{(0)}$ from $p(\Theta)$. Then conditional on $\Theta^{(0)}$, this technique samples data y from $p(y|\Theta=\Theta^{(0)})$ and then simulates sampling from the posterior distribution, $p(\Theta|y)$ using the software to be validated. The resulting posterior sample of size L is denoted $(\Theta^{(1)},\Theta^{(2)},\cdots,\Theta^{(L)})$. Finally, for each coordinate of $\Theta^{(0)}$, denoted $\theta^{(0)}$, compute its posterior quantile $\hat{q}(\theta^{(0)})$, with respect to the posterior sample $(\theta^{(1)},\theta^{(2)},\cdots,\theta^{(L)})$. To perform the validation procedure, many replications are required, each drawing $\Theta^{(0)}$ from $p(\Theta)$ and y from $p(y|\Theta^{(0)})$. The simulation output is a collection of estimated posterior quantiles. From these quantiles calculate a test statistic $X_\theta^2 = \sum_{i=1}^{N_{rep}} \left(\Phi^{-1}(q_i)\right)^2$ where, $q_i = \frac{1}{L}\sum_{l=1}^{L} I_{\theta_i^{(0)}>\theta_i^{(l)}}$ is the posterior quantile for the ith replication, N_{rep} is the total number of replications, θ denotes component of Θ and Φ represents the standard normal CDF. If the software is implemented correctly, this test statistic follows a χ^2 distribution with N_{rep} degrees of freedom and also the posterior quantiles will be uniformly distributed. The posterior quantiles' deviation from uniformity can be quantified by calculating the associated p value, that is, p_θ for each X_θ^2. Extremely small p_θ values indicate an error in the software. As an exploratory tool, p_θ values can be transformed into a z_θ statistic ($z_\theta = \Phi^{-1}(p_\theta)$). If all $|z_\theta|$ statistics are not extreme, such as less than 2, the software may be considered validated.

To validate our method according to the above described technique, we generated the parameters τ^2, σ^2, π, μ from the following prior distributions:

$$\tau^2 \sim Inv - gamma(3,3)$$
$$\sigma^2 \sim Inv - gamma(3,3)$$
$$\pi \sim U(0,1)$$
$$\mu \sim N(0,1)$$
$$\theta \sim U(-1,1)$$
$$\psi \sim U(-1,1)$$

We generated 20 sequences from an ARMA model with 20 different segment means where every sequence has length 100. We simulated 5000 draws from the posterior distribution of the model parameters. Then the quantiles of the posterior distributions for each parameter were determined and the whole procedure was repeated 20 times. From these quantiles we determined the absolute values of the z_θ statistic. The absolute z_θ statistics from this simulation are plotted in Fig 3:

In the above plot, since the z_θ statistic for each parameter is less than 2, we conclude the software is correctly written. More precisely, we find no evidence of software errors.

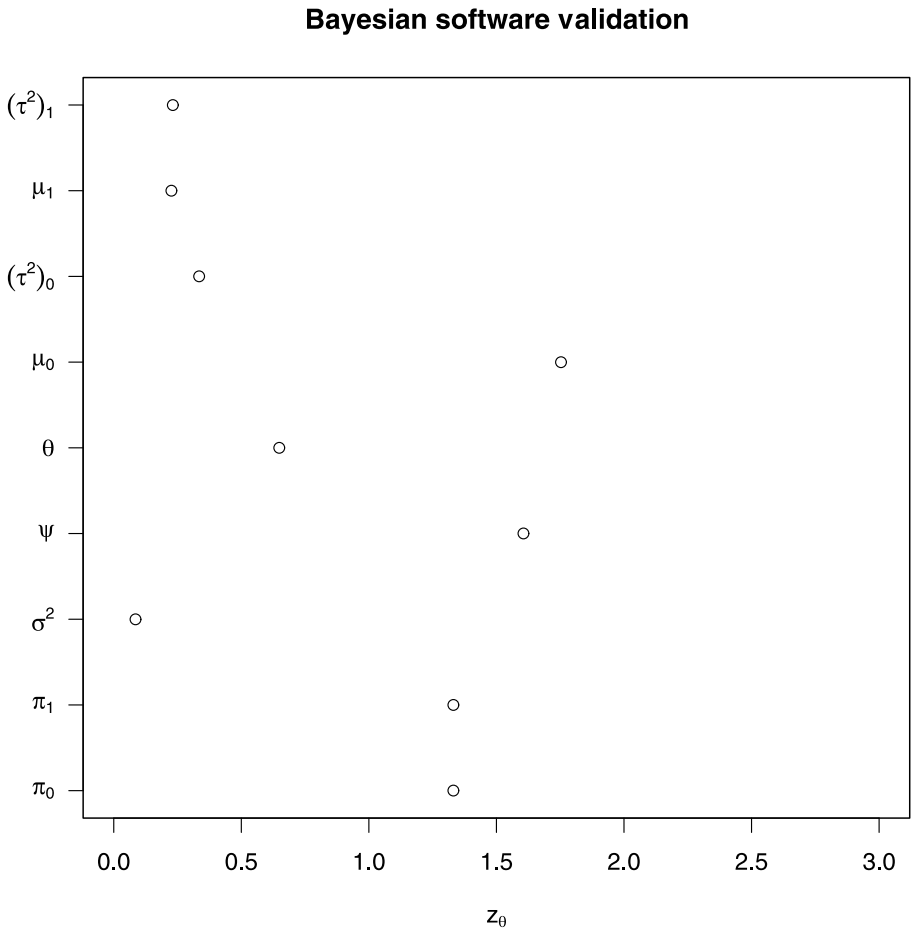


Fig 3. Absolute z_θ statistic plot. Each row shows a parameter of the segmented ARMA model and the $|z_\theta|$ statistics associated with these parameters are displayed as a circle in each row.

https://doi.org/10.1371/journal.pone.0208927.g003

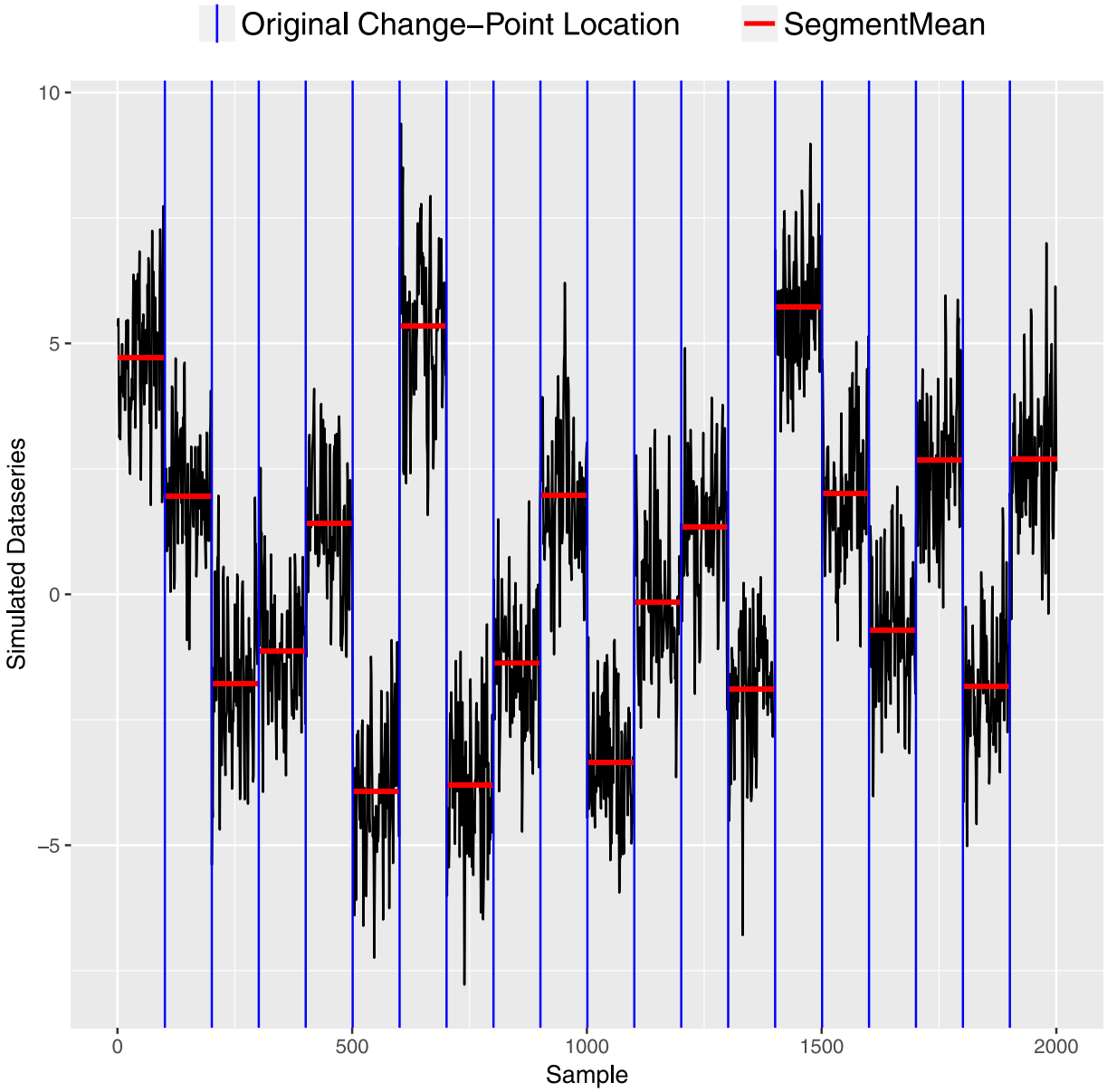


Fig 4. Simulated signal. The true change-point locations are shown as vertical blue lines and segment means are shown as horizontal red lines.

https://doi.org/10.1371/journal.pone.0208927.g004

Illustrative examples

Simulation example

As a test of our method, we applied it to a simulated example in which the number and location of change-points, ARMA parameters, segment means and error variance are known. We analyzed 20 time series, each containing 100 observations, generated from the autoregressive moving average (ARMA(1,1)) model with parameter values $\psi = 0.22$ and $\theta = 0.60$. Each series was generated using $\sigma^2 = 0.96$ and 20 different segment means. The simulated ARMA data with the true segment means and the location of change-points is shown in Fig 4.

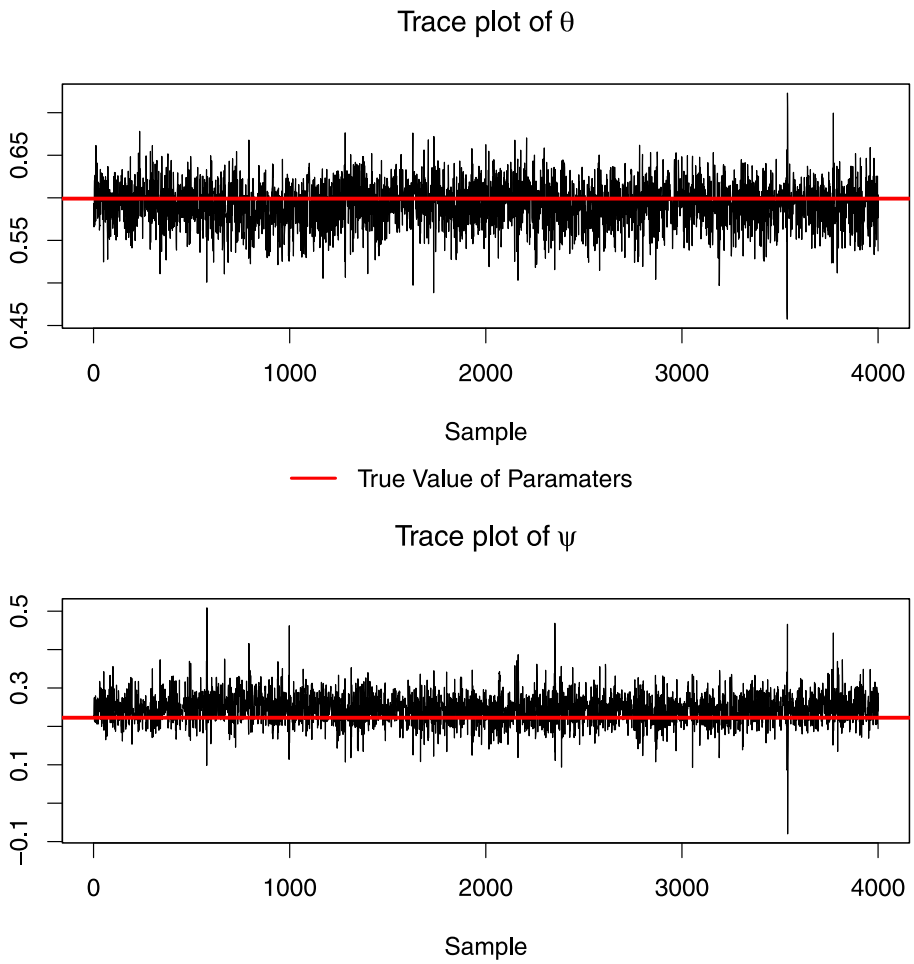


Fig 5. Trace plot of AR and MA parameters. Both parameters converged to the true values.

https://doi.org/10.1371/journal.pone.0208927.g005

We executed 5,000 iterations of the MCMC estimation algorithm and the first 1,000 iterations were treated as a burn in period and discarded. The convergence of AR and MA parameters is evident in Fig 5. Both AR and MA parameters converge, display good mixing and are close to the true values of those parameters.

The top panel of Fig 6 presents the simulated signal with the true change-points (red vertical line). The middle plot shows the posterior distribution of occurrence of change-point locations, that is, the posterior probability of being change-points at each position. The height of the (red) 'spikes' indicates the posterior probability of a change-point being selected at each time point. The top two plots show the locations of estimated change-points and the true change-points are similar. This picture is clearer if we compare actual change-point locations and estimated change-point locations with the simulated data, as in the top and bottom plots of Fig 6. The blue lines indicate time points at which the posterior probabilities of change-points are greater than 0.5. Fig 6 (bottom) identifies 17 change-points out of 19 true change-points.

Fig 7 plots posterior estimators of mean signal level (c) at each position of the simulated signal. This plot clearly indicates 20 segments in the simulated signal detecting a change in mean even where change-points were not detected in Fig 6.

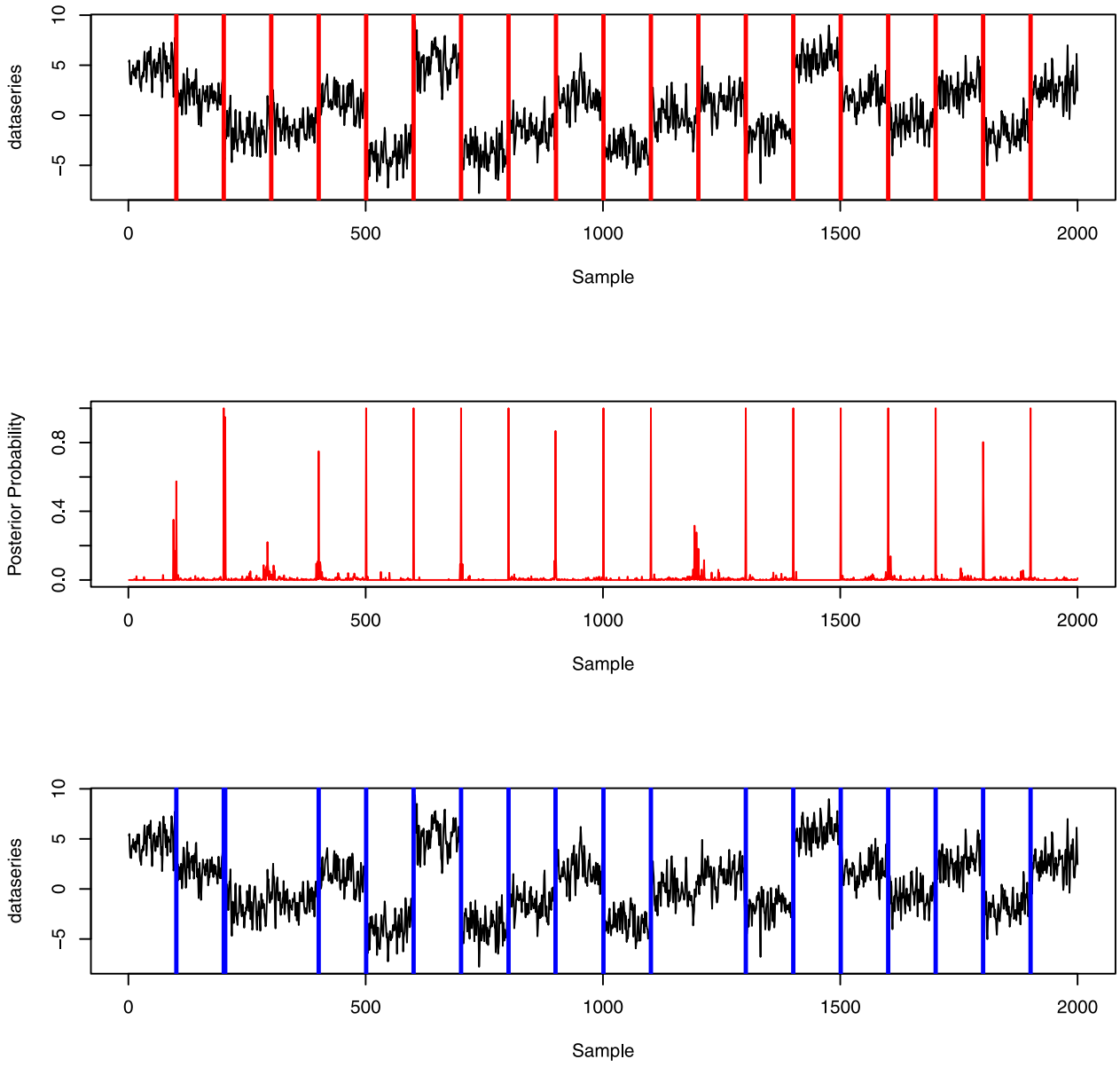


Fig 6. (Top) Segmented signal with the true change-point locations. (Middle) Posterior probabilities of occurrence of change-points. (Bottom) Estimated change-point locations (posterior probability greater than 0.5). The middle plot shows the location of peaks in the probability profile closely follows the true change-points locations but in some positions with low posterior probability. Using a threshold in posterior probability 0.5, we identify 17 change-points out of 19 which match the locations of the true change-points.

https://doi.org/10.1371/journal.pone.0208927.g006

Well-log data

We now apply the segmented ARMA model to identify change-points in a real data set. This data records 4050 measurements of nuclear magnetic response of underground rocks during the drilling of a well. During drilling, data were obtained at discrete time points by lowering a probe into a bore-hole in the Earth's surface. This geophysical data originates from Ó Ruanaidh and Fitzgerald [34] and has been previously analyzed in the context of change-point

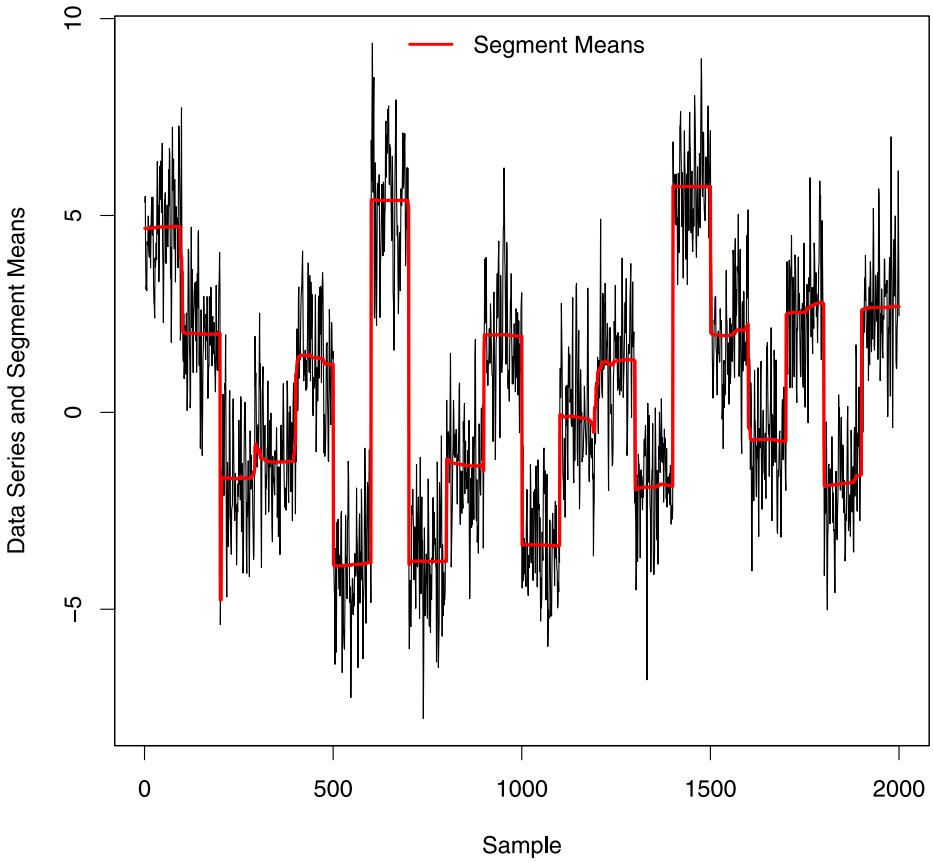


Fig 7. Segment means at each position of simulated data.

https://doi.org/10.1371/journal.pone.0208927.g007

detection by many researchers, for example, by Fearnhead and Clifford [35], Fearnhead [36] and by Whiteley *et al.* [1]. A few outliers present in this data were removed by hand before analyzing the data as in Fearnhead [36]. The data is shown in Fig 8.

The piecewise constant signal of this data indicates information about the geophysical structure of the rocks in the well. Changes in mean occur at each time point whenever a new rock type is found. To identify change-points in this well-log data, Ó Ruanaidh and Fitzgerald [34] used a Gibbs sampler to fit a Bayesian change-point model with a fixed number of change-points; Fearnhead and Clifford [35] used on-line Bayesian analysis of data with a hidden Markov model using particle filters; Fearnhead [36] considered an extension of the model described by Fearnhead and Clifford [35] in which they considered all the parameters of their model to be unknown and they used reversible jump MCMC to fit that model; Whiteley *et al.* [1] considered the same model used by Fearnhead [36] to analyze well-log data but here they used a block Gibbs sampler for generating samples from the posterior distribution.

To find change-points in this data, we attempted to fit an AR(1) model, an MA(1) model and an ARMA(1,1) model. We are the first to investigate this data using a segmented ARMA model. Each model was run for 5000 iterations and then tested for convergence of each parameter using trace plots and an autocorrelation function (ACF) plot which represents the degree of correlation between all pairs of samples separated by progressively larger lags (number of

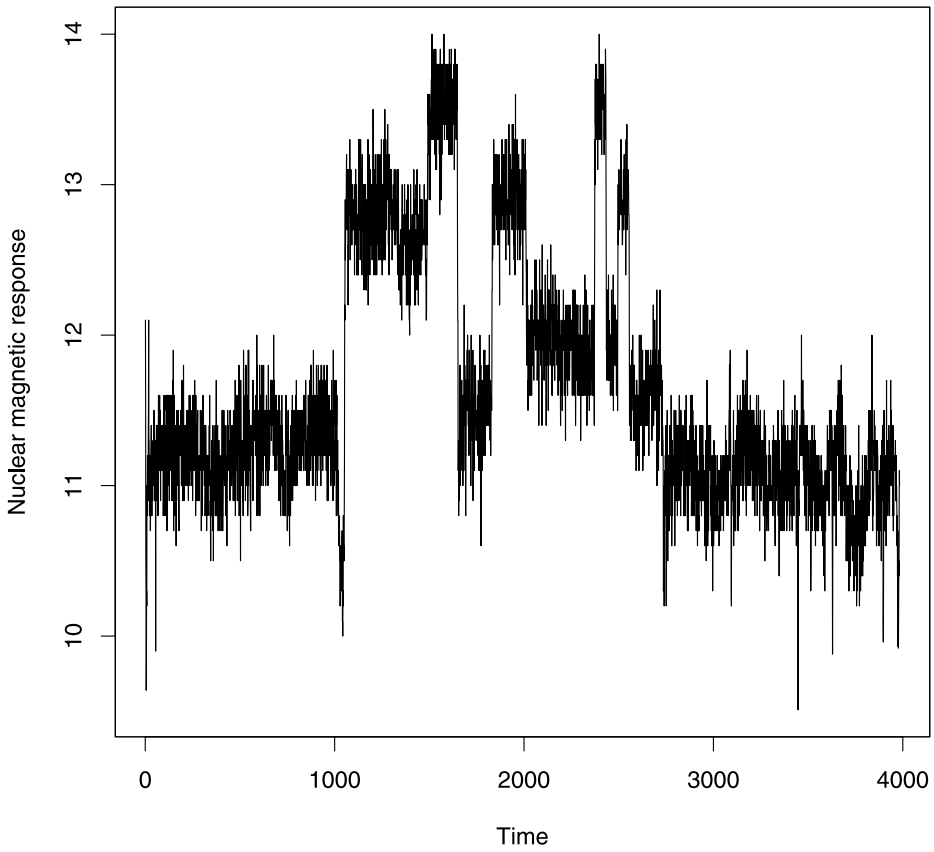


Fig 8. Well-log data. This data provides information about the rock structure of the well. Some change-points are present in the data reflecting the presence of a new rock.

https://doi.org/10.1371/journal.pone.0208927.g008

samples). The change-point profiles and segment mean profiles for each model are given in Figs 9–11.

In our model, the initial segmentation was generated using a probability of starting a new segment $\phi = 0.1$. The initial segmentation was generated by throwing a uniform(0,1) random number for each sequence position except the first, making that position a change-point if the random number is less than ϕ at that position. Note that, this initial segmentation should not affect the stationary distribution of the Markov chain. The posterior probabilities of occurrence of change-points at each position of the input sequence are calculated using the uniform prior probability distribution for ϕ (the probability that any given sequence position is a change-point) and the likelihood probability ($p(K, \mathbf{s}|\boldsymbol{\phi}) = \boldsymbol{\phi}^{K-1}(1-\phi)^{T-K-1}$) of generating a new segmentation with K change-points and $\mathbf{s} = (1 = s_1 < \cdots < s_K \leq T)$ starting positions.

All change-point profile plots (Figs 9–11) show almost the same change-points locations. But at some locations the AR(1) model gives comparatively smaller posterior probability than the MA(1) and ARMA(1,1) models. In addition, the AR(1) model identifies fewer change-points with high posterior probability (when the posterior probability is more than 0.5) than the other two models. These results are somewhat similar to the previous results found in the change-point literature [1, 34–36].

Since the three segmented ARMA models indicate many of the same change-point locations, we compare the change-points locations for which posterior probabilities are greater

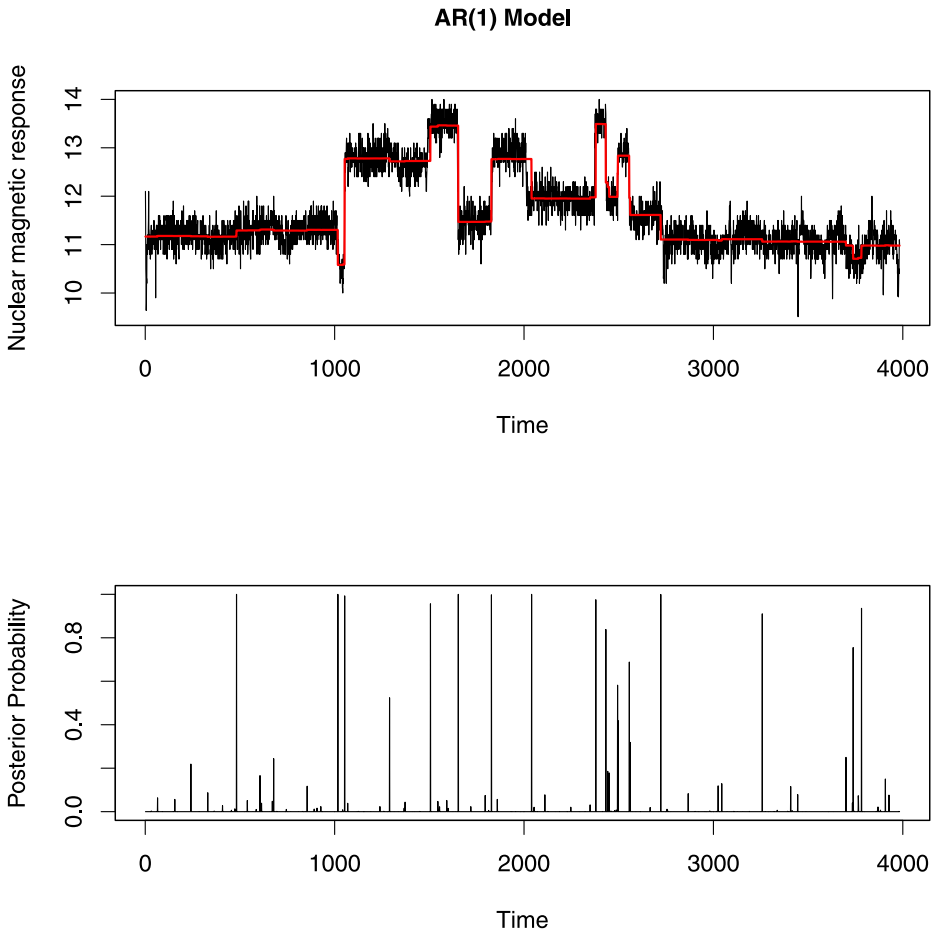


Fig 9. Top plot shows the posterior estimators of mean signal level (c) on each segment of AR(1) model. Bottom plot shows the posterior probability of a change-point at each position. These segment means and change-point positions indicate significant jumps in the original data.

https://doi.org/10.1371/journal.pone.0208927.g009

than 0.5 in Fig 12. Among these three models, the ARMA(1,1) model identifies the largest number of change-points and matches more closely with the number and locations of change-points detectable to the eye. Moreover, the ARMA(1,1) model picks up small changes in mean with high posterior probability whereas the AR(1) and MA(1) models missed change-points at some time points where small jumps occured in the data.

To determine the best model among these three, we compare these models using the deviance information criterion (DICV). The DICV is defined as: DICV $= p_v + \overline{D(\theta)}$, where $\overline{D(\theta)}$ is the mean posterior deviance, $p_v = Var(D(\theta))/2$ and deviance $D(\theta) = -2\ln f(y|\theta)$ (details of DICV are in [37, 38]). The DICV, $\overline{D(\theta)}$ and p_v of these three models are shown in Table 1.

Here, the segmented ARMA(1,1) model gives lower DICV than the other two models. The lower DICV of the ARMA(1,1) model supports the conclusion that the ARMA(1,1) model is the best of the three.

Discussion

In this paper, we have developed a Bayesian change-point segmented ARMA model to segment time series data. The novel features of our approach include: (1) It uses an ARMA model

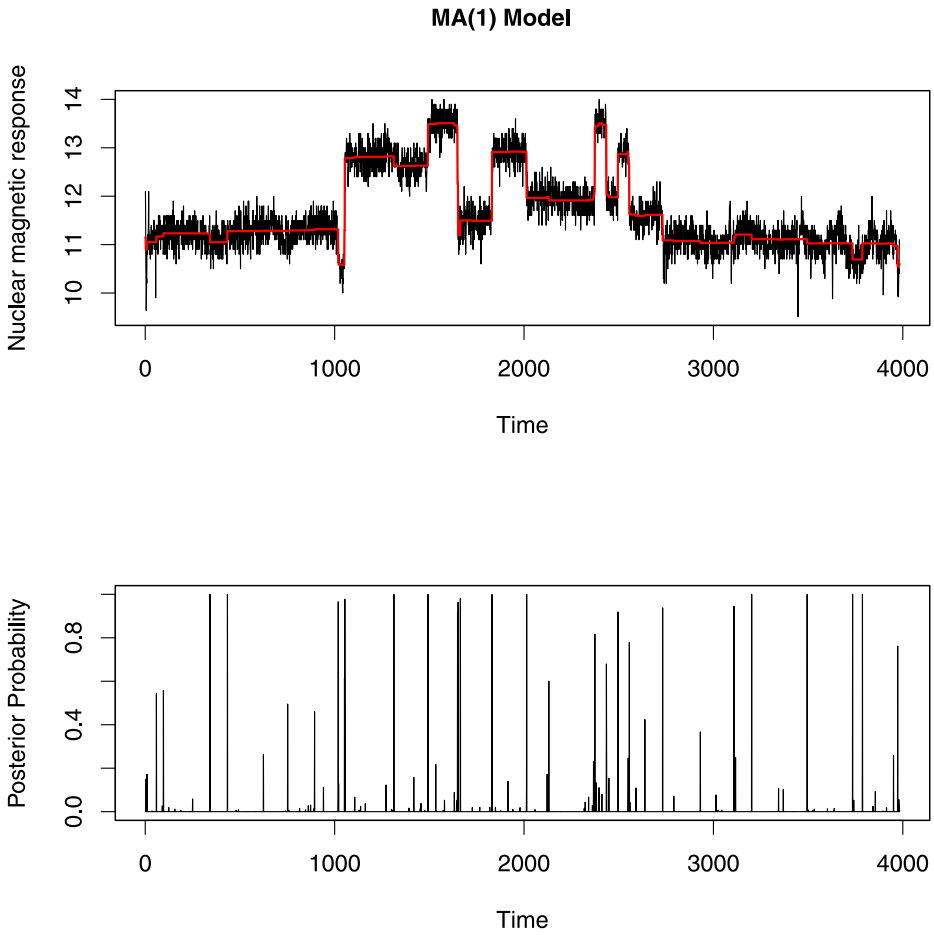


Fig 10. Top and bottom plot show the segment mean profiles and the change-point profiles for MA(1) model respectively. This model identifies more change-points than the AR(1) model.

https://doi.org/10.1371/journal.pone.0208927.g010

in each segment (2) It uses a highly efficient sampling technique (GGS) to generate samples from a posterior distribution. Results for simulated data and real data show that this model achieves high detection accuracy.

The results we obtain for the well-log data seem reasonable when judged by eye. The posterior probabilities of change-point occurrences of the ARMA(1,1) model are somewhat similar to results previously found in the literature for this data. Ó Ruanaidh and Fitzgerald [34] assumed 13 change-points exist in this data (after removing outliers) but our ARMA (1,1) model identifies 27 change-points (considering posterior probability more than 0.5). They missed some small jumps in the data whereas the ARMA(1,1) model identifies small jumps as well as large jumps. Fearnhead and Clifford [35] inferred 16 change-points but since they did not remove the outliers of the data, the results cannot be directly compared. Fearnhead [36] found too many change-points in their piecewise constant model. They used a random walk sampler and found similar change-points to our ARMA(1,1) model. In addition to the change-points found by Fearnhead [36], the ARMA(1,1) model identifies some significant change-points between time points 1 and 1000 and after 2800 (shown in Figs 13 and 14).

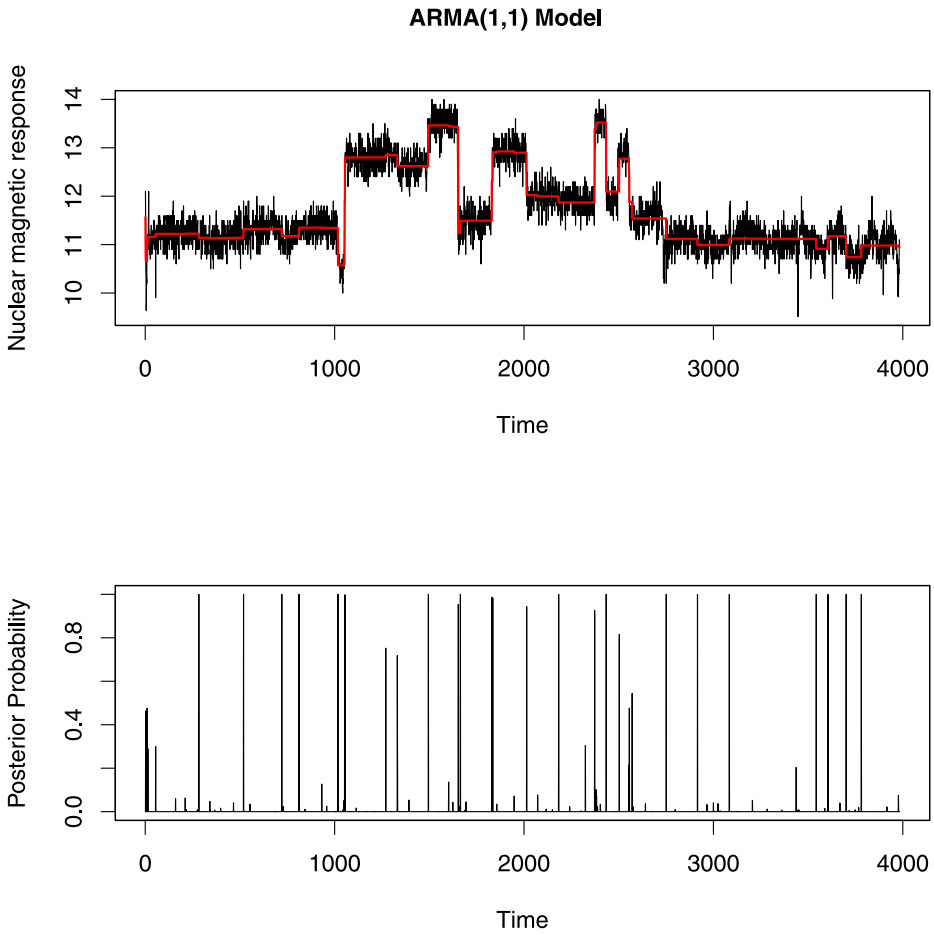


Fig 11. Top and bottom plot show the segment mean profiles and the change-point profiles for ARMA(1,1) model respectively. These change-points and segment means are almost identical to those identified using the MA(1) model.

https://doi.org/10.1371/journal.pone.0208927.g011

The results of the ARMA(1,1) model are similar to the results reported in Whiteley *et al.* [1] but in some time points ARMA(1,1) model shows a higher posterior probability of change-point occurrences.

Overall, this paper has presented a promising new direction for estimation of change-point models by assuming a segment-wise ARMA model. Most of the previous methods discussed in the literature use autoregressive (AR) models in each segment. Adding a moving average component helps to consider the dependence between residual terms which is an advantage over the segmented AR model. Our results obtained using simulated data demonstrate that when the data are generated via an ARMA process, the ARMA model finds more change-points than the AR model, without finding false positives. The ARMA model also finds more change-points in the well log data, but a question remains whether the additional change-points are false positives in this case. As the true locations of change in this data set are unknown, this question cannot be answered definitively. However, we compared the segmented ARMA (1,1) model, segmented AR(1) model and segmented MA (1) model using the deviance information criterion (DICV), and found that the ARMA (1,1) is favoured by this criterion, suggesting the additional change-points reflect a real feature of the data. Since this model assumes the same

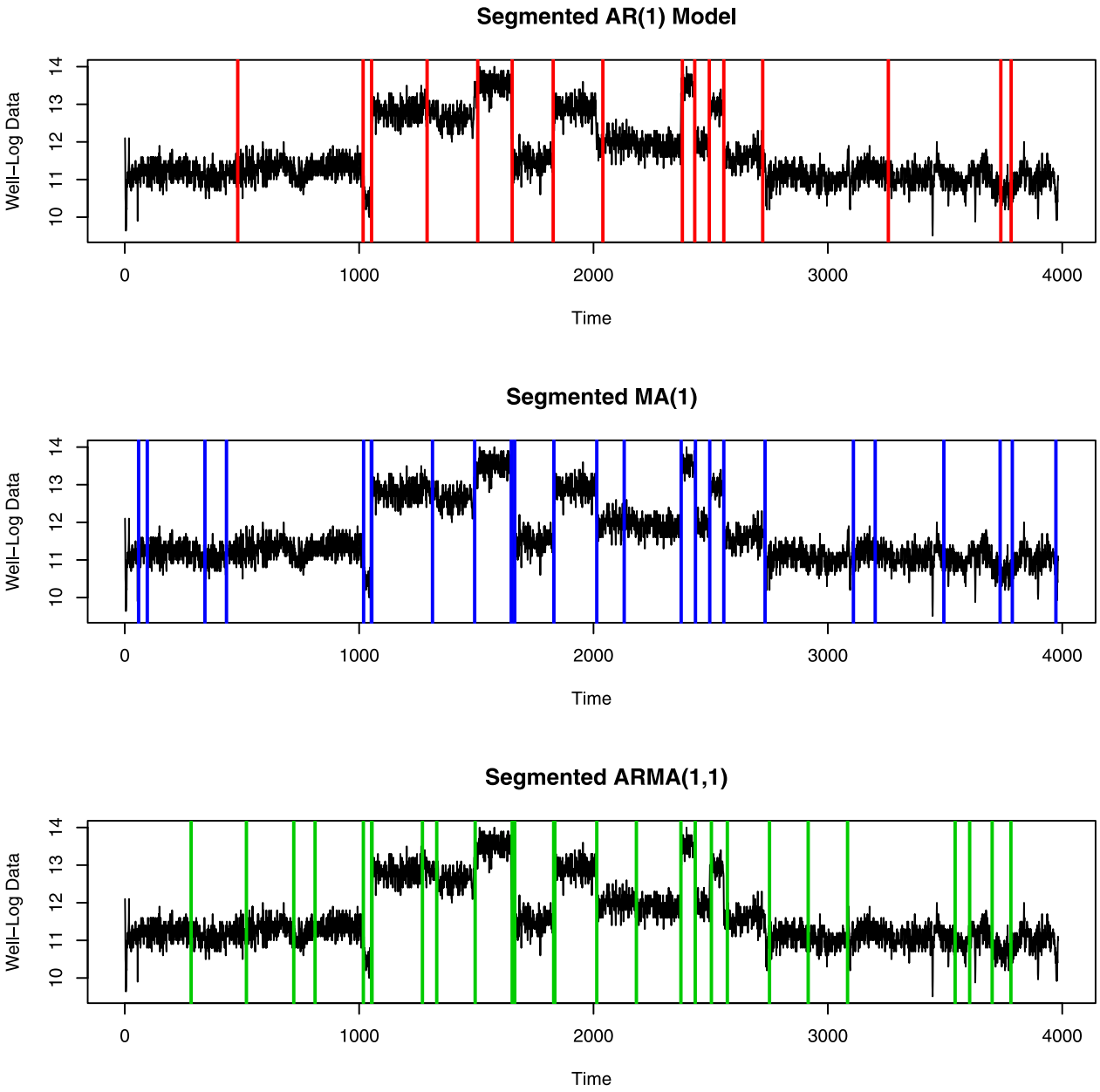


Fig 12. Estimated change-point locations with posterior probability greater than 0.5 for AR(1), MA(1) and ARMA(1,1) model.

https://doi.org/10.1371/journal.pone.0208927.g012

Table 1. DICV of Models.

Model	$\overline{D(\theta)}$	p_v	DICV
Segmented AR(1) Model	958.0043	181306.7	182264.7
Segmented MA(1) Model	509.5514	253045.4	253554.9
Segmented ARMA(1,1) Model	672.8849	169945.6	170618.5

https://doi.org/10.1371/journal.pone.0208927.t001

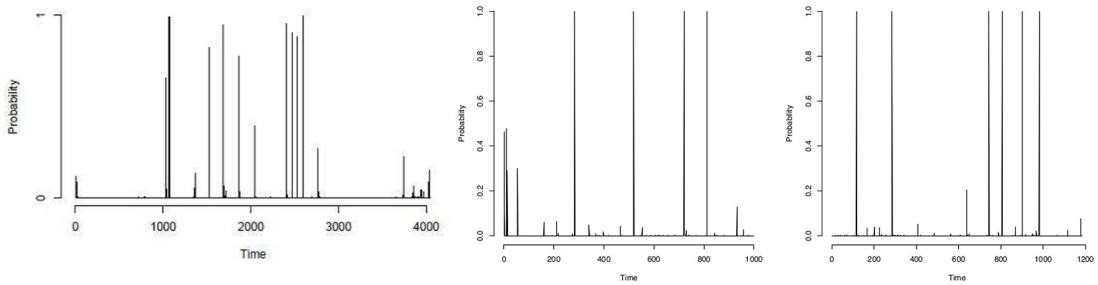


Fig 13. Comparison between the results of a random walk model [36] and our ARMA(1,1) model. (Left) change-point profiles of random walk model [36]; (centre) change-point profiles of ARMA(1,1) at time points 1 to 1000; (right) change-point profiles of ARMA(1,1) after time point 2800.

https://doi.org/10.1371/journal.pone.0208927.g013

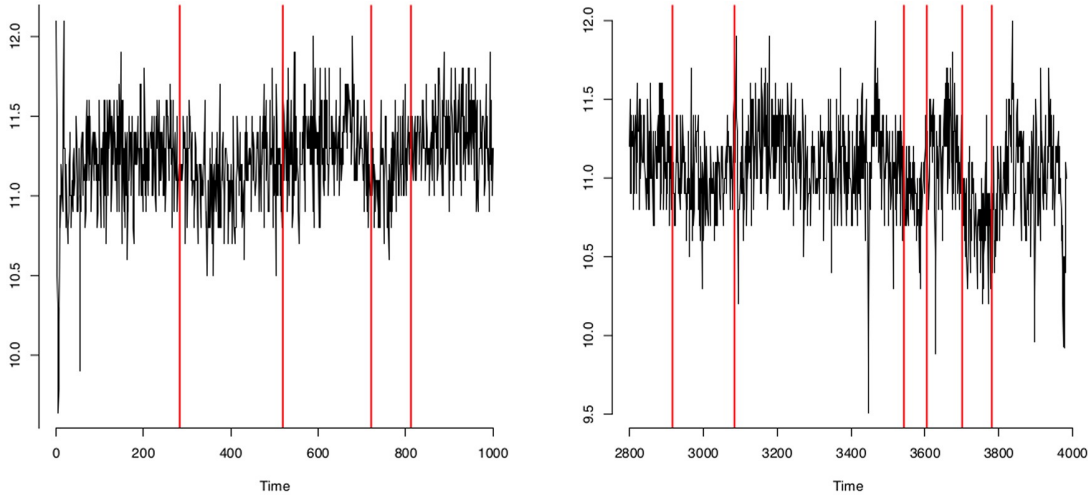


Fig 14. Change-point locations identified using ARMA(1,1) model displayed with the original well-log data at two different time points. (Left) change-point locations using ARMA(1,1) model in well-log data at time points 1 to 1000; (right) change-point locations using ARMA(1,1) model in well-log data after time point 2800.

https://doi.org/10.1371/journal.pone.0208927.g014

variance for all segments, it is not suitable for data sets in which different segments have different variance.

Supporting information

S1 Appendix. Details of posterior distribution.
(PDF)

S2 Appendix. Generalized gibbs sampling.
(PDF)

S3 Appendix. Supplementary material A.
(PDF)

S4 Appendix. Supplementary material B.
(PDF)

Acknowledgments

The authors are grateful to the Australian Research Council for their support of this project (DP1095849). **Data Source**: The data analysed in this paper is a third party data which was collected by the authors. The data can be obtained from https://rss.onlinelibrary.wiley.com/hub/journal/14679868/series-b-datasets/pre_2016a (Vol. 65-4.fearnhead).

Author Contributions

Conceptualization: Farhana Sadia, Sarah Boyd, Jonathan M. Keith.

Data curation: Farhana Sadia.

Investigation: Farhana Sadia.

Methodology: Farhana Sadia, Sarah Boyd, Jonathan M. Keith.

Project administration: Jonathan M. Keith.

Software: Farhana Sadia, Sarah Boyd, Jonathan M. Keith.

Supervision: Jonathan M. Keith.

Validation: Farhana Sadia.

Writing – original draft: Farhana Sadia.

Writing – review & editing: Farhana Sadia, Sarah Boyd, Jonathan M. Keith.

References

1. Whiteley N, Andrieu C, Doucet A. Bayesian computational methods for inference in multiple change-points models; 2011.
2. Carlin BP, Gelfand AE, Smith AF. Hierarchical Bayesian analysis of changepoint problems. Applied statistics. 1992; p. 389–405. https://doi.org/10.2307/2347570
3. Reeves J, Chen J, Wang XL, Lund R, Lu QQ. A review and comparison of changepoint detection techniques for climate data. Journal of Applied Meteorology and Climatology. 2007; 46(6):900–915. https://doi.org/10.1175/JAM2493.1
4. Jensen U, Lütkebohmert C. Change-point models. Encyclopedia of Statistics in Quality and Reliability. 2007.
5. Ko SI, Chong TT, Ghosh P, et al. Dirichlet Process Hidden Markov Multiple Change-point Model. Bayesian Analysis. 2015; 10(2):275–296. https://doi.org/10.1214/14-BA910
6. Chernoff H, Zacks S. Estimating the current mean of a normal distribution which is subjected to changes in time. The Annals of Mathematical Statistics. 1964; p. 999–1018. https://doi.org/10.1214/aoms/1177700517
7. Smith A. A Bayesian approach to inference about a change-point in a sequence of random variables. Biometrika. 1975; 62(2):407–416. https://doi.org/10.1093/biomet/62.2.407
8. Harlé F, Chatelain F, Gouy-Pailler C, Achard S. Bayesian Model for Multiple Change-points Detection in Multivariate Time Series. arXiv preprint arXiv:14073206. 2014;.
9. Green PJ. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika. 1995; 82(4):711–732. https://doi.org/10.1093/biomet/82.4.711
10. Chib S. Estimation and comparison of multiple change-point models. Journal of econometrics. 1998; 86(2):221–241. https://doi.org/10.1016/S0304-4076(97)00115-2
11. Keith JM. Segmenting eukaryotic genomes with the generalized Gibbs sampler. Journal of Computational Biology. 2006; 13(7):1369–1383. https://doi.org/10.1089/cmb.2006.13.1369 PMID: 17037964
12. Sofronov GY, Evans GE, Keith JM, Kroese DP. Identifying change-points in biological sequences via sequential importance sampling. Environmental Modeling & Assessment. 2009; 14(5):577–584. https://doi.org/10.1007/s10666-008-9160-8
13. Cribben I. Detecting Dependence Change Points in Multivariate Time Series with Applications in Neuroscience and Finance. Columbia University; 2012.

14. Punskaya E, Andrieu C, Doucet A, Fitzgerald WJ. Bayesian curve fitting using MCMC with applications to signal segmentation. Signal Processing, IEEE Transactions on. 2002; 50(3):747–758. https://doi.org/10.1109/78.984776
15. Chan KS, Tong H. On estimating thresholds in autoregressive models. Journal of time series analysis. 1986; 7(3):179–190. https://doi.org/10.1111/j.1467-9892.1986.tb00501.x
16. Tong H. On a threshold model. 29. Sijthoff & Noordhoff; 1978.
17. Tong H, Lim KS. Threshold autoregression, limit cycles and cyclical data. Journal of the Royal Statistical Society Series B (Methodological). 1980; p. 245–292. https://doi.org/10.1111/j.2517-6161.1980.tb01126.x
18. Tong H. Threshold models in non-linear time series analysis. vol. 21. Springer Science & Business Media; 2012.
19. Davis RA, Lee TCM, Rodriguez-Yam GA. Structural break estimation for nonstationary time series models. Journal of the American Statistical Association. 2006; 101(473):223–239. https://doi.org/10.1198/016214505000000745
20. Wood S, Rosen O, Kohn R. Bayesian mixtures of autoregressive models. Journal of Computational and Graphical Statistics. 2011; 20(1):174–195. https://doi.org/10.1198/jcgs.2010.09174
21. Chakar S, Lebarbier E, Lévy-Leduc C, Robin S, et al. A robust approach for estimating change-points in the mean of an AR(1) process. Bernoulli. 2017; 23(2):1408–1447. https://doi.org/10.3150/15-BEJ782
22. Albert JH, Chib S. Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts. Journal of Business & Economic Statistics. 1993; 11(1):1–15. https://doi.org/10.2307/1391303
23. Keith JM, Adams P, Stephen S, Mattick JS. Delineating slowly and rapidly evolving fractions of the Drosophila genome. Journal of Computational Biology. 2008; 15(4):407–430. https://doi.org/10.1089/cmb.2007.0173 PMID: 18435570
24. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equation of state calculations by fast computing machines. The journal of chemical physics. 1953; 21(6):1087–1092. https://doi.org/10.1063/1.1699114
25. Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. Biometrika. 1970; 57(1):97–109. https://doi.org/10.1093/biomet/57.1.97
26. Geman S, Geman D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on pattern analysis and machine intelligence. 1984;(6):721–741. https://doi.org/10.1109/TPAMI.1984.4767596 PMID: 22499653
27. Liu JS, Liang F, Wong WH. The multiple-try method and local optimization in Metropolis sampling. Journal of the American Statistical Association. 2000; 95(449):121–134. https://doi.org/10.1080/01621459.2000.10473908
28. Mira A, et al. On Metropolis-Hastings algorithms with delayed rejection. Metron. 2001; 59(3-4):231–241.
29. Keith JM, Kroese DP, Bryant D. A generalized Markov sampler. Methodology and Computing in Applied Probability. 2004; 6(1):29–53. https://doi.org/10.1023/B:MCAP.0000012414.14405.15
30. Oldmeadow C, Mengersen Kerrie, Mattick JS, Keith JM. Multiple evolutionary rate classes in animal genome evolution. Molecular biology and evolution. 2009; 27(4):942–953. https://doi.org/10.1093/molbev/msp299 PMID: 19955480
31. Keith JM, Spring D. Agent-based Bayesian approach to monitoring the progress of invasive species eradication programs. Proceedings of the National Academy of Sciences. 2013; 110(33):13428–13433. https://doi.org/10.1073/pnas.1216146110
32. Neal RM. Slice sampling. Annals of statistics. 2003; p. 705–741. https://doi.org/10.1214/aos/1056562461
33. Cook SR, Gelman A, Rubin DB. Validation of software for Bayesian models using posterior quantiles. Journal of Computational and Graphical Statistics. 2006; 15(3). https://doi.org/10.1198/106186006X136976
34. Ó Ruanaidh JJK, Fitzgerald WJ. Numerical bayesian methods applied to signal processing; 1996.
35. Fearnhead P, Clifford P. On-line inference for hidden Markov models via particle filters. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2003; 65(4):887–899. https://doi.org/10.1111/1467-9868.00421
36. Fearnhead P. Exact and efficient Bayesian inference for multiple changepoint problems. Statistics and computing. 2006; 16(2):203–213. https://doi.org/10.1007/s11222-006-8450-8

37. Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian data analysis. vol. 2. Chapman & Hall/CRC Boca Raton, FL, USA; 2014.

38. Sturtz S, Ligges U, Gelman A, et al. R2WinBUGS: a package for running WinBUGS from R. Journal of Statistical software. 2005; 12(3):1–16. https://doi.org/10.18637/jss.v012.i03

Chapter 4

Event detection in spatio-temporal data using a one-dimensional summary statistic

Chapter Objectives

In this chapter, I explore methods currently available for segmenting parallel sequences, as the thesis is aimed at developing improved methods for parallel segmentation. This chapter describes the second methodological development I introduced in this thesis, which is a simple one-dimensional approach to segment parallel time series. To achieve this, spatio-temporal data was pre-processed to produce a single summary sequence, which was then analysed using the change-point model developed in the previous chapter. The single summary sequence was produced using alternative statistics based on the maximum across sequences or principal component analysis (PCA). The goal of this work was to detect an event of interest in parallel sequences partially obscured by different levels of background noise rather than segmenting the two-dimensional data. This method was tested on synthetic data and real-world data and also compared with a simple event extraction method. Results for the synthetic data provide evidence of high accuracy for detecting events. For the real-world data, the proposed method was successful in detecting the boundary of the events with a low background noise. However, when the background noise level is high and the variance of background noise is not uniform, this method could identify the events of the data but also showed some false positive change-points. The maximum-based approach performed better than the PCA-based approach in our examples although maximum in-

volves loss of information. However, our results recommend that both PCA and maximum are suitable for initial exploratory analysis.

Authorship

Farhana Sadia[1], Sevvandi Kandanaarachchi[2], Kate Smith-Miles[3] Jonathan M. Keith[1]

1 School of Mathematics, Monash University, Clayton, VIC 3800, Australia

2 Department of Econometrics & Business Statistics, Monash University, Clayton, VIC 3800, Australia

3 School of Mathematics and Statistics, University of Melbourne, Parkville, VIC 3010, Australia

Reference

Sadia F, Kandanaarachchi S, Smith-Miles K, Keith JM. (2018). Event detection in spatio-temporal data using a one-dimensional summary statistic. Submitted.

4.1 Abstract

Event detection in spatio-temporal data has recently received increased attention, particularly in the study of anomaly detection. An interesting yet challenging problem is the detection of events in noisy backgrounds. In this paper, a Bayesian change-point segmented ARMA model is used to detect events in two-dimensional (time and one spatial dimension) data and we demonstrate how a technique designed for one-dimensional time-series can be used to segment parallel spatially correlated time-series. Our aim is to detect an event of interest in parallel sequences partially obscured by background noise, without explicitly segmenting the two-dimensional data. We reduce data dimensionality using alternative maximum and principal component analysis (PCA) based summary statistics before change-point detection. We test our model on synthetic data as well as real world data and find the model effective in detecting relatively simple events in backgrounds with various noise levels using an approach based on the maximum across spatial locations. The results of our proposed method are also compared with a simple event extraction method for both synthetic data and real world data. Although the maximum-based approach produces better results than the PCA-based approach in our examples, using the maximum as a low-dimensional summary statistic is not recommended, because of the information loss it involves. However, our results suggest PCA performs no better, and possibly worse, than the maximum, suggesting both are useful only in preliminary exploratory analysis.

Keywords: Event detection; Spatio-temporal data; Bayesian change-point segmented ARMA model; Dimension reduction; Principal component analysis.

4.2 Introduction

Change-point detection is the problem of identifying abrupt changes in the characteristics of a signal at unknown points of time in stochastic processes [1]. Change-point detection methods are used in a broad range of real world applications such as climate change detection [2], genetic time-series analysis [3], signal segmentation [4] and outlier detection [5].

Development of change detection techniques has recently received considerable attention in a variety of fields (for reviews, see [1, 2, 6, 7]). Anomalies and outliers occur at unknown time-points giving rise to changes in some of the statistical properties of the time series.

By correctly detecting these changes we can detect the anomalies. Therefore, the event and anomaly detection problem can be formulated as a change-point detection problem. The goal of change-point detection with particular emphasis on event detection is to detect an abrupt change, subject to false alarm constraints [8].

There is a wide variety of change-point detection methods in the literature used for event detection (for example, [9–12]). Tartakovsky *et al.* [8] proposed an adaptive sequential method and a batch-sequential method for early detection of attacks in computer network traffic. This method considers the fluctuations of the false alarm rate and identifies changes in the statistical model. To detect anomalies in one-dimensional time-series data, Burnaev and Ishimtsev [13] used a non-parametric approach involving a probabilistic interpretation of an anomaly score. To detect an anomaly in a single time series and across multiple time series, Qiao *et al.* [14] proposed a novel anomaly detection algorithm for multiple heterogeneous correlated time series data. They introduced a new clustering-based compression method to deal with large time series data.

A challenging task is to detect change-points in multidimensional data because increasing dimensionality makes change-point estimation computationally expensive. Some existing approaches use dimensionality reduction techniques to project the multidimensional data into a lower dimensional space. Qahtan *et al.* [15] projected the multidimensional data on selected principal components using PCA and created multiple univariate data streams. They used an unsupervised change detection procedure where they measured the variation in each univariate data stream by a change-score. They determined the final change-score by aggregating selected components. Other studies also used PCA [16–23] to reduce the dimension of the data in change-point problems. PCA-based change-detection approaches are effective to reduce the number of features in the data but they produce a large number of false alarms [24]. Moreover, these approaches may fail to detect complex changes over correlations of dimensions as it assumes statistical independence in the transformed space [23]. Besides PCA, some other dimensionality reduction techniques are used in change-point problems. Lévy-Leduc e*t al.* [25] proposed a data reduction technique named record filtering to detect change-points in high-dimensional network traffic data. Record filtering selects the heavy-hitters with high probability from network traffic data, which are collected at several points of an internet network (heavy-hitters are the entities which account for a pre-determined proportion of traffic data from total unusual traffic activity and are measured in terms of number of packets, bytes, or connections [26]). Some authors used an-

other data reduction technique called random aggregation (or sketch) in network anomaly detection problems [26–28]. This technique chooses several linear combinations at random from all the flows and processes them before applying a change-point detection approach. A low-dimensional summary statistic, known as least absolute shrinkage and selection operator (LASSO) estimator, is also used in high dimensional change-point problems in some studies [29–32]. Zou and Qiu [32] used shifted mean components in a multivariate statistical process control (SPC) problem using LASSO estimators. They proposed a LASSO-based multivariate test statistic which uses both variable selection and regularization to improve the interpretability and prediction accuracy of the regression model they use.

In this paper, we reduce the dimension of the data using maximum and PCA based summary statistics before applying a one-dimensional change-point detection algorithm. Our Bayesian change-point segmented ARMA model is used to identify events of interest. This model considers the problem of segmenting a time series where each segment of the time series follows an ARMA model with distinct segment means. It assumes an unknown number of change-points and locations and the main aim of this model is to infer the most probable segment locations and model parameters.

We test our proposed method using two synthetic two-dimensional data sets with different shaped hidden events in the presence of noise, and some real world data from a sensitive security application where we have de-identified the data. We successfully identify the events of interest in synthetic data sets. For the real data set, the maximum-based approach identifies the boundary of the events whereas a PCA-based approach provides a cluster of change-points covering the event in the presence of low background noise. When high background noise exists, the maximum-based approach detects the events of interest and some false positive events with low probability whereas the PCA-based approach fails to identify one real event and also provides false positive events with high probability. These indicate that in some cases, PCA provides no advantage over the maximum.

4.3 Methods

In earlier work, we developed a hierarchical Bayesian model for identifying change-points in time series data [33]. The method models a time series by segmenting the series into blocks of autoregressive moving average (ARMA) processes. Our Bayesian change-point segmented ARMA model was originally designed for one-dimensional time-series data. In this paper,

we consider a collection of parallel time-series collected at roughly evenly spaced points along a line or curve, with correlations between time series that diminish with separation distance along the line. Here we treat the data as a time-series of vectors, where the dimension of each vector is the number of spatial locations at which data are collected. As all our data sets (synthetic and real) are two-dimensional, it is necessary to pre-process the data using a summary statistic before applying the ARMA model. We have used maximum as the summary statistic. An alternative summary statistic, based on principal component analysis (PCA), is alternatively used to reduce the dimension of the data in our real life examples. As a summary statistic, the maximum gives better results than PCA in our examples.

4.3.1 Proposed method for event detection in spatio-temporal data

We consider the problem of detecting events in a spatio-temporal dataset with N parallel time-series collected at distinct spatial locations (with time series corresponding to rows of a data matrix) and M time-points (arranged in columns). We denote by $\mathbf{X_n} = (X_{n1}, \ldots, X_{nM}), n = 1, \ldots, N$ the data samples, where X_{ij} is the observation in row i and column j. We summarise the data to form a single time-series by first computing the maximum values over all locations for each time-point (that is, we find the maximum in each column). Then Bayesian segmented MA(1), AR(1) and ARMA(1,1) change-point models [33] are applied to the maximum values.

The main idea behind the Bayesian change-point segmented ARMA model is to assume a MA(1) or AR(1) or ARMA(1,1) model with a different mean in each segment of the time-series. The aim is to infer the most possible segment locations and model parameters that describe each segment. The model also assumes an unknown number of change-points and unknown locations of those change-points. The Bayesian change-point segmented ARMA model consists of 4 main steps: (1) write down the likelihood function of a model in which the series within each segment is generated by an ARMA process; (2) compute the conditional posterior distribution of the unobserved parameters of interest, given the observed data; (3) generate samples from the conditional posterior distribution of every parameter using a Markov chain Monte Carlo technique known as the Generalized Gibbs Sampler, or GGS (4) estimate and evaluate the model (Further details of this model are found in Appendix B.1 and in the paper [33]).

The best model from these three models (MA(1), AR(1), ARMA(1,1)) is chosen based on the deviance information criterion (DICV) approximation values. The DICV is defined as: DICV= $p_v + \overline{D(\Theta)}$, where $\overline{D(\Theta)}$ is the mean posterior deviance, $p_v = \text{Var}(D(\Theta))/2$ and deviance $D(\Theta) = -2\text{lnf}(y|\Theta)$ (See [34, 35]). In our model, $f(y|\Theta)$ is the probability density of the observed signal, which is a product over all segments. Here, Θ is the set of parameters which are $K, \mathbf{s}, \boldsymbol{\theta}, \boldsymbol{\psi}, \mathbf{c}, \sigma^2$. This probability is normally distributed with mean $c_k + \epsilon_t + \sum_{i=1}^{a} \psi_i(x_{t-i} - c_k) + \sum_{i=1}^{m} \theta_i \epsilon_{t-i}$ and variance σ^2. Here, K is the total number of segments; $\mathbf{s}$ is a vector of the starting positions of the segments; $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are the parameters of the MA and AR model; $\mathbf{c}$ is the mean of the signal. All the parameters are defined in Appendix B.1. We obtain the posterior distribution of occurrence of change-point locations from the selected best model. The posterior probability of being a change-point at each position is $\mathbf{t_r} = [t_{r1}, \ldots, t_{rN}]'$, which is a Monte Carlo estimator (see Equation B.2 in Appendix B.1).

In our segmented model, each segment is assigned to a group. We denote the group to which segment k is assigned by $g_k \in \{1, \cdots, J\}$, where $\mathbf{g} = (g_1, \cdots, g_K)$. For each position in the input sequence, we also estimate the probability that the position belongs to a given group of the selected best model (see Equation B.3 in Appendix B.1). The result is a profile for that group that can be plotted across the time points or locations.

It is important to note that we take the logarithm of the maximum values as the input sequence for the real data sets. These data sets mainly show instability in variance as they do not have the same variance across the signal. The variance is large where the signal is large and small where it is small. The variance of a time series can be stabilized using transformations such as logarithms whereas the mean of a time series can be stabilized using differencing (like ARIMA model) by removing changes in the level of a time series [36]. The logarithm of the maximum values has variance that appears to be approximately constant. This means applying logarithm doesn't indicate ARIMA model as it only helps to stabilize the variance, not the mean as the differencing technique does in an ARIMA model.

If the N spatial locations are arranged sequentially, we can use the same technique to distinguish the locations involved in an event from those that are not. The idea is to swap the roles of rows and columns. To identify the Monte Carlo estimate of the posterior probability of being a change-point at column locations, that is, $\mathbf{t_c} = [t_{c1}, \ldots, t_{cM}]'$, we repeat the above described method on a sequence formed by finding the maximum of each

row.

We summarize the overall algorithm as follows,

input : Dataset with N rows and M columns
output: Probabilities of row change-point locations $\mathbf{t_r} = [t_{r1}, \ldots, t_{rN}]'$, probabilities of row group profile $\mathbf{p_r} = [p_{r1}, \ldots, p_N]'$, probabilities of column change point locations $\mathbf{t_c} = [t_{c1}, \ldots, t_{cM}]'$ and probabilities of column group profiles $\mathbf{p_c} = [p_{c1}, \ldots, p_{cM}]'$

1 Take maximum values of columns;
2 Fit Bayesian segmented MA(1), AR(1) and ARMA(1,1) change-point model on the logarithm of the maximum values;
3 Calculate posterior_Prob_Of_CP_MA, posterior_Prob_Of_CP_AR and posterior_Prob_Of_CP_ARMA;
4 Calculate posterior_Prob_Of_Group_profile_MA, posterior_Prob_Of_Group_profile_AR and posterior_Prob_Of_Group_profile_ARMA;
5 $Best\_model$ = Model with smallest DICV, from models $MA(1)$, $AR(1)$ and $ARMA(1,1)$;
6 $\mathbf{t_r}$=$Posterior\_Prob\_Of\_Change\_Points = Posterior\_Prob\_Of\_CP\_Best\_model$ where $Best\_model \in (MA, AR, ARMA)$;
7 $\mathbf{p_r}$ =$Posterior\_Prob\_Of\_Group\_Profiles = Posterior\_Prob\_Of\_Group\_profile\_Best\_model$ where $Best\_model \in (MA, AR, ARMA)$;
8 Swap rows and columns in the original data and repeat from line 1 to line 6 to get $\mathbf{t_c}$ and line 7 to get $\mathbf{p_c}$.

Algorithm 1: Event detection with Bayesian change-point segmented ARMA model

4.3.2 A comparison method

For comparison we use a simple event extraction method detailed in Kandanaarachchi *et al.* [37], which focuses on early event classification. This algorithm considers points with high signal values, and clusters them using the DBSCAN algorithm [38].

input : a 2 dimensional array $X_{n\times m}$, and parameters α, ϵ and *minPts*.
output : events and event ids

1 Let x_{ij} be the signal value at (i, j) position of X.
2 Let q denote the α-percentile of the signal values of X.
3 $S = \{(i, j) \mid x_{ij} > q\}$. S gives locations of X, which have signal values greater than the α^{th} percentile.
4 Let $X(S)$ be signal values of X in S locations.
5 Using DBSCAN [38] cluster $X(S)$ using ϵ and *minPts*.
6 This clustering gives each $x \in X(S)$ a cluster id.
7 Consider each cluster as an event.

Algorithm 2: *Extract 2-dimensional events from a 2D array using DBSCAN.*

We use the default settings of Algorithm 2 with $\alpha = 0.95$, $\epsilon = 5$ and $minPts = 10$ in our analysis.

4.4 Results

4.4.1 Synthetic data

To evaluate the performance of the proposed method for event detection, we applied it to some synthetic data sets with background noise. The first data set consists of an array of Gaussian noise superimposed on a triangle event and the second one is an example of an odd quadrangle event.

Test 1

The diagram (Figure 4.1) below shows a 300×200 array filled with Gaussian noise with mean 0 and standard deviation 1. A triangle shaped event is hidden among the noise with vertices at positions (100,50), (140,60) and (150,100) where the first coordinate represents spatial location and the second is time. The mean value of the signal in the triangle is 2 while the background has a mean of 0 and the variance in the triangle is same as the background. The change-points on the x axis take place at 50, 60 and 100 while the change-points on the y axis take place at 100, 140 and 150.

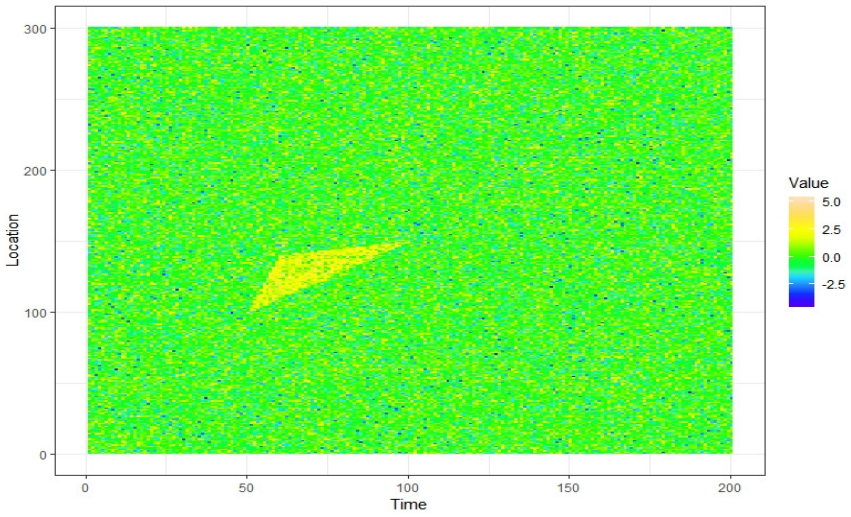


Figure 4.1: Triangle event hidden in the Gaussian background noise

Proposed method

We applied Algorithm 1 to our synthetic data set and fitted Bayesian segmented MA(1), AR(1) and ARMA(1,1) change-point models to the maximum values across row and column locations respectively. The MA(1) model at column locations was chosen as the best model among the three models according to DICV values (DICV of MA(1)= 327.165, DICV of AR(1)= 465.137 and DICV of ARMA(1,1)=654.331). At row locations, we chose AR(1) as the best model (DICV of AR(1)= 265.414, DICV of MA(1)= 317.221 and DICV of ARMA(1,1)= 520.631). The row and column change-point profile plots of the selected best models are given in Figure 4.2a and Figure 4.2c. The change-point profile plot gives the posterior probability of each position being a change point. As seen from the graph, we detect x coordinates 50 and 100 and y coordinates 100 and 150 as change points. These positions correspond to the boundary locations of the triangle event. The change-point profile plots also show two other change-points with low probability at position (126,71) within the boundaries of the event. The coordinate (71,126) represents the actual coordinate (140,60) which is the top corner of this event. To get a clear picture of the boundary location of the event, we also look at the group profile plots given in Figure 4.2b and Figure 4.2d. These plots exhibits the correct boundary locations. This validates the claim that our proposed method can identify the boundary location of events buried in background noise.

Comparison method

To compare with the proposed method we extract events using Algorithm 2. Figure 4.3 shows the events that are extracted, with each event depicted in a different colour. In addition to the triangle we see three small noisy events extracted by Algorithm 2, which were correctly not flagged by Algorithm 1. This is a strength of Algorithm 1 compared to Algorithm 2. However, Algorithm 2 extracts events in a 2-dimensional setting, without considering each dimension separately, which is an advantage of Algorithm 2.

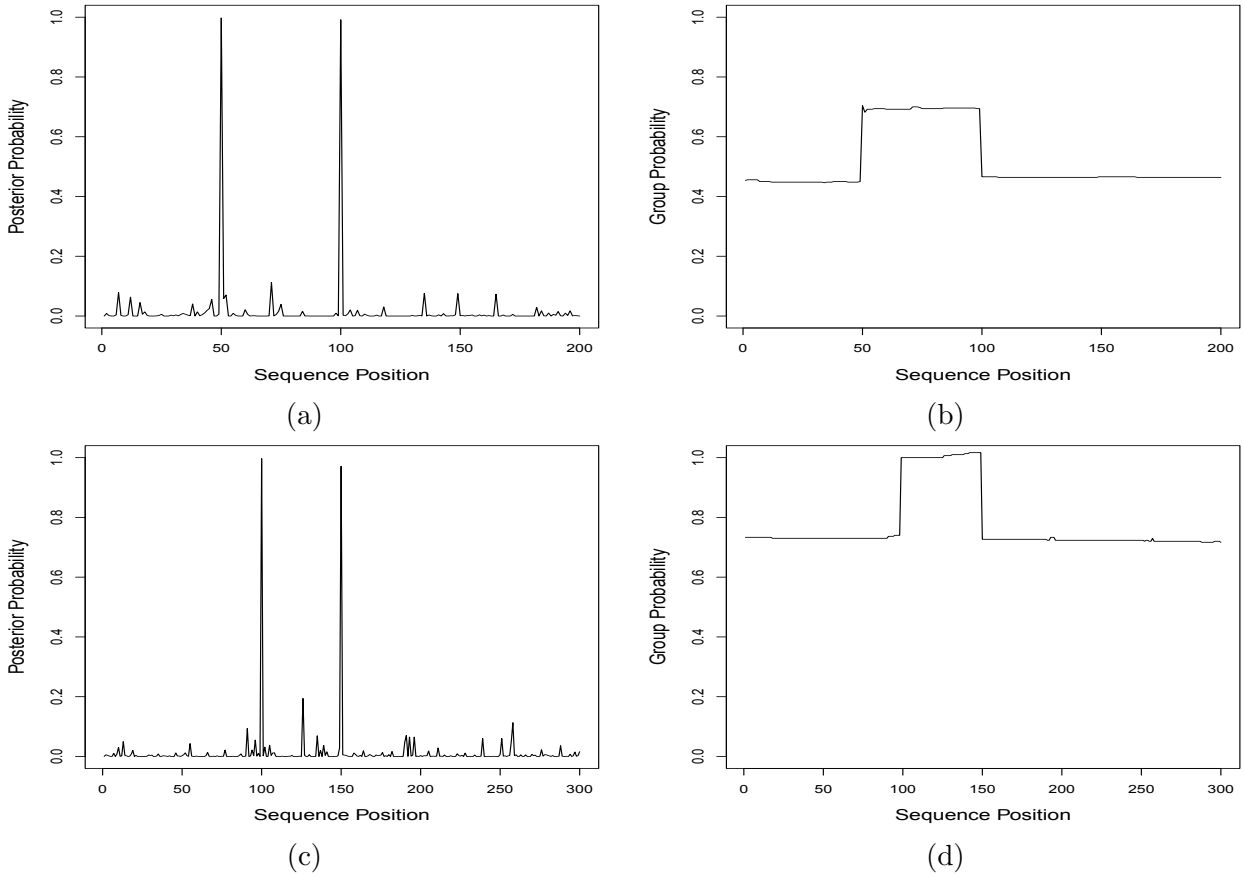


Figure 4.2: Profile plots: (a) Change-point profile plot across time points, (b) Group profile plot across time points, (c) Change-point profile plot across locations, and (d) Group profile plot across locations

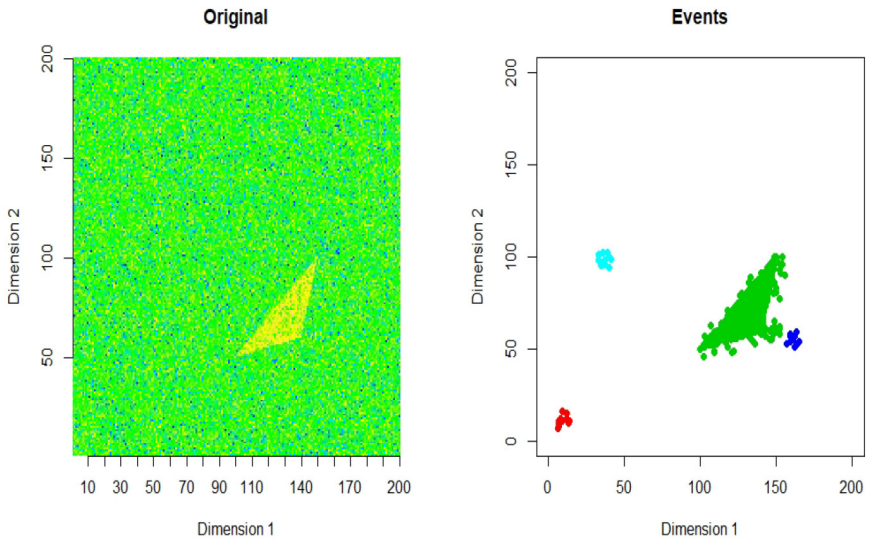


Figure 4.3: Original data and events extracted using Algorithm 2, with each event depicted in a different colour.

Test 2

We also generated 300×200 array of Gaussian noise with mean 0 and standard deviation 1. In this example, a quadrangle is hidden among the noise (Figure 4.4) with a different mean (of 2) from the background and the variance is the same as for the background. The change-points on the x axis occur at 50, 60, 80, 100 while the change-points on the y axis take place at 100, 115, 140 and 150.

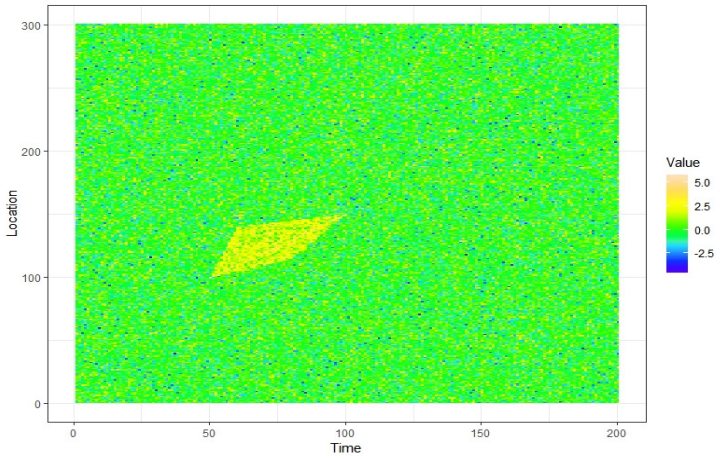


Figure 4.4: Odd quadrangle event hidden in the Gaussian background noise

Proposed method

After applying Algorithm 1 to this synthetic example, we chose the best models as in the previous example. Figure 4.5 plots the row and column change-point profiles (Figure 4.5a, Figure 4.5c) and group profiles of the best models at row and column locations (Figure 4.5b, Figure 4.5d). All the plots accurately detect x coordinates 50,100 and y coordinates 100, 150 as row and column change points. These locations represent the boundary locations of the hidden events.

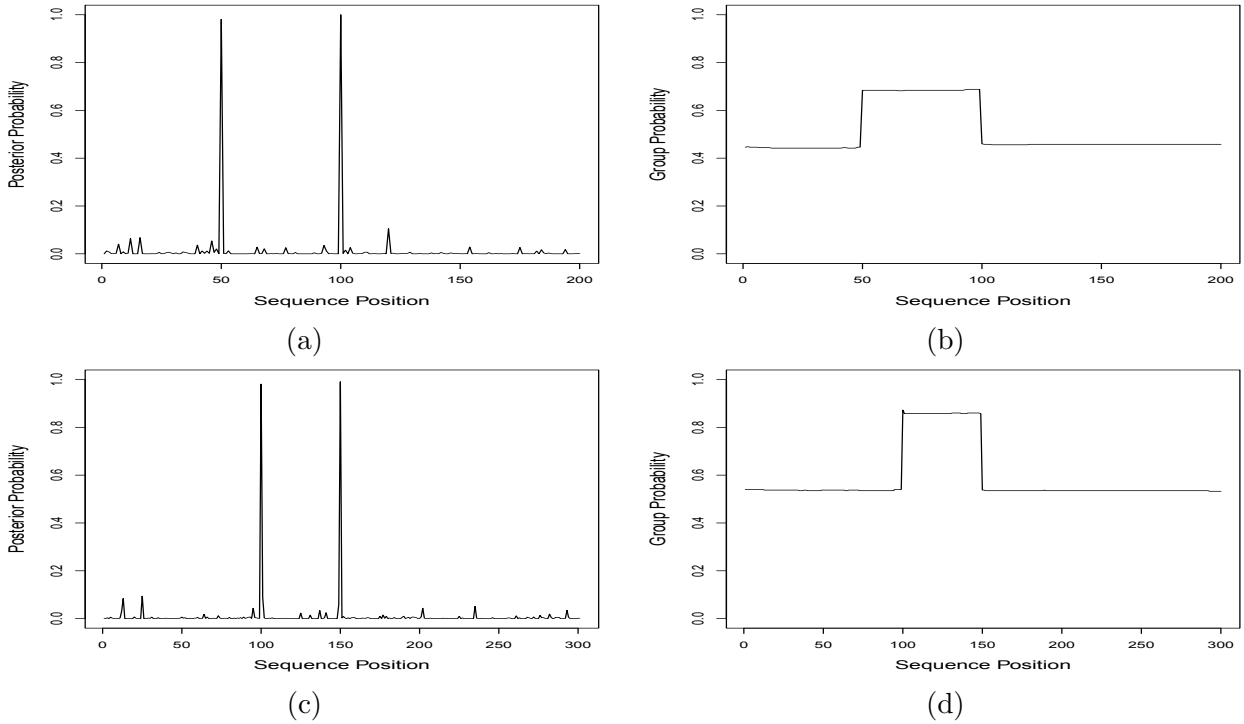


Figure 4.5: Profile plots: (a) Change-point profile plot across time points, (b) Group profile plot across time points, (c) Change-point profile plot across locations, and (d) Group profile plot across locations

Comparison method

Figure 4.6 shows the extracted events from the original dataset with each colour denoting a separate event. Again, we see two additional events in colours blue and green just outside the extracted irregular quadrilateral, which is depicted in red. While both these additional events are small, the blue noisy event is quite close to the red event, signifying a possibility that it can be merged with the red event.

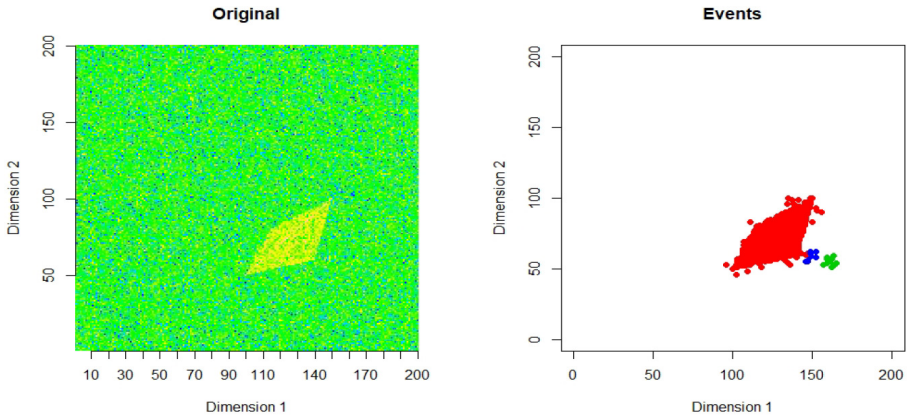


Figure 4.6: Original data and events extracted using Algorithm 2, with each extracted event depicted in a different colour.

4.4.2 Real World Application

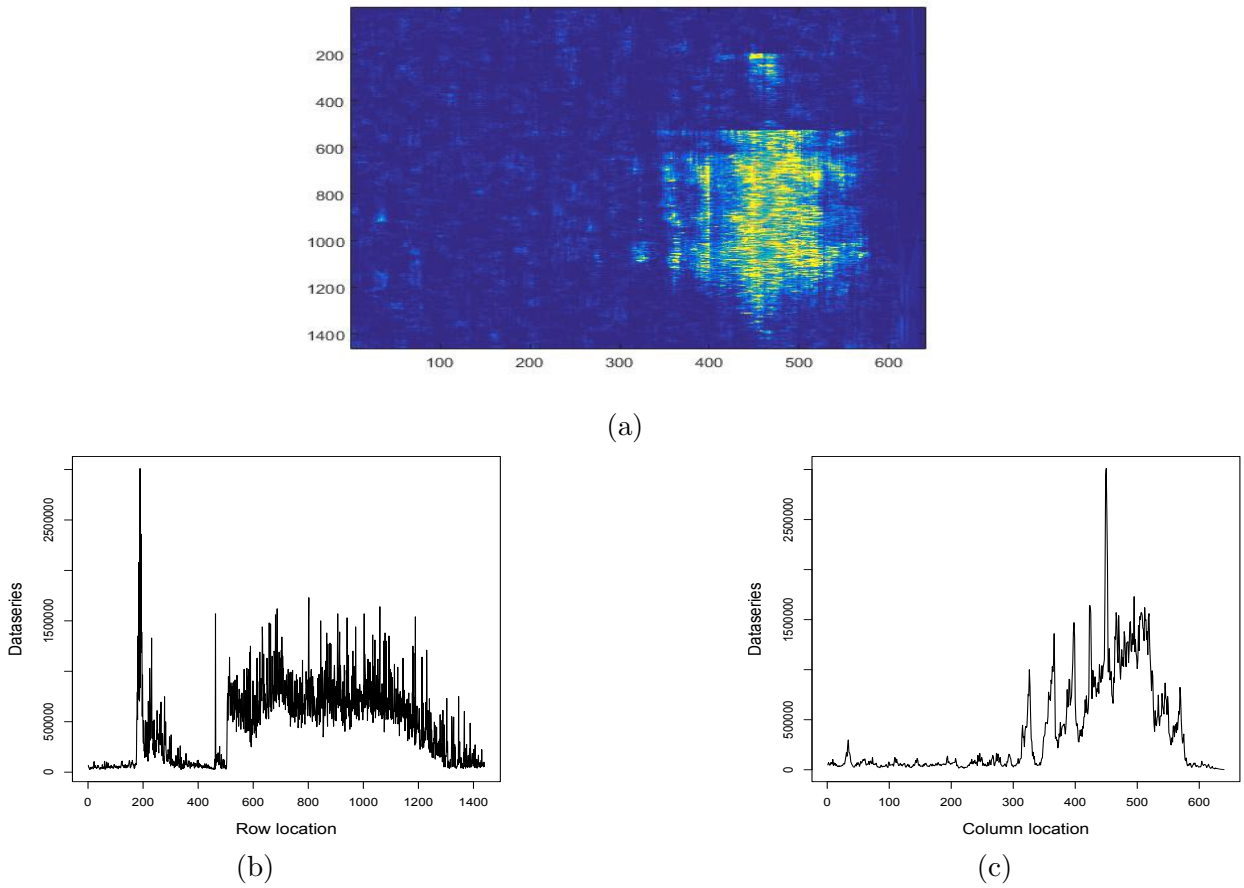
We have used three data sets from a real application that uses fibre-optic cables. Due to the sensitive nature of the industry, we have de-identified this data. All three data sets contain anomalies. The background noise level of the first two data-sets are relatively low, while the third data set contains a high level of background noise. We apply Algorithm 1 to these data sets.

Dataset 1

The heat map in Figure 4.7a represents an event of interest depicted in yellow surrounded by relatively low background noise depicted in blue. Figure 4.7b and Figure 4.7c show the time series plots of maximum values across row and column locations respectively which seems non-linear. Figure 4.7b shows change in means and variance approximately in rows 200-300 and 500-1300. Figure 4.7c shows both change in means and volatilities with an increasing trend between columns 350-600. To check the non-linearity of those time series, Keenan's 1-degree test [39] was used against the null hypothesis that the time series follows some AR process. This test has p-value less that a significance level of 0.05 for both time series, therefore we can reject the null hypothesis and conclude that the time series plotted in Figure 4.7b and Figure 4.7c are nonlinear.

Proposed method

As the data presents in Figure 4.7b and Figure 4.7c show unstable volatility, the logarithm of the maximum values has been used as the input sequence. The non-linearity of logarithm of the maximum values has been checked again using Keenan's 1-degree test. Both time series have shown linearity according to their p-values which can be used now for segmented ARMA model. The row and column change point profiles of the selected best models among the segmented MA(1), AR(1) and ARMA(1,1) models (obtained using the same procedure as in the synthetic examples) are given in Figure 4.8a and Figure 4.8c. By examining the change point profile plots in Figure 4.8 we notice that certain row and column locations have much larger posterior probability than others. There are four spikes in row change-point locations, they are at 140, 338, 502 and 1247. Similarly, the column change point profile plot has spikes at locations 315 and 589. We note that these change-point locations correspond to the rows and columns where the event of interest lies in the heat map.



(a)

(b)

(c)

Figure 4.7: (a) Heat map of an event of interest surrounded by relatively low background noise, (b) Time series plot of maximum values across row locations, and (c) Time series plot of maximum values across column locations

The group profiles plots in Figure 4.8b and Figure 4.8d draw a clear picture of the boundary of the events. Figure 4.8b shows the boundary of the two events at time points 140-338 and at 502-1247 which correspond to the actual two events seen in the heat map across time points. Figure 4.8d also displays the boundary of an event at locations 315-589 that accurately corresponds to the event visible in the heat map across location. The results show that the proposed method finds the boundaries of the events of interest for this real world data set.

Comparison method

Figure 4.9 shows the extracted events from this real-world dataset. The original dataset was converted to the log scale for better visualization purposes. We have repeated the experiment in the original scale and obtained the same events. This is not surprising as Algorithm 2 works with percentiles. We obtain the desired events in addition to one small

noise event, which is depicted in red.

Again, Algorithm 1 does not flag this noise event with a high posterior probability, demonstrating its robustness to noise. However, Algorithm 2 shows the two-dimensional shape of the event, which is advantageous for certain applications.

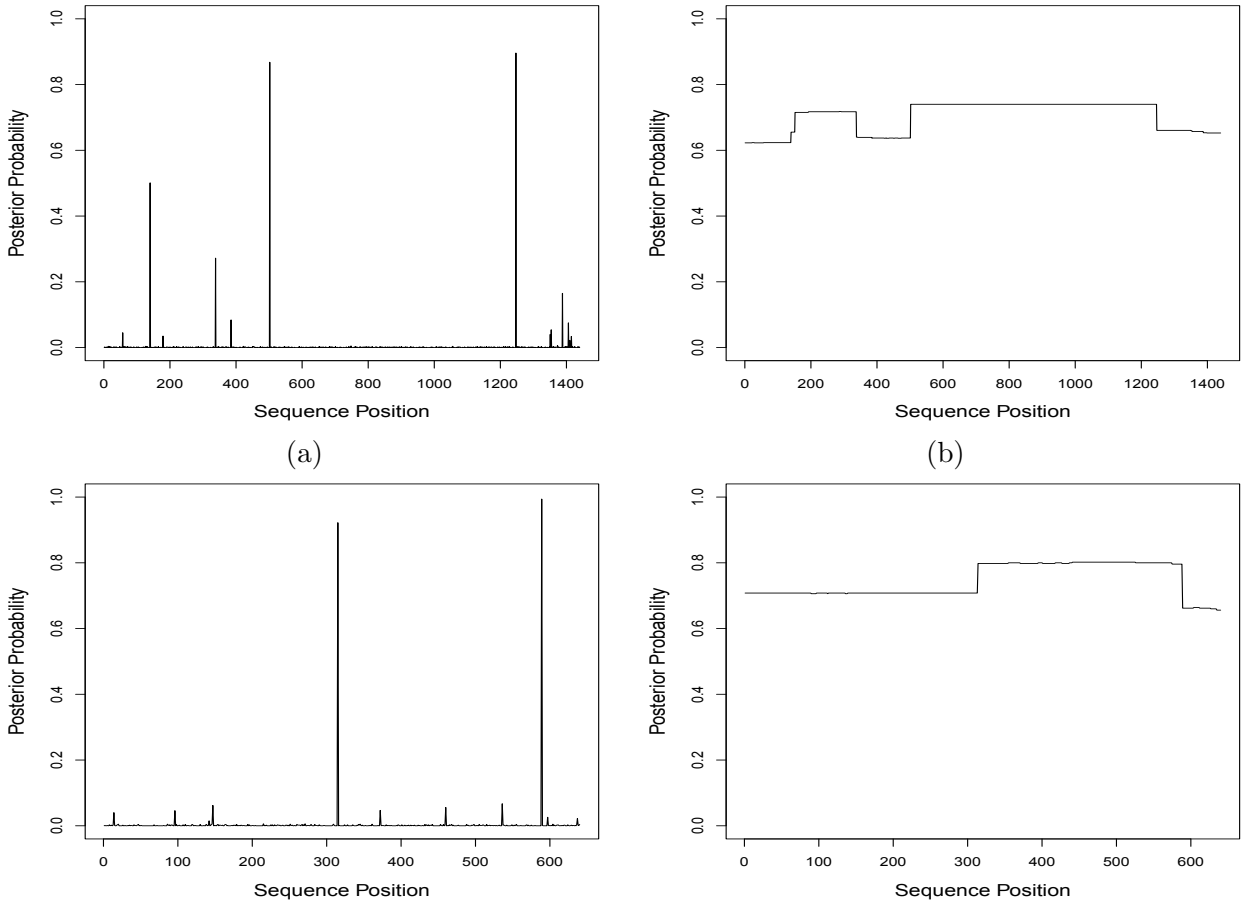


Figure 4.8: Profile plots: (a) Change-point profile plot across row locations, (b) Group profile plot across row locations, (c) Change-point profile plot across column locations, and (d) Group profile plot across column locations

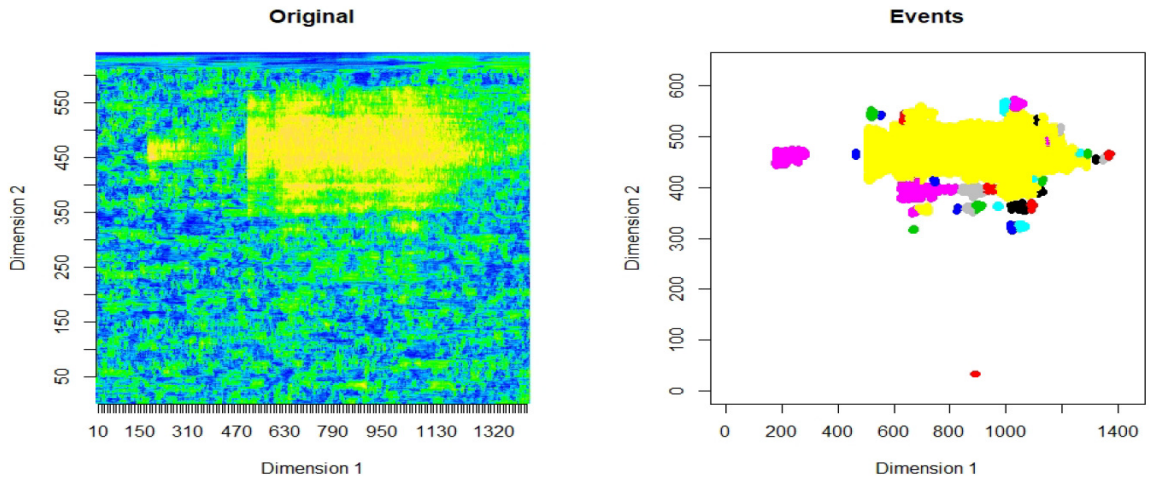


Figure 4.9: Original data and events extracted using Algorithm 2, with each extracted event depicted in a different colour.

PCA-based approach

Our proposed method for event detection can also be applied using PCA as a summary statistic. In this method, first we perform PCA on the data samples in rows and columns respectively and we select the number of principal components (PCs) that corresponds to the "elbow" point of the PCA scree plot. Then the Bayesian segmented MA(1), AR(1) and ARMA(1,1) change-point models are applied to each principal component. A summary of the algorithm using PCA is given in Appendix B.2 and Figure 4.10 presents the change-point profile plots across row and column locations.

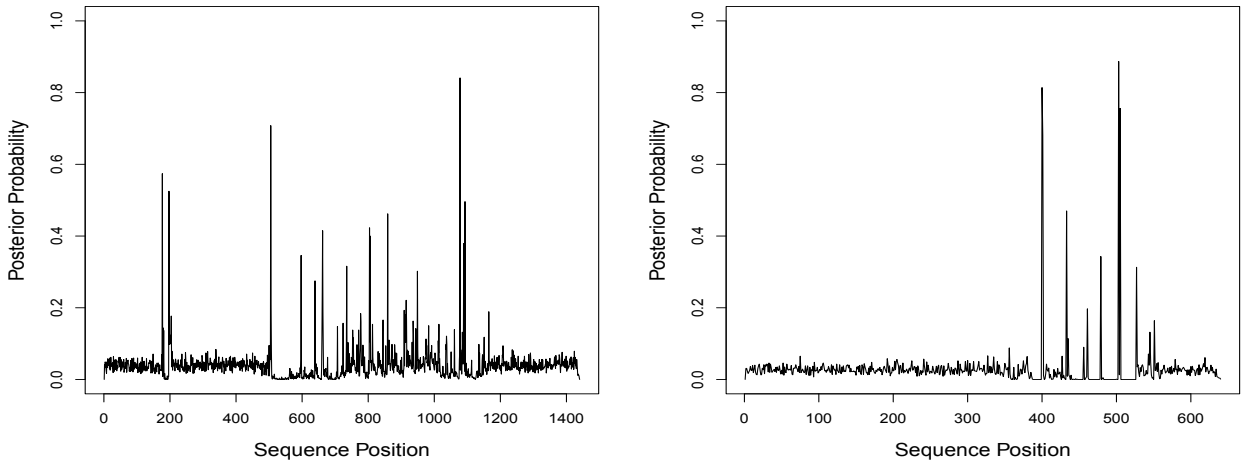


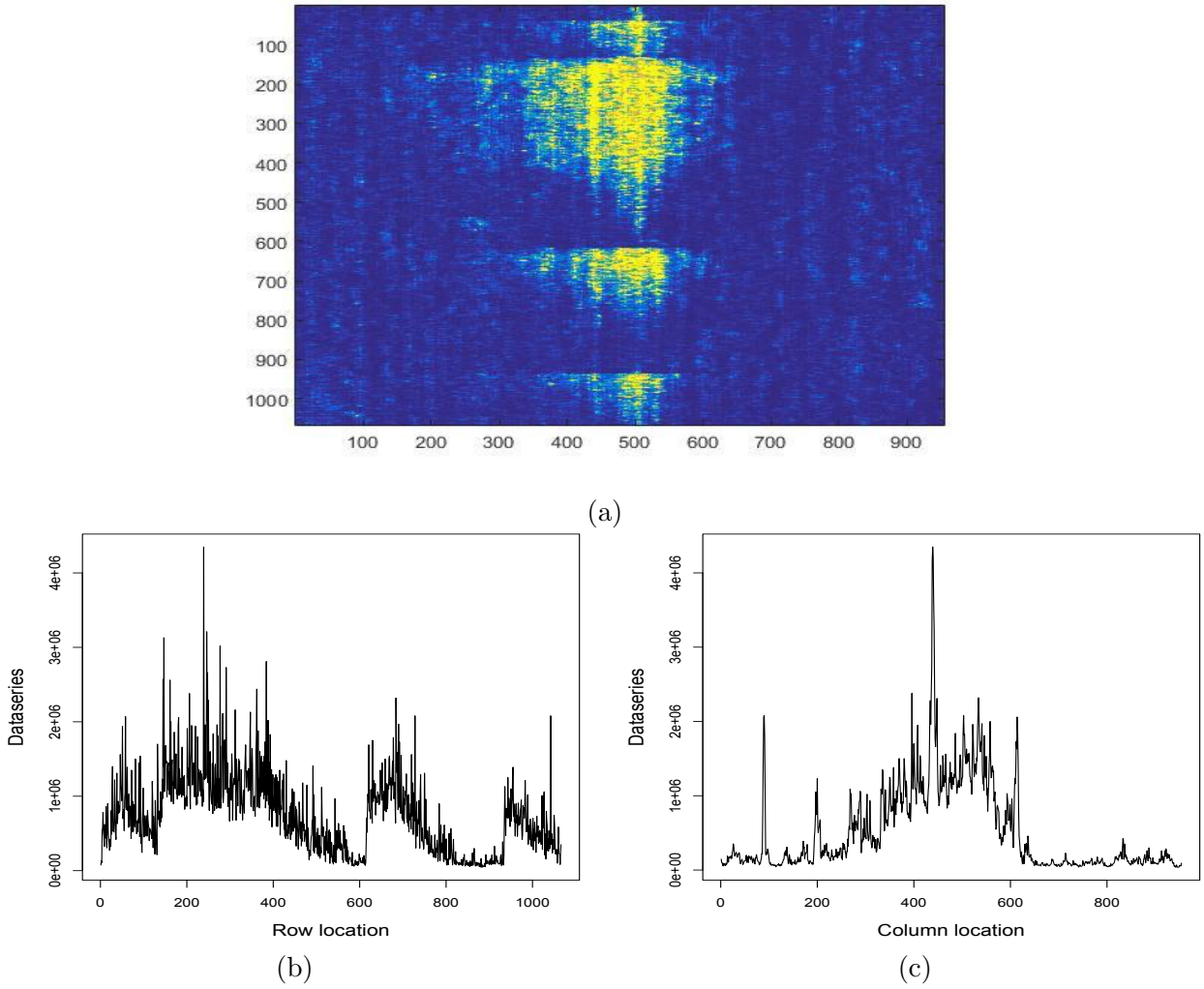
Figure 4.10: (a) Change-point profile plot across row locations, and (b) Change-point profile plot across column locations using PCA based approach

There are two sets of spikes in row change-point locations (Figure 4.10a), one around

200 and the other from 500 to 1100. Similarly, the column change point profile plot (Figure 4.10b) has spikes around locations 400 - 550. These change-point locations correspond to the rows and columns where the event of interest lies in the heat map (Figure 5.7a). The results show that this method finds a cluster of change-points covering the event for this real world data set.

Dataset 2

Figure 4.11a, 4.11b and 4.11c show the heat map, the time series plots of maximum values across row and column locations for the second data respectively. Again, the event of interest is depicted in yellow in a background of relatively low noise. The event seems to occur in rows 40 - 450, 620 - 750 and 930 - 1050 and columns 250 - 650.



(a)

(b)

(c)

Figure 4.11: (a) Heat map of an event of interest surrounded by relatively low background noise, (b) Time series plot of maximum values across row locations, and (c) Time series plot of maximum values across column locations

Proposed method

The row and column change point profiles and group profiles are given in Figure 4.12. We see two spikes of change-points appearing around 50 - 470 in the row change point profile plot (Figure 4.12a), which corresponds to the first part of the event (rows 40 - 450 on the heat map in Figure 4.11a) and some high posterior probability values around 615-794 and 936-1-38 which correspond to the next two contiguous yellow blocks. Similarly, the column change point profile plot (Figure 4.12c) gives two change-points with high probability at locations 286 and 631 which correspond to the columns of the event. The row group profiles plot (Figure 4.12b) clearly exhibits an event with three parts at those locations where the event lies in the heat map. The column group profiles plot (Figure 4.12d) displays the boundary of the event from 286-631. These are consistent with the apparent change-point locations observed in the heat map for this data set.

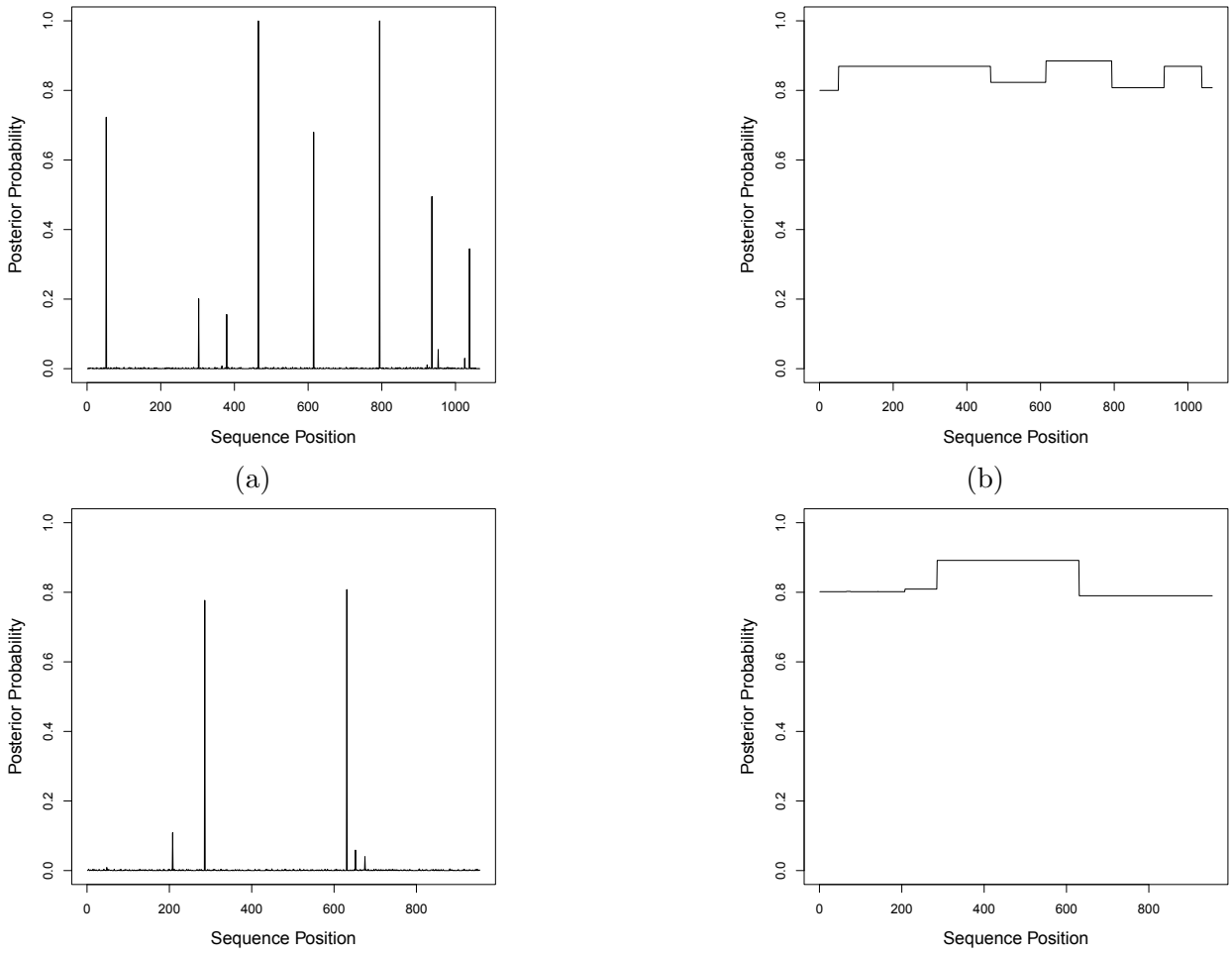


Figure 4.12: Profile plots: (a) Change-point profile plot across row locations, (b) Group profile plot across row locations, (c) Change-point profile plot across column locations, and (d) Group profile plot across column locations

Comparison method

Figure 4.13 shows the events extracted from this dataset using Algorithm 2. In this instance there is one small noise event depicted in light blue approximately at coordinates $(600, 300)$. All other events apart from this event, are real.

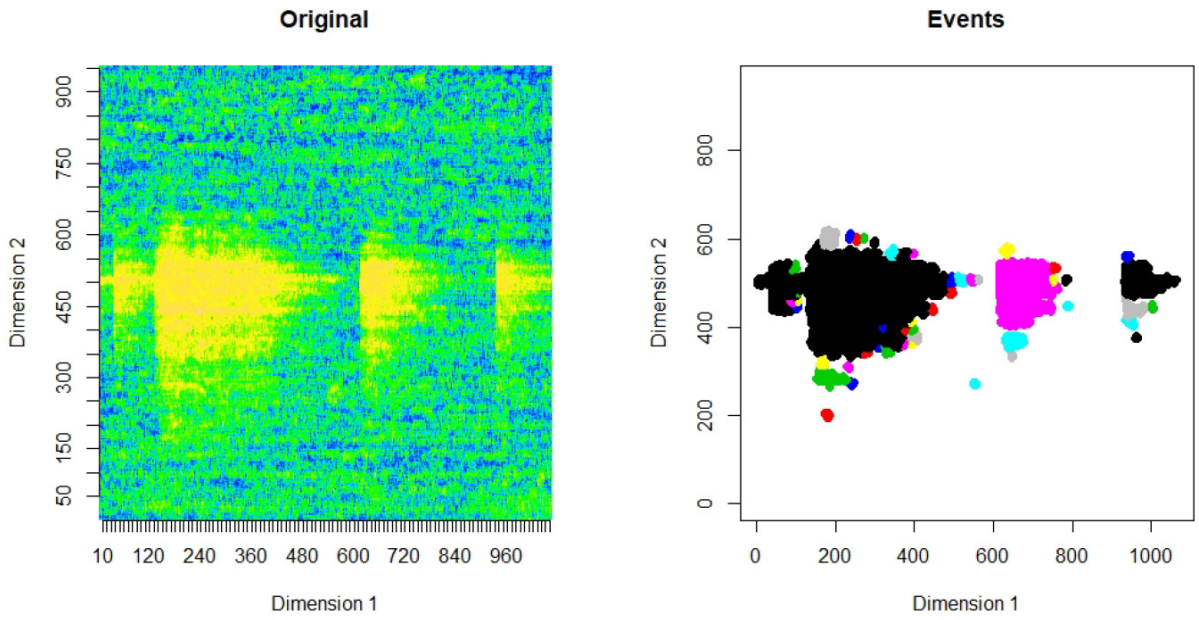
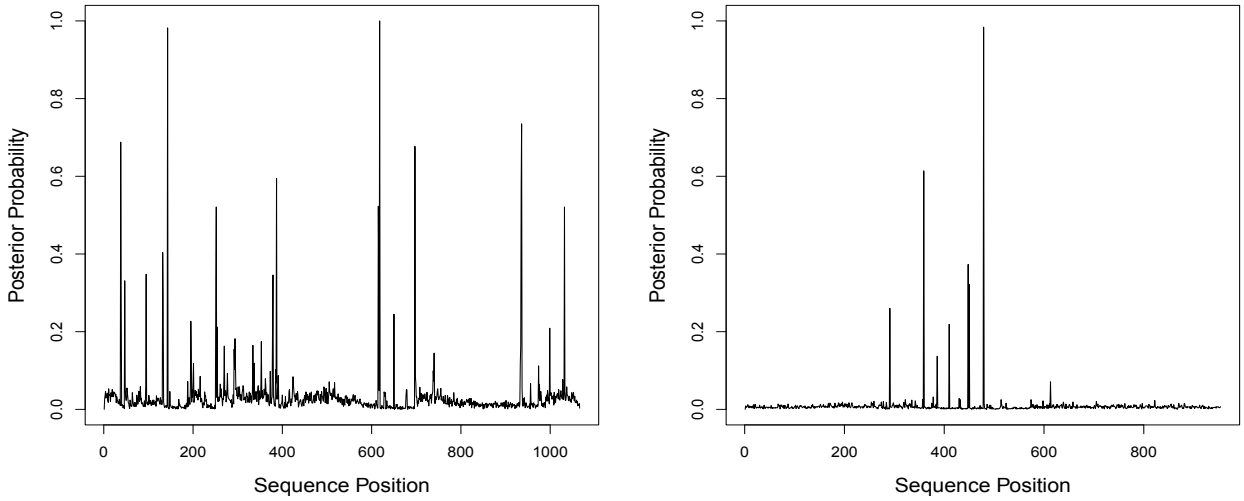


Figure 4.13: Original data and events extracted using Algorithm 2, with each extracted event depicted in a different colour.

PCA-based approach

A cluster of change-points is appearing from 40 - 400 in the row change point profile plot (Figure 4.14a), which corresponds to the first part of the event (rows 40 - 400 on the heat map in Figure 4.11a) and some high posterior probability values around 620 and 930 correspond to the remaining events in the heat map. The column change point profile plot (Figure 4.14b) gives a cluster of change-points from 350- 500, corresponding to the columns of the event.



(a) (b)

Figure 4.14: (a) Change-point profile plot across row locations, and (b) Change-point profile plot across column locations using PCA based approach

Dataset 3

The third dataset represents two events of interest in high background noise. The heat map in Figure 4.15a corresponds to this dataset. The events of interest are the triangular shaped blocks depicted in a lighter shade of blue than the background. This is a particularly challenging example not only because the background noise level is high, but also because the variance of the background noise level is not uniform. For example the background noise level in columns 600 - 800 has very high variance, and the background noise level in columns 1400 - 2000 has relatively low variance. Also, the event of interest occurs in the midst of this transition from high noise variance to low noise variance. The time series plots of maximum values across row and column locations (Figure 4.15b and Figure 4.15c) also present the non uniformity of the variance in this dataset.

Proposed method

Our Bayesian change-point segmented ARMA model is developed on the premise of constant noise variance. As such, we do not expect the same level of accuracy from the proposed method as in the last two examples. By examining the row change-point profile plot in Figure 4.16a, we see two higher peaks around row 199-353, which correspond to the first real event. We also see two other spikes with high probability values around row 544-687 which correspond to the second real event. This change-point profile plot also shows some other change-points within the events as well as outside the events, which are actually false

positives because of non-uniformity of the background noise level. We can detect the real two distinct events from the group profiles plots in Figure 4.16b regardless of the presence of some false positive change-points. The column change point profile plot (Figure 4.16c) gives two spikes with high probability in columns 1140 - 1358 which correspond to the real event and one spike at 718 location. This spike is a false positive change-point because of the high background noise. Figure 4.16d exhibits two events: the second one indicates the real event but the first one is actually a false positive. This opens up avenues for exploration, in particular, ways to improve our model to adapt to changing noise variance.

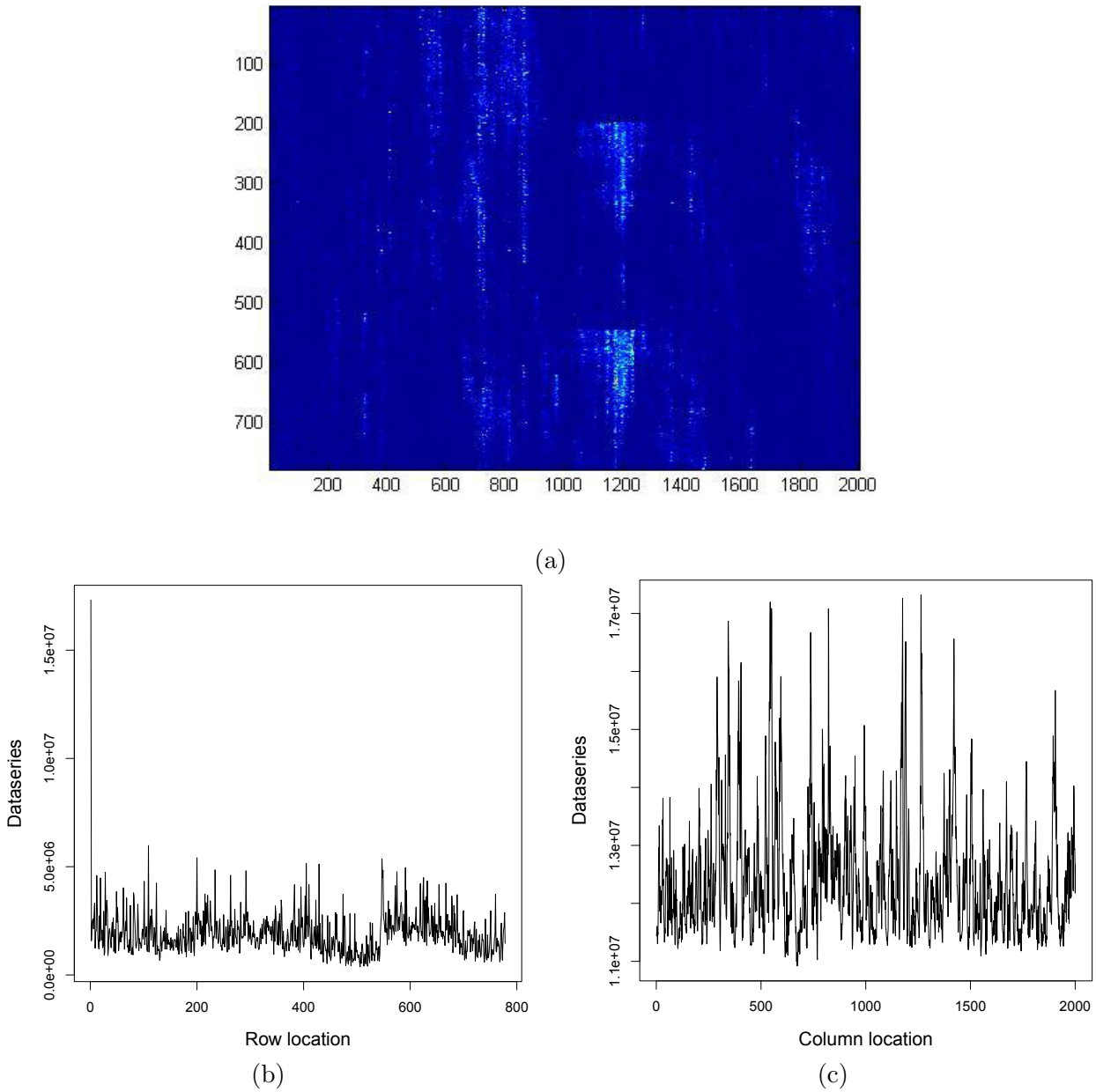


Figure 4.15: (a) Heat map of an event of interest surrounded by high background noise, (b) Time series plot of maximum values across row locations, and (c) Time series plot of maximum values across column locations

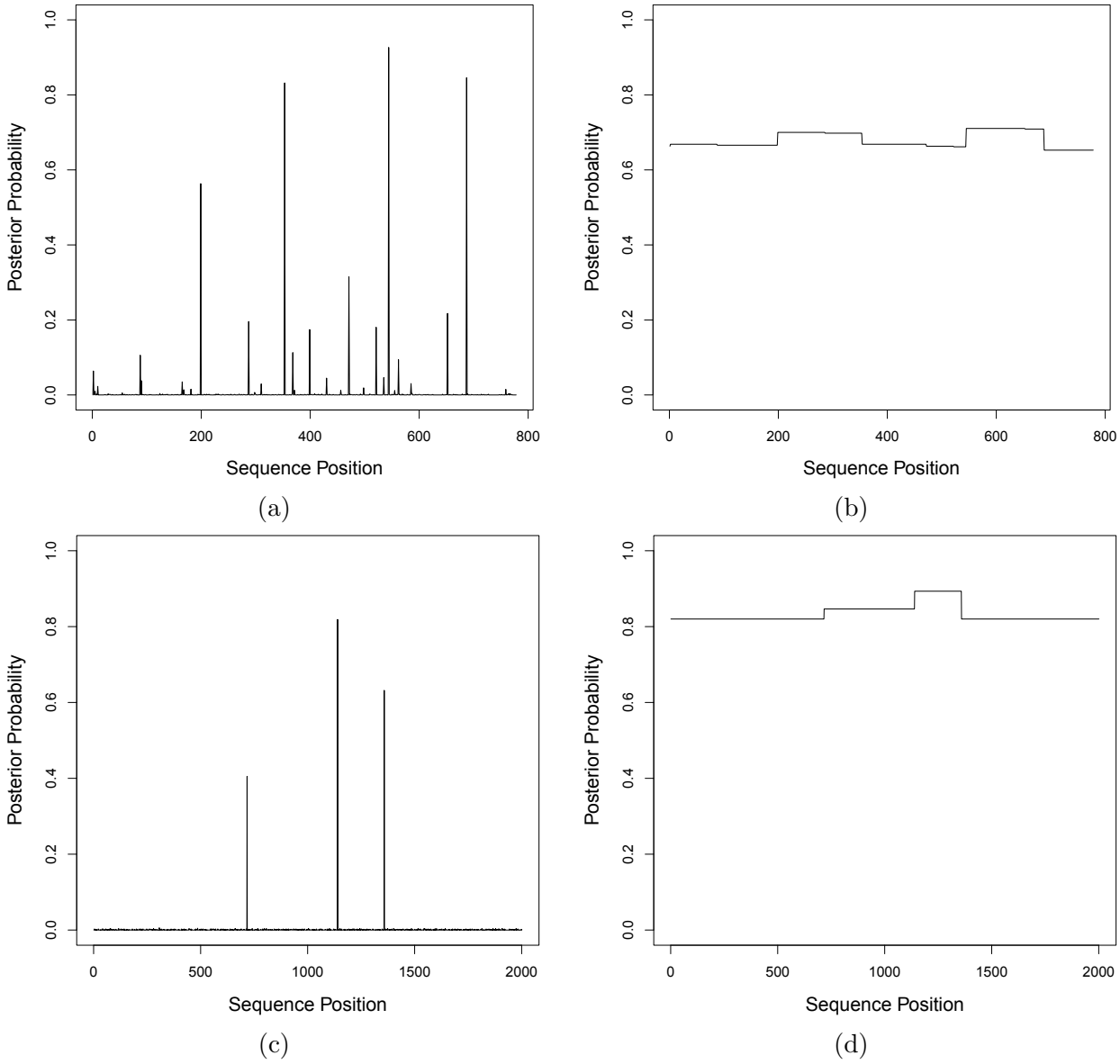


Figure 4.16: Profile plots: (a) Change-point profile plot across row locations, (b) Group profile plot across row locations, (c) Change-point profile plot across column locations, and (d) Group profile plot across column locations

Comparison method

Figure 4.17 shows the extracted events using Algorithm 2 . In addition to the true events, there are a lot of false positive events in this example. This is a particularly challenging dataset due to the non-constant variance in noise. If noise were removed prior to extracting events, we may expect better performance from Algorithm 2.

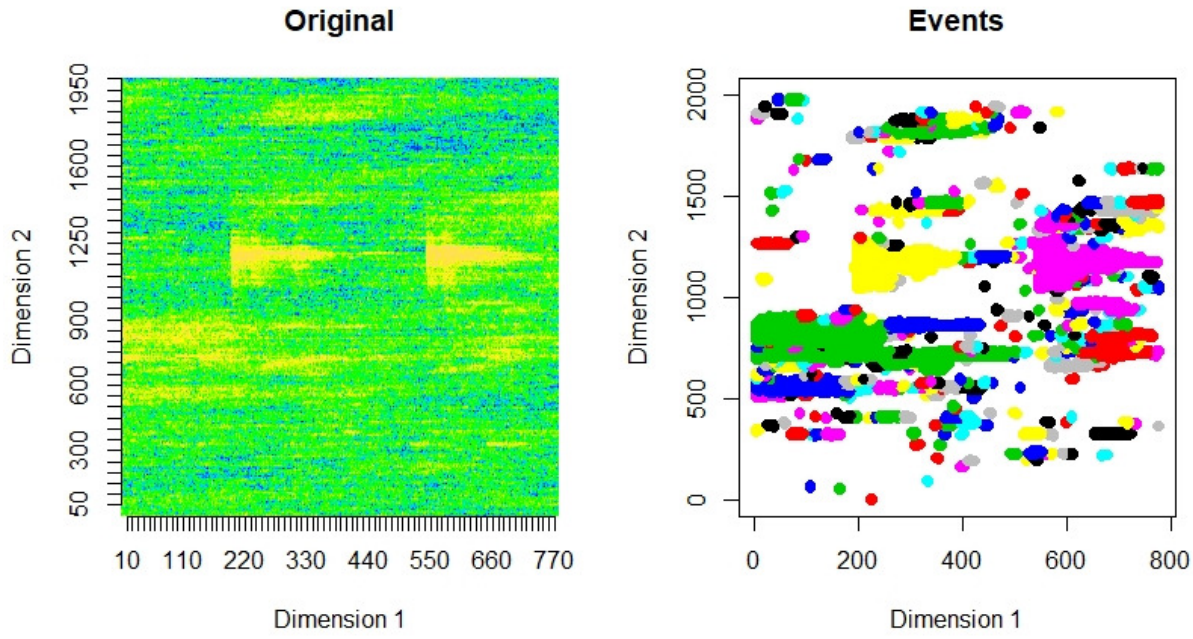
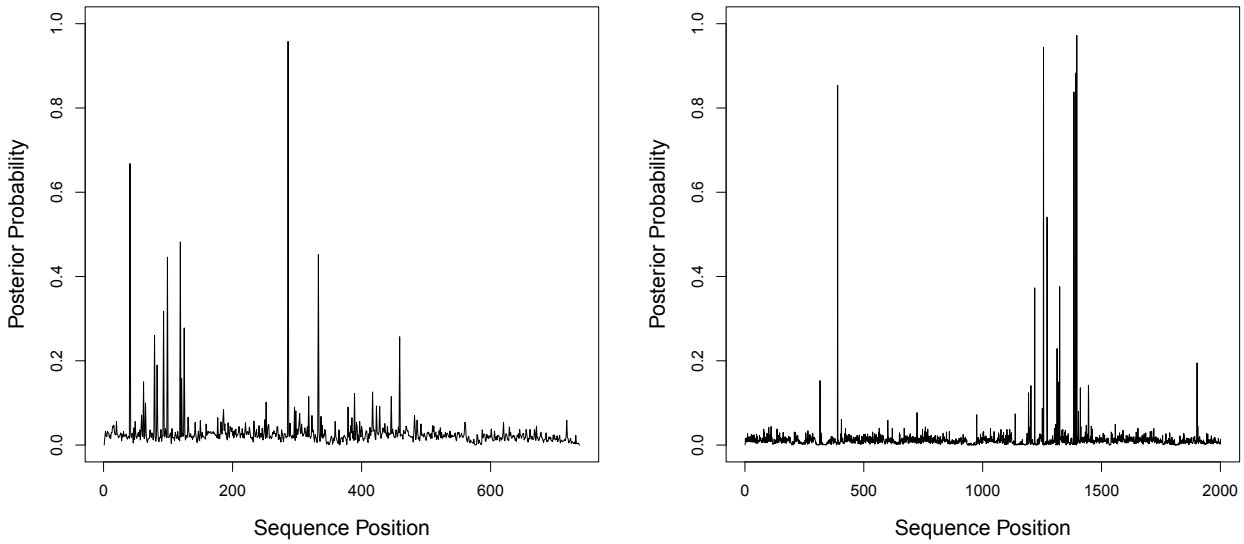


Figure 4.17: Original data and events extracted using Algorithm 2, with each extracted event depicted in a different colour.

PCA-based approach



(a) (b)

Figure 4.18: (a) Change-point profile plot across row locations, and (b) Change-point profile plot across column locations using PCA based approach

By examining the row change point profile plot (Figure 4.18a) we see that there are

high posterior probability values in rows 50- 125, which are actually false positives and a much higher peak around row 300 which corresponds to the first real event in the heat map (Figure 4.15a). The column change point profile plot (Figure 4.18b) gives a cluster of change-points in columns 1219 - 1400 which correspond to the real events and a single high peak around column 390 which is actually a false positive.

4.5 Discussion

We proposed an event detection approach for spatio-temporal data in presence of different levels of background noise. Here we are not segmenting a two-dimensional image, rather it is a complementary approach to identify events in parallel sequences. This method was tested on synthetic data as well as on real world data. Two types of synthetic data are used for event detection: One has a triangular event in presence of Gaussian noise and the other has an odd quadrangle event in presence of Gaussian noise. Both the triangle and odd quadrangle event have different mean from the background but they have the same variance as the background. The results obtained using synthetic data provide evidence of high accuracy of our method for the purpose of detecting the boundary of the events. The results have been compared with a simple event extraction method. For both synthetic data sets, the comparison method found some small noisy events in addition to the real events in spite of extracting events in a two-dimensional setting.

We also detected events in real world data: two examples with relatively low background noise and one with high background noise. Our proposed method finds the boundary of the event with high probability when the data has a low level of background noise. But it also found some other change-points as well with low probability. In this scenario, we identified the boundaries of the event by plotting group profiles of each position of the signal length. The existence of segment groups helped a lot to distinguish between 'background' and 'event' segments. By taking logarithms of the maximum values, the volatility in each segment appears to be relatively constant for the first two examples. But for the third example, the method was less successful not only because of the high background noise but also for the non-uniform variance of the background. This indicates a limitation of the Bayesian change-point segmented ARMA model in the presence of high noise variance. This model cannot be used in a situation where data shows non-uniform volatility. That's why it is important to stabilise the variance of the data by applying any types of transformation

before fitting segmented ARMA. The results of the proposed method were also compared with a simple event extraction method. The comparison method found one noisy event in addition to the real event for the first two examples, whereas our proposed method found the desired boundary of the true events. This comparison method also found more false positive events for the third example than our proposed method.

Detection of changes only in the mean value of the sequence of observations is known as additive change-point detection and detection of changes in the variance, correlation and spectral characteristics of a stochastic process is known as non-additive change-point detection [4]. Our Bayesian change-point segmented ARMA model assumes different means for different segments and the same error variance for each class into which a segment is assigned. It is therefore designed for additive change detection. In future work, we aim to generalize the Bayesian change-point segmented ARMA model to incorporate changes in the variance of the original process by allowing distinct variances for each class of the segments but constant variance with a class.

An alternative approach has been conducted for our real data examples using PCA as a summary statistic instead of maximum. This method finds a cluster of change-points covering the event when the data has a low level of background noise. However, the approach using maximum finds the boundary of the event with high probability. For the third example with high background noise, both approaches found some false positive change-points in addition to the real events. The approach using maximum found the false positive change-points with low posterior probability whereas the PCA-based approach found false positives with high posterior probability. Moreover, the approach using PCA could not find the second real event in rows whereas the maximum approach found both real events in rows. These results indicate that the maximum performs better than PCA in our examples. The data sets of real world examples do have heterogeneous variance across the signal. We used the logarithm of the maximum values as the input sequence in the maximum approach which produces almost the same variance across the signal. The PCA approach used the principal components as the input sequence which exhibit the same heteroscedasticity as in the original data. This may explain why the approach using maximum as a summary statistics exhibits better results than the PCA approach for these examples.

The reduction of dimensionality using any summary statistic involves information loss, and is therefore only recommended in initial exploratory analyses. In our findings, the fact that maximum performs better than PCA casts doubt on the value of the numerous

existing PCA-based approaches ([15–23]). The maximum is intended to be a very crude summary, and the fact that it outperforms PCA even in these simple examples suggests that PCA is also very crude. This approach could certainly be extended to other summary statistics depending on the nature of the data (fail of linearity, heteroscedasticity or lack of normality in error terms) and the type of events to be detected. To stabilize variance like our examples, we could also use Fisher transformation (for the sample correlation coefficient) , square root transformation (for count data) or Box-Cox transformation (for regression analysis) or angular transformation (for binomial data) [40, 41].

Fisher transformation for the sample correlation coefficient, the square root transformation or Anscombe transform for Poisson data (count data), the Box–Cox transformation for regression analysis, and the arcsine square root transformation or angular transformation for proportions (binomial data).

Segmentation of multiple sequences in parallel is an important current line of research. The results of this study highlight the importance of developing efficient algorithms for segmenting parallel sequences instead of using a one-dimensional approach based on dimension reduction. Consequently, our ongoing work is centered on extending the methodology proposed here for segmenting multiple sequences in parallel. This can address the shortcomings of methods based on dimension reduction. For example, segmentation of parallel multiple sequence can pool change-point information across the series to allow for more efficient detection. Moreover, this can consider dependency between multiple series.

Acknowledgements

This work was funded by the Australian Research Council (http://www.arc.gov.au/) grant DP1095849. The authors are grateful to the Australian Research Council (ARC) and the ARC Centre of Excellence for Mathematical and Statistical Frontiers for their support of this project (DP1095849, CE140100049).

Data availability

The first two real datasets of this article are publicly available in the R package oddstream: https://github.com/pridiltal/oddstream. The third dataset is available in https://github.com/farhanas-monashuniversity/Dataset-3.

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. Khodadadi, A. & Asgharian, M. Change-point problem and regression: an annotated bibliography. *COBRA Preprint Series,* 44 (2008).

2. Reeves, J., Chen, J., Wang, X. L., Lund, R. & Lu, Q. Q. A review and comparison of changepoint detection techniques for climate data. *Journal of Applied Meteorology and Climatology* **46,** 900–915 (2007).

3. Wang, Y., Wu, C., Ji, Z., Wang, B. & Liang, Y. Non-parametric change-point method for differential gene expression detection. *PloS one* **6,** e20060 (2011).

4. Basseville, M., Nikiforov, I. V., *et al. Detection of abrupt changes: theory and application* (Prentice Hall Englewood Cliffs, 1993).

5. Naveen, N., Natarajan, S. & Srinivasan, R. *Application of Change Point Outlier Detection Methods in Real Time Intrusion Detection* in *2012 International Conference on Advanced Computer Science Applications and Technologies (ACSAT)* (2012), 110–115.

6. Algama, M. & Keith, J. M. Investigating genomic structure using changept: A Bayesian segmentation model. *Computational and structural biotechnology journal* **10,** 107–115 (2014).

7. Aminikhanghahi, S. & Cook, D. J. A survey of methods for time series change point detection. *Knowledge and information systems* **51,** 339–367 (2017).

8. Tartakovsky, A. G., Rozovskii, B. L., Blazek, R. B. & Kim, H. A novel approach to detection of intrusions in computer networks via adaptive sequential and batch-sequential change-point detection methods. *IEEE Transactions on Signal Processing* **54,** 3372–3382 (2006).

9. Akoglu, L. & Faloutsos, C. *Event detection in time series of mobile communication graphs* in *Army science conference* **1** (2010).

10. Batal, I., Fradkin, D., Harrison, J., Moerchen, F. & Hauskrecht, M. *Mining recent temporal patterns for event detection in multivariate time series data* in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (2012), 280–288.

11. Guralnik, V. & Srivastava, J. *Event detection from time series data* in *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining* (1999), 33–42.

12. Ihler, A., Hutchins, J. & Smyth, P. *Adaptive event detection with time-varying poisson processes* in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (2006), 207–216.

13. Burnaev, E. & Ishimtsev, V. Conformalized density-and distance-based anomaly detection in time-series data. *arXiv preprint arXiv:1608.04585* (2016).

14. Qiao, Z., He, J., Cao, J., Huang, G. & Zhang, P. *Multiple time series anomaly detection based on compression and correlation analysis: a medical surveillance case study* in *Asia-Pacific Web Conference* (2012), 294–305.

15. Qahtan, A. A., Alharbi, B., Wang, S. & Zhang, X. *A pca-based change detection framework for multidimensional data streams: Change detection in multidimensional data streams* in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2015), 935–944.

16. Deng, J., Wang, K., Deng, Y. & Qi, G. PCA-based land-use change detection and analysis using multitemporal and multisensor satellite data. *International Journal of Remote Sensing* **29,** 4823–4838 (2008).

17. Jin-Song, D., Ke, W., Jun, L. & Yan-Hua, D. Urban land use change detection using multisensor satellite images. *Pedosphere* **19,** 96–103 (2009).

18. Kuncheva, L. I. & Faithfull, W. J. PCA feature extraction for change detection in multidimensional unlabeled data. *IEEE transactions on neural networks and learning systems* **25,** 69–80 (2013).

19. Lakhina, A., Crovella, M. & Diot, C. *Diagnosing network-wide traffic anomalies* in *ACM SIGCOMM computer communication review* **34** (2004), 219–230.

20. Li, X. & Yeh, A. Principal component analysis of stacked multi-temporal images for the monitoring of rapid urban expansion in the Pearl River Delta. *International Journal of Remote Sensing* **19,** 1501–1518 (1998).

21. Lu, D., Mausel, P., Batistella, M. & Moran, E. Land-cover binary change detection methods for use in the moist tropical region of the Amazon: a comparative study. *International Journal of Remote Sensing* **26,** 101–114 (2005).

22. MAHMOUDZADEH, H. Digital change detection using remotely sensed data for monitoring green space destruction in Tabriz (2007).

23. Nguyen, H.-V. & Vreeken, J. *Linear-time detection of non-linear changes in massively high dimensional time series* in *Proceedings of the 2016 SIAM International Conference on Data Mining* (2016), 828–836.

24. Sethi, T. S. & Kantardzic, M. On the reliable detection of concept drift from streaming unlabeled data. *Expert Systems with Applications* **82,** 77–99 (2017).

25. Lévy-Leduc, C., Roueff, F., *et al.* Detection and localization of change-points in high-dimensional network traffic data. *The Annals of Applied Statistics* **3,** 637–662 (2009).

26. Zhang, Y., Singh, S., Sen, S., Duffield, N. & Lund, C. *Online identification of hierarchical heavy hitters: algorithms, evaluation, and applications* in *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement* (2004), 101–114.

27. Krishnamurthy, B., Sen, S., Zhang, Y. & Chen, Y. *Sketch-based change detection: methods, evaluation, and applications* in *Proceedings of the 3rd ACM SIGCOMM conference on Internet measurement* (2003), 234–247.

28. Li, X. *et al. Detection and identification of network anomalies using sketch subspaces* in *Proceedings of the 6th ACM SIGCOMM conference on Internet measurement* (2006), 147–152.

29. Ciuperca, G. & Maciak, M. Changepoint Detection by the Quantile LASSO Method. *Journal of Statistical Theory and Practice* **14,** 11 (2020).

30. Meinshausen, N., Yu, B., *et al.* Lasso-type recovery of sparse representations for high-dimensional data. *The annals of statistics* **37,** 246–270 (2009).

31. Rojas, C. R. & Wahlberg, B. On change point detection using the fused lasso method. *arXiv preprint arXiv:1401.5408* (2014).

32. Zou, C. & Qiu, P. Multivariate statistical process control using LASSO. *Journal of the American Statistical Association* **104,** 1586–1596 (2009).

33. Sadia, F., Boyd, S. & Keith, J. M. Bayesian change-point modeling with segmented ARMA model. *PloS one* **13,** e0208927 (2018).

34. Gelman, A. *et al. Bayesian data analysis* (Chapman and Hall/CRC, 2013).

35. Sturtz, S., Ligges, U. & Gelman, A. E. R2WinBUGS: a package for running WinBUGS from R (2005).

36. Hyndman, R. J. & Athanasopoulos, G. Forecasting: principles and practice. OTexts, 2014. *There is no corresponding record for this reference.[Google Scholar]* (2017).

37. Kandanaarachchi, S., Hyndman, R. J. & Smith-Miles, K. Early classification of spatio-temporal events using partial information (2019).

38. Hahsler, M., Piekenbrock, M., Arya, S. & Mount, D. dbscan: Density Based Clustering of Applications with Noise (DBSCAN) and Related Algorithms. *R package version,* 1– (2017).

39. Keenan, D. M. A Tukey nonadditivity-type test for time series nonlinearity. *Biometrika* **72,** 39–44 (1985).

40. McDonald, J. H. *Handbook of biological statistics* (sparky house publishing Baltimore, MD, 2009).

41. Warton, D. I. & Hui, F. K. The arcsine is asinine: the analysis of proportions in ecology. *Ecology* **92,** 3–10 (2011).

Chapter 5

Segmenting multiple sequences in parallel: three paradigms

Chapter Objectives

This chapter addresses the second part of the objective of this thesis, which is to develop methods to segment multiple parallel time series. To achieve this, three alternative models were developed for simultaneous segmentation of multiple parallel series. These models were the generalizations of the Bayesian change-point segmented ARMA model with different assumptions. Unknown parameters and hyperparameters in each model were jointly estimated using a highly efficient Generalized Gibbs Sampling strategy. These models were tested by applying them on a synthetic example and a real life example. All generalizations showed encouraging results for a synthetic example by identifying all change-point locations for a data set for which the segmentation of a single time series detected only some change-points. For the real life example, all the generalizations showed different results. Consequently, these alternative models were compared using three different approximated information criteria. The suitability of the three generalizations was also investigated by comparing the results with change-point locations detected in previous studies of this data. The results obtained with the new method were consistent with those from previous studies.

Authorship

Farhana Sadia[1], Anna Lintern[2], David McCarthy[2] Jonathan M. Keith[1]

1 School of Mathematics, Monash University, Clayton, VIC 3800, Australia

2 Department of Civil Engineering, Monash University, Clayton, VIC 3800, Australia

Reference

Sadia F, Lintern A, McCarthy D, Keith JM. (2018). Segmenting multiple sequences in parallel: three paradigms. Submitted.

5.1 Abstract

The objective of linear segmentation methods is to detect abrupt changes in a signal. Segmentation in multiple sequences simultaneously is a significant and complex task arising in many domains. This paper approaches such problems using a hierarchical Bayesian framework. We consider three generalizations of a Bayesian change-point model in order to deal with multiple sequences. A Generalized Gibbs Sampling strategy allows joint estimation of the unknown parameters and hyperparameters in each model. The performances of these approaches are assessed via a simulation study and a real world sediment core data set. The latter application involved simultaneous segmentation of multiple physical and geochemical properties of sediments in parallel to identify the trends in and main causes of historical change in the hydrology of rivers and their floodplains. The results obtained for simulated data are encouraging and show that these models achieve high detection accuracy. All generalizations produced similar results in the simulated example but in the real life example, these generalizations produced different results. We propose a number of model selection criteria for selecting the most appropriate generalization.

Keywords: Segmentation; parallel sequences; hierarchical Bayesian analysis; Generalized Gibbs sampling; change-point detection..

5.2 Introduction

Segmentation, also known as change-point detection, divides a sequence of observations into segments, in each of which the sequence behaves as an approximately stationary time series. The aim of segmentation is to identify and localize abrupt changes and use this information to detect events of interest in the nonstationary time series[1]. Segmentation for a single series has been the subject of comprehensive research in recent decades [2–8].

One of the key challenges in segmentation is to segment multiple sequences at a time. The problem of jointly analysing several series arises in many application fields. For example, this problem arises in the simultaneous analysis of multiple genomic profiles associated with several patients[9], detection of trends in hydro-climatic variables observed in different locations[10], joint segmentation of multivariate astronomical photon counting time series data [11], or the analysis of human activity recognition through joint segmentation of multidimensional time series of acceleration data [12].

Segmentation of two or more series in parallel typically involves one of three distinct types of statistical problems. The first is the simultaneous segmentation of multiple series by assuming common change-points among all series [10, 13, 14]. The second is joint segmentation where each series has its own specific number and location of change-points [9, 11, 15]. The third type of problem involves observing several series independently and comparing the change-point location between those series[14].

Zhang *et al.* [13] considered simultaneous segmentations in multiple sequences to identify the DNA copy number variants in multiple samples. They considered the problem of detecting local signals that occur at the same location in multiple one-dimensional noisy sequences, giving special attention to relatively weak signals that may occur in only a fraction of the sequences. They suggested simple scan and segmentation algorithms. The algorithms depend on summing chi-square statistics across samples. With the help of these algorithms, they investigated the particular problem of detecting a shared abrupt jump in mean by assuming the noise within each profile to be independent and identically distributed Gaussian variables.

Harlé *et al.* [16] considered the problem of joint segmentation in multivariate time series data by proposing a Bayesian model. To model the presence or absence of change-points at different time points, this Bayesian model defined indicator variables and modelled the change-points using Bernoulli variables. This model combines a robust non-parametric statistical test acting on individual time segments in a Bayesian framework.

To compare change-points locations between several series which had been observed independently, Cleynen and Robin [14] developed two exact approaches in a Bayesian framework. The first approach compared the locations of change-points in both series. The second approach compared more than two series and estimated the posterior probability for a given change-point to have the same location in all series. To make exact inference on the change-point model, both approaches used a Bayesian segmentation model introduced by Rigaill *et al.* [17] with conjugate priors. The authors used several alternative data models in their Bayesian segmentation model, including a Gaussian heteroscedastic distribution, a Gaussian homoscedastic distribution with known variance, a Poisson, and a negative binomial homoscedastic distribution with known dispersion.

Dobigeon *et al.* [11] proposed a joint segmentation algorithm, based on a hierarchical Bayesian model. This algorithm was developed for piecewise constant autoregressive (AR) processes with fixed order on each segment. The authors identified a suitable prior for

considering the correlations between change-point locations of the observed signals. The resulting posterior distribution was sampled using Gibbs sampling strategy. They also considered an extension of this model, assuming unknown model order of the piecewise constant AR processes.

Collilieux *et al.* [15] considered another joint segmentation of correlated time-series data involving spatio-temporal data. They modelled the dependency between series to avoid false change-point detection. They assumed each series to be affected by changes at series-specific breakpoints and also assumed the observed sequences at each spatial location were correlated. They proposed a model for correlated Gaussian series based on a segmentation model combined with a factor model. The authors also developed a heuristic model selection procedure combining two BIC criteria: the classical BIC to determine the number of factors and a modified BIC criterion in the context of segmentation to determine the number of segments. Some other methods to segment parallel sequence include [18–22].

In this article, we consider the simultaneous analysis of multiple sequences in parallel, where the sequences in question may have different data types and characteristics. To deal with two or more sequences at the same time, we generalize the Bayesian change-point segmented ARMA model developed for the one dimensional series described in our earlier article [23]. Here, we propose three generalizations, which we define with reference to the following assumptions:

- Change-points occurs in corresponding locations in all series.
- All time series have the same probability of assigning segments to segment classes.
- Corresponding segments in each time-series belong to a common segment class.

The first generalization makes the first assumption, but not the other two, the second generalization makes the first two, but not the last, and the third generalization makes all three assumptions. In these generalizations, we are interested only in change-points for the mean and we assume variance and ARMA parameters do not change, although they may differ for different sequences. These generalizations consider parallel time series as independent. Dealing with multiple correlated time series requires to account for the correlations between these parallel time series without imposing that the changes occur simultaneously.

5.3 Methodology

5.3.1 First generalization

Our Bayesian change-point segmented ARMA model [23] models a time series by segmenting it into blocks of autoregressive moving average (ARMA) processes. Here we generalize this model to consider multiple parallel time series. In this generalization, we assume change-points occur at corresponding times (or locations) in all time series. We assume that a time series is a realization of a stochastic process i.e., a sequence of random variables (x_t), where the values of the index t correspond to ordered times. The time-series represents observations made at regular intervals indexed by $t = 1, \cdots, T$, which in many cases correspond to time intervals of equal length. Here, T is the total number of observations per sequence, which is the same for all sequences. In this generalization, the time series signal that we want to segment will be $\mathbf{X}_s = (x_{s,t})_{t=1}^{T}$ where s is the index of the sequences, that is, $s \in \{1, \cdots, S\}$. We divide all sequences into segments and fit ARMA models to each segment. For each segment, an ARMA model is given by:

$$x_{s,t} = c_{s,k} + \epsilon_{s,t} + \sum_{i=1}^{a} \psi_{s,i}(x_{s,t-i} - c_{s,k}) + \sum_{i=1}^{m} \theta_{s,i}\epsilon_{s,t-i}. \tag{5.1}$$

where $c_{s,k}$ is the mean signal level or mean of the ARMA model of the sequence $\mathbf{X}_s$ for segment k, $\epsilon_{s,t}, \epsilon_{s,t-1}, \cdots$ are error or white noise terms and $\boldsymbol{\epsilon_s} = (\epsilon_{s,1}, \cdots, \epsilon_{s,T})$ represents the vector of error terms for $t = 1, \cdots, T$. Further, the error terms are assumed to be sampled from a normal distribution with mean 0 and variance σ_s^2 where the subscript s indicates a different variance for each sequence. Here, $\psi_{s,1}, \cdots, \psi_{s,a}$ are the parameters of an AR(a) model of order a and $\theta_{s,1}, \cdots, \theta_{s,m}$ represent the parameters of an MA(m) model of order m in the sequence $\mathbf{X}_s$. Note that, the variance and the ARMA parameters are held constant for all segments. Since we assume the same number and locations of change-points in all sequences, the probability of generating a segmentation with K segments starting at positions $\mathbf{p} = (1 = p_1 < \cdots < p_k < T)$ is:

$$p(K, \mathbf{p}|\phi) = \phi^{K-1}(1-\phi)^{T-K-1}. \tag{5.2}$$

Here, ϕ is the probability of starting a new segment. We assume that the first segment always starts at the beginning of the signal ($p_1 = 1$) and the last segment always finishes at the end of the signal ($d_K = T$) where $\mathbf{d} = (d_1, \cdots, d_K)$ are the right hand end points

of the segments. In each sequence, each segment is assigned to one of N_s groups with probabilities $\boldsymbol{\pi}_s = (\pi_{s,1}, \cdots, \pi_{s,N_s})$ where π_{s,n_s} is the probability of assigning a segment of the sequence $\mathbf{X}_s$ to group $n \leq N_s$. The group of segment k in the sequence $\mathbf{X}_s$ is denoted by $g_{s,k} \in \{1, \cdots, N_s\}$. Let, $\mathbf{g}_s = (g_{s,1}, \cdots, g_{s,K})$ be a vector containing the group assignments of the segments. Then the probability that $b_{s,1}$ segments are assigned to group '1', $\cdots, b_{s,N_s}$ segments are assigned to group 'N_s' is:

$$p(\mathbf{g}_s|K, \boldsymbol{\pi}_s) = \pi_{s,1}^{b_{s,1}} \cdots \pi_{s,N_s}^{x_{b,N_s}} = \prod_{n_s=1}^{N_s} \boldsymbol{\pi}_{s,n_s}^{b_{s,n_s}} = \prod_{k=1}^{K} \pi_{s,g_{s,k}}. \tag{5.3}$$

Each segment's mean $c_{s,k}$ is normally distributed with mean $\mu_{s,g_{s,k}}$ and variance $\tau^2_{s,g_{s,k}}$ for a segment in group $g_{s,k}$, that is, the probability of the ARMA means for all segments is:

$$p(\mathbf{c}_s|\mathbf{g}_s, \boldsymbol{\mu}_s, \boldsymbol{\tau}_s) = \prod_{k=1}^{K} norm(c_{s,k}|\mu_{s,g_{s,k}}, \tau^2_{s,g_{s,k}}). \tag{5.4}$$

where, $\mathbf{c}_s = (c_{s,1}, \cdots, c_{s,K})$ is the mean signal level for the segment of the sequence $\mathbf{X}_s$, $\boldsymbol{\mu}_s = (\mu_{s,g_{s,1}}, \cdots, \mu_{s,g_{s,K}})$ and $\boldsymbol{\tau}^2_s = (\tau^2_{s,g_{s,1}}, \cdots, \tau^2_{s,g_{s,K}})$. Next we define $\Lambda_s = (\lambda_{s,1}, \cdots, \lambda_{s,T})$ where $\lambda_{s,T}$ is given by

$$\lambda_{s,T} = c_{s,k} + \sum_{i=1}^{a} \psi_{s,i}(x_{s,t-i} - c_{s,k}) + \sum_{i=1}^{m} \theta_{s,i}\epsilon_{s,t-i}.$$

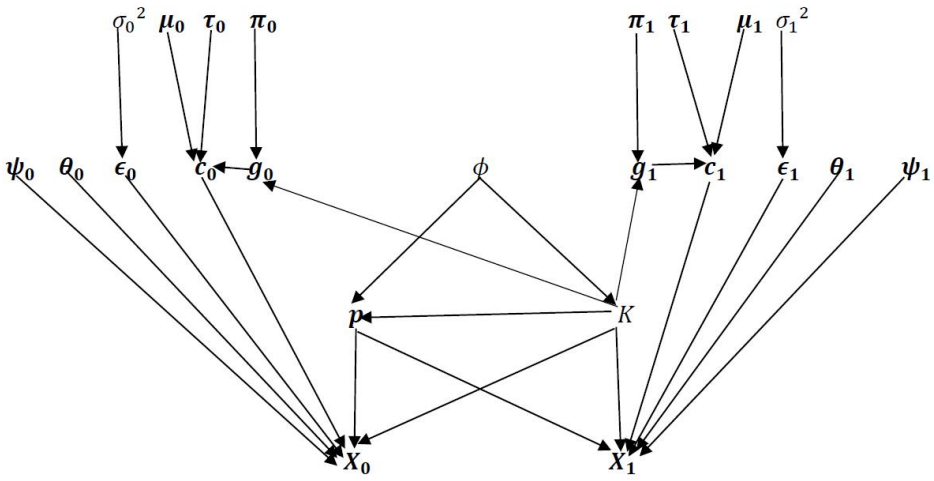
Now we can write $x_{s,t}$ as a function of $\lambda_{s,t}$ and $\epsilon_{s,t}$:

$$x_{s,t} = \lambda_{s,t} + \epsilon_{s,t}.$$

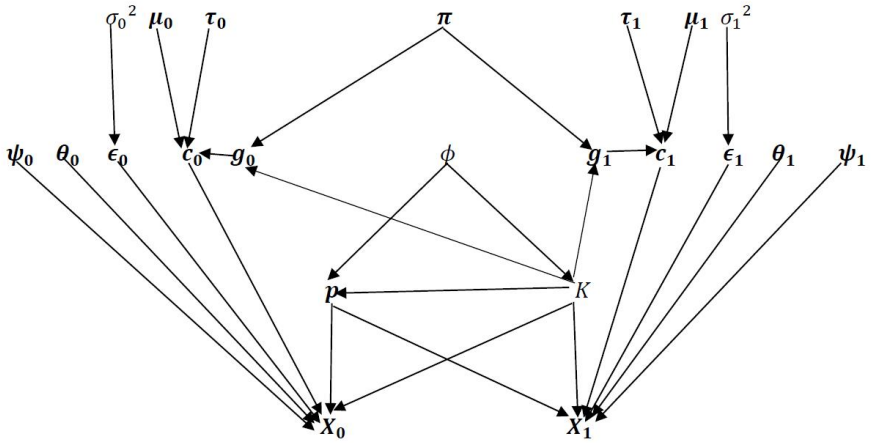
Figure 5.1a shows the parameters of the model and their conditional dependencies for two sequences. A parameter at the head of the arrow is conditionally dependent on the parameter at the tail. Finally the probability of the observed signal X_s conditioning on parameters $K, \mathbf{p}_s, \boldsymbol{\theta}_s, \boldsymbol{\psi}_s, \mathbf{c}_s$ and σ^2_s is expressed as a product of normal distributions with mean $\lambda_{s,t}$ and variance σ^2_s as follows:

$$\begin{aligned} p(\mathbf{X}_s|K, \mathbf{p}, \boldsymbol{\theta}_s, \boldsymbol{\psi}_s, \mathbf{c}_s, \sigma^2_s) &= \prod_{t=1}^{T} p(x_{s,t}|K, p, \theta_s, \psi_s, c_s, \sigma^2_s, x_{s,<t}) \\ &= \prod_{t=1}^{T} norm(x_{s,t}|\lambda_{s,t}, \sigma^2_s). \end{aligned} \tag{5.5}$$

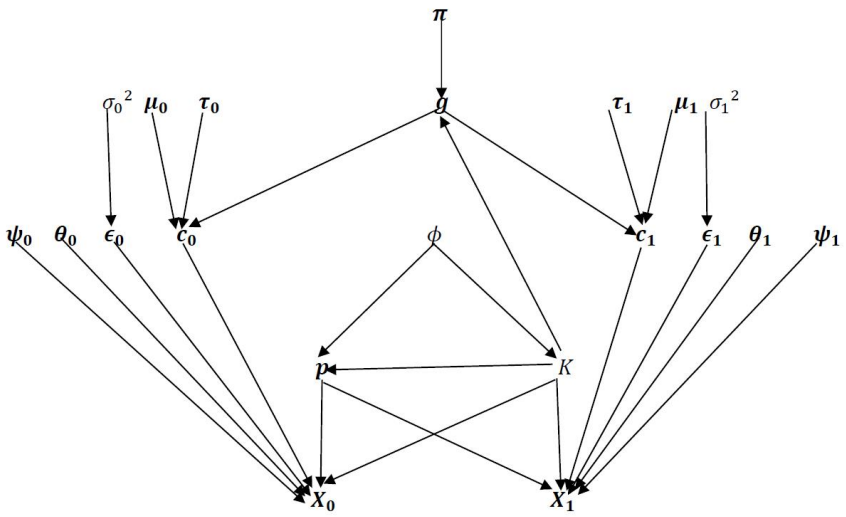
Here, $x_{s,<t} = (x_{s,1}, \cdots, x_{s,t-1})$.



(a) First generalization



(b) Second generalization



(c) Third generalization

Figure 5.1: The conditional dependencies of the parameters

As shown in Figure 5.1a, the joint distribution of $\mathbf{X}_s, K, \mathbf{p}, \mathbf{g}_s$ and $\mathbf{c}_s$ can be written as:

$$\begin{aligned} &p(\mathbf{X}_s, K, \mathbf{p}, \mathbf{g}_s, \mathbf{c}_s | \phi, \boldsymbol{\pi}_s, \boldsymbol{\theta}_s, \boldsymbol{\psi}_s, \sigma_s^2, \boldsymbol{\mu}_s, \boldsymbol{\tau}_s) = \\ &p(K, \mathbf{p}|\phi) \times \prod_{s=1}^{S} p(\mathbf{X}_s | K, \mathbf{p}, \boldsymbol{\theta}_s, \boldsymbol{\psi}_s, \mathbf{c}_s, \sigma_s^2) \times p(\mathbf{c}_s | \mathbf{g}_s, \boldsymbol{\mu}_s, \boldsymbol{\tau}_s) \times p(\mathbf{g}_s | K, \boldsymbol{\pi}_s) \end{aligned} \tag{5.6}$$

Prior and Posterior Distribution

To complete the model we must specify prior probability distributions for parameters $\phi, \boldsymbol{\pi}, \boldsymbol{\psi}, \boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\tau}^2$ and σ^2. We assign a beta prior to ϕ (probability of starting a new segment) with parameters $a = 1.0$ and $b = 1.0$, a uniform Dirichlet prior to $\boldsymbol{\pi}_s = (\pi_{s,1}, \cdots, \pi_{s,N_s})$ (probabilities of assigning segments to groups) where $\sum_{n_s=1}^{N_s} \pi_{s,n_s} = 1$. We also assign uniform prior distributions on the interval (-1,1) for parameters $\boldsymbol{\psi}_s$ and $\boldsymbol{\theta}_s$ and assume they are independent of each other. A weakly informative normal prior with mean 0.0 and variance 1.0 is chosen for the mean μ_{s,n_s} of the distribution of ARMA means in segment class N_s. Inverse gamma prior distributions with parameters $\alpha = 3.0$ and $\beta = 3.0$ are assigned to σ_s^2 and τ_{s,n_s}^2. We have chosen conjugate priors of σ_s^2 and τ_{s,n_s}^2 purely for mathematical convenience and we have chosen uninformative, or very weakly informative priors for ϕ, $\boldsymbol{\pi}_s$, μ_{s,n_s}, $\boldsymbol{\psi}_s$ and $\boldsymbol{\theta}_s$, because we have little prior knowledge of these parameters. The forms of these hyper-priors were chosen to reflect the degree of prior belief about their respective parameters. Further we treat a, the order of the AR model and m, the order of the MA model as fixed parameters of known value.

Using Bayes' theorem, we obtain the joint posterior distribution of all data and parameters by the following formula:

$$\begin{aligned} &p(\mathbf{X}_s, K, \mathbf{p}, \mathbf{g}_s, \mathbf{c}_s, \phi, \boldsymbol{\pi}_s, \boldsymbol{\theta}_s, \boldsymbol{\psi}_s, \sigma_s^2, \boldsymbol{\mu}_s, \boldsymbol{\tau}_s) = p(\phi) \times \\ &\prod_{s=1}^{S} p(\mathbf{X}_s, K, \mathbf{p}, \mathbf{g}_s, \mathbf{c}_s | \phi, \boldsymbol{\pi}_s, \boldsymbol{\theta}_s, \boldsymbol{\psi}_s, \sigma_s^2, \boldsymbol{\mu}_s, \boldsymbol{\tau}_s) p(\boldsymbol{\pi}_s) p(\boldsymbol{\theta}_s) p(\boldsymbol{\psi}_s) p(\sigma_s^2) p(\boldsymbol{\mu}_s) p(\boldsymbol{\tau}_s) \end{aligned} \tag{5.7}$$

The details of the calculation of the conditional posterior distribution of each parameter are provided in Appendix C.1.1.

Sampling

A Markov chain Monte Carlo technique called the Generalized Gibbs Sampler (GGS) is used to sample the posterior distribution described in supplementary material. (See [24] and [23] for a full description of the GGS and its application to a single sequence.) Here we describe the different move-types of GGS and the order in which these move-types are carried out for our first generalization for multiple sequences.

Move types

Figure 5.2 illustrates the following defined move-types:

- (I, k): Decide whether to insert a new change-point in segment k, and at what position.
- (D, k): Decide whether to remove change-point k or move it to a new position (for each change-point except the first).
- $c_{s,k}$: Update mean signal level $c_{s,k}$ in segment $k = 1, 2, 3, \cdots, K$ for the sequence $\mathbf{X}_s$.
- $g_{s,k}$: Update segment group assignments $g_{s,k}$ in segment $k = 1, 2, 3, \cdots, K$ for the sequence $\mathbf{X}_s$.
- $\mu_{s,\mathbf{g}_s}$: Update group mean $\mu_{s,\mathbf{g}_s}$ for group g of the sequence $\mathbf{X}_s$.
- $\tau^2_{s,\mathbf{g}_s}$: Update group variance $\tau^2_{s,\mathbf{g}_s}$ for group g of the sequence $\mathbf{X}_s$.
- $(\boldsymbol{\theta}_s, \boldsymbol{\psi}_s)$: Update $\boldsymbol{\theta}_s$ and $\boldsymbol{\psi}_s$.
- $(\boldsymbol{\pi}_s, \boldsymbol{\sigma^2}_s, \phi)$: Update all other parameters, $\boldsymbol{\pi}_s$, $\boldsymbol{\sigma^2}_s$ and ϕ.

The total number of moves for s sequences with K segments is:

$$T(K) = 2(s+1)K - 1 + 2sN + sa + sm + (2s+1), \quad s = 1, 2, \cdots, S$$

where N is the number of groups, a is the order of the AR model, m is the order of the MA model and s is the index of the sequence.

5.3.2 Second generalization

This generalization retains the condition used in the first generalization that $K, \mathbf{P}, \phi, T$ and $\boldsymbol{\pi}$ are common to all sequences. The only change from the first generalization is to assume the same probability of assigning segments to groups ($\boldsymbol{\pi}$) for all sequences. Then the probability of a specific assignment of the segments of the sequences $\mathbf{X_s}$ is,

$$p(g_s|K,\boldsymbol{\pi}) = \prod_{k=1}^{K} \boldsymbol{\pi}_{g_{s,k}} = \prod_{n_s=1}^{N_s} \boldsymbol{\pi}_{n_s}^{b_{s,n_s}}. \tag{5.8}$$

Figure 5.1b illustrates the dependencies of the parameters in this generalization for two sequences. The other notations are the same as in the previous section. Then the joint distribution of this model is:

$$\begin{aligned} &p(\mathbf{X}_s, K, \mathbf{p}, \mathbf{g}_s, \mathbf{c}_s|\phi, \boldsymbol{\pi}_s, \boldsymbol{\theta}_s, \boldsymbol{\psi}_s, \sigma_s^2, \boldsymbol{\mu}_s, \boldsymbol{\tau}_s) = p(K, \mathbf{p}|\phi)\times \\ &\prod_{s=1}^{S} p(\mathbf{X}_s|K, \mathbf{p}, \boldsymbol{\theta}_s, \boldsymbol{\psi}_s, \mathbf{c}_s, \sigma_s^2) p(\mathbf{c}_s|\mathbf{g}_s, \boldsymbol{\mu}_s, \boldsymbol{\tau}_s) p(\mathbf{g}_s|K, \boldsymbol{\pi}) \end{aligned} \tag{5.9}$$

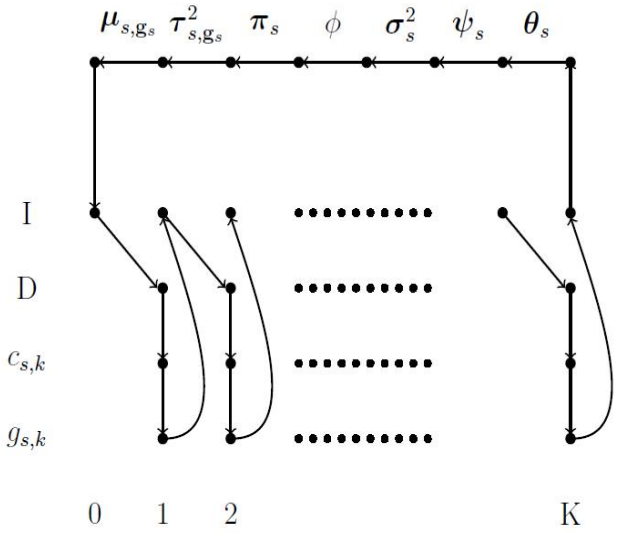


Figure 5.2: Order of move-types for the sampler of multiple sequences in first generalization. For the second generalization, all moves are same, except only we update $\boldsymbol{\pi}$ instead of updating $\boldsymbol{\pi}_s$, since it is same for all sequences in this generalization. For the third generalization, all moves are same as the first generalization described in Section 5.3.1 except $\boldsymbol{\pi}$, $\mathbf{g}_k$, $\mu_{s,\mathbf{g}}$ and $\tau^2_{s,\mathbf{g}}$ are updated in place of $\boldsymbol{\pi}_s$, $\mathbf{g}_{s,k}$, $\mu_{s,\mathbf{g}_s}$ and $\tau^2_{s,\mathbf{g}_s}$, as this generalization assume common segment class $\mathbf{g}$ and common $\boldsymbol{\pi}$ for all sequences.

Prior and Posterior Distribution

We use a Dirichlet prior distribution with parameters $(\alpha_1, \cdots \alpha_N) = (1.0, \cdots, 1.0)$ for $\boldsymbol{\pi}$. The same priors as in the first generalization are used for all the other parameters. The conditional posterior distributions of all parameters except $\boldsymbol{\pi}$ and $\mathbf{g}_s$ are the same as in the previous generalization. The conditional posterior distribution of $\boldsymbol{\pi}$ and $\mathbf{g}_s$ are given in Appendix C.2.

Sampling

Move types

As this generalization assumes all time series have the same probability of assigning segments to groups ($\boldsymbol{\pi}$), all the move types described in Sect. 5.3.1 and Figure 5.2 are the same for the second except only we update $\boldsymbol{\pi}$ instead of updating $\boldsymbol{\pi}_s$. Consequently, the total number of moves for s sequences with K segments becomes:

$$T(K) = 2(s+1)K - 1 + 2sN + sa + sm + (s+2), \quad s = 1, 2, \cdots, S$$

5.3.3 Third generalization

In addition to the assumptions of the first and second generalizations, this generalization assumes the corresponding segments in each time-series belong to a common segment class $\mathbf{g}$ for all sequences. We set $\mathbf{g} = (g_1, \cdots, g_k)$ where g_k is the group of segment k for all sequences. Then the probability of a specific assignment of the segments for all sequences is,

$$p(\mathbf{g}|K, \boldsymbol{\pi}) = \prod_{k=1}^{K} \boldsymbol{\pi}_{g_k} = \prod_{n=1}^{N} \boldsymbol{\pi_n}^{b_n} \tag{5.10}$$

Now each segment's mean $c_{s,k}$ is normally distributed with mean μ_{s,g_k} and variance τ^2_{s,g_k} for the group g_k and sequences X_s, that is, the probability of the ARMA mean for all segments is:

$$p(\mathbf{c}_s|\mathbf{g}, \boldsymbol{\mu}_s, \boldsymbol{\tau}_s) = \prod_{k=1}^{K} norm(c_{s,k}|\mu_{s,g_k}, \tau^2_{s,g_k}). \tag{5.11}$$

The other notations are the same as second generalization. Figure 5.1c illustrates the dependencies of the parameters for two sequences in this generalization. The joint distribution of this model is,

$$\begin{aligned} &p(\mathbf{X}_s, K, \mathbf{p}, \mathbf{g}_s, \mathbf{c}_s|\phi, \boldsymbol{\pi}_s, \boldsymbol{\theta}_s, \boldsymbol{\psi}_s, \sigma_s^2, \boldsymbol{\mu}_s, \boldsymbol{\tau}_s) = p(K, \mathbf{p}|\phi)\times \\ &p(\mathbf{g}|K, \boldsymbol{\pi}) \times \prod_{s=1}^{S} p(\mathbf{X}_s|K, \mathbf{p}, \boldsymbol{\theta}_s, \boldsymbol{\psi}_s, \mathbf{c}_s, \sigma_s^2) p(\mathbf{c}_s|\mathbf{g}_s, \boldsymbol{\mu}_s) \end{aligned} \tag{5.12}$$

Prior and Posterior Distribution

We use the same priors as in the second generalization for all parameters. The conditional posterior distributions of all parameters are the same as in the second generalization except for $\mathbf{g}, \boldsymbol{\mu}_s, \boldsymbol{\tau}_s^2$ and $\mathbf{c}_s$ and $\boldsymbol{\epsilon}_s$. All the conditional posterior distributions are given in Appendix C.1.3.

Sampling

Move types

As this generalization assumes common segment class $\mathbf{g}$ for all sequences, all the move types described in Sect. 5.3.1 and Figure 5.2 are the same for the third except we update $\boldsymbol{\pi}$, g_k, $\mu_{s,\mathbf{g}}$ and $\tau^2_{s,\mathbf{g}}$ instead of updating $\boldsymbol{\pi}_s$, $g_{s,k}$, $\mu_{s,\mathbf{g}_s}$ and $\tau^2_{s,\mathbf{g}_s}$. The total number of moves for s sequences with K segments becomes:

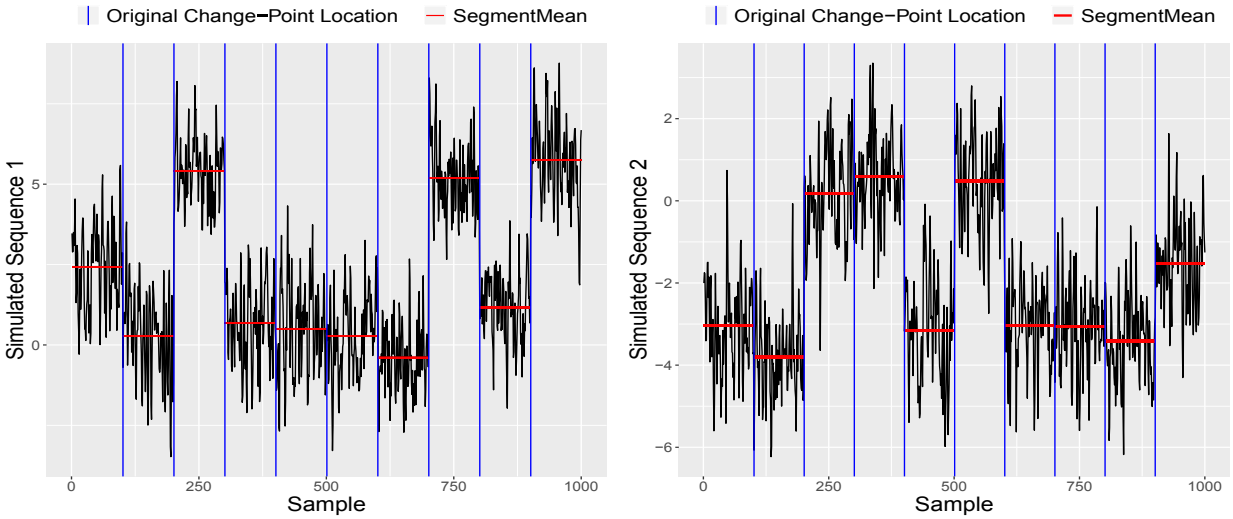
$$T(K) = (s+3)K - 1 + 2sN + sa + sm + (s+3), \quad s = 1, 2, \cdots, S$$

5.4 Illustrative Examples

5.4.1 Synthetic Example

To test the feasibility and measure the performances of the three generalizations, we generated two sequences from the autoregressive moving average(ARMA) model with parameter values $\psi_1 = 0.23$ and $\theta_1 = 0.60$ and $\psi_2 = 0.21$ and $\theta_2 = 0.77$. Here the number and location of change-points, ARMA parameters, segment means and error variance are known. Each

sequence contains 100 observations and then both sequences were replicated 10 times. The first and second sequences were generated using $\sigma_1^2 = 0.96$, $\sigma_2^2 = 0.6$ and 10 different segment means. The simulated ARMA data with the true segment means and the location of change-points for first and second sequences are given in Figure 5.3a and Figure 5.3b, respectively .

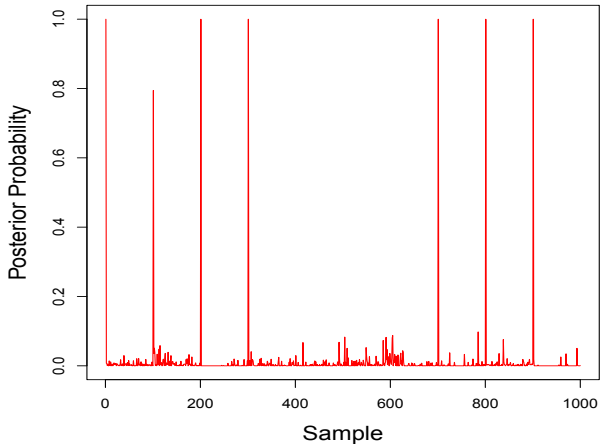


(a) First Sequence

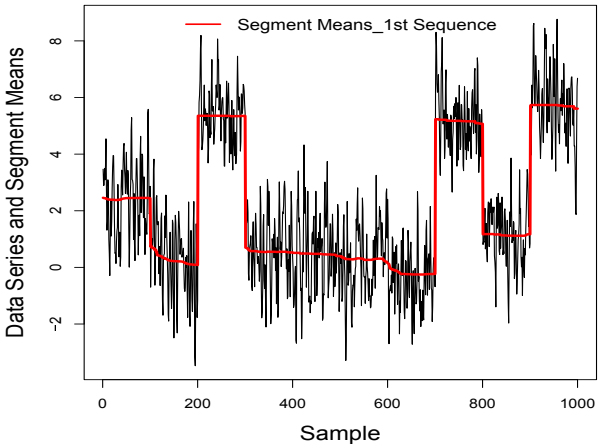
(b) Second Sequence

Figure 5.3: Simulated signal with true change-point locations and segment means. The change-point locations are shown as a vertical blue line and segment means are shown as a horizontal red line.

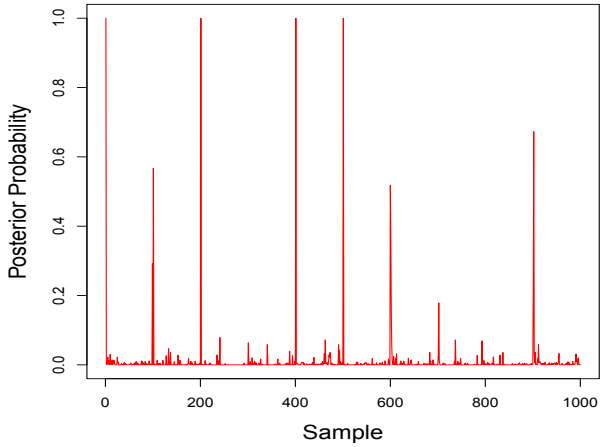
Before applying segmented ARMA for the above two sequences in parallel, we segmented each sequence separately. Figure 5.4 presents the change-point profiles plot and segment means plot. The change-point profile shows the estimation of the posterior distribution of the change-point positions and segment means at each position of the simulated signal. Figure 5.4a and Figure 5.4b point out six change-points out of nine true change-points in the first sequence, thus failing to find any change-point in the range of times 300-700. Figure 5.4c and Figure 5.4d find seven change-points out of nine true change-points in the second sequence, but couldn't detect the change-points in the ranges 200-400 and 700-900.



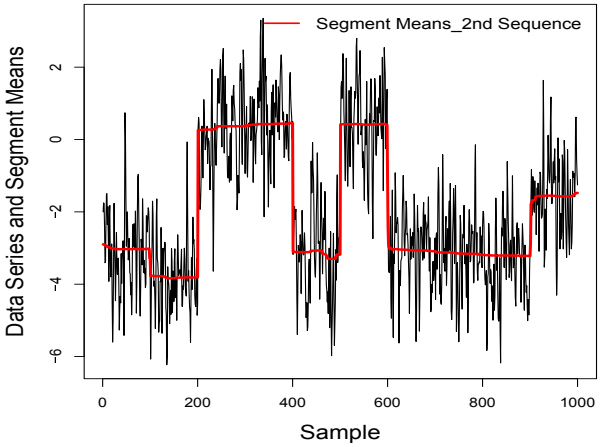
(a) Change-point profile plot: First sequence



(b) Segment mean plot: First sequence



(c) Change-point profile plot: Second sequence



(d) Segment mean plot: Second sequence

Figure 5.4: Change-point profile plots and segment mean plots

Results of three generalizations

The first generalization was used to generate the data and to infer the parameters and change-point locations. The convergence of AR and MA parameters of both sequences was investigated by plotting time series for the ARMA model parameters, as shown in Figure 5.5. Here we performed 5,000 iterations of the MCMC estimation algorithm. We discarded the first 1,000 iterations as a burn-in period. AR and MA parameters of both sequences show well-conditioned behaviour and converge rapidly. They are not only close to the true values of the parameters but also display rapid mixing. Here, we present only the trace plot for the first generalization, as the other two generalizations show similar behaviour with rapid convergence.

Figure 5.6 presents the estimated change-point locations for all the generalizations. The change-point profile plots correctly identified all the change-point locations where segmenting each sequence separately detected only some change-points. All three generalizations

produced similar results although for the second generalization, a change-point is detected at time point 600 with lower posterior probability than for the other generalizations.

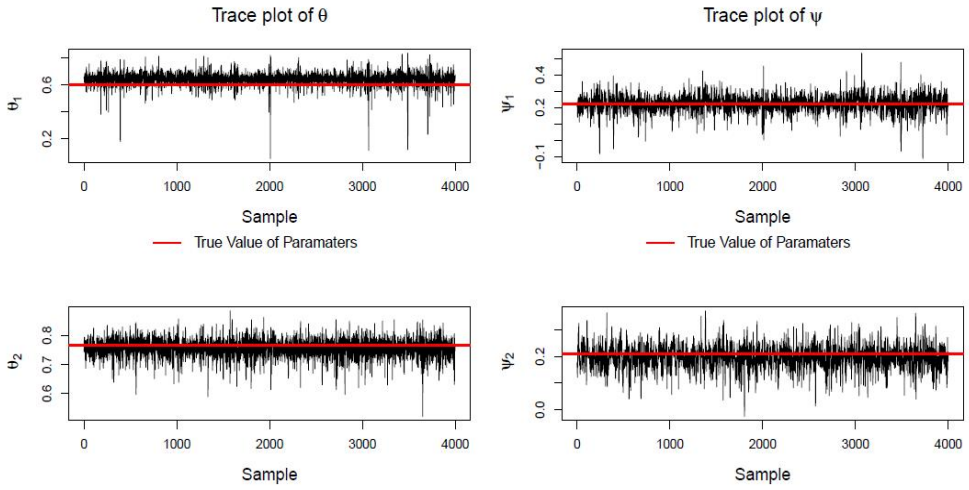


Figure 5.5: Trace plot of AR and MA parameters. Both parameters converged to a neighbourhood of the true parameters.

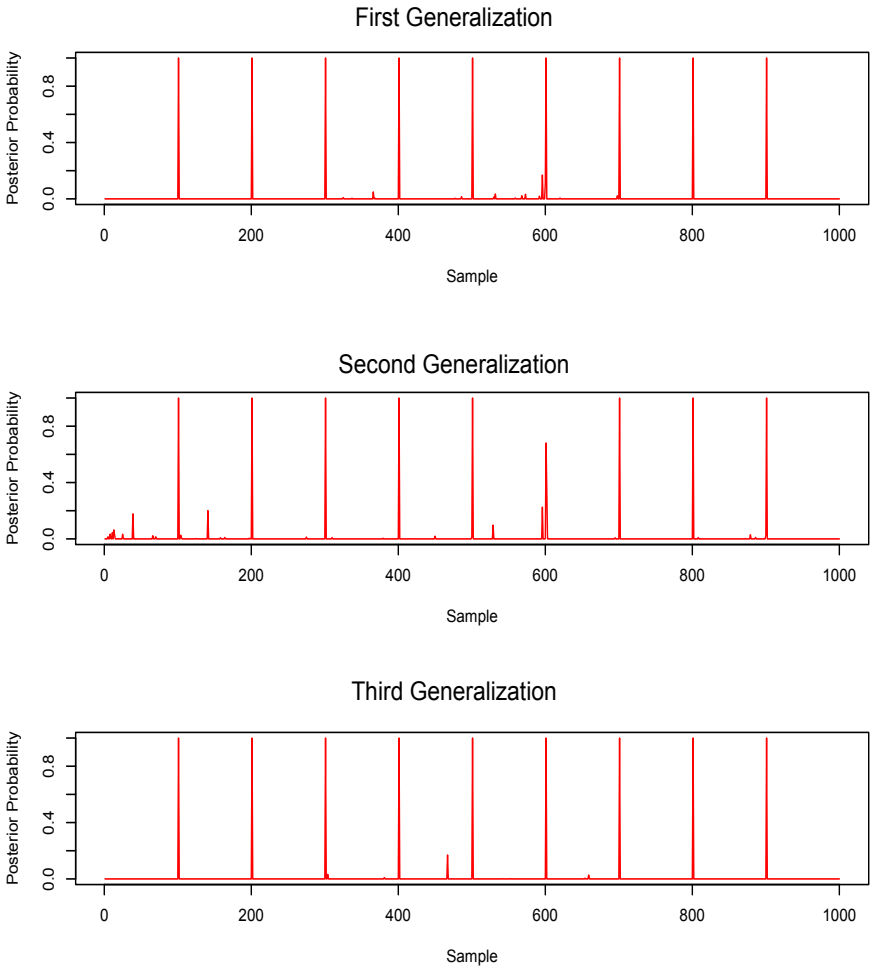


Figure 5.6: Posterior distribution of occurrence of change-point locations for the synthetic example.

Figure 5.7 plots segment means at each position of the simulated first and second signal with the true change-point locations and true segment means for all generalizations. All the plots clearly indicate 10 segments in both sequences, and are very close to the true signal. All generalizations produce similar segment means for both simulated signals. Figure 5.6 and Figure 5.7 demonstrate that the locations of estimated change-points and true change-points are similar in both sequences.

Instead of generating sequences from a stable ARMA model, the simulation studies had also been conducted for two more data sets which were simulated from AR model using different AR parameters. In both data sets, we generated 10 time series of 100 observations, each with same 10 different segment mean and same variance $\sigma_1^2 = 0.49, \sigma_1^2 = 0.6$. The two data sets differed in the value of the AR parameter used to generate the data: in the first we used $\psi_1 = 0.23$ and $\psi_2 = 0.21$ and in the second we used $\psi_1 = 0.91$ and $\psi_2 = 0.96$. We applied our methods for the segmented ARMA model and segmented AR model to these data using all the generalizations and compared the location of change points and number of change points found by these two models in all the generalizations. Detailed results are provided in Appendix C.2. These results suggest that if we generate data from an AR model with high parameter coefficient instead of ARMA model, segmented ARMA works better than segmented AR model in finding true change-points.

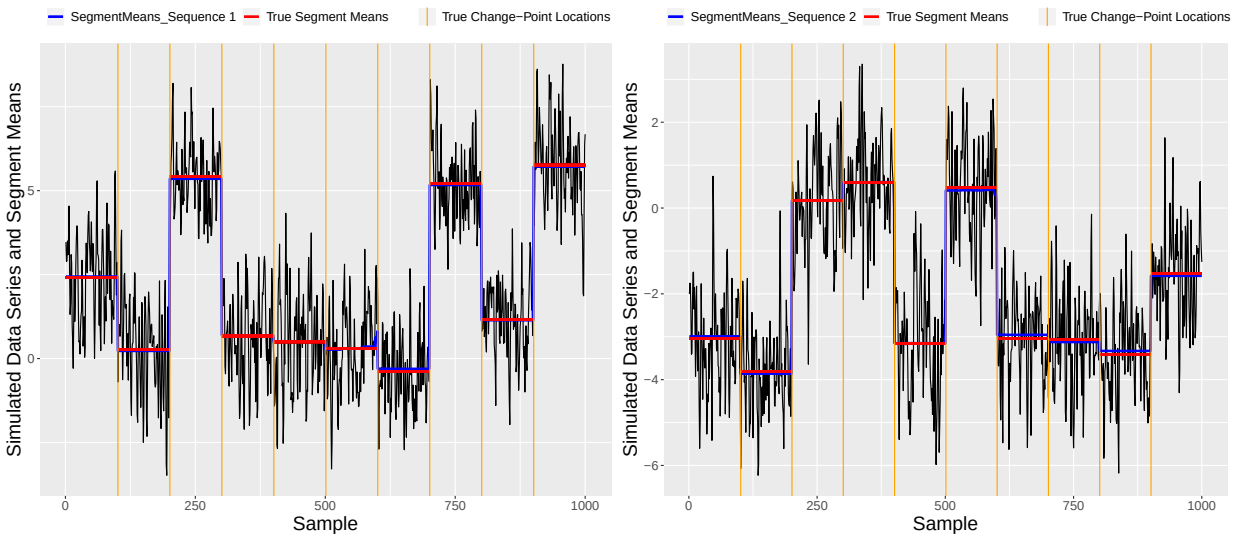
5.4.2 Application to a real data set: Results and Discussions

In this section we use the three generalizations developed in Sect. 5.3 to analyze a sediment core data set which includes concentrations of metal elements detected at each depth (in mm) in the cores. The cores are collected from a floodplain lake (Willsmere Billabong) located in the urbanised Yarra River catchment, in Victoria, Australia. Coring and analytical methods for this core have been previously published in Lintern *et al.* [25, 26].

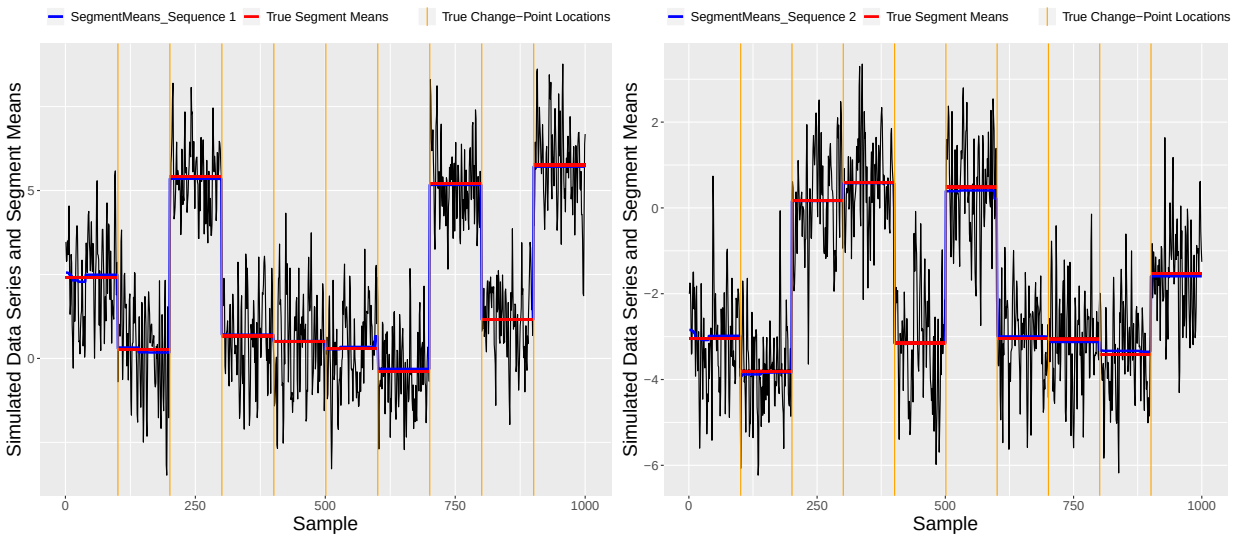
Sediment cores from aquatic environments can give important information about the presence of hydrological changes over time (at decadal, centennial and millenial scales) in the river and its floodplain lakes [27, 28]. Changes in the hydrology of a river reflect river and floodplains ecology and affect the safety of human society within the river catchment. Consequently, identification of these historical environmental changes is desirable, in order to better design environmental management strategies and to protect human life and infrastructure from fluvial floods. Specific sediment characteristics (magnetic susceptibility, organic matter, sediment particle size and elemental composition of the sediments) in the

cores need to be assessed in parallel for identifying changes in historical river hydrology [26, 28]. On that account, we are interested in simultaneous segmentation of multiple profiles in parallel for this example. We omitted some time series of elemental composition of sediments containing zeroes, as our approach is not designed for zero-inflated data. Zeros in the time series indicate that the elements were not detected in the sediments. Only ten elements with non-zero values, that also represent important sedimentological processes in waterways are included in this analysis. Figure 5.8 represents the intensity of ten elements detected at each depth (in mm) in the cores. Time series plots of Cr, Ni, Pd in Figure 5.8 clearly shows change in means and relatively constant variances in different segments. Other time series plots display heteroscedasticy in volatilities and trend, which indicates the presence of nonlinearity or nonstationarity in the data. Further descriptions about these data can be found from [26, 29, 30].

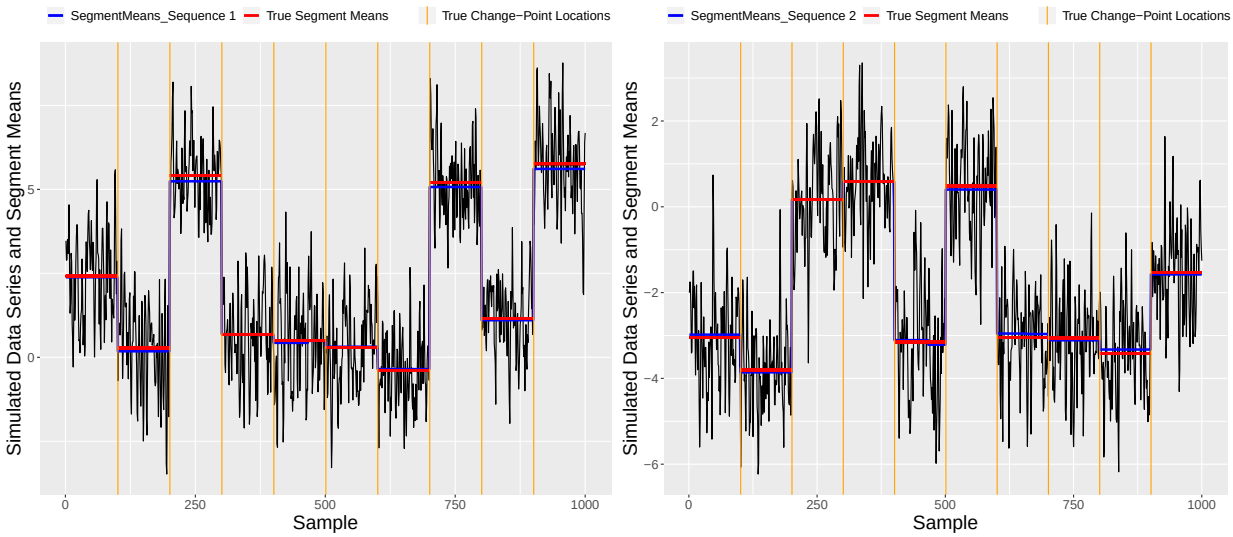
For each generalization of our Bayesian segmented ARMA model, we fitted ARMA(1,1) models to segments of every parallel sequence. Each model was run for 2000 iterations and we tested the convergence of each parameter as in the synthetic examples. To reduce the effect of heteroscedasticity in the time-series for some elements, all data were logarithmically transformed before segmentation. In all MCMC runs we used an initial value of $\phi = 0.1$ (which does not affect the stationary distribution of the Markov chain, but may affect speed of convergence). We segmented ten sequences in parallel and plotted change-point profiles for each generalization with the posterior probabilities of occurrence of change-points at each position of the input sequence (Figure 5.9). These probabilities are computed using the uniform prior probability distribution for ϕ and the likelihood probability $p(K, \mathbf{p}|\phi) = \phi^{K-1}(1-\phi)^{T-K-1}$ of generating a new segmentation with K change-points and $\mathbf{p} = (1 = p_1 < \cdots < p_k < T)$ starting positions.



(a) First Generalization



(b) Second Generalization



(c) Third Generalization

Figure 5.7: Segment mean plots with the true change-point locations and true segment means for both sequence.

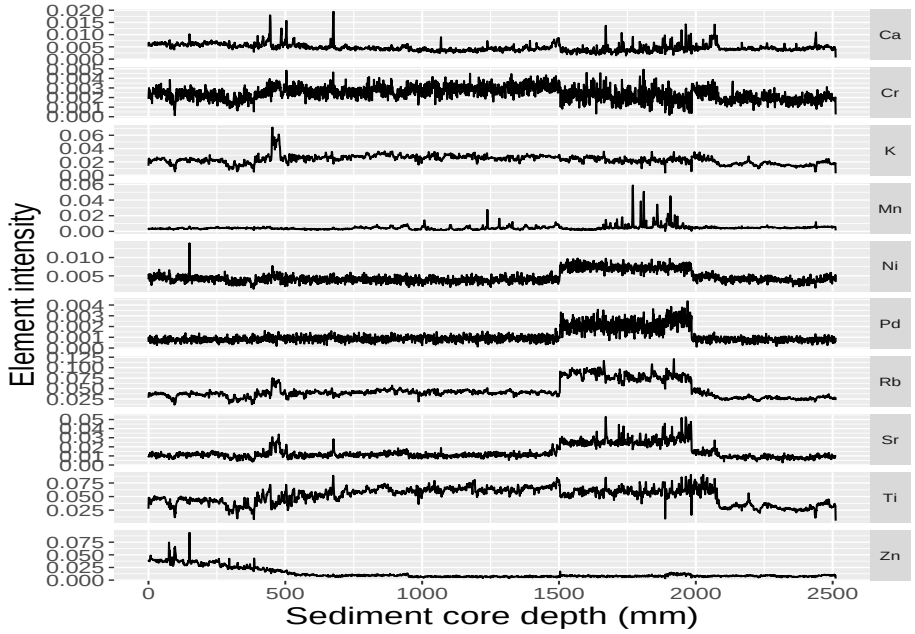


Figure 5.8: Sediment Core Data

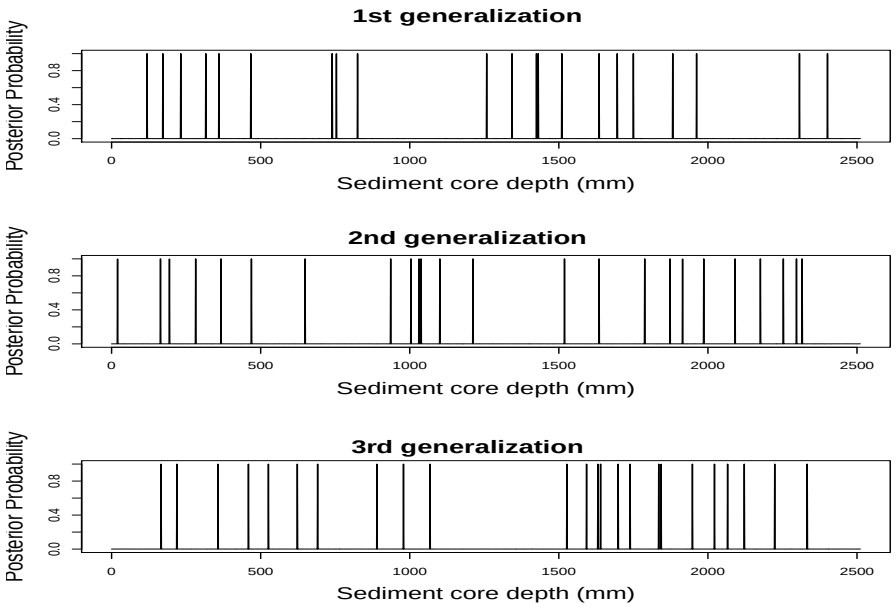


Figure 5.9: Posterior distributions of occurrence of change-point locations for the sediment data.

As noted above, the three generalizations produced similar results for the synthetic data example, but this was not the case for the real life data. It is not clear whether this difference is due to a lack of model fit, the larger number of sequences being segmented in parallel, or the presence of heteroscedasticity for some time series. I tested heteroscedasticity of sediment core data using Bartlett's test [31]. This statistic tests of the null that the variances in each of the samples are the same. This test has a p-value 0.002257614 less that a significance level of 0.05, therefore we can reject the null hypothesis that the variances in

each of the samples are constant and infer that heteroscedasticity is indeed present. The results of the three generalizations highlight the necessity of finding a way to decide which generalization is most appropriate.

A challenging statistical problem is to generate an effective tool for model comparison, especially for computationally demanding hierarchical models. In this example, our preferred model was selected with reference to three information criteria which were approximated using posterior samples. We have introduced these information criteria approximations elsewhere [32–35]. These approximations are based on MCMC sampled values. Three information criterion are used in this example, namely Deviance information criterion (DICV), and approximations to the Akaike information criterion (AIC) and Bayesian information criterion (BIC). The DICV is defined as: DICV=$p_v + \overline{D(\Theta)}$ where $\overline{D(\Theta)}$ is the mean posterior deviance, $p_v = Var(D(\Theta)/2)$ and the deviance is $\overline{D(\Theta)} = -2\ln f(\mathbf{X}|\Theta)$. The AIC approximation is defined as AIC=$2\overline{K} - 2\overline{\ln f(\mathbf{X}|\Theta)}$ where $\overline{K}$ is the average number of segments over the set of segmentations sampled by MCMC. The BIC approximation is defined as: BIC=$-2\overline{\ln f(\mathbf{X}|\Theta)} + \overline{K}\ln T$ where T is the total length of the signal. Further details of these information criteria are provided in [33–36]. The model with the lowest value of an information criterion is considered the preferred model. The computed values of these information criteria are presented in Table 5.1.

Table 5.1: Information criterions of three generalizations

Information Criterion	Generalizations		
	First	Second	Third
DICV	-87535.79	-88494.98	-88039.09
AIC	-92351.19	-92153.52	-92501.67
BIC	-92199.64	-91955.34	-92256.86

In light of these values, the third generalization appears preferable to the others according to AIC and BIC, although DICV favours the second generalization. To further investigate the suitability of three generalizations, we next consider how the change-point profiles (Figure 5.9) compare to known changes in the sediment core data.

All generalizations identify change-points between points 1500 and 2000. These correspond to a known problem in the element readings (complete temporal patterns were not obtained) [29]. These change-points stand out clearly in Figure 5.8. Moreover, all three generalizations identify a change-point close to time point 1000. This is a section where a part of the sediment core was hypothesised to have been lost in coring during collection

(due to stiffness in some sedimentary units) [29].

A boundary between pre-European and post-European sediment deposition had been previously hypothesized around the point 1000 [29]. This is plausible given the date of this time point-before 1870 but after the 1700s. This is also based on an analysis of the pollen species within the sediments, which shows a change in pollen from bush land to a surrounding catchment that has been increasingly cleared. It's encouraging that the second and the third generalization exhibit change-points here.

The sediments in the core change in the proportion of allochthonous (soil from outside the floodplain lake) to autochthonous (algae and particles created within the lake environment) deposition at around time-point 250-300. The first generalization identifies a change-point at point 233, the second generalization identifies one at point 283 and the third generalization identifies one at point 220. These change-points support the occurrence of a fundamental change occurring at around this time. All the change-point profiles plots identify some change-points around time point 358-370, indicating a sudden change in the sediment type. There is a change between high density sediment and low density sediments here, which is indicated by changes in other characteristics measured in the same core. Again it is encouraging that all three generalizations find change-points in this region of the core.

The region around 650-710 has been dated to the 1940s, when the types of sediments coming into the floodplain lake changed due to urban development in the catchment [29]. It is interesting that the second and the third generalization identify a change-point in this region whereas the first generalization does not.

There is a unique section (time points 450-480) of the core where there was distinct textural change in the sediments. This is the section of the core where clear changes occur (visually) in Figure 5.8. The elements in these sections were also different (due to the construction material deposited during the 1970s) [29]. All generalizations found change-points around 460-470. The change-point profiles also exhibit some change-points at time points 120-200. This may correspond to either a flood - or a drought - or both occurring in a short time.

These results demonstrate that the second and the third generalization find a greater number of change-points than the first generalization, which supports the conclusion drawn using the information criteria.

5.5 Conclusion

This paper proposes a general and flexible method for the segmentation of multiple sequences in parallel. We considered three alternative models, all of which are generalizations of the Bayesian change-point segmented ARMA model described in our earlier paper [23]. All generalizations are based on the posterior probability of having a common change-point. The second generalization makes an additional assumption that all time series have the same probability of assigning segments to a number of segment classes. The third generalization adds the further assumption that corresponding segments in each time-series belong to a common segment class.

The main advantage of using an ARMA model over an AR model is that it can consider the dependency between the residual terms of the model. This can help to take into account autocorrelation in the time series in a more flexible way, thus permitting the fitting of more flexible change-point models.

Our method samples from a varying dimensional space (since the number of change-points is unknown) using an efficient sampling technique called the Generalized Gibbs Sampler.

All generalizations produced similar and very encouraging results in a synthetic example. They successfully identified all change-point locations for a data set for which the segmentation of a single time series detected only some change-points. When applied to the real life example, the three generalizations produced different results and these alternative models are compared using approximated AIC, BIC and DICV values. The change-points identified in the real-life example were consistent with previous studies on the dataset [25, 26, 29, 30]. Based on these and other considerations, the fit of the third generalization was preferred. However, the accuracy of the segmentation of the real data is expected to be lower than for the synthetic example, due to the presence of heteroscedasticity in some time series, and other discrepancies between the model and the underlying process.

The three generalizations presented here work well when the change-points pertain to the mean of the signal in each segment, since the model assumes different segment means but constant variance within segments. This constant variance model assumption is also likely not valid for the real data set. Further work is needed to incorporate changes in the variance of the original process by allowing distinct variances for each class of the segments but constant variance with a class. Furthermore, these generalizations also assume the ARMA parameters are the same for all segments. There might also be change-points in the

ARMA parameters.

Acknowledgements

The authors thank Benjamin Goursaud and Rachel Créhange for helpful discussions. This work was funded by the Australian Research Council (http://www.arc.gov.au/) grant DP1095849. The authors are grateful to the Australian Research Council (ARC) and the ARC Centre of Excellence for Mathematical and Statistical Frontiers for their support of this project (DP1095849, CE140100049).

Data availability

Sediment core dataset used in Sect. 5.4.2 is available at https://doi.org/10.26180/5e16d5aadf912.

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. Badagián, A. L., Kaiser, R. & Peña, D. in *Empirical Economic and Financial Research* 45–59 (Springer, 2015).
2. Basseville, M., Nikiforov, I. V., *et al. Detection of abrupt changes: theory and application* (Prentice Hall Englewood Cliffs, 1993).
3. Algama, M. & Keith, J. M. Investigating genomic structure using changept: A Bayesian segmentation model. *Computational and structural biotechnology journal* **10,** 107–115 (2014).
4. Jensen, U. & Lütkebohmert, C. Change-point models. *Encyclopedia of Statistics in Quality and Reliability* **1** (2008).
5. Reeves, J., Chen, J., Wang, X. L., Lund, R. & Lu, Q. Q. A review and comparison of changepoint detection techniques for climate data. *Journal of Applied Meteorology and Climatology* **46,** 900–915 (2007).

6. Aminikhanghahi, S. & Cook, D. J. A survey of methods for time series change point detection. *Knowledge and information systems* **51,** 339–367 (2017).

7. Rodionov, S. A brief overview of the regime shift detection methods. *Large-scale disturbances (regime shifts) and recovery in aquatic ecosystems: challenges for management toward sustainability,* 17–24 (2005).

8. Truong, C., Oudre, L. & Vayatis, N. Selective review of offline change point detection methods. *Signal Processing,* 107299 (2019).

9. Picard, F., Lebarbier, É., Budinská, E. & Robin, S. Joint segmentation of multivariate Gaussian processes using mixed linear models. *Computational Statistics & Data Analysis* **55,** 1160–1170 (2011).

10. Ehsanzadeh, E., Ouarda, T. B. & Saley, H. M. A simultaneous analysis of gradual and abrupt changes in Canadian low streamflows. *Hydrological Processes* **25,** 727–739 (2011).

11. Dobigeon, N., Tourneret, J.-Y. & Davy, M. Joint segmentation of piecewise constant autoregressive processes by using a hierarchical model and a Bayesian sampling approach. *Signal Processing, IEEE Transactions on* **55,** 1251–1263 (2007).

12. Chamroukhi, F., Mohammed, S., Trabelsi, D., Oukhellou, L. & Amirat, Y. Joint segmentation of multivariate time series with hidden process regression for human activity recognition. *Neurocomputing* **120,** 633–644 (2013).

13. Zhang, N. R., Siegmund, D. O., Ji, H. & Li, J. Z. Detecting simultaneous changepoints in multiple sequences. *Biometrika* **97,** 631–645 (2010).

14. Cleynen, A. & Robin, S. Comparing change-point location in independent series. *Statistics and Computing* **26,** 263–276 (2016).

15. Collilieux, X., Lebarbier, E. & Robin, S. A factor model approach for the joint segmentation with between-series correlation. *Scandinavian Journal of Statistics* **46,** 686–705 (2019).

16. Harlé, F., Chatelain, F., Gouy-Pailler, C. & Achard, S. Bayesian Model for Multiple Change-points Detection in Multivariate Time Series. *arXiv preprint arXiv:1407.3206* (2014).

17. Rigaill, G., Lebarbier, É. & Robin, S. Exact posterior distributions and model selection criteria for multiple change-point detection problems. *Statistics and computing* **22,** 917–929 (2012).

18. Barigozzi, M., Cho, H. & Fryzlewicz, P. Simultaneous multiple change-point and factor analysis for high-dimensional time series. *Journal of Econometrics* **206,** 187–225 (2018).

19. Dobigeon, N., Tourneret, J.-Y. & Scargle, J. D. Joint segmentation of multivariate astronomical time series: Bayesian sampling with a hierarchical model. *IEEE Transactions on Signal Processing* **55,** 414–423 (2007).

20. Hoai, M., Lan, Z.-Z. & De la Torre, F. *Joint segmentation and classification of human actions in video* in *In: Proceedings of CVPR, IEEE.* (2011), 3265–3272.

21. Picard, F. *et al.* Joint segmentation, calling, and normalization of multiple CGH profiles. *Biostatistics* **12,** 413–428 (2011).

22. Vert, J.-P. & Bleakley, K. *Fast detection of multiple change-points shared by many signals using group LARS* in *Advances in neural information processing systems* (2010), 2343–2351.

23. Sadia, F., Boyd, S. & Keith, J. M. Bayesian change-point modeling with segmented ARMA model. *PloS one* **13,** e0208927 (2018).

24. Keith, J. M., Kroese, D. P. & Bryant, D. A generalized Markov sampler. *Methodology and Computing in Applied Probability* **6,** 29–53 (2004).

25. Lintern, A. *et al.* Identifying heavy metal levels in historical flood water deposits using sediment cores. *Water research* **105,** 34–46 (2016).

26. Lintern, A. *et al.* Sediment cores as archives of historical changes in floodplain lake hydrology. *Science of the Total Environment* **544,** 1008–1019 (2016).

27. Ferrand, E. *et al.* Historical levels of heavy metals and artificial radionuclides reconstructed from overbank sediment records in lower Rhône River (South-East France). *Geochimica et Cosmochimica Acta* **82,** 163–182 (2012).

28. Wolfe, B. B. *et al.* Reconstruction of multi-century flood histories from oxbow lake sediments, Peace-Athabasca Delta, Canada. *Hydrological Processes: An International Journal* **20,** 4131–4153 (2006).

29. Lintern, A., Deletic, A., Leahy, P. & McCarthy, D. Digging up the dirty past: evidence for stormwater's contribution to pollution of an urban floodplain lake. *Marine and Freshwater Research* **66,** 596–608 (2015).

30. Lintern, A. *et al.* Uncertainties in historical pollution data from sedimentary records from an Australian urban floodplain lake. *Journal of hydrology* **560,** 560–571 (2018).

31. Bartlett, M. S. Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences* **160,** 268–282 (1937).

32. Keith, J. M. Segmenting eukaryotic genomes with the generalized Gibbs sampler. *Journal of Computational Biology* **13,** 1369–1383 (2006).

33. Keith, J. M., Adams, P., Stephen, S. & Mattick, J. S. Delineating slowly and rapidly evolving fractions of the Drosophila genome. *Journal of Computational Biology* **15,** 407–430 (2008).

34. Oldmeadow, C., Mengersen, K., Mattick, J. S. & Keith, J. M. Multiple evolutionary rate classes in animal genome evolution. *Molecular biology and evolution* **27,** 942–953 (2009).

35. Oldmeadow, C. & Keith, J. M. Model selection in Bayesian segmentation of multiple DNA alignments. *Bioinformatics* **27,** 604–610 (2011).

36. Gelman, A. *et al. Bayesian data analysis* (Chapman and Hall/CRC, 2013).

Chapter 6

A Study of Models for Zero-inflated Time Series Data

Chapter Objectives

This chapter addresses a particular objective of this thesis which emerged during the data analysis of the previous chapter. The real world data I used in the previous chapter was sediment core data, which includes concentrations of thirty-seven metal elements detected at each depth (in mm) in the cores. However, not all the elements were used for simultaneous segmentation in the previous chapter, as some time series contained an excessive number of zeros. Zeros indicate that some elements were not detected at all depths in the core. These zeros motivated me to investigate the literature concerning zero-inflated time series data. In this chapter, I reviewed models for zero-inflated semicontinuous time series data. Generalization of the Bayesian change-point segmented ARMA model by incorporating an appropriate zero-inflated model will provide an interesting direction for future research.

Authorship

Farhana Sadia[1], Jonathan M. Keith[1]

1 School of Mathematics, Monash University, Clayton, VIC 3800, Australia

Reference

Sadia F, Keith JM. (2018). A Study of Models for Zero-inflated Time Series Data. Submitted. In Preparation.

6.1 Abstract:

In this chapter I consider variables with probability distributions that have continuous densities on the entire sample space with the exception of one value at which there is a positive point mass (such variables are said to be semicontinuous). Applications in which data are distributed with a right-skewed continuous positive density and a point mass at zero occur in many disciplines. When a variable can be modelled as having a standard distribution apart from a large proportion of zero values, this phenomenon is known as zero-inflation. In this chapter, I review proposed models for handling zero-inflated semicontinuous time series data.

6.2 Introduction

Time series data with excess zero values are often encountered by researchers in numerous fields. These data are characterized by a right-skewed continuous distribution of positive values with a high proportion of zeros. Because of the substantial proportion of zero observations, these data are referred to as semicontinuous data, point mass mixture data or zero-inflated data [1]. The unique features of this type of data make standard statistical methods, designed for continuously distributed data, either invalid or inaccurate. Failure to account for zero-inflation in the data may make the resulting estimators of quantities of interest and their standard errors biased and may result in misleading inferences [2, 3]. Furthermore, data transformation does not typically render the data any easier to analyse if excess zeros are present [4]. Consequently, excessive zero values need to be accounted for and require specialised methodology when performing analyses [5].

Semicontinuous time series data with excess zeros arise in many research areas such as epidemiology, social sciences, health, environment, economics, life sciences, engineering and others. For example, in a study of household expenditures on certain commodities during a period of time, it was found that the amount spent on some commodities follows a semicontinuous distribution because some households spend nothing on that commodity during the study period [6, 7]. Semicontinuous data also arise in studies involving cloud seeding data, where a zero indicates no rain and positive values indicate the amount of rainfall when there is rain [8]. As another example, Figure 6.1 shows the distribution of annual mental health expenditures among federal employees. More than 80% of the employees had zero annual expenditures, represented by a vertical line at zero, whereas the

other employees spent more than 1000 USD in total during the period of investigation [9, 10].

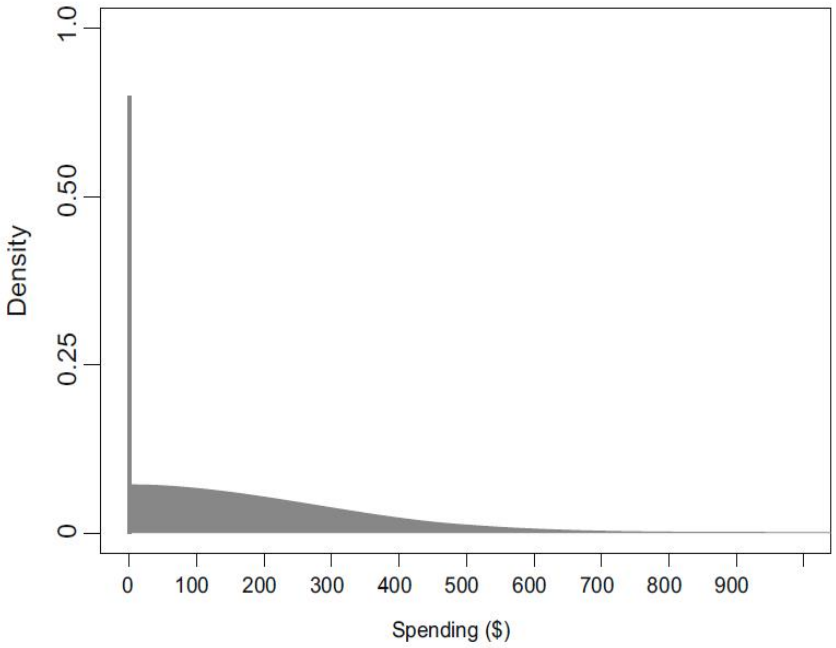


Figure 6.1: Distribution of annual mental health expenditures among federal employees

Zero values encountered in such data can be designated as true zeros, censored zeros or a mixture of true and censored zeros. A true zero is generally indicated by something not happening and a censored zero indicates its occurrence below a certain threshold [1]. A related phenomenon occurs in ecological studies of suitability of a given habitat for a particular species. In such studies, true zeroes correspond to sites where the species is genuinely absent, whereas censored zeros occur due to detection limits or observer effects (for example due to observers being more likely to visit and record sightings at some locations than others) [11, 12]. This chapter focuses on models of semicontinuous time series data with both types of zeros.

The difficulty with semicontinuous data analysis is that the large proportion of zero values makes conventional continuous probability distributions such as the normal, gamma or log-normal inappropriate for the analysis of zero-inflated data. Aitchison [13] first expressed concern about statistical analysis of zero-inflated data. Subsequently, several methodologies for such data have been developed, due to the increasing interest of dealing with zero-inflated data. In this article, I survey some methods developed for modeling zero-inflated time series data and discuss the main ideas behind each method.

6.3 Models for zero-inflated time series data

6.3.1 Delta distribution

Aitchison [13] proposed an approach to model a positive random variable by introducing a mixture consisting of a point mass at zero and a lognormal distribution or some other well known distribution with a continuous density. The resulting distribution was also referred to as a delta distribution. Mathematically, let X denote a random variable and α denote the probability that X is zero, Then the distribution of X conditional on $X \neq 0$ is some well-known distribution of a positive variable written as:

$$P\{X \subset (x + dx)|x > 0\} = f(x)dx, \text{and}$$
$$P\{X \subset (x + dx)\} = (1 - \alpha)f(x)dx, \quad x > 0.$$

where $f(x)$ is the conditional probability density given $x > 0$. If β and γ are the mean and variance respectively of $f(x)$ and θ and δ are the corresponding parameters of X, then

$$\theta = (1 - \alpha)\beta$$
$$\delta = (1 - \alpha)\gamma + \alpha(1 - \alpha)\beta^2.$$

Aitchison also developed estimators of the mean and the variance of the mixture distribution. He identified some general results to obtain the best unbiased estimators of θ and δ under certain conditions. Suppose a random sample S is drawn from the population at time $t = 1, 2, \cdots, T$ where a of the sample values are zero and $T - a$ samples $x_1, x_2, \cdots, x_{T-a}$ are non-zero. Aitchison considered a random sample with independent observations in his article [13].

- If a sufficient unbiased estimator of β, denoted $b_{(m)}$, exists for a sample of size m from the non-zero population, then

$$c = \begin{cases} \left(1 - \frac{a}{T}\right) b_{(T-a)}, & a < T \\ 0, & a = T \end{cases}$$

is a best unbiased estimator of θ. Here, $b_{(T-a)}$ is not defined for $a = T$, so the two-fold definition of c is needed. If $b_{(m)}$ is the arithmetic mean of m sample values, then c

becomes the mean of the sample S including zero values, which is,

$$c = \frac{1}{T}\sum_{t=1}^{T-a} x_t.$$

and the variance of c is

$$var(c) = \frac{\delta}{T}.$$

- If $p_{(m)}$ and $q_{(m)}$ are jointly sufficient unbiased estimators of β^2 and γ respectively for a sample of size m, then

$$d = \begin{cases} \left(1 - \frac{a}{T}\right) q_{(T-a)} + \frac{a}{T}\left(1 - \frac{a-1}{T-1}\right) p_{(T-a)} & \text{if } a < T \\ 0 & \text{if } a = T. \end{cases}$$

  is a best unbiased estimator of δ. Such jointly efficient estimators of β and γ occur seldomly. If γ depends on β so that $b_{(m)}$ is sufficient for both β and γ, with $\gamma = K\beta^2$, then the following property holds.

- If $\gamma = K\beta^2$ and $b_{(m)}$, the sufficient unbiased estimator of β, is the sample mean then

$$\delta = (1-\alpha)(K+\alpha)\beta^2.$$

  and

$$d = \begin{cases} \frac{\left\{K+(1-K)\frac{a}{n} - \frac{a(a-1)}{n(n-1)}\right\} b^2_{(n-a)}}{\left\{1+\frac{K}{n-a}\right\}}, & \text{if } a < T \\ 0, & \text{if } a = T \end{cases}$$

  is a best unbiased estimator of δ.

Advantages

This model is motivated by economic studies in which the data consists of a large number of zeroes, and the positive portion of the data fits very well to some well known distribution. Not only does this model have the ability to consider excess zeros, but also it provides the potential for modeling positive skewness in the nonzero observations. For instance, in health economics studies, a large proportion of the population may not incur any medical cost while relatively few gravely ill patients impose very high costs [6].

Drawbacks

Depending on the endpoint of the study, data analysis using the delta model may not be straightforward. For instance, a two-step procedure (separate modeling of zeros and non-zeros) may no longer be appropriate in many applications. In an air pollution example, for instance, the concentration level of air contaminant's at a given industrial site were modelled using a delta distribution [14]. But this approach didn't work for this example because the main interest of this example was in the overall mean contamination level, rather than in the mean of the non-zero values. Indeed, the focus of the inference is almost always on the mean of the entire mixture in most studies of this type [14].

6.3.2 Tobit models

A Tobit model is any of a class of regression models where the observed range of the dependent variable is censored in some way [15]. This model was first proposed by James Tobin in 1958 [16] to diminish the problem of zero-inflated data for observations of household expenditure on durable goods. The main idea of the Tobit model is to modify the likelihood function so that it estimates the unequal sampling probability for each observation conditioning on whether the latent dependent variable occurs above or below the determined threshold [17]. Here, the likelihood function is a combination of probability densities (for the observed or uncensored part of the distribution) and cumulative densities (for the censored part of the distribution) [18].

Mathematically, Let x_t denote the observation at time t, $t = 1, 2, \cdots, T$, of a random variable X. This model assumes that the random variable X can be expressed in terms of a latent variable X^* and can be defined as:

$$x_t = \begin{cases} x_t^*, & x_t^* > 0 \\ 0, & x_t^* \leq 0 \end{cases}$$

When, $x_t^* \leq 0$, its value is unobserved. The latent variable x_t^* can be expressed as a linear combination of a number of explanatory variables, that is,

$$x_t^* = \mathbf{g}_t' \boldsymbol{\beta} + \epsilon_t.$$

where, $\mathbf{g}_t'$ is a row vector containing the explanatory variables which are observable, $\boldsymbol{\beta}$ is a column vector containing the corresponding coefficients describing the linear dependency

of x_t^* on $\mathbf{g}_t$. The error terms ϵ_t are assumed to be independently and identically distributed and to follow a normal $N(0, \sigma^2)$ distribution.

The probability that X takes the value zero is given by

$$\begin{aligned} p(X_t = 0) = P(X_t^* \leq 0) &= P(\mathbf{g}_t'\boldsymbol{\beta} + \epsilon_t \leq 0) \\ &= P(\epsilon \leq -\mathbf{g}_t'\boldsymbol{\beta}) \\ &= P(\frac{\epsilon}{\sigma} \leq \frac{-\mathbf{g}_t'\boldsymbol{\beta}}{\sigma}) \\ &= \Phi\left(\frac{-\mathbf{g}_t'\boldsymbol{\beta}}{\sigma}\right) = 1 - \Phi\left(\frac{\mathbf{g}_t'\boldsymbol{\beta}}{\sigma}\right) \end{aligned}$$

where $\Phi(.)$ denotes the cumulative distribution function (cdf) of the $N(0, 1)$ distribution. The above equation is similar to the so-called Probit model. Probit model is a regression model which takes response variable with only two possible outcomes (0 and 1). In this model, the response variable can be expressed in terms of the latent variable as follows [18]:

$$x_t = \begin{cases} 1, & x_t^* > 0 \\ 0, & \text{otherwise} \end{cases}.$$

Probit model considers binary response with the following conditional Bernoulli probabilities:

$$\begin{aligned} P(X_t = 1|\mathbf{g}) &= \Phi(\mathbf{g}_t'\boldsymbol{\beta}) \\ P(X_t = 0|\mathbf{g}) &= 1 - \Phi(\mathbf{g}_t'\boldsymbol{\beta}). \end{aligned}$$

Tobit model is based on the above Probit model. Conditional on $x_t > 0$, the likelihood function L of the uncensored positive values of X is given by the probability density function of the latent variable X^*, that is,

$$f(x_t; \boldsymbol{\beta}, \sigma) = \sigma^{-1}\phi\left(\frac{x_t - \mathbf{g}_t'\boldsymbol{\beta}}{\sigma}\right)$$

where $\phi(.)$ denotes the probability distribution function of the $N(0, 1)$ distribution. Thus, the likelihood function for a sample of n independent observations is

$$l(\boldsymbol{\beta}, \sigma) = \left[\prod_{x_t=0}\left\{1 - \boldsymbol{\Phi}\left(\frac{\mathbf{g}_t'\boldsymbol{\beta}}{\sigma}\right)\right\}\right]\left[\prod_{x_t>0}\sigma^{-1}\phi\left(\frac{x_t - \mathbf{g}_t'\boldsymbol{\beta}}{\sigma}\right)\right]$$

Tobin estimated the maximum likelihood (ML) estimates of $\boldsymbol{\beta}$ and σ using a Newton-Raphson algorithm. This model assumes a normal distribution with constant variance for the error term but it is unrealistic in many applications. The ML estimators becomes inconsistent when the distribution of ϵ_t is not normal [19]. Powell [20] suggested semi-parametric estimation for the Tobit model by using a symmetrically trimmed least squares (STLS) estimator with an assumption that the ϵ_t are symmetrically distributed about zero. The STLS estimator is defined as [21]:

$$\hat{\boldsymbol{\beta}}_{STLS} = \arg\min_{\boldsymbol{\beta}} \sum_{t=1}^{T} I(\mathbf{g}_t'\boldsymbol{\beta} > 0)[\min(x_t, 2\mathbf{g}_t'\boldsymbol{\beta}) - \mathbf{g}_t'\boldsymbol{\beta}]^2$$

where I is the indicator function. The sum in this expression eliminates the observations for a given $\boldsymbol{\beta}$ with $\mathbf{g}_t'\boldsymbol{\beta} \leq 0$. The lower tail of the distribution of X_t is censored at zero for $\mathbf{g}_t'\boldsymbol{\beta} > 0$; symmetrically censoring the upper tail of the distribution (essentially by interchanging x_t by $\min\{x_t, 2\mathbf{g}_t'\boldsymbol{\beta}\}$) restores the symmetry of distribution of X^*. The final estimator $\hat{\boldsymbol{\beta}}_{STLS}$ is consistent and asymptotically normal under the symmetrical distribution assumption [20]. $\hat{\boldsymbol{\beta}}_{STLS}$ is obtained by an iterative procedure.

To estimate the covariance matrix of $\hat{\boldsymbol{\beta}}_{STLS}$, Yoo *et al.* [22] used a method involving the bootstrap. For M bootstrap replications with estimate $\hat{\boldsymbol{\beta}}_j$ in replication j, their estimate is

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{M} \sum_{j=1}^{M} (\hat{\boldsymbol{\beta}}_j - \overline{\boldsymbol{\beta}}_{STLS})(\hat{\boldsymbol{\beta}}_j - \overline{\boldsymbol{\beta}}_{STLS})'$$

where $\overline{\boldsymbol{\beta}}_{STLS} = (1/M) \sum_{j=1}^{M} \hat{\boldsymbol{\beta}}_j$.

Advantages

The Tobit model is suitable to model a truncated or censored response variable, as it assumes an underlying normal random variable that is censored by a random mechanism. Another advantage is that the probability of a zero observation depends on the same random variable that determines the magnitude of the observation given that it is positive.

Drawbacks

Although the Tobit model is sometimes used for semicontinuous data, it is not appropriate for "true zeroes" as defined above because zeros in this model do not represent actual responses. The underlying normal assumption becomes dubious if the zeros in the outcome variable are true zeros.

6.3.3 Sample selection models

The sample selection model is an extension of the Tobit model that was first proposed by J. Heckman in 1979 [23]. It is a statistical model to fix bias from non-randomly selected samples or otherwise incidentally truncated dependent variables. This is achieved by explicitly modelling the individual sampling probability of each observation along with the conditional expectation of the dependent variable. The likelihood function is mathematically similar to the Tobit model for censored dependent variables [23]. Many variants of sample selection models are available but here the version of Ven *et al.* [24] is illustrated. This model is based on two latent variables X_1^* and X_2^* which are defined as:

$$x_{1t}^* = \mathbf{g}_{1t}'\boldsymbol{\beta}_1 + \epsilon_{1t},$$
$$x_{2t}^* = \mathbf{g}_{2t}'\boldsymbol{\beta}_2 + \epsilon_{2t},$$

Sample selection models thus allow for the latent variables to depend on different covariates. Now, the observed variable is defined as:

$$x_t = \begin{cases} \exp(x_{2t}^*) & \text{if } x_{1t}^* > 0, \\ 0 & \text{if } x_{1t}^* \leq 0 \end{cases}$$

The error terms ϵ_1 and ϵ_2 are assumed to be independent of the regressors $\mathbf{g}_1$ and $\mathbf{g}_2$, and follow a bivariate normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$ distribution, where

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$$

If $x_{1t}^* > 0$, $x_t > 0$ is observed and $x_{2t}^* = \log(x_t)$, whereas if $x_{1t}^* \leq 0$, $x_t = 0$ is observed and x_{2t}^* is missing. The covariates and parameter vectors $(\mathbf{g}_{1t}', \boldsymbol{\beta}_1)$ for x_{1t}^* may differ from $(\mathbf{g}_{2t}', \boldsymbol{\beta}_2)$ for x_{2t}^*. Heckman used two alternative estimation methods with this model, specifically ML and a two-step procedure. For ML estimation, the likelihood function of the model is given by [21]

$$
\begin{aligned}
l(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\Sigma}) &= \left[\prod_{x_t=0} P(x_{1t}^* \leq 0)\right]\left[\prod_{x_t>0} f(x_t^*|x_{1t}^* > 0)P(x_{1t}^* > 0)\right] \\
&= \left[\prod_{x_t=0} P(x_{1t}^* \leq 0)\right]\left[\prod_{x_t>0} \int_0^\infty f(x_t^*, x_{1t}^*)dx_{1t}^*\right] \\
&= \left[\prod_{x_t=0} \left\{1 - \Phi\left(\frac{\mathbf{g}_{1t}'\boldsymbol{\beta}_1}{\sigma_1}\right)\right\}\right] \\
&\times \left[\prod_{x_t>0} \Phi\left\{\left(\frac{\mathbf{g}_{1t}'\boldsymbol{\beta}_1}{\sigma_1} + \frac{\log(x_t) - \mathbf{g}_{2t}'\boldsymbol{\beta}_2}{\sigma_{12}^{-1}\sigma_1\sigma_2^2}\right) \times (1 - \sigma_{12}^2\sigma_1^{-2}\sigma_2^{-2})^{-\frac{1}{2}}\right\} \sigma_2^{-1}\phi\left(\frac{\log(x_t) - \mathbf{g}_{2t}'\boldsymbol{\beta}_2}{\sigma_2}\right)\right]
\end{aligned}
$$

ML estimates are found by an iterative method.

Heckman's two-step procedure is very simple and easy to implement but it does not work as well as the ML estimators. With the two-step procedure, the subsample regression function for X_t^* is

$$
E[X_t^*|x_{1t}^* > 0] = \mathbf{g}_{2t}'\boldsymbol{\beta}_2 + E[\epsilon_{2t}|\epsilon_{1t} > -\mathbf{g}_{1t}'\boldsymbol{\beta}_1] = \mathbf{g}_{2t}'\boldsymbol{\beta}_2 + \frac{\sigma_{12}}{\sigma_1}\lambda_t. \tag{6.1}
$$

where $\lambda_t = \phi(z_t)/\Phi(z_t)$, and $z_t = \mathbf{g}_{1t}'\boldsymbol{\beta}_1/\sigma_1$. Then,

$$
\begin{aligned}
\log(X_t) &= E[X_t^*|\mathbf{g}_{2t}, X_1^* > 0] + \epsilon_t \\
&= \mathbf{g}_{2t}'\boldsymbol{\beta}_2 + \frac{\sigma_{12}}{\sigma_1}\lambda_t + \epsilon_t
\end{aligned}
$$

where Heckman (1979) showed that ϵ_t has mean 0 and variance $\sigma_2^2[(1-\rho^2)+\rho^2(1+z_t\lambda_t-\lambda_t^2)]$ with $\rho^2 = \sigma_{12}^2/(\sigma_1^2\sigma_2^2)$. The parameters $\boldsymbol{\beta}_1$ and σ_1 can be estimated by a Probit model using the full sample. Therefore, z_t and hence λ_t can be easily estimated. In Equation 6.1, the estimated value of λ_t is used as a regressor and $\boldsymbol{\beta}_2$ can be estimated using least squares.

Advantages

The sample-selection model is also suitable for censored data as it is an extension of the Tobit model. But in the Tobit model, the latent variable performs both the role of determining whether the data is censored and the outcome of interest. The sample selection model allows the process of participation (selection) and the outcome of interest to be independent, conditional on observable data.

Drawbacks

The sample selection model has some disadvantages:

- Like the Tobit model, the sample selection model also assumes truncation or censoring in the outcome measure, therefore is not appropriate for semicontinuous data.
- The two-step estimator discussed above is a limited information maximum likelihood (LIML) estimator. The full information (FIML) estimator displays better statistical properties in asymptotic theory and in finite samples as demonstrated by Monte Carlo simulations [25].
- The covariance matrix generated by OLS (a method for estimating the unknown parameters in a linear regression model by minimizing the sum of the squares of the differences between the observed dependent variable in a model and those predicted by the linear function) estimation of the second stage is inconsistent [26].
- The estimator is generally inconsistent if the assumption that the errors are jointly normally distributed fails, and can give confusing inference in small samples [27].
- Duan [28] and Duan et al. [29–31] discussed the strongest criticisms against the sample selection model. They argued that the selection models are intrinsically limited as they depend on untestable assumptions and have poor statistical and numerical properties and therefore may be inappropriate for any applications including either conditional (actual) or unconditional (potential) outcomes.

6.3.4 Cragg's double hurdle model

The following model was proposed by Cragg [32] and is also based on two latent variables X_1^* and X_2^*. The idea of Cragg's double hurdle model is exemplified for a data set that concerns cigarette consumption. Let $\mathbf{X}$ denote the amount of cigarette consumed by an individual during a certain period of time. The first latent variable X_1^* ascertains whether an individual is a smoker or non-smoker and it depends on some socioeconomic factors which can be accounted for by the dependency of X_1^* on $\mathbf{g}_1$. The second latent variable X_2^* ascertains how much cigarette is consumed by an individual if the individual is a smoker and it may depend on other covariates in addition to those that influenced the probability of the individual being a smoker in the first place. During the investigation period, it is possible for a smoker not to consume any cigarette, that is, $X_2^* \leq 0 | X_1^* > 0$. To observe

positive values of $\mathbf{X}$, two hurdles need to be overcome: one is the individual must be a smoker and the other is the individual has to smoke during the period of the study. Hence this model is named the double hurdle model. In this model, the observed variable can be defined as:

$$X = \begin{cases} X_2^*, & X_1^* > 0 \text{ and } X_2^* > 0 \\ 0, & \text{otherwise} \end{cases}.$$

Both hurdles are assumed to be linear in the parameters $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$, with additive error terms ϵ_1 and ϵ_2 randomly distributed with a bivariate normal distribution, where $\mathbf{g}_1$ and $\mathbf{g}_2$ are the regressors that influence participation and consumption [33].

- **Observed variable (consumption)**: $x_t = dx_{2t}^*$.
- **Participation equation**:

$$\begin{aligned} x_{1t}^* = \mathbf{g}_{1t}'\boldsymbol{\beta}_1 + \epsilon_{1t}, \quad d &= 1 \text{ if } x_{1t}^* > 0 \\ &= 0 \text{ otherwise.} \end{aligned}$$

- **Consumption equation**:

$$\begin{aligned} x_{2t}^{**} &= \max[0, x_{2t}^*] \\ x_{2t}^* &= \mathbf{g}_{2t}'\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}_{2t}. \end{aligned}$$

To compute the likelihood of the model, the sample is divided into those with zero consumption (denoted $x_t = 0$) and those with positive consumption (denoted $x_t > 0$). Then the likelihood for Cragg's double-hurdle model considering dependence between ϵ_1 and ϵ_2 is:

$$\begin{aligned} L &= \prod_{x_t=0} [1 - p(d=1)p(x_2^* > 0 | d=1)] \prod_{x_t>0} p(d=1)p(x_{2t}^* > 0 | d=1) f(x_{2t}^* | x_{2t}^* > 0, d=1) \\ &= \prod_{x_t=0} [1 - p(\epsilon_{1t} > -\mathbf{g}_{1t}'\boldsymbol{\beta}_1) p(\epsilon_{2t} > -\mathbf{g}_{2t}'\boldsymbol{\beta}_2 | \epsilon_{1t} > -\mathbf{g}_{1t}'\boldsymbol{\beta}_1)] \prod_{x_t>0} p(\epsilon_{1t} > -\mathbf{g}_{1t}'\boldsymbol{\beta}_1) \\ &\qquad p(\epsilon_{2t} > -\mathbf{g}_{2t}'\boldsymbol{\beta}_2 | \epsilon_{1t} > -\mathbf{g}_{1t}'\boldsymbol{\beta}_1) f(x | \epsilon_{2t} > -\mathbf{g}_{2t}'\boldsymbol{\beta}_2, \epsilon_{1t} > -\mathbf{g}_{1t}'\boldsymbol{\beta}_1). \end{aligned}$$

The above expression includes the density and distribution functions of the truncated

bivariate normal distribution [33]. If ϵ_1 and ϵ_2 are assumed independent [34], then the above likelihood reduces to:

$$
\begin{aligned}
L &= \prod_{x_t=0}[1-p(\epsilon_{1t}>-\mathbf{g}_{1t}'\boldsymbol{\beta}_1)p(\epsilon_{2t}>-\mathbf{g}_{2t}'\boldsymbol{\beta}_2]\prod_{x_t>0}p(\epsilon_{1t}>-\mathbf{g}_{1t}'\boldsymbol{\beta}_1)p(\epsilon_{2t}>-\mathbf{g}_{2t}'\boldsymbol{\beta}_2)f(x|\epsilon_{2t}>-\mathbf{g}_{2t}'\boldsymbol{\beta}_2) \\
&= \prod_{x_t=0}\left[1-\Phi\left(\frac{\mathbf{g}_{1t}'\boldsymbol{\beta}_1}{\sigma_1}\right)\Phi\left(\frac{\mathbf{g}_{2t}'\boldsymbol{\beta}_2}{\sigma_2}\right)\right]\prod_{x_t>0}\Phi\left(\frac{\mathbf{g}_{1t}'\boldsymbol{\beta}_1}{\sigma_1}\right)\Phi\left(\frac{\mathbf{g}_{2t}'\boldsymbol{\beta}_2}{\sigma_2}\right)\times \\
&\qquad \prod_{x_t>0}\left[\left(\frac{1}{\sigma_2}\right)\phi\left(\frac{x_t-\mathbf{g}_{2t}'\boldsymbol{\beta}_2}{\sigma_2}\right)\Phi\left(\frac{\mathbf{g}_{2t}'\boldsymbol{\beta}_2}{\sigma_2}\right)\right].
\end{aligned}
$$

where, $\phi(.)$ and $\Phi(.)$ are the standard normal density and distribution functions.

Advantages

Cragg's double hurdle model generalises both the Tobit model and the sample selection model and hence is more flexible than the other two models. The Tobit model assumes that the participation and consumption decision can be modelled as one equation whereas Cragg's model eases this assumption and models both decisions separately. In the sample selection model, zero observations arise due to nonparticipation solely whereas Cragg's model eases this assumption and allows zero observations to arise in both the participation hurdle and consumption hurdle [35].

The difference between the sample selection model and the double hurdle model is that the sample selection model assumes that in the second stage there will be no zero observations, whereas the double hurdle model allows a possibility of a zero observation, which may arise from the individual's choice or random circumstances.

Drawbacks

A limitation of Cragg's double hurdle model is that it depends on the assumption of bivariate normality of the error terms. If the normality assumption is violated, then the maximum likelihood estimates of the model will not be consistent [36].

6.3.5 Two-part models

Zero-inflated log-normal two-part models

Two-part models incorporate two distinct stochastic process: one determining the occurrence of zero values and the other determining the distribution of the continuous non-zero values. Models of this type consider zero values as true observed zeros. Mathematically,

suppose that X_t is the semicontinuous response variable at time t; g_t is a vector of covariates at time t; β is the parameters used to model the probability of positive responses and θ presents mean and dispersion parameters of the conditional distribution of the positive responses. This model uses the same covariates (g_t) in both parts of the model. The probability of a positive response is $p_\beta = P(X_t > 0|g_t)$ and the conditional distribution of the positive responses has density $f_\theta(X_t|X_t > 0)$. Then an indicator function $\mathbf{J}(X_t > 0)$ is defined, that is, $\mathbf{J}(X_t > 0) = 1$ if $X_t > 0$ and $\mathbf{J}(X_t = 0) = 0$, otherwise.

Duan *et al.* [29] developed the two-part model without assuming an underlying normal distribution. This model consists of two parts described by two equations. The first part determines whether the response outcome is positive and is a binary model for the dichotomous event of having zero or positive values, such as the logistic regression model, that is,

$$\text{logit}[P(X_t = 0)] = \mathbf{g}'_{1t}\boldsymbol{\beta}_1. \tag{6.2}$$

The second part determines the level of response conditional on its being positive and assumes a log-normal distribution.

$$\log(x_t|x_t > 0) = \mathbf{g}'_{2t}\boldsymbol{\beta}_2 + \epsilon_t.$$

where, $\epsilon_t \sim N(0, \sigma^2)$. Then the likelihood function of the model is [21]

$$\begin{aligned} L(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma) &= \left[\prod_{x_t=0} p(x_t = 0)\right]\left[\prod_{x_t>0} p(x_t > 0) f(x_t|x_t > 0)\right] \\ &= \left[\prod_{x_t=0} \frac{e^{\mathbf{g}'_{1t}\boldsymbol{\beta}_1}}{1 + e^{\mathbf{g}'_{1t}\boldsymbol{\beta}_1}}\right]\left[\prod_{x_t>0} \frac{1}{1 + e^{\mathbf{g}'_{1t}\boldsymbol{\beta}_1}} \sigma^{-1}\phi\left(\frac{\log x_t - \mathbf{g}'_{2t}\boldsymbol{\beta}_2}{\sigma}\right)\right]. \end{aligned}$$

The maximum likelihood estimators of the above likelihood function are relatively easy to determine as the likelihood function factors into two components. The first component has only the logit model parameter,

$$L_1(\boldsymbol{\beta}_1) = \left[\prod_{x_t=0} e^{\mathbf{g}'_{1t}\boldsymbol{\beta}_1}\right]\left[\prod_{t=1}^{T} \frac{1}{1 + e^{\mathbf{g}'_{1t}\boldsymbol{\beta}_1}}\right].$$

The second component involves only the parameters of the second model part

$$L_2(\boldsymbol{\beta}_2, \sigma) = \prod_{x_t>0} \sigma^{-1}\phi\left(\frac{\log x_t - \mathbf{g}'_{2t}\boldsymbol{\beta}_2}{\sigma}\right).$$

The maximum likelihood estimates can be obtained by separately maximizing the two components.

Zero-inflated gamma two-part models

This model differs from the previous model only in that the non-zero part assumes a gamma distribution. Under this regression framework, the first part of the model follows Equation 6.2 and in the second part, X_t is modeled using gamma regression with a log link where $X_t \sim \Gamma(\exp(\mathbf{g}'_{2t}\boldsymbol{\beta}_2), \nu)$. Here, $\nu = \text{CoV}^2$, and CoV is the coefficient of variation. Combining these two pieces of the model gives a zero-inflated gamma distributions (ZIG) [1].
Here, $p(x_t = 0)$ follows Equation 6.2 and $f(x_t|x_t > 0; \mathbf{g}'_{2t}, \boldsymbol{\beta_2}, \nu^{-1}) = \frac{1}{\Gamma(\nu)}\left(\frac{\nu}{\mu_t}\right)^{\nu} x_t^{\nu-1}\exp\left(\frac{-\nu x_t}{\mu_t}\right)$, with μ_t modeled as $\log(\mu_t) = \mathbf{g}'_{2t}\boldsymbol{\beta}_2$. The likelihood is:

$$\begin{aligned} L(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \nu) &= \left[\prod_{x_t=0} p(x_t = 0)\right]\left[\prod_{x_t>0} p(x_t > 0) f(x_t|x_t > 0)\right] \\ &= \left[\prod_{x_t=0} \frac{e^{\mathbf{g}'_{1t}\boldsymbol{\beta}_1}}{1 + e^{\mathbf{g}'_{1t}\boldsymbol{\beta}_1}} \prod_{x_t>0} \frac{1}{1 + e^{\mathbf{g}'_{1t}\boldsymbol{\beta}_1}}\right]\left[\prod_{x_t>0} \frac{1}{\Gamma(\nu)}\left(\frac{\nu}{e^{\mathbf{g}'_{2t}\boldsymbol{\beta}_2}}\right)^{\nu} x_t^{\nu-1}\exp\left(\frac{-\nu x_t}{e^{\mathbf{g}'_{2t}\boldsymbol{\beta}_2}}\right)\right]. \end{aligned}$$

Maximizing two parts of the likelihood separately will also maximize the overall likelihood. This maximization can be performed via a Newton-Raphson algorithm for each part.

Advantages

The two-part model has several attractive properties:

- When the zeros are true zeros, this model provides a well-behaved likelihood function and more appropriate interpretations than the Tobit and sample selection models.
- These models do not assume an underlying normal distribution, hence can be used for a wider range of data than Tobit models.
- These models don't have any latent variable and do not involve censoring [37]. Consequently, two-part models are more appropriate than the other models discussed above

for modelling semicontinuous data.

- These models are very flexible in the sense that the covariates can be included in the zero and nonzero parts of the model using conventional generalized linear modelling techniques [11].
- These models don't require the addition of a constant which can introduce a bias [11].
- These models provide a flexible parametric approach for analysis of zero-inflated semi-continuous data. Such flexibility can produce improved model fit over traditional one-part models in many cases [9].

Drawbacks

Two-part models rely on parametric assumptions, which can be a liability. Erroneous assumptions about the response distribution can naturally lead to misleading inferences. Specifically, to achieve unbiased parameter estimates in any regression analysis, careful attention should be given to modeling assumptions [9].

6.3.6 Exponential dispersion models/Tweedie models

Instead of handling zero and non-zero values separately, Jorgensen [38] developed another approach which uses a positive distribution that simultaneously includes zeros and positive quantities, known as exponential dispersion models for generalized linear models with a power variance function. These models are also known as Tweedie distributions or models, and can handle zero-inflated data without treating the zero and nonzero values separately.

Any exponential dispersion model can be characterized by its variance function $V(.)$. This function determines the mean-variance relationship of the distribution when the dispersion is held constant. If X follows an exponential dispersion model distribution with mean μ, variance function $V(.)$ and dispersion ϕ, then the variance of X can be written as:

$$V(X) = \phi V(\mu).$$

Tweedie distributions are a special case of the exponential dispersion family for which $V(\mu) = \mu^p$ and $V(X) = \phi\mu^p$ [31]. The distribution is defined for all values of p except values in the open interval (0, 1). Many conventional known distributions such as normal ($p = 0$), Poisson ($p = 1$), gamma ($p = 2$), and inverse Gaussian ($p = 3$) are a special case of

Tweedie distributions. The probability density function (pdf) of the Tweedie distribution does not have any analytical expression except in these special cases. For $p > 1$, the pdf is:

$$f(x;\mu,\phi,p) = c(x,\phi)\exp\left[\frac{1}{\phi}\left(\frac{x\mu^{1-p}}{1-p} - \kappa(\mu,p)\right)\right].$$

where, $\kappa(\mu,p) = \frac{\mu^{2-p}}{2-p}$ for $p \neq 2$ and $\kappa(\mu,p) = \log(\mu)$ for $p = 2$. The function $c(x,\phi)$ doesn't have an analytic expression in general, and is usually estimated by using the series expansion methods that are described in Dunn and Smyth [31] as it doesn't have an analytical expression. The Tweedie distribution is a compound Poisson-gamma mixture distribution for $1 < p < 2$, which is the distribution of Z defined as:

$$Z(t) = \begin{cases} \sum_{i=1}^{N_t} X_i, & N_t = 1,2,3,\cdots \\ 0, & N_t = 0 \end{cases}$$

where $N_t = \sum_{t\geq 1} 1_{[t,\infty)}(t) \sim \text{Poisson}(\lambda)$, $X_i \sim \text{gamma}(\alpha,\theta)$ are independently and identically distributed gamma random variables with the shape parameter α and the scale parameter θ and the probability density function of X_i is [39]:

$$f(X) = \begin{cases} \frac{\alpha^\theta X^{\theta-1}e^{-\alpha X}}{\Gamma(\theta)} & \text{if } X > 0, \\ 0 & \text{if } X \leq 0. \end{cases}$$

The density of Z is governed by a Poisson distribution at $X = 0$, that is, $P(Z = 0) = \exp(-\lambda) = p_0 = 1 - q_0$. For $X > 0$, the probability density function of Z is a mixture of gamma variates with Poisson mixing probability which is:

$$P(Z > 0) = \sum_{i=1}^{\infty} \frac{p_i}{q_0}\left(\frac{\alpha^{t\theta}Z^{i\theta-1}e^{-\alpha Z}}{\Gamma(i\theta)}\right) = \frac{Z^{-1}e^{-\alpha Z}}{e^\lambda - 1} r_\theta(vZ^\theta).$$

where $p_i = \frac{e^{-\lambda}\lambda^i}{i!}$, $v = \lambda\alpha^\theta$, and $r_\theta(vZ^\theta) = \text{Poisson mixing probability} = \sum_{i=1}^{\infty} \frac{vZ^\theta}{i!\Gamma(i\theta)}$ [39, 40].

The parameters λ, α, and θ are related to the natural parameters μ, ϕ and p of the Tweedie distribution as follows

$$\lambda = \frac{\mu^{2-p}}{\phi(2-p)}$$
$$\alpha = \frac{2-p}{p-1}$$
$$\theta = \phi(p-1)\mu^{p-1}$$

The mean of a Tweedie distribution is positive for $p > 1$. For the mean response, a model can be specified as $f(\mu_t) = \mathbf{g}'\boldsymbol{\beta}$ with link function f. The maximum likelihood estimator for $\boldsymbol{\beta}$ doesn't involve $c(x_t, \phi)$. This model can be fitted with software for generalized linear models when p is known. Generally, p is unknown and needs to be estimated, but the estimation can be intricate as it involves (through α) an infinite sum and a gamma function in $c(x_t, \phi)$ (Jorgensen [38]). As an alternative, moment based estimation may perform well. Tweedie [41] proposed an estimate of p based on a single random sample as $\hat{p} = \hat{k}_1\hat{k}_3\hat{k}_2^{-2}$, where $\hat{k}_i$ is an estimate of cumulant i of the distribution. Jorgensen [38] developed a possible generalization of this approach for a regression model [21]. Let, $\mathbf{x}$ and $\hat{\mu}$ be the vectors of observations and fitted values, respectively. A moment estimator for ϕ is $\hat{\phi} = \mathbf{g}^2/n - k$, where k is the number of unknown parameters and $\mathbf{g}^2 = (\mathbf{x} - \hat{\boldsymbol{\mu}})^T V(\hat{\boldsymbol{\mu}})^{-1}(\mathbf{x} - \hat{\boldsymbol{\mu}})$ [21].

Advantages

The exponential dispersion model is relatively simple in the sense that it analyzes data with a single model including both aspects described in the two-part model. This model is easy to fit when the power p of the variance function is given.

Drawbacks

This model seems problematic when the power of the variance function is not known. Moreover, this model does not seem to have received attention in practice other than in Jorgensen's work.

6.3.7 Threshold model

Saei *et al.* [42] suggested recoding the continuous response into an ordinal scale by grouping the positive values into intervals and then developed a threshold model to analyze outcomes of a methadone randomized controlled trial and to relate the ordinal outcome variable to covariates. This model groups the possible outcome values into k ordered categories. Let X_t be the grouped response variable. The threshold model for an ordinal response involves an

unobserved continuous random variable Z, such that $Z+\mathbf{g}'\boldsymbol{\beta}$ has the cumulative distribution function G. If Z stays in the interval $\theta_{j-1} < Z \leq \theta_j$, then response $X_t = j, \quad j \in 1, 2 \cdots, k$ is observed, where θ_j are threshold parameters [43]. Then,

$$P(X_t \leq j|\mathbf{g}) = P(Z \leq \theta_j|\mathbf{g}) = P(Z + \mathbf{g}'\boldsymbol{\beta} \leq \theta_j + \mathbf{g}'\boldsymbol{\beta}) = G(\theta_j + \mathbf{g}'\boldsymbol{\beta}).$$

The threshold model is then written as:

$$G^{-1}[P(X_t \leq j; \mathbf{g})] = \theta_j + \mathbf{g}'\boldsymbol{\beta}, \quad j = 1, 2, \cdots, k-1.$$

The link function is the inverse of the cdf. The applications of semicontinuous data with a point mass at zero consider the first category to be the zero outcome and select cutpoints or thresholds on the positive outcome scale to define the other $k-1$ categories. If G is assumed to be a logistic function, this leads to a logit model for the cumulative probabilities, called a cumulative logit model [43]. Then the threshold model can be expressed as:

$$P(X_t \leq j|\mathbf{g}) = \frac{\exp(\theta_j + \mathbf{g}'\boldsymbol{\beta})}{1 + \exp(\theta_j + \mathbf{g}'\boldsymbol{\beta})}.$$

McCullagh [44] assume that G is normal, resulting in a cumulative Probit model. This approach assumes same covariate effects (the relationship between the cumulative probabilities of the ordinal categories of the outcome variable and the covariates) for each category of the outcome variable, known as proportional odds assumption. Score test or Wald test are used to check this important assumption of this approach [43, 45]. Chang and Pocock [43] used the cumulative logit model for a data set concerning the amount of personal care for the elderly.

Advantages

This model is conceptually simple in that it uses a single model to deal with zero-inflated continuous data and hence is generally easy to fit. Elements of $\boldsymbol{\beta}$ summarize overall effects, rather than conditional on the response being positive. To make a comparison between different groups with different levels of the explanatory variables, $\boldsymbol{\beta}$ can be used directly. This is in contrast to two-part models, in which one needs to average the results from the two components of the model to make an unconditional comparison (e.g., to estimate $E(X)$ for the groups).

Drawbacks

One disadvantage of this model is that it can lose information about the data because of grouping. Moreover, it uses an arbitrary way of collapsing the positive scales into categories.

6.4 Summary and Future directions

One difficulty statisticians often confront is in analyzing data that have a substantial proportion of zero values. Zero-inflated data need special statistical methodology to avoid biases and inappropriate decisions. The data set analysed in Chapter 5 is part of a larger data set that includes time-series with zero-inflation. The data was sediment core data including concentrations of metal elements detected at each depth (in mm) in the cores. Different elements indicate different characteristics of the sediments such as magnetic susceptibility, organic matter, sediment particle size and elemental composition of the sediments. Zeros in the time series indicate that the elements were not detected at all depths in the sediment core. In that chapter, the time series containing zeros were ignored, as the approach was not designed for zero-inflated data. To handle this type of data for segmentation purposes, the Bayesian change-point segmented ARMA model needs to be generalized. On that account, I surveyed the available models for zero-inflated time series data. Each model above is discussed with its benefits and drawbacks. To the best of my knowledge, these models have not previously been applied to change-point problems. A summary of the reviewed models is listed in Table 6.1.

Table 6.1: Summary of the reviewed models

Models	True zero	Censored zero	Two separate model	Single model	Normality assumption
Delta distribution	✓		✓		
Tobit models		✓	✓		✓
Sample selection models		✓	✓		✓
Cregg's double hurdle model		✓	✓		✓
Two-part models	✓		✓		
Exponential dispersion models	✓			✓	
Threshold model	✓			✓	

The appropriate choice of model depends on the nature of the data used in change-point problems. The major difference between these models is whether they treat zeros as true zeros or censored zeros. Some models also assume an underlying normal distribution, hence can not be used for a wider range of data. Moreover, some models use two separate stochastic processes (one determines the occurrence of zero values and the other determines the distribution of the continuous non-zero values) for handling extra zeros in the data and

others model zeros and non-zeros as a single model.

Generalization of the Bayesian change-point segmented ARMA model by incorporating an appropriate zero-inflated model can be considered an interesting direction for future research. The zeros which were omitted from the analysis in Chapter 5 were true zeros. Consequently, Bayesian change-point segmented ARMA model can be generalized either by using any of the four models mentioned above in Table 6.1 which work with true zeros or by introducing an entirely new model. Two-part models can be used to handle extra zeros in the sediment core data. The generalization model can consist of two parts in each segment of the time series. The first part will determine whether the time series is positive and hence hold a binary model for the dichotomous event of having zero or positive values. The second part will determine the level of time series conditional on its being positive and use an ARMA model for the non-zero values of the time series. Administering the interruption of the ARMA process by zeros in this generalization model can be considered as an area for future research.

References

1. Mills, E. D. Adjusting for covariates in zero-inflated gamma and zero-inflated log-normal models for semicontinuous data (2013).

2. Möller, T. A., Weiß, C. H., Kim, H.-Y. & Sirchenko, A. Modeling zero inflation in count data time series with bounded support. *Methodology and Computing in Applied Probability* **20,** 589–609 (2018).

3. Yang, M. Statistical models for count time series with excess zeros (2012).

4. Xu, L., Paterson, A. D., Turpin, W. & Xu, W. Assessment and selection of competing models for zero-inflated microbiome data. *PloS one* **10,** e0129606 (2015).

5. Pimentel, R. S., Niewiadomska-Bugaj, M. & Wang, J.-C. Association of zero-inflated continuous variables. *Statistics & Probability Letters* **96,** 61–67 (2015).

6. Tu, W. & Zhou, X.-H. A Wald test comparing medical costs based on log-normal distributions with zero valued costs. *Statistics in medicine* **18,** 2749–2761 (1999).

7. Xiao-Hua, Z. & Tu, W. Comparison of several independent population means when their samples contain log-normal and possibly zero observations. *Biometrics* **55,** 645–651 (1999).

8. Feuerverger, A. On some methods of analysis for weather experiments. *Biometrika* **66,** 655–658 (1979).

9. Neelon, B., O'Malley, A. J. & Normand, S.-L. T. A Bayesian two-part latent class model for longitudinal medical expenditure data: assessing the impact of mental health and substance abuse parity. *Biometrics* **67,** 280–289 (2011).

10. Neelon, B. & O'Malley, A. J. Two-Part Models for Zero-Modified Count and Semicontinuous Data. *Health Services Evaluation,* 695–716 (2019).

11. Lecomte, J.-B. *et al.* Compound P oisson-gamma vs. delta-gamma to handle zero-inflated continuous data under a variable sampling volume. *Methods in Ecology and Evolution* **4,** 1159–1166 (2013).

12. Warton, D. I., Renner, I. W. & Ramp, D. Model-based control of observer bias for the analysis of presence-only data in ecology. *PloS one* **8,** e79168 (2013).

13. Aitchison, J. On the distribution of a positive random variable having a discrete probability mass at the origin. *Journal of the american statistical association* **50,** 901–908 (1955).

14. Tu, W. & Liu, H. Zero-inflated data. *Wiley StatsRef: statistics reference online,* 1–7 (2014).

15. Hayashi, F. Econometrics. 2000. *Princeton University Press. Section* **1,** 60–69 (2000).

16. Tobin, J. Estimation of relationships for limited dependent variables. *Econometrica: journal of the Econometric Society,* 24–36 (1958).

17. Kennedy, P. *A guide to econometrics* (MIT press, 2003).

18. Bierens, H. J. *Introduction to the mathematical and statistical foundations of econometrics* (Cambridge University Press, 2004).

19. Robinson, P. M. On the asymptotic properties of estimators of models containing limited dependent variables. *Econometrica: Journal of the Econometric Society,* 27–41 (1982).

20. Powell, J. L. Symmetrically trimmed least squares estimation for Tobit models. *Econometrica: journal of the Econometric Society,* 1435–1460 (1986).

21. Min, Y. & Agresti, A. Modeling nonnegative data with clumping at zero: a survey. *Journal of the Iranian Statistical Society* **1,** 7–33 (2002).

22. Yoo, S.-H., Kim, T.-Y. & Lee, J.-K. Modeling zero response data from willingness to pay surveys: A semi-parametric estimation. *Economics Letters* **71,** 191–196 (2001).

23. Heckman, J. J. Sample selection bias as a specification error. *Econometrica: Journal of the econometric society,* 153–161 (1979).

24. Van de Ven, W. P. & Van Praag, B. M. The demand for deductibles in private health insurance: A probit model with sample selection. *Journal of econometrics* **17,** 229–252 (1981).

25. Puhani, P. The Heckman correction for sample selection and its critique. *Journal of economic surveys* **14,** 53–68 (2000).

26. Cameron, A. C. & Trivedi, P. K. *Microeconometrics: methods and applications* (Cambridge university press, 2005).

27. Goldberger, A. S. in *Studies in econometrics, time series, and multivariate statistics* 67–84 (Elsevier, 1983).

28. Duan, N. Smearing estimate: a nonparametric retransformation method. *Journal of the American Statistical Association* **78,** 605–610 (1983).

29. Duan, N., Manning, W. G., Morris, C. N. & Newhouse, J. P. A comparison of alternative models for the demand for medical care. *Journal of business & economic statistics* **1,** 115–126 (1983).

30. Duan, N., Manning, W. G., Morris, C. N. & Newhouse, J. P. Choosing between the sample-selection model and the multi-part model. *Journal of Business & Economic Statistics* **2,** 283–289 (1984).

31. Dunn, P. K. & Smyth, G. K. Series evaluation of Tweedie exponential dispersion model densities. *Statistics and Computing* **15,** 267–280 (2005).

32. Cragg, J. G. Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica (pre-1986)* **39,** 829 (1971).

33. Jones, A. M. A double-hurdle model of cigarette consumption. *Journal of applied econometrics* **4,** 23–39 (1989).

34. Atkinson, A. B., Gomulka, J. & Stern, N. *Household expenditure on tobacco 1970-1980: evidence from the family expenditure survey* **57** (ESRC Programme on Taxation, Incentives and the Distribution of Income, 1984).

35. Eakins, J. An application of the double hurdle model to petrol and diesel household expenditures in Ireland. *Transport Policy* **47,** 84–93 (2016).

36. Aristei, D. & Pieroni, L. A double-hurdle approach to modelling tobacco consumption in Italy. *Applied Economics* **40,** 2463–2476 (2008).

37. Eggers, J. *On statistical methods for zero-inflated models* 2015.

38. Jørgensen, B. Exponential dispersion models. *Journal of the Royal Statistical Society: Series B (Methodological)* **49,** 127–145 (1987).

39. Dzupire, N. C., Ngare, P. & Odongo, L. A poisson-gamma model for zero inflated rainfall data. *Journal of Probability and Statistics* **2018** (2018).

40. Withers, C. & Nadarajah, S. On the compound Poisson-gamma distribution. *Kybernetika* **47,** 15–37 (2011).

41. Tweedie, M. C. *An index which distinguishes between some important exponential families* in *Statistics: Applications and new directions: Proc. Indian statistical institute golden Jubilee International conference* **579** (1984), 579–604.

42. Saei, A., Ward, J. & McGilchrist, C. Threshold models in a methadone programme evaluation. *Statistics in Medicine* **15,** 2253–2260 (1996).

43. Chang, B.-H. & Pocock, S. Analyzing data with clumping at zero: an example demonstration. *Journal of clinical epidemiology* **53,** 1036–1043 (2000).

44. McCullagh, P. Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)* **42,** 109–127 (1980).

45. Peterson, B. & Harrell Jr, F. E. Partial proportional odds models for ordinal response variables. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **39,** 205–217 (1990).

Chapter 7

Discussion, Conclusion and Future Directions

This thesis considers the problem of modeling a time series by segmenting the series into blocks which can be fitted by approximately stationary ARMA processes. The main objective of this thesis is to extend and develop a Bayesian change-point segmented ARMA model for time series data and develop methods to segment multiple parallel time series.

The Bayesian change-point segmented ARMA methodology presented here segments a time series so that the resulting segments of the time series are consistent with an ARMA model with distinct segment means. In the past, some authors modeled nonstationary time series by segmenting the series into AR processes [1–9]. Autoregressive models operate under the premise that past values have an effect on current values, which makes this statistical model popular for dealing with autocorrelation in the time series. However, the presence of autocorrelation may sometimes lead to excessive proliferation of change-points [10].

One reason this thesis proposes ARMA models in each segment is that ARMA models consider the dependence between residual terms by adding a moving average component instead of only considering the dependence within time series like the AR model. Moreover, a Bayesian approach with a segmented ARMA model allows the fitting of more flexible change-point models than is possible with an AR model. Another advantage of this model is that it uses a highly efficient sampling technique (GGS) [11] for generating samples from a posterior distribution. These reasons make the Bayesian change-point segmented ARMA model presented in this thesis a promising new direction in the field of time series segmentation.

In this thesis, at first I introduced and validated the Bayesian change-point segmented ARMA model and segmented one-dimensional time series synthetic data as well as a real world data set (Chapter 3). When the data was generated via an ARMA process with different segment means and the same variance, the results showed that an ARMA model detects more true change-points than either an AR or MA model, for simulated data. Moreover, in some locations, the segmented ARMA model found higher posterior probability of occurrence of true change-points than the segmented AR model. The posterior estimators of the mean signal level at each position of the simulated signal correctly identified the number of actual segments in the simulated data. A change in mean was evident even where change-points were detected with very low probability.

For real life data, I fitted AR(1), MA(1) and ARMA(1,1) models to detect change-points in the data and all models identified almost the same change-point locations. But at some locations the AR(1) model showed relatively smaller posterior probability than the MA(1) and ARMA(1,1) models. Since the three models found high probabilities of change-point occurrence at many of the same locations, I compared the change-point locations using a threshold of 0.5 in the posterior probabilities. By this method, the AR(1) model identified fewer change-points with high posterior probability than the other two models. Among these three models, the ARMA(1,1) model identified the largest number of change-points and matched more closely with the number and locations of actual change-points detectable to the eye. In addition, the ARMA(1,1) model picked up small changes in mean with high posterior probability whereas the AR(1) and MA(1) models missed change-points at some time points where small jumps occurred in the data. The results of the ARMA(1,1) model were somewhat similar to previous results found in the change-point literature for this data set but a question remains whether the additional change-points are false positives in this case. I can not answer this question definitively as the true locations of change in this data set are unknown. However, the ARMA(1,1) model was found to be the best of the three models based on DICV values, suggesting the additional change-points found in segmented ARMA model reflect a real feature of the data. Arguably, it is not justified to compare MA -which is always stable/stationary- with AR and ARMA. The AR(1) terms in AR and ARMA introduce dependency in the processes through the relationship between X_t and X_{t1}. This may be why the DICV of MA(1) is very high compared to the other two in Table 1 of Chapter 3. The results of simulated data and real data suggest high detection accuracy for this model. This method is especially beneficial in analysing segmentation patterns in the

data where changes are present in the mean. It is not suitable for data sets in which different segments have different variance, as this model assumes the same variance for all segments. In future work, the Bayesian segmented ARMA change-point model can be generalized to incorporate changes in the variance of the original process by allowing distinct variances for different segments or classes of segments. There are some other limitations of this model, such as this model is a univariate segmentation method which assumes stationarity within each segment and this model considers the minimum number of observations in each segment. Also, this Bayesian change-point segmented ARMA model assumes known and fixed orders of AR and MA submodels on each segement. Selecting higher orders of AR or MA model will make this methodology more complex. The selection of the order of the AR or MA model is a challenging task. This problem can be addressed in two ways: i) orders of AR or MA model can be considered as unknown and a joint prior distribution can be set over the orders of the models within each segment ii) different model orders can be considered for different segments and can be estimated from the data.

Developing a simple, essentially one-dimensional, approach to segmenting parallel time series was the second main methodological development I introduced in this thesis. In this phase of the project, an event detection approach was proposed for segmenting spatio-temporal data in presence of different levels of background noise (Chapter 4). For the detection of such events in parallel spatially correlated time series data, I pre-processed the data to produce a single time series amenable to analysis with the segmented ARMA model instead of segmenting two-dimensional data. A single summary sequence was produced based on the maximum signal over locations for each time point. The proposed approach was tested by segmenting two synthetic data sets and three real life data sets with different shaped hidden events in the presence of noise. Both the synthetic data sets contained events in the Gaussian noise with a different mean from the background and the same variance as the background. One contained a triangular event and the other contained an odd quadrangle event. AR(1), MA(1) and ARMA(1,1) models were fitted for both the data sets across row and column locations and the best model was chosen based on investigating DICV values. The change-point profiles and the group profiles of the selected models accurately identified the boundary locations of events obscured by background noise. A simple event extraction method was also used to make a comparison with the results of the proposed method. In spite of extracting events in a two-dimensional setting, this comparison method found some small noisy events in addition to the real events for both

synthetic data sets. Instead of simulating data with Gaussian noise, a nonlinear time series can be simulated and then an ARMA model can be fitted within each segment. This work can be done in future as it will take more time and more coding.

In this chapter, I considered two real life examples with relatively low background noise and one with high background noise. The change-point profiles showed the boundary of the event with high probability for the data sets with a low level of background noise. It also detected some other change-points with low probability. In this situation, the boundaries of the event were identified by plotting group profiles of each position of the input signal, which is the probability that these positions belong to a given group in the selected model. The third real life example contained two events of interest in high background noise. This was a particularly challenging example, not only because of the high background noise level, but also because of the presence of non-uniformity in the variance of the background noise level. The proposed method didn't produce the same level of accuracy as in the other two real life examples. Instead, it found some false positive change-points in addition to the real events. This example indicates a limitation of this simple approach in the presence of high noise variance. The results of real life examples were also compared with the event extraction method. The comparison method identified one false positive event in addition to the real event for the first two examples, whereas the proposed method identified only the desired boundary of the true events. The comparison method also found more false positive events for the third example than the proposed method.

An alternative summary statistic, principal component analysis (PCA) was also used as the summary statistic in the proposed method instead of the maximum. This approach found a cluster of high probability change-points covering the event in the data with a low level of background noise where the true boundary of the event was located. Both approaches found some false positive change-points in the third example. The difference is that the maximum approach found false positive change-points with low posterior probability whereas the PCA approach found false positive events with high posterior probability. Moreover, the PCA approach missed a second real event whereas the maximum approach found both real events. These results indicate that the maximum performs better than PCA in our examples.

This proposed method of segmenting parallel time series is useful to detect the boundaries of events in two-dimensional data in spite of being designed for one-dimensional time series data. However, this dimension reduction approach involves information loss using

any summary statistic and hence is only recommended in initial exploratory analysis. The proposed method assumes univariate ARMA model in each segment, which may be of limited value for segmenting two-dimensional time series data. The orders of AR and MA submodels are being fixed on each segment, which is another limitation of this method (as discussed in Chapter 3). In future, a prior distribution can be assigned to the unknown orders of the AR and MA submodels, or different model orders can be considered from one segment to another and can be estimated from the data. This work opens up avenues for exploration, in particular, in three ways: one is to improve the Bayesian change-point segmented ARMA model to adapt to changing noise variance (as already noted in Chapter 3); another is to generalise the Bayesian change-point segmented ARMA model to handle multi-dimensional data and the last one is to extend this model by assuming a multivariate ARMA model in each segment.

The above results demonstrate the need to devise better models for segmenting parallel sequences instead of using a one-dimensional approach. Consequently, the next step of this thesis was to propose a more general and flexible method for the segmentation of multiple sequences in parallel (Chapter 5). I considered simultaneous segmentation of parallel sequences and presented three alternative models. The first generalization assumes common change-point locations in all time series, the second assumes the same probabilities of assigning segments to a number of segment classes, in addition to the assumption of the first generalization and the third assumes the segments in each time-series belong to a common segment class, in addition to the assumptions of the first and second generalizations. The feasibility and performance of these three alternative models were assessed by applying them to a simulated data as well as to a real world data set.

Results of the synthetic example show that all generalizations achieve similar results and identified all true change-points where the segmentation of a single time series detected only some change-points. The results obtained for the real life example showed different results for the three generalizations. Consequently, the results were compared using approximated AIC, BIC and DICV values. AIC and BIC prefer the third generalization and the DICV prefers the second generalization as the most favourable model. The performance of the three models were again investigated by comparing the locations of estimated change-points with the known change-point locations in the real life example. The estimated change-points were consistent with the true change-point in this example. The third generalization was preferred based on these results which also support the conclusion drawn using the

information criteria. However, because of the presence of heteroscedasticity in some time series and other discrepancies between the model and the underlying process, the accuracy of the results using real life example is lower than the for the synthetic example. The three alternative models work well when change-points were present in the mean of the signal as all these models assume different means for different segment. Change-points are also present in the variance of the signal in each segment but these models assume constant variance within segments. Consequently, future work is required to assume distinct variances for each class of the segments but constant variance with in a class (as noted in Chapter 3 and Chapter 4). All the three alternative models assume a univariate ARMA model in each segment, which can be considered as a limitation in segmenting correlated parallel time series data (as mentioned in Chapter 4). Moreover, these generalizations also assume the same ARMA parameters in each segment but change-points might be exist in the ARMA parameters.

Chapter 5 performed simultaneous segmentation of ten parallel time series corresponding to ten distinct elemental concentrations. However, the data set actually includes time-series for thirty-seven elements. Most of the elements were not detected at all depths in the core and thus their time series contained zeroes over extended time-periods. These time series with many zeroes motivated me to investigate the literature regarding zero-inflated data (Chapter 6). Zero-inflated semicontinuous data has a distribution with a continuous density except for a probability mass at 0. In Chapter 6, I reviewed models for zero-inflated data. In future work, the Bayesian change-point segmented ARMA model can be generalized for handling zero-inflated data, incorporating insights obtained from this review.

In summary, future works following on from this thesis may include:

- Generalize the Bayesian change-point segmented ARMA model by allowing distinct variances for each segment or segment class.
- Generalize the three alternative models of segmenting parallel sequences for changing variance.
- Generalize the Bayesian change-point segmented ARMA model for multi-dimensional data.
- Generalize the Bayesian change-point segmented ARMA model and the three alternative models of segmenting parallel sequences by assuming different ARMA parameters for each segment.

- Generalize the Bayesian change-point segmented ARMA model for handling zero-inflated data.
- Generalize the Bayesian change-point segmented ARMA model using multivariate ARIMA in each segment for incorporating correlated parallel time series.

References

1. Davis, R. A., Lee, T. C. M. & Rodriguez-Yam, G. A. Structural break estimation for nonstationary time series models. *Journal of the American Statistical Association* **101,** 223–239 (2006).
2. Ho, P.-G. *Image segmentation* (BoD–Books on Demand, 2011).
3. Joseph, L., Vandal, A. C. & Wolfson, D. B. Estimation in the multipath change point problem for correlated data. *Canadian Journal of Statistics* **24,** 37–53 (1996).
4. Lai, T. L., Liu, H. & Xing, H. Autoregressive models with piecewise constant volatility and regression parameters. *Statistica Sinica,* 279–301 (2005).
5. Lai, T. L. & Xing, H. A simple Bayesian approach to multiple change-points. *Statistica Sinica,* 539–569 (2011).
6. Punska, O., Doucet, C. A. A., Fitzgerald, W., Andrieu, C. & Doucet, A. Bayesian segmentation of piecewise constant autoregressive processes using MCMC methods (1999).
7. Punskaya, E., Andrieu, C., Doucet, A. & Fitzgerald, W. J. Bayesian curve fitting using MCMC with applications to signal segmentation. *Signal Processing, IEEE Transactions on* **50,** 747–758 (2002).
8. Ruanaidh, J. J. O. & Fitzgerald, W. J. *Numerical Bayesian methods applied to signal processing* (Springer Science & Business Media, 2012).
9. Wood, S., Rosen, O. & Kohn, R. Bayesian mixtures of autoregressive models. *Journal of Computational and Graphical Statistics* **20,** 174–195 (2011).
10. Ruggieri, E. A Bayesian approach to detecting change points in climatic records. *International Journal of Climatology* **33,** 520–528 (2013).
11. Keith, J. M., Kroese, D. P. & Bryant, D. A generalized Markov sampler. *Methodology and Computing in Applied Probability* **6,** 29–53 (2004).

Appendix A

Appendix Chapter 3

A.1 Details of Posterior Distribution

Conditional Posterior Distribution of $\boldsymbol{\pi}$

The conditional posterior distribution of $\boldsymbol{\pi}$ is:

$$p(\boldsymbol{\pi}|\mathbf{X}, K, \mathbf{s}, \mathbf{g}, \mathbf{c}, \phi, \boldsymbol{\theta}, \boldsymbol{\psi}, \sigma^2, \boldsymbol{\mu}, \boldsymbol{\tau}) = \frac{p(\boldsymbol{\pi}, \mathbf{X}, K, \mathbf{s}, \mathbf{g}, \mathbf{c}, \phi, \boldsymbol{\theta}, \boldsymbol{\psi}, \sigma^2, \boldsymbol{\mu}, \boldsymbol{\tau})}{\int_{\boldsymbol{\pi}} p(\boldsymbol{\pi}, \mathbf{X}, K, \mathbf{s}, \mathbf{g}, \mathbf{c}, \phi, \boldsymbol{\theta}, \boldsymbol{\psi}, \sigma^2, \boldsymbol{\mu}, \boldsymbol{\tau}) d\boldsymbol{\pi}}$$

Using Eq (8), we get,

$$\frac{p(\mathbf{X}, K, \mathbf{s}, \mathbf{g}, \mathbf{c}|\phi, \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\psi}, \sigma^2, \boldsymbol{\mu}, \boldsymbol{\tau}) p(\phi) p(\boldsymbol{\pi}) p(\boldsymbol{\theta}) p(\boldsymbol{\psi}) p(\sigma^2) p(\boldsymbol{\mu}) p(\boldsymbol{\tau})}{\int_{\boldsymbol{\pi}} p(\mathbf{X}, K, \mathbf{s}, \mathbf{g}, \mathbf{c}|\phi, \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\psi}, \sigma^2, \boldsymbol{\mu}, \boldsymbol{\tau}) p(\phi) p(\boldsymbol{\pi}) p(\boldsymbol{\theta}) p(\boldsymbol{\psi}) p(\sigma^2) p(\boldsymbol{\mu}) p(\boldsymbol{\tau}) d\boldsymbol{\pi}}$$

$$= \frac{p(\mathbf{X}, K, \mathbf{s}, \mathbf{g}, \mathbf{c}|\phi, \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\psi}, \sigma^2, \boldsymbol{\mu}, \boldsymbol{\tau}) p(\boldsymbol{\pi})}{\int_{\boldsymbol{\pi}} p(\mathbf{X}, K, \mathbf{s}, \mathbf{g}, \mathbf{c}|\phi, \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\psi}, \sigma^2, \boldsymbol{\mu}, \boldsymbol{\tau}) p(\boldsymbol{\pi}) d\boldsymbol{\pi}}$$

Then, using Eq (6) and cancelling same terms from the numerator and the denominator the conditional posterior distribution of $\boldsymbol{\pi}$ becomes

$$\frac{p(\mathbf{X}|\boldsymbol{\lambda}, \sigma^2) p(\mathbf{c}|\mathbf{g}, \boldsymbol{\mu}, \boldsymbol{\tau}) p(K, \mathbf{s}|\phi) p(\mathbf{g}|K, \boldsymbol{\pi}) p(\boldsymbol{\pi})}{\int_{\boldsymbol{\pi}} p(\mathbf{X}|\boldsymbol{\lambda}, \sigma^2) p(\mathbf{c}|\mathbf{g}, \boldsymbol{\mu}, \boldsymbol{\tau}) p(K, \mathbf{s}|\phi) p(\mathbf{g}|K, \boldsymbol{\pi}) p(\boldsymbol{\pi}) d\boldsymbol{\pi}} = \frac{p(\mathbf{g}|K, \boldsymbol{\pi}) p(\boldsymbol{\pi})}{\int_{\boldsymbol{\pi}} p(\mathbf{g}|K, \boldsymbol{\pi}) p(\boldsymbol{\pi}) d\boldsymbol{\pi}}$$

Now, Using Eq (3), this becomes:

$$\frac{\prod_{n=1}^{N} \boldsymbol{\pi}_n^{b_n}}{\int_{\boldsymbol{\pi}} \prod_{n=1}^{N} \boldsymbol{\pi}_n^{b_n} d\boldsymbol{\pi}} = \frac{\boldsymbol{\pi}_1^{(b_1+1)-1} \times \cdots \times \boldsymbol{\pi}_1^{(b_N+1)-1}}{\int_{\boldsymbol{\pi}} \boldsymbol{\pi}_1^{(b_1+1)-1} \times \cdots \times \boldsymbol{\pi}_1^{(b_N+1)-1} d\boldsymbol{\pi}} = Dirichlet(\boldsymbol{\pi}|((b_1+1), \cdots, (b_N+1))$$

Conditional Posterior Distribution of ϕ

The conditional posterior distribution of ϕ is,

$$\frac{p(\mathbf{X}|\boldsymbol{\lambda}, \sigma^2)p(\mathbf{c}|\mathbf{g}, \boldsymbol{\mu}, \boldsymbol{\tau})p(K, \mathbf{s}|\phi)p(\mathbf{g}|K, \phi)p(\boldsymbol{\pi})}{\int_\phi p(\mathbf{X}|\boldsymbol{\lambda}, \sigma^2)p(\mathbf{c}|\mathbf{g}, \boldsymbol{\mu}, \boldsymbol{\tau})p(K, \mathbf{s}|\phi)p(\mathbf{g}|K, \boldsymbol{\pi})p(\phi)d\phi} = \frac{p(K, \mathbf{s}|\phi)p(\phi)}{\int_\phi p(K, \mathbf{s}|\phi)p(\phi)d\phi}$$

Using Eq (1), the conditional posterior distribution of ϕ becomes

$$\frac{\phi^{K-1}(1-\phi)^{T-K-1}}{\int_\phi \phi^{K-1}(1-\phi)^{T-K-1}d\phi} = Beta(\phi|K, T-K)$$

where, K = Number of segments and $T - K$ = Length of signal-number of segments.

Conditional Posterior Distribution of $\boldsymbol{\mu}$

The conditional posterior distribution of $\boldsymbol{\mu}$ is:

$$\frac{p(\mathbf{X}|\boldsymbol{\lambda}, \sigma^2)p(\mathbf{c}|\mathbf{g}, \boldsymbol{\mu}, \boldsymbol{\tau})p(K, \mathbf{s}|\phi)p(\mathbf{g}|K, \boldsymbol{\pi})p(\boldsymbol{\mu})}{\int_{\boldsymbol{\mu}} p(\mathbf{X}|\boldsymbol{\lambda}, \sigma^2)p(\mathbf{c}|\mathbf{g}, \boldsymbol{\mu}, \boldsymbol{\tau})p(K, \mathbf{s}|\phi)p(\mathbf{g}|K, \boldsymbol{\pi})p(\boldsymbol{\mu})d\boldsymbol{\mu}} = \frac{p(\mathbf{c}|\mathbf{g}, \boldsymbol{\mu}, \boldsymbol{\tau})p(\boldsymbol{\mu})}{\int_{\boldsymbol{\mu}} p(\mathbf{c}|\mathbf{g}, \boldsymbol{\mu}, \boldsymbol{\tau})p(\boldsymbol{\mu})d\boldsymbol{\mu}}$$

Note that, μ or μ_n will (currently) not be updated if there is less than one segment in group n. Then using Eq (7), the conditional distribution for μ_n (holding all the other μ's constant) becomes

$$\frac{\prod_{k:g_k=n} \mathcal{N}(c_k|\mu_n, \tau_n^2)}{\int_\mu \prod_{k:g_k=n} \mathcal{N}(c_k|\mu_n, \tau_n^2)d\mu}$$

.

Let m be the number of segments that have $g_k = n$. Then simplifying the product $\prod_{k:g_k=n} \mathcal{N}(c_k|\mu_n, \tau_n^2) = \prod_{k:g_k=n} \frac{1}{\sqrt{2\pi\tau_n^2}}\ exp\left[\left(\frac{(c_k-\mu_n)^2}{2\tau_n^2}\right)\right]$, the conditional posterior distribution of $\boldsymbol{\mu}$ becomes a normal distribution with mean $\frac{\sum_{k:g_k=n} c_k}{m}$ and variance $\frac{\tau_n^2}{m}$.

Conditional Posterior Distribution of σ^2

The conditional posterior distribution of σ^2 is:

$$\frac{p(\mathbf{X}|\boldsymbol{\lambda},\sigma^2)p(\mathbf{c}|\mathbf{g},\boldsymbol{\mu},\boldsymbol{\tau})p(K,\mathbf{s}|\phi)p(\mathbf{g}|K,\boldsymbol{\pi})p(\sigma^2)}{\int_{\sigma^2} p(\mathbf{X}|\boldsymbol{\lambda},\sigma^2)p(\mathbf{c}|\mathbf{g},\boldsymbol{\mu},\boldsymbol{\tau})p(K,\mathbf{s}|\phi)p(\mathbf{g}|K,\boldsymbol{\pi})p(\sigma^2)d\sigma^2} = \frac{p(\mathbf{X}|\boldsymbol{\lambda},\sigma^2)p(\sigma^2)}{\int_{\sigma^2} p(\mathbf{X}|\boldsymbol{\lambda},\sigma^2)p(\sigma^2)d\sigma^2}$$

The prior distribution for σ^2 is: $p(\sigma^2,u_0,v_0) = \frac{v_0^{u_0}}{\Gamma(u_0)}(\sigma^2)^{-u_0-1}exp\left(-\frac{v_0}{\sigma^2}\right).$
where u_0 and v_0 are the prior parameters. Now using $\boldsymbol{\epsilon} \sim \mathcal{N}(0,\sigma^2)$ and the above prior distribution and simplifying, the conditional posterior distribution of σ^2 takes the form of an inverse gamma distribution with parameter u and v, where $u = u_0 + \frac{T}{2}$ and $v = \frac{2v_0+\sum_{t=1}^{T}\epsilon_t^2}{2} = v_0 + \frac{1}{2}\sum_{t=1}^{T}\epsilon_t^2$.

Conditional Posterior Distribution of $\boldsymbol{\tau_n^2}$

Let $\boldsymbol{\tau}$ be the vector of $\tau_1^2,\cdots\tau_N^2$, then the conditional posterior distribution of $\boldsymbol{\tau}$ is:

$$\frac{p(\mathbf{X}|\boldsymbol{\lambda},\sigma^2)p(\mathbf{c}|\mathbf{g},\boldsymbol{\mu},\boldsymbol{\tau})p(K,\mathbf{s}|\phi)p(\mathbf{g}|K,\boldsymbol{\pi})p(\boldsymbol{\tau})}{\int_{\boldsymbol{\tau}} p(\mathbf{X}|\boldsymbol{\lambda},\sigma^2)p(\mathbf{c}|\mathbf{g},\boldsymbol{\mu},\boldsymbol{\tau})p(K,\mathbf{s}|\phi)p(\mathbf{g}|K,\boldsymbol{\pi})p(\boldsymbol{\tau})d\boldsymbol{\tau}} = \frac{p(\mathbf{c}|\mathbf{g},\boldsymbol{\mu},\boldsymbol{\tau})p(\boldsymbol{\tau})}{\int_{\boldsymbol{\tau}} p(\mathbf{c}|\mathbf{g},\boldsymbol{\mu},\boldsymbol{\tau})p(\boldsymbol{\tau})d\boldsymbol{\tau}}$$

The prior distribution for $\boldsymbol{\tau}$ is: $p(\boldsymbol{\tau},\alpha_0,\beta_0) = \prod_{n=1}^{N}\frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}(\tau_n^2)^{-\alpha_0-1}exp\left(-\frac{\beta_0}{\tau_n^2}\right).$
where α_0 and β_0 are the prior parameters. Note that τ_n^2 will (currently) only be updated if there is more than one segment in group n. Now using Eq (7) and the above prior distribution and simplifying, the conditional posterior distribution of $\boldsymbol{\tau}$ takes the form of an inverse gamma distribution with parameter α and β, where $\alpha = \alpha_0 + \frac{|(k:g_k=n)|}{2}$ and $\beta = \beta_0 + \frac{1}{2}\sum_{k:g_k=n}(c_k-\mu_n)^2$.

Conditional Posterior Distribution of g

The conditional posterior distribution of $\mathbf{g}$ is:

$$\frac{p(\mathbf{X}|\boldsymbol{\lambda},\sigma^2)p(\mathbf{c}|\mathbf{g},\boldsymbol{\mu},\boldsymbol{\tau})p(K,\mathbf{s}|\phi)p(\mathbf{g}|K,\boldsymbol{\pi})p(\mathbf{g})}{\int_{\mathbf{g}} p(\mathbf{X}|\boldsymbol{\lambda},\sigma^2)p(\mathbf{c}|\mathbf{g},\boldsymbol{\mu},\boldsymbol{\tau})p(K,\mathbf{s}|\phi)p(\mathbf{g}|K,\boldsymbol{\pi})p(\mathbf{g})d\mathbf{g}} = \frac{p(\mathbf{c}|\mathbf{g},\boldsymbol{\mu},\boldsymbol{\tau})p(\mathbf{g}|K,\boldsymbol{\pi})p(g)}{\int_{\mathbf{g}} p(\mathbf{c}|\mathbf{g},\boldsymbol{\mu},\boldsymbol{\tau})p(\mathbf{g}|K,\boldsymbol{\pi})p(g)d\mathbf{g}}$$

Using Eq (2) and Eq (7), The conditional posterior distribution of $\mathbf{g}$ becomes

$$\frac{\prod_{k=1}^{K} \mathcal{N}(c_k|\mu_{g_k}, \tau_{g_k}^2) \times \prod_{k=1}^{K} \pi_{g_k}}{\int_g (\prod_{k=1}^{K} \mathcal{N}(c_k|\mu_{g_k}, \tau_{g_k}^2) \times \prod_{k=1}^{K} \pi_{g_k}) dg}.$$

The g_k 's are independent, so can be updated one at a time with the conditional posterior distribution of g which is a discrete distribution with parameter $\frac{\mathcal{N}(c_k|\mu_{g_k}, \tau_{g_k}^2) \times \pi_{g_k}}{\int_g (\mathcal{N}(c_k|\mu_{g_k}, \tau_{g_k}^2) \times \pi_{g_k}) dg}$.

Conditional Posterior Distribution of c and $\boldsymbol{\epsilon}$

The conditional posterior distribution of $\mathbf{c}$ and $\boldsymbol{\epsilon}$ is:

$$\frac{p(\mathbf{X}|\boldsymbol{\lambda}, \sigma^2) p(\mathbf{c}|\mathbf{g}, \boldsymbol{\mu}, \boldsymbol{\tau}) p(K, \mathbf{s}|\phi) p(\mathbf{g}|K, \boldsymbol{\pi}) p(\mathbf{c})}{\int_{\mathbf{c}} p(\mathbf{X}|\boldsymbol{\lambda}, \sigma^2) p(\mathbf{c}|\mathbf{g}, \boldsymbol{\mu}, \boldsymbol{\tau}) p(K, \mathbf{s}|\phi) p(\mathbf{g}|K, \boldsymbol{\pi}) p(\mathbf{c}) d\mathbf{c}} = \frac{p(\mathbf{X}|\boldsymbol{\lambda}, \sigma^2) p(\mathbf{c}|\mathbf{g}, \boldsymbol{\mu}, \boldsymbol{\tau}) p(\mathbf{c})}{\int_{\mathbf{c}} p(\mathbf{X}|\boldsymbol{\lambda}, \sigma^2) p(\mathbf{c}|\mathbf{g}, \boldsymbol{\mu}, \boldsymbol{\tau}) p(\mathbf{c}) d\mathbf{c}}$$

Using Eq (5) and Eq (7) and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2)$, the above equation becomes

$$\frac{\prod_{t=1}^{T} p(x_t|\lambda_t, \sigma^2) \times \prod_{k=1}^{K} \mathcal{N}(c_k|\mu_{g_k}, \tau_{g_k}^2)}{\int_c (\prod_{t=1}^{T} p(x_t|\lambda_t, \sigma^2) \times \prod_{k=1}^{K} \mathcal{N}(c_k|\mu_{g_k}, \tau_{g_k}^2)) dc}.$$

The c_k's and the corresponding $\epsilon_{s_k}, \cdots, \epsilon_{d_k}$ can be updated one segment at a time with the conditional distribution given by:

$$\frac{\prod_{t=s_k}^{d_k} p(x_t|\lambda_t, \sigma^2) \times \mathcal{N}(c_k|\mu_{g_k}, \tau_{g_k}^2)}{\int_c (\prod_{t=s_k}^{d_k} p(x_t|\lambda_t, \sigma^2) \times \mathcal{N}(c_k|\mu_{g_k}, \tau_{g_k}^2)) dc}.$$

Conditional Posterior Distribution of $\boldsymbol{\psi}$ and $\boldsymbol{\epsilon}$

The conditional posterior distribution of $\boldsymbol{\psi}$ and $\boldsymbol{\epsilon}$ is:

$$\frac{p(\mathbf{X}|\boldsymbol{\lambda}, \sigma^2) p(\mathbf{c}|\mathbf{g}, \boldsymbol{\mu}, \boldsymbol{\tau}) p(K, \mathbf{s}|\phi) p(\mathbf{g}|K, \boldsymbol{\pi}) p(\boldsymbol{\psi})}{\int_{\boldsymbol{\psi}} p(\mathbf{X}|\boldsymbol{\lambda}, \sigma^2) p(\mathbf{c}|\mathbf{g}, \boldsymbol{\mu}, \boldsymbol{\tau}) p(K, \mathbf{s}|\phi) p(\mathbf{g}|K, \boldsymbol{\pi}) p(\boldsymbol{\psi}) d\boldsymbol{\psi}} = \frac{p(\mathbf{X}|\boldsymbol{\lambda}, \sigma^2) p(\boldsymbol{\psi})}{\int_{\boldsymbol{\psi}} p(\mathbf{X}|\boldsymbol{\lambda}, \sigma^2) p(\boldsymbol{\psi}) d\boldsymbol{\psi}}$$

Using Eq (5) and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2)$, the conditional posterior distribution of $\boldsymbol{\psi}$ and $\boldsymbol{\epsilon}$ given by:

$$\frac{\prod_{t=1}^{T} p(x_t|\lambda_t, \sigma^2) \times p(\psi)}{\int_{\psi}(\prod_{t=1}^{T} p(x_t|\lambda_t, \sigma^2) \times p(\psi))d\psi}.$$

Conditional Posterior Distribution of $\boldsymbol{\theta}$ and $\boldsymbol{\epsilon}$

The conditional posterior distribution of $\boldsymbol{\theta}$ and $\boldsymbol{\epsilon}$ is:

$$\frac{p(\mathbf{X}|\boldsymbol{\lambda}, \sigma^2)p(\mathbf{c}|\mathbf{g}, \boldsymbol{\mu}, \boldsymbol{\tau})p(K, \mathbf{s}|\phi)p(\mathbf{g}|K, \boldsymbol{\pi})p(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} p(\mathbf{X}|\boldsymbol{\lambda}, \sigma^2)p(\mathbf{c}|\mathbf{g}, \boldsymbol{\mu}, \boldsymbol{\tau})p(K, \mathbf{s}|\phi)p(\mathbf{g}|K, \boldsymbol{\pi})p(\boldsymbol{\theta})d\boldsymbol{\theta}} = \frac{p(\mathbf{X}|\boldsymbol{\lambda}, \sigma^2)p(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} p(\mathbf{X}|\boldsymbol{\lambda}, \sigma^2)p(\boldsymbol{\theta})d\boldsymbol{\theta}}$$

Using Eq (5) and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2)$, the conditional posterior distribution of $\boldsymbol{\theta}$ and $\boldsymbol{\epsilon}$ given by:

$$\frac{\prod_{t=1}^{T} p(x_t|\lambda_t, \sigma^2) \times p(\theta)}{\int_{\theta}(\prod_{t=1}^{T} p(x_t|\lambda_t, \sigma^2) \times p(\theta))d\theta}.$$

A.2 Generalized Gibbs Sampling

This section summarizes the method of the MCMC sampler of Keith *et al.*[1]. In this part, we want to sample from a distribution f over a space $\mathcal{X}$, called the target space.

We define a set $\mathcal{I}$, which will be referred to, in the following, as the index set. We also define $\mathcal{U} \subset \mathcal{I} \times \mathcal{X}$, so that projection of $\mathcal{U}$ onto $\mathcal{X}$ and $\mathcal{I}$ are surjective.

For each $x \in \mathcal{X}$, let $\mathcal{Q}(x)$ be the set $\{(k, z) \in \mathcal{U} : z = x\}$. $\mathcal{Q}(x)$ is a catalogue of the types of transitions available from x.

For every $x \in \mathcal{X}$, we define a transition matrix Q_x on $\mathcal{Q}(x)$. We denote by q_x the density of the distribution which is stationary with respect to Q_x. We denote Q the global transition matrix on $\mathcal{U}$:

$$Q((i,x),(j,y)) = \begin{cases} Q_x((i,x),(j,y)) & \text{for}(j,y) \in \mathcal{Q}(x) \\ 0 & \text{otherwise} \end{cases}$$

For each $(i, x) \in \mathcal{U}$, let $\mathcal{R}(i, x)$ be the set of possible transitions. These sets are required to be a partition of $\mathcal{U}$:

$$(j,y) \in \mathcal{R}(i,x) \Leftrightarrow (i,x) \in \mathcal{R}(j,y)$$

$$\left.\begin{array}{l} (j,y) \in \mathcal{R}(i,x) \\ (k,z) \in \mathcal{R}(j,y) \end{array}\right\} \Rightarrow (k,z) \in \mathcal{R}(i,x)$$

We also have $(i, x) \in \mathcal{R}(i, x)$. On $\mathcal{R}(i, x)$, we define a transition matrix $R(i, x)$ as follows :

$$R_{(i,x)}((i,x),(j,y)) = \frac{f(y)q_y(j,y)}{\sum_{(k,z)\in\mathcal{R}(i,x)} f(z)q_z(k,z)}$$

We also define a global transition matrix R on $\mathcal{U}$:

$$R((i,x),(j,y)) = \begin{cases} R_{(i,x)}((i,x),(j,y)) & \text{for}(j,y) \in \mathcal{R}(i,x) \\ 0 & \text{otherwise} \end{cases}$$

We can generalize this formula by replacing $R((i, x), (j, y))$ with $R((i, x), (j, y))S((i, x), (j, y))$ in the definition of the matrix R, given that $S((i, x), (j, y)) = S((j, y), (i, x))$ and adjusting S so that for all (i, x), $\sum_{(j,y)} R((i, x), (j, y))S((i, x), (j, y)) = 1$. This can be useful to divide

the $\mathcal{R}(i,x)$ set into two (or more) subsets.

We consider a Markov chain $\{U_1, U_2, \cdots\}$ on $\mathcal{U}$ with a transition matrix $P = QR$.

$$P((i,x),(j,y)) = \sum_{(k,z)\in\mathcal{Q}(x)\cap\mathcal{R}(j,y))} Q((i,x),(k,z))R((k,z),(j,y))$$

Let μ be the distribution defined by $\mu(i,x) = f(x)q_x(i,x)$.

We have

$$\mu(i,x)R((i,x),(j,y)) = \begin{cases} \frac{f(x)q_x(i,x)f(y)q_y(j,y)}{\sum_{(k,z)\in\mathcal{R}(i,x)} f(z)q_z(k,z)} & \text{if}(j,y)\in\mathcal{R}(i,x) \\ 0 & \text{otherwise} \end{cases}$$

$$\mu(j,y)R((j,y),(i,x)) = \begin{cases} \frac{f(x)q_x(i,x)f(y)q_y(j,y)}{\sum_{(k,z)\in\mathcal{R}(j,y)} f(z)q_z(k,z)} & \text{if}(i,x)\in\mathcal{R}(i,x) \\ 0 & \text{otherwise} \end{cases}$$

We know that $(j,y)\in\mathcal{R}(i,x), (k,z)\in\mathcal{R}(j,y) \Rightarrow (k,z)\in\mathcal{R}(i,x)$, so

$$\mu(i,x)R((i,x),(j,y)) = \mu(j,y)R((j,y),(i,x))$$

(note that if $S((i,x),(j,y)) = S((j,y),(i,x))$ and

$$\begin{aligned}\sum_{(i,x)\in\mathcal{U}} \mu(i,x)Q((i,x),(j,y)) &= \sum_{(i,y)\in\mathcal{Q}(y)} f(y)q_y(i)Q((i,y),(j,y)) \\ &= f(y)q_y(j) = \mu(j,y)\end{aligned}$$

So, μ is stationary with respect to Q and to R and with respect to P. If P is irreducible and aperiodic, μ is the limiting distribution of the process P. So, we have the generalized Gibbs sampler :

1. Q-step : Given $U_n = (i,x)$, generate $V\in\mathcal{Q}(x)$ by drawing from the distribution with density $Q((i,x),.)$.

2. R-step : Given $V = (j,y)$, generate $W\in\mathcal{R}(j,y)$ by drawing from the distribution with density $R((j,y),.)$.

3. Let $U_{n+1} = W$.

Now we redefine R for the generalisation of the GGS sampler by including Metropolis' sampler, Hastings' generalisations and the reversible jump sampler, that is,

$$R((i,x),(j,y)) = \begin{cases} \frac{s((i,x),(j,y))f(y)q_y(j)}{\sum_{(k,z)\in\mathcal{R}(i,x))} f(z)q_z(k)} & \text{if } (j,y) \in R((i,x))\backslash\{(i,x)\} \\ 1 - \sum_{(k,z)\in\mathcal{R}(i,x))\backslash\{(i,x)\}} R((i,x),(k,z)) & \text{if } (j,y) = (i,x) \\ 0 & \text{otherwise} \end{cases}$$

Here s is a non-negative and symmetric function such that,

$$\sum_{w\in\mathcal{R}(i,x)\backslash\{(i,x)\}} R((i,x),(k,z)) \leq 1$$

In case of involving non-denumerable target space, a measure ζ exits on the set $\mathcal{R} = \{\mathcal{R}(i,x) : (i,x) \in \mathcal{U}\}$ and measures η_r on r for each $r \in \mathcal{R}$ such that

$$\xi(A) = \int_{\mathcal{R}} \eta_r(A \cap r) d\zeta(r)$$

where, ξ is a reference measure on $\mathcal{U}$. Now, we can define $R_{(i,x)}((i,x),.)$ to be the density with respect to η_r on $r = \mathcal{R}(i,x)$ given by

$$R_{(i,x)}((i,x),(j,y)) = \frac{f(y)q_y(j,y)}{\int_r f(z)q_z(k,z)d\eta_r(k,z)}$$

We can also define $R_{(i,x)}((i,x),.)$ to be the density with respect to η_r on $r\backslash\{(i,x)\}$ and assign probability mass to (i,x) given by

$$\begin{aligned} R_{(i,x)}((i,x),(j,y)) &= \frac{s((i,x),(j,y))f(y)q_y(j,y)}{\int_r f(z)q_z(k,z)d\eta_r(k,z)} \\ &= 1 - \frac{\int_{r\backslash(i,x)} s((i,x),(j,y))f(y)q_y(j,y)d\eta_r(j,y)}{\int_r f(z)q_z(k,z)d\eta_r(k,z)} \end{aligned}$$

The function s must satisfy $\frac{\int_{r\backslash(i,x)} s((i,x),(j,y))f(y)q_y(j,y)d\eta_r(j,y)}{\int_r f(z)q_z(k,z)d\eta_r(k,z)} \leq 1$.

In case of reversible jump MCMC sampler instance of the GGS, we define a transition density function $Q_x((m_0,y_0,x_0),(m,y,x)) = q_x(m,y,x)$ in Q-step, where $q_x(m,y,x) = \sigma_x(m)A_m(x,y)$. Here, m is a move-type; $m \in \mathcal{M}$ ($\mathcal{M}$ is the countable set of move-types), $\sigma_x(m)$ is the probability of selecting move-type m at proposed new element x and $A_m(x,y)$ is the density for move-type m at x.

In R-step, the density of μ is $\sigma_x(m)f_m(x,y)$ at (m,y,x) with respect to ξ and we define a partition of $\mathcal{U}$ which consists of $\mathcal{R}(m,y,x) = \{(m,y,x),(m,x,y)\}$ for all $(m,y,x) \in \mathcal{U}$. Here,

$$s((m,y,x),(m,x,y)) = \min\left(1+\frac{\sigma_m(x)f_m(x,y)}{\sigma_m(y)f_m(y,x)}, 1+\frac{\sigma_m(y)f_m(y,x)}{\sigma_m(x)f_m(x,y)}\right)$$

This s matrix gives the transition matrix

$$R_{(i,x)}((i,x),(j,y)) = \begin{cases} \alpha_m(x,y) & \text{if } (j,y) = (i,x) \equiv (m,y,x) \\ 1-\alpha_m(x,y) & \text{if } (j,y) = (m,y,x) \end{cases}$$

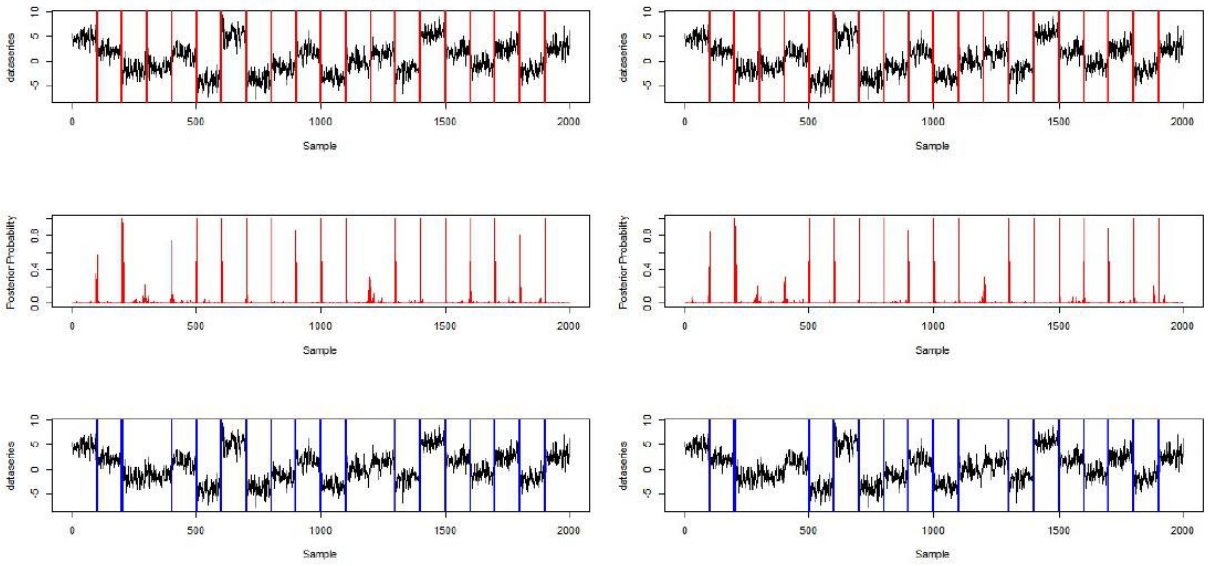
where,

$$\alpha_m(x,y) = \min\left(1, \frac{\sigma_m(y)f_m(y,x)}{\sigma_m(x)f_m(x,y)}\right).$$

Here, new element(m,y,x) is selected with the transition matrix R which is equivalent to accepting the proposed y with probability $\alpha_m(x,y)$.

A.3 Supplementary Material A

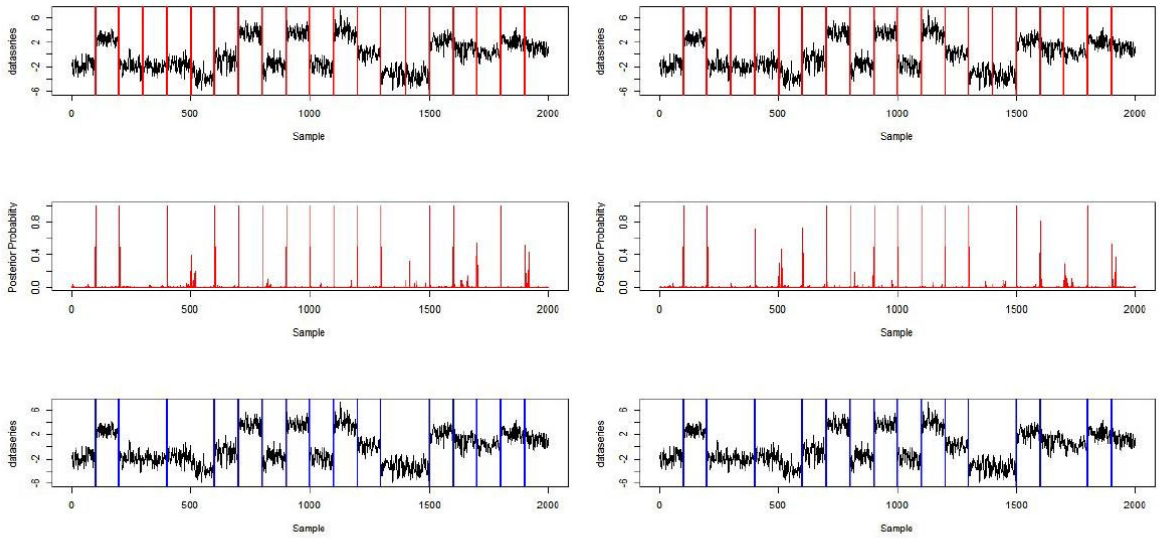
To test our method, we generated two simulated data sets. In both data sets, we generated 20 time series of 100 observations, each with a different segment mean, from the autoregressive moving average (ARMA (1,1)) model with parameter values $\psi = 0.22$ and $\theta = 0 : 60$. The two data sets differed in the value of the parameter σ^2 used to generate the data: in the first we used $\sigma^2 = 0.96$ and in the second we used $\sigma^2 = 0.5$. We applied our methods for the segmented ARMA model and segmented AR model to these data and compared the location of change points and number of change points found by these two models.



(a) Segmented ARMA with $\sigma^2 = 0.96$ (b) Segmented AR with $\sigma^2 = 0.96$

S3 Fig A.1: Segmented ARMA model and segmented AR model with $\sigma^2 = 0.96$

The top panels of S3 Fig A.1 present the simulated signal for the first data set ($\sigma^2 = 0.96$) with the true change-points marked as red vertical lines. The middle plots show the posterior probabilities of occurrence of change-point locations and the bottom plots show the estimated change-point locations using a threshold (0.5) in the posterior probabilities of occurrence of change-points. If we compare the above plots, it is clear that the segmented ARMA and segmented AR models find almost the same number of change points and locations. But in some locations, the segmented ARMA model gives higher posterior probability than the segmented AR model. When we apply a threshold (the posterior probability of occurrence of change-points is greater than 0.5), the segmented ARMA model identifies 17 change-points out of 19 true change-points whereas the segmented AR model identifies 16 change-points.



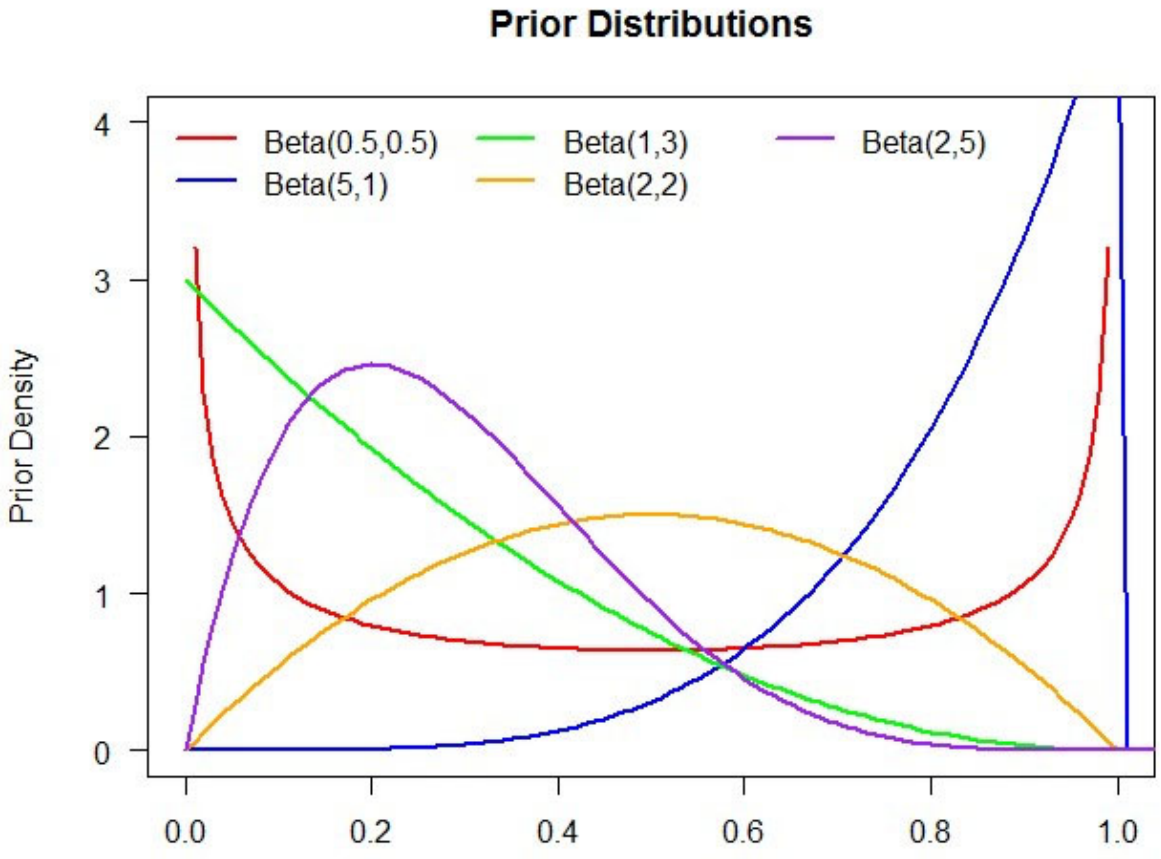
(a) Segmented ARMA with $\sigma^2 = 0.5$ (b) Segmented AR with $\sigma^2 = 0.5$

S3 Fig A.2: Segmented ARMA model and segmented AR model with $\sigma^2 = 0.5$

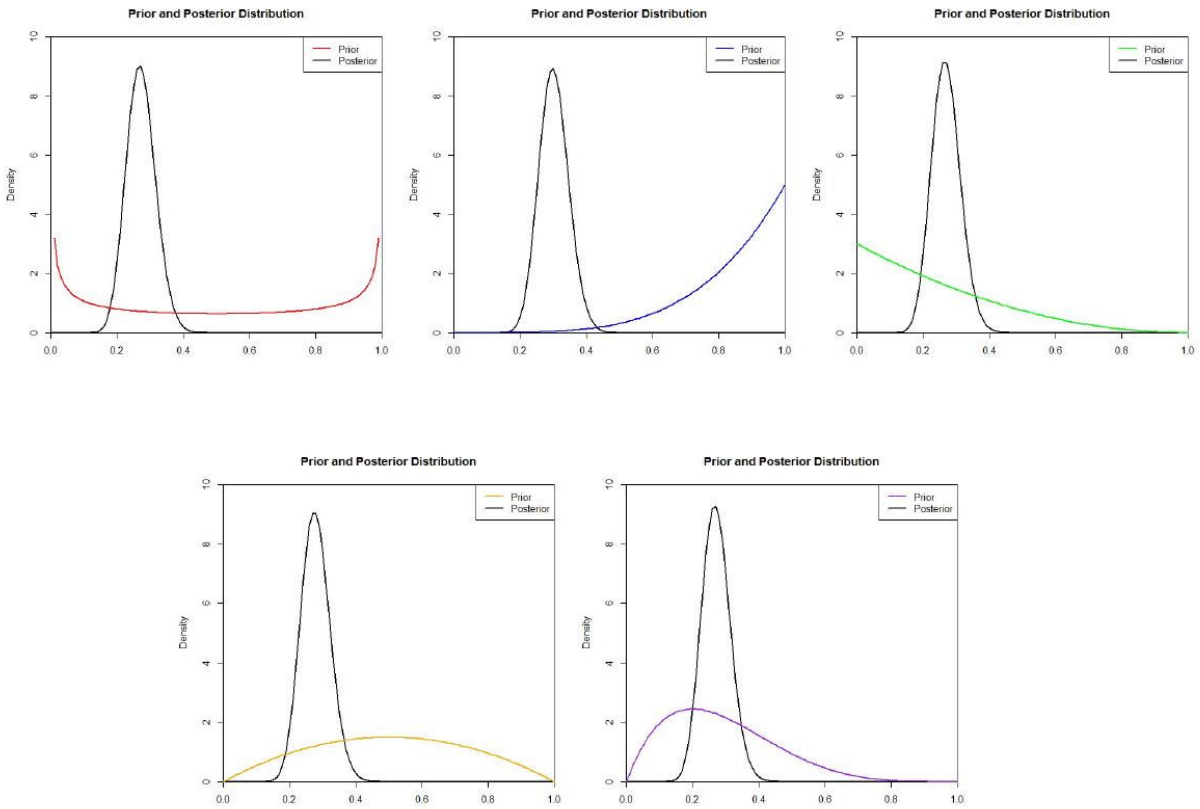
Results for the second simulated data set ($\sigma^2 = 0.5$) are shown in S3 Fig A.2. These plots also demonstrate that the ARMA and AR models find the same number of change-points and the same locations, but in some locations the posterior probability of occurrence of change-points is lower using the segmented AR model than the segmented ARMA model. The bottom panel of the segmented AR plot indicates 15 change-points were identified out of 19 but the segmented ARMA model identifies 16 change-points. These results suggest that the segmented ARMA model identifies significant change-points with higher probability of occurrence than the AR model, when the data are generated using an ARMA model, regardless of the value of σ^2.

A.4 Supplementary Material B

Our computational algorithm is not sensitive to the choice of prior for ϕ. The posterior distribution of ϕ is a beta distribution that is scarcely changed by the choice of prior, as the following graphs indicate. The top panel in S4 Fig A.3 below shows five low information prior distributions for ϕ, with different means and variances. S4 Fig A.4 show the posterior distributions obtained using each of the five priors. Despite the large discrepancies in the prior information, the posterior distributions are indistinguishable.



S4 Fig A.3: Different prior distributions for ϕ



S4 Fig A.4: Posterior distributions for different prior distributions

Appendix B

Appendix Chapter 4

B.1 Summary of Bayesian change-point modeling with segmented ARMA model

This section summarizes the Bayesian segmented ARMA model of [2]. This model segments the input time series into blocks of autoregressive moving average (ARMA) processes. The joint distribution of the data conditional on the other parameters of the model is given by:

$$\begin{aligned} p(\mathbf{X}, K, \mathbf{s}, \mathbf{g}, \mathbf{c}|\phi, \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\psi}, \sigma^2, \boldsymbol{\mu}, \boldsymbol{\tau}) = \\ p(\mathbf{X}|K, \mathbf{s}, \boldsymbol{\theta}, \boldsymbol{\psi}, \mathbf{c}, \sigma^2)\times \\ p(K, \mathbf{s}|\phi)\times \\ p(\mathbf{c}|\mathbf{g}, \boldsymbol{\mu}, \boldsymbol{\tau})\times \\ p(\mathbf{g}|K, \boldsymbol{\pi}) \end{aligned} \tag{B.1}$$

Here, the probabilities of Equation B.1 are:

$$\begin{aligned} p(K, \mathbf{s}|\phi) = &\text{ The probability of generating a segmentation with } K \text{ segments that} \\ &\text{have starting positions } \mathbf{s} \\ &= \phi^{K-1}(1-\phi)^{T-K}. \end{aligned}$$

where, K=Number of segments; $\mathbf{s} = (1 = s_1 < \cdots < s_K \leq T)$ is a vector of the starting positions of the segments; $T - K$ = Length of signal-number of segments and ϕ is the

probability of starting a new segment at each position in the signal except the first.

$$p(\mathbf{g}|K,\boldsymbol{\pi}) = \text{The probability of a specific assignment of the } K \text{ segments to the } N \text{ groups} = \prod_{n=1}^{N} \boldsymbol{\pi}_n^{b_n}.$$

where, $\mathbf{g} = (g_1, \cdots, g_K)$ is a vector containing the assignment of each segment to a group where $g_k \in \{1, \cdots, J\}$ is the group to which segment k is assigned; $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_J)$ are the probabilities of assigning segments to groups where π_j is the probability of assigning any segment to group j and b_j is the number of segments with $g_k = j$.

$$p(\mathbf{c}|\mathbf{g},\boldsymbol{\mu},\boldsymbol{\tau}) = \text{Probability of the ARMA mean for all segments} = \prod_{k=1}^{K} \mathcal{N}(c_k|\mu_{g_k}, \tau_{g_k}^2).$$

where, c_k is the mean signal level for segment k, that is, the mean of the ARMA model and $c_k \sim \mathcal{N}(\mu_{g_k}, \tau_{g_k}^2)$, where μ_{g_k} and $\tau_{g_k}^2$ are the mean and variance of the distribution of these means for the group g_k.

$$\begin{aligned} p(\mathbf{X}|K,\mathbf{s},\boldsymbol{\theta},\boldsymbol{\psi},\mathbf{c},\sigma^2) &= \text{The probability density of the observed signal} \\ &= \prod_{t=1}^{T} p(x_t|K,s,\theta,\psi,c,\sigma^2,x_{<t}) \\ &= \prod_{t=1}^{T} \mathcal{N}(x_t|\lambda_t,\sigma^2). \end{aligned}$$

with

$$x_t = c_k + \epsilon_t + \sum_{i=1}^{a} \psi_i(x_{t-i} - c_k) + \sum_{i=1}^{m} \theta_i \epsilon_{t-i}.$$

where, $\mathbf{X} = (x_t)_{t=1}^{T}$ represent the vector of time series or the signal that we want to segment; $\psi_1, \cdots, \psi_a$ are the parameters of the AR sub-model and $\theta_1, \cdots, \theta_m$ are the parameters of the MA sub-models where a and m denote the order of the AR and MA sub-models; $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_T)$ is the vector of error terms and $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$, where σ^2 is the variance of error terms.

Details of the posterior distributions for every parameter are given in [2]. To detect the boundaries of the event in Section 4.3.1, we plotted the change-point profiles and group

profiles. Change-point profiles are the posterior probability of being a change point at each position of the input series and it can be calculated using the following equation (Equation B.2)

$$\prod_{k=1}^{K}(s_k - 1) \times \frac{\phi^{K-1}(1-\phi)^{T-K-1}}{\int_\phi \phi^{K-1}(1-\phi)^{T-K-1}d\phi} \tag{B.2}$$

A running total of the position of change-points is kept in each segmentation. We used the start position of each segment to record the location of the change-points and incremented the count at that position.

Group profiles gives the posterior probability that each position in the input sequence belongs to one of the segment classes in the model and it can be determined using the following equation (Equation B.3)

$$\prod_{t=s_k}^{d_k} p(x_t|\lambda_t, \sigma^2) \times \frac{\prod_{k=1}^{K} \mathcal{N}(c_k|\mu_{g_k}, \tau_{g_k}^2) \times \prod_{k=1}^{K} \pi_{g_k}}{\int_g (\prod_{k=1}^{K} \mathcal{N}(c_k|\mu_{g_k}, \tau_{g_k}^2) \times \prod_{k=1}^{K} \pi_{g_k}) dg} \tag{B.3}$$

where, d_k is end position of the segment k and $\lambda_t = c_k + \sum_{i=1}^{a} \psi_i(x_{t-i} - c_k) + \sum_{i=1}^{m} \theta_i \epsilon_{t-i}$.

The above two posterior probabilities, that is, change-point profiles and group profiles are estimated by Monte Carlo integration. A sample is drawn from the posterior distribution, with each element in the sample consisting of values for K and π. The posterior probabilities are then calculated and averaged over the sample.

B.2 Algorithm of the proposed method using PCA as a summary statistic

input : Dataset with N rows and M columns

output: Probabilities of row change-point locations $\mathbf{t_r} = [t_{r1}, \ldots, t_{rN}]'$, probabilities of row group profile $\mathbf{p_r} = [p_{r1}, \ldots, p_N]'$, probabilities of column change point locations $\mathbf{t_c} = [t_{c1}, \ldots, t_{cM}]'$ and probabilities of column group profiles $\mathbf{p_c} = [p_{c1}, \ldots, p_{cM}]'$

1 Treat rows as observations;

2 Perform PCA ;

3 PCs $\rightarrow$ Get principal components that explain a significant percentage of variance ;

4 `/* We looked for the ''elbow'' point on the PCA scree plot and used the number of PCs associated with the elbow point. */`

5 l = number of PCs ;

6 Posterior_Prob_Of_CP_MA = zero matrix of $N \times 1$;

7 Posterior_Prob_Of_CP_AR = zero matrix of $N \times 1$;

8 Posterior_Prob_Of_CP_A.MA = zero matrix of $N \times 1$;

9 Posterior_Prob_Of_Change_Points = zero matrix of $N \times l$;

10 **for** *each PC* k *from* $1 : l$ **do**

11 $Posterior\_Prob\_Of\_CP\_MA = MA(1)\_Model(PC[k])$;

12 $Posterior\_Prob\_Of\_CP\_AR = AR(1)\_Model(PC[k])$;

13 $Posterior\_Prob\_Of\_CP\_ARMA = ARMA(1,1)\_Model(PC[k])$;

14 $Best\_model$ = Model with smallest DICV, from models $MA(1)$, $AR(1)$ and $ARMA(1,1)$;

15 $Posterior\_Prob\_Of\_Change\_Points$[column k] = $Posterior\_Prob\_Of\_CP\_Best\_model$ where $Best\_model \in (MA, AR, ARMA)$;

16 **end**

17 $\mathbf{t_r}$= $Max\_Posterior\_Prob\_Of\_Change\_Pt = \max(Posterior\_Prob\_Of\_Change\_Points)$ row-wise ;

18 `/* Here we get the maximum posterior probability of change-points for each row, from the posterior probability of all selected components. */`

19 Set *threshold* ;

20 `/* We used a threshold of 0.5 */`

21 Boundary of the event = the locations where $Max\_Posterior\_Prob\_Of\_Change\_Pt > threshold$;

22 Treat columns as observations;

23 Replace N with M and repeat from line 2 to line 21 to get $\mathbf{t_c}$.

Algorithm 3: Event detection with Bayesian change-point segmented ARMA model using PCA as a summary statistic

Appendix C

Appendix Chapter 5

C.1 Posterior Distributions

C.1.1 First Generalizations

Conditional Posterior Distribution of $\boldsymbol{\pi}_s$

The conditional posterior distribution of $\boldsymbol{\pi}_s$ is:

$$p(\boldsymbol{\pi}_s|\mathbf{X}_s, K, \mathbf{p}, \mathbf{g}_s, \mathbf{c}_s, \phi, \boldsymbol{\theta}_s, \boldsymbol{\psi}_s, \sigma_s^2, \boldsymbol{\mu}_s, \boldsymbol{\tau}_s) = \frac{p(\boldsymbol{\pi}_s, \mathbf{X}_s, K, \mathbf{p}, \mathbf{g}_s, \mathbf{c}_s, \phi, \boldsymbol{\theta}_s, \boldsymbol{\psi}_s, \sigma_s^2, \boldsymbol{\mu}_s, \boldsymbol{\tau}_s)}{\int_{\boldsymbol{\pi}_s} p(\boldsymbol{\pi}_s, \mathbf{X}_s, K, \mathbf{p}, \mathbf{g}_s, \mathbf{c}_s, \phi, \boldsymbol{\theta}_s, \boldsymbol{\psi}_s, \sigma_s^2, \boldsymbol{\mu}_s, \boldsymbol{\tau}_s) d\boldsymbol{\pi}_s}$$

Using Equation (7) from the main article, we get,

$$\frac{p(\phi) \prod_{s=1}^{S} p(\mathbf{X}_s, K, \mathbf{p}, \mathbf{g}_s, \mathbf{c}_s|\phi, \boldsymbol{\pi}_s, \boldsymbol{\theta}_s, \boldsymbol{\psi}_s, \sigma_s^2, \boldsymbol{\mu}_s, \boldsymbol{\tau}_s) p(\boldsymbol{\pi}_s) p(\boldsymbol{\theta}_s) p(\boldsymbol{\psi}_s) p(\sigma_s^2) p(\boldsymbol{\mu}_s) p(\boldsymbol{\tau}_s)}{\int_{\boldsymbol{\pi}_s} \prod_{s=1}^{S} p(\mathbf{X}_s, K, \mathbf{p}, \mathbf{g}_s, \mathbf{c}_s|\phi, \boldsymbol{\pi}_s, \boldsymbol{\theta}_s, \boldsymbol{\psi}_s, \sigma_s^2, \boldsymbol{\mu}_s, \boldsymbol{\tau}_s) p(\phi) p(\boldsymbol{\pi}_s) p(\boldsymbol{\theta}_s) p(\boldsymbol{\psi}_s) p(\sigma_s^2) p(\boldsymbol{\mu}_s) p(\boldsymbol{\tau}_s) d\boldsymbol{\pi}_s}$$

$$= \frac{p(\phi) p \prod_{s=1}^{S} (\mathbf{X}_s, K, \mathbf{p}, \mathbf{g}_s, \mathbf{c}_s|\phi, \boldsymbol{\pi}_s, \boldsymbol{\theta}_s, \boldsymbol{\psi}_s, \sigma_s^2, \boldsymbol{\mu}_s, \boldsymbol{\tau}_s) p(\boldsymbol{\pi}_s)}{\int_{\boldsymbol{\pi}_s} \prod_{s=1}^{S} p(\mathbf{X}_s, K, \mathbf{p}, \mathbf{g}_s, \mathbf{c}_s|\phi, \boldsymbol{\pi}_s, \boldsymbol{\theta}_s, \boldsymbol{\psi}, \sigma_s^2, \boldsymbol{\mu}_s, \boldsymbol{\tau}_s) p(\boldsymbol{\pi}_s) d\boldsymbol{\pi}_s}$$

Then, using Equation (6) and cancelling same terms from the numerator and the denominator the conditional posterior distribution of $\boldsymbol{\pi}_s$ becomes

$$\frac{p(K,\mathbf{p}|\phi)\prod_{s=1}^{S}p(\mathbf{X}_s|\boldsymbol{\lambda}_s,\sigma_s^2)p(\mathbf{c}_s|\mathbf{g}_s,\boldsymbol{\mu}_s,\boldsymbol{\tau}_s)p(\mathbf{g}_s|K,\boldsymbol{\pi}_s)p(\boldsymbol{\pi}_s)}{p(K,\mathbf{p}|\phi)\int_{\boldsymbol{\pi}_s}\prod_{s=1}^{S}p(\mathbf{X}_s|\boldsymbol{\lambda}_s,\sigma_s^2)p(\mathbf{c}_s|\mathbf{g}_s,\boldsymbol{\mu}_s,\boldsymbol{\tau}_s)p(\mathbf{g}_s|K,\boldsymbol{\pi}_s)p(\boldsymbol{\pi}_s)d\boldsymbol{\pi}_s} = \frac{\prod_{s=1}^{S}p(\mathbf{g}_s|K,\boldsymbol{\pi}_s)p(\boldsymbol{\pi}_s)}{\int_{\boldsymbol{\pi}_s}\prod_{s=1}^{S}p(\mathbf{g}_s|K,\boldsymbol{\pi}_s)p(\boldsymbol{\pi}_s)d\boldsymbol{\pi}_s}$$

Now, Using Equation (3), this becomes:

$$\begin{aligned}
&\frac{\prod_{s=1}^{S}\prod_{n_s=1}^{N_s}\boldsymbol{\pi}_{s,n_s}^{b_{s,n_s}}}{\int_{\boldsymbol{\pi}_s}\prod_{s=1}^{S}\prod_{n_s=1}^{N_s}\boldsymbol{\pi}_{s,n_s}^{b_{s,n_s}}d\boldsymbol{\pi}_s}\\
&= \frac{\prod_{s=1}^{S}\boldsymbol{\pi}_{s,1}^{(b_{s,1}+1)-1}\times\cdots\times\boldsymbol{\pi}_{s,N_s}^{(b_{s,N_s}+1)-1}}{\int_{\boldsymbol{\pi}_s}\prod_{s=1}^{S}\boldsymbol{\pi}_{s,1}^{(b_{s,1}+1)-1}\times\cdots\times\boldsymbol{\pi}_{s,N_s}^{(b_{s,N_s}+1)-1}d\boldsymbol{\pi}_s}\\
&= Dirichlet(\boldsymbol{\pi}_s|((b_{s,1}+1),\cdots,(b_{s,N_s}+1))
\end{aligned}$$

Conditional Posterior Distribution of ϕ

The conditional posterior distribution of ϕ is,

$$\frac{p(\phi)p(K,\mathbf{p}|\phi)\prod_{s=1}^{S}p(\mathbf{X}_s|\boldsymbol{\lambda}_s,\sigma_s^2)p(\mathbf{c}_s|\mathbf{g}_s,\boldsymbol{\mu}_s,\boldsymbol{\tau}_s)p(\mathbf{g}_s|K,\boldsymbol{\pi}_s)}{\int_{\phi}p(\phi)p(K,\mathbf{p}|\phi)d\phi\prod_{s=1}^{S}p(\mathbf{X}_s|\boldsymbol{\lambda}_s,\sigma_s^2)p(\mathbf{c}_s|\mathbf{g}_s,\boldsymbol{\mu}_s,\boldsymbol{\tau}_s)p(\mathbf{g}_s|K,\boldsymbol{\pi}_s)} = \frac{p(K,\mathbf{p}|\phi)p(\phi)}{\int_{\phi}p(K,\mathbf{p}|\phi)p(\phi)d\phi}$$

Using Equation (1), the conditional posterior distribution of ϕ becomes

$$\frac{\phi^{K-1}(1-\phi)^{T-K-1}}{\int_{\phi}\phi^{K-1}(1-\phi)^{T-K-1}d\phi} = Beta(\phi|K,T-K)$$

where, K = Number of segments and $T-K$ = Length of signal-number of segments.

Conditional Posterior Distribution of $\boldsymbol{\mu}_s$

The conditional posterior distribution of $\boldsymbol{\mu}_s$ is:

$$\frac{p(K,\mathbf{p}|\phi)\prod_{s=1}^{S}p(\mathbf{X}_s|\boldsymbol{\lambda}_s,\sigma_s^2)p(\mathbf{c}_s|\mathbf{g}_s,\boldsymbol{\mu}_s,\boldsymbol{\tau}_s)p(\mathbf{g}_s|K,\boldsymbol{\pi}_s)p(\boldsymbol{\mu}_s)}{\int_{\boldsymbol{\mu}_s}p(K,\mathbf{p}|\phi)\prod_{s=1}^{S}p(\mathbf{X}_s|\boldsymbol{\lambda}_s,\sigma_s^2)p(\mathbf{c}_s|\mathbf{g}_s,\boldsymbol{\mu}_s,\boldsymbol{\tau}_s)p(\mathbf{g}_s|K,\boldsymbol{\pi}_s)p(\boldsymbol{\mu}_s)d\boldsymbol{\mu}_s}=\frac{\prod_{s=1}^{S}p(\mathbf{c}_s|\mathbf{g_s},\boldsymbol{\mu_s},\boldsymbol{\tau_s})p(\boldsymbol{\mu}_s)}{\int_{\boldsymbol{\mu}_s}\prod_{s=1}^{S}p(\mathbf{c}_s|\mathbf{g_s},\boldsymbol{\mu_s},\boldsymbol{\tau_s})p(\boldsymbol{\mu}_s)d\boldsymbol{\mu}_s}$$

Note that, μ_s or $\mu_{s,n}$ will (currently) not be updated if there is less than one segment in group n_s. Then using Equation (4), the conditional distribution for $\mu_{s,n}$ (holding all the other μ_s's constant) becomes

$$\frac{\prod_{s,k:g_{s,k}=n_s}\mathcal{N}(c_{s,k}|\mu_{s,n_s},\tau^2_{s,n_s})}{\int_{\mu_s}\prod_{s,k:g_{s,k}=n_s}\mathcal{N}(c_{s,k}|\mu_{s,n_s},\tau^2_{s,n_s})d\mu_s}$$

.

Let m be the number of segments that have $g_{s,k}=n_s$. Then simplifying the product $\prod_{s,k:g_{s,k}=n_s}\mathcal{N}(c_{s,k}|\mu_{s,n_s},\tau^2_{s,n_s})=\prod_{s,k:g_{s,k}=n_s}\frac{1}{\sqrt{2\pi_s\tau^2_{s,n_s}}}\,exp\left[\left(\frac{(c_{s,k}-\mu_{s,n_s})^2}{2\tau^2_{s,n_s}}\right)\right]$, the conditional posterior distribution of $\boldsymbol{\mu_s}$ becomes a normal distribution with mean $\frac{\sum_{s,k:g_{s,k}=n}c_{s,k}}{m}$ and variance $\frac{\tau^2_{s,n_s}}{m}$.

Conditional Posterior Distribution of σ_s^2

The conditional posterior distribution of σ_s^2 is:

$$\frac{p(K,\mathbf{p}|\phi)\prod_{s=1}^{S}p(\mathbf{X}_s|\boldsymbol{\lambda}_s,\sigma_s^2)p(\mathbf{c}_s|\mathbf{g}_s,\boldsymbol{\mu}_s,\boldsymbol{\tau}_s)p(\mathbf{g}_s|K,\boldsymbol{\pi}_s)p(\sigma_s^2)}{\int_{\sigma_s^2}p(K,\mathbf{p}|\phi)\prod_{s=1}^{S}p(\mathbf{X}_s|\boldsymbol{\lambda}_s,\sigma_s^2)p(\mathbf{c}_s|\mathbf{g}_s,\boldsymbol{\mu}_s,\boldsymbol{\tau}_s)p(\mathbf{g}_s|K,\boldsymbol{\pi}_s)p(\sigma_s^2)d\sigma_s^2}=\frac{\prod_{s=1}^{S}p(\mathbf{X}_s|\boldsymbol{\lambda}_s,\sigma_s^2)p(\sigma_s^2)}{\int_{\sigma_s^2}\prod_{s=1}^{S}p(\mathbf{X}_s|\boldsymbol{\lambda}_s,\sigma_s^2)p(\sigma_s^2)d\sigma_s^2}$$

The prior distribution for σ_s^2 is: $p(\sigma_s^2,u_0,v_0)=\frac{v_0^{u_0}}{\Gamma(u_0)}(\sigma_s^2)^{-u_0-1}exp\left(-\frac{v_0}{\sigma_s^2}\right)$.
where u_0 and v_0 are the prior parameters. Now using $\boldsymbol{\epsilon}_s\sim\mathcal{N}(0,\sigma_s^2)$ and the above prior distribution and simplifying, the conditional posterior distribution of σ^2 takes the form of an inverse gamma distribution with parameter u and v, where $u=u_0+\frac{T}{2}$ and $v=\frac{2v_0+\sum_{t=1}^{T}\epsilon^2_{s,t}}{2}=v_0+\frac{1}{2}\sum_{t=1}^{T}\epsilon^2_{s,t}$.

Conditional Posterior Distribution of $\boldsymbol{\tau}^2_{s,n_s}$

Let $\boldsymbol{\tau}_s$ be the vector of $\tau^2_{s,1}, \cdots \tau^2_{s,N_s}$, then the conditional posterior distribution of $\boldsymbol{\tau}_s$ is:

$$\frac{p(K,\mathbf{p}|\phi)\prod_{s=1}^{S} p(\mathbf{X}_s|\boldsymbol{\lambda}_s,\sigma_s^2)p(\mathbf{c}_s|\mathbf{g}_s,\boldsymbol{\mu}_s,\boldsymbol{\tau}_s)p(\mathbf{g}_s|K,\boldsymbol{\pi}_s)p(\boldsymbol{\tau}_s)}{\int_{\boldsymbol{\tau}_s} p(K,\mathbf{p}|\phi)\prod_{s=1}^{S} p(\mathbf{X}_s|\boldsymbol{\lambda}_s,\sigma_s^2)p(\mathbf{c}_s|\mathbf{g}_s,\boldsymbol{\mu}_s,\boldsymbol{\tau}_s)p(\mathbf{g}_s|K,\boldsymbol{\pi}_s)p(\boldsymbol{\tau}_s)d\boldsymbol{\tau}_s} = \frac{\prod_{s=1}^{S} p(\mathbf{c}_s|\mathbf{g}_s,\boldsymbol{\mu}_s,\boldsymbol{\tau}_s)p(\boldsymbol{\tau}_s)}{\int_{\boldsymbol{\tau}_s}\prod_{s=1}^{S} p(\mathbf{c}_s|\mathbf{g}_s,\boldsymbol{\mu}_s,\boldsymbol{\tau}_s)p(\boldsymbol{\tau}_s)d\boldsymbol{\tau}_s}$$

The prior distribution for $\boldsymbol{\tau}_s$ is: $\quad p(\boldsymbol{\tau}_s,\alpha_0,\beta_0) = \prod_{n_s=1}^{N_s}\frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}(\tau^2_{s,n_s})^{-\alpha_0-1}exp\left(-\frac{\beta_0}{\tau^2_{s,n_s}}\right)$.
where α_0 and β_0 are the prior parameters. Note that τ^2_{s,n_s} will (currently) only be updated if there is more than one segment in group n_s. Now using Equation (4) and the above prior distribution and simplifying, the conditional posterior distribution of $\boldsymbol{\tau}_s$ takes the form of an inverse gamma distribution with parameter α and β, where $\alpha = \alpha_0 + \frac{|(s,k:g_{s,k}=n_s)|}{2}$ and $\beta = \beta_0 + \frac{1}{2}\sum_{s,k:g_{s,k}=n_s}(c_{s,k}-\mu_{s,n_s})^2$.

Conditional Posterior Distribution of $\mathbf{g}_s$

The conditional posterior distribution of $\mathbf{g}_s$ is:

$$\frac{p(K,\mathbf{p}|\phi)\prod_{s=1}^{S} p(\mathbf{X}_s|\boldsymbol{\lambda}_s,\sigma_s^2)p(\mathbf{c}_s|\mathbf{g}_s,\boldsymbol{\mu}_s,\boldsymbol{\tau}_s)p(\mathbf{g}_s|K,\boldsymbol{\pi}_s)p(\mathbf{g}_s)}{\int_{\mathbf{g}_s} p(K,\mathbf{p}|\phi)\prod_{s=1}^{S} p(\mathbf{X}_s|\boldsymbol{\lambda}_s,\sigma_s^2)p(\mathbf{c}_s|\mathbf{g}_s,\boldsymbol{\mu}_s,\boldsymbol{\tau}_s)p(\mathbf{g}_s|K,\boldsymbol{\pi}_s)p(\mathbf{g}_s)d\mathbf{g}_s}$$
$$= \frac{\prod_{s=1}^{S} p(\mathbf{c}_s|\mathbf{g}_s,\boldsymbol{\mu}_s,\boldsymbol{\tau}_s)p(\mathbf{g}_s|K,\boldsymbol{\pi}_s)p(\mathbf{g}_s)}{\int_{\mathbf{g}_s}\prod_{s=1}^{S} p(\mathbf{c}_s|\mathbf{g}_s,\boldsymbol{\mu}_s,\boldsymbol{\tau}_s)p(\mathbf{g}_s|K,\boldsymbol{\pi}_s)p(\mathbf{g}_s)d\mathbf{g}_s}$$

Using Equation (3) and Equation (4), The conditional posterior distribution of $\mathbf{g}_s$ becomes

$$\frac{\prod_{s=1}^{S}(\prod_{k=1}^{K}\mathcal{N}(c_{s,k}|\mu_{s,g_{s,k}},\tau^2_{s,g_{s,k}})\times\prod_{k=1}^{K}\pi_{s,g_{s,k}})}{\int_{g_s}\prod_{s=1}^{S}(\prod_{k=1}^{K}\mathcal{N}(c_{s,k}|\mu_{s,g_{s,k}},\tau^2_{s,g_{s,k}})\times\prod_{k=1}^{K}\pi_{s,g_{s,k}})dg_s}.$$

The $g_{s,k}$'s are independent, so can be updated one at a time with the conditional posterior distribution of g_s which is a discrete distribution with parameter $\frac{\mathcal{N}(c_{s,k}|\mu_{s,g_{s,k}},\tau^2_{s,g_{s,k}})\times\pi_{s,g_{s,k}}}{\int_{g_s}(\mathcal{N}(c_{s,k}|\mu_{s,g_{s,k}},\tau^2_{s,g_{s,k}})\times\pi_{s,g_{s,k}})dg_s}$.

Conditional Posterior Distribution of $\mathbf{c}_s$ and $\boldsymbol{\epsilon}_s$

The conditional posterior distribution of $\mathbf{c}_s$ and $\boldsymbol{\epsilon}_s$ is:

$$\frac{p(K,\mathbf{p}|\phi)\prod_{s=1}^{S} p(\mathbf{X}_s|\boldsymbol{\lambda}_s,\sigma_s^2)p(\mathbf{c}_s|\mathbf{g}_s,\boldsymbol{\mu}_s,\boldsymbol{\tau}_s)p(\mathbf{g}_s|K,\boldsymbol{\pi}_s)p(\mathbf{c}_s)}{\int_{\mathbf{c}_s} p(K,\mathbf{p}|\phi)\prod_{s=1}^{S} p(\mathbf{X}_s|\boldsymbol{\lambda}_s,\sigma_s^2)p(\mathbf{c}_s|\mathbf{g}_s,\boldsymbol{\mu}_s,\boldsymbol{\tau}_s)p(\mathbf{g}_s|K,\boldsymbol{\pi}_s)p(\mathbf{c}_s)d\mathbf{c}_s}$$
$$= \frac{\prod_{s=1}^{S} p(\mathbf{X}_s|\boldsymbol{\lambda}_s,\sigma_s^2)p(\mathbf{c}_s|\mathbf{g}_s,\boldsymbol{\mu}_s,\boldsymbol{\tau}_s)p(\mathbf{c}_s)}{\int_{\mathbf{c}_s}\prod_{s=1}^{S} p(\mathbf{X}_s|\boldsymbol{\lambda}_s,\sigma_s^2)p(\mathbf{c}_s|\mathbf{g}_s,\boldsymbol{\mu}_s,\boldsymbol{\tau}_s)p(\mathbf{c}_s)d\mathbf{c}_s}$$

Using Equation (4) and Equation (5) and $\boldsymbol{\epsilon} \sim \mathcal{N}(0,\sigma_s^2)$, the above equation becomes

$$\frac{\prod_{s=1}^{S}(\prod_{t=1}^{T} p(x_{s,t}|\lambda_{s,t},\sigma_s^2)\times\prod_{k=1}^{K}\mathcal{N}(c_{s,k}|\mu_{s,g_{s,k}},\tau_{s,g_{s,k}}^2))}{\int_{c_s}\prod_{s=1}^{S}(\prod_{t=1}^{T} p(x_{s,t}|\lambda_{s,t},\sigma_s^2)\times\prod_{k=1}^{K}\mathcal{N}(c_{s,k}|\mu_{s,g_{s,k}},\tau_{s,g_{s,k}}^2))dc_s}.$$

The $c_{s,k}$'s and the corresponding $\epsilon_{p_{s,k}},\cdots,\epsilon_{d_{s,k}}$ can be updated one segment at a time with the conditional distribution given by:

$$\frac{\prod_{s=1}^{S}(\prod_{t=p_k}^{d_k} p(x_{s,t}|\lambda_{s,t},\sigma_s^2)\times\mathcal{N}(c_{s,k}|\mu_{s,g_{s,k}},\tau_{s,g_{s,k}}^2))}{\int_{c_s}\prod_{s=1}^{S}(\prod_{t=p_k}^{d_k} p(x_{s,t}|\lambda_{s,t},\sigma_s^2)\times\mathcal{N}(c_{s,k}|\mu_{s,g_{s,k}},\tau_{s,g_{s,k}}^2))dc_s}.$$

Conditional Posterior Distribution of $\boldsymbol{\psi}_s$ and $\boldsymbol{\epsilon}_s$

The conditional posterior distribution of $\boldsymbol{\psi}_s$ and $\boldsymbol{\epsilon}_s$ is:

$$\frac{p(K,\mathbf{p}|\phi)\prod_{s=1}^{S} p(\mathbf{X}_s|\boldsymbol{\lambda}_s,\sigma_s^2)p(\mathbf{c}_s|\mathbf{g}_s,\boldsymbol{\mu}_s,\boldsymbol{\tau}_s)p(\mathbf{g}_s|K,\boldsymbol{\pi}_s)p(\boldsymbol{\psi}_s)}{\int_{\boldsymbol{\psi}_s} p(K,\mathbf{p}|\phi)\prod_{s=1}^{S} p(\mathbf{X}_s|\boldsymbol{\lambda}_s,\sigma_s^2)p(\mathbf{c}_s|\mathbf{g}_s,\boldsymbol{\mu}_s,\boldsymbol{\tau}_s)p(\mathbf{g}_s|K,\boldsymbol{\pi}_s)p(\boldsymbol{\psi}_s)d\boldsymbol{\psi}_s} = \frac{\prod_{s=1}^{S} p(\mathbf{X}_s|\boldsymbol{\lambda}_s,\sigma_s^2)p(\boldsymbol{\psi}_s)}{\int_{\boldsymbol{\psi}_s}\prod_{s=1}^{S} p(\mathbf{X}_s|\boldsymbol{\lambda}_s,\sigma_s^2)p(\boldsymbol{\psi}_s)d\boldsymbol{\psi}_s}$$

Using Equation (5) and $\boldsymbol{\epsilon}_s \sim \mathcal{N}(0,\sigma_s^2)$, the conditional posterior distribution of $\boldsymbol{\psi}_s$ and $\boldsymbol{\epsilon}_s$ given by:

$$\frac{\prod_{s=1}^{S}(\prod_{t=1}^{T} p(x_t|\lambda_{s,t},\sigma_s^2)\times p(\psi_s))}{\int_{\psi_s}\prod_{s=1}^{S}(\prod_{t=1}^{T} p(x_{s,t}|\lambda_{s,t},\sigma_s^2)\times p(\psi_s))d\psi_s}.$$

Conditional Posterior Distribution of $\boldsymbol{\theta}_s$ and $\boldsymbol{\epsilon}_s$

The conditional posterior distribution of $\boldsymbol{\theta}_s$ and $\boldsymbol{\epsilon}_s$ is:

$$\frac{p(K,\mathbf{p}|\phi)\prod_{s=1}^{S} p(\mathbf{X}_s|\boldsymbol{\lambda}_s,\sigma_s^2)p(\mathbf{c}_s|\mathbf{g}_s,\boldsymbol{\mu}_s,\boldsymbol{\tau}_s)p(\mathbf{g}_s|K,\boldsymbol{\pi}_s)p(\boldsymbol{\theta}_s)}{\int_{\boldsymbol{\theta}_s} p(K,\mathbf{p}|\phi)\prod_{s=1}^{S} p(\mathbf{X}_s|\boldsymbol{\lambda}_s,\sigma_s^2)p(\mathbf{c}_s|\mathbf{g}_s,\boldsymbol{\mu}_s,\boldsymbol{\tau}_s)p(\mathbf{g}_s|K,\boldsymbol{\pi}_s)p(\boldsymbol{\theta}_s)d\boldsymbol{\theta}_s} = \frac{\prod_{s=1}^{S} p(\mathbf{X}_s|\boldsymbol{\lambda}_s,\sigma_s^2)p(\boldsymbol{\theta}_s)}{\int_{\boldsymbol{\theta}_s}\prod_{s=1}^{S} p(\mathbf{X}_s|\boldsymbol{\lambda}_s,\sigma_s^2)p(\boldsymbol{\theta}_s)d\boldsymbol{\theta}_s}$$

Using Equation (5) and $\boldsymbol{\epsilon}_s \sim \mathcal{N}(0,\sigma_s^2)$, the conditional posterior distribution of $\boldsymbol{\theta}_s$ and $\boldsymbol{\epsilon}_s$ given by:

$$\frac{\prod_{s=1}^{S}(\prod_{t=1}^{T} p(x_{s,t}|\lambda_{s,t},\sigma_s^2)\times p(\theta_s))}{\int_{\theta_s}\prod_{s=1}^{S}(\prod_{t=1}^{T} p(x_{s,t}|\lambda_{s,t},\sigma_s^2)\times p(\theta_s))d\theta_s}.$$

C.1.2 Second Generalizations

All conditional posterior distributions for the first generalization are the same for the second except for $\boldsymbol{\pi}$ and $\mathbf{g}_s$ as this generalization assumes all time series have the same probability of assigning segments to groups ($\boldsymbol{\pi}$).

Conditional Posterior Distribution of $\boldsymbol{\pi}$

The conditional posterior distribution of $\boldsymbol{\pi}$ is:

$$\frac{p(K,\mathbf{p}|\phi)p(\boldsymbol{\pi})\prod_{s=1}^{S} p(\mathbf{X}_s|\boldsymbol{\lambda}_s,\sigma_s^2)p(\mathbf{c}_s|\mathbf{g}_s,\boldsymbol{\mu}_s,\boldsymbol{\tau}_s)p(\mathbf{g}_s|K,\boldsymbol{\pi})}{p(K,\mathbf{p}|\phi)\int_{\boldsymbol{\pi}}\prod_{s=1}^{S} p(\boldsymbol{\pi})p(\mathbf{X}_s|\boldsymbol{\lambda}_s,\sigma_s^2)p(\mathbf{c}_s|\mathbf{g}_s,\boldsymbol{\mu}_s,\boldsymbol{\tau}_s)p(\mathbf{g}_s|K,\boldsymbol{\pi})d\boldsymbol{\pi}} = \frac{p(\boldsymbol{\pi})\prod_{s=1}^{S} p(\mathbf{g}_s|K,\boldsymbol{\pi})}{\int_{\boldsymbol{\pi}} p(\boldsymbol{\pi})\prod_{s=1}^{S} p(\mathbf{g}_s|K,\boldsymbol{\pi})p(\boldsymbol{\pi})d\boldsymbol{\pi}}$$

Now, Using Equation (3), this becomes:

$$\frac{\prod_{s=1}^{S}\prod_{n_s=1}^{N_s}\boldsymbol{\pi}_{n_s}^{b_{s,n_s}}}{\int_{\boldsymbol{\pi}}\prod_{s=1}^{S}\prod_{n_s=1}^{N_s}\boldsymbol{\pi}_{n_s}^{b_{s,n_s}}d\boldsymbol{\pi}}$$
$$= \frac{\prod_{s=1}^{S}\boldsymbol{\pi}_1^{(b_{s,1}+1)-1}\times\cdots\times\boldsymbol{\pi}_{N_s}^{(b_{s,N_s}+1)-1}}{\int_{\boldsymbol{\pi}}\prod_{s=1}^{S}\boldsymbol{\pi}_1^{(b_{s,1}+1)-1}\times\cdots\times\boldsymbol{\pi}_{N_s}^{(b_{s,N_s}+1)-1}d\boldsymbol{\pi}}$$
$$= Dirichlet(\boldsymbol{\pi}|((b_{s,1}+1),\cdots,(b_{s,N_s}+1))$$

Conditional Posterior Distribution of $\mathbf{g}_s$

The conditional posterior distribution of $\mathbf{g}_s$ is:

$$\frac{\prod_{s=1}^{S}(\prod_{k=1}^{K}\mathcal{N}(c_{s,k}|\mu_{s,g_{s,k}},\tau^2_{s,g_{s,k}})\times\prod_{k=1}^{K}\pi_{g_{s,k}})}{\int_{g_s}\prod_{s=1}^{S}(\prod_{k=1}^{K}\mathcal{N}(c_{s,k}|\mu_{s,g_{s,k}},\tau^2_{s,g_{s,k}})\times\prod_{k=1}^{K}\pi_{g_{s,k}})dg_s}.$$

The $g_{s,k}$'s are independent, so can be updated one at a time with the conditional posterior distribution of g_s which is a discrete distribution with parameter $\frac{\mathcal{N}(c_{s,k}|\mu_{s,g_{s,k}},\tau^2_{s,g_{s,k}})\times\pi_{g_{s,k}}}{\int_{g_s}(\mathcal{N}(c_{s,k}|\mu_{s,g_{s,k}},\tau^2_{s,g_{s,k}})\times\pi_{g_{s,k}})dg_s}$.

C.1.3 Third Generalizations

All conditional posterior distributions of the second generalization are the same for the third except for $\mathbf{g}$, $\boldsymbol{\pi}$, $\boldsymbol{\mu}_s$, $\boldsymbol{\tau}^{\mathbf{2}}_{\boldsymbol{s,n}}$ and $\mathbf{c}_s$ and $\boldsymbol{\epsilon}_s$ as this generalization assumes common segment class $\mathbf{g}$ for all sequences.

Conditional Posterior Distribution of $\boldsymbol{\pi}$

The conditional posterior distribution of $\boldsymbol{\pi}$ is

$$\frac{\prod_{n=1}^{N}\boldsymbol{\pi}_n^{b_n}}{\int_{\boldsymbol{\pi}}\prod_{n=1}^{N}\boldsymbol{\pi}_n^{b_n}d\boldsymbol{\pi}} = \frac{\boldsymbol{\pi}_1^{(b_1+1)-1}\times\cdots\times\boldsymbol{\pi}_1^{(b_N+1)-1}}{\int_{\boldsymbol{\pi}}\boldsymbol{\pi}_1^{(b_1+1)-1}\times\cdots\times\boldsymbol{\pi}_1^{(b_N+1)-1}d\boldsymbol{\pi}} = Dirichlet(\boldsymbol{\pi}|((b_1+1),\cdots,(b_N+1))$$

Conditional Posterior Distribution of $\boldsymbol{\mu}_s$

The conditional posterior distribution of $\boldsymbol{\mu}_s$ is:

$$\frac{\prod\limits_{s,k:g_k=n} \mathcal{N}(c_{s,k}|\mu_{s,n},\tau^2_{s,n})}{\int_{\mu_s} \prod\limits_{s,k:g_k=n} \mathcal{N}(c_{s,k}|\mu_{s,n},\tau^2_{s,n})d\mu_s}$$

.

Let m be the number of segments that have $g_k = n$. Then simplifying the product $\prod\limits_{s,k:g_k=n} \mathcal{N}(c_{s,k}|\mu_{s,n},\tau^2_{s,n}) = \prod\limits_{s,k:g_k=n} \frac{1}{\sqrt{2\pi\tau^2_{s,n}}}\ exp\left[\left(\frac{(c_{s,k}-\mu_{s,n})^2}{2\tau^2_{s,n}}\right)\right]$, the conditional posterior distribution of $\boldsymbol{\mu_s}$ becomes a normal distribution with mean $\frac{\sum\limits_{s,k:g_k=n} c_{s,k}}{m}$ and variance $\frac{\tau^2_{s,n}}{m}$.

Conditional Posterior Distribution of $\boldsymbol{\tau^2_{s,n}}$

Let $\boldsymbol{\tau}_s$ be the vector of $\tau^2_{s,1},\cdots\tau^2_{s,N}$, then the conditional posterior distribution of $\boldsymbol{\tau}_s$ is:

$$\frac{\prod\limits_{s=1}^{S} p(\mathbf{c}_s|\mathbf{g},\boldsymbol{\mu}_s,\boldsymbol{\tau}_s)p(\boldsymbol{\tau}_s)}{\int_{\boldsymbol{\tau}_s} \prod\limits_{s=1}^{S} p(\mathbf{c}_s|\mathbf{g},\boldsymbol{\mu}_s,\boldsymbol{\tau}_s)p(\boldsymbol{\tau}_s)d\boldsymbol{\tau}_s}$$

The prior distribution for $\boldsymbol{\tau}_s$ is: $\quad p(\boldsymbol{\tau}_s,\alpha_0,\beta_0) = \prod\limits_{n=1}^{N} \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}(\tau^2_{s,n})^{-\alpha_0-1} exp\left(-\frac{\beta_0}{\tau^2_{s,n}}\right)$.
where α_0 and β_0 are the prior parameters. The conditional posterior distribution of $\boldsymbol{\tau}_s$ takes the form of an inverse gamma distribution with parameter α and β, where $\alpha = \alpha_0 + \frac{|(s,k:g_k=n)|}{2}$ and $\beta = \beta_0 + \frac{1}{2}\sum\limits_{s,k:g_k=n}(c_{s,k}-\mu_{s,n})^2$.

Conditional Posterior Distribution of g

The conditional posterior distribution of $\mathbf{g}$ is:

$$\frac{\prod\limits_{s=1}^{S}(\prod_{k=1}^{K}\mathcal{N}(c_{s,k}|\mu_{s,g_k},\tau^2_{s,g_k}) \times \prod\limits_{k=1}^{K}\pi_{g_k})}{\int_g \prod\limits_{s=1}^{S}(\prod_{k=1}^{K}\mathcal{N}(c_{s,k}|\mu_{s,g_k},\tau^2_{s,g_k}) \times \prod\limits_{k=1}^{K}\pi_{g_k})dg}.$$

The g_k 's are independent, so can be updated one at a time with the conditional posterior distribution of g which is a discrete distribution with parameter $\frac{\mathcal{N}(c_{s,k}|\mu_{s,g_k},\tau^2_{s,g_k}) \times \pi_{g_k}}{\int_g(\mathcal{N}(c_{s,k}|\mu_{s,g_k},\tau^2_{s,g_k}) \times \pi_{g_k})dg}$.

Conditional Posterior Distribution of $\mathbf{c}_s$ and $\boldsymbol{\epsilon}_s$

The $c_{s,k}$'s and the corresponding $\epsilon_{p_{s,k}}, \cdots, \epsilon_{d_{s,k}}$ can be updated one segment at a time with the conditional distribution given by:

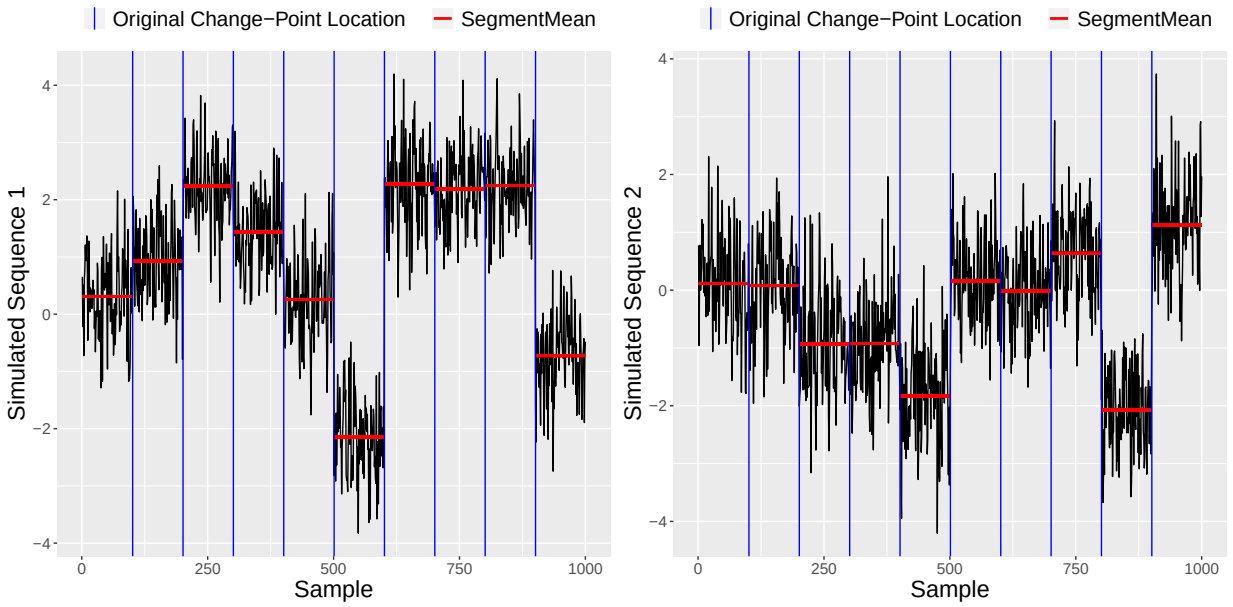
$$\frac{\prod_{s=1}^{S}(\prod_{t=p_k}^{d_k} p(x_{s,t}|\lambda_{s,t},\sigma_s^2) \times \mathcal{N}(c_{s,k}|\mu_{s,g_k},\tau_{s,g_k}^2))}{\int_{c_s}\prod_{s=1}^{S}(\prod_{t=p_k}^{d_k} p(x_{s,t}|\lambda_{s,t},\sigma_s^2) \times \mathcal{N}(c_{s,k}|\mu_{s,g_k},\tau_{s,g_k}^2))dc_s}.$$

C.2 Supplementary material

To test our method, we generated two more simulated data sets from AR model using different AR parameters. The first data set was simulated using $\psi_1 = 0.23$ and $\psi_2 = 0.21$ and the second data set was simulated using $\psi_1 = 0.91$ and $\psi_2 = 0.96$. Both sequences in those data sets were generated using same variance ($\sigma_1^2 = 0.49, \sigma_1^2 = 0.6$) and same 10 different segment means. Next we fitted AR model and ARMA model in each segment using all the generalizations and compared the location of change points and number of change points found by these two models in all generalizations.

First Dataset:

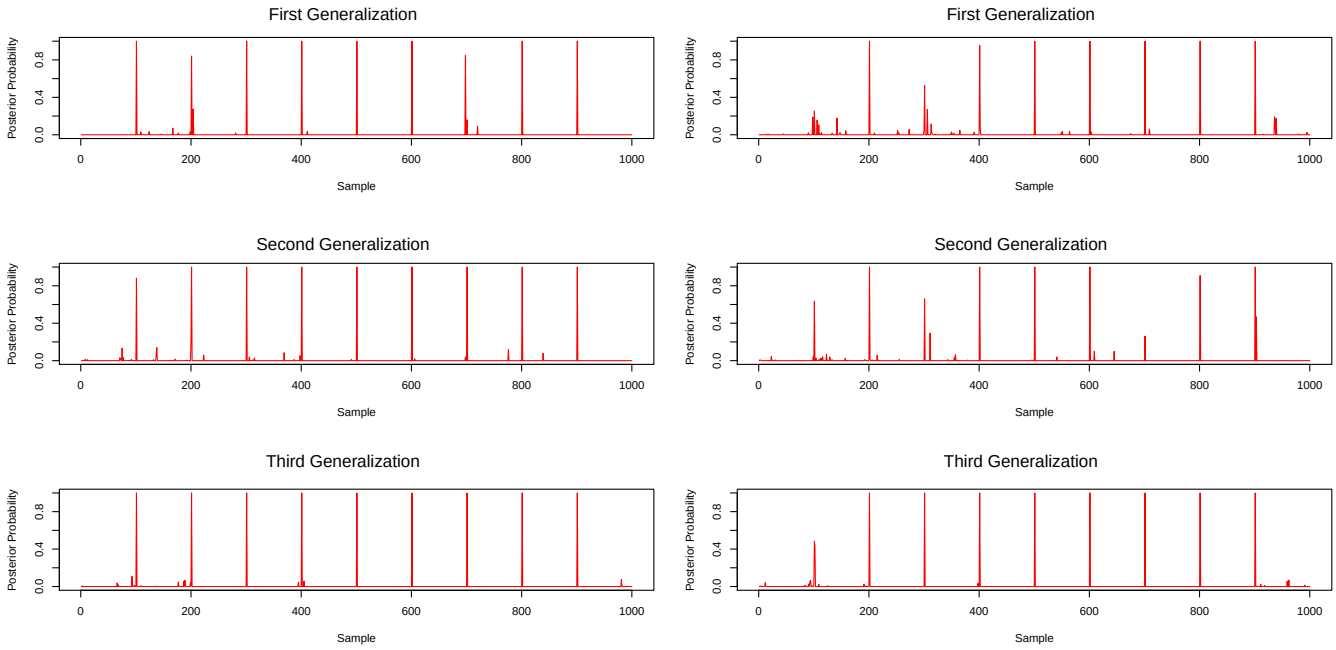
The simulated AR data (for $\psi_1 = 0.23$ and $\psi_2 = 0.21$) with the true segment means and the location of change points of both sequences are given in Figure C.1.



S4 Fig C.1: Simulated AR data with the true segment means and true change-point locations ($\psi_1 = 0.23$ and $\psi_2 = 0.21$)). The change-point locations are shown as a vertical blue line and segment means are shown as a horizontal red line.

Next I applied segmented AR(1) model and segmented ARMA(1,1) model in all generalizations. Figure **??** presents the estimated change-point locations for all the generalizations. If we compare the above plots, it is clear that the Segmented AR(1) model identified more change-points than segmented ARMA(1,1) model in all generalizations. And if we compare between all the generalizations in the segmented AR model, it is clear that the third

generalization identified all change-points with higher posterior probability.

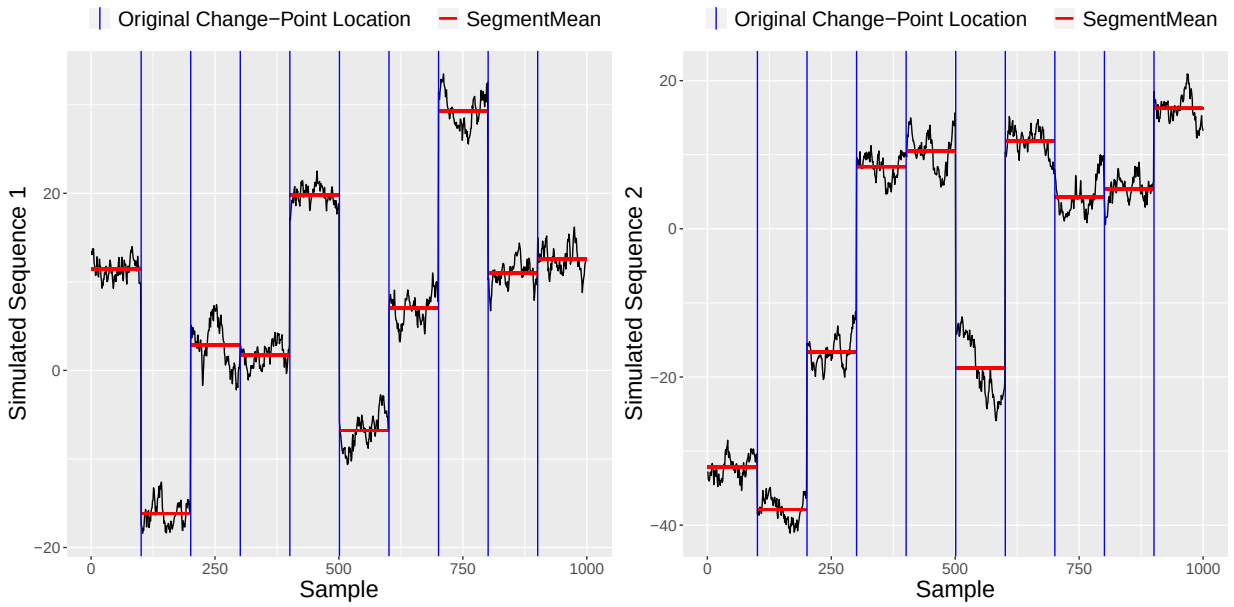


(a) Segmented AR(1) model (b) Segmented ARMA(1,1) model

S4 Fig C.2: Posterior distribution of occurrence of change-point locations.

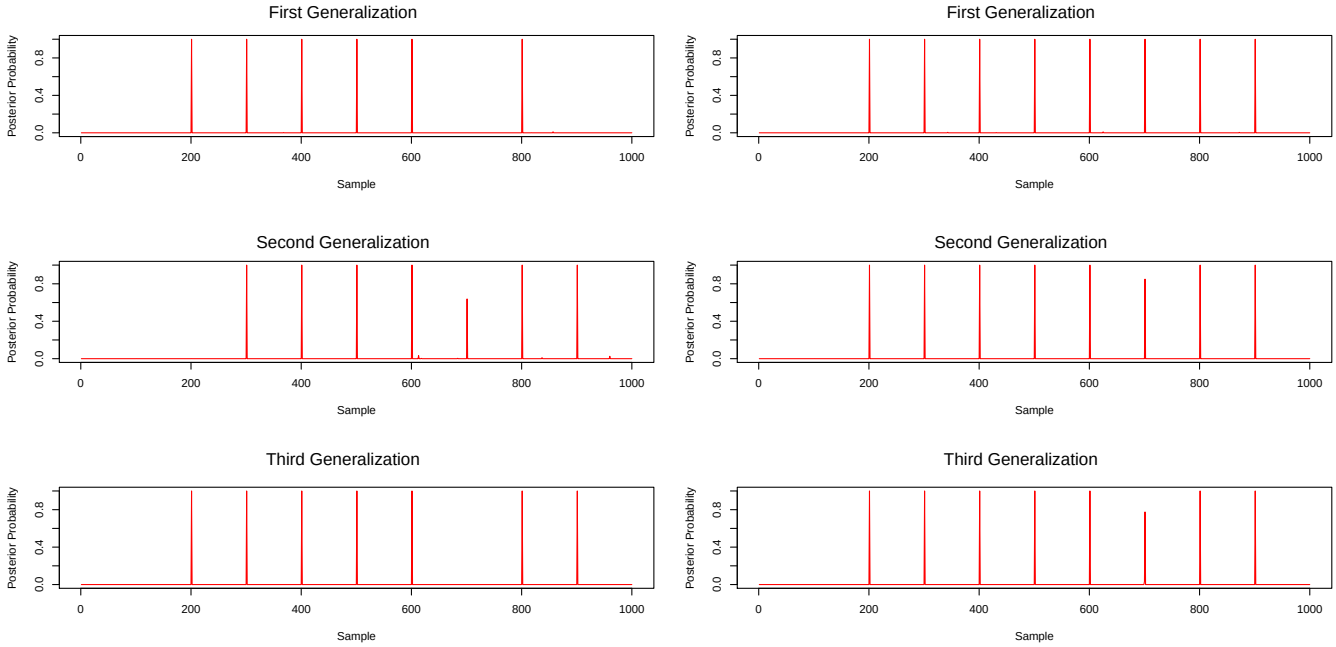
Second Dataset:

The simulated AR data (for $\psi_1 = 0.91$ and $\psi_2 = 0.96$) with the true segment means and the location of change points of both sequences are given in Figure C.3.



S4 Fig C.3: Simulated AR data with the true segment means and true change-point locations ($\psi_1 = 0.91$ and $\psi_2 = 0.96$)). The change-point locations are shown as a vertical blue line and segment means are shown as a horizontal red line.

Figure C.4 presents the estimated change-point locations for all the generalizations. Although this data was simulated from AR(1) model but here segmented ARMA(1,1) model identified more change-points than segmented AR(1) model in all generalizations. In segmented ARMA(1,1) model, the third and first generalization identified all change-points with higher posterior probability. These results indicate that the segmented ARMA model identifies more change-points than the AR model when the data are generated from an AR model with high parameter coefficient.



(a) Segmented AR(1) model (b) Segmented ARMA(1,1) model

S4 Fig C.4: Posterior distribution of occurrence of change-point locations.