Imperial College London
Faculty of Natural Sciences
Department of Mathematics

Bayesian changepoint models motivated by cyber-security applications

Karl Hallgren

July 2022

Supervised by Professor Nicholas Heard and Professor Niall Adams

Submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy in Mathematics of Imperial College London
and the Diploma of Imperial College London

Statement of originality

I certify that this thesis, and the research to which it refers, are the product of my own work, and that any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with the standard referencing practices of the discipline.

Karl Hallgren
July, 2022

Copyright

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a *Creative Commons Attribution - NonCommercial 4.0 International Licence (CC BY-NC).*

Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Acknowledgements

First, I would like to thank my supervisors, Nick Heard and Niall Adams, for their guidance and encouragement throughout my PhD. I am grateful for the many engaging discussions with them that have shaped my approach to research, for their eagerness to help with all aspects of the PhD, and for creating fantastic opportunities for my professional development, including an internship at Los Alamos National Laboratory (LANL).

I would also like to thank Melissa Turcotte who supervised me during the internship at LANL. Her expertise and enthusiasm made the internship a great experience.

I am thankful for the financial support received from the Engineering and Physical Sciences Research Council (EPSRC) and the Department of Mathematics at Imperial College London.

Finally, I would like to thank my family for their support throughout my PhD.

Abstract

Changepoint detection has an important role to play in the next generation of cyber security defenses. A cyber attack typically changes the behaviour of the target network. Therefore, to detect the presence of a network intrusion, it can be informative to monitor for changes in the high-volume data sources that are collected inside an enterprise computer network. However, most traditional changepoint detection methods are not adapted to characterise what cyber security analysts mean by a change, and consequently raise too many false alerts but also overlook weak signals that are suggestive of a real attack. This thesis will present three novel Bayesian changepoint models that address some challenges raised by cyber data: the first model combines evidence across a graph of time series to identify patterns of changepoints that are *a priori* more likely to correspond to an attack; the second model offers robustness to non-exchangeable data within segments so that normal dynamic phenomena observed in cyber data can be captured; and, the third model relaxes the standard assumption that changes are instantaneous, so that time intervals where cyber data may be subject to non-instantaneous changes can be identified.

Table of Contents

page

Acknowledgements **3**

Abstract **4**

Table of Contents **5**

1 Introduction **9**

1.1 Los Alamos National Laboratory (LANL) computer network data 11

1.1.1 Authentication event data 11

1.1.2 Network event data 12

1.2 Outline of the thesis 14

1.3 List of publications 17

2 Bayesian changepoint analysis **18**

2.1 Standard Bayesian changepoint modelling 20

2.2 Monte Carlo integration 22

2.2.1 Classical Monte Carlo integration 22

2.2.2 Importance sampling 22

2.2.3 Sequential importance sampling (SIS) 23

2.3 Inference via Markov chain Monte Carlo (MCMC) methods 24

2.3.1 Metropolis-Hastings algorithm 25

2.3.2 Gibbs sampling 25

2.3.3 Reversible jump Markov Chain Monte Carlo (RJMCMC) 26

2.3.4 Auxiliary variable methods for MCMC 28

2.4 Loss function for changepoint parameter estimation 30

2.5 Changepoint detection in network authentication data 32
2.6 Conclusion 33
3 Graph of time series 35
3.1 Motivation: encoding prior beliefs on cyber attacks 37
3.1.1 Monitoring the authentication activity of users 37
3.1.2 Motivation for an informative graph-based changepoint prior 39
3.2 Informative prior for changepoints on a graph of time series 41
3.2.1 Synchronous dependent changepoints across time series 42
3.2.2 Examples of graphical dependence structures for changepoints 44
3.2.3 Extension to asynchronous changepoint dependence 46
3.3 Inference via RJMCMC with auxiliary variables 48
3.3.1 Sampler for synchronous dependent changepoints 49
3.3.2 Sampler for asynchronous dependent changepoints 51
3.3.3 Time complexity and initialisation of the sampler 53
3.3.4 Possible sampling algorithm extensions for segment parameters 54
3.4 Simulation study 55
3.4.1 Synthetic data 55
3.4.2 Changepoint inference 56
3.4.3 MCMC diagnostics and runtime 57
3.4.4 Changepoint estimation results 59
3.5 Red team detection in network authentication data from LANL 63
3.5.1 Presence of a red team 63
3.5.2 Changepoint modelling 63
3.5.3 Results 64
3.6 Conclusion 67
4 Non-exchangeable data 69
4.1 Moving-sum model 72
4.1.1 Model definition 72
4.1.2 Bayesian changepoint analysis with moving-sums 74
4.1.3 Simulation from the model 74
4.2 Conditional likelihood 75
4.2.1 Relationship between the observed data and the latent variables 75
4.2.2 Conditional likelihood of the data within a segment 77
4.3 Analysis of the latent parameter space 78
4.3.1 Characterisation of the parameter space in terms of the observed data 79
4.3.2 Adding and removing data 81

4.3.3 Different ranges of dependence . . . 81
4.3.4 Asymptotic properties of the parameter space . . . 82
4.4 Inference via MCMC 84
4.4.1 Description of the sampler . . . 85
4.4.2 Sampler initialisation . . . 88
4.4.3 Parameter estimation . . . 88
4.4.4 Validating the estimated orders of dependence . . . 88
4.5 Simulations 88
4.5.1 Synthetic data . . . 88
4.5.2 Model comparison . . . 89
4.5.3 Sampler convergence . . . 90
4.5.4 Runtime . . . 91
4.5.5 Results . . . 92
4.6 Applications 93
4.6.1 Enterprise-wide computer network traffic monitoring . . . 93
4.6.2 Financial time series segmentation . . . 95
4.7 Conclusion 97
5 Change interval detection 98
5.1 From changepoints to change intervals 100
5.1.1 Model definition . . . 101
5.1.2 Simulation from the model . . . 102
5.2 Posterior distribution 103
5.3 Reversible jump MCMC 105
5.3.1 Update segment parameters . . . 106
5.3.2 Birth and death of zero-length intervals . . . 107
5.3.3 Update interval lengths . . . 108
5.3.4 Shift move . . . 108
5.3.5 Merge and split of change intervals . . . 108
5.3.6 Initialisation of the sampler . . . 108
5.3.7 Change interval estimates . . . 109
5.4 Simulation study 109
5.4.1 Synthetic data . . . 109
5.4.2 Model comparison . . . 109
5.4.3 Runtime . . . 110
5.4.4 Sampler convergence . . . 111
5.4.5 Results . . . 111
5.5 Computer network traffic monitoring 112

5.6 Conclusion 115

6 Conclusion 116

6.1 Summary 116

6.2 Future work 118

Bibliography 120

A Simulation studies 126

A.1 Graph of time series 126

A.1.1 MCMC diagnostics 126

A.1.2 Detection of clusters of changepoints on a graph of time series 127

A.1.3 Extension for asynchronous changepoints 127

A.2 Non-exchangeable data 131

A.3 Change intervals 133

B Proofs 135

B.1 Proof of Proposition 1 135

B.2 Proof of Proposition 3 136

B.3 Proof of Proposition 4 137

B.4 Order of dependence estimation 138

1

Introduction

Changepoint detection is concerned with identifying the times when structural changes occur in time series. The field originated when Shewhart (1931) proposed control charts to monitor industrial processes, that could be *in control* or *out of control*. Since then, change detection has emerged as a relevant problem in several domains, including epidemiology (Sparks, Keighley, and Muscatello, 2011), oceanography (Killick et al., 2010), network security (Tartakovsky, Polunchenko, and Sokolov, 2012) and genetics (Olshen et al., 2004).

What is meant by a "change" depends on the context. Consequently, many interesting developments in the field have resulted from the need to design changepoint detection methods that are adapted to new areas of application. For example, Fearnhead and Rigaill (2019) consider applications where abrupt changes in the data occurring for a short period of time are not of interest, such as the analysis of well-log data. Well-log data, which are analysed to detect changes in rock strata, are subject to outliers resulting from measurement errors of the probe used to collect data into a borehole. Fearnhead and Rigaill (2019) show that standard methods for detecting changepoints are not adapted to well-log data because they are not robust to outliers, and therefore fit more changepoints than necessary. Motivated by this limitation of standard changepoint models, Fearnhead and Rigaill (2019) propose a novel method for changepoint detection in the presence of outliers, resulting in suitable segmentations of well-log data. Similarly, changepoint detection methods proposed in Bardwell and Fearnhead (2017), Wang and Samworth (2018), and Fisch, Eckley, and Fearnhead (2022) are motivated by an application that challenges standard changepoint detection methods for multivariate time series, namely the detection of copy number variants (CNV's) in DNA sequences. This thesis proposes changepoint models that are motivated by cyber-security applications.

Since organisations and individuals increasingly rely on computers, cyber security is a pressing concern. Network intruders typically leave evidence for malicious activities in computer network

data (Morgan et al., 2016). In computer networks, multiple high-volume data sources can be leveraged to prevent and detect cyber-attacks. Examples of data that can be collected in computer networks include records of connections between computers, user authentication events, Domain Name System (DNS) logs and web traffic logs. Traditional attack detection methods focus on identifying predetermined signatures for known attacks, such as file hashes, IP addresses and domains previously associated with malicious activity. Signature-based detection methods are vulnerable to previously unobserved attacks, known as zero-day attacks. To complement signature-based detection methods, anomaly detection methods, which aim to detect deviations from normal behaviour, can be considered. Chandola, Banerjee, and Kumar (2009) provide a survey of various anomaly detection techniques. In the context of cyber-security, it is of interest to identify unusual behaviours of computer networks that may be evidence of attacks. For example, Neil et al. (2013) consider scan statistics to detect unusual network traversals using legitimate credentials. Marchette and Wegman (2012), Adams and Heard (2014), Adams and Heard (2016) discuss other examples of statistical anomaly detection methods for cyber security.

In the context of cyber-security, anomaly detection can be seen as a changepoint detection problem. A cyber attack typically changes the behaviour of some aspects of the target network. Therefore, to detect the presence of a network intrusion, it can be informative to monitor for changes in the high-volume data sources that are collected inside an enterprise computer network. For example, consider the WannaCry ransomware attack that affected important organisations such as the NHS in 2017. The attack relied on exploiting a vulnerability in port 445 in machines running the Microsoft operating system, resulting in significant spikes of activity on port 445 (CERT-EU, 2017). Consequently, monitoring network traffic for changes within a targeted enterprise network could have revealed the attack, making it possible to trigger timely responses to stop or limit the impact of the ransomware. Examples of analyses demonstrating changepoint detection methods can be instrumental in identifying cyber-attacks include Wang, Zhang, and Shin (2004), which detect Denial of Service (DoS) attacks by monitoring changes in the number of connections made to an IP address, and Turcotte, Heard, and Neil (2014), which propose a two-stage approach to detect anomalous behaviours in computer networks, relying on changepoint detection in hierarchical Bayesian models.

However, most traditional changepoint detection methods are not adapted to characterise what cyber-security analysts mean by a change, and consequently raise too many false alerts and overlook weak but coordinated signals that are suggestive of real attacks. This thesis identifies key features in cyber data that challenge traditional changepoint detection methods, and proposes changepoint models that are adapted to these features.

1.1. Los Alamos National Laboratory (LANL) computer network data

This section presents computer network data that will be used in this thesis to motivate and demonstrate the benefits of the proposed changepoint models. Los Alamos National Laboratory (LANL) released multiple data sets consisting of the typical data sources that are available in an enterprise computer network for the discovery and prevention of cyber-attacks. Two data sets collected from LANL enterprise network will be considered in this thesis: LANL2015 (Kent, 2015) and LANL2017 (Turcotte, Kent, and Hash, 2017).

1.1.1. Authentication event data

Kent (2015) presents a comprehensive data set summarising 58 days of traffic on the enterprise computer network of LANL, which is available online at `https://lanl.ma.ic.ac.uk/data/cyber1`. The network authentication data consist of events describing the authentication activity of users connecting from one computer to another. Each event contains the fields given in Table 1.1.

Different methods for the detection of network intrusions in computer networks have been demonstrated on the authentication logs from LANL. For example, based on the observation that new edges, defined as connections between a source computer and a destination computer not previously observed, might suggest the presence of intruders or malicious implants, Metelli and Heard (2019a) propose an anomaly detection method based on a Bayesian model for simultaneously characterising existing network structure and modelling likely new edge formation, which is demonstrated on LANL data.

To motivate a novel changepoint model on a graph of time series, Chapter 3 considers network authentication data. With the aim of modelling network activity that is human driven rather than high frequency and automated, the network authentication data were filtered as follows: events involving

Field Name	**Description**
Time	Time of the authentication event (one second resolution)
SrcUser@Domain	User initiating the authentication event
DstUser@Domain	User that the authentication event is mapping to
SrcComputer	Source computer
DstComputer	Destination computer
AuthType	Type of authentication (e.g. Negotiate, Kerberos etc.)
LogonType	Type of logon (e.g. Batch, System service etc.)
Orientation	Orientation
Status	Status of the request: success or failure

Table 1.1.: *Description of the fields of the authentication data recorded at LANL (Kent, 2015).*

Field Name	Description
Time	Start time of the event in epoch time format
Duration	Duration of the event in seconds
SrcDevice	Device which likely initiated the communication event
DstDevice	Receiving device
Protocol	Protocol number
SrcPort	Port used by *SrcDevice*
DstPort	Port used by *DstDevice*
SrcPackets	Number of packets the SrcDevice sent during the event
DstPackets	Number of packets the DstDevice sent during the event

Table 1.2.: *Description of the network event data recorded at LANL (Turcotte, Kent, and Hash, 2017).*

the same pair of user and source computer more than 5 000 times were removed; computers that act as source computers for more than 300 distinct users were discarded. The resulting data involves 11 985 distinct users and 12 947 distinct source computers.

The application consists of monitoring for changes in the authentication activity of each user in the network to identify the presence of a network intrusion. Let V denote the set of users in the enterprise. To detect occurrences of malicious activity in the network, the authentication activity of each user $i \in V$ is monitored via hourly counts of network logons per source computer. Let M denote the number of distinct source computers in the network. For each user $i \in V$, let

$$x_{i,t} = (x_{i,t,1}, \ldots, x_{i,t,M}), \tag{1.1}$$

where $x_{i,t,\ell}$ denotes the number of network logons initiated by user i from source computer ℓ during the t-th hour of the 58 day data collection period. Figure 1.1 displays the authentication data for two users. For each user, it is of interest to detect temporal changes in the distribution of network logons across source computers as possible evidence for malicious activity.

1.1.2. Network event data

Turcotte, Kent, and Hash (2017) present a data set summarising 90 days of network and computer (host) events collected from the LANL enterprise network, which is available online at `http://lanl.ma.ic.ac.uk/data/2017/`. The network event data originated from many of the internal enterprise routers within the LANL enterprise network. Each recorded network event gives the start time and the duration of a transfer of packets from a network device to another. Each event contains the fields given in Table 1.2. The data contain 1 286 000 events per day on average. Note that, a *destination port* is associated to each network event, which describes the purpose of the transfer of packets: for example, web, email, remote login or file transfer. For example, the port with associated number 80 usually corresponds to a web browser communication (HTTP) and the

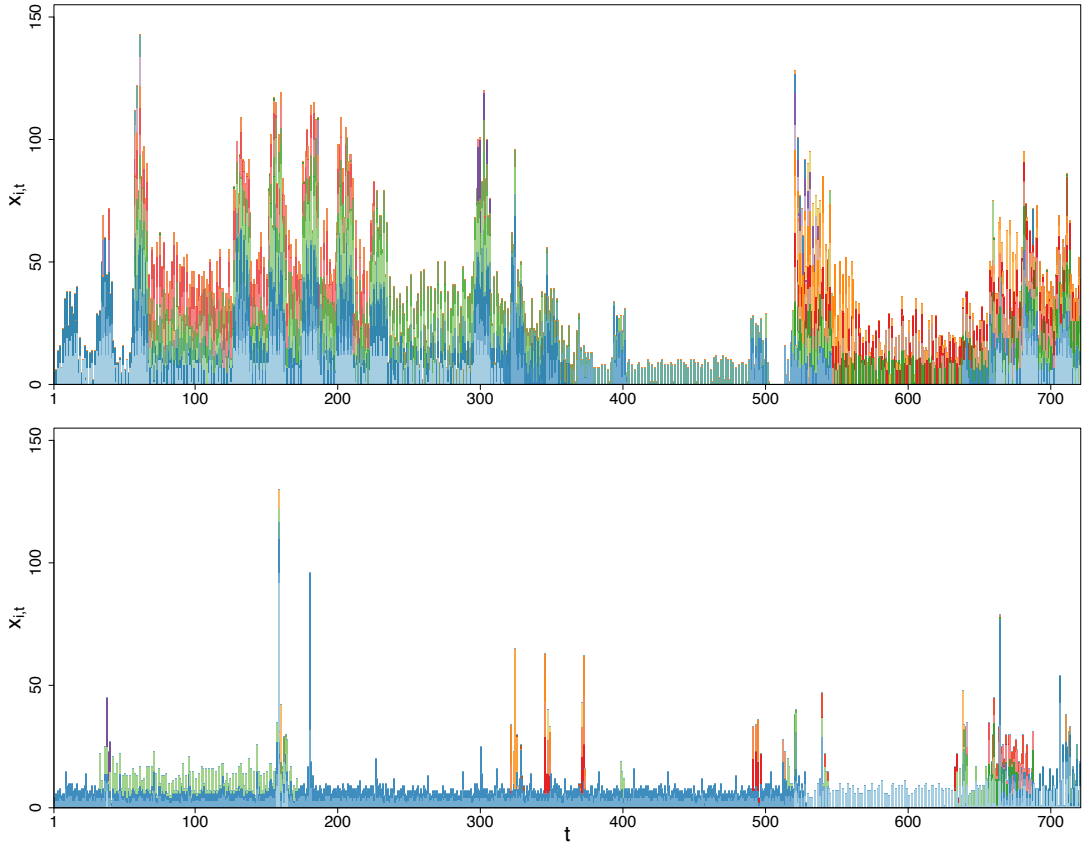


Figure 1.1.: *Hourly counts of network logons per source computer for two users over the first month of data collection. Source computers are identified by distinct colours. Top panel corresponds to user U342@DOM1, and bottom panel to user U86@DOM1.*

port 443 to a secure web browser communication (HTTPS).

Chapter 4 and Chapter 5 consider network event data from LANL, which are described in Example 1.1 and Example 1.2, to demonstrate novel changepoint models. In both cases, it is of interest to monitor for changes in the volumes of different types of traffic passing through a network over time to detect the presence of a network intrusion.

Example 1.1 (Traffic passing through the most common destination ports)**.** To restrict the analysis to the most common types of network activities, events that do not correspond to the 100 most recurrent destination ports in the data can be discarded. Consider the data $x_1, \ldots, x_T$ where, for all t, x_t denotes the number of network events, corresponding to the 100 most recurrent destination ports, that are in progress across the network during the t-th second between 10:00 and 10:20 on day 22 of the data collection period. The data are displayed in Figure 1.2.

Example 1.2 (Traffic passing through a single destination port)**.** The focus may be on a single type of activity. For example, consider the data $x_1, \ldots, x_T$ where, for all t, x_t denotes the number of

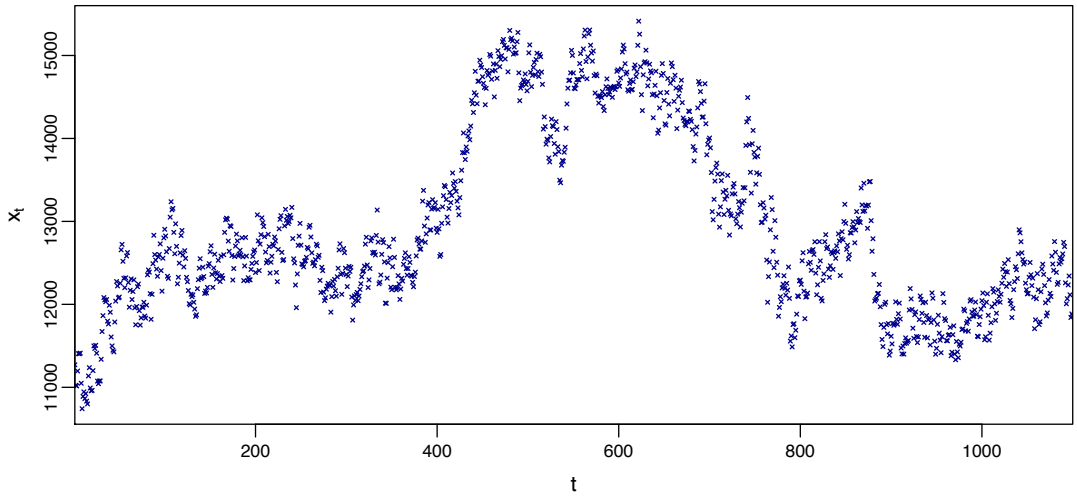


Figure 1.2.: *Counts of computer network events for the* 100 *most common destination ports each second between 10:00 and 10:20 on day* 22 *of the data collection period.*

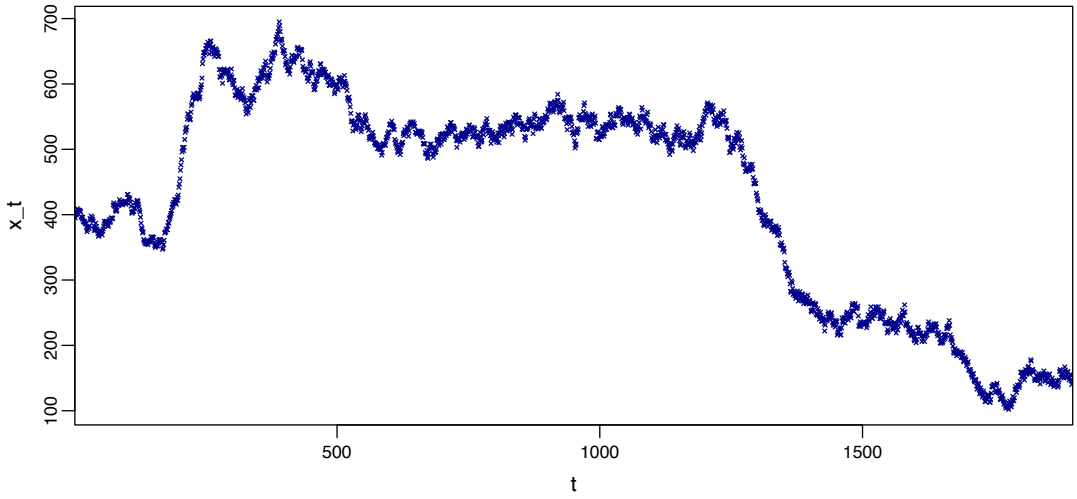


Figure 1.3.: *Counts of computer network events for the anonymised destination port 'Port20990' each second between 11:40 and 12:30 on day* 22 *of the data collection period.*

network events for the anonymised destination port 'Port20990' that are in progress across the network during the t-th second between 11:45 and 12:15 on day 22 of the data collection period. The data are displayed in Figure 1.3.

1.2. Outline of the thesis

This section describes the structure of the thesis and it summarises the main contributions of our work.

Chapter 2 provides some background for the work presented in the thesis. After a discussion on different approaches to changepoint modelling, a standard Bayesian changepoint model is presented, which provides the basis for the novel changepoint models proposed in the thesis. Inference via simulations is required for the proposed models, and therefore some Markov Chain Monte Carlo methods, which are used in this thesis, are reviewed. Moreover, standard approaches to give an account of the posterior distribution of changepoints are discussed, and a novel loss function for changepoints is proposed, which takes into account both the number and the positions of changepoints.

Chapters 3, 4 and 5 propose three different novel changepoint models motivated by cyber-security applications. The three chapters, which may be read independently, build upon the Bayesian framework for changepoint analysis discussed in Chapter 2. Chapter 3 considers a setting with multiple time series subject to changepoints, whereas Chapters 4 and 5 focus on a setting with a single time series subject to changepoints.

The work discussed in Chapter 3 is motivated by the commonly held belief of security experts that cyber attacks tend to correspond to coordinated sequences of events across multiple connected endpoints on the network (Sexton, Storlie, and Neil, 2015). For example, some attacks are most likely to be identified through a chain of behavioural changes across related activity types on the same machine, or across machines linked by network connectivity. In other words, there is a requirement for changepoint detection methods that combine evidence from multiple sources to detect changepoints, whilst taking into account which structures of changepoints are *a priori* likely to be of interest. To address this challenge, Chapter 3 considers the problem of jointly analysing multiple time series that may be subject to changepoints, with a focus on settings where it is sometimes possible to specify *a priori*, by means of a graph, which pairs of time series are likely to be impacted by simultaneous changepoints. Chapter 3 proposes a novel informative prior for changepoints that encodes the information contained in the graph, inducing a changepoint model for multiple time series that borrows strength across clusters of connected time series to detect weak signals for synchronous changepoints. The graphical model for changepoints is further extended to allow dependence between nearby but not necessarily synchronous changepoints across neighbouring time series in the graph. A novel reversible jump MCMC algorithm making use of auxiliary variables is proposed to sample from the graphical changepoint model. The merit of the proposed approach is demonstrated through a changepoint analysis of computer network authentication logs from LANL, demonstrating an improvement at detecting weak signals for network intrusions across users linked by network connectivity, whilst limiting the number of false alerts.

Chapter 4 originates from the lack of robustness of traditional changepoint detection methods to normal dynamic phenomena one may observe in temporal data. Cyber data are often subject to gradual population changes, seasonal variations and other temporal trends that are unlikely to be

evidence for malicious activity. Changepoint models typically assume the data within each segment are independent and identically distributed conditional on some parameters that change across segments. This construction may be inadequate when data are subject to local correlation patterns, often resulting in many more changepoints fitted than preferable. Chapter 4 proposes a Bayesian changepoint model that relaxes the assumption of exchangeability within segments. The proposed model supposes data within a segment are m-dependent for some unknown $m \geqslant 0$ that may vary between segments, resulting in a model suitable for detecting clear discontinuities in data that are subject to different local temporal correlations. The approach is suited to both continuous and discrete data. A novel reversible jump MCMC algorithm is proposed to sample from the model; in particular, a detailed analysis of the parameter space is exploited to build proposals for the orders of dependence. Two applications demonstrate the benefits of the proposed model: computer network monitoring via change detection in network data from LANL, and segmentation of financial time series.

Chapter 5 proposes another changepoint model motivated by cyber-security applications. It emanates from the observation that cyber data may be subject to gradual changes that occur over some time intervals. For example, volumes of traffic passing through the entire network over time may be impacted by gradual changes, as attackers traverse the network, progressively taking control of more computers. Changepoint models typically assume that changes occur instantly at changepoints, resulting in many changepoints fitted throughout time intervals where the distribution of the data changes. To characterise changes impacting the distribution of the data, it would be preferable to identify time intervals where changes gradually occur. Chapter 5 proposes an extension of the standard changepoint model, that relaxes the assumption that changes occur abruptly from one segment to the next. Changepoint parameters are augmented with time window parameters, thereby defining change intervals of possibly unequal lengths: each changepoint location identifies the position of a time interval, where a segment parameter transitions to the next segment parameter via a random walk, which is not necessarily smooth and may begin or end with a clear discontinuity, thereby encompassing various forms of transitions. The proposed model is suitable for both continuous and count data. A standard reversible jump MCMC algorithm for parametric changepoint models is extended to sample from the change interval model. An application in cyber-security is discussed to demonstrate the benefits of the proposed model: computer network monitoring via change detection in network data collected at LANL.

Chapter 6 concludes the thesis with a summary of the main contributions of our work, and discusses future work.

1.3. List of publications

Three articles will result from the work presented in this thesis. Two articles have been accepted for publication in academic journals: Hallgren, Heard, and Turcotte (2023) is based on Chapter 3 and also propose the novel loss function defined in Section 2.4, and Hallgren, Heard, and Adams (2022) is based on Chapter 4. The third article, yet to be submitted, is based on Chapter 5.

Python implementations of the samplers proposed in Chapter 3 and in Chapter 4 are available in the GitHub repositories `https://github.com/karl-hallgren/cp_on_graph_of_timeseries` and `https://github.com/karl-hallgren/mvsum`, restrospectively. Code for the sampler proposed in Chapter 5 is yet to be published.

.

2

Bayesian changepoint analysis

Early works on changepoint detection include Page (1954), Shiryaev (1963) and Hinkley (1970). Since then, the field has evolved rapidly, resulting in a vast literature on changepoint modelling. Changepoint models assume a possibly unknown number of changepoints define disjoint segments, where the data observed within each segment are assumed to arise from a single generative model, that changes from one segment to the next. Changepoint detection is concerned with finding the number of these changes and their locations in time.

The literature on changepoint detection is extensive both for the frequentist and for the Bayesian approaches. This thesis proposes novel Bayesian changepoint models and therefore the literature review will predominantly focus on Bayesian changepoint analysis. Due to the broad nature of the changepoint detection literature and the existence of numerous, engaging surveys on the topic (Eckley, Fearnhead, and Killick, 2011; Horváth and Rice, 2014; Aminikhanghahi and Cook, 2017; Truong, Oudre, and Vayatis, 2020), this chapter is meant to be a selective review of methods that will be built upon in the remainder of the thesis, rather than an exhaustive description of changepoint detection methods. Each subsequent chapter will also contain some element of review material, pertinent to the novel changepoint model being proposed.

Changepoint models can be classified into parametric and nonparametric models. In the parametric setting, the distribution of the data within each segment typically has the same functional form dependent on some unknown parameter that changes across segments. In the nonparametric setting, the data across segments follow unrelated, unknown probability distributions. Another classification of changepoint detection methods distinguishes online from offline detection methods. For online detection methods, inference of changepoints is made sequentially as new observations are made available. For offline detection methods, the inference is made by retrospective study based on all data available. Our work focuses on offline, parametric changepoint models.

Multiple frequentist approaches have been proposed for the changepoint detection problem,

which may be classified according to their estimation procedures. A popular approach is based upon dynamic programming with the use of constrained and penalised algorithms (for a review, see Maidstone et al. (2017)): a cost function is defined to measure the heterogeneity of data within segments, and the segmentation that achieves the smallest cost is obtained recursively. In particular, Killick, Fearnhead, and Eckley (2012) introduced the Pruned Exact Linear Time (PELT) algorithm that solves exactly the minimisation problem with an expected computational cost that can be linear in the number of observations. Another class of popular methods relies on Binary Segmentation (BS): originating from Scott and Knott (1974) and Sen and Srivastava (1975), the BS procedure aims to obtain a segmentation through an iterative divide-and-conquer approach. As pointed out in Killick, Fearnhead, and Eckley (2012), BS does not guarantee the most optimal changepoints are obtained. To gain statistical power upon the BS approach, Fryzlewicz (2014) proposed the Wild Binary Segmentation (WBS) procedure that computes the BS procedure multiple times on random subsequences of the observed data. Other frequentist approaches to the changepoint detection problem include sequential testing approaches, such as the MOSUM approach (Eichinger and Kirch, 2018). In subsequent chapters, some frequentist methods, which will be considered in simulation studies for comparison purposes, will be described in the context of the work being put forward.

The Bayesian approach, in contrast to frequentist approach, assumes changepoints are random: a prior distribution is specified for the changepoint parameters, namely the number and the locations of changepoints, and it is of interest to estimate the posterior distribution of changepoint parameters. Shiryaev (1963) gave one of the earliest approach to consider changepoints random with a prior distribution. Recent research on changepoint models rely on Monte Carlo Markov chain (MCMC) methods to sample changepoints. Stephens (1994) and Chib (1996) proposed an inference procedure based on MCMC methods for settings where the number of changepoints is known. The reversible jump MCMC algorithm introduced in Green (1995) can be used to infer both the number and the positions of changepoints, and it has been widely used in various settings (Bolton and Heard, 2018; Turcotte, 2014). Chib (1998) proposed an alternative approach based on MCMC methods for settings where the number of changepoints is not known *a priori*, which relies on the comparison of models via their marginal likelihoods. Fearnhead (2006) proposed an algorithm for direct simulation of the number and positions of changepoints based on recursions for product partition models (Barry and Hartigan, 1993). One of the advantages of Bayesian changepoint modelling is that the posterior distribution of changepoint parameters encapsulates information about uncertainty for both the number and the locations of changepoints.

This chapter presents a standard Bayesian parametric changepoint model that is considered both in Green (1995) and in Fearnhead (2006), and that will provide the basis for the discussion in the remainder of the thesis. Moreover, Monte Carlo methods for numerical integration and MCMC methods for sampling that are used in this thesis to perform inference for the proposed changepoint models are reviewed. Finally, different approaches to reporting the posterior distribution of

changepoints are discussed, and a novel loss function for changepoints is proposed that takes into account both the number and the locations of changepoints.

2.1. Standard Bayesian changepoint modelling

Let $x_{1:T} = (x_1, \ldots, x_T)$ denote time-indexed data. The changepoint model supposes that there are $k \geqslant 0$ changepoints that partition the time series $x_{1:T}$ into $k+1$ segments. The ordered locations of the changepoints, denoted by $\tau_{1:k} = (\tau_1, \ldots, \tau_k)$, belong to the set $\mathcal{T}_k$, where

$$\mathcal{T}_k = \left\{ (\tau_1, \ldots, \tau_k) \in \mathbb{N}^k;\ 1 \equiv \tau_0 < \tau_1 < \cdots < \tau_k < \tau_{k+1} \equiv T+1 \right\}. \tag{2.1}$$

The data $x_{\tau_{j-1}:\tau_j - 1} = (x_{\tau_{j-1}}, \ldots, x_{\tau_j - 1})$ in each segment j are assumed to be drawn from a distribution from the same parametric family $L(\cdot \mid \theta_j)$, with a segment specific parameter $\theta_j \in \Theta$. Conditional on changepoint parameters $(k, \tau_{1:k})$ and segment parameters $\theta_{1:(k+1)}$, the likelihood of the data is

$$\mathcal{L}\left(x_{1:T} \mid k, \tau_{1:k}, \theta_{1:(k+1)}\right) = \prod_{j=1}^{k+1} L\left(x_{\tau_{j-1}:\tau_j - 1} \mid \theta_j\right).$$

The domain of the model parameters $(k, \tau_{1:k}, \theta_{1:(k+1)})$ is $\bigcup_{k=0}^{\infty} \{k\} \times \mathcal{T}_k \times \Theta^{k+1}$. Given a prior distribution for the changepoint parameters and the segment parameters, denoted $\pi(k, \tau_{1:k}, \theta_{1:(k+1)})$, the posterior density of the model parameters is

$$\pi(k, \tau_{1:k}, \theta_{1:(k+1)} \mid x_{1:T}) = \frac{\mathcal{L}(x_{1:T} \mid k, \tau_{1:k}, \theta_{1:(k+1)}) \pi(k, \tau_{1:k}, \theta_{1:(k+1)})}{Z},$$

where Z is some normalising constant which is typically unknown. The prior density for the model parameters is usually chosen to have the form

$$\pi(k, \tau_{1:k}, \theta_{1:(k+1)}) = \pi(k, \tau_{1:k}) \prod_{j=1}^{k+1} \pi(\theta_j),$$

assuming segment parameters are independently drawn from the same prior density. A Bernoulli process is often assumed *a priori* for changepoint parameters (Fearnhead, 2006), such that

$$\pi(k, \tau_{1:k} \mid p) = p^k (1-p)^{T-k} \mathbf{1}_{\mathcal{T}_k}(\tau_{1:k}), \tag{2.2}$$

for some Bernoulli parameter $0 < p < 1$, that encodes prior belief on the expected number of changepoints.

The parameters of interest are the changepoint parameters $(k, \tau_{1:k})$, and therefore, when possible,

segment parameters are marginalised to obtain an expression of the posterior density of changepoint parameters

$$\pi(k, \tau_{1:k} \mid x_{1:T}) = \frac{\mathcal{L}(x_{1:T} \mid k, \tau_{1:k})\pi(k, \tau_{1:k})}{Z},$$

where

$$\mathcal{L}(x_{1:T} \mid k, \tau_{1:k}) = \prod_{j=1}^{k+1} \int_{\Theta} L(x_{\tau_{j-1}:\tau_j - 1} \mid \theta_j)\pi(\theta_j)\mathrm{d}\theta_j. \tag{2.3}$$

Examples of changepoint models where segment parameters may be marginalised include models for independent and identically distributed data within segments (Fearnhead, 2006; Denison et al., 2002), changing linear regressions (Punskaya et al., 2002; Carlin, Gelfand, and Smith, 1992), models for time-dependent data within segments, such as Markov models with time-varying transition matrices (Bolton and Heard, 2018), zero-mean and heteroscedastic processes with changing variance (Johnson, Elashoff, and Harkema, 2003), and changepoint models with segment parameters subject to seasonal effects (Turcotte, 2014).

In particular, consider the popular class of changepoint models where, within each segment, the data are assumed to be independent and identically distributed such that

$$x_t \sim f(\cdot \mid \theta_j), \quad \tau_{j-1} \leqslant t < \tau_j, \tag{2.4}$$

for some parametric density $f(\cdot \mid \theta_j)$ dependent on some segment parameter $\theta_j \sim \pi(\cdot)$. The integrals in (2.3) can be calculated analytically when π is chosen to be conjugate to f; and for non-conjugate cases, (2.3) may be calculated numerically for low-dimensional segment parameters.

Figure 2.1 displays simulated data $x_{1:T}$ from a changepoint model assuming the segment model

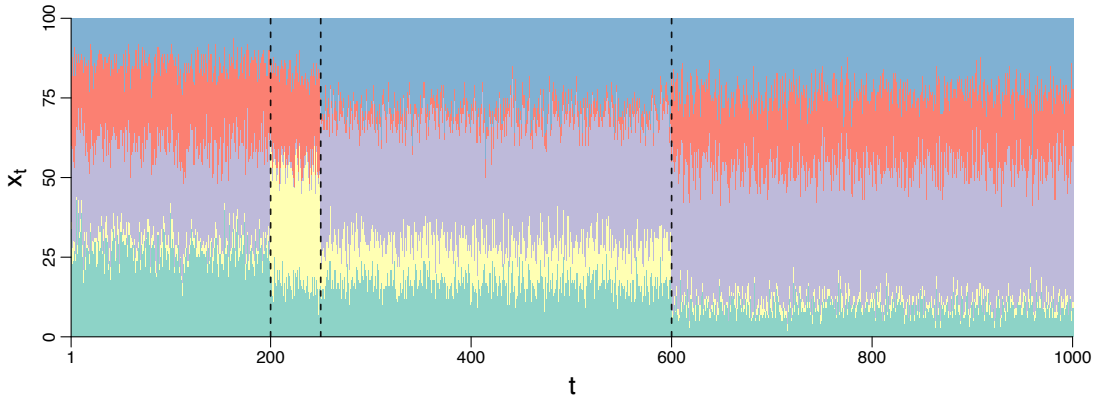


Figure 2.1.: *Multinomial data subject to three changepoints (τ_1, τ_2 and τ_3) indicated by vertical dashed lines.*

(2.4), where $x_t = (x_{t,1}, \ldots, x_{t,5}) \sim \text{Multinomial}_5(100, \theta_j)$ for each t in segment j. The probability parameter vectors θ_j are draws from a conjugate Dirichlet($\mathbf{1}^5$) prior, where $\mathbf{1}^M$ denotes an M-dimensional vector of ones. At each changepoint, it is apparent that the observed proportions of the categories changes.

2.2. Monte Carlo integration

For settings where the conditional likelihood of the data (2.3) is not tractable due to integrals that cannot be computed analytically, numerical integration methods can be used. In particular, Monte Carlo methods are popular numerical integration methods. Robert and Casella (2005) provide an in-depth presentation of Monte Carlo methods. This section provides a succinct review of some Monte Carlo methods, with a focus on sequential importance sampling (SIS) that is used in Chapter 5 to estimate the conditional likelihood of the data for a novel changepoint model.

2.2.1. Classical Monte Carlo integration

Let $z_{1:n} = (z_1, \ldots, z_n)$ be some random sequence of observations with joint density

$$\pi(z_{1:n}) = \frac{\upsilon(z_{1:n})}{Z}, \tag{2.5}$$

where Z is some unknown normalising constant. Suppose it is of interest to compute the integral

$$E_\pi[h(z_{1:n})] = \int h(z_{1:n})\pi(z_{1:n})dz_{1:n}, \tag{2.6}$$

where h is some finite function. Given a sample $z_{1:n}^{(1)}, \ldots, z_{1:n}^{(M)}$ from (2.5), the classical Monte Carlo approximation to (2.6) is the empirical average

$$\hat{E}_\pi^M[h(z_{1:n})] = \frac{1}{M}\sum_{i=1}^{M} h(z_{1:n}^{(i)}).$$

By the Strong Law of Large Numbers, $\hat{E}_\pi^M[h(z_{1:n})]$ converges almost surely to $E_\pi[h(z_{1:n})]$ as M tends to infinity.

2.2.2. Importance sampling

The method of importance sampling is relevant to settings where it is not directly possible to generate a sample from the density π, but it is possible to generate a sample from some density function q, whose support includes the support of π. Consider the alternative representation of

(2.6),

$$E_\pi[h(z_{1:n})] = E_q\left[h(z_{1:n})\frac{\pi(z_{1:n})}{q(z_{1:n})}\right] = \int h(z_{1:n})\frac{\pi(z_{1:n})}{q(z_{1:n})}q(z_{1:n})dz_{1:n}. \tag{2.7}$$

It follows that, given a sample $z_{1:n}^{(1)}, \ldots, z_{1:n}^{(M)}$ from the importance distribution with density q, the empirical average

$$\hat{E}_{IS}^{M}\left[h(z_{1:n})\right] = \hat{E}_{q}^{M}\left[h(z_{1:n})\frac{\pi(z_{1:n})}{q(z_{1:n})}\right] = \frac{1}{M}\sum_{i=1}^{M} h(z_{1:n}^{(i)})\frac{\pi(z_{1:n}^{(i)})}{q(z_{1:n}^{(i)})}$$

converges almost surely to (2.7) as M increases. In particular, note that if $h(z_{1:n}) = Z$ then

$$\hat{E}_{IS}^{M}\left[h(z_{1:n})\right] = \frac{1}{M}\sum_{i=1}^{M}\omega(z_{1:n}^{(i)}),$$

is an estimator for the normalising constant Z, where

$$\omega(z_{1:n}^{(i)}) = \frac{\upsilon(z_{1:n}^{(i)})}{q(z_{1:n}^{(i)})}$$

is the unnormalised importance weight for the ith sample.

2.2.3. Sequential importance sampling (SIS)

Sequential importance sampling refers to a special case of importance sampling for the target distribution (2.5) such that

$$\upsilon(z_{1:n}) = \prod_{t=1}^{n}\upsilon(z_t \mid z_{t-1}), \tag{2.8}$$

where $\upsilon(z_t \mid z_{t-1})$ denotes an unnormalised density function of z_t conditional on z_{t-1}, and z_0 is some known constant. To exploit the sequential structure of υ, the importance proposal distribution is chosen such that

$$q(z_{1:n}) = \prod_{t=1}^{n} q(z_t \mid z_{t-1}), \tag{2.9}$$

where $q(z_t \mid z_{t-1})$ denotes the importance proposal density function of z_t conditional on z_{t-1}. Let

$$\omega(z_t \mid z_{t-1}) = \frac{\upsilon(z_t \mid z_{t-1})}{q(z_t \mid z_{t-1})}$$

Algorithm 2.1: Sequential importance sampling

Input: Unnormalised density (2.8), importance proposal distribution (2.9)

Result: M samples $\left(z_{1:n}^{(i)}\right)_{i=1,\ldots,M}$ from target density (2.5) and importance weights $\left(\omega(z_{1:n}^{(i)})\right)_{i=1,\ldots,M}$

Initialisation: for all $i = 1, \ldots, M$, $z_0^{(i)} = z_0$

for $t = 1, \ldots, n$ **do**

  for $i = 1, \ldots, M$ **do**

    sample $z_t^{(i)} \sim q(\cdot \mid z_{t-1}^{(i)})$

    set

$$\omega\left(z_{1:t}^{(i)}\right) = \omega\left(z_{1:(t-1)}^{(i)}\right) \omega\left(z_t^{(i)} \mid z_{t-1}^{(i)}\right) \tag{2.10}$$

  end

end

for all z_t and z_{t-1}, so that $\omega(z_{1:n}) = \prod_t \omega(z_t \mid z_{t-1})$. Algorithm 2.1 gives the sequential importance sampling to obtain M samples from (2.5). Note that sequential importance sampling is adequate for recursive sampling and estimation as more data are observed: for example, when new data z_{n+1} become available, the importance weights for $z_{1:(n+1)}$ can be efficiently updated using the importance weights computed for the sequence $z_{1:n}$ according to (2.10).

2.3. Inference via Markov chain Monte Carlo (MCMC) methods

For the Bayesian changepoint model discussed in Section 2.1, the parameters of interest, namely the changepoint parameters and possibly the segment parameters, have a dimension that is not fixed and a posterior distribution that is known only up to a multiplication constant. Recent Bayesian changepoint detection methods typically rely on inference via simulations. In particular, Markov chain Monte Carlo (MCMC) methods have been widely used to sample from the posterior distribution of changepoints. The reversible jump Markov Chain Monte Carlo (RJMCMC) algorithm, which is a Metropolis-Hastings algorithm used to sample from target distributions of varying dimension, has been a popular algorithm to sample changepoints since being introduced in Green (1995). The sampling algorithm proposed in Green (1995) can be used to infer both the number and the positions of changepoints, in contrast with methods proposed earlier that require the number of changepoints to be known *a priori* (Carlin, Gelfand, and Smith, 1992).

This section reviews RJMCMC algorithms and other popular MCMC methods that are considered in this thesis to sample from the proposed changepoint models.

Algorithm 2.2: Metropolis-Hastings algorithm

Input: unnormalised target density υ, proposal density $q(\cdot \mid x)$, Markov chain length M
Result: Markov chain $x^{(1)}, \ldots, x^{(M)}$ with stationary distribution π

Initialisation set starting value $x^{(0)}$
for $t = 1, \ldots, M$ **do**
  propose $y \sim q(\cdot \mid x^{(t-1)})$,
  compute the acceptance probability of the proposal

$$\alpha = 1 \wedge \frac{\upsilon(y) q(x^{(t-1)} \mid y)}{\upsilon(x^{(t-1)}) q(y \mid x^{(t-1)})},$$

  sample $u \sim \text{Unif}(0, 1)$,
  if $u < \alpha$ **then** $x^{(t)} = y$ **else** $x^{(t)} = x^{(t-1)}$.
end

Algorithm 2.3: Gibbs sampling

Input: full conditional distributions $\pi(\cdot \mid x_{-i})$ for all $i = 1, \ldots, d$, Markov chain length M
Result: Markov chain $x^{(1)}, \ldots, x^{(M)}$ with stationary distribution π

Initialisation set starting value $x^{(0)} = (x_1^{(0)}, \ldots, x_d^{(0)})$
for $t = 1, \ldots, M$ **do**
  for $i = 1, \ldots, d$ **do**
    set $x_i^{(t)} \sim \pi(\cdot \mid x_{-i}^{(t)})$, where $x_{-i}^{(t)} = (x_1^{(t)}, \ldots, x_{i-1}^{(t)}, x_{i+1}^{(t-1)}, \ldots, x_d^{(t-1)})$
  end
end

2.3.1. Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm is a Markov chain Monte Carlo method for obtaining a sample $x^{(1)}, \ldots, x^{(M)}$ from a probability distribution with density π when direct simulation is not straightforward. The Metropolis-Hastings algorithm assumes that a function υ proportional to the density π is available and it relies on choosing a conditional distribution with density $q(\cdot \mid x)$ to propose a candidate for the next element of the chain given the latest element x. Robert and Casella (2005) show that the Metropolis-Hastings algorithm, which is given in Algorithm 2.2, may be used to produce a Markov chain $x^{(1)}, \ldots, x^{(M)}$ with stationary distribution π. Details of the methodology may be found in Robert and Casella (2005).

2.3.2. Gibbs sampling

Suppose it is of interest to obtain a sample of some multivariate random variable $x = (x_1, \ldots, x_d)$, for some $d > 1$, from a probability distribution with density π. Gibbs sampling assumes that, for all $i = 1, \ldots, d$, it is possible to sample from the full conditional distribution of x_i, denoted $\pi(\cdot \mid x_{-i})$

where $x_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d)$. The Gibbs sampler given in Algorithm 2.3, which is a special case the Metropolis–Hastings algorithm where all moves are accepted with probability 1, produces a Markov chain $x^{(1)}, \ldots, x^{(M)}$, where $x^{(t)} = (x_1^{(t)}, \ldots, x_d^{(t)})$ for all t, with stationary distribution π.

2.3.3. Reversible jump MCMC

Consider a dimension varying model defined as a collection of models

$$M_k = \{\pi(\cdot \mid \phi_k), \quad \phi_k \in \Phi_k\}, \quad k = 0, \ldots, K,$$

where model M_k has parameter space Φ_k with dimension d_k. The RJMCMC algorithm is a Markov chain Monte Carlo method used to sample from the joint distribution of the model index and the associated model parameters with joint density $\pi(k, \phi_k)$. The RJMCMC requires that a function υ proportional to the density π is available.

Moves between models of different dimensions, which can be reversed, are made possible by dimension matching proposals between pairs of models (Green, 1995). For a move from model M_{k_1} to model M_{k_2}, which is proposed with probability $r(k_1, k_2)$, the parameter spaces Φ_{k_1} and Φ_{k_2} are augmented by sets U_1 and U_2, respectively, such that $\dim(\Phi_{k_1}) + \dim(U_1) = \dim(\Phi_{k_2}) + \dim(U_2)$. Then, (k_2, ϕ_{k_2}) is proposed from (k_1, ϕ_{k_1}) as follows: sample $u_1 \in U_1$ and $u_2 \in U_2$ from the densities $g_{k_1,k_2}(\cdot)$ and $g_{k_2,k_1}(\cdot)$, respectively, and set $(\phi_{k_2}, u_2) = \mathrm{T}_{k_1,k_2}(\phi_{k_1}, u_1)$, where T_{k_1,k_2} is

Algorithm 2.4: Reversible Jump MCMC algorithm

Input: target density υ, Markov chain length M, model jump probabilities and proposal densities

Result: Markov chain $(k^{(1)}, \phi^{(1)}_{k^{(1)}}), \ldots, (k^{(M)}, \phi^{(M)}_{k^{(M)}})$ with stationary distribution $\pi(k, \phi_k)$

Initialisation set starting value $(k^{(1)}, \phi^{(1)}_{k^{(1)}})$

for $t = 2, \ldots, M$ **do**

- propose a move from model $M_{k^{(t-1)}}$ to model $M_{k'}$
- sample $u_{k'} \sim g_{k^{(t-1)},k'}(\cdot)$ and $u_{k^{(t-1)}} \sim g_{k',k^{(t-1)}}(\cdot)$
- set $(\phi_{k'}, u_{k'}) = \mathrm{T}_{k^{(t-1)},k'}(\phi^{(t-1)}_{k^{(t-1)}}, u_{k^{(t-1)}})$
- compute the acceptance probability of the proposal

$$\alpha = 1 \wedge \frac{\upsilon(k', \phi_{k'})}{\upsilon(k^{(t-1)}, \phi^{(t-1)}_{k^{(t-1)}})} \frac{r(k', k^{(t-1)})}{r(k^{(t-1)}, k')} \frac{g_{k',k^{(t-1)}}(u_{k^{(t-1)}})}{g_{k^{(t-1)},k'}(u_{k'})} \left| \frac{\partial \mathrm{T}_{k^{(t-1)},k'}(\phi^{(t-1)}_{k^{(t-1)}}, u_{k^{(t-1)}})}{\partial(\phi^{(t-1)}_{k^{(t-1)}}, u_{k^{(t-1)}})} \right|$$

- sample $u \sim \mathrm{Unif}(0, 1)$,
- **if** $u < \alpha$ **then** $(k^{(t)}, \phi^{(t)}_{k^{(t)}}) = (k', \phi_{k'})$ **else** $(k^{(t)}, \phi^{(t)}_{k^{(t)}}) = (k^{(t-1)}, \phi^{(t-1)}_{k^{(t-1)}})$.

end

a diffeomorphism.

Green (1995) shows that the RJMCMC algorithm, which is given in Algorithm 2.4, may be used to produce a Markov chain $(k^{(1)}, \phi^{(1)}_{k^{(1)}}), \ldots, (k^{(M)}, \phi^{(M)}_{k^{(M)}})$ with stationary distribution $\pi(k, \phi_k)$.

Green (1995) demonstrates the RJMCMC is suitable to sample from the posterior distribution of the number and the locations of changepoints. For the changepoint model discussed in Section 2.1, the model M_k corresponds to a changepoint model with k changepoints, and the parameter space $\Phi_k = \mathcal{T}_k \times \Theta^{k+1}$, or $\Phi_k = \mathcal{T}_k$ if segment parameters can be marginalised. The RJMCMC algorithm proposed in Green (1995), which assumes continuous time changepoints, is adapted to discrete time changepoints in Denison et al. (2002) for the setting where segment parameters are not marginalised. This thesis focuses on discrete time changepoints, and therefore we describe below the RJMCMC proposed in Denison et al. (2002). At each iteration of the algorithm, given a model with changepoint parameters $(k, \tau_{1:k})$ and segment parameters $\theta_{1:(k+1)}$, one of the following four moves is proposed:

- *Death move.* The death move proposes to remove an existing changepoint. An index j of one element of $\tau_{1:k}$ is uniformly chosen and changepoint parameters $(k', \tau'_{1:k'})$ are proposed, where $k' = k - 1$ and $\tau'_{1:k'} = (\tau_{1:(j-1)}, \tau_{(j+1):k})$. Segment parameters $\theta'_{1:(k'+1)} = (\theta_{1:(j-1)}), \theta', \theta_{(j+2):(k+1)})$ are proposed, where θ' is sampled conditional on $(k', \tau'_{1:k'})$.
- *Birth move*. The birth move proposes to add a new changepoint. Draw τ_h uniformly from $\{2, \ldots, T+1\} \setminus \tau_{1:k}$ and propose the changepoint parameters $(k', \tau'_{1:k'})$, where $k' = k + 1$ and $\tau'_{1:k'}$ is obtained by inserting τ_h into $\tau_{1:k}$. Let $j = \max\{i; \tau_i < \tau_h\}$ and propose the segment parameters $\theta'_{1:(k'+1)} = (\theta_{1:j}, \theta', \theta'', \theta_{(j+2):(k+1)})$, where θ' and θ'' are sampled conditional on $(k', \tau'_{1:k'})$.
- *Shift move.* The shift move proposes to modify the position of one randomly chosen changepoint. The index j is uniformly sampled from $\{1, \ldots, k\}$, the changepoint position τ_h is uniformly sampled from $\{\tau_{j-1} + 1, \ldots, \tau_{j+1} - 1\}$ and changepoint parameters $(k, \tau'_{1:k'})$ are proposed, where $\tau'_{1:k'} = (\tau_{1:(j-1)}, \tau_h, \tau_{(j+1):k})$. Segment parameters $\theta'_{1:(k'+1)} = (\theta_{1:(j-1)}, \theta', \theta'', \theta_{(j+2):(k+1)})$ are proposed, where θ' and θ'' are sampled conditional on $(k', \tau'_{1:k'})$.
- *Update segment parameters.* One of the segment parameters is sampled conditional on changepoint parameters.

If segment parameters can be marginalised, it is sufficient to consider the *birth*, *death* and *shift* moves to sample changepoint parameters $(k, \tau_{1:k})$, and it is not needed to propose segment parameters.

There have been many implementations of the RJMCMC proposed in the context of Bayesian changepoint analysis. For example, Bolton and Heard (2018) propose a reversible jump Markov chain Monte Carlo for changepoint detection that incorporates regime-switching, allowing regimes to be inferred from malware instruction traces. Another example is Lebre et al. (2010) that propose a RJMCMC algorithm to sample changepoints impacting a graphical model motivated by biological

applications.

2.3.4. Auxiliary variable methods for MCMC

For some models exhibiting multimodality or strong dependence between variables, standard MCMC algorithms may suffer from poor mixing and slow convergence to the target distribution (Besag and Green, 1993). To improve sampling performance, one approach is to use auxiliary variables, that are introduced to facilitate the exploration of the parameter space. Chapter 3 will propose the use of auxiliary variables in a RJMCMC algorithm to sample dependent changepoints across multiple time series.

The general principle of auxiliary variable methods for MCMC is the following. Suppose it is of interest to generate realisations from a distribution with density $\pi(x)$ that exhibits some characteristics making the exploration of the state space difficult for standard MCMC methods. The variable of interest $x \in \mathcal{X}$ is augmented by some variable $u \in \mathcal{U}$. A conditional distribution $\pi(u \mid x)$ is defined, thereby specifying the joint distribution $\pi(x, u) = \pi(u \mid x)\pi(x)$. The auxiliary variable u and $\pi(u \mid x)$ are chosen such that it is possible to build an MCMC algorithm to sample from $\pi(x, u)$ on $\mathcal{X} \times \mathcal{U}$ which has good sampling performance. Given a sample from $\pi(x, u)$, a sample from the marginal distribution with density $\pi(x)$ is readily available.

The MCMC algorithm with auxiliary variables that is proposed in Chapter 3 in the context of Bayesian changepoint analysis is related to an MCMC method for binary Markov random fields proposed in Higdon (1998), which is discussed below.

Let $G = (V, E)$ be some undirected graph where V is the node set and $E \subset V \times V$ is the edge set. Suppose we observe binary random variables indexed on the graph $x_V = (x_i)_{i \in V}$ that follow the binary Markov random field with density

$$\pi(x_V) \propto \exp\left\{ \sum_{i \in V} \alpha_i(x_i) + \beta \sum_{(i,i') \in E} \mathbf{1}(x_i, x_{i'}) \right\}, \quad x_V \in \{0, 1\}^{|V|},$$

where

$$\mathbf{1}(x_i, x_{i'}) = \begin{cases} 1 & \text{if } x_i = x_{i'} \\ 0 & \text{otherwise,} \end{cases}$$

for some external field functions α_i and some interaction parameter $\beta > 0$. The larger β, the more likely for two variables x_i and x_i' to be equal if $(i, i') \in E$. If G is a lattice graph and α_i is the zero function for all i, then x_V follow the Ising model. Gibbs sampling from $\pi(x_V)$ can be considered since, for all $i \in V$, it is straightforward to sample from the full conditional distribution

of $x_i \in \{0, 1\}$,

$$\pi(x_i \mid \{x_{i'} \mid i' \in V \setminus \{i\}\}) = \frac{q_i^{x_i}}{1 + q_i}, \tag{2.11}$$

where

$$q_i = \exp\left\{\frac{\alpha_i(1)}{\alpha_i(0)} + \beta \sum_{i':(i,i')\in E} x_{i'}\right\}.$$

However, if β is large, due to the strong dependence between connected variables in G, Gibbs sampling may suffer from poor mixing and slow convergence as illustrated in Higdon (1998). This motivates a sampler that can simultaneously update multiple variables that are connected in G.

To improve the performance of the sampler, Higdon (1998) proposes to augment the parameter space with auxiliary variables $u_E = (u_{i,i'})_{(i,i')\in E}$ such that, independently for all $(i, i') \in E$, the conditional density of $u_{i,i'} \in \{0, 1\}$ is

$$\pi(u_{i,i'} \mid x_V) = (1 - \exp\{-\beta\delta\mathbf{1}(x_i, x_{i'})\})^{u_{i,i'}} \exp\{-\beta\delta\mathbf{1}(x_i, x_{i'})\}^{1-u_{i,i'}}, \tag{2.12}$$

for some decoupling parameter $0 \leqslant \delta \leqslant 1$ whose role will be discussed below. Consequently, for $x_V \in \{0, 1\}^{|V|}$,

$$\pi(x_V \mid u_E) \propto \exp\left\{\sum_{i\in V} \alpha_i(x_i) + \beta(1-\delta) \sum_{(i,i')\in E} \mathbf{1}(x_i, x_{i'})\right\} \prod_{(i,i')\in E} (\exp\{\beta\delta\mathbf{1}(x_i, x_{i'})\} - 1)^{u_{i,i'}}, \tag{2.13}$$

If $x_i \neq x_{i'}$ then $\exp\{\beta\delta\mathbf{1}(x_i, x_{i'})\} = 1$. Consequently, according to (2.13), if $u_{i,i'} = 1$ then $x_i = x_{i'}$. Let $H = (V, E')$ be the graph such that $(i, i') \in E'$ if and only if $u_{i,i'} = 1$. It follows that, with $\mathcal{C}$ denoting the set of connected components in H, for all connected component $\gamma \in \mathcal{C}$, $x_i = x_{i'}$ for all $i, i' \in \gamma$.

To sample from the joint distribution of x_V and u_E, consider the Gibbs sampler such that, at each iteration of the algorithm, the auxiliary variables u_E are updated conditional on x_V according to (2.12), and for each connected component $\gamma \in \mathcal{C}$ the variables x_γ are updated from their full conditional distribution (2.13) so that either $x_i = 0$ for all $i \in \gamma$ or $x_i = 1$ for all $i \in \gamma$. If $\delta = 1$, the algorithm corresponds to the Swendsen algorithm (Swendsen and Wang, 1987): for each cluster $\gamma \in \mathcal{C}$, the full conditional distribution of x_γ is such that

$$\pi\left(x_\gamma \mid u_E\right) = \left(\frac{q_\gamma}{1+q_\gamma}\right)^{\prod_{i\in\gamma} x_i} \left(\frac{1}{1+q_\gamma}\right)^{1-\prod_{i\in\gamma} x_i} \prod_{i,i'\in\gamma} \mathbf{1}(x_i, x_{i'}),$$

where $q_\gamma = \exp\{\sum_{i\in\gamma} \alpha_i(1)/\alpha_i(0)\}$, so that, conditional on u_E, the variables x_γ are updated independently of $x_{V\setminus\gamma}$. The auxiliary variables decouple the external field term from the interaction term, thereby providing a means of moving through the state space more efficiently. Higdon (1998) demonstrates that there are settings where fixing some $0 < \delta < 1$ can lead to further improvement of the performance of the sampler. For example, for settings where the external field functions are uneven, the Swendsen algorithm might produce clusters of variables that are unlikely to be updated becausethey are too large. The smaller δ, the smaller the expected size of clusters induced by auxiliary variables. Therefore, some $0 < \delta < 1$ can be fixed to control the expected size of clusters induced by auxiliary variables to improve the performance of the sampler. Finally, note that if $\delta = 0$, then $u_{i,i'} = 0$ for all $(i, i') \in E$, and therefore the sampler corresponds to the Gibbs sampler with single variable updating according to (2.11).

2.4. Loss function for changepoint parameter estimation

Bayesian inference for changepoints quantifies uncertainty about the number and the positions of changepoints. To give an account of the posterior distribution of changepoint parameters, following Green (1995), one may consider the posterior marginal distribution of the number of changepoints k, and the posterior distribution of the changepoint positions $\tau_{1:k}$ conditional on k.

However, in some applications such as cyber-security, it will also be necessary to report a point estimate $(\hat{k}, \hat{\tau}_{1:\hat{k}})$ for changepoint parameters $(k, \tau_{1:k})$. Bayesian decision theory (Berger, 1985) provides a framework for defining optimal estimates via the principle of minimum expected loss. In the context of changepoint estimation, a loss function ℓ is defined to quantify the quality of estimated changepoints, such that $\ell(\hat{k}, \hat{\tau}_{1:\hat{k}} \mid k, \tau_{1:k})$ is the cost associated with the estimate $(\hat{k}, \hat{\tau}_{1:\hat{k}})$ of $(k, \tau_{1:k})$. Then, an optimal Bayes estimate for changepoints is defined to minimise the expected loss

$$(\hat{k}, \hat{\tau}_{1:\hat{k}}) = \underset{(\hat{k}, \hat{\tau}_{1:\hat{k}})}{\operatorname{argmin}} \mathbb{E}_{\pi(k,\tau_{1:k})}[\ell(\hat{k}, \hat{\tau}_{1:\hat{k}} \mid k, \tau_{1:k})]$$

with respect to the posterior distribution of changepoints $\pi(k, \tau_{1:k})$. When inference via simulations is required, given a sample $(k^{(1)}, \tau^{(1)}_{1:k^{(1)}}), \ldots, (k^{(M)}, \tau^{(M)}_{1:k^{(M)}})$ from the posterior distribution of changepoints, the optimal Bayes estimate $(\hat{k}, \hat{\tau}_{1:\hat{k}})$ can be approximately identified by numerically finding within the sample the changepoint parameters that minimise the estimated posterior expected loss,

$$(\hat{k}, \hat{\tau}_{1:\hat{k}}) = \underset{(\hat{k}, \hat{\tau}_{1:\hat{k}}) \in S}{\operatorname{argmin}} \sum_{i=1}^{M} \ell(\hat{k}, \hat{\tau}_{1:\hat{k}} \mid k^{(i)}, \tau^{(i)}_{1:k^{(i)}}), \tag{2.14}$$

where $S = \{(k^{(i)}, \tau^{(i)}_{1:k^{(i)}}), i = 1, \ldots, M\}$.

To the best of our knowledge, there are few loss functions suitable for changepoint parameters and none seem to be adapted to cyber-security applications of interest. For example, Yau and Holmes (2013) and Bardwell and Fearnhead (2017) consider loss functions for settings where the task is to segment and classify data according to an unobserved discrete state sequence, assuming a finite number of states, for example 'normal' and 'abnormal'. For the cyber-security applications of interest, the focus is on the estimation of the number of changepoints and their locations, not the classification of the data into discrete states.

A common approach to defining changepoint estimates, which has been applied in the context of cyber-security (Bolton and Heard, 2018), is to use *maximum a posteriori* (MAP) changepoint estimates $(\hat{k}, \hat{\tau}_{1:\hat{k}})$, defined such that $\hat{k}$ is the MAP number of changepoints and $\hat{\tau}_{1:\hat{k}}$ are the MAP changepoint locations of dimension $\hat{k}$. It is unclear how to define a loss function ℓ justifying the MAP changepoint estimates according to the Bayesian decision theory, due to the two-step formulation that separates the estimation of the number of changepoints from the estimation of their locations.

This section proposes a novel loss function for changepoint parameters that is suitable for taking into account both the number and the positions of changepoints. To address this challenge, we use matchings in graphs, as defined in Definition 1 and Definition 2, to define a loss function ℓ for changepoint estimates in Definition 3.

Definition 1. *(Maximum matching in a graph (Bondy and Murty, 1976)).* Let $G = (V, E)$ be a graph where V is a vertex set and $E \subseteq V \times V$ is an edge set. A matching M in G is a subset of E such that no two edges in M share a common vertex. A maximum matching in G is a matching that is not a subset of a larger matching in G.

Definition 2. *(Minimum weight maximum matching in a graph (Bondy and Murty, 1976)).* Let $G = (V, E)$ be a graph with weights $w_{i,j} \geqslant 0$ for all $(i, j) \in E$. A minimum weight maximum matching in G is a maximum matching in G for which the sum of weights of the edges is minimised.

When $B = (V, E)$ is a bipartite graph, the Kuhn–Munkres algorithm, also known as the Hungarian algorithm, finds a minimum weight maximum matching in B (Bondy and Murty, 1976). Consider the complete bipartite graph B with independent vertex sets $V = \{0, \ldots, k\}$, $\hat{V} = \{0, \ldots, \hat{k}\}$ and weights (2.15). A minimum weight maximum matching M in B gives a matching of the elements of $\hat{\tau}_{1:\hat{k}}$ and $\tau_{1:k}$ that minimises the sum of distances (2.15) between matched changepoints.

Definition 3. *(Loss function ℓ for changepoint estimates).* Let $\gamma \geqslant 0$. For all $k, \hat{k} \geqslant 0$, $\tau_{1:k} = (\tau_1, \ldots, \tau_k) \in \mathcal{T}_k$ and $\hat{\tau}_{1:\hat{k}} = (\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{k}}) \in \mathcal{T}_{\hat{k}}$, let B be the weighted complete bipartite graph

with vertex sets $V = \{0, \ldots, k\}$ and $\hat{V} = \{0, \ldots, \hat{k}\}$, and weights

$$w_{i,j} = \min\{\gamma, |\tau_i - \hat{\tau}_j|\} \tag{2.15}$$

for all $i \in V$ and $j \in \hat{V}$. Given a minimum weight maximum matching M in B, for all $i \in V$ and $j \in \hat{V}$, let $m_{i,j} = 1$ if i and j are matched, that is $(i, j) \in M$, and $m_{i,j} = 0$ otherwise. Then, define the loss to be

$$\ell\left[(\hat{k}, \hat{\tau}_{1:\hat{k}}), (k, \tau_{1:k})\right] = \gamma|\hat{k} - k| + \frac{1}{2}\sum_{i \in V}\sum_{j \in V'} m_{i,j} w_{i,j}. \tag{2.16}$$

Given M, according to the loss function (2.16), the cost associated with the estimate $(\hat{k}, \hat{\tau}_{1:\hat{k}})$ of $(k, \tau_{1:k})$ is obtained by adding the cost γ for each unmatched changepoint and the total distance between matched changepoints. Note that according to (2.15), the cost of matching two changepoint positions, that are separated by more than γ time units, is equal to the cost of an unmatched changepoint, namely γ. Therefore, the loss function ℓ takes into account both the number and the positions of changepoints, and the cost γ is chosen to be the maximum acceptable distance between a changepoint position and its estimated position.

2.5. Changepoint detection in network authentication data

This section gives an example of Bayesian changepoint analysis of real computer network data. Consider the LANL authentication data discussed in Section 1.1.1. The top panel of Figure 2.2 displays the hourly counts of network logons per source computer $x_{i,1}, \ldots, x_{i,T}$ defined in (1.1) for user $i =$ 'U342@DOM1'. Assume the changepoint model defined in (2.4) where $f(\cdot \mid \theta_j)$ denotes the density of the multinomial distribution with unknown probability parameter vectors θ_j with an uninformative, conjugate prior Dirichlet($\mathbf{1}^M$). Each changepoint τ_j would correspond to a temporal change in the distribution of counts of logons initiated by the user $i \in V$ across M host computers in the network. A sample of size 50,000 is obtained from the posterior distribution of changepoints via the RJMCMC algorithm discussed in Section 2.3.3 with a burn-in of size 10,000. Figure 2.3 displays estimated the posterior distribution of the number of changepoints k obtained from the sample. The bottom panel of Figure 2.2 displays the posterior distribution of the changepoint positions conditional on $k = 27$. The optimal Bayes estimate for changepoint parameters with respect to the loss function ℓ with $\gamma = 48$ were obtained according to (2.14). The optimal Bayes estimate is identical to the MAP changepoints, which are displayed in the top panel of Figure 2.2. Note that this example will be considered in Chapter 3 to motivate a novel changepoint model.

2.6. Conclusion

This chapter presented a standard Bayesian changepoint model that provides the basis for the novel changepoint models motivated by cyber-security applications that are introduced in Chapters 3, 4 and 5. Monte Carlo methods for numerical integration and MCMC methods for sampling that are used in this thesis to perform inference for the proposed changepoint models were reviewed. Moreover, different approaches to summarise the posterior distribution of changepoints were dicussed, and a novel loss function for changepoints was proposed, which takes into account both the number and the positions of changepoints. The next chapter considers a setting with multiple time series subject to changepoints, where it is of interest to encode prior beliefs on the dependence structure of changepoints across time series.

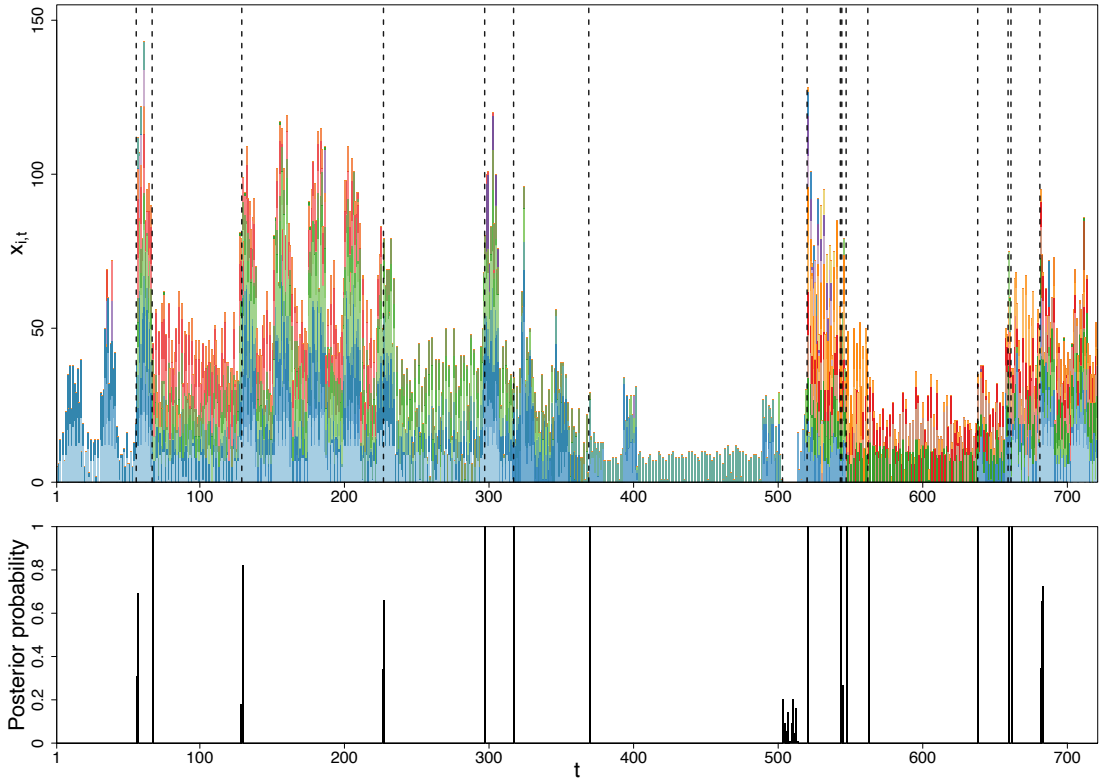


Figure 2.2.: *Top panel: hourly counts of network logons per source computer defined in* (1.1) *for a user in the LANL network. Dashed vertical lines indicate MAP changepoints. Bottom panel: Posterior distribution of the changepoint positions of dimension* $k = 27$.

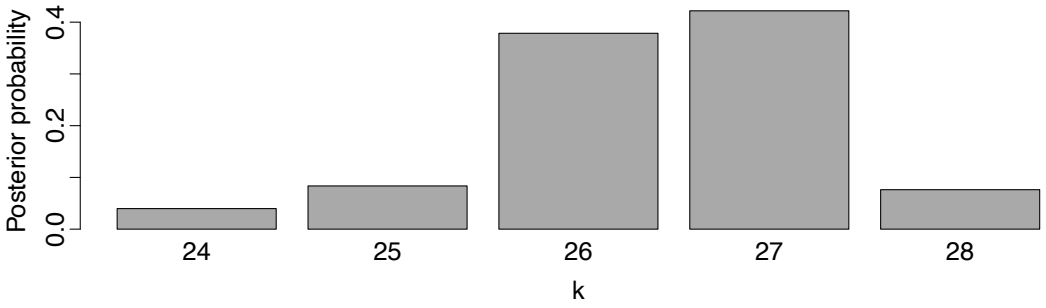


Figure 2.3.: *Posterior distribution of the number of changepoints* k *for the authentication data displayed in Figure 2.2.*

3

Graph of time series

Consider $N > 1$ time series of random observations

$$\{x_{i,t} \mid 1 \leqslant i \leqslant N, t \geqslant 0\} \tag{3.1}$$

that are subject to changepoints. This chapter will suppose the existence of an underlying graph G on N nodes corresponding to each of the time series, such that changepoints are believed to occur simultaneously or closely together in time for time series connected by edges in G.

This setting is motivated by cyber-security applications. To identify the presence of a network intrusion, it is informative to monitor for changes in the authentication activity of each user in the network as discussed in Section 1.1.1. However, cyber data often exhibit much variability and apparent changes are not guaranteed to correspond to an attack. As a result, to limit the number of false alerts and yet not overlook weak signals from genuine, small footprints of attacks, it is key to incorporate expert knowledge in the change detection procedure. A commonly held belief of security experts is that attacks are *a priori* likely to be identified through quasi-simultaneous changes in the behaviour of users that are linked by network connectivity (Sexton, Storlie, and Neil, 2015). Hence, it is of interest to encode a changepoint prior by means of a graph G representing the network of users, such that pairs of connected users in G are *a priori* more likely to be affected by quasi-simultaneous behavioural changes.

Limited attention has previously been given to encoding prior beliefs on graph-based dependence structure of discrete-time changepoints across multiple time series. Existing changepoint model for multiple time series, which admit changepoints may simultaneously affect a subset of the time series, typically assume *a priori* changepoint locations are exchangeable across time series (Jeng, Cai, and Li, 2012; Bardwell and Fearnhead, 2017; Bolton and Heard, 2018; Wang and Samworth, 2018; Bardwell et al., 2019; Grundy, Killick, and Mihaylov, 2020). Moreover, with the exception of

Fisch, Eckley, and Fearnhead (2022), dependent changepoints across time series are often assumed to perfectly align, which is a limiting assumption in cyber-security monitoring where attacks may span a substantial period of time.

More generally, graphical models provide a useful framework for characterising joint distributions for random variables: the nodes of the graph identify the random variables and the edges characterise dependencies among these variables (Lauritzen, 1996). In particular, graphical models have been employed to encode prior beliefs, for example, in the context of Bayesian variable selection for regression models. Li and Zhang (2010) assume that covariates lie on an undirected graph and formulates an Ising model prior on the covariate space to incorporate structural information.

This chapter proposes a novel informative, graphical model-based prior for changepoints that encodes beliefs on the dependence structure of changepoints across time series (3.1). For practical purposes, changepoints are represented in discrete time by a binary matrix $\boldsymbol{S} = (S_{i,t})$, such that $S_{i,t}$ indicates whether the time point t is a changepoint for the time series with index i. Then, extending the standard memoryless prior for changepoints (Fearnhead, 2006), independent and identical Markov random fields (Lauritzen, 1996) with respect to G are assumed *a priori* for the columns of $\boldsymbol{S}$. As a result, the model assumes that clusters of time series (according to G) are likely to be simultaneously affected by changepoints. Conditional on changepoints, the time series data are assumed to be independent of G and to follow a standard parametric changepoint model (Fearnhead, 2006). A key consequence of the graphical model is that stronger evidence from data is required to infer scattered synchronous changepoints than synchronous changepoints clustered according to G. Furthermore, a more general model is proposed that admits related changepoints not occurring at exactly the same time; the extended model supposes that changepoints may cluster according to G within some finite time windows of possibly unknown lengths, which are specific to each series.

As discussed in Chapter 2, a common approach to sampling changepoints for a single time series is that of Green (1995), using a reversible jump MCMC algorithm to explore the state space of changepoints: at each iteration of the algorithm, a new changepoint is proposed, or else an existing changepoint is either deleted or shifted to a new position. Specifying a joint model for dependent changepoints across multiple time series introduces additional computational challenges that are not present when changepoints are inferred for each time series independently. A simulation study will demonstrate that it can be impractical to simply propose updates to the changepoints of a randomly chosen time series via one of the moves of Green (1995). To efficiently explore the state space of dependent changepoints, it is necessary to consider joint proposals for changepoints across multiple time series.

This chapter proposes an MCMC algorithm making use of auxiliary variables (Besag and Green, 1993) to sample from the posterior distribution. As discussed in Chapter 2, Swendsen and Wang (1987) and Higdon (1998) provide notable examples of use of auxiliary variables in MCMC

schemes that improve mixing and convergence for undirected graphical models. In brief, our sampling strategy is the following. The changepoint parameter space is augmented with auxiliary variables that induce clusters of time series according to the dependence graph G. Then, the MCMC algorithm of Green (1995) is extended to sample from the augmented parameter space, such that, at each iteration of the algorithm, a new cluster of changepoints may be proposed or an existing cluster of changepoints may be deleted or shifted.

The practical benefits of the proposed graphical model are demonstrated via a changepoint analysis of real computer network authentication data from LANL presented in Chapter 2, where a subset of the data relating to a 'red team' exercise provide a proxy for intruder behaviour (Kent, 2015). The challenge consists of monitoring for temporal changes in the authentication activity of network users to detect the presence of red team actors. The proposed changepoint prior is used to encode beliefs that signals for network intrusions are *a priori* likely to occur at nearby times for users historically linked by previous network connectivity. We show that, as a consequence, the proposed model can detect weak signals for red team activity in the network, whilst limiting the number of false alerts, in contrast with a standard model assuming independence of behavioural changes across users.

The remainder of the chapter is organised as follows. Section 3.1 presents the motivating application in cyber-security. Section 3.2 introduces a novel, graph-based informative prior for changepoints. Section 3.3 proposes an auxiliary variable MCMC sampling strategy. Section 3.4 presents a simulation study to demonstrate the advantages of the approach. Section 3.5 presents results of a changepoint analysis of network authentication data, illustrating the practical benefits of the proposed model. Section 3.6 summarises the main contributions of this chapter.

3.1. Motivation: encoding prior beliefs on cyber attacks

To motivate an informative graph-based prior for changepoints, we consider an application of changepoint detection in cyber-security. A cyber-attack typically changes the behaviour of connected endpoints on the target computer network (Sexton, Storlie, and Neil, 2015). Therefore, to detect the presence of a network intrusion, it is informative to monitor for synchronous, or quasi-synchronous, changes in the behaviour of entities that are *a priori* known to be linked by network connectivity.

3.1.1. Monitoring the authentication activity of users

Consider the authentication data from LANL (Kent, 2015) that are discussed in Section 1.1.1. The set of users in the computer network is denoted V. For each user, it is of interest to detect temporal changes in the distribution of network logons across source computers as possible evidence for

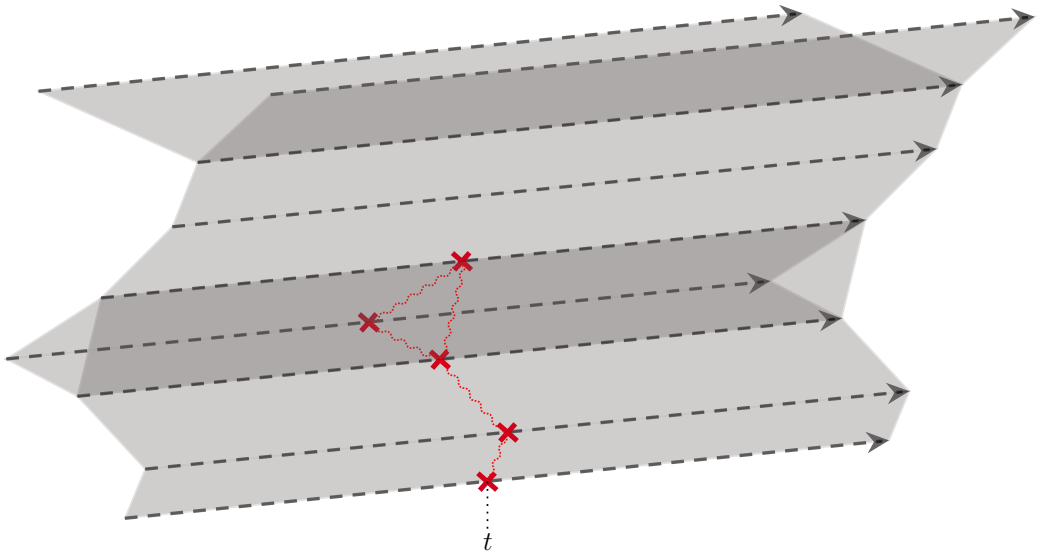


Figure 3.1.: *Cartoon representation of a cluster of synchronous changepoints (red crosses) on a graph of time series. Arrows represent time series, and shaded rectangles indicate which pairs of time series are likely to be impacted by simultaneous changepoints.*

malicious activity. For each user $i \in V$, let

$$\boldsymbol{x}_i = (x_{i,0}, \ldots, x_{i,T}) \tag{3.2}$$

denote the time series of hourly counts of network logons per source computer defined in (1.1). Figure 1.1 displays the authentication data for two users. For each user $i \in V$, it is assumed that $\boldsymbol{x}_i$ follow the changepoint model defined in Section 2.1. The changepoint parameters impacting $\boldsymbol{x}_i$ are denoted by $(k_i, \boldsymbol{\tau}_i)$, where $k_i \geqslant 0$ and $\boldsymbol{\tau}_i = (\tau_{i,1}, \ldots, \tau_{i,k_i+1}) \in \mathcal{T}_{k_i}$. The segment model (2.4) is assumed such that, independently for all t,

$$x_{i,t} \sim f_i(\cdot \mid \theta_{i,j}), \quad \tau_{i,j-1} \leqslant t < \tau_{i,j}, \tag{3.3}$$

where f_i denotes the probability mass function of the multinomial distribution with unknown probability parameter vectors $\theta_{i,j}$ with an uninformative, conjugate prior Dirichlet($\mathbf{1}^M$), where $\mathbf{1}^M$ denotes the M-dimensional vector of ones. Note that segment parameters can be marginalised due to the conjugacy of their prior distribution. As a result, each changepoint $\tau_{i,j}$ corresponds to a temporal change in the distribution of counts of logons initiated by the user $i \in V$ across M host computers in the network.

3.1.2. Motivation for an informative graph-based changepoint prior

The authentication data (3.2) exhibit much variability, and some observed changes can correspond to legitimate activity. Therefore, to limit the number of false alerts and yet not overlook weak signals from genuine attack footprints, it is key to incorporate prior knowledge in the change detection procedure.

When attackers penetrate a network, they rarely gain access to the target users directly; instead, they typically take control of a vulnerable user, for example via email phishing, and then they move laterally through the network, gaining access and compromising additional users, to achieve their objectives (Sexton, Storlie, and Neil, 2015). Attackers are typically constrained in the way they can navigate the network, and it will often be possible for cyber-security experts to specify a graph $G = (V, E)$, where an edge $(i, i') \in E \subseteq V \times V$ indicates it is believed *a priori* that attackers may switch credentials between user i and user i' at any time during the data collection period. Therefore, it is of interest to encode in the changepoint prior that cyber-attacks are *a priori* likely to result in quasi-synchronous changes in the authentication activity of multiple users that are connected in G. In this chapter, we consider the following specification of G for demonstration purposes: $(i, i') \in E$ if and only if both user i and user i' successfully initiated a network logon from the same source computer on the same day. This choice follows from the following considerations. In Windows operating systems, when a user logs on with their credentials (username and password hash) to a device on the domain, these credentials are cached locally on the device. Credential caching prevents users from continuously having to re-authenticate (single sign-on), and enables them to log on to the device even if the device is disconnected from the network. Attackers will exploit credentials that are cached on devices to upgrade their privileges and move laterally through the network. How long credentials may be cached on devices depends on the enterprise's network settings. In the absence of precise knowledge about the enterprise's network settings, it is reasonable to assume that if both user i and user i' have logged into a device on the same day then both those credentials may be cached on that device during the data collection period. As a result, if attackers had access to that device then they would have the ability to exploit cached credentials to switch credentials between user i and user i'.

In Figure 3.1, for the application of interest, each arrow corresponds to the authentication activity of a user on the network, and shaded rectangles indicate which pairs of users are connected in G and therefore likely to be impacted by simultaneous changes during an attack. It is of interest to encode in the changepoint prior, by means of the graph G, that pairs of users $(i, i') \in E$ are likely to be simultaneously affected by malicious behavioural changes, thereby inducing a changepoint model for the authentication data that borrows strength across connected users in G to detect signals for clusters of synchronous changes, as sketched in Figure 3.1.

In contrast with recent intrusion detection methods (Chen and Zhang, 2015; Chen, 2019; Metelli and Heard, 2019b; Passino, Turcotte, and Heard, 2022), the focus of this chapter is not the temporal

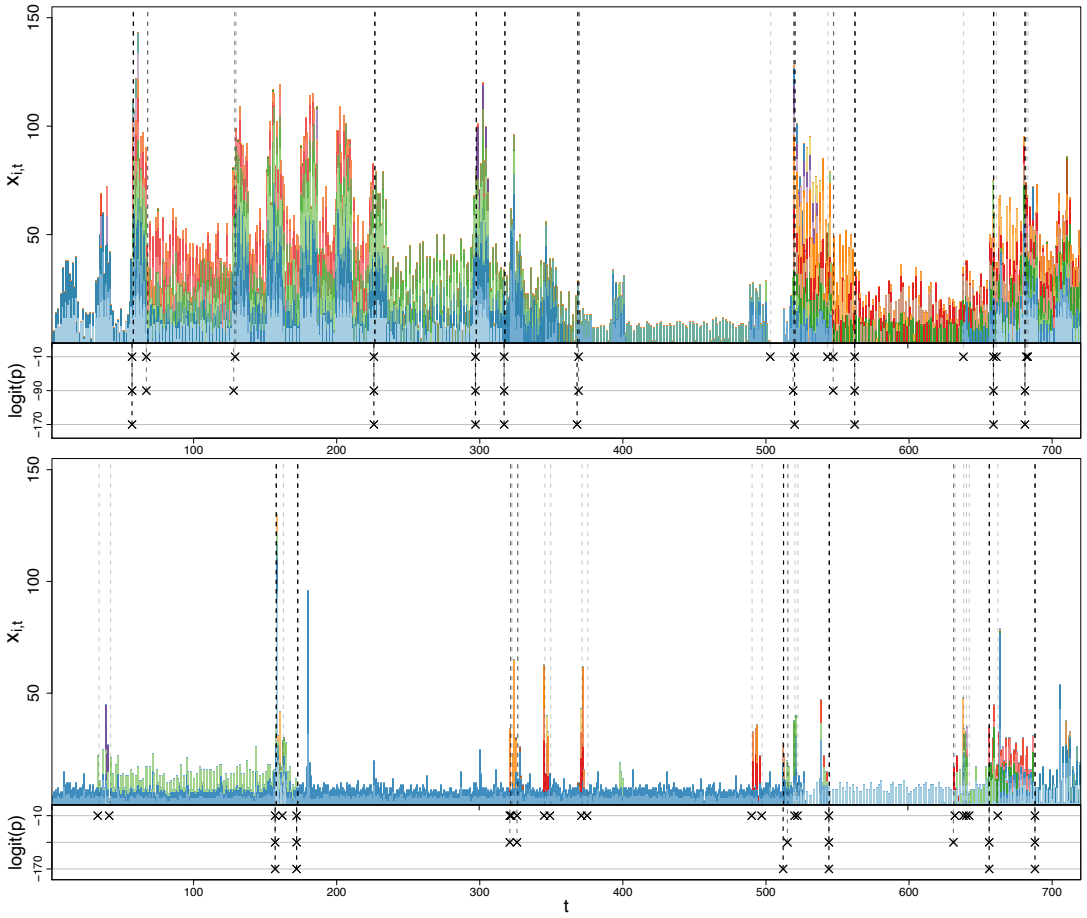


Figure 3.2.: *Hourly counts of network logons per source computer for two users over the first month of data collection; top panel corresponds to user U342@DOM1 and bottom panel to user U86@DOM1. Crosses indicate positions of Bayes changepoint estimates, according to the loss function L with $\gamma = 40$, corresponding to three different Bernoulli parameters p.*

evolution of a graph representing a network, and both V and E are constant in time. The graph G represents the best available static characterisation of the network that can be used to guide change detection in the authentication activity of users (3.2), and it is assumed to be readily available prior to running network intrusion detection methods; note that, in practice, the edge set could be derived from historic data. Section 3.6 discusses possible model extensions for settings where prior beliefs on which time series are likely to be impacted by simultaneous changepoints may be time-dependent.

A standard approach to specify a prior distribution for changepoint parameters $(\boldsymbol{k}, \boldsymbol{\tau})$, where $\boldsymbol{k} = (k_i)_{i \in V}$ and $\boldsymbol{\tau} = (\boldsymbol{\tau}_i)_{i \in V}$, would be to assume that, independently for each user $i \in V$,

changepoint parameters $(k_i, \boldsymbol{\tau}_i)$ follow the Bernoulli process (2.2), such that

$$\pi(\boldsymbol{k}, \boldsymbol{\tau} \mid p) = \prod_{i \in V} p^{k_i}(1-p)^{T-k_i} \mathbf{1}_{\mathcal{T}_{k_i}}(\boldsymbol{\tau}_i) \tag{3.4}$$

for some Bernoulli parameter $0 < p < 1$, which encodes prior belief on the expected number of changepoints. For the cyber-security application where G represents a network of users, the prior (3.4) cannot fully encode prior beliefs on changepoints. Figure 3.2 displays the authentication data for two users. For a selection of Bernoulli parameters p, independently for each user, a sample of size 50,000 from the posterior distribution of changepoints was obtained via the reversible jump MCMC algorithm discussed in Section 2.3.3 after a burn-in of length 10,000. In Figure 3.2, for different Bernoulli parameters, the crosses and the red lines indicate the positions of Bayes changepoint estimates corresponding to the loss function ℓ given in Definition 3 with $\gamma = 40$; it is noticeable that p controls *a priori* the level of granularity of the segmentation of the data. The smaller p, the stronger the evidence required from the data to suggest *a posteriori* inferred changes. No choice of p seems satisfactory: choosing a small value for p will limit the number of false alerts due to noise in user-specific legitimate activity; yet it will also prevent the detection of weak signals for changes shared by different users which are linked in the network, that may be of great interest. It would be preferable to specify *a priori* that changepoints are more likely to occur simultaneously across time series that are linked in G, in order to require strong evidence from the data for changes impacting a single user, or possibly weak signals for changes that impact multiple users linked in the network.

3.2. Informative prior for changepoints on a graph of time series

We begin with a description of the setting of interest. Let $G = (V, E)$ be a graph with node set $V = \{1, \ldots, N\}$ and edge set $E \subseteq V \times V$. For each node $i \in V$ we observe a time series $\boldsymbol{x}_i = (x_{i,0}, \ldots, x_{i,T})$ that is assumed to follow the parametric changepoint model presented in Section 2.1. Let $(k_i, \boldsymbol{\tau}_i)$, where $k_i \geqslant 0$ and $\boldsymbol{\tau}_i = (\tau_{i,1}, \ldots, \tau_{i,k_i+1}) \in \mathcal{T}_{k_i}$, denote the changepoint parameters for the time series i. The distribution of the data conditional on changepoints may differ from one time series to another. Conditional on changepoints, the data are independent of E. The edge set E indicates which pairs of time series are *a priori* likely to be impacted by quasi-simultaneous changepoints.

This section proposes a novel prior for the changepoint parameters $(\boldsymbol{k}, \boldsymbol{\tau})$, where $\boldsymbol{k} = (k_i)_{i \in V}$ and $\boldsymbol{\tau} = (\boldsymbol{\tau}_i)_{i \in V}$, which encodes prior information contained in E. Changepoints are modelled by means of an undirected graphical model encoding that pairs of time series $(i, i') \in E$ are *a priori* likely to be simultaneously affected by changepoints. The graphical model is further extended by relaxing the assumption that dependent changepoints across time series are synchronous; the

extended model assumes dependent changepoints across time series correspond to nearby but not necessarily identical time points.

3.2.1. Synchronous dependent changepoints across time series

3.2.1.1. Model definition

So far, changepoints were most simply defined in terms of their number and locations, $(\boldsymbol{k}, \boldsymbol{\tau})$. Subsequently, it will be useful to represent changepoints by means of a binary matrix. For changepoint parameters $(\boldsymbol{k}, \boldsymbol{\tau})$, let $\boldsymbol{S} = (S_{i,t})$ be the corresponding binary matrix such that, for all $i \in V$ and $t = 1, \ldots, T$,

$$S_{i,t} = \begin{cases} 1 & \text{if } \exists j \in \{1, \ldots, k_i\} \text{ s.t. } t = \tau_{i,j} \\ 0 & \text{otherwise,} \end{cases} \tag{3.5}$$

so that $(\boldsymbol{k}, \boldsymbol{\tau})$ and $\boldsymbol{S}$ are equivalent representations of the changepoints. Moreover, let $S_{i,0} = S_{i,T+1} = 1$ for all i.

To encode the dependence structure of synchronous changepoints across time series in $G = (V, E)$, let $\boldsymbol{\lambda} = (\lambda_{i,i'})$ be a symmetric matrix of non-negative edge weights for the graph satisfying $\lambda_{i,i'} > 0$ if and only if $(i, i') \in E$ for all $i, i' \in V$. Then, conditional on $\boldsymbol{\lambda}$, changepoints are assumed to have a prior distribution described by the weighted, undirected graph G such that, for all $(\boldsymbol{k}, \boldsymbol{\tau})$,

$$\pi(\boldsymbol{k}, \boldsymbol{\tau} \mid p, \boldsymbol{\lambda}) = \frac{1}{Z(p, \boldsymbol{\lambda})} \prod_{t=1}^{T} \exp\left\{ \bar{p} \sum_{i \in V} S_{i,t} + \sum_{i<i'} \lambda_{i,i'} S_{i,t} S_{i',t} \right\}, \tag{3.6}$$

for some $0 < p < 1$, where $\bar{p} = \text{logit}(p) = \log\{p/(1-p)\}$ and the normalising constant $Z(p, \boldsymbol{\lambda})$ has no convenient closed form in general.

If the edge set E is the empty set, implying $\lambda_{i,i'} = 0$ for all i and i', then the prior distribution in (3.6) is equivalent to the standard prior for independent changepoints (2.2); for all changepoint parameters $(\boldsymbol{k}, \boldsymbol{\tau})$ and for all $0 < p < 1$,

$$\pi(\boldsymbol{k}, \boldsymbol{\tau} \mid p, \boldsymbol{0}) = \prod_{i \in V} p^{\sum_{t=1}^{T} S_{i,t}} (1-p)^{T - \sum_{t=1}^{T} S_{i,t}}, \tag{3.7}$$

where $\boldsymbol{0}$ is the null matrix. The memoryless property of the standard prior (Fearnhead, 2006) is maintained by the extended prior (3.6), conditional on fixed value of p. The latter assumes independent and identical Markov random fields (Lauritzen, 1996) for the columns of $\boldsymbol{S}$. The memoryless property would be lost if (3.6) were marginalised over a prior distribution for p.

The graphical prior distribution (3.6) takes into account both the number of changepoints across

time series and their relative positions; the parameter p controls prior belief on the sparsity of changepoints, and the edge weight parameters $\boldsymbol{\lambda}$ control the synchronisation of changepoints between time series. For all pairs (i, i'), the larger the edge weight $\lambda_{i,i'} > 0$, the higher the probability for time series i and i' to be simultaneously affected by changepoints. Hence, the prior in (3.6) may specify changepoints are likely to occur simultaneously across clusters of time series according to G.

To understand how to set the changepoint prior parameters p and $\boldsymbol{\lambda}$ in practice, it is instructive to consider the conditional prior distribution of the components of the binary matrix $\boldsymbol{S}$. Under (3.6), the conditional distribution of $S_{i,t}$ given $\boldsymbol{S}_{-(i,t)} = \{S_{i',t'} : (i', t') \neq (i, t)\}$ is

$$\pi(S_{i,t} \mid \boldsymbol{S}_{-(i,t)}, p, \boldsymbol{\lambda}) \propto \exp\left\{ S_{i,t} \left(\bar{p} + \sum_{i' \colon (i,i') \in E} \lambda_{i,i'} S_{i',t} \right) \right\}, \quad S_{i,t} \in \{0, 1\}.$$

Therefore, for all i and t, the hyperparameter p corresponds to the prior probability that t is a changepoint for the ith time series given that no changepoints occur at time t for the graph neighbour time series of i; and, for all i' such that $(i, i') \in E$, the interaction parameter $\lambda_{i,i'}$ governs how much the conditional prior probability increases if the neighbour time series i' is impacted by a changepoint at time t. Moreover, to perceive the influence of the changepoint prior parameters on the posterior distribution of changepoints, it is helpful to consider the full conditional distribution of $S_{i,t}$ given $\boldsymbol{S}_{-(i,t)} = \{S_{i',t'};\ (i', t') \neq (i, t)\}$,

$$\begin{aligned}
\pi(S_{i,t} \mid \boldsymbol{S}_{-(i,t)}, p, \boldsymbol{\lambda}, \boldsymbol{x}) &\propto \left(\frac{\mathcal{L}_i(\tau_t', t)\mathcal{L}_i(t, \tau_t'')}{\mathcal{L}_i(\tau_t', \tau_t'')} \right)^{S_{i,t}} \pi(S_{i,t} \mid \boldsymbol{S}_{-(i,t)}, p, \boldsymbol{\lambda}) \\
&\propto \exp\left\{ S_{i,t} \left(\log\left\{ \frac{\mathcal{L}_i(\tau_t', t)\mathcal{L}_i(t, \tau_t'')}{\mathcal{L}_i(\tau_t', \tau_t'')} \right\} + \bar{p} + \sum_{i' \colon (i,i') \in E} \lambda_{i,i'} S_{i',t} \right) \right\},
\end{aligned}$$

where $\tau_t' = \max\{t' : t' < t, S_{i,t'} = 1\}$, $\tau_t'' = \min\{t' : t' > t, S_{i,t'} = 1\}$, and

$$\mathcal{L}_i(t_1, t_2) = \int_\Theta L_i(x_{i,t_1}, \ldots, x_{i,t_2-1} \mid \theta_{i,j}) \pi_i(\theta_{i,j}) d\theta_{i,j} \tag{3.8}$$

for all $t_1 < t_2$. In essence, p determines the level of evidence required from the data to suggest a changepoint, and the edge weight parameters control, relative to p, how weak signals for synchronous changepoints can be combined across time series.

3.2.1.2. A special case: identical edge weight parameters

In practice, it will often be natural to assume that, for all $(i, i') \in E$, $\lambda_{i,i'} = \lambda$ for some fixed value $\lambda > 0$. For all i and t, let

$$n_{i,t} = \sum_{i':(i,i')\in E} S_{i',t}$$

be the number of neighbour time series of i that are affected by a changepoint at time t. Then, under (3.6), the conditional prior distribution of $S_{i,t}$ given $\boldsymbol{S}_{-(i,t)} = \{S_{i',t'} : (i', t') \neq (i, t)\}$ is

$$\pi(S_{i,t} \mid \boldsymbol{S}_{-(i,t)}, p, \lambda) = \frac{\exp\{S_{i,t}(\bar{p} + \lambda n_{i,t})\}}{\exp\{\bar{p} + \lambda n_{i,t}\} + 1}, \quad S_{i,t} \in \{0, 1\}. \tag{3.9}$$

Moreover, λ will typically be chosen relative to $\bar{p}$ and the degree distribution of the nodes in G. For example, it can be convenient to assume $\lambda = \lambda_s|\bar{p}|/n$, where n denotes the maximum degree of the nodes in G, for some $\lambda_s > 0$.

3.2.2. Examples of graphical dependence structures for changepoints

The prior distribution (3.6) is suitable for a wide variety of settings. This section provides graph motifs that can be regarded as building blocks to encode the dependence structure of changepoints across multiple time series. For these examples, we assume identical non-zero edge weights as considered in Section 3.2.1.2 and provide some insight on how to choose the changepoint prior parameters p and λ. These exemplar dependence structures for changepoints are explored via a simulation study in Section 3.4.

3.2.2.1. Lattices

It might be natural to choose the edge set E to induce an $N_1 \times N_2$ lattice graph when the number of time series is $N = N_1 N_2$ for some $N_1, N_2 > 0$. For all i, let $0 \leqslant i_1 \leqslant N_1 - 1$ and $0 \leqslant i_2 \leqslant N_2 - 1$ be the unique natural numbers such that $i = i_2 N_1 + i_1 + 1$. Then the $N_1 \times N_2$ lattice graph is such that $(i, i') \in E$ if and only if $|i_1 - i'_1| + |i_2 - i'_2| = 1$. For example, suppose the data $\boldsymbol{x}$ are recorded for the analysis of some spatio-temporal phenomenon such that $x_{i,t}$ denotes the observation at time t and at the coordinate i of some $N_1 \times N_2$ grid over a map of the region of interest, and it is of interest to detect the times and the coordinates at which the distribution of the data changes.

Figure 3.3 illustrates the dependence structure for changepoints induced by the graphical changepoint prior (3.6) given a lattice graph G on 20 time series of length T. The larger the edge weight $\lambda > 0$, the higher the probability for pairs of time series connected on the lattice graph G to be simultaneously impacted by changepoints. As a result, in Figure 3.3, changepoints at time t', which are connected by edges, are *a priori* more likely than isolated changepoints at time t. The

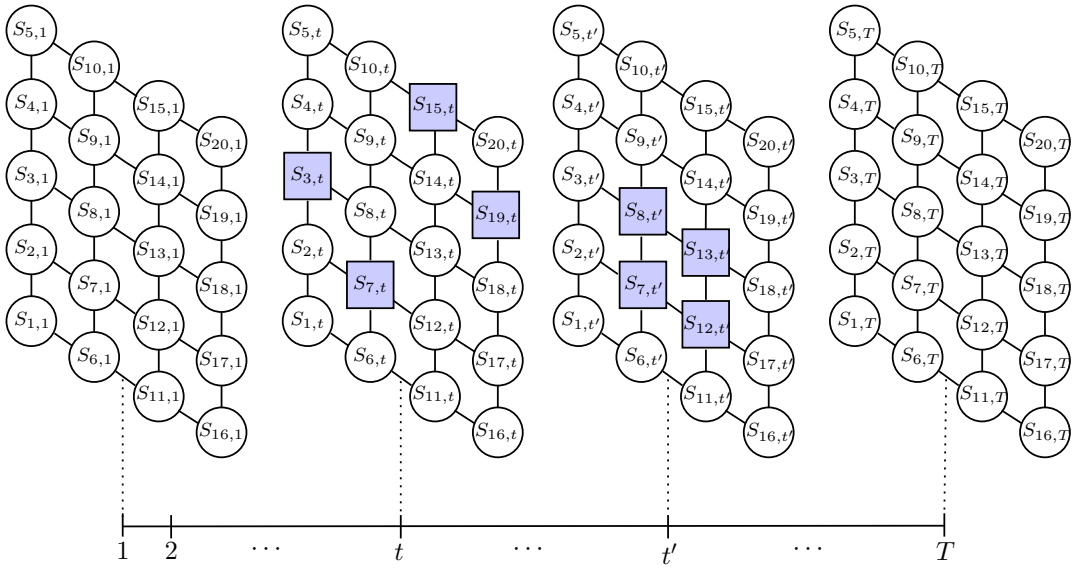


Figure 3.3.: *Cartoon representation of a changepoint matrix $\boldsymbol{S} = (S_{i,t})$, defined in (3.5), for 20 time series of length T that lie on a 5×4 lattice graph G. Edges indicate dependence between components of $\boldsymbol{S}$ according to the prior (3.6) given the lattice graph G. Blue squares indicate $S_{i,t} = 1$ and white circles indicate $S_{i,t} = 0$ for all i and t.*

prior (3.6) can therefore specify that changepoints are likely to occur as clusters of simultaneous changepoints on the lattice. The conditional probability (3.9) specifies that p is the prior probability that a changepoint occurs in isolation on the lattice, and is constrained such that $n_{i,t} \in \{0, \ldots, 4\}$.

3.2.2.2. r-chains

Another dependence structure of interest arises when there is a natural ordering of the time series, which is encoded by the time series indices $1 < \cdots < N$, and changepoints are *a priori* likely to occur as chains of simultaneous changepoints across consecutive time series. For instance, suppose the data consist of multiple time series that are recorded to monitor various aspects of a system; and, it is of interest to detect some event that evolves through multiple phases, such that each phase is likely to manifest through the perturbation of one aspect of the system. In such a setting, it is appropriate to consider the following graph for the time series indices, which we call an r-chain graph: let $(i, i') \in E$ if and only if $1 \leqslant |i - i'| \leqslant r$, for some $r > 0$ chosen to allow gaps of length $r - 1$ within chains of changepoints. Panel (b) of Figure 3.4 gives a cartoon representation of a 2-chain graph, which will be considered in a simulation study. For r-chain graphs, $n_{i,t} \in \{0, \ldots, 2r\}$.

3.2.2.3. Complete graphs

Suppose a complete graph for the time series indices, that is $(i, i') \in E$ for all $i \neq i'$, so that, according to Section 3.2.1.2, $\lambda_{i,i'} = \lambda > 0$ for all $i \neq i'$. In such a setting, the prior given in (3.6) assumes changepoint locations are exchangeable across time series, like the MVCAPA model proposed in Fisch, Eckley, and Fearnhead (2022), and therefore solely takes into account the number of time series impacted by a changepoint at time t, for all t.

3.2.2.4. Unknown graph

This chapter assumes that the graph G is known *a priori* and contains useful information concerning the dependence structure of changepoints across time series. Future work could reverse this idea and consider applications where estimating G is one of the inferential objectives. It might often be computationally unrealistic to specify an unconstrained prior for G admitting that $\lambda_{i,j} \geqslant 0$ for all $(i, i') \in V \times V$. However, in some settings it might be appropriate to consider a class of possible graphs G, such as those considered in the previous two subsections; for example, it may be assumed *a priori* that G is an r-chain with $r \geqslant 0$ unknown.

3.2.3. Extension to asynchronous changepoint dependence

The model in Section 3.2.1 assumes changepoints are likely to simultaneously affect clusters of time series according to the dependence graph G. In this section, we relax this model to allow dependence between changepoints in different time series at nearby points in time. The extended model relies on representing changepoints as lagged realisations of simultaneous but unobserved latent changepoints. The latent changepoints are distributed according to the model introduced in Section 3.2.1 and, conditional on these latent changepoints, time series-specific lags are assumed to be uniformly distributed over some small time window.

Let $(\boldsymbol{k}, \boldsymbol{\tau})$ be changepoint parameters for multiple time series as defined in Section 3.2.1, where it is assumed that the ith time series is subject to k_i changepoints whose positions are denoted $\boldsymbol{\tau}_i = (\tau_{i,1}, \ldots, \tau_{i,k_i}) \in \mathcal{T}_{k_i}$ as defined in (2.1). The asynchronous model further assumes that, for all time series $i = 1, \ldots, N$, there exist latent changepoint positions $\tilde{\boldsymbol{\tau}} = (\tilde{\boldsymbol{\tau}}_1, \ldots, \tilde{\boldsymbol{\tau}}_N)$, $\tilde{\boldsymbol{\tau}}_i = (\tilde{\tau}_{i,1}, \ldots, \tilde{\tau}_{i,k_i}) \in \mathcal{T}_{k_i}$, and lags $\boldsymbol{d} = (\boldsymbol{d}_1, \ldots, \boldsymbol{d}_N)$, $\boldsymbol{d}_i = (d_{i,1}, \ldots, d_{i,k_i}) \in \{0, \ldots, w_i\}^{k_i}$, for $\boldsymbol{w} = (w_1, \ldots, w_N)$, where $w_i \geqslant 0$ is an upper bound for the lags, such that, for all $j = 1, \ldots, k_i$, the jth changepoint for time series i is

$$\tau_{i,j} = \tilde{\tau}_{i,j} + d_{i,j}. \tag{3.10}$$

Let $\tilde{\tau}_{i,0} = \tau_{i,0} = 0$ and $\tilde{\tau}_{i,k_i+1} = \tau_{i,k_i+1} = T + 1$. For all $(k_i, \boldsymbol{\tau}_i)$ and $w_i \geqslant 0$, if $\tilde{\boldsymbol{\tau}}_i = \boldsymbol{\tau}_i$ and $\boldsymbol{d}_i$ is the zero vector then (3.10) holds, and therefore the existence of a corresponding pair $(\tilde{\boldsymbol{\tau}}_i, \boldsymbol{d}_i)$ with $\boldsymbol{\tau}_i \in \mathcal{T}_{k_i}$ is guaranteed. If $w_i = 0$ then the latent changepoints and the changepoints must be

identical; but, in general, given changepoints $(k_i, \boldsymbol{\tau}_i)$ and $w_i > 0$, there are multiple distinct pairs of latent changepoints $(k_i, \tilde{\boldsymbol{\tau}}_i)$ and lags $\boldsymbol{d}_i$ satisfying (3.10) and $\tilde{\boldsymbol{\tau}}_i \in \mathcal{T}_{k_i}$.

For some applications the upper bounds for the lags, $\boldsymbol{w}$, may be fixed. In particular, for some reference time series $i \in V$, it can be set that $w_i = 0$, implying that $\tau_{i,j} = \tilde{\tau}_{i,j}$ for all j, so that changepoints for time series i' with $w_{i'} \geqslant 0$ are lagged relative to changepoints for time series i. However, in general, upper bounds for the lags will not be known. For example, in the motivational application in cyber-security, no user $i \in V$ can be assumed to be the first user to be affected by an attack, making it awkward to pick a reference time series i, and the exact duration of attacks is not known *a priori*. It will be assumed that, independently for all time series i, $w_i \sim \text{Geometric}(\eta)$ for some value $0 < \eta < 1$ chosen to reflect the expected duration of an attack.

Suppose the latent changepoints $(\boldsymbol{k}, \tilde{\boldsymbol{\tau}})$ are distributed according to the prior distribution (3.6) given some $0 < p < 1$ and graph edge weight parameters $\boldsymbol{\lambda}$. Then, independently for all time series i, conditional on w_i and $(k_i, \tilde{\boldsymbol{\tau}}_i)$, the lags $\boldsymbol{d}_i = (d_{i,1}, \ldots, d_{i,k_i})$ are assumed to be uniformly distributed on the set

$$\mathcal{D}(w_i, k_i, \tilde{\boldsymbol{\tau}}_i) = \left\{(d_{i,1}, \ldots, d_{i,k_i}) \in \{0, \ldots, w_i\}^{k_i} : (\tilde{\tau}_{i,1} + d_{i,1}, \ldots, \tilde{\tau}_{i,k_i} + d_{i,k_i}) \in \mathcal{T}_{k_i}\right\}, \quad (3.11)$$

such that, for all $\boldsymbol{d} = (\boldsymbol{d}_1, \ldots, \boldsymbol{d}_N)$,

$$\pi(\boldsymbol{d} \mid \boldsymbol{k}, \tilde{\boldsymbol{\tau}}, \boldsymbol{w}) = \prod_{i=1}^{N} \frac{\mathbf{1}_{\mathcal{D}(w_i, k_i, \tilde{\boldsymbol{\tau}}_i)}(\boldsymbol{d}_i)}{|\ \mathcal{D}(w_i, k_i, \tilde{\boldsymbol{\tau}}_i)\ |}.$$

Proposition 1 gives a recursion to derive the cardinality of (3.11).

Proposition 1. *(Cardinality of $\mathcal{D}$).* Let $w \geqslant 0$, $k \geqslant 0$, $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_k) \in \mathcal{T}_k$ with $\tau_0 = 1$ and $\tau_{k+1} = T + 1$, and let $\mathcal{D}(w, k, \boldsymbol{\tau})$ be the set defined in (3.11).

(i) For all $j \geqslant 1$ and $l \geqslant 0$, let $\rho(j, l) = \min\{w + 1, T + 1 - \tau_j\} - (\tau_{j+l} - \tau_j)$ and

$$Q(j, l) = \frac{(\rho(j, l) + l)!}{(\rho(j, l) - 1)!(l + 1)!} \mathbf{1}_{\{0, \ldots, w\}}(\tau_{j+l} - \tau_j). \quad (3.12)$$

Additionally, let $Z(0) = 1$, $Z(1) = Q(1, 0)$ and, recursively for all $k > 1$,

$$Z(k) = \sum_{j=1}^{k} (-1)^{k-j} Z(j - 1) Q(j, k - j).$$

Then

$$|\mathcal{D}(w, k, \boldsymbol{\tau})| = Z(k). \quad (3.13)$$

In particular, if $k = 0$ then $\boldsymbol{\tau}$ is the empty sequence and $\mathcal{D}(w, k, \boldsymbol{\tau})$ contains a unique element, namely the empty sequence.

(ii) $|\mathcal{D}(w, k, \boldsymbol{\tau})| \leqslant (w+1)^k$ and the equality holds if and only if $\tau_{j+1} - \tau_j > w$ for all j.

Proof. See Appendix B.1. □

Consequently, the joint prior density for $(\boldsymbol{k}, \tilde{\boldsymbol{\tau}}, \boldsymbol{d})$ is

$$\pi(\boldsymbol{k}, \tilde{\boldsymbol{\tau}}, \boldsymbol{d} \mid p, \boldsymbol{\lambda}, \boldsymbol{w}) = \frac{\pi(\boldsymbol{k}, \tilde{\boldsymbol{\tau}} \mid p, \boldsymbol{\lambda})}{\prod_{i=1}^{N} |\mathcal{D}(w_i, k_i, \boldsymbol{\tau}_i)|} \tag{3.14}$$

and the induced changepoint prior distribution for $(\boldsymbol{k}, \boldsymbol{\tau})$ is

$$\pi(\boldsymbol{k}, \boldsymbol{\tau} \mid p, \boldsymbol{\lambda}, \boldsymbol{w}) = \sum_{(\tilde{\boldsymbol{\tau}}, \boldsymbol{d}) \in \Upsilon(\boldsymbol{k}, \boldsymbol{\tau}, \boldsymbol{w})} \pi(\boldsymbol{k}, \tilde{\boldsymbol{\tau}}, \boldsymbol{d} \mid p, \boldsymbol{\lambda}, \boldsymbol{w}), \tag{3.15}$$

where $\Upsilon(\boldsymbol{k}, \boldsymbol{\tau}, \boldsymbol{w})$ denotes the set of pairs of latent changepoints and lags, $(\boldsymbol{k}, \tilde{\boldsymbol{\tau}}, \boldsymbol{d})$, that identify the changepoints $(\boldsymbol{k}, \boldsymbol{\tau})$ according to (3.10).

3.3. Inference via RJMCMC with auxiliary variables

Joint sampling of changepoints across time series is required when the assumption of independence for changepoints is relaxed. In this section, we propose a reversible jump MCMC algorithm (Green, 1995) to sample changepoints in multiple time series, $(\boldsymbol{k}, \boldsymbol{\tau})$. The changepoint parameter space is augmented with auxiliary variables (Besag and Green, 1993; Higdon, 1998) that induce clusters of time series indices according to G. Then, the reversible jump MCMC algorithm of Denison et al. (2002) is extended to sample from the augmented parameter space, thereby providing a means to efficiently explore the changepoint parameter space. At each iteration of the algorithm, a new cluster of changepoints may be proposed or an existing cluster of changepoints may be deleted or shifted. The validity of the proposed MCMC algorithm follows immediately from the reversibility of the proposed moves.

For notational simplicity it is assumed there are no missing data, but even with data that are not independent and identically distributed within segments, any missing observations would present no methodological complication, since missing data can be sampled from their predictive distribution within the proposed MCMC scheme (Gelman et al., 2004).

The sampling strategy proposed in this section will assume that segment parameters can be marginalised. Section 3.3.4 discusses how the sampling strategy could be adapted to settings where segment parameters cannot be marginalised.

3.3.1. Sampler for synchronous dependent changepoints

We begin by proposing an MCMC algorithm to sample from the posterior distribution of changepoints when changepoint parameters are *a priori* distributed according to the prior introduced in Section 3.2.1, $\pi(\boldsymbol{k}, \boldsymbol{\tau} \mid p, \boldsymbol{\lambda})$, given $0 < p < 1$ and some interaction parameters $\boldsymbol{\lambda}$.

To sample changepoints for multiple time series, consider the following adaptation of the standard MCMC algorithm to sample changepoints for a unique time series (Denison et al., 2002), which will be called the "single site updating" MCMC algorithm thereafter. At each iteration of the algorithm, with $(\boldsymbol{k}, \boldsymbol{\tau})$ denoting the latest particle of the sample chain, propose one of the following two moves: for a uniformly chosen index (i, t), propose $S_{i,t}$ to be updated to $1 - S_{i,t}$, thereby allowing birth or death of a changepoint; alternatively, the position of a randomly chosen changepoint, $\tau_{i,j}$, is sampled uniformly from $\{\tau_{i,j-1} + 1, \ldots, \tau_{i,j+1} - 1\}$.

With the graphical prior distribution (3.6), synchronous changepoints can be correlated across time series, and the single site updating MCMC algorithm can become impractical, as illustrated in Section 3.4.4.3 through a simulation study. Instead, it will be necessary to propose moves that allow birth, death or shift of clusters of synchronous changepoints according to the graph induced by $\boldsymbol{\lambda}$.

3.3.1.1. Augmenting the parameter space with auxiliary variables

To provide a means of moving efficiently through the state space of the changepoint parameters, the parameter space is augmented with binary auxiliary variables $\boldsymbol{u} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_T)$ such that, for all t, $\boldsymbol{u}_t$ is an $N \times N$ symmetric binary graph adjacency matrix with (i, i') element $u_t(i, i')$. For all t, the prior density of $\boldsymbol{u}_t$ is assumed to take the conditionally independent form

$$\pi(\boldsymbol{u}_t \mid \boldsymbol{\lambda}, \delta, \boldsymbol{k}, \boldsymbol{\tau}) = \prod_{i<i'} q_t(i, i')^{u_t(i,i')} \{1 - q_t(i, i')\}^{1-u_t(i,i')}, \tag{3.16}$$

where

$$q_t(i, i') = 1 - \exp\{-\delta \lambda_{i,i'} (1 - \mid S_{i,t} - S_{i',t} \mid)\} \tag{3.17}$$

is the conditional probability that $u_t(i, i') = 1$, given a partial decoupling parameter $\delta \geqslant 0$ (Higdon, 1998) whose role will be discussed in Section 3.3.1.2. After observing data $\boldsymbol{x}$ distributed according to (3.3), the joint posterior density of the augmented parameters $(\boldsymbol{k}, \boldsymbol{\tau}, \boldsymbol{u})$ is

$$\pi(\boldsymbol{k}, \boldsymbol{\tau}, \boldsymbol{u} \mid p, \boldsymbol{\lambda}, \boldsymbol{x}, \delta) = \pi(\boldsymbol{u} \mid \boldsymbol{\lambda}, \delta, \boldsymbol{k}, \boldsymbol{\tau}) \pi(\boldsymbol{k}, \boldsymbol{\tau} \mid p, \boldsymbol{\lambda}, \boldsymbol{x}), \tag{3.18}$$

where

$$\pi(\boldsymbol{u} \mid \boldsymbol{\lambda}, \delta, \boldsymbol{k}, \boldsymbol{\tau}) = \prod_{t=1}^{T} \pi(\boldsymbol{u}_t \mid \boldsymbol{\lambda}, \delta, \boldsymbol{k}, \boldsymbol{\tau}).$$

For all t, consider the graph $H_t = (V, E_t)$ with vertex set $V = \{1, \ldots, N\}$ and edge set E_t such that $(i, i') \in E_t$ if and only if $u_t(i, i') = 1$ for all $i, i' \in V$. According to (3.17), if $u_t(i, i') = 1$ then $q_t(i, i') > 0$ and, consequently, $S_{i,t} = S_{i',t}$. As a result, with $\mathcal{C}_t$ denoting the set of connected components of H_t, for all clusters of time series $\gamma \in \mathcal{C}_t$, $S_{i,t} = S_{i',t}$ for all $i, i' \in \gamma$. In other words, the auxiliary variables $\boldsymbol{u}_t$ induce a partition of the time series, $\mathcal{C}_t$, such that, for each cluster $\gamma \in \mathcal{C}_t$, either all time series or no time series in γ are affected by a changepoint at time t. Moreover, according to (3.17), if $S_{i,t} = S_{i',t}$ then the conditional probability that $u_t(i, i') = 1$ increases with $\lambda_{i,i'} > 0$, so that clusters induced by $\boldsymbol{u}_t$ will tend to be clusters on the graph induced by the edge weight parameters.

3.3.1.2. MCMC algorithm

To generate realisations from the posterior distribution of the changepoints, we consider a "cluster updating" MCMC algorithm that samples from the extended joint posterior density (3.18). By inducing clusters of time series determined by the edge weight parameters for each time point t, the auxiliary variables $\boldsymbol{u}$ provide a means to efficiently explore the state space of $(\boldsymbol{k}, \boldsymbol{\tau})$.

The parameter δ is a tuning parameter for the cluster updating MCMC algorithm. The size of clusters will tend to increase with δ; in particular, if $\delta = 0$ then, for all t, each cluster corresponds to a unique time series index, even if the edge weights of the dependence graph are large, so that the "cluster updating" MCMC algorithm reduces to the "single site updating" algorithm. Typically δ is fixed (Higdon, 1998) to control the probabilities in (3.16) and therefore the expected size of clusters. However, we propose to treat $\delta \geqslant 0$ as an unknown parameter with prior distribution $\pi(\delta)$, so that expected sizes of cluster may vary in the sample; specifically, we assume that $\delta = 0$ with probability $0 \leqslant \delta_0 \leqslant 1$ and otherwise δ is drawn from Beta(δ_1, δ_2) for $\delta_1, \delta_2 > 0$.

For the cluster updating MCMC algorithm, at each iteration of the algorithm, with $(\boldsymbol{k}, \boldsymbol{\tau}, \boldsymbol{u}, \delta)$ denoting the latest particle of the sample chain, one of the following moves is proposed.

Birth/death move

Conditional on the auxiliary variables, the birth/death move proposes the birth or death of a cluster of synchronous changepoints. Sample t' uniformly from $\{1, \ldots, T\}$. A cluster γ of time series indices is randomly chosen from $\mathcal{C}_{t'}$, the set of clusters of time series induced by the auxiliary variables $\boldsymbol{u}_{t'}$. Then, leaving the auxiliary variables unchanged, propose changepoint parameters

$(\boldsymbol{k}', \boldsymbol{\tau}')$ such that, for all $i = 1, \ldots, N$ and $t = 1, \ldots, T$,

$$S'_{i,t} = \begin{cases} 1 - S_{i,t} & \text{if } i \in \gamma \text{ and } t = t' \\ S_{i,t} & \text{otherwise,} \end{cases}$$

where $\boldsymbol{S}$ and $\boldsymbol{S}'$ are binary matrix representations of $(\boldsymbol{k}, \boldsymbol{\tau})$ and $(\boldsymbol{k}', \boldsymbol{\tau}')$ according to (3.5), respectively.

Shift move

The shift move proposes to shift the position of a cluster of synchronous changepoints. First, a time unit t is uniformly chosen from $\{t : \sum_{i=1}^{N} S_{i,t} > 0\}$. Let $\mathcal{C}_t^* \subseteq \mathcal{C}_t$ denote the set of clusters of time series indices γ induced by $\boldsymbol{u}_t$ such that, for all $i \in \gamma$, $S_{i,t} = 1$. A cluster γ is uniformly chosen from $\mathcal{C}_t^*$. For all $i \in \gamma$, let j_i be the index of the changepoint with position t for the ith time series, that is $\tau_{i,j_i} = t$. Then, sample uniformly t' from $\bigcap_{i \in \gamma} \{\tau_{i,j_i-1} + 1, \ldots, \tau_{i,j_i+1} - 1\}$ and propose changepoint parameters $(\boldsymbol{k}, \boldsymbol{\tau}')$ that are identical to $(\boldsymbol{k}, \boldsymbol{\tau})$ but with $\tau'_{i,j_i} = t'$, for all $i \in \gamma$.

In parallel, it is required to propose auxiliary variables $\boldsymbol{u}'$ that are adapted to $(\boldsymbol{k}, \boldsymbol{\tau}')$. The updated auxiliary variables differ from $\boldsymbol{u}$ as follows. For all $i, i' \in \gamma$, $u'_{t'}(i, i') = u_t(i, i')$ and $u'_t(i, i') = u_{t'}(i, i')$; for all $i \in \gamma$ and $i' \notin \gamma$ such that $t' \notin \boldsymbol{\tau}_{i'}$, $u'_{t'}(i, i') = 0$; and, ensuring reversibility of the move, for all $i \in \gamma$ and $i' \notin \gamma$ such that $t \notin \boldsymbol{\tau}_{i'}$, $u'_t(i, i')$ is sampled conditionally on $(\boldsymbol{k}, \boldsymbol{\tau}')$ according to the Bernoulli $\big(1 - \exp\{-\delta\lambda_{i,i'}(1 - |S_{i,t} - S_{i',t}|)\}\big)$ target distribution implied by (3.16).

Update of auxiliary variables

Changepoints are left unchanged, δ is sampled from its prior distribution, and auxiliary variables are sampled from the full conditional distribution given in (3.16), thereby proposing an updated clustering of time series indices for all t.

3.3.2. Sampler for asynchronous dependent changepoints

According to the changepoint model (3.15) introduced in Section 3.2.3, changepoints do not need to occur at the same time to be related. Consequently, to explore the changepoint parameter space it will be required to propose the birth, death or shift of clusters of asynchronous changepoints. This section extends the MCMC algorithm from Section 3.3.1 to sample from the posterior distribution of changepoints when changepoint parameters are *a priori* distributed according to $\pi(\boldsymbol{k}, \boldsymbol{\tau} \mid p, \boldsymbol{\lambda}, \boldsymbol{w})$ from (3.15).

Recall that under the asynchronous model, changepoints $(\boldsymbol{k}, \boldsymbol{\tau})$ are deterministically specified by latent changepoints $(\boldsymbol{k}, \tilde{\boldsymbol{\tau}})$ and unknown lags $\boldsymbol{d}$ according to (3.10). Therefore, a sample from $(\boldsymbol{k}, \boldsymbol{\tau})$ can be obtained from a sample from $(\boldsymbol{k}, \tilde{\boldsymbol{\tau}}, \boldsymbol{d})$. Next, we propose a sampler from the joint

posterior distribution of $(\boldsymbol{k}, \tilde{\boldsymbol{\tau}}, \boldsymbol{d})$, updated from the prior density (3.14) by the observed data $\boldsymbol{x}$, thereby providing a means to obtain a sample from the posterior distribution of $(\boldsymbol{k}, \boldsymbol{\tau})$.

As in Section 3.3.1, the parameter space is augmented with auxiliary variables $\boldsymbol{u} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_T)$ to facilitate the exploration of the state space of the parameters of interest $(\boldsymbol{k}, \tilde{\boldsymbol{\tau}}, \boldsymbol{d})$. Conditionally on latent changepoints $(\boldsymbol{k}, \tilde{\boldsymbol{\tau}})$ and independently of the lags and the data, for all t and $i < i'$, $u_t(i, i')$ is assumed to be distributed according to (3.16), such that $u_t(i, i') \sim \text{Bernoulli}\left(1 - \exp\{-\delta\lambda_{i,i'}(1 - |\tilde{S}_{i,t} - \tilde{S}_{i',t}|)\}\right)$, where $\tilde{\boldsymbol{S}}$ is the binary matrix representation of $(\boldsymbol{k}, \tilde{\boldsymbol{\tau}})$ according to (3.5). As described in Section 3.3.1.1, it follows that the auxiliary variables $\boldsymbol{u}_t$ induce a partition of the time series, $\mathcal{C}_t$, such that, for each cluster $\gamma \in \mathcal{C}_t$, either all time series or no time series in γ are affected by a latent changepoint at time t.

The joint posterior density of the augmented parameters $(\boldsymbol{k}, \tilde{\boldsymbol{\tau}}, \boldsymbol{d}, \boldsymbol{u})$ is

$$\pi(\boldsymbol{k}, \tilde{\boldsymbol{\tau}}, \boldsymbol{d}, \boldsymbol{u} \mid p, \boldsymbol{\lambda}, \boldsymbol{w}, \boldsymbol{x}, \delta) = \pi(\boldsymbol{u} \mid \boldsymbol{\lambda}, \delta, \boldsymbol{k}, \tilde{\boldsymbol{\tau}})\pi(\boldsymbol{k}, \tilde{\boldsymbol{\tau}}, \boldsymbol{d} \mid p, \boldsymbol{\lambda}, \boldsymbol{w}, \boldsymbol{x}).$$

To sample from the posterior distribution of $(\boldsymbol{k}, \tilde{\boldsymbol{\tau}}, \boldsymbol{d}, \boldsymbol{u})$, or $(\boldsymbol{k}, \tilde{\boldsymbol{\tau}}, \boldsymbol{d}, \boldsymbol{u}, \boldsymbol{w})$ if upper bounds for the lags are *a priori* unknown, the MCMC algorithm discussed in Section 3.3.1 is extended as follows: the birth/death and shift moves are adapted to pairs of latent changepoints and lags; and additional moves are introduced for updating the lags. For the lags, note that according to (3.11), for all $i = 1, \ldots, N$ and $j = 1, \ldots, k_i$, to maintain monotonicity in the changepoints, the lag associated to the jth changepoint of the ith time series must satisfy

$$d_{i,j} \in \mathcal{D}_j(w_i, k_i, \tilde{\boldsymbol{\tau}}_i) = \{\ell \in \mathbb{N};\, d_- \leqslant \ell \leqslant d^+\},$$

where $d_- = \max(0, d_{i,j-1} + \tau_{i,j-i} - \tau_{i,j} + 1)$ and $d^+ = \min(w_i, d_{i,j+1} + \tau_{i,j+1} - \tau_{i,j} - 1)$; and the full conditional probability distribution of $d_{i,j}$ is such that, for all $d_{i,j} \in \mathcal{D}_j(w_i, k_i, \tilde{\boldsymbol{\tau}}_i)$,

$$\pi\left(d_{i,j} \mid \boldsymbol{d}_{-(i,j)}, k_i, \tilde{\boldsymbol{\tau}}_i, w_i, \boldsymbol{x}_i\right) \propto \mathcal{L}_i\left(\tilde{\tau}_{i,j-1} + d_{i,j-1}, \tilde{\tau}_{i,j} + d_{i,j}\right) \mathcal{L}_i\left(\tilde{\tau}_{i,j} + d_{i,j}, \tilde{\tau}_{i,j+1} + d_{i,j+1}\right) \tag{3.19}$$

where $\boldsymbol{d}_{-(i,j)} = \{d_{i',j'};\, (i', j') \neq (i, j)\}$ and $\mathcal{L}_i$ is defined in (3.8).

For the extended cluster updating MCMC algorithm, at each iteration of the algorithm, with $(\boldsymbol{k}, \tilde{\boldsymbol{\tau}}, \boldsymbol{d}, \boldsymbol{u}, \boldsymbol{w})$ denoting the latest particle of the sample chain, one of the following moves is proposed.

Extended birth/death move

The extended birth/death move proposes the birth or death of a cluster of asynchronous changepoints. First, conditionally on the auxiliary variables, latent changepoints $(\boldsymbol{k}', \tilde{\boldsymbol{\tau}}')$ are proposed according to the birth/death move detailed in Section 3.3.1: for all time series $i \in \gamma \subseteq \{1, \ldots, N\}$, the

birth or death of latent changepoint with position t is proposed. Then, updated lags are proposed conditional on $(\boldsymbol{k}', \tilde{\boldsymbol{\tau}}')$: If the birth of changepoints is proposed, then, for all time series $i \in \gamma$, there is j_i such that $\tilde{\tau}'_{i,j_i} = t$, and the lags $\boldsymbol{d}'_i = (d_{i,1}, \ldots, d_{i,j_i-1}, d'_{i,j_i}, d_{i,j_i} \ldots, d_{i,k_i})$ are proposed for the ith time series, where d'_{i,j_i} is sampled from the full conditional distribution (3.19); otherwise, if the death of changepoints is proposed, then, for all $i \in \gamma$, there is j_i such that $\tilde{\tau}_{i,j_i} = t$, and the lags $\boldsymbol{d}'_i = (d_{i,1}, \ldots, d_{i,j_i-1}, d_{i,j_i+1} \ldots, d_{i,k_i})$ are proposed for the ith time series.

Extended shift move

The extended shift move proposes to shift the positions of a cluster of asynchronous changepoints. First, latent changepoints $(\boldsymbol{k}', \tilde{\boldsymbol{\tau}}')$ and auxiliary variables $\boldsymbol{u}'$ are proposed according to the shift move discussed in Section 3.3.1: for all time series with index $i \in \gamma$, the position t' is proposed for latent changepoint with position t. Then, for all $i \in \gamma$, letting j_i denote the index such that $\tilde{\tau}'_{i,j_i} = t'$, propose d'_{i,j_i} from the full conditional distribution (3.19).

Update of auxiliary variables

δ is sampled from its prior distribution and, conditional on latent changepoints $(\boldsymbol{k}, \tilde{\boldsymbol{\tau}})$, auxiliary variables are sampled from their full conditional distribution (3.16).

Update of lags

A pair (i, j) is uniformly chosen from $\{(i, j);\ i = 1, \ldots, N \text{ and } j = 1, \ldots, k_i\}$, and the lag $d_{i,j}$ is sampled from the full conditional distribution given in (3.19).

Update of upper bounds for lags

If the maximal lags $\boldsymbol{w}$ are *a priori* unknown, for a randomly chosen time series with index i it is proposed to update w_i to $w'_i = w_i + \sigma$ with probability $1/2$, and to update w_i to $w'_i = |w_i - \sigma|$ otherwise, where σ is drawn from Geometric(ρ) for some $0 < \rho < 1$. Proposing to update w_i requires proposing updated lags $\boldsymbol{d}'_i = (d'_{i,1}, \ldots, d'_{i,k_i}) \in \mathcal{D}(w'_i, k_i, \tilde{\boldsymbol{\tau}}_i)$ for the ith time series. For $j = 1, \ldots, k_i$, given w'_i and $(d'_{i,1}, \ldots, d'_{i,j-1}, d_{i,j+1} \ldots d_{i,k_i})$, the lag $d'_{i,j}$ is sampled from the full conditional distribution (3.19).

3.3.3. Time complexity and initialisation of the sampler

The move proposing to update auxiliary variables $\boldsymbol{u} = (\boldsymbol{u}_t)$ from their full conditional distribution (3.16) is independent of preceding auxiliary variables in the sample chain and is accepted with probability 1. Moreover, the auxiliary variables $\boldsymbol{u}$ provide a means to explore the changepoint parameter space but it is not of interest to store a sample from the posterior distribution of $\boldsymbol{u}$. Therefore, for all t, the sampling of the auxiliary parameters $\boldsymbol{u}_t$ is performed only when a subsequent

move depends upon $\boldsymbol{u}_t$, for example when a move proposes to update a cluster of changepoints at time t. As a result, the cost of updating auxiliary variables is at most $\mathcal{O}(|E|)$ per iteration. Moreover, for the shift and the birth/death moves that propose to update changepoints for a cluster of time series C, an $\mathcal{O}(T)$ Bayes factor calculation for the model in (3.3) is required for each time series in C. If $\delta = 0$ then $|C| = 1$; otherwise, the expected size of the cluster $1 \leqslant |C| \leqslant N$ will tend to increase with N. Hence, the computational cost of moves proposing to update changepoints is at most $\mathcal{O}(TN)$.

To speed-up the convergence of the sampler, we may consider the following initialisation of the sample chain: latent changepoints are set to be the optimal Bayesian estimates, according to the loss function later specified in Definition 3, corresponding to the standard model with independent changepoints across time series with prior $\pi(\boldsymbol{k}, \boldsymbol{\tau} \mid p)$, obtained in parallel for each time series; initial lags, upper bounds for lags and auxiliary variables are set to 0. As a result, the burn-in for the sampler begins with a sensible positioning of changepoints at a limited cost: no iterations for the burn-in of the joint sampling of changepoints across time series are wasted identifying strong signals for changepoints, relative to $\bar{p}$.

Section 3.4.3 presents results of a simulation study demonstrating the feasibility of the proposed inference via MCMC for a range of N and T that can correspond to realistic scenarios. For high-dimensional applications where the time until reaching suitable convergence can be prohibitively long, future work could examine analytical approximations to the posterior distribution of changepoints, for example via variational methods (Blei, Kucukelbir, and McAuliffe, 2017).

3.3.4. Possible sampling algorithm extensions for segment parameters

The proposed sampling strategy assumes that segment parameters can be marginalised to compute the conditional likelihood of the data given changepoints, $\mathcal{L}(\boldsymbol{x} \mid \boldsymbol{k}, \boldsymbol{\tau})$. Although this assumption is appropriate in many applications, as discussed in Section 3.1, it will not always be the case; for example when segment parameters are dependent across segments (Chib, 1998; Peluso, Chib, and Mira, 2019; Fearnhead and Liu, 2011). If segment parameters cannot be marginalised, the proposed reversible jump MCMC algorithm can be adapted to sample segment parameters alongside changepoint parameters as in Green (1995): a move is introduced to sample segment parameters conditional on changepoints; the moves proposing updated changepoint parameters are extended to propose suitable parameters for the segments that are affected by the proposed change of changepoint parameters; note that the proposal distributions for segment parameters depend on the segment model of interest.

Moreover, an interesting model extension of the changepoint model in (3.3) consists in specifying that segment parameters may be shared across segments: each segment parameter $\theta_{i,j}$ can take one of a finite, but unknown, number of states. Segments that share the same parameter, possibly across time series, are said to be under the same regime, and regime parameters can still be marginalised

for conjugate probability models (Bolton and Heard, 2018). In this setting, the proposed MCMC algorithm would need to be adapted to sample the changepoint parameters, the number of regime and a regime for each segment following Bolton and Heard (2018).

3.4. Simulation study

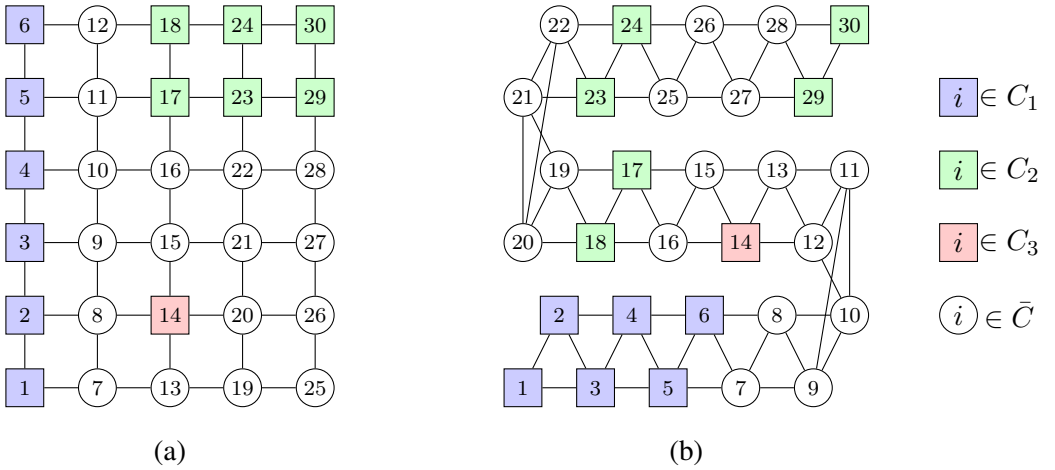
A simulation study is now presented to demonstrate the model for dependent changepoints proposed in Section 3.2 and the sampling strategy discussed in Section 3.3. In particular, various graphs and changepoint parameters are considered to illustrate the flexibility of the proposed model. Moreover, the proposed model is compared with the standard model for independent changepoints, which assumes all edge weight parameters are null (3.7), and with MVCAPA (Fisch, Eckley, and Fearnhead, 2022), which can borrow strength across multiple time series to detect synchronous changepoints across a subset of the time series but assumes *a priori* changepoint locations are exchangeable across time series.

Fisch, Eckley, and Fearnhead, 2022)

3.4.1. Synthetic data

Synthetic data for the simulation study were sampled according to the changepoint model for N time series each of length T defined in (3.3) with, for all i, $f_i(\cdot \mid \theta_{i,j})$ corresponding to Poisson$(\theta_{i,j})$, for a range of values of $N \in \{30, 130, 230, 330\}$ and $T \in \{300, 3\,300, 6\,300, 9\,300\}$.

For simulating data, let $C \subset \{1, \ldots, N\}$ be a non-empty subset of the time series indices, and let $\bar{C}$ be the set complement such that $C \cup \bar{C} = \{1, \ldots, N\}$. Then let $k_i = k$ if $i \in C$, where k is uniformly sampled from $\{1, \ldots, 7\}$, and $k_i = 0$ otherwise, such that C denotes those series which experience changepoints while those in $\bar{C}$ are not impacted by changepoints. The impact of encoding prior information on the dependence structure of changepoints across time series on a graph will be investigated by means of further divisions of the set C into subsets. We fixed $C = C_1 \cup C_2 \cup C_3$ for some disjoint sets $C_1 = \{1, 2, \ldots, \lfloor N/5 \rfloor\}$, $C_2 = \{N, N-1\} \cup \{N - \lfloor N/5 \rfloor, N - \lfloor N/5 \rfloor - 1\} \cup \{N - 2\lfloor N/5 \rfloor, N - 2\lfloor N/5 \rfloor - 1\}$ and $C_3 = \{\lfloor N/2 \rfloor - 1\}$. As illustrated in Figure 3.4 for $N = 30$ time series, according to both the 6×5 lattice graph or the 2-chain graph, the elements of C_1 and C_2 cluster according to the graph, whereas the single element of C_3 has no neighbours in C so that it is isolated in the graph. Moreover, different scenarios for changepoint positions were considered: for all $i \in C$, we set $\tau_{i,j} = \lfloor jT/(k_i + 1) \rfloor + v_{i,j}$, where $v_{i,j}$ is sampled from $\{-v, 0, v\}$, for increasing levels of asynchrony for the changepoints $v \in \{0, 5, 10, 15\}$. For the scenario with $v = 0$, changepoints are synchronous. Finally, for all $i \in C$, the signal strength



(a) (b)

Figure 3.4.: *Graphs of time series indices $i \in \{1, \ldots, 30\}$ representing different dependence structures of changepoints discussed in Section 3.2.2; panels (a) and (b) correspond to a 6×5 lattice and a 2-chain, respectively. For all i, circles indicate $i \in \bar{C}$ and squares indicate $i \in C$; colours blue, green and red indicate that $i \in C_1$, C_2 and C_3, respectively.*

for changepoints may be quantified by

$$\mu_i = \frac{1}{k_i} \sum_{j=1}^{k_i} \log \left\{ E \left(\frac{\mathcal{L}_i(\tau_{j-1}, \tau_j) \mathcal{L}_i(\tau_j, \tau_{j+1})}{\mathcal{L}_i(\tau_{j-1}, \tau_{j+1})} \right) \right\},$$

where $\mathcal{L}_i$ is defined in (3.8). For all $i \in V$ and $j = 1, \ldots, k_i + 1$, let $\theta_{i,j} = 1000$ if j is odd and $\theta_{i,j} = \theta(\mu, k_i)$ if j is even, where $\theta(\mu, k_i)$ is such that $\mu_i = \mu$ for some $\mu > 0$ if $i \in C$. The greater μ, the greater $\theta(\mu, k_i)$. A range of values for μ was considered and Monte Carlo estimations were computed for the corresponding parameters $\theta(\mu, k_i)$. Ten simulations were performed for each combination of the parameters N, T, v and μ.

3.4.2. Changepoint inference

For each simulation, different models were used to infer changepoint estimates from the data: the proposed changepoint model assuming different prior beliefs on the dependence structure of changepoints, and MVCAPA (Fisch, Eckley, and Fearnhead, 2022). MVCAPA detects collective anomalies in multiple time series such that, by means of lags, anomalies are not necessarily aligned, and it relies on the choice of a penalty parameter that controls the level of evidence required to flag a changepoint. We used the implementation from the *R* package *anomaly* to fit MVCAPA, without lags and with lags (fixing the maximal lag to 30), with the default penalty rescaled by a constant ϕ for various $\phi > 0$.

For the proposed changepoint model (3.3), it is assumed that, for all i, $f_i(\cdot \mid \theta_{i,j})$ corresponds

to Poisson$(\theta_{i,j})$ and $\theta_{i,j} \sim \Gamma(100, 0.1)$, given the proposed changepoint prior for a collection of changepoint prior parameters. Different dependent structures for changepoints across time series were used, as illustrated in Figure 3.4: the dependence structure corresponding to the r-chain graph for time series indices, for $r \in \{2, 4, 6, \ldots, N/2\}$, given in Section 3.2.2.2; the dependence structure corresponding to a 6×5 lattice graph, given in Section 3.2.2.1, for scenarios with $N = 30$ time series; and the dependence structure corresponding to a complete graph given in Section 3.2.2.3. Moreover, different changepoint prior parameters $\bar{p} \in \{-60, -70, \ldots, -130\}$ and $\lambda = \lambda_s |\bar{p}|/n$ with $\lambda_s \in \{0, 0.2, \ldots, 0.8\}$, where n denotes the maximum degree of G. Note that models with $\lambda = 0$ correspond to the standard model for independent changepoints (3.7). Furthermore, three different scenarios are considered for the upper bounds $\boldsymbol{w}$ for the lags: for the *fixed window* scenario, we fix $w_i = 30$ for all i; for the *variable window* scenario, it is assumed *a priori* that $w_i \sim$ Geometric(0.9) for all i; for the *zero window* scenario, we fix $w_i = 0$ for all i, which is equivalent to assuming the prior for synchronous dependent changepoints defined in (3.6).

For each combination of changepoint prior parameters, five independent samples of size $2\,000N$ were obtained from the posterior distribution of changepoints, via the MCMC algorithm proposed in Section 3.3.1, with a burn-in of $300N$ iterations. Samples were thinned at a rate of one per N. Note that, when $\lambda > 0$, we set $\delta_0 = 0.5$, $\delta_1 = 1$, $\delta_2 = 30$ for the prior of the parameter δ, so that linked time series indices are expected to bond with probability 0.5 when $\delta > 0$, according to (3.16).

It is of interest to compare the role of graph-based hyperparameters for the proposed model with the role of ϕ for MVCAPA when estimating changepoints. MVCAPA and the proposed model provide changepoint estimates for each time series. To compare inferred changepoints with respect to the changepoints used to simulate data for each time series, both the mean squared error (MSE) for the number of changepoints and the loss ℓ defined in Definition 3, fixing $\gamma = 40$, which takes into account both the number and the positions of changepoints, are considered.

3.4.3. MCMC diagnostics and runtime

The Gelman-Rubin test (Gelman and Rubin, 1992) is used to assess the convergence of the sampler by comparing between-chain and within-chain variances of multiple chains. For each simulation and choice of changepoint prior parameters, the Gelman and Rubin test statistic, which was computed for the five independent chains of the value of the loss function (2.16) obtained via MCMC, was less than 1.2, suggesting that the sampler has converged; see Section A.1.1 for a breakdown of results. Moreover, to illustrate the good mixing properties of the sampler, one simulation for the scenario with $N = 30$, $T = 300$, $k = 1$, $v = 10$ and $\mu = 90$ is considered. Figure 3.5 displays five independent chains of changepoint loss obtained via the proposed MCMC algorithm assuming the 2-chain graph for time series with $\bar{p} = -80$, $\lambda_s = 0.6$; it is apparent that the chains converge to the same changepoints with good mixing properties.

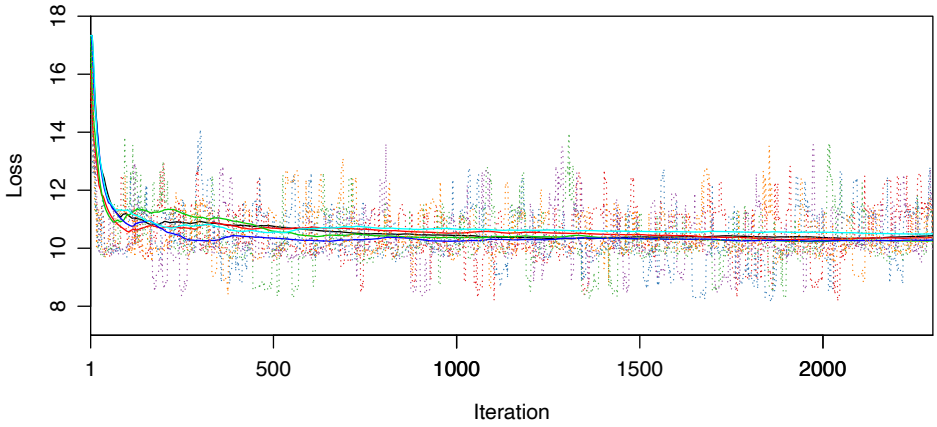


Figure 3.5.: *Five independent chains of the value of the loss function obtained via MCMC (dotted lines), with corresponding average loss function value (solid lines).*

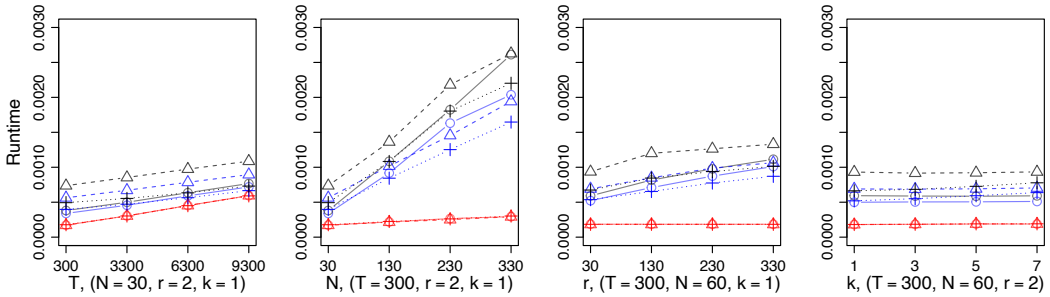


Figure 3.6.: *Average runtime in seconds per iteration for the MCMC algorithm for a collection of scenarios:* $\lambda_s = 0$ *(Red);* $\lambda_s = 0.2$ *(Blue);* $\lambda_s = 0.6$ *(Black); zero window (Cross); fixed window (Circle); variable window (Triangle).*

Figure 3.6 displays the average runtime of an iteration of the MCMC algorithm, implemented in *Python* and run on a 2.6GHz Intel Core i7 processor, for a collection of scenarios assuming an r-graph for the dependence structure of the changepoints. Results support the discussion in Section 3.3.3. The runtime is linear in the number of observations T. When changepoints are dependent with $\lambda > 0$ and $\delta > 0$, the runtime is linear in the number of time series N and in the number of edges in G, which is $2r$ for the r-chain graph. However, when $\lambda = 0$, we update a single time series at each iteration of the algorithm, so that the runtime is independent of N and r. The runtime does not increase with the number of changepoints k. Moreover, the runtime per iteration increases for the extended model with lags, for fixed and unknown windows, since lags must be proposed alongside latent changepoints.

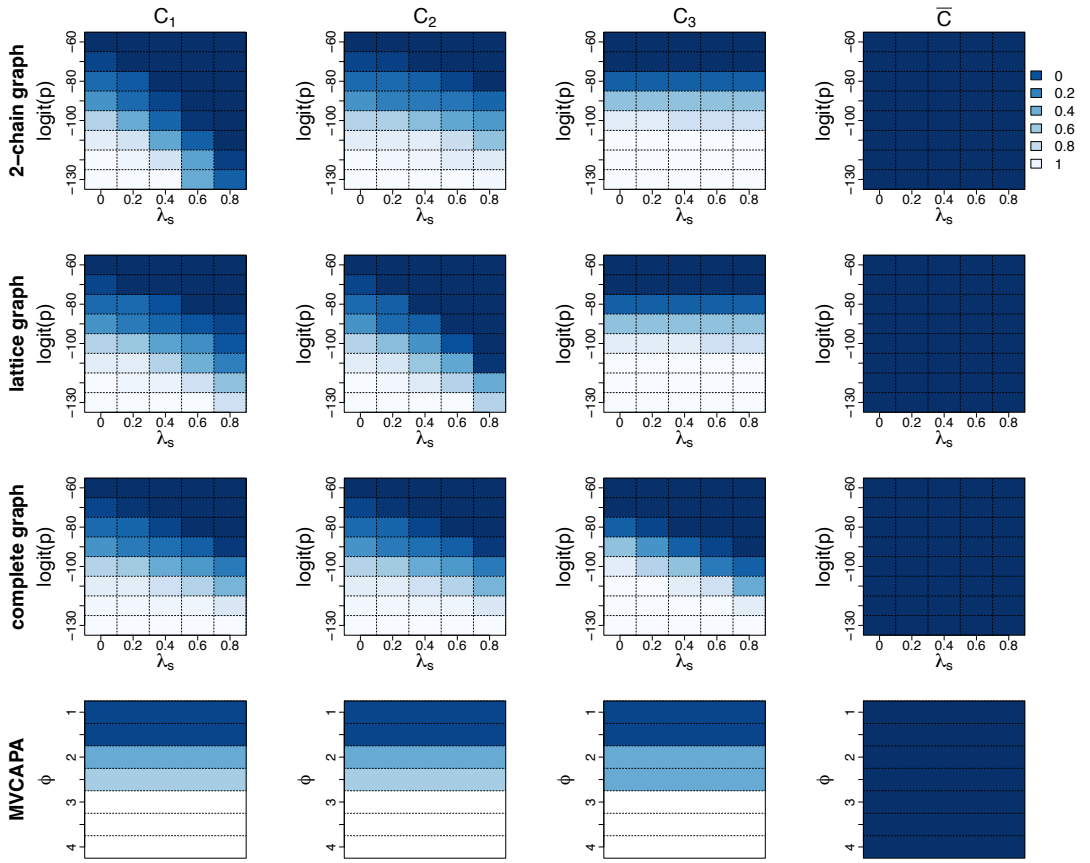


Figure 3.7.: *Mean squared error for the number of changepoints k_i for time series i in clusters $C_1, C_2, C_3, \bar{C}$. Top rows: estimates for the graphical changepoint model assuming the 2-chain graph, the lattice graph and the complete graph, as a function of $\bar{p} = logit(p)$ and λ_s. Bottom row: estimates for MVCAPA as function of ϕ. Results correspond to scenarios with $v = 0$ and $\mu = 90$.*

3.4.4. Changepoint estimation results

This section discusses results of the simulation study for scenarios with $N = 30$ and $T = 300$ to demonstrate the merit of the graphical changepoint prior that induces a changepoint model combining weak signals for changepoints across clusters of time series in G, in comparison with the standard model for independent changepoints and MVCAPA.

3.4.4.1. Detection of clusters of changepoints on a graph of time series

First, we focus on demonstrating the role of the graph G and the changepoint parameters p and λ using simulations corresponding to scenarios with $v = 0$ for synchronous changepoints and with $\mu = 90$. Figure 3.7 displays the MSE for k_i for time series i in clusters C_1, C_2, C_3, for the graphical changepoint models and for MVCAPA as a function of changepoint prior parameters.

For each graphical changepoint model, as $\bar{p}$ decreases, stronger evidence is required to infer changepoints and therefore the MSE for k_i increases for $i \in C_1, C_2, C_3$. The impact of λ depends on the graph-based dependence structure. Consider the 2-chain and the 6×5 lattice graphs for $N = 30$ time series as illustrated in Figure 3.4: for $i \in C_3$, the simulated changepoint is isolated on the graphs, whereas for $i \in C_1, C_2$ the simulated changepoints cluster on the graphs. As a result, as λ increases, the interaction parameter λ has no impact on the MSE for k_i for $i \in C_3$, and the MSE for k_i decreases for $i \in C_1, C_2$ because weaker signals for changepoints, relative to $\bar{p}$, are combined across time series which cluster according to the graphs. The MSE for k_i for $i \in \bar{C}$ is close to 0 for all scenarios, showing that an increase in λ does not lead to changepoint overfitting.

However, both for the graphical changepoint model with the complete graph and for MVCAPA, changepoint locations are assumed to be exchangeable across time series, and therefore the impact of changepoint prior parameters are identical for all time series $i \in C_1, C_2$ and C_3. The simulated changepoints are all connected on the complete graph, and consequently, as λ increases, the MSE for k_i decreases for $i \in C_1, C_2, C_3$ for the graphical changepoint model with a complete graph. For MVCAPA, as the penalty term ϕ increases, the MSE for k_i decreases for time series $i \in C_1, C_2, C_3$.

For further evidence that the posterior distribution adapts to the dependence structure for changepoints specified *a priori* via the graph G, observe that, as λ increases, for the dependence structure induced by the 2-chain, the MSE for k_i is lower for $i \in C_1$ than for $i \in C_2$, since time series $i \in C_1$ have a greater proportion of neighbour time series impacted by changepoints than time series $i \in C_2$, as illustrated in Figure 3.4. For the dependence structure induced by the lattice, however, time series $i \in C_1$ have a lower proportion of neighbour time series impacted by changepoints than time series $i \in C_2$, so that the MSE for k_i is greater for $i \in C_1$ than for $i \in C_2$. Moreover, as λ increases, the MSE for k_i for $i \in C_1$ is lower for the 2-chain graph than for the complete graph, whereas the MSE for k_i for $i \in C_2$ is greater for the 2-chain graph than for the complete graph. This follows because the proportion of neighbour time series impacted by changepoints is lower for the complete graph than the 2-chain graph for time series $i \in C_1$, but larger for time series $i \in C_2$.

Section A.1.2 demonstrates that the roles of the graph G and the changepoint parameters p and λ are similar for other signal strengths $\mu \in \{55, 70, 105, 120\}$ and for scenarios with lagged changepoints, and it also presents results in terms of the loss ℓ defined in Definition 3.

3.4.4.2. Detection of quasi-synchronous changepoints

This section discusses results for scenarios with increasing levels of asynchrony for changepoints $v \in \{0, 5, 10, 15\}$ and with $\mu = 90$. Recall changepoints are synchronous for scenarios with $v = 0$.

Figure 3.8 displays the MSE for k_i for $i \in V$ for the graphical changepoint model assuming different assumptions for the upper bounds for the lags, different graphs G and edge weight parameters λ, and $\bar{p} = -90$. As the level of asynchrony of the changepoints v increases, the MSE for k_i tends to be greater for the *zero window* scenario than for the *fixed window* or the *variable*

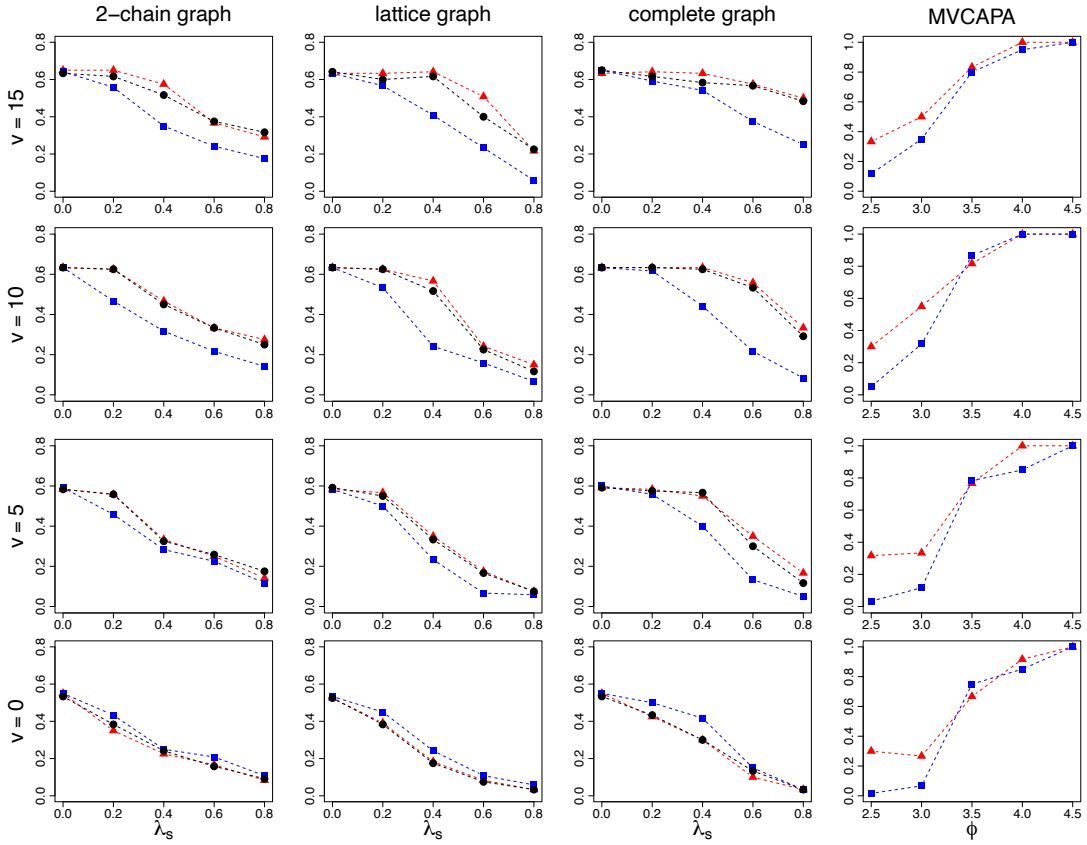


Figure 3.8.: *Mean squared error for k_i for time series $i \in V$ for increasing levels of changepoint asynchrony v. Left columns: graphical changepoint models for different assumptions for the upper bounds for the lags* - fixed window *(blue squares),* variable window *(black circles), and* zero window *(red triangles). Right column: MVCAPA with lags (blue squares) and without lags (red triangles).*

window scenarios. Hence, there is merit in relaxing the assumption that signals for changepoints must be synchronous. In particular, the MSE for k_i tends be greater for the *variable window* than for the *fixed window* scenarios, therefore encouraging practioners to specify fixed time windows when possible. Moreover, as the interaction parameter λ increases, the decrease in MSE tends to be greater for the 2-chain graph and the lattice graph than for the complete graph. This follows because for time series $i \in C$ the proportion of neighbour time series impacted by changepoints is greater for the lattice graph and the 2-chain graph than for the complete graph. Figure 3.8 also displays results for MVCAPA with and without lags. As the level of asynchrony for changepoints v increases, the MSE tends to increase for MVCAPA without lags. Results also show that, in contrast with the graphical changepoint models, specifying fixed lags when simulated changepoints are synchronous may adversely affect the performance of MVCAPA.

Section A.1.3 shows the role of lags is similar when assuming different values of $\bar{p}$. Moreover, it presents results in terms of the expected loss ℓ, showing the model extensions have merit both in

terms of detecting asynchronous signals for changepoints and correctly estimating the positions of these signals.

3.4.4.3. Importance of auxiliary variables

We demonstrate that it is pertinent to use auxiliary variables to sample from the posterior distribution of dependent changepoints, as discussed in Section 3.3. For the simulated data discussed in Section 3.4.4.1, which correspond to scenarios with $v = 0$ and $\mu = 90$, changepoints were sampled via MCMC as described in Section 3.4.2, but now with $\delta = 0$, meaning the parameter space was not augmented with auxiliary variables (3.17). Figure 3.9 compares the MSE for k_i from samples obtained with and without auxiliary variables, for $i \in C_1, C_2$, and for the 2-chain and the lattice graph based dependence structures for changepoints. Without auxiliary variables, the MCMC algorithm fails to explore the state space of changepoints so that the number of changepoints tend to be underestimated; in other words, clusters of weak signals for changepoints across time series tend to be overlooked. The difference in performance, in favour of the MCMC algorithm making

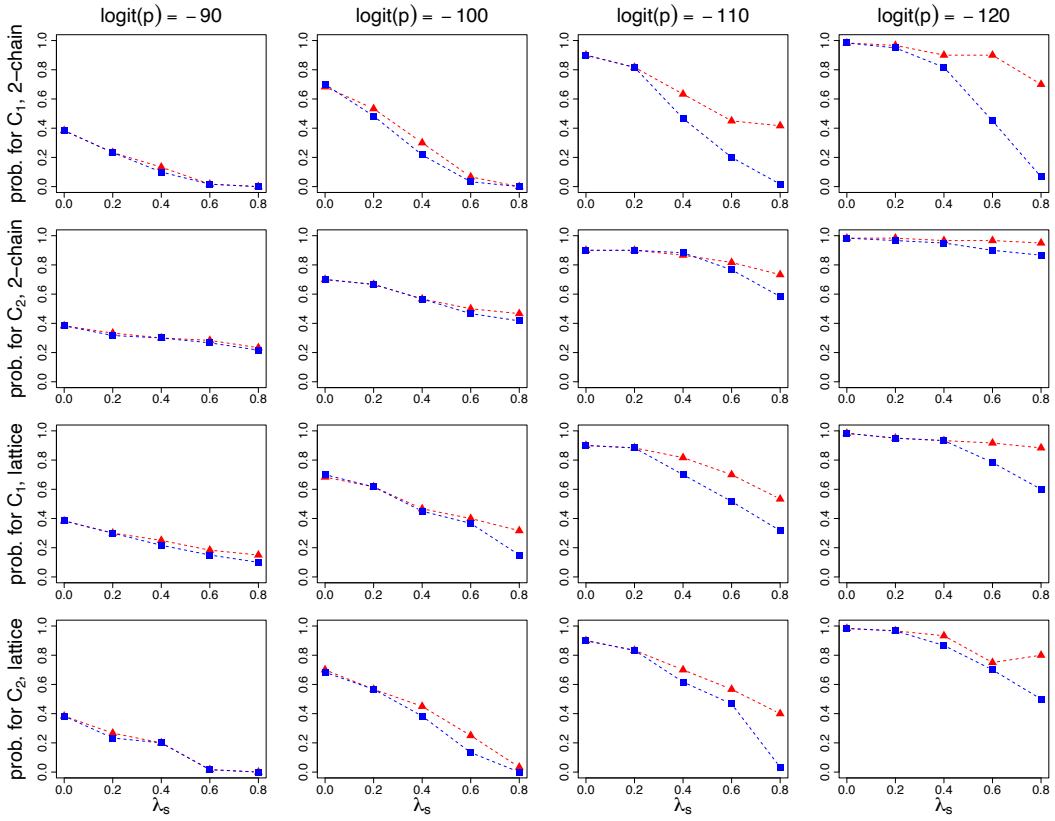


Figure 3.9.: *Mean squared error for k_i for time series $i \in C_1, C_2$, obtained via simulations with auxiliary variables (blue squares) and without auxiliary variables (red triangles), for the 2-chain and the lattice graph based dependence structures, and for a collection of changepoint prior parameters p and λ_s.*

use of auxiliary variables, is particularly important when interaction parameters are large and weak signals for changepoints correspond to time series whose indices induce subgraphs with a large number of edges, as for $i \in C_1$ given the 2-chain dependence structure in Figure 3.9.

3.5. Red team detection in network authentication data from LANL

This section presents results of an analysis of the LANL network authentication data presented in Section 3.1 that demonstrates the utility of the graphical changepoint model proposed in Section 3.2.

3.5.1. Presence of a red team

The occurrence of a red team exercise during the first month of the data collection provides surrogate intruder behaviour in the authentication data (Kent, 2015). In particular, 103 user IDs are known to have been used by the red team. We show the graphical model for dependent changepoints can combine evidence from multiple users that are linked in the network, to detect chains of quasi-synchronous weak signals for changes in the authentication activity of red team users, whilst limiting the number of false alerts.

For our demonstration purposes, it suffices to examine a subset of the full LANL network of users, which is represented by the graph $G = (V, E)$ defined in Section 3.1.2. Let R denote the set of red team users and let $B \subseteq \{u \in V \setminus R : \exists r \in R \text{ s.t. } (u, r) \in E\}$ denote $|R|$ randomly selected users that are not labelled as red team users in the data but are linked to red team users on the network. The focus is on the network corresponding to the subgraph G' induced in G by the set of users $V' = R \cup B$. Figure 3.10 shows the degree distribution of the 206 users in G'. Red team users tend to have a greater degree in G' than legitimate users; to traverse the network towards high value targets, intruders tend to take control of users that are highly linked on the network.

3.5.2. Changepoint modelling

Recall from Section 3.1 that, for each user $i \in V'$, the data $x_{i,0}, \ldots, x_{i,T}$ consist of hourly counts of network logons per source computer for the first month of data collection as defined in (1.1), which are now assumed to follow the model specified in (2.4) for multinomial data. Different graphical changepoint priors are considered to demonstrate the benefits of encoding prior beliefs about cyber-attacks. To encode prior belief that signals for changes resulting from an attack are likely to occur at similar times across users that are linked in the network G', the graphical changepoint prior specified in (3.15) is considered with an identical edge weight parameter $\lambda > 0$, as defined in Section 3.2.1.2, for all pairs of time series corresponding to users that are linked in G'. Moreover, for comparison purposes, the graphical changepoint prior (3.15) is also considered assuming the complete graph defined in Section 3.2.2.3 such that $\lambda_{i,j} = \lambda > 0$ for all pairs of users

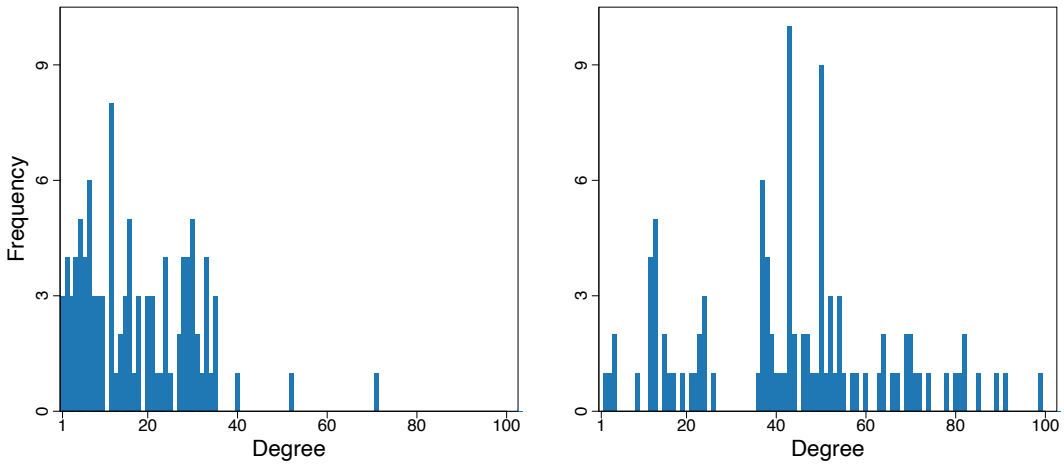


Figure 3.10.: *Degree distribution of users in the LANL network represented by the graph G'. Left panel: counts for legitimate users. Right panel: counts for red team users.*

$(i, j) \in V' \times V'$. With $\lambda = 0$, the two graphical changepoint priors of interest correspond to the standard changepoint model assuming independence of changepoints across time series (3.7).

For comparison purposes and to illustrate the flexibility of the proposed model, a collection of changepoint prior parameters are considered: $\bar{p} \in \{-30, -50, -70\}$ and $\lambda = \lambda_s |\bar{p}|/n$ with $\lambda_s \in \{0, 0.5, 0.6, 0.7, 0.8\}$, where n denotes the average node degree in the graph. Moreover, different assumptions for the upper bounds $\boldsymbol{w}$ for the lags are compared: the *zero window* assumption with $w_i = 0$ for all i, implying signals for attacks are assumed to be synchronous across users; and, the *variable window* assumption with $w_i \sim \text{Geometric}(0.9)$ for all i, admitting signals for attacks may be asynchronous across users.

For each simulation, a sample of size $1\,000\,000$ was obtained from the posterior distribution of the changepoints via the MCMC algorithm proposed in Section 3.3.2, with a burn-in of $300\,000$ iterations; the Bayes estimate for changepoints corresponding to the loss function (2.16) with $\gamma = 48$ was then derived from the sample.

3.5.3. Results

For each user in the network, each estimated changepoint represents a piece of evidence for possible malicious behaviour that might require further investigation by cyber analysts. Identifying inferred changepoints for time series corresponding to legitimate users $i \in B$ as false alerts, it is meaningful to compare models in terms of the estimated number of changepoints per time series $i \in V'$,

$$\bar{k} = \frac{1}{|V'|} \sum_{i \in V'} k_i,$$

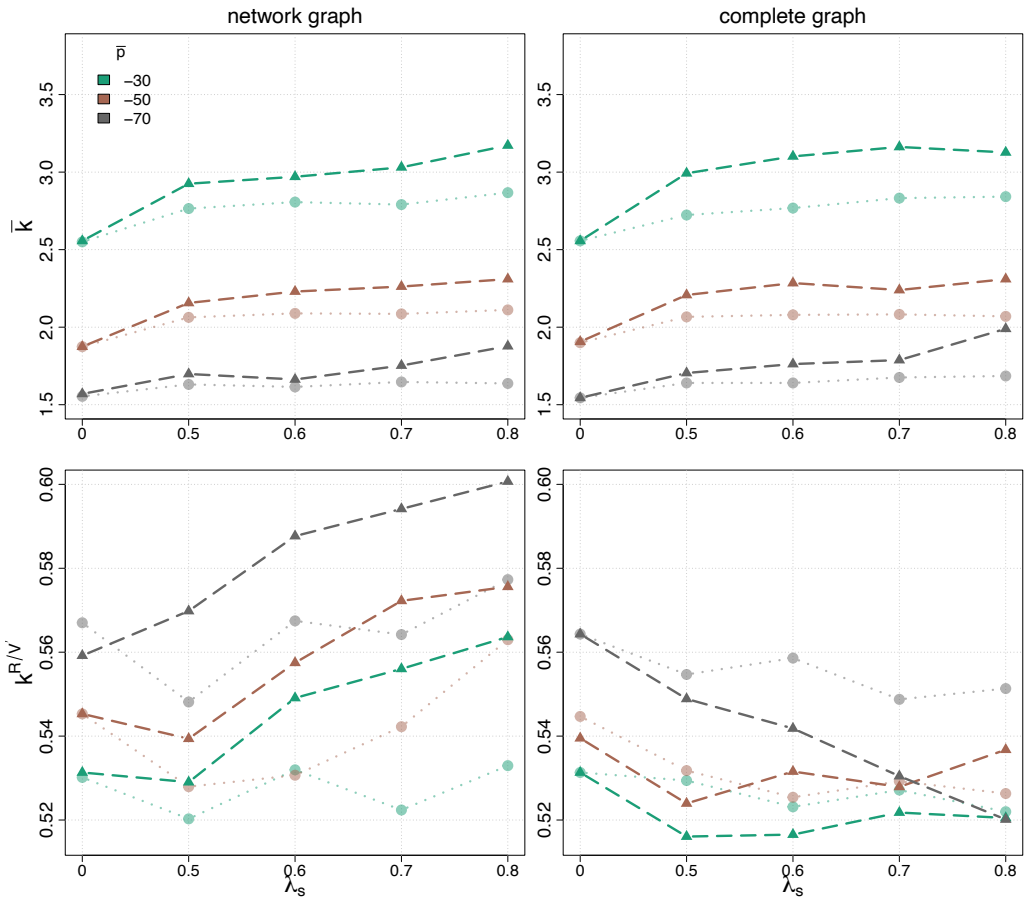


Figure 3.11.: *Average number of changepoints per time series $\bar{k}$ (top row), and proportion of changepoints impacting redteam users $k^{R/V'}$ (bottom row), assuming the network graph (left column) and the complete graph (right column), for different assumptions for the upper bounds for the lags - zero window (circles) and variable window (triangles), and for a collection of prior parameters $\bar{p}$ (identified by distinct colours) and λ_s.*

and the proportion of estimated changepoints that impact redteam users,

$$k^{R/V'} = \frac{\sum_{i \in R} k_i}{\sum_{i \in V'} k_i}.$$

Moreover, since cyber-attacks tend to be identified through clusters of behavioural changes across machines that are linked on the network, it is of interest to prioritise for investigation the estimated changepoints that belong to clusters of quasi-synchronous changepoints on the network. Given some time window $\varpi \geqslant 0$, let the weight of $\tau_{i,j}$ be

$$c_{i,j}^{\varpi} = \frac{n_{i,j}^{\varpi} + 1}{n_i + 1},$$

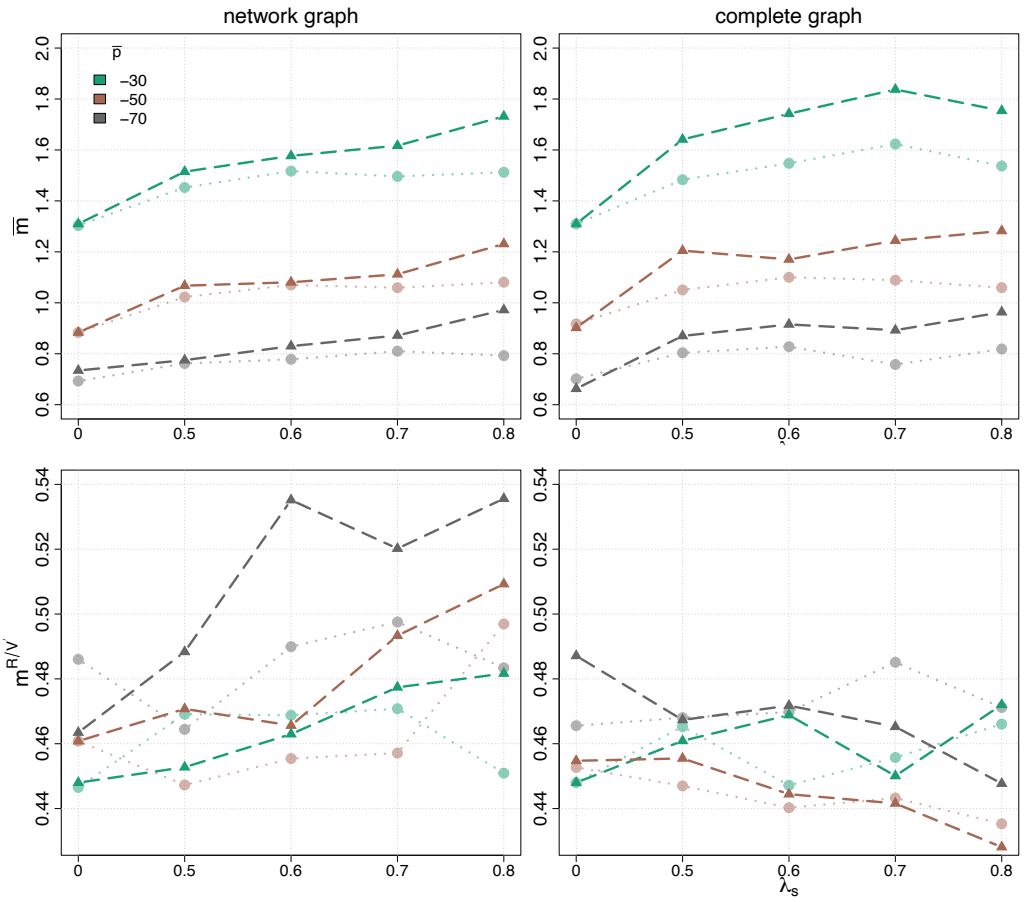


Figure 3.12.: *Estimated $\bar{m}$ and $m^{R/V'}$ (3.5.3), assuming the network graph (left column) and the complete graph (right column), for different assumptions for the upper bounds for the lags - zero window (circles) and variable window (triangles), and for a collection of prior parameters $\bar{p}$ (identified by distinct colours) and λ_s.*

where

$$n^{\varpi}_{i,j} = |\{(i,i') \in E : \exists j' \in \{1,\ldots,k_{i'}\} \text{ s.t. } |\tau_{i,j} - \tau_{i',j'}| \leqslant \varpi\}|$$

denote the number of users linked to user i in G' that are impacted by a changepoint within ϖ hours of $\tau_{i,j}$, and where $n_i = |\{(i,i') \in E\}|$ is the degree of node i in G', such that $(n_i+1)^{-1} \leqslant c_{i,j} \leqslant 1$. The larger the weight $c_{i,j}$, the more connected $\tau_{i,j}$ to other changepoints across the network. To take into account both the number of changepoints and their connectedness, for each user i, changepoint estimates are also compared via the sum of weights

$$m_i \equiv m^{\varpi}_i = \sum_{j=1}^{k_i} c^{\varpi}_{i,j},$$

such that $0 \leqslant m_i \leqslant k_i$. Note that for each user i, m_i increases with both the number of changepoints and their weights. Let

$$\bar{m} = \frac{1}{|V'|} \sum_{i \in V'} m_i, \quad m^{R/V'} = \frac{\sum_{i \in R} m_i}{\sum_{i \in V'} m_i}$$

be the average sum of changepoint weights per user and the proportion of changepoint weights associated to redteam users, respectively.

For each choice of graph and changepoint prior parameters, Figure 3.11 displays the estimated values of $\bar{k}$ and $k^{R/V'}$, and Figure 3.12 displays the estimated values of $\bar{m}$ and $m^{R/V'}$ assuming $\varpi = 48$. As $\bar{p}$ increases, weaker evidence is required to infer changepoints, and therefore, for each graph, $\bar{k}$ and $\bar{m}$ increase. As λ increases, the estimates for $\bar{k}$ and $\bar{m}$ tend to increase for each graph, but the estimates for $k^{R/V'}$ and $m^{R/V'}$ tend to increase only when assuming the network graph. This follows because the graphical changepoint model assuming the network graph successfully encodes prior knowledge that cyber attacks tend to correspond to coordinated activity across multiple users linked by network connectivity, and consequently, as λ increases, it detects weaker signals for behavioural changes that correspond to red team activity, whilst crucially limiting the number of false alerts. For the proposed model with the complete graph, all time series are connected, and therefore, as λ increases, weaker signals for changes are detected for red team activity but also for legitimate activity, which would impede fast identification of the attack.

Moreover, results show the benefits of the model extension which relaxes the assumption that signals for attacks are synchronous across users. As λ increases, when assuming the network graph, the increase of the estimates for $k^{R/V'}$ and $m^{R/V'}$ tend to be greater for the *variable window* scenario than for the *zero window* scenario. In contrast with the *zero window* scenario, the *variable window* scenario admits attacks may result in quasi-synchronous behavioural changes across the network, and consequently clusters of nearby but not necessarily synchronous weak signals for changepoints are detected across red team users.

The results show that, in comparison with the standard model for independent changepoints across time series, the proposed graphical changepoint model provides a flexible tool for cyber-analysts to incorporate expert knowledge in changepoint analysis for network monitoring, thereby facilitating network intrusion detection.

3.6. Conclusion

This chapter considered settings with N time series (3.1) subject to changepoints, where it is desirable to encode in the changepoint prior, by means of a graph $G = (V, E)$ on N nodes corresponding to each of the time series, that pairs of time series $(i, i') \in E$ are *a priori* more likely to be impacted by simultaneous changepoints. A novel informative prior for changepoints was proposed, which encodes the information contained in the graph G, inducing a changepoint model

for multiple time series that borrows strength across clusters of connected time series to detect weak signals for synchronous changepoints. The graphical model for changepoints was further extended to allow dependence between nearby but not necessarily synchronous changepoints across neighbouring time series in the graph. A novel reversible jump MCMC algorithm making use of auxiliary variables was proposed to sample from the graphical changepoint model.

The novel informative prior for changepoints was motivated by the commonly held belief of security experts that cyber attacks tend to correspond to coordinated sequences of events across multiple connected endpoints on the network (Sexton, Storlie, and Neil, 2015). For demonstration purposes, this chapter discussed an analysis of real computer network authentication data from LANL presented in Section 1.1.1: it was of interest to monitor for temporal changes in the authentication activity of network users to detect the presence of red team actors. The proposed changepoint prior was used to encode beliefs that signals for network intrusions are *a priori* likely to occur at nearby times for users historically linked by previous network connectivity. We showed that, as a consequence, the proposed model can detect weak signals for red team activity in the network, whilst limiting the number of false alerts, in contrast with a standard model assuming independence of behavioural changes across users.

The setting of interest in this chapter may be relevant to other domains of applications. For some applications, it might be restrictive to assume that prior beliefs on which time series are likely to be impacted by simultaneous changepoints do not vary over time. For such settings, it would be possible to specify a time series of graphs $\{G_t = (V, E_t) \mid E_t \subseteq V \times V, t \geqslant 0\}$, such that pairs of time series $(i, i') \in E_t$ are *a priori* more likely to be impacted by simultaneous changepoints at time t. A further model extension could consider relaxing the assumption V is fixed, specifying a time series of graphs $\{G_t = (V_t, E_t) \mid E_t \subseteq V_t \times V_t, V_t \subset \mathbb{N}, t \geqslant 0\}$. With the introduction of time-dependent edge weight parameters $\boldsymbol{\lambda} = (\lambda_{i,i',t})$ such that $\lambda_{i,i',t} > 0$ if and only if $(i, i') \in E_t$, it would present no theoretical complication to adapt the graphical changepoint prior and the proposed sampling strategy.

In the next chapter, we revert back to a setting with a single time series to focus the discussion on a novel segment model, that is motivated by temporal dynamics observed cyber data.

4

Non-exchangeable data

As discussed in Chapter 2, standard changepoint models for single time series rely on partitioning the passage of time into segments, and fitting relatively simple models within each segment. In particular, the data within each segment are often assumed to be independent and identically distributed conditional on some segment-specific parameter, as specified in (2.4). This construction assumes the data within each segment are exchangeable, rendering the order in which the data are observed irrelevant when calculating their joint likelihood (Bernardo and Smith, 1993).

Definition 4 (Infinite exchangeability)**.** A potentially infinite sequence $x_1, x_2, \ldots$ is said to be infinitely exchangeable under a probability measure F if, for every n and any set of indices $I_n = \{i_1, \ldots, i_n\}$, we have

$$F(x_{i_1}, \ldots, x_{i_n}) = F(x_{\sigma(i_1)}, \ldots, x_{\sigma(i_n)}),$$

for any permutation σ on I_n.

There are many examples of applications where this assumption is suitable; see, for example, Olshen et al. (2004), Fryzlewicz (2014) and Fearnhead and Rigaill (2019) for the detection of changes in the mean or variance of time series. However, for some applications, such as cyber-security, it can be reductive to assume data are exchangeable within segments. As stated in Chapter 1, to detect the presence of a network intrusion, it might be informative to monitor for changes in the volumes of different types of traffic passing through a network over time. Yet, cyber data are often subject to population drifts, seasonal variations and other temporal trends that are unlikely to be evidence for cyber-attacks. As a result, traditional changepoint detection methods, which assume data are exchangeable within segments, will fail to capture temporal dynamics and consequently fit many more changepoints than necessary. For example, consider the network event data recorded at

LANL, which were presented in Section 1.1.2. It is of interest to detect temporal changes in the distribution of the data $x_{1:T} = (x_1, \ldots, x_T)$ where, for all t, x_t denotes the number of network events that are in progress across the network during the t-th second between 10:00 and 10:20 on day 22 of the data collection period. Suppose the data $x_{1:T}$ follow the standard parametric changepoint model assuming data are iid within each segment (2.4), where $f(\cdot \mid \theta_j)$ denotes the density function of the negative binomial distribution with unknown segment specific success probability $\theta_j \sim \text{Beta}(\alpha, \beta)$, for some $\alpha > 0$ and $\beta > 0$, and some fixed number of failures $r > 0$. A sample from the posterior distribution of changepoints is obtained via the standard reversible jump MCMC algorithm discussed in Section 2.3.3. Figure 4.1 displays the network data $x_{1:T}$ and the MAP changepoints are indicated by vertical lines. No labels are available to indicate the optimal segmentation for these data, but we argue that more changepoints are fitted than would be preferable for network monitoring. It is desirable for abrupt changes, such as the ones observed near the 420th and 880th seconds, to be detected because they correspond to behavioural changes of the network that are sufficiently drastic that cyber-analysts may want to investigate whether they correspond to malicious activities. However, changes due to small fluctuations, such as the ones between the 100th and 380th seconds, may not be relevant to cyber-analysts because they result from local temporal dynamics in traffic volumes that are not suggestive of behavioural changes of the network. There are various possible causes for such small fluctuations, including data transfer mechanisms in networks. For example, when large data are transferred across a network, they are often divided up into batches following protocols that involve routing and queueing steps, which may lead to temporal correlations in traffic volumes. Hence, a changepoint model to detect clear discontinuities in the presence of non-exchangeable data is needed. Moreover, since temporal dynamics for cyber data may change when a clear discontinuity occurs, it would not be satisfactory to assume the dependence structure of the data is the same for each segment. For example, an attack may involve reconnaissance techniques such as port scanning that could change temporal dynamics in traffic volumes.

Existing models to detect clear discontinuities in the presence of non-exchangeable data typically assume the dependence structure is identical for each segment (Albert and Chib, 1993; Sparks, Keighley, and Muscatello, 2011; Chakar et al., 2017; Romano et al., 2021; Cho and Fryzlewicz, 2021). In particular, it is often assumed the data within each segment are Markov conditional on some segment parameter. Moreover, changepoint models for dependent data are often designed for a specific marginal distribution, for example the normal distribution (Chakar et al., 2017; Romano et al., 2021; Cho and Fryzlewicz, 2021), the negative binomial distribution (Sparks, Keighley, and Muscatello, 2011; Yu, Baron, and Choudhary, 2013) or the Poisson distribution (Weiß, 2011; Franke, Kirch, and Kamgaing, 2012).

This chapter extends a standard changepoint model (Fearnhead, 2006), relaxing the assumption that data are exchangeable within segments. The proposed changepoint model, named the

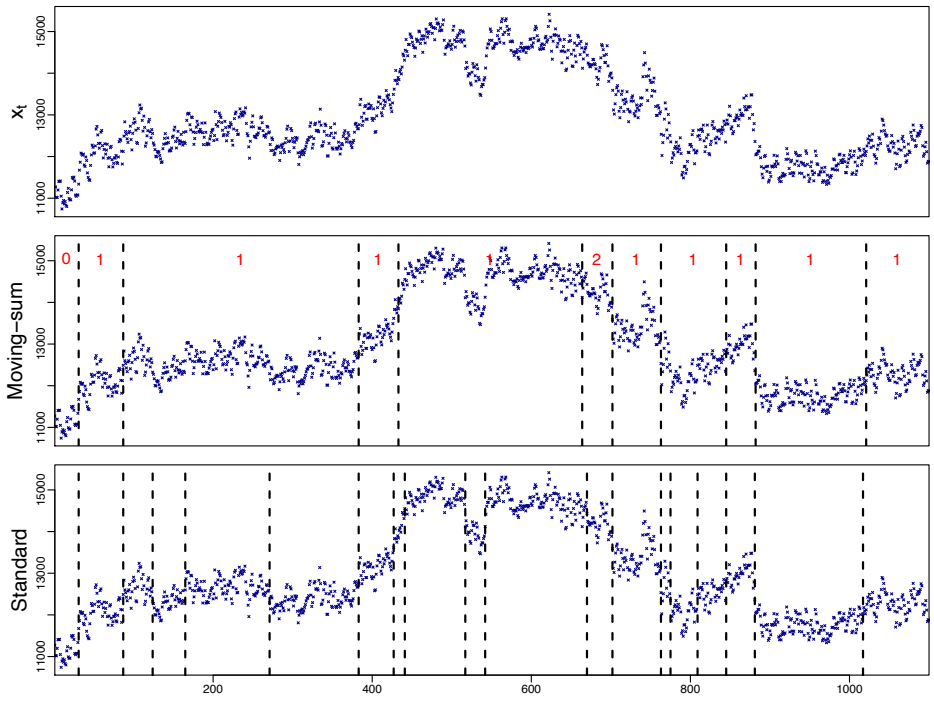


Figure 4.1.: *Counts of computer network events each second between 10:00 and 10:20 on day 22 of the data collection period. Vertical lines indicate estimated changepoints for the proposed changepoint model (middle panel) and for the standard changepoint model (Fearnhead, 2006) (bottom panel). Numbers in red indicate the MAP order of dependence m for each segment for the moving-sum model.*

moving-sum changepoint model, supposes a segment model that is related to a model for m-dependent, stationary data discussed in Joe (1996); a sequence $x_1, x_2, \ldots$ is m-dependent if $(x_{t+m+1}, x_{t+m+2}, \ldots)$ is unconditionally independent of $(x_1, x_2, \ldots, x_t)$ for all $t \geqslant 1$. Within each segment, our model assumes the data are m-dependent and identically distributed conditional on some parameter θ, where both θ and $m \geqslant 0$ are unknown and change from one segment to the next. Whilst θ denotes a parameter of the marginal distribution of the data such as the mean or the variance, m corresponds to the level of dependency within the segment. The dependence structure is built upon moving sums over independent and identically distributed latent variables. Note that the proposed moving sum changepoint model is not related to the MOSUM estimation procedure in Eichinger and Kirch (2018), which focuses on statistical properties of changepoint estimators based on moving sum statistics. To maintain tractability, the marginal distribution of the observed data is assumed to belong to the class of convolution-closed infinitely divisible distributions, which includes, for example, the normal, negative binomial and Poisson distributions; see Definition 6. Therefore, the moving-sum changepoint model is suitable for various settings where it is of interest to detect clear discontinuities in the presence of non-exchangeable data. For example,

consider the MAP changepoints obtained with the moving-sum model for the counts of network events displayed in the middle panel of Figure 4.1. In comparison with the standard changepoint model, the moving-sum changepoint model captures temporal dynamics of the network behaviour, resulting in a segmentation of the data that is more adapted to network monitoring.

As discussed in Section 2.3, a standard approach to sampling changepoints for a time series is that of Green (1995), using a reversible jump MCMC algorithm to explore the state space of changepoints. At each iteration of the algorithm, one of the following move types is proposed: sample a segment parameter, propose a new changepoint, or delete or shift an existing changepoint to a new position. This chapter proposes a sampling strategy within that framework to sample from the moving-sum changepoint model. In particular, our approach exploits an analysis of the constraints of the parameter space; for example, for non-negative data such as count data, the constraints of the parameter space depend on the observed data, and this must be understood to build proposals for segment specific dependency levels.

The remainder of the chapter is organised as follows. Section 4.1 introduces a novel changepoint model for non-exchangeable data. Section 4.2 gives an approach for deriving the likelihood of the data conditional on proposed changepoints, characterising the segment model in terms of a stochastic difference equation with an unknown initial condition. Section 4.3 provides a detailed analysis of the constraints to the parameter space, along with asymptotic results on the behaviour of segment parameters. A reversible jump MCMC sampling strategy is given in Section 4.4. Section 4.5 presents results demonstrating the benefits of the proposed changepoint model, via a comparison with the standard model (Fearnhead, 2006), PELT (Killick, Fearnhead, and Eckley, 2012), DeCAFS (Romano et al., 2021) and WCMgSA (Cho and Fryzlewicz, 2021). Section 4.6 considers two applications of changepoint detection showing the benefits of the proposed changepoint model: computer network monitoring via change detection in count data, and detection of breaks in daily prices of a stock.

4.1. Moving-sum model

This section introduces the moving-sum model that is used as a segment model to define a novel changepoint model for non-exchangeable data.

4.1.1. Model definition

A moving-sum model assumes that observed data $x_1, \ldots, x_n$ satisfy

$$x_t = \sum_{i=0}^{m} y_{t-i}, \tag{4.1}$$

for $t = 1, \ldots, n$, where $y_{-(m-1)}, \ldots, y_n$ are $m + n$ iid latent random variables with common parametric density $f_m(\cdot \mid \theta)$ for some unknown parameters $\theta \in \Theta$ and $m \geqslant 0$. If $m = 0$, the construction in (4.1) implies that, for all $t = 1, \ldots, n$,

$$x_t \overset{\text{iid}}{\sim} f_0(\cdot \mid \theta),$$

and, consequently, is equivalent to exchangeability in the data. Yet, if $m > 0$, the sequence of observed data (4.1) is m-dependent and therefore non-exchangeable.

Definition 5 (m-dependence)**.** For $m \geqslant 0$, the sequence $x_1, x_2, \ldots$ is m-dependent if $(x_{t+m+1}, x_{t+m+2}, \ldots)$ is unconditionally independent of $(x_1, x_2, \ldots, x_t)$ for all $t \geqslant 1$. Note that if a sequence is m-dependent, then it is also Markov of order m.

For all t, x_t in (4.1) is the sum of $m + 1$ latent random variables, leading to m-dependence. Noting this duality, for simplicity of presentation in the following discussion we use the notational convention

$$\bar{m} = m + 1.$$

It will be helpful to identify a class of distributions for which the construction in (4.1) gives rise to a tractable marginal distribution of the observed data. Recall the distribution of a random variable x is *infinitely divisible* if, for all $m \geqslant 0$, there exists a sequence of iid random variables $y_0, \ldots, y_m$ such that $\sum_{i=0}^{m} y_i$ has the same distribution as x. For all infinitely divisible marginal distributions F for x_t (4.1), there exists a distribution F_m for the latent random variables for all m, and F_m is known if F is closed under convolution. In this chapter, it will be assumed that the marginal distribution of x_t is an infinitely divisible distribution that is closed under convolution, so that the corresponding density $f_m(\cdot \mid \theta)$ of the iid latent variables is available for all m.

Definition 6 (Infinitely divisible distribution)**.** The distribution of x is infinitely divisible if, for all $m \geqslant 0$, there exists a sequence of iid random variables $y_0, \ldots, y_m$ such that $\sum_{i=0}^{m} y_i$ has the same distribution as x.

We consider in detail three instances of the moving-sum segment model based on such distributions, one for continuous data with unbounded domain given in Example 4.1, one for continuous data with bounded domain given in Example 4.2, and one for discrete data with bounded domain given in Example 4.3, which we will refer back to throughout the chapter for illustration. Conjugate priors for θ are given for each example.

Example 4.1 (Normal marginal distribution)**.** Suppose that $f_m(\cdot \mid \theta)$ corresponds to density of the normal distribution with mean $\mu/\bar{m}$ and variance $\sigma^2/\bar{m}$, for some $\theta = (\mu, \sigma)$ where $\mu \in \mathbb{R}$ and $\sigma > 0$. It follows that (x_t) is marginally $N(\mu, \sigma^2)$ with m-dependence. Moreover, *a priori* $\sigma^{-2} \sim \text{Gamma}(\alpha, \beta)$, for some $\alpha > 0$ and $\beta > 0$, and $\mu \sim N(\mu_0, \sigma^2/\lambda)$ for some $\mu_0 \in \mathbb{R}$ and $\lambda > 0$.

Example 4.2 (Gamma marginal distribution)**.** Suppose that $f_m(\cdot \mid \theta)$ corresponds to density of the gamma distribution with shape parameter $\lambda/\bar{m}$ and rate θ, where $\lambda > 0$ and $\theta > 0$. It follows that (x_t) is marginally $\Gamma(\lambda, \theta)$ with m-dependence. The prior for θ is assumed to be $\Gamma(\alpha, \beta)$ for some $\alpha > 0$ and $\beta > 0$.

Example 4.3 (Negative binomial marginal distribution)**.** Suppose that $f_m(\cdot \mid \theta)$ corresponds to density of the negative binomial distribution with number of failures $r/\bar{m}$ and success probability $\theta \in [0, 1]$, for some fixed $r > 0$. It follows that (x_t) is marginally $NB(r, \theta)$ with m-dependence. Moreover, *a priori* $\theta \sim \text{Beta}(\alpha, \beta)$, for some $\alpha > 0$ and $\beta > 0$.

Other examples of such distributions include the Poisson, Cauchy and chi-squared distributions.

4.1.2. Bayesian changepoint analysis with moving-sums

Suppose we observe real-valued discrete time data $x_{1:T} = (x_1, \ldots, x_T)$. The parametric changepoint model defined in Section 2.1 is assumed, such that $k \geqslant 0$ changepoints with ordered positions $\tau_{1:k} = (\tau_1, \ldots, \tau_k) \in \mathcal{T}_k$ partition the passage of time into $k + 1$ independent segments. Changepoints are *a priori* assumed to follow a Bernoulli process as specified in (2.2).

For the moving-sum changepoint model, within each segment j, the data $x_{\tau_{j-1}}, \ldots, x_{\tau_j - 1}$ are assumed to follow the moving-sum model (4.1) conditional on some unknown dependency level $m_j \geqslant 0$ and parameter $\theta_j \in \Theta$, which both change from one segment to the next. Dependency levels $m_1, \ldots, m_{k+1}$ and segment parameters $\theta_1, \ldots, \theta_{k+1}$ are assumed to be independent. For all j, it is assumed *a priori* that m_j is drawn from a geometric distribution with parameter $0 < \rho < 1$, so that

$$\pi(m_j) = \rho(1 - \rho)^{m_j},$$

meaning that for each segment the order of dependence may be increased or decreased at a fixed cost. Moreover, motivated by computational considerations, the prior for θ_j is chosen to be conjugate for f_0. For notational simplicity, we denote by π the density of the prior distribution of both m_j and θ_j.

4.1.3. Simulation from the model

Figure 4.2 displays data generated from the moving-sum changepoint model, given a fixed sequence of changepoints, for negative binomial data given in Example 4.3 with $r = 200$, $\alpha = 20$ and $\beta = 10$. It is apparent that the changepoints correspond to changes in both the dependence structure and the mean of the data. In particular, we note that for larger values of m_j, the data tend to be smoother in the corresponding segment. For segments with $m_j > 0$, it is reductive to judge the data to be exchangeable since there are clear temporal dynamics.

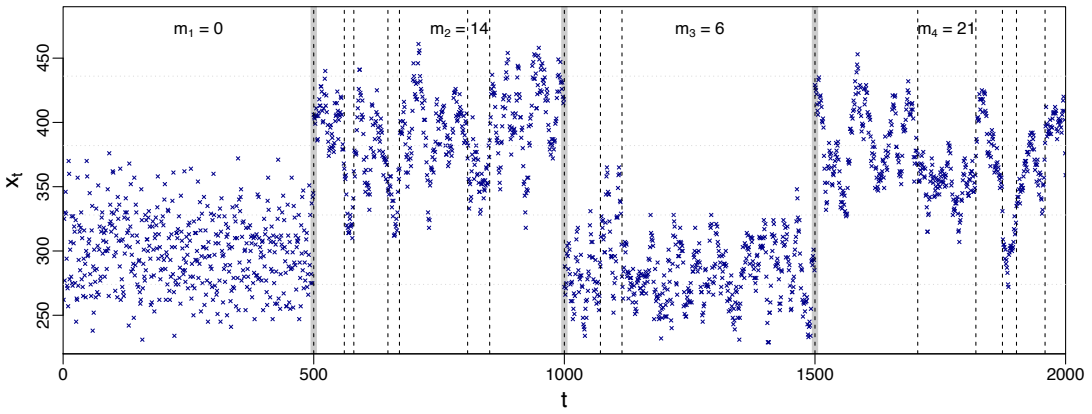


Figure 4.2.: *Data generated from the moving-sum changepoint model for negative binomial data given in Example 4.3 with three changepoints (τ_1, τ_2 and τ_3) indicated by thick grey lines. The black dashed lines indicate the positions of the MAP changepoints obtained by fitting the standard changepoint model (2.4) to the data.*

Figure 4.2 displays the positions of the MAP changepoints obtained by fitting the standard changepoint model for exchangeable data given in (2.4) to the simulated data using Metropolis-Hastings sampling of the changepoints as described in Denison et al. (2002). Within segments where the data are not exchangeable, the standard model cannot capture the temporal dynamics and therefore the data are inferred to be more segmented than preferable.

4.2. Conditional likelihood

Since changepoints split the data into independent segments, the joint posterior density of changepoints is tractable, up to a normalising constant, if the conditional likelihood of data within each segment can be computed. This section discusses the computation of the conditional likelihood of some data $x_1, \ldots, x_n$ observed within a single segment, assuming the moving-sum model defined in Section 4.1.1 for some $m \geqslant 0$ and θ.

4.2.1. Relationship between the observed data and the latent variables

This section gives insights on the relationship between the observed data and the latent variables, which is exploited to compute the conditional likelihood of the observed data within a generic segment in Section 4.2.2. We show that, for the latent m-dependence framework (4.1), there are m free latent variables subject to some constraints, and then all further latent variables are implied by the observed data sequence. Some of the notation introduced in this section will be useful in the remainder of the chapter.

Consider the backward difference operator ∇ and the forward difference operator $\triangle$ defined such

that, for all t,

$$\begin{aligned} \nabla x_t &= x_t - x_{t-1}, \\ \triangle x_t &= x_t - x_{t+1} = -\nabla x_{t+1}. \end{aligned} \tag{4.2}$$

The equation given in (4.1) may be equivalently expressed as

$$y_t = y_{t-\bar{m}} + \nabla x_t, \tag{4.3}$$

for $t = 2, \ldots, n$, and $y_1 = x_1 - (y_{-m+1} + \cdots + y_0)$. Iterating the expression in (4.3) shows that, given the initial m latent random variables $y_{-m+1}, \ldots, y_0$, there is a one-to-one relationship between the finite differences of $x_{1:n}$ (4.2) and the remaining latent variables $y_{1:n}$. Let the first m latent variables be $\gamma_{1:m} = (\gamma_1, \ldots, \gamma_m)$, with

$$\gamma_r = y_{-m+r}$$

for $r = 1, \ldots, m$. Explicitly, for all $t = 1, \ldots, n$, letting r be the remainder and q the quotient of the Euclidean division of $t-1$ by $\bar{m}$ such that $t = q\bar{m} + r + 1$, we have

$$y_t = \begin{cases} (\sum_{i=0}^{q} \nabla x_{i\bar{m}+1}) - \gamma^\bullet, & r = 0 \\ \gamma_r - \sum_{i=0}^{q} \triangle x_{i\bar{m}+r}, & r = 1, \ldots, m, \end{cases} \tag{4.4}$$

where $x_0 \equiv 0$ and

$$\gamma^\bullet = \sum_{r=1}^{m} \gamma_r. \tag{4.5}$$

The role played by $\gamma_{1:m}$ is akin to the role played by the unknown initial conditions of a stochastic difference equation.

The choice to condition on the first m latent random variables is arbitrary; given any sequence of m consecutive latent random variables, there is a one-to-one relationship between the other latent variables and the observed data. The following definition gives a transformation that may be used to obtain $\gamma_{1:m}$ from any m consecutive latent random variables, and thus it is sufficient to consider conditioning on the first m latent variables in all subsequent discussion. The transformation is also useful in the later sections of the chapter, where, for example, we need to obtain the initial latent variables when some data are added to, or removed from, an edge of a segment.

Definition 7. Let $x_1, \ldots, x_n$ be data observed within a segment, assuming the model (4.1) for some $m \geqslant 0$. Let $S[y_t, \ldots, y_{t+m-1}] = (y_{t+1}, \ldots, y_{t+m})$ denote the 'shift' map, where $y_{t+m} = x_{t+m} - \sum_{i=0}^{m-1} y_{t+i}$, for all suitable t. Clearly, S is iterable and invertible, and for all sequences

of m consecutive latent random variables $y_{(-m+1+u):u}$, with $0 \leqslant u \leqslant n$, $S^{-u}[y_{(-m+1+u):u}] = y_{(-m+1):0} = \gamma_{1:m}$.

4.2.2. Conditional likelihood of the data within a segment

Given m and $\gamma_{1:m}$, (4.4) provides a one-to-one deterministic mapping between $x_{1:n}$ and $y_{1:n}$ with unit Jacobian. Hence, if we treat the sequence $\gamma_{1:m}$ as an additional unknown segment parameter, whose elements are independent and identically distributed with density $f_m(\cdot \mid \theta)$, then the conditional likelihood of the observed data within a segment is

$$\begin{aligned} L(x_{1:n} \mid \theta, m, \gamma_{1:m}) &= L(y_{1:n} \mid \theta, m, \gamma_{1:m}) \\ &= \prod_{i=1}^{n} f_m(y_i \mid \theta). \end{aligned} \tag{4.6}$$

Thus, using the notation introduced in Section 4.1.2, but ignoring the subscripts corresponding to the indices of segments, the unknown segment parameters are $(\theta, m, \gamma_{1:m})$, with prior density $\pi(\theta)\pi(m)\pi(\gamma_{1:m} \mid \theta, m)$ where $\pi(\gamma_{1:m} \mid \theta, m) = \prod_{r=1}^{m} f_m(\gamma_r \mid \theta)$.

Recall it is assumed that the prior for θ is chosen to be conjugate for $f_m(\cdot \mid \theta)$ conditional on m. Consequently, the joint likelihood of the data and the initial latent variables conditional on m can be derived by invoking Bayes' theorem,

$$L(x_{1:n}, \gamma_{1:m} \mid m) = \int L(y_{1:n} \mid \theta, m, \gamma_{1:m}) \pi(\gamma_{1:m} \mid \theta, m) \pi(\theta) d\theta. \tag{4.7}$$

Let $\mathcal{F}$ denote the support of $f_m(\cdot \mid \theta)$ and, given m, $x_{1:n}$ and (4.4), let

$$\begin{aligned} \mathcal{Y}_m &\equiv \mathcal{Y}_m(x_{1:n}) \\ &= \{\gamma_{1:m} \in \mathcal{F}^m; y_t \in \mathcal{F} \text{ for all } t = 1, \ldots, n\}. \end{aligned} \tag{4.8}$$

$\mathcal{Y}_m$ is the set of sequences $\gamma_{1:m}$ such that the joint conditional probability density (4.7) is non-zero. As stated in Remark 1, it is not guaranteed that, for all m, $\mathcal{Y}_m$ will be non-empty. An expression for (4.7) is given below for the three examplar segment models introduced in Section 4.1.1.

Remark 1 (Set $\mathcal{Y}_m$)**.** Note that in the case of $m = 0$, where the sequence $x_{1:n}$ is assumed to be exchangeable, then $\gamma_{1:m}$ is the empty sequence, and the expression in (4.6) is always well defined. Now, if $m > 0$, two cases need to be considered separately. If $\mathcal{F}$ is unbounded, for all m and sequence $x_{1:n}$, the set $\mathcal{Y}_m$ is $\mathcal{F}^m$. However, if $\mathcal{F}$ is bounded then $\mathcal{Y}_m$ is a proper subset of $\mathcal{F}^m$ and is not necessarily non-empty for all $m > 0$ and $x_{1:n}$. For example, if $\mathcal{F}$ is bounded below by 0, for any non-negative sequence $x_{1:n}$ with $x_2 > x_1 + x_3$, the set $\mathcal{Y}_m$ is empty for all $m > 0$.

Example 4.1 (Normal marginal distribution, continued)**.** Given parameters $(m, \gamma_{1:m})$ and a known hyperparameters $\lambda, \alpha, \beta > 0$, it follows that

$$L(x_{1:n}, \gamma_{1:m} \mid m) = \left(\frac{\bar{m}}{2\pi}\right)^{(n+m)/2} \left(\frac{\lambda}{\lambda'}\right)^{1/2} \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha')}{(\beta')^{\alpha'}},$$

where $\lambda' = \frac{(n+m+\bar{m}\lambda)}{\bar{m}}$, $\alpha' = \frac{(n+m)}{2} + \alpha$ and $\beta' = \beta + \frac{\bar{m}}{2}(\sum_{t=-m+1}^{n} y_t^2) + \frac{\lambda}{2}\mu_0^2 - \frac{(\lambda\mu_0 + \sum_{t=-m+1}^{n} y_t)^2}{2\lambda'}$.

Example 4.2 (Gamma marginal distribution, continued)**.** Given parameters $(m, \gamma_{1:m})$ and known hyperparameters $\lambda, \alpha, \beta > 0$, it follows that

$$L(x_{1:n}, \gamma_{1:m} \mid m) = \frac{\Gamma(\alpha + (n+m)\lambda_m)}{\Gamma(\alpha)\Gamma(\lambda_m)^{n+m}} \frac{\beta^\alpha y_\bullet^{(n+m)(\lambda_m - 1)}}{(\beta + y^\bullet)^{\alpha + (n+m)\lambda_m}} \, \mathbf{1}_{\mathcal{Y}_m}(\gamma_{1:m}),$$

where $\lambda_m = \lambda/\bar{m}$, $y_\bullet = \prod_{t=-m+1}^{n} y_t$ and $y^\bullet = \sum_{t=-m+1}^{n} y_t$.

Example 4.3 (Negative binomial marginal distribution, continued)**.** Given parameters $(m, \gamma_{1:m})$ and known hyperparameters $r, \alpha, \beta > 0$, it follows that

$$L(x_{1:n}, \gamma_{1:m} \mid m) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \left(\prod_{t=-m+1}^{n} \frac{\Gamma(y_t + r_m)}{\Gamma(y_t + 1)\Gamma(r_m)} \right) \frac{\Gamma(y^\bullet + \alpha)\Gamma((n+m)r_m + \beta)}{\Gamma(y^\bullet + \alpha + (n+m)r_m + \beta)} \mathbf{1}_{\mathcal{Y}_m}(\gamma_{1:m}),$$

where $r_m = r/\bar{m}$ and $y^\bullet = \sum_{t=-m+1}^{n} y_t$.

The segment parameters $(m, \gamma_{1:m})$ cannot be marginalised, and consequently to sample from the posterior distribution of the changepoints, we need to sample $(m, \gamma_{1:m})$ for each segment. Different challenges arise when attempting to do so: the dimension of the segment parameter space is unknown; as first hinted in Remark 1, the parameter space depends on the observed data when the support of $f_m(\cdot \mid \theta)$ is bounded. Hence, in order to develop a sampling strategy that is computationally realistic, it is necessary to characterise the parameter space we are seeking to navigate.

4.3. Analysis of the latent parameter space

In Section 4.2.2 we defined $\mathcal{Y}_m$ to be set of sequences $\gamma_{1:m}$ such that the joint conditional probability density of $\gamma_{1:m}$ and $x_{1:n}$ is non-zero within a generic segment. We now define

$$\mathcal{M} \equiv \mathcal{M}(x_{1:n}) = \{m \in \mathbb{N}_0 \mid \mathcal{Y}_m \neq \emptyset\}$$

to be the set of $m \geqslant 0$ for which $\mathcal{Y}_m$ is non-empty. In other words, given some observed data $x_1, \ldots, x_n$ within a generic segment, $\mathcal{M}$ and $\mathcal{Y}_m$ provide the values of m and $\gamma_{1:m}$ for which the segment model (4.1) is valid.

As stated in Remark 1, it is always possible to assume the sequence $x_{1:n}$ to be exchangeable, and hence $0 \in \mathcal{M}$. If $\mathcal{F}$, the support of $f_m(\cdot \mid \theta)$, is unbounded, such as in Example 4.1, then it is immediate that $\mathcal{M} = \mathbb{N}_0$, and $\mathcal{Y}_m$ is $\mathcal{F}^m$ for all $m > 0$. However, if $\mathcal{F}$ is bounded, such as in Example 4.2 and Example 4.3, then both $\mathcal{M}$ and $\mathcal{Y}_m$ depend on the observed data $x_{1:n}$, and $\mathcal{M}$ and $\mathcal{Y}_m$ are proper subsets of $\mathbb{N}_0$ and $\mathcal{F}^m$, respectively. Note that infinitely divisible distributions with support bounded from below and from above have zero variance (Steutel, 1975), and therefore we only consider the case where $\mathcal{F}$ is bounded below but not above, without loss of generality.

Section 4.3.1 explicitly states the relationship between the observed data $x_{1:n}$ and the sets $\mathcal{M}$ and $\mathcal{Y}_m$ for $m \in \mathcal{M}$ when $\mathcal{F}$ is bounded below. Section 4.3.2 and Section 4.3.3 provides results on the structure of the constrained parameter space that are used in Section 4.4 to design a sampling strategy for the model parameters. Section 4.3.4 discusses the asymptotic behaviour of the parameter space as the number of observations increase; in particular, for the segment parameters $\gamma_{1:m}$, which may be considered to be the unknown initial conditions of a process governed by a difference equation as argued in Section 4.2.1, it is shown that, asymptotically, the uncertainty on the initial conditions vanishes.

4.3.1. Characterisation of the parameter space in terms of the observed data

Suppose that $\mathcal{F}$ is unbounded above but bounded below by a constant, which can be set to 0 without loss of generality. It follows that, given a sequence $x_{1:n}$ and some $m > 0$, the set $\mathcal{Y}_m$ consists of those $\gamma_{1:m} \in \mathcal{F}^m$ such that $y_t \geqslant 0$ for all $t = 1, \ldots, n$. According to (4.4) and (4.5), defining

$$\begin{aligned} U^m &\equiv U^m(x_{1:n}) = \min\{\nabla x_1, \ldots, \textstyle\sum_{q=0}^{\kappa_{\bar{m}}} \nabla x_{q\bar{m}+1}\}, \\ L_r^m &\equiv L_r^m(x_{1:n}) = \max\{0, \triangle x_r, \ldots, \textstyle\sum_{q=0}^{\kappa_r} \triangle x_{q\bar{m}+r}\}, \end{aligned} \tag{4.9}$$

where $x_0 \equiv 0$ and $\kappa_r = \kappa_r(n)$ is the largest $q \in \mathbb{N}_0$ such that $q\bar{m} + r + 1 \leqslant n$ for all r, then

$$\mathcal{Y}_m = \mathbb{Y}[U^m, L_{1:m}^m], \tag{4.10}$$

where

$$\mathbb{Y}[U^m, L_{1:m}^m] = \{\gamma_{1:m};\, U^m \geqslant \gamma^\bullet \text{ and } \gamma_r \geqslant L_r^m \text{ for all } r = 1, \ldots, m\}.$$

It will be useful both in the remainder of this section and in Section 4.4 to characterise $\mathcal{Y}_m$ (4.8) in terms of $\mathbb{Y}$ (4.10). Moreover, for $m \geqslant 0$, $m \in \mathcal{M}$ and $\mathcal{Y}_m$ is non-empty if and only if

$$D^m \equiv D^m(x_{1:n}) = U^m - \sum_r L_r^m \geqslant 0. \tag{4.11}$$

Example 4.4 gives expanded expressions of the bounds of $\mathcal{Y}_m$ for a small sequence of observed

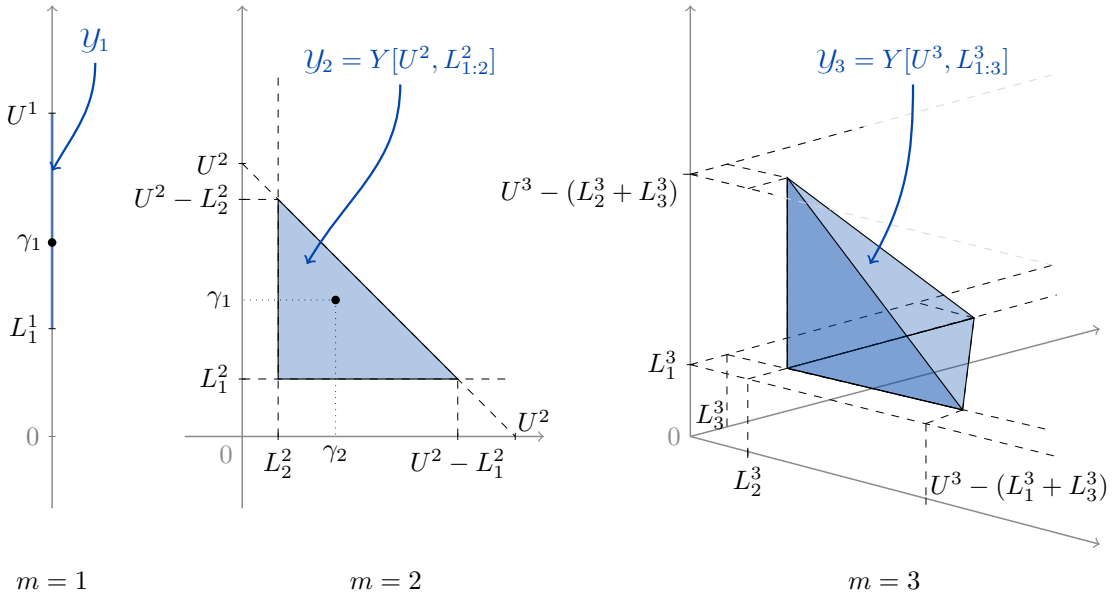


Figure 4.3.: *Cartoons of the set $\mathcal{Y}_m$ for $m = 1, 2, 3$ with $\mathcal{F}$ assumed to be continuous.*

data.

Example 4.4 (Bounds of $\mathcal{Y}_m$)**.** Let $x_1, \ldots, x_7$ be count data. If $m = 1$ ($\bar{m} = 2$), then

$$\begin{cases} U^1 = \min\{x_1, x_1 + (x_3 - x_2), x_1 + (x_3 - x_2) + (x_5 - x_4), x_1 + (x_3 - x_2) + (x_5 - x_4) + (x_7 - x_6)\} \\ L_1^1 = \max\{0, (x_1 - x_2), (x_1 - x_2) + (x_3 - x_4), (x_1 - x_2) + (x_3 - x_4) + (x_5 - x_6)\}, \end{cases}$$

whilst if $m = 3$ ($\bar{m} = 4$), then

$$\begin{cases} U^3 = \min\{x_1, x_1 + (x_5 - x_4)\} \\ L_1^3 = \max\{0, (x_1 - x_2), (x_1 - x_2) + (x_5 - x_6)\} \\ L_2^3 = \max\{0, (x_2 - x_3), (x_2 - x_3) + (x_6 - x_7)\} \\ L_3^3 = \max\{0, (x_3 - x_4)\}. \end{cases}$$

Figure 4.3 displays cartoon representations of $\mathcal{Y}_m$ for $m = 1, 2$ and 3. For all $m > 0$, if non-empty, $\mathcal{Y}_m$ is a convex polyhedron in $\mathcal{F}^m$, whose vertices are determined by the bounds U^m and $L^m_{1:m}$, and whose size only depends on m and D^m for all m.

Given a sequence of observed data $x_{1:n}$ within a segment, it was shown in (4.10) that the sets $\mathcal{M}$ and $\mathcal{Y}_m$ may be expressed in terms of the first observation $\nabla x_1 = x_1 - x_0 = x_1$ and all of the subsequent the finite differences of the data. Essentially, for all m, the larger the finite differences, or *jumps*, are relative to x_1, the smaller $\mathcal{Y}_m$ becomes unless the jumps happen to exhibit negative $\bar{m}$-lagged autocorrelations. In particular, if the observed data $x_{1:n}$ consist of a succession of gradual drifts, whose directions flip every $\bar{m}$ time points, then $\mathcal{Y}_m$ is large with respect to x_1.

Furthermore, in order to stress that the size of the jumps must be considered relative to x_1, we note that for all sequences $x_{1:n}$ such that $\mathcal{Y}_m$ is non empty for some m, and for all constant μ, $D^m(x_{1:n} + \mu) = D^m(x_{1:n}) + \mu$, where $x_{1:n} + \mu$ denotes the sequence $(x_t + \mu)_{t=1}^n$.

4.3.2. Adding and removing data

To develop a sampling strategy for the latent variables, it is important to determine how $\mathcal{M}$ and $\mathcal{Y}_m$, $m \in \mathcal{M}$, may change when data are added to, or removed from, an edge of a segment. Proposition 2 guarantees that, for all sequences of data, if there exist segment parameters $\gamma_{1:m}$ for some m, then we can obtain valid parameters of the same dimension for all sequences of data that could be obtained by removing data from the beginning or the end of the original sequence.

Proposition 2. Let $x_{1:n}$ be observed data and $1 \leqslant s \leqslant t \leqslant n$. Then $\mathcal{M}(x_{1:n}) \subseteq \mathcal{M}(x_{s:t})$, and $S^{s-1}[\mathcal{Y}_m(x_{1:n})] \subseteq \mathcal{Y}_m(x_{s:t})$ for all $m \in \mathcal{M}(x_{1:n})$, where the shift map S is defined in Definition 7. In particular, if $s = 1$ we have $\mathcal{Y}_m(x_{1:n}) \subseteq \mathcal{Y}_m(x_{1:t})$.

Proof. Suppose $\gamma_{1:m} = y_{(-m+1):0} \in \mathcal{Y}_m(x_{1:n})$ for some $m \in \mathcal{M}(x_{1:n})$. Let $y_{1:n}$ denote the latent variables obtained from $\gamma_{1:m}$ and $x_{1:n}$ via (4.4). By definition of the shift map, the latent variables obtained from $S^{s-1}[y_{(-m+1):0}] = y_{(-m+s):(s-1)}$ and $x_{s:t}$ via (4.4) are equal to $y_{s:t}$. Hence $S^{s-1}[y_{(-m+1):0}] \in \mathcal{Y}_m(x_{s:t})$, and therefore $m \in \mathcal{M}(x_{s:t})$. □

However, if some data are added to the beginning or the end of a sequence of data for which we currently have valid parameters $\gamma_{1:m}$ for some m, then it is not guaranteed that $m \in \mathcal{M}$ for the extended sequence of data. For example, it follows directly from (4.9) that, as more data are added to the end of a segment, for all m, U^m may only decrease and L_m^r may only increase for all r, and therefore $\mathcal{Y}_m = \mathbb{Y}[U^m, L_{1:m}^m]$ may only shrink.

4.3.3. Different ranges of dependence

Given a segment of data, there is a non-negligible computational cost in verifying via (4.11) that some integer m belongs to the corresponding set $\mathcal{M}$. Hence, it is instructive to determine whether knowing that some m belongs to $\mathcal{M}$ can inform whether some $m' \neq m$ is also an element of $\mathcal{M}$.

Suppose $m \in \mathcal{M}$ and $0 \leqslant m' \leqslant m$ such that $\bar{m}'$ divides $\bar{m}$ (written $\bar{m}' \mid \bar{m}$), meaning $\bar{m} = \ell\bar{d}$ for some $\ell > 0$. Then define a mapping

$$J_{m'}(\gamma_{1:m}) = \left(\sum_{j=0}^{\ell-1} \gamma_{j\bar{m}'+1}, \ldots, \sum_{j=0}^{\ell-1} \gamma_{j\bar{m}'+m'} \right)$$

for aggregating the latent variables. Figure 4.4 displays a cartoon representation of the mapping. This mapping is required for the following proposition.

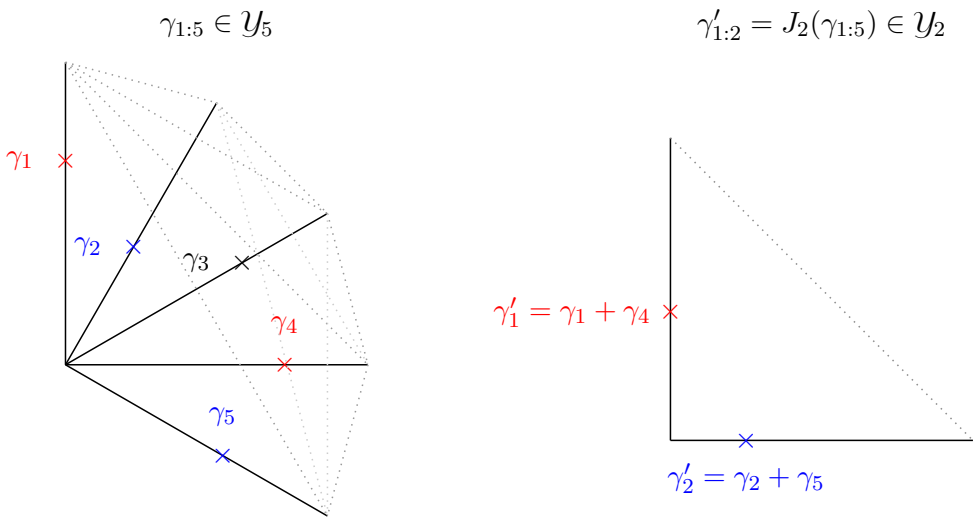


Figure 4.4.: *Cartoon representation of the mapping of* $\gamma_{1:m} \in \mathcal{Y}_m$ *to* $\gamma'_{1:m'} = J_{m'}(\gamma_{1:m}) \in \mathcal{Y}_{m'}$ *with* $m = 5$ *and* $m' = 2$.

Proposition 3. For all $m \in \mathcal{M}$, $\gamma_{1:m} \in \mathcal{Y}_m$ and $m' < m$ such that $\bar{m}' \mid \bar{m}$,

(i) $m' \in \mathcal{M}$ and $J_{m'}(\gamma_{1:m}) \in \mathcal{Y}_{m'}$.

(ii) $J_{m'}(\mathcal{Y}_m) = \mathbb{Y}[\tilde{U}, J_{m'}(L^m_{1:m})] \subset \mathbb{Y}[U^{m'}, L^{m'}_{1:m'}] = \mathcal{Y}_{m'}$, where $\tilde{U} = U^m - \sum_{j=1}^{\ell-1} L^m_{j\bar{m}'}$.

Proof. See Appendix B.2. □

To provide some intuition for Proposition 3, it is helpful to consider the bounds given in Example 4.4, noting that $3 + 1 = 2(1 + 1)$ and that the finite differences appearing in L^3_1 and L^3_3 for $m = 3$ coincide with the finite differences in L^1_1 for $m = 1$.

Proposition 3 says that for all $m \in \mathcal{M}$ it follows immediately that $\mathcal{D}(m) \subset \mathcal{M}$, where

$$\mathcal{D}(m) = \{m' \in \mathbb{N}_0;\ \bar{m}' \mid \bar{m}\} \tag{4.12}$$

consists of the integers m' such that $\bar{m}'$ divides $\bar{m}$.

4.3.4. Asymptotic properties of the parameter space

In Section 4.3.2 it was shown that the parameter space may only shrink as more data are observed within a segment. Proposition 4 sheds further light upon the asymptotic properties of $\mathcal{M}$ and $\mathcal{Y}_m$ for all m.

Proposition 4. For $m \in \mathbb{N}_0$, suppose a sequence of latent variables $y_{-m+1}, \ldots, y_n$, and a sequence of observed data $x_1, \ldots, x_n$ are generated from model (4.1), assuming some density $f_m(\cdot \mid \theta)$ with

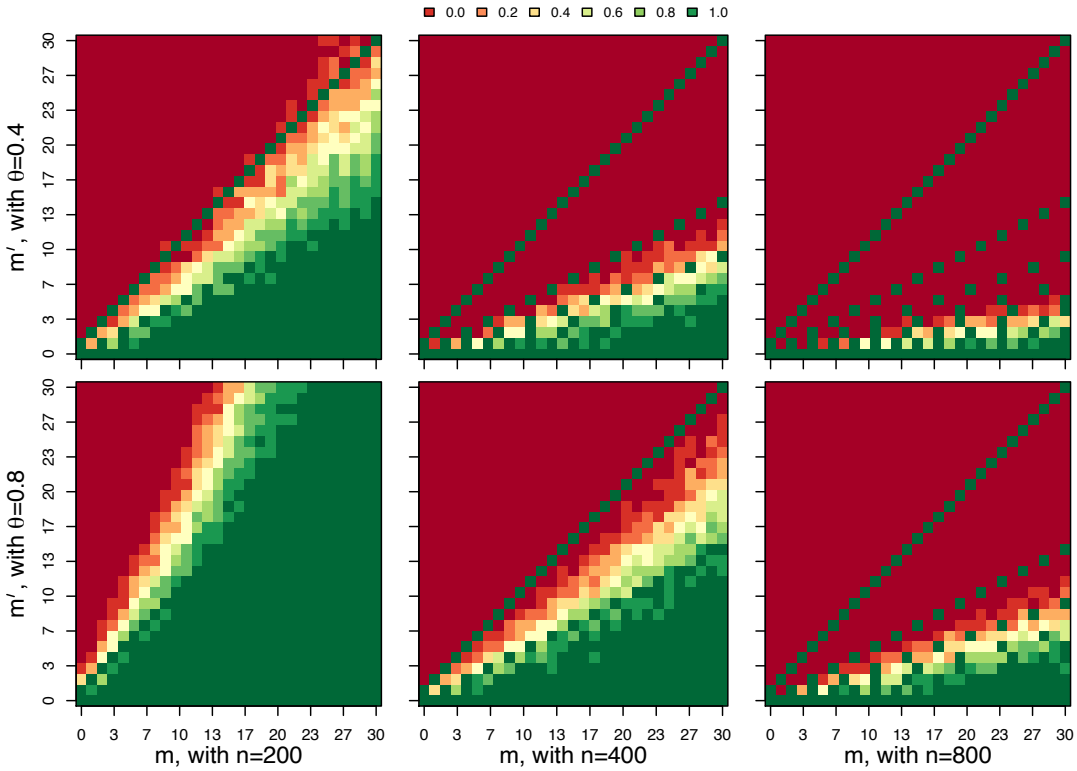


Figure 4.5.: *The colours indicate the proportion of simulations for which $m' \in \mathcal{M}$, denoted $Q(m' \in \mathcal{M} \mid n, m, \theta)$, based on 50 simulations from the moving-sum segment model for negative binomial data given in Example 4.3 for $m = 0, \ldots, 30$, $n = 200, 400, 800$ and $\theta = 0.4, 0.8$.*

support $\mathbb{N}_0$ or $\mathbb{R}^+$. As $n \to \infty$, for all $m' > 0$,

$$U^{m'}(x_{1:n}) \xrightarrow{\text{a.s.}} \begin{cases} \sum_{r=1}^{m'} \gamma'_r & \text{if } m' \in \mathcal{D}(m) \\ -\infty & \text{otherwise} \end{cases}$$

$$L_r^{m'}(x_{1:n}) \xrightarrow{\text{a.s.}} \begin{cases} \gamma'_r & \text{if } m' \in \mathcal{D}(m) \\ \infty & \text{otherwise} \end{cases}$$

for all $r = 1, \ldots, m'$, with $\gamma'_{1:m'} = J_{m'}(\gamma_{1:m})$, and thus, almost surely, $\mathcal{M}(x_{1:n})$ converges to $\mathcal{D}(m)$ (4.12) and, for all $m' \in \mathcal{D}(m)$, $\mathcal{Y}_{m'}(x_{1:n}) = \mathbb{Y}[U^{m'}, L_{1:m'}^{m'}]$ converges to $\{\gamma'_{1:m'}\}$.

Proof. See Appendix B.3. □

Proposition 4 tells us that if some data are generated from model (4.1) for some $m > 0$ and some initial latent variables $\gamma_{1:m}$, then almost surely, as the number of observations tends to infinity, $\mathcal{M}$ converges to $\mathcal{D}(m)$ and $\mathcal{Y}_{m'}$ converges to a set containing a unique sequence, namely the transformation of $\gamma_{1:m}$ by $J_{m'}$, for all $m' \in \mathcal{D}(m)$. We note that the result shows that both the

structure in $\mathcal{M}$ and the transformation identified in Proposition 3 are fundamental, and increasingly important as the length of a changepoint segment increases.

Within the changepoint detection framework, it is appealing that, the more data are observed within a segment, the more information the parameter space gives us about the nature of the dependence within the segment, so that a changepoint may be forced immediately upon observing some data generated from a different dependency structure.

Recall that in Section 4.2.1 it was argued that the segment parameters $\gamma_{1:m}$ may be considered as the unknown initial conditions of a process governed by a difference equation. From this point of view, Proposition 4 says that, asymptotically, the uncertainty on the initial conditions vanishes.

An experiment was performed to illustrate Proposition 4. For six different parameter configurations corresponding to different fixed values of n and θ, and for all m in $\{0, \ldots, 30\}$, we performed 50 simulations from the moving-sum segment model for negative binomial data given in Example 4.3 with $r = 300$. For all $m' \in \{0, \ldots, 30\}$ we computed the proportion of simulations for which $m' \in \mathcal{M}$, denoted $Q(m' \in \mathcal{M} \mid n, m, \theta)$. Figure 4.5 displays the results of the experiment. As expected from Proposition 4, for all parameter choices it is apparent that $m' \in \mathcal{M}$ with estimated probability 1 for all $m' \in \mathcal{D}(m)$, and that $Q(m' \in \mathcal{M} \mid n, m, \theta)$ tends to decrease as n increases for all $m' \notin \mathcal{D}(m)$. Moreover, three other trends are worth mentioning. First, $Q(m' \in \mathcal{M} \mid n, m, \theta)$ tends to increase as m increases for all $m' \notin \mathcal{D}(m)$ given n and θ. Second, it tends to be more likely for m'_1 to be in $\mathcal{M}$ than for m'_2 to be in $\mathcal{M}$ for all $m'_1, m'_2 \notin \mathcal{D}(m)$ such that $m'_1 < m'_2$. Third, one may observe that $Q(m' \in \mathcal{M} \mid n, m, \theta)$ tends to increase as θ increases for all $m' \notin \mathcal{D}(m)$ given all m and n.

4.4. Inference via MCMC

For the Bayesian changepoint model given in Section 4.1.2, the reversible jump MCMC algorithm (Green, 1995), which is a Metropolis-Hastings algorithm suitable for target distributions of varying dimension, may be used to sample from the posterior distribution of the positions of an unknown number of changepoints. Four types of moves are considered to explore the support of the target distribution: *shift* of a randomly selected changepoint; change of a randomly chosen segment parameter; *birth* of a new changepoint chosen uniformly over the time period; and *death* of a randomly selected changepoint.

In this section, within the framework given in Green (1995), we propose a strategy to sample from the posterior distribution of changepoints when the moving-sum model defined in Section 4.1 is assumed for each segment. To address the challenges that the dimensions of the segment parameters are unknown and that the segment parameter space depends on the observed data within each segment, we exploit the analysis of the segment parameter space in Section 4.3.

4.4.1. Description of the sampler

Suppose that the latest particle of the sample chain consists of k changepoints, whose positions are $\tau_{1:k}$, and $(k+1)$ segment parameters. For segment j, recall from Section 4.2.2 that we assume the latent variable density parameter θ_j may be marginalised, and thus the segment parameters consist of the order of dependence m_j and the initial latent variables $\gamma_{j,:} \equiv \gamma_{j,1:m_j} = (\gamma_{j,1}, \ldots, \gamma_{j,m_j})$. To explore the support of the target distribution, we propose the next element of the chain via one of the following moves.

4.4.1.1. Shift move

The shift move proposes to modify the position of one randomly chosen changepoint. The index j is uniformly chosen from $\{1, \ldots, k\}$, and a new position τ'_j is uniformly sampled from $\{\tau_{j-1} + 1, \ldots, \tau_{j+1} - 1\}$. The parameters m_j, m_{j+1} and $\gamma_{j,:}$ are not modified and we replace $\gamma_{j+1,:}$ by $\gamma'_{j+1,:} = S^u(\gamma_{j+1,:})$, with $u = \tau'_j - \tau_j$.

As noted in Section 4.3.2, when the support of $f_m(\cdot \mid \theta)$ is bounded, the move may be rejected because the updated latent variables are invalid: If the length of the j-th segment is reduced by the shift move, i.e. $\tau'_j - \tau_j$ is negative, then it is guaranteed that $\gamma_{j,:}$ remains a valid sequence of initial latent variables but not that $\gamma'_{j+1,:}$ is valid for the extended $(j+1)$-th segment; one the other hand, if $\tau'_j - \tau_j$ is positive then the sequence $\gamma'_{j+1,:}$ is valid but it must be checked that $\gamma_{j,:}$ is valid for the extended j-th segment.

4.4.1.2. Sampling a segment parameter

A segment j is uniformly chosen amongst the $k+1$ segments, and the corresponding segment parameters are changed: Either sample the initial latent variables conditional on the order of dependence that is left unchanged; or the order of dependence is sampled and therefore the initial latent variables must be adapted. Here, the focus is on one segment only, and therefore we temporarily drop the segment index j from the notation as in Section 4.3, and the data observed within the segment are denoted by $x_1, \ldots, x_n$, where $n = \tau_j - \tau_{j-1}$.

4.4.1.2.1. An approximation to the posterior distribution of θ First, we consider an approximation to the posterior distribution of θ that will be useful to build proposals for m and $\gamma_{1:m}$. In the absence of knowledge on m and $\gamma_{1:m}$, motivated by computational considerations, it is interesting to consider the posterior distribution of θ conditional on $m = 0$. When the data are assumed to be exchangeable, by conjugacy of the prior for θ, the posterior distribution of θ is tractable,

$$\hat{\pi}(\theta \mid x_{1:n}) = \pi(\theta \mid m = 0, x_{1:n}).$$

Based on this approximation, a natural estimator for θ is

$$\hat{\theta} = \arg\max_\theta \hat{\pi}(\theta \mid x_{1:n}).$$

4.4.1.2.2. Updating $\gamma_{1:m}$ conditional on m The move consists in proposing $\gamma'_{1:m}$ conditional on m and $\gamma_{1:m}$. Recall from Section 4.3.1 that, conditional on the observed data, the support of the initial latent variables is $\mathcal{Y}_m$, and note that $\gamma_{1:m} \in \mathcal{Y}_m$ if and only if, for all $r = 1, \ldots, m$, $\gamma_r \in \mathcal{Y}_m^r$ where

$$\mathcal{Y}_m^r = \left\{ \gamma_r \in \mathcal{F} \mid L_r^m \leqslant \gamma_r \leqslant U^m - \sum_{i \neq r} \gamma_i \right\},$$

if $\mathcal{F}$ is bounded and $\mathcal{Y}_m^r = \mathcal{F}$ if $\mathcal{F}$ is unbounded.

We consider two distinct scenarios based on the nature of the latent variables. If the latent variables are discrete valued, then $\gamma'_{1:m}$ are proposed via Gibbs sampling. It follows from the discussion in Section 4.2.2 that, for all $r = 1, \ldots, m$, the full conditional distribution of γ_r is

$$\pi(\gamma_r \mid x_{1:n}, \gamma_{-r}, m) \propto L(x_{1:n}, \gamma_{1:m} \mid m) \mathbf{1}_{\mathcal{Y}_m^r}(\gamma_r),$$

where $\gamma_{-r} = (\gamma_1, \ldots, \gamma_{r-1}, \gamma_{r+1}, \ldots, \gamma_m)$. If the domain of the latent variables is continuous, then Gibbs sampling is not possible in general; instead, for all $r = 1, \ldots, m$, sample γ'_r from the distribution with step function density

$$q(\gamma_r \mid x_{1:n}, \gamma_{-r}, m) \propto \sum_{i=1}^{N} L(\gamma_{1:m}, y_{1:n} \mid m) \mathbf{1}_{[\gamma^{(i)}, \gamma^{(i+1)})}(\gamma_r), \tag{4.13}$$

where $\gamma^{(1)} < \cdots < \gamma^{(N)}$ form an equally spaced grid on the largest interval $\mathcal{Y}_m^{r*} \subseteq \mathcal{Y}_m^r$ satisfying

$$\int_{y_r \in \mathcal{Y}_m^{r*}} f_m(\gamma_r \mid \hat{\theta}) \mathrm{d}\gamma_r \leqslant \eta.$$

The greater $N > 1$ and $0 < \eta < 1$, the more accurate the step function approximation of the full conditional distribution of γ'_r in (4.13). The tuning parameters N and η can be chosen via pilot runs investigating the trade-off between precision and computational cost.

4.4.1.2.3. Updating m and $\gamma_{1:m}$ When m is replaced by some m', it is necessary that $m' \in \mathcal{M}$ for the move to be accepted, and we must propose some revised initial latent variables $\gamma'_{1:m'}$.

To sample m', whose full conditional distribution is not tractable in general, we consider a proposal distribution that relies on the following observations: the joint likelihood of the jumps (∇x_t) defined in (4.2) is not tractable in general due the dependence of the jumps; yet, the jumps

are identically distributed with mean $E[\nabla x_t] = 0$ and variance $V[\nabla x_t] = 2g(\theta, m)$ for some function g that depends on the marginal distribution of the latent variables $f_m(\cdot \mid \theta)$; and therefore the approximation of the likelihood of the jumps

$$\hat{L}(\nabla x_{1:n} \mid \hat{\theta}, m) = \prod_t \phi\left(\nabla x_t \mid 0, 2g(\hat{\theta}, m)\right),$$

where $\phi(\cdot \mid \mu, \sigma^2)$ is the density function of the normal distribution with mean μ and variance σ^2, is tractable and depends on m. The proposed order of dependence m' is sampled from the distribution with probability mass function

$$q(m' \mid \nabla x_{1:n}) \propto \hat{L}(\nabla x_{1:n} \mid \hat{\theta}, m')\pi(m')\mathbf{1}_{\mathcal{M}}(m').$$

Then, according to Section 4.4.1.2.2, $\gamma'_{1:m'}$ is proposed conditional on m' and $\gamma^*_{1:m'}$, where

$$\gamma^*_{1:m} = \arg\max_{\gamma^*_{1:m'} \in \mathcal{Y}_{m'}} \prod_r f_{m'}(\gamma^*_r \mid \hat{\theta})$$

is an estimator of the initial latent variables in $\mathcal{Y}_{m'}$ that can be derived efficiently but does not take into account the dependence of the data.

4.4.1.3. Death and birth moves

If a death move is proposed, an index j of one element of $\tau_{1:k}$ is uniformly chosen, and the corresponding changepoint is removed, resulting in $k' = k - 1$ changepoints with positions $\tau'_{1:k'} = (\tau_1, \ldots, \tau_{j-1}, \tau_{j+1}, \ldots, \tau_k)$. The parameters corresponding to the segments that are not impacted by the move are re-indexed but kept unchanged, and it is natural to propose the parameters for the j-th segment resulting from $\tau'_{1:k'}$, namely $(m'_j, \gamma'_{j,:})$, based on the parameters of either the original j-th or $(j+1)$-th segment. Specifically, let i be either the index j or $j+1$ with probability proportional to the length of the segment with index i, and then set m'_j to m_i, and $\gamma'_{j,:}$ to $S^{(\tau_j - \tau_i)}(\gamma_{i,:})$. Note that the death move may then be seen as the extension of one of the segments on either side of the deleted changepoint.

The above death move may be reversed by the following birth move. Draw τ'_j uniformly from $\{2, \ldots, T\} \setminus \tau_{1:k}$, and obtain $\tau'_{1:k'}$ by inserting τ'_j into $\tau_{1:k}$ at the j-th position, resulting in $k' = k+1$ changepoints. Let i be either the index j or $j+1$ with probability proportional to $\tau'_i - \tau'_{i-1}$. Set m'_i to m_j and $\gamma'_{i,:}$ to $S^{(\tau'_i - \tau'_j)}(\gamma_{j,:})$, and finally propose the segment parameters of the new segment using the approach given in Section 4.4.1.2.

4.4.2. Sampler initialisation

To speed-up the convergence of the sampler for the moving-sum changepoint model, the sample chain is initialised as follows: the changepoint parameters are set to be the changepoint estimates corresponding to the standard changepoint model; and, for each segment, the order of dependence is set to 0. Hence, the sampler begins with a sensible positioning of the changepoints obtained at a limited computational cost.

4.4.3. Parameter estimation

Following Section 2.4, in this chapter, the point estimate $(\hat{k}, \hat{\tau}_{1:\hat{k}}, \hat{m}_{1:\hat{k}+1})$ for the model parameters $(k, \tau_{1:k}, m_{1:k+1})$ is defined as follows: $\hat{k}$ is the MAP number of changepoints; $\hat{\tau}_{1:\hat{k}}$ are the MAP changepoint positions of dimension $\hat{k}$; and $\hat{m}_{1:\hat{k}+1}$ are the MAP orders of dependence conditional on MAP changepoint parameters $(\hat{k}, \hat{\tau}_{1:\hat{k}})$.

4.4.4. Validating the estimated orders of dependence

Appendix B.4 details an importance sampling approach to estimating the posterior distribution of the order of dependence within a segment. Conditional on MAP changepoint parameters obtained via MCMC, it can be used to validate the MAP orders of dependence.

4.5. Simulations

This section describes a simulation study that demonstrates the benefits of the moving-sum changepoint model in comparison to changepoint models that assume the data are exchangeable within segments, namely the standard Bayesian changepoint model (2.4) and PELT (Killick, Fearnhead, and Eckley, 2012), and models that detect abrupt mean changes in normal data subject to local fluctuations and autocorrelated noise, namely DeCAFS (Romano et al., 2021) and WCMgSa (Cho and Fryzlewicz, 2021).

4.5.1. Synthetic data

Different scenarios were assumed to sample time series of length $T \in \{1200, 3200, 5200, 7200, 9200\}$ from the moving-sum changepoint model with segment model defined in Example 4.1: within each segment, the data are m-dependent and marginally normally distributed with mean μ and variance σ^2 for some segment specific parameters m, μ and σ^2. Different scenarios were considered for the number of changepoints $k \in \{0, 1, 3, 5, 7, 9\}$, and changepoint positions $\tau_{1:k}$ were specified such that, for all $j = 1, \ldots, k$, $\tau_j = \lfloor jT/(k+1) \rfloor + u_j$ for some u_j drawn uniformly from $\{-\lfloor T/(2k+2) \rfloor, \ldots, \lfloor T/(2k+2) \rfloor\}$, so that segment lengths may be uneven. The orders of dependence for each segment were sampled independently from Geometric(ν) for some $0 \leqslant \upsilon \leqslant 1$;

let $\bar{\upsilon} = (1 - \upsilon)/\upsilon$ denote the expected order of dependence for each segment. The segment mean parameters are set such that, for all $j = 1, \ldots, k + 1$, $\mu_j = \mu$ if j is odd and $-\mu$ otherwise, for some $\mu \in \mathbb{R}$. The segment parameters $\sigma_1^{-2}, \ldots, \sigma_{k+1}^{-2}$ were sampled independently from Gamma$(\alpha_0, 100)$ for some $\alpha_0 > 0$. A grid of parameters such that $\upsilon \in \{1, 0.33, 0.2, 0.1, 0.06\}$, $\mu \in \{0.5, 1, 2, 4, 8\}$ and $\alpha_0 \in \{1, 5, 10, 25, 50, 75\}$ was considered for the experiments. For each scenario, 10 simulations were performed.

4.5.2. Model comparison

For each simulation, different models were used to infer changepoint estimates from the data: the moving-sum changepoint model assuming different hyperparameters, the standard changepoint model (2.4), DeCAFS, PELT with different penalties, and WCMgSA.

For the moving-sum changepoint model, the data within segments are assumed to follow the model given in Example 4.1 with $\alpha = \alpha_0$, $\beta = 100$ and $\lambda = 0.05\beta/\alpha$; and the prior for the orders of dependence is assumed to be Geometric(ρ) for $\rho \in \{0.01, 0.1, 0.5\}$. Ten independent sample chains of size 20000, after a burn-in of size 10000, were obtained via the MCMC algorithm described in Section 4.4 with tuning parameters $N = 100$ and $\eta = 0.99$ for the proposal of the initial latent variables (4.13). Changepoint estimates obtained from each independent sample chain will be compared to assess the convergence of the sampler. Moreover, for the standard Bayesian changepoint model, changepoint estimates are obtained as for the moving-sum changepoint model with the only difference that, for each segment, we fix the order of dependence $m = 0$, ensuring the data are assumed to be exchangeable. For PELT, we used the default implementation in the R package *changepoint* (Killick and Eckley, 2014) with the MBIC penalty rescaled by a constant $\phi \in \{0.5, 1, 1.5\}$; as the penalty increases, more evidence is required from the data to flag changepoints. Finally, for DeCAFS and WCMgSA, changepoints were estimated with the default implementations described in Romano et al. (2021) and in Cho and Fryzlewicz (2021), respectively.

To compare changepoint estimations, we use the $F1$ score. A changepoint τ is said to be detected if there is an estimated changepoint $\hat{\tau}$ such that $\mid \tau - \hat{\tau} \mid \leqslant \epsilon$ for some error tolerance of size $\epsilon = 5$. Given some changepoints, the $F1$ score of the changepoint estimates is

$$F1 = \frac{2PR}{P + R} \in [0, 1],$$

where R and P denote the recall, the proportion of simulated changes that are detected, and the precision, the proportion of detected changes that are correct, respectively. The greater the $F1$ score, the better the estimation. To compare the three changepoint models of interest, for each simulation the $F1$ score was computed for each estimation of the simulated changepoints $(k, \tau_{1:k})$.

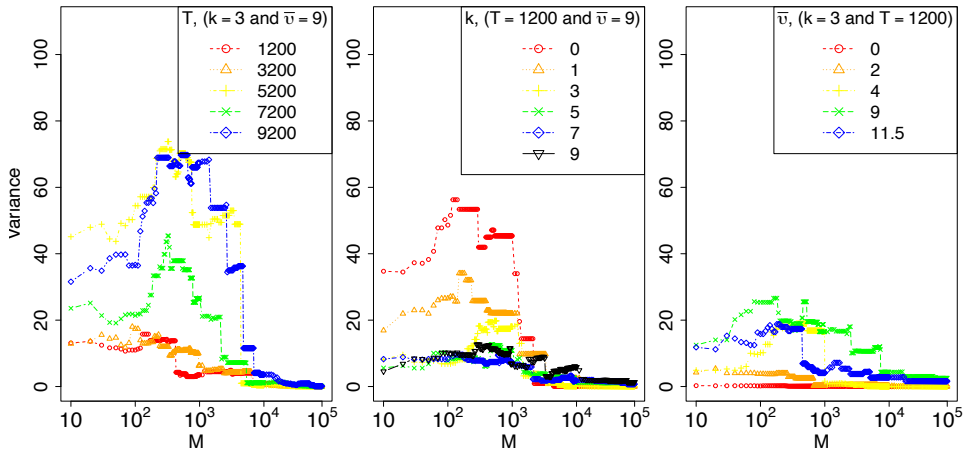


Figure 4.6.: *Variance of the MAP number of changepoints for the moving-sum model computed using the first M elements of sample chains obtained via the proposed MCMC algorithm for varying number of observations T, number of changepoints k and expected order of dependence $\bar{v}$ for each segment.*

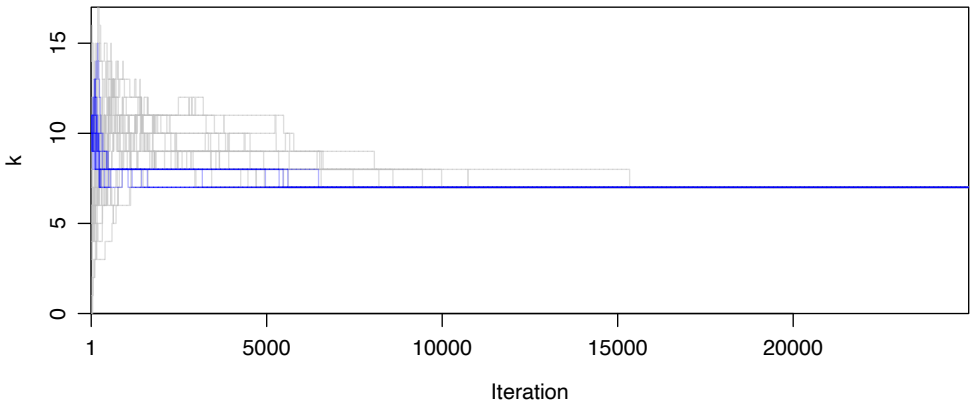


Figure 4.7.: *Thirty independent sample chains for the number of changepoints, obtained via the MCMC algorithm for the moving-sum changepoint model, for one selected time series with $k = 7$. In blue: ten sample chains initialised with the proposed initialisation strategy. In grey: twenty sample chains initialised with randomly selected changepoint parameters such that $0 \leqslant k \leqslant 20$.*

4.5.3. Sampler convergence

Recall from Section 4.5.1 that ten independent sample chains for the moving-sum changepoint model were obtained via the proposed MCMC algorithm for each time series analysed. Figure 4.6 displays the variance of the MAP number of changepoints obtained for the moving-sum model computed using the first M elements of the sample chains for a collection of scenarios;

results are similar for other scenarios. As M increases, the variances tend to 0, suggesting the samplers converge; a burn-in of size 10000 seems adapted to all scenarios considered. The rate of convergence tends to decrease with the number of observations T and with the expected order of dependence $\bar{v}$ for each segment, but tends to increase with the number of changepoints k.

Moreover, to illustrate the benefits of the sampler initialisation proposed in Section 4.4.2 we assess the convergence of the sampler for one randomly selected simulation with $k = 7$. Following Section 4.4.2, the ten independent sample chains for the moving-sum changepoint model were initialised with the changepoint estimates corresponding to the standard changepoint model. For comparison purposes only, twenty extra independent sample chains, initialised with randomly selected changepoint parameters such that $0 \leqslant k \leqslant 20$, were computed for the moving-sum model, as described in Section 4.5.2. Figure 4.7 displays the thirty independent sample chains for the number of changepoints k. All sample chains converge to the same number of changepoints, namely $k = 7$, and the ten sample chains corresponding to our proposed initialisation, indicated in blue, converge faster, illustrating the benefits of our proposed sampling strategy.

4.5.4. Runtime

The moving-sum changepoint model is computationally more demanding than the standard changepoint model due to the need to propose orders of dependence and initial latent variables. For our Python implementations, which could be optimised, the runtime is on average 8 times greater for the moving-sum model than for the standard changepoint model, but this could greatly vary for different implementations. Here it is of interest to demonstrate how runtimes are affected by different model parameters. Figure 4.8 displays the average runtime of an iteration of the MCMC algorithm for a collection of scenarios for the moving-sum changepoint model. The runtime increases linearly with the number of observations T and with the expected order of dependence $\bar{v}$ for each segment, but

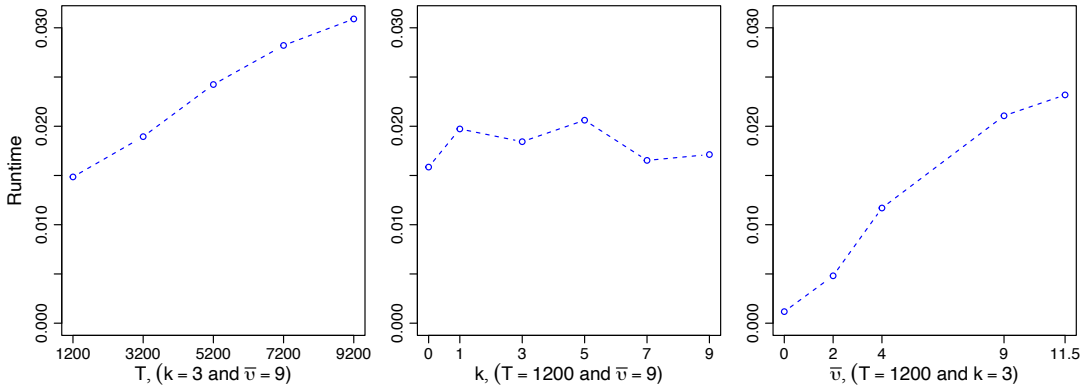


Figure 4.8.: *Average runtime in seconds per iteration for the MCMC algorithm for the moving-sum changepoint model for a collection of scenarios with varying number of observations T, number of changepoints k and expected order of dependence $\bar{v}$ for each segment.*

does not tend to increase with the number of changepoints k. Hence, settings with large segments where the order of dependence is high may be computationally demanding.

4.5.5. Results

Figure 4.9 displays the average $F1$ score of the changepoint models of interest for different scenarios considered in the study. The corresponding average precisions and recalls are displayed in Figure A.5 and Figure A.6. For each scenario, the moving-sum model outperforms the other changepoint models of interest. Similar results for each ρ demonstrate the model is robust to the choice of ρ. Results show that a decrease in performance can be observed when μ is low, that is when the magnitude of mean changes is small; when σ_0 is low, meaning the variance of the data is large; and when υ is low, that is when the mean and variance of orders of dependence is high.

The standard changepoint model and PELT assume data are exchangeable within segments, and consequently their segmentations of the simulated data are not satisfactory. In particular, $F1$ scores decrease as the orders of dependence increase because more changepoints are fitted than necessary.

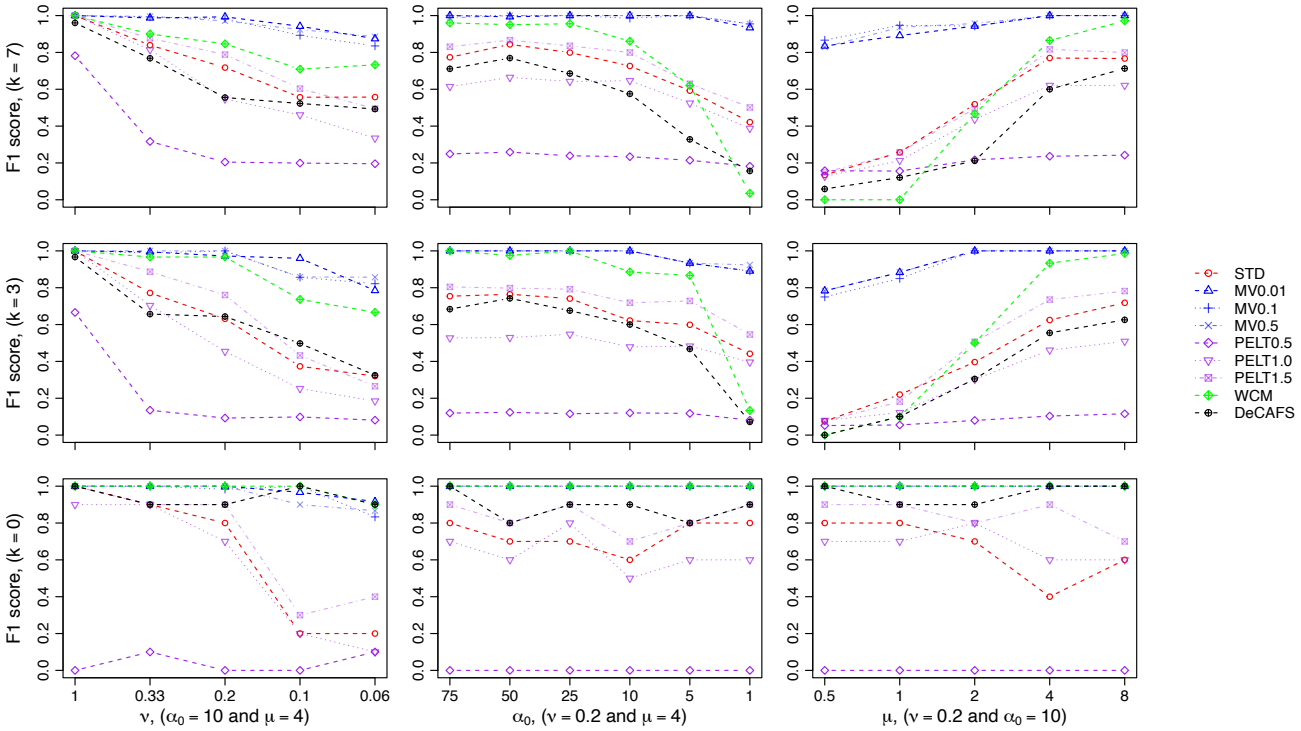


Figure 4.9.: *Average $F1$ scores of changepoint estimates for the changepoint models of interest (STD for the standard changepoint model* (2.4)*; MVρ for the moving-sum changepoint model with hyperparameter ρ for orders of dependence; PELTϕ for PELT with MBIC penalty scaled by ϕ; WCM for WCMgSa; and DeCAFS) for a collection of scenarios considered in the study.*

For PELT, as the penalty increases, $F1$ scores tend to increase because more evidence is required from the data to flag changepoints. For the standard changepoint model, $F1$ scores are satisfactory when $\upsilon = 1$, that is when simulated data are exchangeable within segments.

The performances of DeCAFS and WCMgSA, which admit data may not be exchangeable within segments, tend to be relatively good when $k = 0$, but they tend to deteriorate as the number of segments increases. In particular, for $k > 0$, F1 scores decrease as υ or σ_0 decrease, because DeCAFS and WCMgSA assume that the dependence structure of the data is the same for each segment and that the mean of the data, but not their variance, change from one segment to the next.

Results demonstrate the moving-sum changepoint model may capture various temporal dynamics of m-dependent data within segments that challenge the other changepoint models considered in the simulation study.

4.6. Applications

Two applications are considered to demonstrate the benefits of the proposed changepoint model for non-exchangeable data: computer network monitoring via change detection in count data, and detection of breaks in daily prices of a stock.

For each application, the moving-sum changepoint model is compared with the standard changepoint model for exchangeable data within segments (2.4). For each model, MAP changepoint estimates are derived from a sample of size $50\,000$ obtained via the MCMC algorithm proposed in Section 4.4, with a burn-in of $10\,000$ iterations. Moreover, since DeCAFS is not suitable for count data (Romano et al., 2021), the moving-sum model is compared with DeCAFS for the daily stock prices only.

4.6.1. Enterprise-wide computer network traffic monitoring

A cyber attack typically changes the behaviour of the target network. Therefore, to detect the presence of a network intrusion, it can be informative to monitor for temporal changes in counts of network events.

Consider the network event data recorded at LANL, which were presented in Section 1.1.2. It is of interest to detect temporal changes in the distribution of the data $x_{1:T} = (x_1, \ldots, x_T)$ where, for all t, x_t denotes the number of network events that are in progress across the network during the t-th second between 10:00 and 10:20 on day 22 of the data collection period. The data are displayed in Figure 4.10. It is of interest to detect temporal changes in the distribution of $x_1, \ldots, x_T$.

Two models are used to estimate changepoints for the network data: the moving-sum and the standard changepoint model. For each changepoint model, the segment model for negative binomial data defined in Example 4.3 is assumed with $\alpha = 3$, $\beta = 1$ and $r = 3000$. For each segment, $m \sim \text{Geometric}(0.1)$ for the moving-sum model, and $m = 0$ for the standard changepoint model.

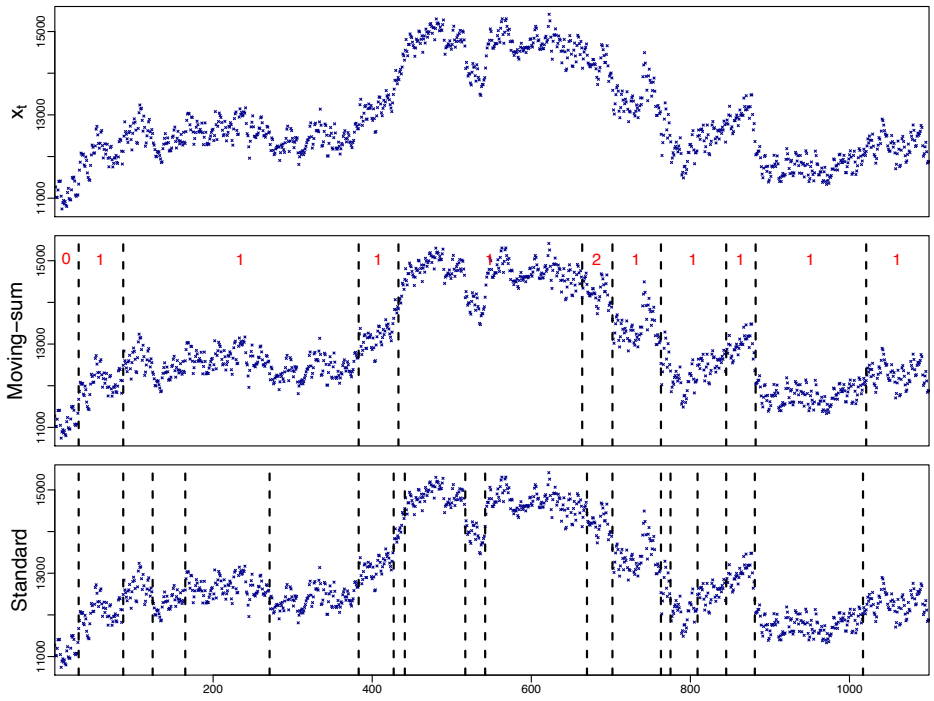


Figure 4.10.: *Counts of computer network events each second between 10:00 and 10:20 on day* 22 *of the data collection period. Vertical lines indicate estimated changepoints for the proposed changepoint model (middle panel) and for the standard changepoint model (Fearnhead, 2006) (bottom panel). Numbers in red indicate the MAP order of dependence* m *for each segment for the moving-sum model.*

Figure 4.10 displays the changepoint estimates for each model, and the MAP segment orders of dependence for the moving-sum changepoint model. No labels are available to indicate the optimal segmentation for these data, but we argue that the moving-sum model gives a more suitable segmentation of the data than the standard changepoint model. Both changepoint models detect clear discontinuities, such as the ones observed near the 420th and 880th seconds, that correspond to behavioural changes of the network that are sufficiently drastic that cyber-analysts may want to investigate whether they correspond to malicious activities. However, local temporal correlations and small fluctuations, such as the ones between the 100th and 380th seconds, and the one between the 500th and 550th seconds, which is characterised by a small decrease quickly followed by a symmetric increase in the mean, give rise to changepoints for the standard model but not the moving-sum changepoint model. Such local temporal correlations and small fluctuations may be caused by normal temporal dynamics of the network behaviour that are less of interest to cyber-analysts, for example standard transfers of large data across networks that involve dividing the data into batches, routing and queueing steps, which may lead to temporal correlations in traffic volumes. Hence, the proposed changepoint model results in a segmentation of the data that is more

adapted to network monitoring.

4.6.2. Financial time series segmentation

For economists and investors, it can be of interest to detect changepoints in financial time series, such as daily prices of a stock. Changes can be monitored to assess the impact of economic policies, or can indicate shifts in market behaviours. It is often reductive to assume the data are exchangeable within segments.

For demonstration purposes, this chapter considers the price of the stock Sainsbury plc (SBRY.L). For all t, let x_t be the closing price of the stock SBRY.L at the t-th day between 20$^{\text{th}}$ February 2006

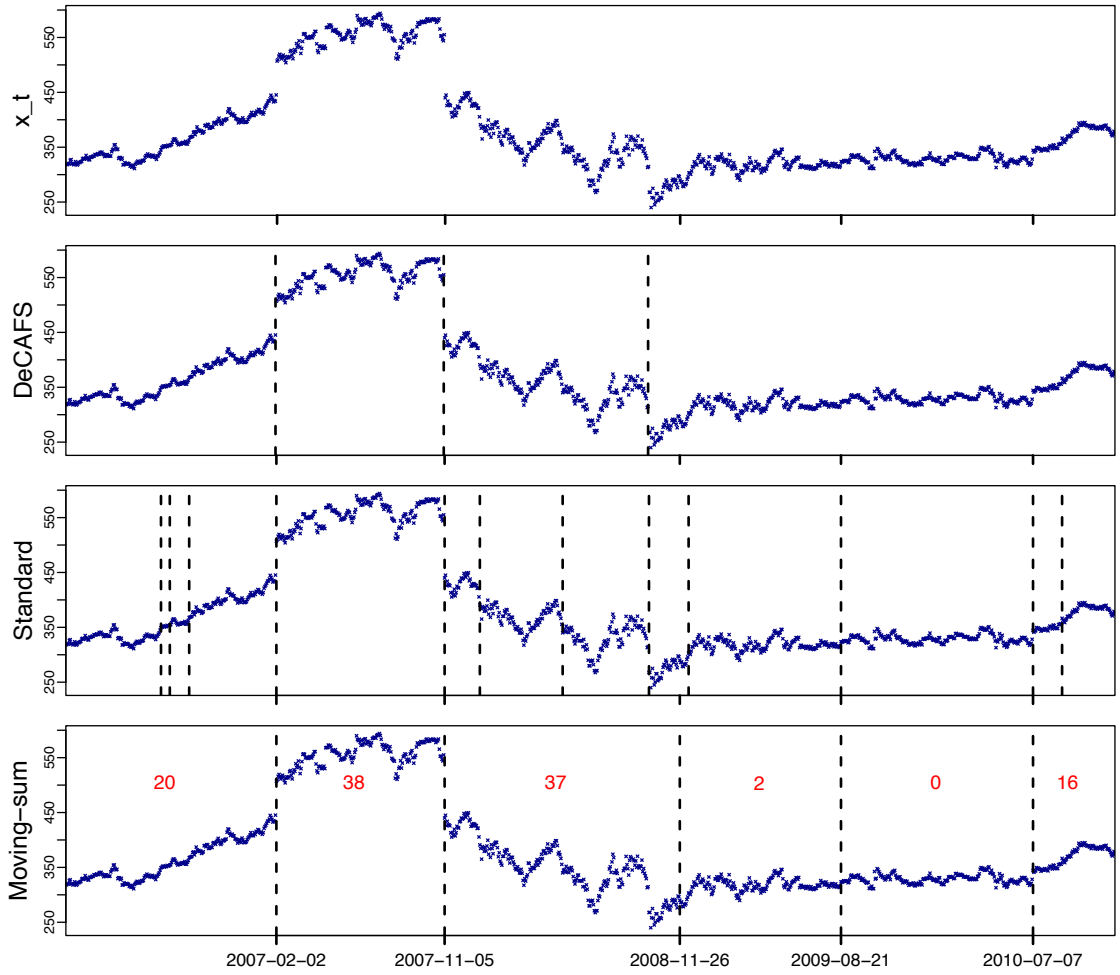


Figure 4.11.: *Daily prices of the stock SBRY.L from* 20^{th} *February 2006 to* 16^{th} *November 2010. Vertical lines indicate estimated changepoints for three models: moving-sum, standard and DeCAFS changepoint models. Numbers in red indicate the MAP order of dependence* m *for each segment for the moving-sum changepoint model.*

to 16th November 2010. The data $x_1, \ldots, x_T$, which are available online at `finance.yahoo.com`, are displayed in Figure 4.11. The stock data are subject to changes that may be associated with takeover bids and the 2008 financial crisis: on 2nd February 2007 major private equity groups reveal they are considering a bid for Sainsbury plc; on 5th November 2007 takeover bids are abandoned in the wake of the global credit crunch; November 2008 ends a period of major uncertainty in financial markets; in August 2009 major decisions to develop the business are announced, including acquisitions of new shops and important increase in workforce; and, on 7th July 2010 a possible takeover bid is announced by Qatari investors.

Three models are considered for the data: the moving-sum changepoint model, the standard changepoint model, and DeCAFS using an inflated BIC beta penalty as suggested in Romano et al. (2021). For the moving-sum and the standard changepoint models, the segment model for normal data defined in Example 4.1 is assumed with $\alpha = 7$, $\beta = 1$, $\mu_0 = 350$ and $\lambda = 1$. For each segment, $m \sim \text{Geometric}(0.1)$ for the moving-sum model, and $m = 0$ for the standard changepoint model.

Figure 4.11 displays the changepoint estimates for each model of interest. All models detect changes associated with takeover bids on 2nd February 2007 and 5th November 2007. In contrast with the standard changepoint model, the moving-sum changepoint model captures temporal dynamics of the stock price, and therefore market trends are not unnecessarily segmented. For example, smooth drifts of the stock price, such as the ones before 2nd February 2007 and between 5th November 2007 and 26th November 2008, give rise to multiple changepoints for the standard model but not for the moving-sum model. Moreover, the proposed model detects changes in the mean as well as changes in the level of dependency of the data, so that it detects both shifts in price levels and changes in the temporal dynamics of prices, which are both of interest to financial analysts. In contrast with DeCAFS, the moving-sum changepoint model detects the change associated with the takeover bid on 7th July 2017, and perhaps less importantly, the change in August 2009 associated with business developments after the financial crisis. The estimated orders of dependence for the moving-sum model vary greatly across segments, suggesting prices have been subject to distinct market dynamics. In particular, the order of dependence decreases drastically in November 2008, marking the end of strong market dynamics associated with the 2008 financial crisis, and it increases again in July 2017 as the announcement of a takeover bid leads to increased market activity. MAP orders of dependence obtained via MCMC were validated with the method proposed in Section 4.4.4 conditional on MAP changepoint parameters obtained via MCMC: estimates are identical for each segment except for the second one where the importance sampling approach gives 39 instead of 38.

4.7. Conclusion

This chapter proposed a novel segment model for changepoint models that is motivated by temporal dynamics observed in cyber data. Cyber data can be subject to gradual population changes, seasonal variations and other temporal trends that are unlikely to be evidence for malicious activity. However, standard changepoint models, which typically assume data are exchangeable within each segment, may be inadequate when data are subject to local correlation patterns, often resulting in many more changepoints being fitted than is preferable. This chapter proposed a Bayesian changepoint model that relaxes the assumption of exchangeability within segments. The proposed model supposes that within each segment the data are m-dependent and identically distributed conditional on some parameter θ, where both θ and $m \geqslant 0$ are unknown and vary between segments, resulting in a model suitable for detecting clear discontinuities in data that are subject to different local temporal correlations. A novel reversible jump MCMC algorithm was proposed to sample from the model; in particular, a detailed analysis of the parameter space was exploited to build proposals for the orders of dependence. The benefits of the proposed model were demonstrated with real network event data recorded at LANL.

Future work could consider a model extension relaxing the assumption that both m and θ change at each changepoint; changepoint parameters could be augmented with indicator vectors specifying for each changepoint if only m, only θ or both m and θ change. The MCMC algorithm would need to be adapted to sample indicator vectors alongside changepoint parameters.

In the next chapter, extending the standard changepoint model, we propose a novel model relaxing the assumption that changes occur instantaneously at changepoints, motivated by cyber data where changes are gradual over a time interval.

5

Change interval detection

As discussed in Chapter 2, standard changepoint models for a single time series suppose that, at each changepoint, the distribution of the data undergoes a sudden change. In particular, for parametric changepoint models, such as the models proposed in Chapter 3 and Chapter 4, it is typically assumed that the distribution of the data within each segment has the same functional form dependent on some unknown parameter that changes *abruptly* from one segment to the next.

This construction is adequate for various applications, such as the segmentation of well-log data segmentation (Fearnhead and Rigaill, 2019) or the detection of copy number variants in DNA sequences (Bardwell and Fearnhead, 2017), where changes may be assumed to occur instantly at changepoints. However, such an assumption on the abruptness of changes can be limiting for some cyber-security applications, where each change in the distribution of data can occur over an interval of time. For example, as discussed in Section 1.1.2, volumes of traffic passing through the entire network over time may be impacted by gradual changes, as attackers traverse the network, progressively taking control of more computers. Network entities are often subject to similar behavioural changes, such as reduced activity during lunch time, but at slightly different times, resulting in gradual behavioural changes for the entire network. For demonstration purposes, consider the second-by-second counts of computer network events recorded at LANL for the anonymised destination port 'Port20990', which are defined in Example 1.2. Suppose the data $x_1, \ldots, x_t$ follow the standard changepoint model (2.4) where the parametric density $f(\cdot \mid \theta)$ corresponds to the Poisson distribution with unknown mean $\theta \sim \Gamma(1, 0.1)$. Figure 5.1 displays the network data and the MAP changepoints obtained via the standard RJMCMC algorithm given in Section 2.3. It is apparent that multiple changepoints are fitted in time intervals where the mean of the data gradually changes. To characterise the changes impacting the distribution of the data, it would be preferable to identify time intervals where changes in the mean gradually occur. More generally, gradual changes in cyber data may also begin, or end, with a clear discontinuity. For

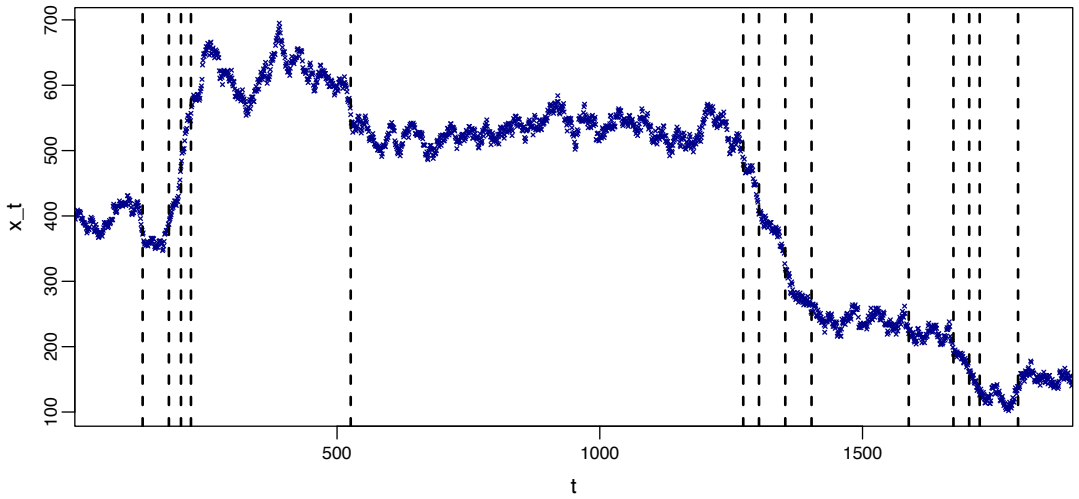


Figure 5.1.: *Counts of computer network events for the anonymised destination port 'Port20990' each second between 11:40 and 12:30 on day* 22 *of the data collection period. Vertical lines indicate MAP changepoints for the standard changepoint model* (2.4).

example, attackers may start to exfiltrate sensitive data from multiple compromised network entities at the same time, and then progressively end command-and-control for each network entity when data have been fully exfiltrated, resulting a clear discontinuity in traffic volumes followed by a gradual transition to a post-attack stationary regime. Moreover, it is of interest to build models that are suitable for count data, which are often encountered in cyber security applications as discussed in Section 1.1.2.

Modelling non-instantaneous changes occurring in time series has received limited attention in the changepoint modelling literature. Existing changepoint models that relax the assumption changes are sudden typically require strong assumptions on gradual transitions between stationary regimes. For example, motivated by an application in medicine, Gao et al. (2019) propose a model where the variance of data may be subject to significant jumps while the mean changes in a smooth fashion. Herlands et al. (2019) require the prior specification of a deterministic function characterising transitions between regimes. These assumptions are too restrictive for cyber data where transitions between regimes may take various forms. Moreover, existing approaches typically identify changepoints where the distribution starts to change but cannot identify when the transition has ended, which may represent valuable information for forensic investigation in cyber data. For example, Vogt and Dette (2015) provide a flexible approach based on a nonparametric method to detect only the start of smooth changes in a locally stationary processes. Other approaches that can be considered to segment data subject to non-instantaneous changes include the detection of changepoints in smooth curves (Gijbels, Hall, and Kneip, 1999; Loader, 1996; Grégoire and Hamrouni, 2002), or the detection of abrupt changes in normally distributed data with local fluctuations and autocorrelated noise (Romano et al., 2021), but these approaches are seldom

suitable for count data, which are often encountered in cyber security, and they do not assume stationary regimes for data between changes.

This chapter reports an initial exploration of modelling the gradual changes that may be observed in cyber data, and we hope it encourages further research in this direction. An extension of the standard parametric changepoint model (2.4) is proposed, that relaxes the assumption that changes occur abruptly at changepoints. Within each segment the data has the same functional form dependent on some unknown parameter that may change instantaneously or over some time interval from one segment to the next. Changepoint parameters are augmented with time window parameters, thereby defining change intervals of possibly unequal lengths: each changepoint location identifies the beginning of a time interval, where a segment parameter transitions to the next segment parameter via a random walk, which is not necessarily smooth and may begin or end with a clear discontinuity, thereby encompassing various forms of transitions. Note that change interval lengths may be equal to 0; zero-length change intervals indicate the locations of abrupt changes. The model is suitable for both continuous and count data.

The parameters of interest are the change interval parameters, consisting of changepoint parameters and change interval length parameters. The RJMCMC algorithm sampling changepoint parameters discussed in Section 2.3 is extended to sample change interval parameters. At each iteration of the algorithm, one of the following move types is proposed: sample a segment parameter, shift an existing changepoint to a new position, propose to add or remove a zero-length change interval, propose to update change interval length parameters, split an existing change interval, or merge two existing change intervals. For demonstration purposes, conjugacy of the prior distribution of segment parameters will be assumed (2.4), but the sampler could be adapted to settings with non-conjugate prior distributions; see Section 2.2 for a discussion on Monte Carlo integration methods based on sequential importance sampling (SIS). The proposed RJMCMC algorithm relies on a SIS based approximation to the conditional likelihood of the data within change intervals.

The remainder of the chapter is organised as follows. Section 5.1 defines the change interval model. Section 5.2 proposes an approach to compute the posterior distribution of change interval parameters. Section 5.3 gives an RJMCMC algorithm to sample from the proposed model. Section 5.4 discusses simulation results that demonstrate the benefits of the model, in comparison to standard changepoint models. Section 5.5 illustrates the practical benefits of the change interval model using the LANL computer network data introduced in Section 1.2. Section 5.6 concludes.

5.1. From changepoints to change intervals

This section proposes an extension of the standard changepoint model that relaxes the assumption that changes occur abruptly at changepoints. Changepoint parameters are augmented with time

window parameters, thereby defining change intervals where the distribution of the data transitions from one segment to the next. The concept of change intervals has the advantage to explicitly characterise both the beginning and the end of non-instantaneous changes in contrast with models that focus solely on the beginning of changes, such as the model proposed in Vogt and Dette (2015). Various approaches are possible to model data within change intervals. This section proposes one approach that is motivated by the cyber application presented in the introduction of this chapter, and Section 6.2 outlines other possible approaches, hoping it encourages further research in this direction.

5.1.1. Model definition

Let $x_{1:T} = (x_1, \ldots, x_T)$ denote observed data. Recall the standard changepoint model (2.4) supposes that there are $k \geqslant 0$ changepoints, whose ordered positions are denoted by $\tau_{1:k} = (\tau_1, \ldots, \tau_k) \in \mathcal{T}_k$, that partition the time series into $k+1$ segments. Within each segment j, the data are assumed to be iid with density function $f(\cdot \mid \theta_j)$ conditional on some parameter θ_j that changes abruptly from one segment to the next. Note that $\tau_0 = 1$ and $\tau_{k+1} = T + 1$.

The extended model proposed in this section relies on introducing the concept of change intervals. The proposed model supposes, for each $j = 1, \ldots, k$, an interval length parameter $w_j \geqslant 0$ is associated to the changepoint location τ_j, constrained such that

$$\tau_j + w_j < \tau_{j+1}. \tag{5.1}$$

Set $w_0 = w_{k+1} = 0$. For each segment j, which corresponds to the time points $\tau_{j-1}, \ldots, \tau_j - 1$, the pair (τ_{j-1}, w_{j-1}) defines a change interval

$$I_j = \{t \mid \tau_{j-1} \leqslant t < \tau_{j-1} + w_{j-1}\},$$

that indicates time points where the distribution of the data progressively transitions between the distributions corresponding to segment $j-1$ and segment j, as described below. The standard prior distribution (2.2) for changepoint parameters $(k, \tau_{1:k})$ is assumed, and, conditional on $(k, \tau_{1:k})$, interval length parameters $w_{1:k} = (w_1, \ldots, w_k)$ are assumed to be independent with prior distribution

$$\pi(w_{1:k} \mid k, \tau_{1:k}) \propto \prod_{j=1}^{k} \rho(1-\rho)^{w_j} \mathbf{1}_{\{0,1,\ldots,\tau_{j+1}-\tau_j-1\}}(w_j), \tag{5.2}$$

for some $0 < \rho \leqslant 1$, meaning each interval length may be increased or decreased at a fixed cost subject to the constraints (5.1).

For each segment j, in addition to the independent segment parameter θ_j with prior distribution

$\pi(\cdot)$, independent segment parameters γ_j and h_j are introduced to specify the distribution of the data in I_j: if $w_{j-1} > 0$ then $\gamma_j \sim \text{Bernoulli}(\xi)$, for some $0 < \xi < 1$, and $h_j \sim \text{Beta}(a, b)$, for some $a, b > 0$, and if $w_{j-1} = 0$ then $\gamma_j = h_j = 0$. Let $\pi(\cdot \mid w_{j-1})$ denote both the prior distribution of γ_j and h_j.

The model assumes that, independently for all $t = 1, \ldots, T$,

$$x_t \sim f(\cdot \mid \bar{\theta}_t) \tag{5.3}$$

given some sequence of latent variables $\bar{\theta}_{1:T} = (\bar{\theta}_1, \bar{\theta}_2, \ldots, \bar{\theta}_T)$, such that, within each segment j,

$$\begin{cases} \bar{\theta}_t = \theta_j & \text{for all } t = \tau_{j-1} + w_{j-1}, \ldots, \tau_j - 1, \\ \bar{\theta}_t \sim N(\mu_t, \sigma_j^2) & \text{for all } t \in I_j, \end{cases} \tag{5.4}$$

where $\sigma_j = \sigma|\theta_j - \theta_{j-1}|$ for some $\sigma > 0$, and

$$\mu_t = \begin{cases} \bar{\theta}_{t-1} + h_j \delta_j & \text{if } \gamma_j = 1, \\ \bar{\theta}_{t+1} - h_j \delta_j & \text{otherwise,} \end{cases}$$

where $\delta_j = (\theta_j - \theta_{j-1})/(w_j + 1)$. Note that zero-length change intervals indicate the locations of abrupt changes. In particular, if $\rho = 1$ then (5.2) imposes $w_j = 0$ for all j, and the change interval model is the standard changepoint model (2.4).

The transition model characterised by (5.4) assumes segment parameters lie in $\mathbb{R}$. For settings where the domain of the segment parameters is a subset of $\mathbb{R}$, it would be possible to consider a random walk truncated to the segment parameter space, or a random walk of a transformation of the segment parameters, for example $\log\{\bar{\theta}_t/(1 - \bar{\theta}_t)\}$ if the segment parameter space is $[0, 1]$. In practice, it is also possible to use (5.4) when the probability mass for segment parameters in $\mathbb{R}$ outside the segment parameter space is negligible; for example, in this chapter, (5.4) is used for the mean of large count data.

5.1.2. Simulation from the model

Simulations from the proposed model are discussed to demonstrate the role of change intervals and segment parameters. Figure 5.2 displays simulations from the proposed model with $f(\cdot \mid \theta)$ corresponding to the Poisson distribution with unknown mean parameter $\theta \sim \Gamma(10, 0.1)$. Consider a change between segment $j - 1$ and segment j. It is apparent that the change is abrupt if $w_{j-1} = 0$, but occurs progressively over a change interval I_j if $w_{j-1} > 0$. Moreover, if $\gamma_j = 0$ then there is no clear discontinuity between segment $j - 1$ and change interval I_j but there can be an abrupt discontinuity from change interval I_j to segment j; note that the smaller $0 < h_j < 1$, the greater is the discontinuity. However, if $\gamma_j = 1$ then the location of the discontinuity is reversed. Finally, the

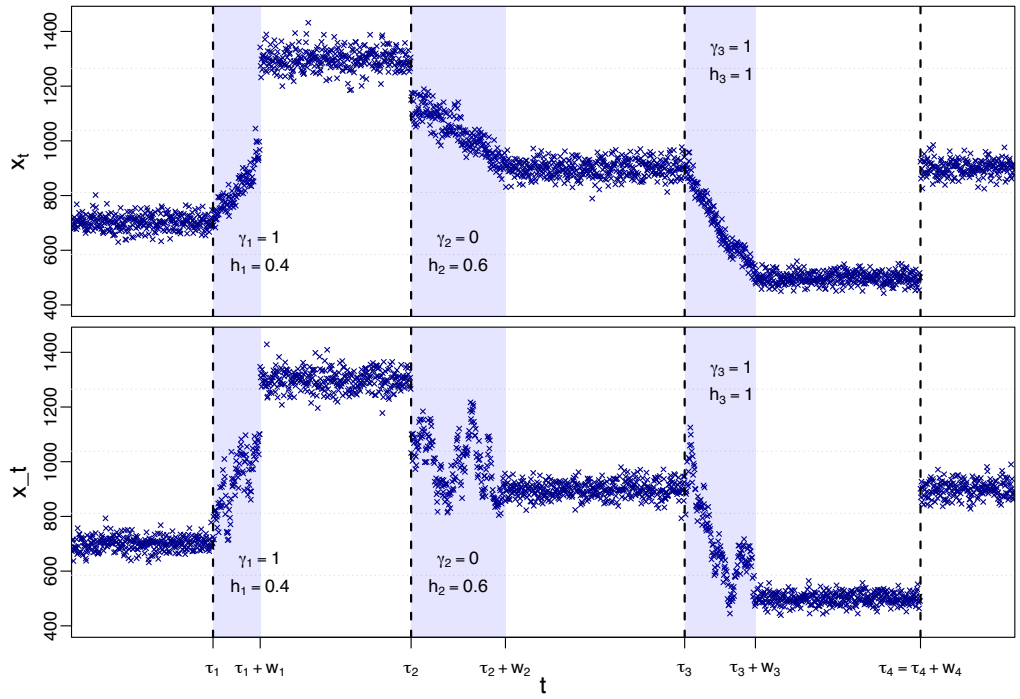


Figure 5.2.: *Data generated from the proposed model for Poisson data with four changepoints, with positions* τ_1, τ_2, τ_3 *and* τ_4 *indicated by dashed lines, and interval lengths* $w_1, w_2, w_3 > 0$ *and* $w_4 = 0$ *specifying change intervals indicated by blue rectangles. Top:* $\sigma = 0.1$. *Bottom:* $\sigma = 0.4$

smaller σ, the smoother is the change in the change interval I_j.

Figure 5.3 displays MAP changepoints for the standard changepoint model, which assumes $\rho = 0$ so that $w_j = 0$ for all j, obtained via the RJMCMC algorithm discussed in Section 2.3. More changepoints are fitted than preferable in change intervals because the standard changepoint model assumes abrupt changes between segments.

5.2. Posterior distribution

The parameters of interest are the changepoint parameters $(k, \tau_{1:k})$ and the interval length parameters $w_{1:k}$, that define change interval parameters $(k, \tau_{1:k}, w_{1:k})$. Let $\psi_j = (\theta_j, \gamma_j, h_j)$ for all j, so that $\psi_{1:(k+1)}$ denote the segment parameters. The joint posterior distribution of the model

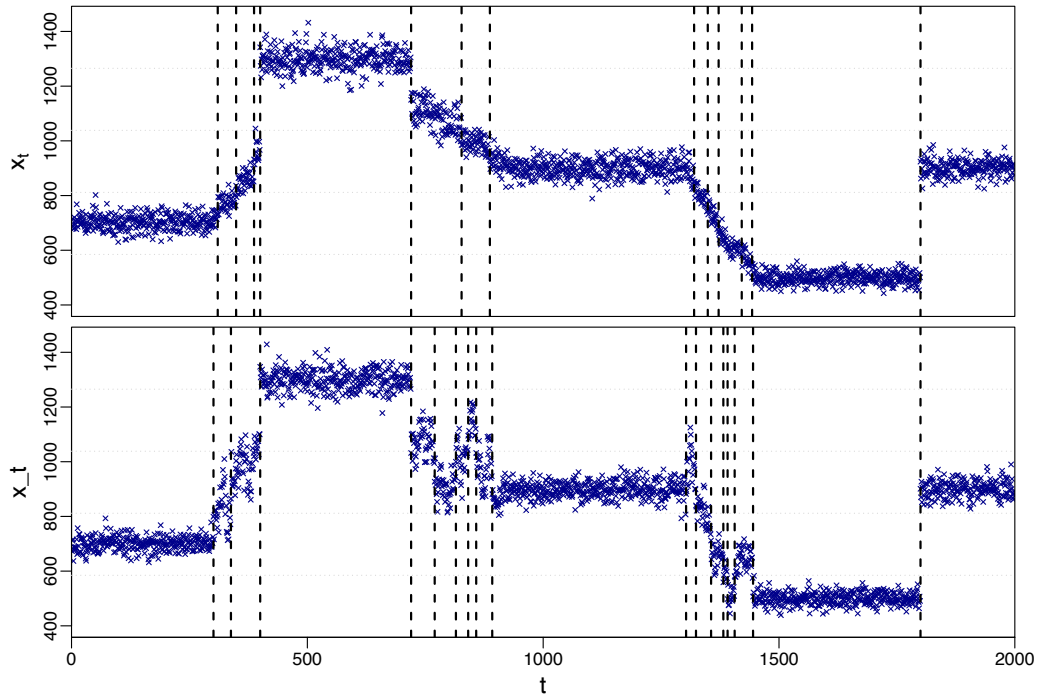


Figure 5.3.: *Data generated from the proposed model for Poisson data displayed in Figure 5.2. The black dashed lines indicate the positions of the MAP changepoints obtained by fitting the standard changepoint model with abrupt changes to the data.*

parameters is such that, for all $k \geqslant 0$, $\tau_{1:k} \in \mathcal{T}_k$, $w_{1:k}$ satisfying (5.1), and $\psi_{1:(k+1)}$,

$$\pi(k, \tau_{1:k}, w_{1:k}, \psi_{1:(k+1)} \mid x_{1:T}) = \pi(k, \tau_{1:k})\pi(w_{1:k} \mid k, \tau_{1:k}) \prod_{j=1}^{k+1} \pi(\theta_j)\pi(h_j \mid w_{j-1})\pi(\gamma_j \mid w_{j-1})$$

$$\times \prod_{j=1}^{k+1} L(x_{\tau_{j-1}:(\tau_j-1)} \mid w_{j-1}, \theta_{j-1}, \theta_j, h_j, \gamma_j),$$

where $L(x_{\tau_{j-1}:(\tau_j-1)} \mid w_{j-1}, \theta_{j-1}, \theta_j, h_j, \gamma_j)$ denotes the conditional likelihood of the data within segment j. Segment parameters $\psi_{1:(k+1)}$ cannot be marginalised in general.

Within segment j, the conditional likelihood of the data is

$$L(x_{\tau_{j-1}:(\tau_j-1)} \mid w_{j-1}, \theta_{j-1}, \theta_j, h_j, \gamma_j) = \upsilon(x_{I_j}) \left(\prod_{t=\tau_{j-1}+w_{j-1}}^{\tau_j-1} f(x_t \mid \theta_j) \right),$$

where $\upsilon(x_{I_j}) \equiv \upsilon(x_{I_j} \mid w_{j-1}, \theta_{j-1}, \theta_j, h_j, \gamma_j)$ denotes the joint density of the data $x_{I_j} = (x_t)_{t \in I_j}$

within the change interval I_j. It follows from (5.4) that

$$\upsilon(x_{I_j}) = \int \upsilon(x_{I_j}, \bar{\theta}_{I_j}) d\bar{\theta}_{I_j}, \tag{5.5}$$

where

$$\upsilon(x_{I_j}, \bar{\theta}_{I_j}) = \prod_{t \in I_j} f(x_t \mid \bar{\theta}_t) \phi(\bar{\theta}_t \mid \mu_t, \sigma_j^2), \tag{5.6}$$

with $\bar{\theta}_{I_j} = (\bar{\theta}_t)_{t \in I_j}$ and $\phi(\cdot \mid \mu, \sigma^2)$ denoting the density of the normal distribution with mean μ and variance $\sigma^2 > 0$.

The integral (5.5) is not analytically tractable in general, but Monte Carlo integration is possible. See Section 2.2 for a brief presentation of standard Monte Carlo integration methods. In particular, sequential importance sampling from the posterior density of $\bar{\theta}_{I_j}$,

$$\upsilon(\bar{\theta}_{I_j} \mid x_{I_j}) = \frac{\upsilon(x_{I_j}, \bar{\theta}_{I_j})}{\upsilon(x_{I_j})}, \tag{5.7}$$

can be performed since the joint density (5.6) is tractable and has a sequential structure, so that an estimate for the marginal density (5.5) can be computed, as described in Section 2.2.3. To generate N samples $\left(\bar{\theta}_{I_j}^{(i)}\right)_{i=1,\ldots,N}$ from (5.7) via sequential importance sampling, we use the proposal joint density

$$q(\bar{\theta}_{I_j} \mid x_{I_j}) = \prod_{t \in I_j} \phi(\bar{\theta}_t \mid \tilde{\sigma}_j^2 \tilde{\mu}_t, \tilde{\sigma}_j^2),$$

where $\tilde{\mu}_t = \frac{\mu_t}{\sigma_j^2} + \frac{x_t}{\nu_j^2}$, $\tilde{\sigma}_j^2 = \left(\frac{1}{\sigma_j^2} + \frac{1}{\nu_j^2}\right)^{-1}$ and ν_j^2 denotes the empirical variance of $x_{\tau_{j-1}+w_{j-1}:\tau_j-1}$. The proposed algorithm to obtain an approximation of the marginal density (5.5) of x_{I_j} is provided in Algorithm 5.1.

5.3. Reversible jump MCMC

The standard reversible jump MCMC algorithm for changepoint models discussed in Section 2.3 is extended to sample from the proposed change interval model. To explore the parameter space, in addition to moves proposing to update segment parameters, to shift the location of a changepoint, to delete or to add a changepoint (that is, a zero-length change interval), we consider moves to update interval length parameters and to merge or split change intervals.

Suppose the latest particle of the sample chain consists of $k \geqslant 0$ changepoints with ordered locations $\tau_{1:k} \in \mathcal{T}_k$, interval lengths $w_{1:k}$ satisfying (5.1), and segment parameters $\theta_{1:(k+1)}$, $\gamma_{1:(k+1)}$

Algorithm 5.1: Sequential importance sampling based approximation of $\upsilon(\bar{\theta}_{I_j} \mid x_{I_j})$

Input: Data x_{I_j}, number of samples N for the SIS, segment parameters
Result: SIS based estimator $\hat{\upsilon}^N(x_{I_j})$ of the marginal density (5.5) of x_{I_j}

Step 1: Obtain N samples $\left(\bar{\theta}_{I_j}^{(i)}\right)_{i=1,\ldots,N}$ from density (5.7) and weights $\left(\omega_{I_j}^{(i)}\right)_{i=1,\ldots,N}$
Initialisation: for all $i = 1, \ldots, N$, $\bar{\theta}_{\tau_{j-1}-1}^{(i)} = \theta_{j-1}$ and $\bar{\theta}_{\tau_{j-1}+w_{j-1}}^{(i)} = \theta_j$
for all t, from τ_{j-1} to $\tau_{j-1} + w_{j-1} - 1$ if $\gamma_j = 1$, and from $\tau_{j-1} + w_{j-1} - 1$ to τ_{j-1} if $\gamma_j = 0$ **do**
  for $i = 1, \ldots, N$ **do**
    if $\gamma_j = 1$ **then** $\mu_t^{(i)} = \bar{\theta}_{t-1}^{(i)} + h_j\delta_j$ **else** $\mu_t^{(i)} = \bar{\theta}_{t+1}^{(i)} - h_j\delta_j$.
    set $\tilde{\mu}_t^{(i)} = \frac{\mu_t^{(i)}}{\sigma_j^2} + \frac{x_t}{\nu_j^2}$
    sample $\bar{\theta}_t^{(i)} \sim \phi(\cdot \mid \tilde{\sigma}_j^2\tilde{\mu}_t^{(i)}, \tilde{\sigma}_j^2)$
    set

$$\omega_t^{(i)} = \frac{f(x_t \mid \bar{\theta}_t^{(i)})\phi(\bar{\theta}_t^{(i)} \mid \mu_t^{(i)}, \sigma_j^2)}{\phi(\bar{\theta}_t^{(i)} \mid \tilde{\sigma}_j^2\tilde{\mu}_t^{(i)}, \tilde{\sigma}_j^2)}$$

  end
end
Step 2: Compute the estimator of the marginal density (5.5) of x_{I_j},

$$\hat{\upsilon}^N(x_{I_j}) = \frac{1}{N}\sum_{i=1}^{N}\prod_{t\in I_j}\omega_t^{(i)}. \tag{5.8}$$

and $h_{1:(k+1)}$. The next element of the chain is proposed via one of the following moves.

5.3.1. Update segment parameters

To update segment parameters conditional on change interval parameters, two different moves are used: a move to update $\theta_{1:(k+1)}$, and a move to update $\gamma_{1:(k+1)}$ and $h_{1:(k+1)}$ conditional on $\theta_{1:(k+1)}$.

5.3.1.1. Update $\theta_{1:(k+1)}$ conditional on change interval parameters

Draw an index j uniformly from $\{1, \ldots, k+1\}$, and propose to sample $\theta_j \in \Theta$ from the distribution with density

$$q(\theta_j \mid k, \tau_{1:k}, w_{1:k}, x_{1:T}) \propto \pi(\theta_j)\prod_{t=\tau_{j-1}+w_j}^{\tau_j-1} f(x_t \mid \theta_j), \tag{5.9}$$

which is tractable by conjugacy of the prior distribution of θ_j. Note that (5.9) is the full conditional distribution of θ_j if $w_{j-1} = w_j = 0$, and an approximation to the full conditional distribution of θ_j otherwise.

5.3.1.2. Update $\gamma_{1:(k+1)}$ and $h_{1:(k+1)}$ conditional on change interval parameters and $\theta_{1:(k+1)}$

An index j is uniformly chosen from $\{j \mid 1 \leqslant j \leqslant k+1, w_{j-1} > 0\}$, and it is proposed to update γ_j and h_j conditional on the remaining segment parameters. Motivated by the observation that if $x_{\tau_{j-1}} - x_{\tau_{j-1}-1} > x_{\tau_{j-1}+w_{j-1}} - x_{\tau_{j-1}+w_{j-1}-1}$ then it is reasonable to believe that $\gamma_j = 1$ is more likely than $\gamma_j = 0$, propose

$$\gamma_j \mid x_{1:T}, k, \tau_{1:k}, w_{1:k} \sim \text{Bernoulli}\left(\frac{x_{\tau_{j-1}} - x_{\tau_{j-1}-1}}{x_{\tau_{j-1}} - x_{\tau_{j-1}-1} + x_{\tau_{j-1}+w_{j-1}} - x_{\tau_{j-1}+w_{j-1}-1}}\right).$$

Then, conditional on the updated γ_j, propose

$$h_j \mid \gamma_j, x_{1:T}, k, \tau_{1:k}, w_{1:k} \sim \text{Beta}\left(\alpha + 1, \beta + \max\{0, q_j\}\right),$$

where

$$q_j = \begin{cases} \frac{x_{\tau_{j-1}+w_{j-1}-1} - x_{\tau_{j-1}}}{x_{\tau_{j-1}+w_{j-1}} - x_{\tau_{j-1}}} & \text{if } \gamma_j = 1 \\ \frac{x_{\tau_{j-1}+1} - x_{\tau_{j-1}+w_{j-1}}}{x_{\tau_{j-1}} - x_{\tau_{j-1}+w_{j-1}}} & \text{if } \gamma_j = 0. \end{cases}$$

5.3.2. Birth and death of zero-length intervals

To propose to delete a zero-length interval, an index j is uniformly chosen from $\{j \mid 1 \leqslant j \leqslant k,\ w_j = 0\}$, τ_j and w_j are removed from $\tau_{1:k}$ and $w_{1:k}$, resulting in $k' = k-1$ changepoints with positions $\tau'_{1:k'} = (\tau_1, \ldots, \tau_{j-1}, \tau_{j+1}, \ldots, \tau_k)$ and interval lengths $w'_{1:k'} = (w_1, \ldots, w_{j-1}, w_{j+1}, \ldots, w_k)$. For segment parameters, h_{j-1} and γ_{j-1} are removed from $h_{1:(k+1)}$ and $\gamma_{1:(k+1)}$, respectively, and we propose $\theta'_{1:(k'+1)} = (\theta_1, \ldots, \theta_{j-2}, \theta'_j, \theta_{j+1}, \ldots, \theta_{k+1})$, where θ'_j is sampled from (5.9).

The above death move can be reversed by a birth move. To propose to add a changepoint, a new position τ' is drawn uniformly from $\{2, \ldots, T\} \setminus \bigcup_{j=1}^{k+1} I_j$, and τ' is inserted into $\tau_{1:k}$, resulting in $k' = k+1$ changepoints with locations $\tau'_{1:k'} = (\tau_1, \ldots, \tau_{j-1}, \tau', \tau_j, \ldots, \tau_k) \in \mathcal{T}_{k'}$ and interval lengths $w'_{1:k'} = (w_1, \ldots, w_{j-1}, 0, w_j, \ldots, w_k)$, where j is the greatest integer such that $\tau' < \tau_j$. Set $\gamma'_{1:(k'+1)} = (\gamma_1, \ldots \gamma_j, 0, \gamma_{j+1}, \ldots, \gamma_{k+1})$, $h'_{1:(k'+1)} = (h_1, \ldots, h_j, 0, h_{j+1}, \ldots, h_{k+1})$, and $\theta'_{1:(k'+1)} = (\theta_1, \ldots, \theta_{j-1}, \theta'_j, \theta'_{j+1}, \theta_{j+1}, \ldots, \theta_{k+1})$ where θ'_j and θ'_{j+1} are sampled using the proposal distribution (5.9) for the segment j and $j+1$ resulting from $(k', \tau'_{1:k'})$.

5.3.3. Update interval lengths

The move proposes to update the interval length associated with a changepoint. Let j be an index uniformly chosen from $\{1, \ldots, k\}$, sample w_j uniformly from $\{0, \ldots, \tau_{j+1} - \tau_j - 1\}$, and sample $\theta_{j+1}, \gamma_{j+1}$ and h_{j+1} using the proposal distributions given in Section 5.3.1.

5.3.4. Shift move

The shift move proposes to modify the position of a changepoint. The index j is uniformly sampled from $\{1, \ldots, k\}$. If $w_j = 0$ then a new changepoint position τ'_j is sampled uniformly from $\{\tau_{j-1} + w_{j-1} + 1, \ldots, \tau_{j+1} - 1\}$, and the segment parameters θ_j, $\theta_{j+1}, \gamma_j, \gamma_{j+1}, h_j$ and h_{j+1} are sampled using the proposal distributions given in Section 5.3.1. If $w_j > 0$ then a new position τ'_j is sampled uniformly from $\{\tau_{j-1} + w_{j-1} + 1, \ldots, \tau_j + w_j - 1\}$, an updated interval length is set $w'_j = w_j + \tau'_j - \tau_j$, leaving the right endpoint of the change interval unchanged, and segment parameters θ_j, γ_j, γ_{j+1}, h_j and h_{j+1} are sampled using the proposal distributions given in Section 5.3.1.

5.3.5. Merge and split of change intervals

A merge move proposes that, given an index j uniformly chosen from $\{1, \ldots, k-1\}$, the change intervals (τ_j, w_j) and (τ_{j+1}, w_{j+1}) are merged into the change interval $(\tau_j, w_{j+1} + \tau_{j+1} - \tau_j)$: τ_{j+1} and w_{j+1} are removed from $\tau_{1:k}$ and $w_{1:k}$, respectively, and we set $w_j = \tau_{j+1} + w_{j+1} - \tau_j$. The segment parameters, h_j, γ_j and θ_j are removed from $h_{1:(k+1)}, \gamma_{1:(k+1)}$ and $\theta_{1:(k+1)}$, respectively, and γ_{j+1} and h_{j+1} are sampled using the proposal distributions given in Section 5.3.1.

To reverse the above move, a move is proposed to split a change interval into two change intervals. An index j is uniformly chosen from $\{j \mid 1 \leqslant j \leqslant k, w_j > 0\}$, w' is drawn uniformly from $\{0, \ldots, w_j - 1\}$, τ' is drawn uniformly from $\{\tau_j + w' + 1, \ldots, \tau_j + w_j\}$, so that $k' = k + 1$ changepoints with locations $\tau'_{1:k'} = (\tau_1, \ldots, \tau_j, \tau', \tau_{j+1}, \ldots, \tau_k)$ and interval lengths $w'_{1:k'} = (w_1, \ldots, w_j, w', w_j + \tau_j - \tau', w_{j+1}, \ldots, w_k)$ are proposed. Moreover, the segment parameters $\theta'_{1:(k'+1)} = (\theta_1, \ldots, \theta_j, \theta', \theta_{j+1}, \ldots, \theta_{k+1})$, $\gamma'_{1:(k'+1)} = (\gamma_1, \ldots, \gamma_j, \gamma', \gamma'', \gamma_{j+2}, \ldots, \gamma_{k+1})$ and $h'_{1:(k'+1)} = (h_1, \ldots, h_j, h', h'', h_{j+2}, \ldots, h_{k+1})$ are proposed, where θ', h', h'', γ' and γ'' are sampled using the proposal distributions given in Section 5.3.1.

5.3.6. Initialisation of the sampler

To speed-up the convergence of the sampler for the proposed model, the sample chain is initialised as follows: the changepoint parameters $(k, \tau_{1:k})$ and the segment parameters $\theta_{1:k}$ are set to be the MAP changepoint parameters and the MAP segment parameters corresponding to the standard changepoint model, and the interval length parameters are set to zero, implying $h_j = \gamma_j = 0$ for all j.

5.3.7. Change interval estimates

In some applications such as cyber security, given a sample obtained via MCMC as described above, it will be of interest to report a point estimate for the change interval parameters $(\hat{k}, \hat{\tau}_{1:\hat{k}}, \hat{w}_{1:\hat{k}})$. We propose to use the MAP change interval parameters $(\hat{k}, \hat{\tau}_{1:\hat{k}}, \hat{w}_{1:\hat{k}})$, defined such that $\hat{k}$ is the MAP number of changepoints, $\hat{\tau}_{1:\hat{k}}$ denote the MAP changepoint locations of dimensions $\hat{k}$, and $\hat{w}_{1:\hat{k}}$ correspond to the MAP interval lengths associated to $\hat{\tau}_{1:\hat{k}}$.

5.4. Simulation study

A simulation study is presented to demonstrate the change interval model proposed in Section 5.1 and the sampling strategy discussed in Section 5.3. In particular, Section 5.4.5 compares the proposed model with the standard changepoint model, which is defined in Section 2.1, and with DeCAFS (Romano et al., 2021). In contrast with the proposed model, both the standard model and DeCAFS assume changes are abrupt. Data within segments are assumed to be stationary for the standard model but may be subject to non-stationary fluctuations according to DeCAFS. Gradual changes can be interpreted as non-stationary temporal dynamics either at the beginning or at the end of segments. The comparison with DeCAFS aims to investigate to what extent relaxing the stationarity assumption within segments is suitable in presence of gradual changes.

5.4.1. Synthetic data

Time series of observations $x_1, \ldots, x_T$ were sampled according to the change interval model (5.3) with $f(\cdot \mid \theta)$ denoting the density function of the normal distribution with unknown mean θ, which may change from one segment to the next, and known variance $\sigma_x^2 > 0$. Data were sampled for scenarios with $\sigma_x^2 \in \{1, 5, 25, 50, 100\}$ and $\sigma \in \{0.01, 0.025, 0.05, 0.075, 0.1\}$, where σ is defined in (5.4). Different numbers of observations $T \in \{1000, 2000, 3000, 4000, 5000\}$ and numbers of changepoints $k \in \{0, 1, 2, 3, 4\}$ are considered. For scenarios with $k > 0$, changepoint locations $\tau_{1:k}$ and change interval lengths $w_{1:k}$ are set such that, for all j, $\tau_j = jT/(k+1)$ and $w_j = w$, for some $w \geqslant 0$. Different scenarios are considered for the interval lengths $w \in \{0, 50, 100, 150, 200\}$. For segment parameters, it is assumed that, for all j, $\theta_j = \theta > 0$ if j is odd and $-\theta$ otherwise, $h_j = h$, $\gamma_j \sim \text{Bernoulli}(0.5)$; scenarios with $\theta \in \{3, 6, 12, 15, 18\}$, $h \in \{0, 0.25, 0.5, 0.75, 1\}$ are considered. For each scenario, 10 simulations were performed.

5.4.2. Model comparison

For each simulation, three models are used to fit the data: the proposed change interval model, the standard changepoint model (2.4) and DeCAFS. For the change interval model, the parametric density $f(\cdot \mid \theta)$ is chosen to correspond to the normal distribution with unknown mean $\theta \sim N(0, 25)$

and known variance σ_x^2. For the prior parameters defined in Section 5.1.1, we set $p = 1/T$, $\rho = 0.1$, $\xi = 0.5$, $a = b = 1$, and we assume σ is known. Ten independent sample chains of size 20 000, after a burn-in of size 5 000, were obtained via the MCMC algorithm described in Section 5.3 for different values of N, which is the number of samples for the SIS based approximation (5.8). Change interval estimates $(\hat{k}, \hat{\tau}_{1:\hat{k}}, \hat{w}_{1:\hat{k}})$ obtained from each independent sample chain, as described in Section 5.3, will be compared to assess the convergence of the sampler. For the standard changepoint model, changepoint estimates $(\hat{k}, \hat{\tau}_{1:\hat{k}})$ are obtained as for the change interval model with the only difference that we set $\rho = 1$, ensuring that $w_j = 0$ for all j. Finally, for the DeCAFS model, changepoint estimates $(\hat{k}, \hat{\tau}_{1:\hat{k}})$ were obtained with the default implementation described in Romano et al. (2021).

Given some change interval parameters $(k, \tau_{1:k}, w_{1:k})$, let $(k^\star, \tau^\star_{1:k^\star})$ denote the change interval borders, where $k^\star = |\mathcal{C}|$ and $\tau^\star_{1:k^\star} = (\tau^\star_1, \ldots, \tau^\star_{k^\star})$ are the ordered locations in $\mathcal{C} = \bigcup_{j=1}^{k} \{\tau_j, \tau_j + w_j\}$. Models are compared based on change interval border estimations. For the proposed model, let $(\hat{k}^\star, \hat{\tau}^\star_{1:\hat{k}^\star})$ denote the change interval border estimates corresponding to the change interval estimates $(\hat{k}, \hat{\tau}_{1:\hat{k}}, \hat{w}_{1:\hat{k}})$. For the standard changepoint model and DeCAFS, the change interval border estimates are defined to be the changepoint estimates so that $\hat{k}^\star = \hat{k}$ and $\hat{\tau}^\star_j = \hat{\tau}_j$ for all j. To compare change interval border estimations, we use the F1 score. A change interval border $\tau^\star$ is said to be detected if there is an estimated change interval border $\hat{\tau}^\star$ such that $|\tau^\star - \hat{\tau}^\star| \leqslant \epsilon$ for some error tolerance of size $\epsilon = 15$. Given $(k^\star, \tau^\star_{1:k^\star})$, the F1 score of $(\hat{k}^\star, \hat{\tau}^\star_{1:\hat{k}^\star})$ is the geometric mean of the recall, which is the proportion of simulated change interval border that are detected, and the precision, which is the proportion of detected change interval borders that are correct. To compare the three models of interest, for each simulation the F1 score was computed for each estimation of the simulated change interval borders $(k^\star, \tau^\star_{1:k^\star})$.

5.4.3. Runtime

Figure 5.4 displays the average runtime of an iteration of the MCMC algorithm, implemented in Python and run on a 2.6GHz Intel Core i7 processor, for a collection of scenarios for the standard changepoint model and the change interval model. The change interval model is computationally more demanding than the standard changepoint model due to the need to propose interval length parameters and to estimate conditional likelihoods of the data within change intervals. For our Python implementations, which could be optimised, the runtime is on average 8 times greater for the change interval model than for the standard changepoint model, but this could greatly vary for different implementations. Here it is of interest to demonstrate how runtimes are affected by different models parameters. The runtime increases linearly with the number of observations for both models. As the interval length w increases, the runtime increases for the change interval model, but decreases for standard changepoint model because more changepoints are fitted resulting in smaller segments. The runtime increases linearly with the number of samples N for the SIS based

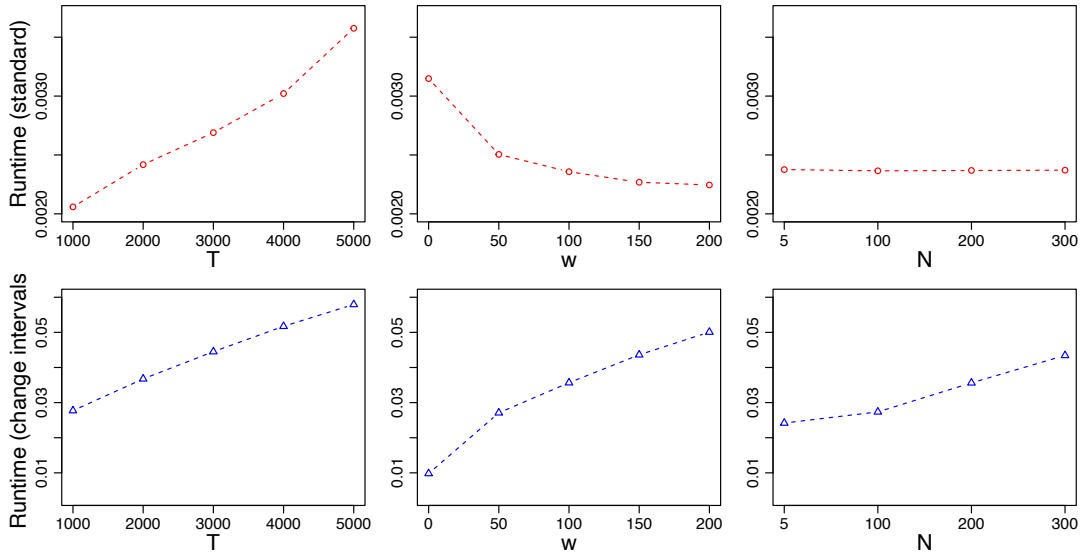


Figure 5.4.: *Average runtime in seconds per iteration for the MCMC algorithm for a collection of scenarios (with various numbers of observations T, interval length parameters w, and numbers of samples N for the SIS) for the standard changepoint model (top panel) and the proposed change interval model (bottom panel).*

numerical integration (5.8) for the change interval model.

5.4.4. Sampler convergence

The convergence of the sampler is assessed for different values of N, which is the number of samples for the SIS based approximation of the conditional likelihood of the data within change intervals. Figure 5.5 displays the variance of the F1 scores obtained from ten independent sample chains for a collection of scenarios. As N increases, the variance of the F1 scores decreases below 0.01, suggesting the sampler converges. As the change interval lengths increases, greater values of N are needed for the sampler to converge. For all scenarios considered in this simulation study, where $N \geqslant 100$, the variance of the F1 scores of changepoint estimates obtained from ten independent sample chains was less than 0.01, suggesting the proposed sampler converges. In the remainder of this section, we focus on results obtained with $N = 100$.

5.4.5. Results

Figure 5.6 displays the average F1 score of the change interval border estimations for each model of interest for a collection of scenarios: key model parameters used to sample synthetic data are varied keeping the remaining parameters fixed to the values $w = 100$, $\sigma = 5$, $h = 0.75$, $\theta = 12$, $\sigma_x = 0.075$, $k = 1$. The corresponding average precisions and recalls are displayed in Figure A.7 and Figure A.8. For each scenario the change interval model outperforms both the standard

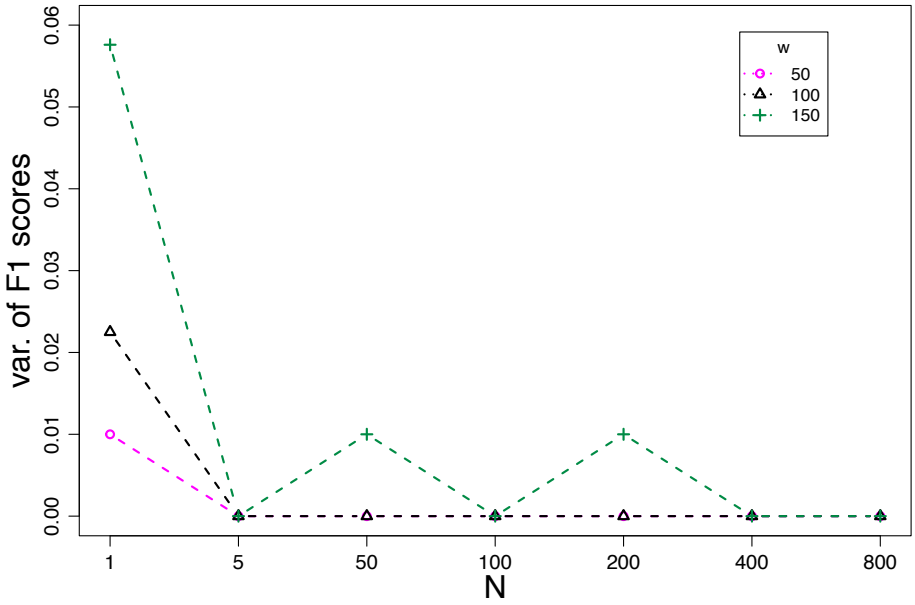


Figure 5.5.: *Variance of F1 scores for the change interval model for different numbers of samples N for the SIS and change interval length parameters w.*

changepoint model and DeCAFS. As the change interval length w increases, more changepoints are fitted within change intervals for the standard changepoint model and DeCAFS, resulting in decreasing F1scores, whereas F1 scores remain close to 1 for the change interval model. As θ decreases, gradual changes in the mean of the data decreases in magnitude, so that the need to characterise changes with time intervals of non-zero length become less relevant, and therefore F1 scores for the change interval model decrease to be equal to the F1 scores of the standard changepoint model. As the variance of the data σ_x increases, the relative magnitude of gradual changes in change interval decreases, so that F1 scores for the change interval model decrease to be equal to the F1 scores of the standard changepoint model.

The average recall and precision of the change interval border estimations for each model are given in Appendix A.3.

5.5. Computer network traffic monitoring

An application in cyber-security is discussed to demonstrate the benefits of the change interval model. As discussed in Section 1.1.2, it can be of interest to monitor for changes in the volumes of different types of traffic passing through a network over time to detect the presence of a network intrusion. For demonstration purposes, the data $x_1, \ldots, x_T$ denote the second-by-second counts of computer network events recorded at LANL for the anonymised destination port 'Port20990', which are defined in Example 1.2.

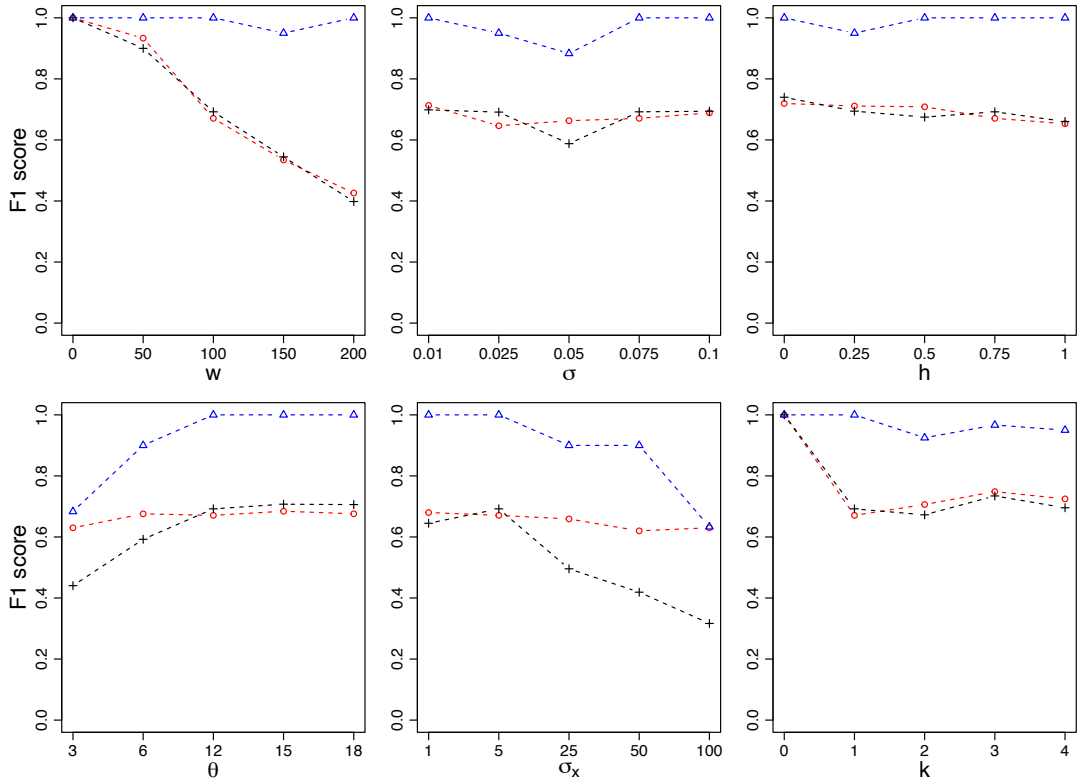


Figure 5.6.: *Average $F1$ score for the three models of interest (change interval model: blue triangles. Standard changepoint model: red circles. DeCAFS: black crosses) for a collection of scenarios considered in the study.*

Two models are used to identify temporal changes in the distribution of the count data $x_1, \ldots, x_T$: the change interval model defined in Section 5.1.1 and the standard changepoint model (2.4). For both models, the parametric density $f(\cdot \mid \theta)$ is chosen to correspond to the Poisson distribution with unknown mean $\theta \sim \Gamma(1, 0.1)$, and we set $p = 1/T$ for the prior distribution of changepoint parameters (2.2). For the change interval model, we consider a collection of prior parameters $\sigma \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and $\rho \in \{0.1, 0.4, 0.7\}$. Recall that the changepoint model is a special case of the change interval model with $\rho = 1$ so that (5.2) imposes $w_j = 0$ for all $j = 1, \ldots, k$.

MAP change interval parameters are obtained as described in Section 5.3.7 given a sample of length 30 000, with a burn-in of length 10 000, obtained via the reversible jump MCMC algorithm proposed in Section 5.3. For the standard changepoint model, MAP changepoint models are obtained given a sample of length 30 000, with a burn-in of length 10 000, obtained via the reversible jump MCMC algorithm proposed in Section 2.3.

Figure 5.7 displays the MAP changepoints for the standard changepoint model and the MAP change intervals for the change interval model with $\rho = 0.4$ and $\sigma = 0.1$. For the standard changepoint model, it is apparent that multiple changepoints are fitted in regions where the mean

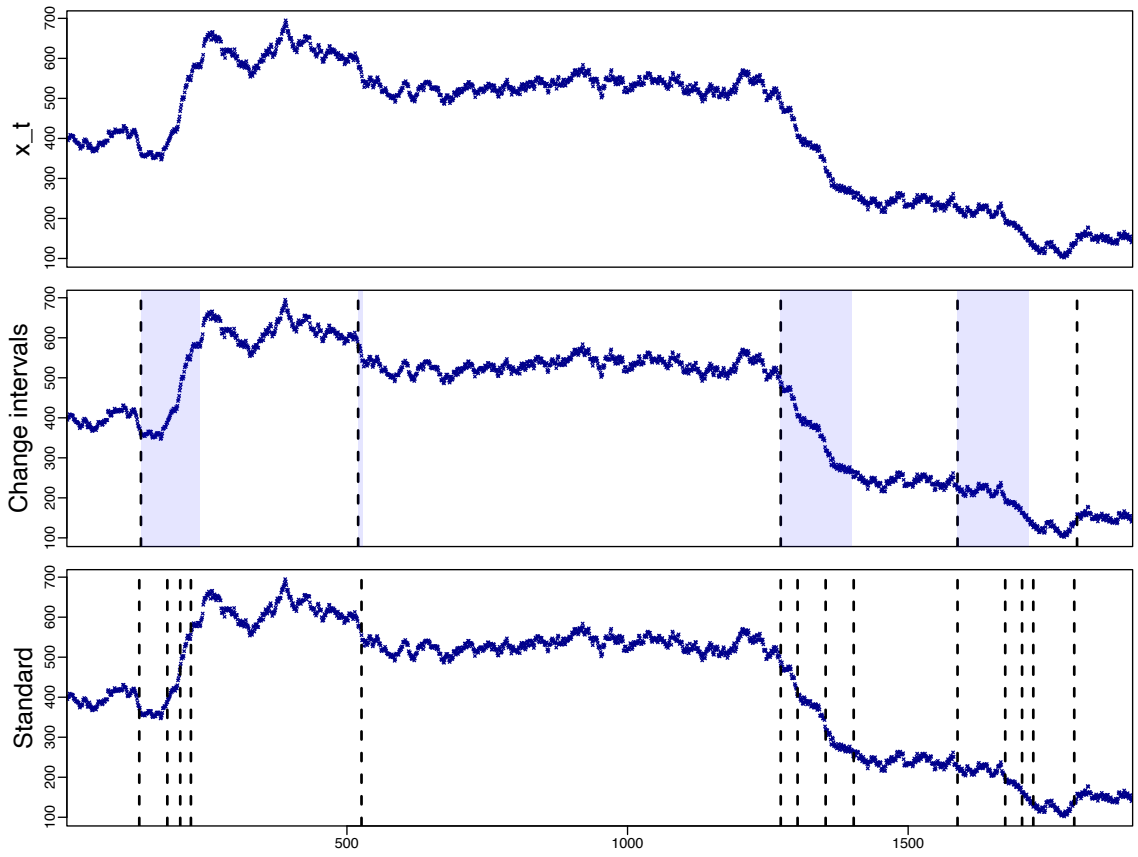


Figure 5.7.: *Counts of computer network events for the anonymised destination port 'Port20990' each second between 11:40 and 12:30 on day 22 of the data collection period. Middle panel: vertical lines indicate the starts of MAP change intervals, which are identified by blue rectangles when they have non-zero length. Bottom panel: vertical lines indicate MAP changepoints for the standard changepoint model* (2.4)*.*

of the data gradually changes. The estimated change intervals for the proposed model better characterise changes impacting the computer network traffic volumes, by identifying time intervals where changes in the mean gradually occur.

Moreover, to demonstrate the impact of prior parameters for the change interval model, Figure 5.8 displays the MAP number of change intervals $\hat{k}$ and the average change interval lengths for the MAP change intervals for a collection of prior parameters σ and ρ. As ρ increases, the cost for increasing the length of a change interval increases, and therefore more change intervals with smaller lengths are fitted to segment the data. As σ decreases, the change interval assumes smoother transitions between segments. Therefore, for the network data of interest where transitions are smooth, as σ increases, segmentations of the data with more change intervals of smaller lengths become more likely. Section 5.1.1 may guide the choice of ρ and σ: it follows from (5.2) that

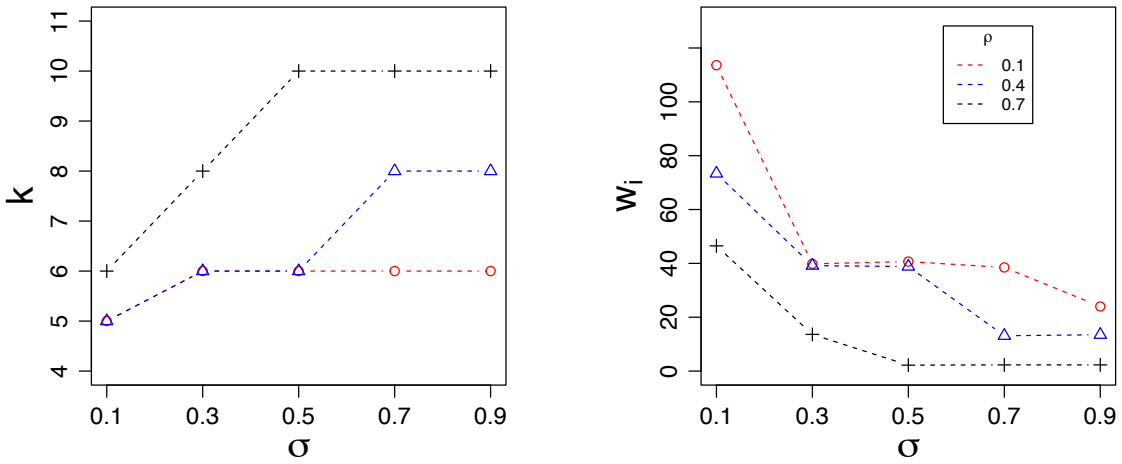


Figure 5.8.: *MAP number of change intervals $\hat{k}$ and the average change interval lengths for the MAP change intervals for a collection of prior parameters σ and ρ for the network count data defined in Example 1.2.*

the expected interval length is $(1-\rho)/\rho$, and according to (5.4) the random walk noise from one segment parameter to the next depends on σ, which may be estimated from previously observed non-abrupt transitions.

5.6. Conclusion

This chapter proposed an extension of the standard changepoint model that relaxes the assumption that changes occur instantaneously at changepoints. Changepoint parameters were augmented with time window parameters, thereby defining change intervals of possibly unequal lengths: each changepoint location identifies the position of a time interval, where a segment parameter transitions to the next segment parameter via a random walk, which is not necessarily smooth and may begin or end with a clear discontinuity, thereby encompasing various forms of transitions. A novel reversible jump MCMC algorithm was proposed to sample from the change interval model.

The change interval model was motivated by cyber data that may be subject to gradual changes that occur over some time intervals. For demonstration purposes, this chapter discussed an analysis of real computer network data collected at LANL. Volumes of traffic passing through the entire network that are subject to gradual changes were segmented according to both the standard changepoint model and the change interval model. We showed that the proposed model gives a better characterisation of the changes impacting the distribution of the data: whilst the proposed model identifies time intervals of non-zero length where gradual changes occur, the standard changepoint model fits many changepoints throughout change intervals.

6
Conclusion

6.1. Summary

This thesis proposed Bayesian changepoint models motivated by cyber-security applications. A cyber attack typically changes the behaviour of the target network. Therefore, to detect the presence of a network intrusion, it can be informative to monitor for changes in the high-volume data sources that are collected inside an enterprise computer network. However, most traditional changepoint detection methods are not adapted to characterise what cyber-security analysts mean by a change, and consequently raise too many false alerts and overlook weak but coordinated signals that are suggestive of real attacks. Chapters 3, 4 and 5 identified key features in cyber data that challenge traditional changepoint detection methods, and proposed three novel changepoint models that are adapted to these features.

Chapter 3 considers the problem of jointly analysing multiple time series that may be subject to changepoints, with a focus on settings where it is sometimes possible to specify *a priori*, by means of a graph, which pairs of time series are likely to be impacted by simultaneous changepoints. This setting was motivated by the commonly held belief of security experts that cyber attacks tend to correspond to coordinated sequences of events across multiple connected endpoints on the network (Sexton, Storlie, and Neil, 2015). A novel informative prior for changepoints was proposed, which encodes the information contained in the graph, inducing a changepoint model for multiple time series that borrows strength across clusters of connected time series to detect weak signals for synchronous changepoints. The graphical model for changepoints was further extended to allow dependence between nearby but not necessarily synchronous changepoints across neighbouring time series in the graph. A novel reversible jump MCMC algorithm making use of auxiliary variables was proposed to sample from the graphical changepoint model. For demonstration purposes, the chapter discussed an analysis of real computer network authentication data from LANL: it was of

interest to monitor for temporal changes in the authentication activity of network users to detect the presence of red team actors. The proposed changepoint prior was used to encode beliefs that signals for network intrusions are *a priori* likely to occur at nearby times for users historically linked by previous network connectivity. We showed that, as a consequence, the proposed model can detect weak signals for red team activity in the network, whilst limiting the number of false alerts, in contrast with a standard model assuming independence of behavioural changes across users.

Chapter 4, focussing on a setting with a single time series, proposes a novel segment model for changepoint models that is motivated by temporal dynamics observed in cyber data. Standard changepoint models, which typically assume data are exchangeable within each segment, may be inadequate when data are subject to local correlation patterns, often resulting in many more changepoints fitted than preferable. Chapter 4 identified this phenomenon in real cyber data that can be subject to gradual population changes, seasonal variations and other temporal trends that are unlikely to be evidence for malicious activity. To address this challenge, Chapter 4 proposed a novel Bayesian changepoint model that relaxes the assumption of exchangeability within segments. The proposed model supposes data within a segment are m-dependent for some unknown $m \geqslant 0$ that may vary between segments, resulting in a model suitable for detecting clear discontinuities in data that are subject to different local temporal correlations. A novel reversible jump MCMC algorithm was proposed to sample from the model; in particular, a detailed analysis of the parameter space was exploited to build proposals for the orders of dependence. The benefits of the proposed model were demonstrated with real network event data recorded at LANL.

Chapter 5 proposed an extension of the standard changepoint model that relaxes the assumption that changes occur instantaneously at changepoints. The extension was motivated by cyber data that may be subject to gradual changes that occur over some time intervals. Changepoint models typically assume that changes occur instantly at changepoints, resulting in many changepoints fitted throughout time intervals where the distribution of the data changes. To characterise changes impacting the distribution of the data, Chapter 5 proposed to augment changepoint parameters with time window parameters, thereby defining change intervals of possibly unequal lengths: each changepoint location identifies the position of a time interval, where a segment parameter transitions to the next segment parameter via a random walk, which is not necessarily smooth and may begin or end with a clear discontinuity, thereby encompasing various forms of transitions. A standard reversible jump MCMC algorithm for parametric changepoint models was extended to sample from the change interval model. For demonstration purposes, Chapter 5 discussed an analysis of real computer network data collected at LANL. Volumes of traffic passing through the entire network that are subject to gradual changes were segmented according to both the standard changepoint model and the change interval model. We showed that the proposed model gives a better characterisation of the changes impacting the distribution of the data: whilst the proposed model identifies time intervals of non-zero length where gradual changes occur, the standard

changepoint model fits many changepoints throughout change intervals.

6.2. Future work

The Bayesian changepoint models proposed in the thesis assume data are observed in discrete time. It could be of interest to consider how the proposed models can be adapted to data observed in continuous time. The inference of changepoints, or change intervals, is made by retrospective study based on all data available. Adapting the samplers to the online setting, where inference is made sequentially as new observations are made available, is of interest. Changepoint inference is performed via RJMCMC, which makes it possible to quantify uncertainty on changepoint parameters but may have a non-negligible computational cost. For settings where it is only of interest to obtain changepoint estimates, it might be beneficial to design posterior distribution mode finder algorithms that would lead to more efficient changepoint estimations. The proposed models extend the standard parametric changepoint model (2.4); investigating how to adapt the models to the nonparametric setting may lead to valuable developments.

The features of cyber-data that were identified as challenging for standard changepoint models can be the motivation for additional extensions of the standard changepoint model. For example, in Chapter 4, the moving-sum model was proposed as one approach to relax the standard assumption that data are exchangeable within segments, but other novel segment models could be designed to capture alternative forms of dependence for data within segments. In particular, since the moving-sum model assumes data are stationary within segments, it could be interesting to propose novel segment models for non-exchangeable and non-stationary data within segments. Moreover, Chapter 5 proposed the change interval model that assumes data that may be subject to gradual changes that occur over some time intervals. A flexible approach based on a random walk was considered to model transitions of segment parameters from one segment to the next, but other transition models could be designed. For example, future work could consider models assuming that, within change intervals, data are subject to increased variance or to outliers. Section 3.6 also discusses possible extensions of the graphical changepoint, which is proposed in Chapter 5, that relax the assumption that prior beliefs on changepoints do not vary over time.

Finally, there are other challenges faced by traditional change detection methods in cyber security applications that can motivate novel changepoint models. For example, a possible extension of the standard changepoint model could stem from the following observations. Cyber data are subject to changes that are heterogeneous in nature; examples of change types include drifts, spikes, seasonal changes, clear discontinuities. To distinguish changes that are evidence for attacks from changes that result from legitimate activities, it would be helpful to identify the types of these changes. Some types of changes may be more likely to correspond to attacks than others; and, more generally, forensic analysis of the data could rely on the detection of patterns of change types that

are surprising. Hence, it might be of interest to consider changepoint models that can characterise multiple distinct types of change.

Bibliography

Adams, N and Heard, N (2014). *Data analysis for network cyber-security*. Imperial College Press: London.

Adams, N and Heard, N (2016). *Dynamic networks and cyber-security*. Imperial College Press: London.

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

Aminikhanghahi, S. and Cook, D. J. (2017). "A survey of methods for time series change point detection". In: *Knowledge and Information Systems* 51(2), pp. 339–367.

Bardwell, L. and Fearnhead, P. (2017). "Bayesian detection of abnormal segments in multiple time series". In: *Bayesian Analysis* 12(1), pp. 193–218.

Bardwell, L., Fearnhead, P., Eckley, I. A., Smith, S., and Spott, M. (2019). "Most recent changepoint detection in panel data". In: *Technometrics* 61(1), pp. 88–98.

Barry, D. and Hartigan, J. (1993). "A Bayesian analysis for change point problems". In: *Journal of the American Statistical Association* 88(421), pp. 309–319.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis; 2nd ed.* Springer series in statistics. Springer: New York.

Bernardo, J. M and Smith, A. F. M. (1993). *Bayesian theory*. Wiley series in probability and mathematical statistics. John Wiley: New York; Chichester.

Besag, J. and Green, P. J. (1993). "Spatial statistics and Bayesian computation". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 55(1), pp. 25–37.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). "Variational inference: a review for statisticians". In: *Journal of the American Statistical Association* 112(518), pp. 859–877.

Bolton, A. D. and Heard, N. A. (2018). “Malware family discovery using reversible jump MCMC sampling of regimes”. In: *Journal of the American Statistical Association* 113(524), pp. 1490–1502.

Bondy, J. A. and Murty, U. S. R. (1976). *Graph theory with applications*. Elsevier: New York.

Carlin, B. P., Gelfand, A. E., and Smith, A. F. M. (1992). “Hierarchical Bayesian analysis of changepoint problems”. In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 41(2), pp. 389–405.

CERT-EU (2017). *WannaCry ransomware campaign exploiting SMB vulnerability*. URL: `https://cert.europa.eu/static/SecurityAdvisories/2017/CERT-EU-SA2017-012.pdf`.

Chakar, S., Lebarbier, E., Lévy-Leduc, C., and Robin, S. (2017). “A robust approach for estimating change-points in the mean of an $\mathrm{AR}(1)$ process”. In: *Bernoulli* 23(2), pp. 1408 –1447.

Chandola, V., Banerjee, A., and Kumar, V. (2009). “Anomaly detection: a survey”. In: *ACM Comput. Surv.* 41.

Chen, H. (2019). “Change-point detection for multivariate and non-Euclidean data with local dependency”. In: *preprint arXiv:1903.01598*.

Chen, H. and Zhang, N. (2015). “Graph-based change-point detection”. In: *The Annals of Statistics* 43(1), pp. 139 –176.

Chib, S. (1996). “Calculating posterior distributions and modal estimates in Markov mixture models”. In: *Journal of Econometrics* 75(1), pp. 79–97.

Chib, S. (1998). “Estimation and comparison of multiple change-point models”. In: *Journal of Econometrics* 86(2), pp. 221–241.

Cho, H. and Fryzlewicz, P. (2021). *Multiple change point detection under serial dependence: Wild contrast maximisation and gappy Schwarz algorithm*. Preprint arXiv:2011.13884.

Denison, D., Holmes, C., Bani, M., and Smith, A. (2002). *Bayesian methods for nonlinear classification and regression*. Wiley Series in Probability and Statistics: Chichester: Wiley.

Eckley, I., Fearnhead, P., and Killick, R. (2011). “Analysis of changepoint models.” In: *Bayesian time-series models*. Ed. by D. Barber, A. Cemgil, and S. Chiappa. Cambridge University Press, pp. 203–224.

Eichinger, B. and Kirch, C. (2018). “A MOSUM procedure for the estimation of multiple random change points”. In: *Bernoulli* 24(1), pp. 526 –564.

Fearnhead, P. and Liu, Z. (2011). “Efficient Bayesian analysis of multiple changepoint models with dependence across segments”. In: *Statistics and Computing* 21, pp. 217–229.

Fearnhead, P. (2006). “Exact and efficient Bayesian inference for multiple changepoint”. In: *Statistics and Computing* 16, pp. 203–213.

Fearnhead, P. and Rigaill, G. (2019). “Changepoint detection in the presence of outliers”. In: *Journal of the American Statistical Association* 114(525), pp. 169–183.

Fisch, A. T. M., Eckley, I. A., and Fearnhead, P. (2022). “Subset multivariate collective and point anomaly detection”. In: *Journal of Computational and Graphical Statistics* 31(2), pp. 574–585.

Franke, J., Kirch, C., and Kamgaing, J. (2012). “Changepoints in time series of counts”. In: *Journal of Time Series Analysis* 33(5), pp. 757–770.

Fryzlewicz, P. (2014). “Wild binary segmentation for multiple change-point detection”. In: *The Annals of Statistics* 42(6), pp. 2243–2281.

Gao, Z., Shang, Z., Du, P., and Robertson, J. L. (2019). “Variance change point detection under a smoothly-changing mean trend with application to liver procurement”. In: *Journal of the American Statistical Association* 114(526), pp. 773–781.

Gelman, A. and Rubin, D. B. (1992). “Inference from iterative simulation using multiple sequences”. In: *Statistical Science* 7(4), pp. 457 –472.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. 2nd ed. Chapman and Hall/CRC.

Gijbels, I., Hall, P., and Kneip, A. (1999). “On the estimation of jump points in smooth curves”. In: *Annals of the Institute of Statistical Mathematics* 51, pp. 231–251.

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

Grundy, T. J., Killick, R., and Mihaylov, G. (2020). “High-dimensional changepoint detection via a geometrically inspired mapping”. In: *Statistics and Computing* 30, pp. 1155–1166.

Grégoire, G. and Hamrouni, Z. (2002). “Change point estimation by local linear smoothing”. In: *Journal of Multivariate Analysis* 83, pp. 56–83.

Hallgren, K. L., Heard, N. A., and Adams, N. M. (2022). “Changepoint detection in non-exchangeable data”. In: *Statistics and Computing* 32(6), p. 110.

Hallgren, K. L., Heard, N. A., and Turcotte, M. J. M. (2023). “Changepoint detection on a graph of time series”. In: *Bayesian Analysis* to appear. Preprint arXiv:2102.04112.

Herlands, W., Neill, D. B., Nickisch, H., and Wilson, A. G. (2019). “Change surfaces for expressive multidimensional changepoints and counterfactual prediction”. In: *Journal of Machine Learning Research* 20(99), pp. 1–51.

Higdon, D. M. (1998). “Auxiliary variable methods for Markov chain Monte Carlo with applications”. In: *Journal of the American Statistical Association* 93(442), pp. 585–595.

Hinkley, D. V. (1970). “Inference About the Change-Point in a Sequence of Random Variables”. In: *Biometrika* 57(1), pp. 1–17.

Horváth, L. and Rice, G. (2014). “Extensions of some classical methods in change point analysis”. In: *TEST* 23, pp. 219–255.

Jeng, X. J., Cai, T. T., and Li, H. (2012). “Simultaneous discovery of rare and common segment variants”. In: *Biometrika* 100(1), pp. 157–172.

Joe, H. (1996). “Time series models with univariate margins in the convolution-closed infinitely divisible class”. In: *Journal of Applied Probability* 33(3), pp. 664–677.

Johnson, T., Elashoff, R., and Harkema, S. (2003). “A Bayesian change-point analysis of electromyographic data: detecting muscle activation patterns and associated applications”. In: *Biostatistics (Oxford, England)* 4, pp. 143–64.

Kent, A. D. (2015). “Cybersecurity data sources for dynamic network research”. In: *Dynamic Networks in Cybersecurity*. Imperial College Press.

Killick, R. and Eckley, I. A. (2014). “changepoint:an R package for changepoint analysis”. In: *Journal of Statistical Software* 58(3), pp. 1–19.

Killick, R., Fearnhead, P., and Eckley, I. A. (2012). “Optimal detection of changepoints with a linear computational cost”. In: *Journal of the American Statistical Association* 107(500), pp. 1590–1598.

Killick, R., Eckley, I., Jonathan, P., and Ewans, K. (2010). “Detection of changes in the characteristics of oceanographic time-series using changepoint analysis.” In: *Ocean Engineering* 37(13), pp. 1120–1126.

Lauritzen, S. L. (1996). *Graphical models*. Oxford University Press: Oxford.

Lebre, S., Becq, J., Devaux, F., Stumpf, M., and Lelandais, G. (2010). “Statistical inference of the time-varying structure of gene-regulation networks”. In: *BMC systems biology* 4, p. 130.

Li, F. and Zhang, N. R. (2010). “Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics”. In: *Journal of the American Statistical Association* 105(491), pp. 1202–1214.

Loader, C. R. (1996). “Change point estimation using nonparametric regression”. In: *The Annals of Statistics* 24(4), pp. 1667–1678.

Maidstone, R., Hocking, T., Rigaill, G., and Fearnhead, P. (2017). “On optimal multiple changepoint algorithms for large data”. In: *Statistics and Computing* 27(2), pp. 519–533.

Marchette, D. and Wegman, E. (2012). “Statistical analysis of network data for cybersecurity”. In: *CHANCE* 17, pp. 8–18.

Metelli, S. and Heard, N. (2019a). “On Bayesian new edge prediction and anomaly detection in computer networks”. In: *The Annals of Applied Statistics* 13(4), pp. 2586 –2610.

Metelli, S. and Heard, N. (2019b). “On Bayesian new edge prediction and anomaly detection in computer networks”. In: *The Annals of Applied Statistics* 13(4), pp. 2586 –2610.

Morgan, M., Sexton, J., Neil, J., Ricciardi, A., and Theimer, J. (2016). “Network attacks and the data they affect.” In: *Dynamic graphs and cyber-security*. Ed. by N. Adams and N. A. Heard. Imperial College Press: London. Chap. 1, 1–36.

Neil, J., Hash, C., Brugh, A., Fisk, M., and Storlie, C. B. (2013). “Scan statistics for the online detection of locally anomalous subgraphs”. In: *Technometrics* 55(4), 403–414.

Olshen, A. B., Venkatraman, E. S., Lucito, R., and Wigler, M. (2004). "Circular binary segmentation for the analysis of array-based DNA copy number data". In: *Biostatistics* 5(4), pp. 557–572.

Page, E. S. (1954). "Continuous Inspection Schemes". In: *Biometrika* 41(1/2), pp. 100–115.

Passino, F. S., Turcotte, M. J. M., and Heard, N. A. (2022). "Graph link prediction in computer networks using Poisson matrix factorisation". In: *The Annals of Applied Statistics* 16(3), pp. 1313 –1332.

Peluso, S., Chib, S., and Mira, A. (2019). "Semiparametric multivariate and multiple change-point modeling". In: *Bayesian Analysis* 14(3), pp. 727 –751.

Punskaya, E., Andrieu, C., Doucet, A., and Fitzgerald, W. (2002). "Bayesian curve fitting using MCMC with applications to signal segmentation". In: *IEEE Transactions on Signal Processing* 50(3), pp. 747–758.

Robert, C. P. and Casella, G. (2005). *Monte Carlo statistical methods (Springer Texts in Statistics)*. Springer-Verlag New York, Inc.: Secausus, NJ, USA.

Romano, G., Rigaill, G., Runge, V., and Fearnhead, P. (2021). "Detecting abrupt changes in the presence of local fluctuations and autocorrelated noise". In: *Journal of the American Statistical Association*, pp. 1–31.

Scott, A. J. and Knott, M. (1974). "A cluster analysis method for grouping means in the analysis of variance". In: *Biometrics* 30(3), pp. 507–512.

Sen, A. and Srivastava, M. S. (1975). "On tests for detecting change in mean". In: *The Annals of Statistics* 3(1), pp. 98 –108.

Sexton, J. O., Storlie, C., and Neil, J. (2015). "Attack chain detection". In: *Statistical analysis and data mining* 8, pp. 353–363.

Shewhart, W. A. (1931). *Economic control of quality of manufactured product*. New York: Van Nostrand Reinhold Co.

Shiryaev, A. N. (1963). "On optimum methods in quickest detection problems". In: *Theory of Probability and Its Applications* 8(1), pp. 22–46.

Sparks, R. S., Keighley, T., and Muscatello, D. (2011). "Optimal exponentially weighted moving average (EWMA) plans for detecting seasonal epidemics when faced with non-homogeneous negative binomial counts". In: *Journal of Applied Statistics* 38(10), pp. 2165–2181.

Stephens, D. A. (1994). "Bayesian Retrospective Multiple-Changepoint Identification". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 43(1), pp. 159–178.

Steutel, F. W. (1975). "Simple tools in finite and infinite divisibility". In: *Advances in Applied Probability* 7(2), pp. 257–259.

Swendsen, R. H. and Wang, J.-S. (1987). "Nonuniversal critical dynamics in Monte Carlo simulations". In: *Phys. Rev. Lett.* 58 (2), pp. 86–88.

Tartakovsky, A., Polunchenko, A., and Sokolov, G. (2012). "Efficient computer network anomaly detection by changepoint detection methods." In: *IEEE Journal of Selected Topics in Signal Processing* PP(99), p. 1.

Truong, C., Oudre, L., and Vayatis, N. (2020). "Selective review of offline change point detection methods". In: *Signal Processing* 167, p. 107299.

Turcotte, M. (2014). "Anomaly detection in dynamic networks". In: *PhD thesis, Imperial College London*.

Turcotte, M., Heard, N., and Neil, J. (2014). "Detecting localised anomalous behaviour in a computer network". In: *Advances in Intelligent Data Analysis XIII*. Springer International Publishing: Cham, pp. 321–332.

Turcotte, M. J. M., Kent, A. D., and Hash, C. (2017). "Unified host and network data set". In: *Data Science for Cyber-Security*. Chap. 1, pp. 1–22.

Vogt, M. and Dette, H. (2015). "Detecting gradual changes in locally stationary proceses". In: *The Annals of Statistics* 43(2), pp. 713–740.

Wang, H., Zhang, D., and Shin, K. (2004). "Change-point monitoring for the detection of DoS attacks." In: *IEEE Transactions on Dependable and Secure Computing* 1(4), 193–208.

Wang, T. and Samworth, R. J. (2018). "High dimensional change point estimation via sparse projection". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80(1), pp. 57–83.

Weiß, C. H. (2011). "Detecting mean increases in Poisson INAR(1) processes with EWMA control charts". In: *Journal of Applied Statistics* 38(2), pp. 383–398.

Yau, C. and Holmes, C. (2013). "A decision-theoretic approach for segmental classification". In: *The Annals of Applied Statistics* 7(3), pp. 1814–1835.

Yu, X., Baron, M., and Choudhary, P. K. (2013). "Change-point detection in Binomial thinning processes, with applications in epidemiology". In: *Sequential Analysis* 32(3), pp. 350–367.

A

Simulation studies

This chapter provides additional results for the simulation studies considered in this thesis. Section A.1, Section A.2 and Section A.3 correspond to the simulation study of Chapter 3, Chapter 4 and Chapter 5, respectively.

A.1. Graph of time series

A.1.1. MCMC diagnostics

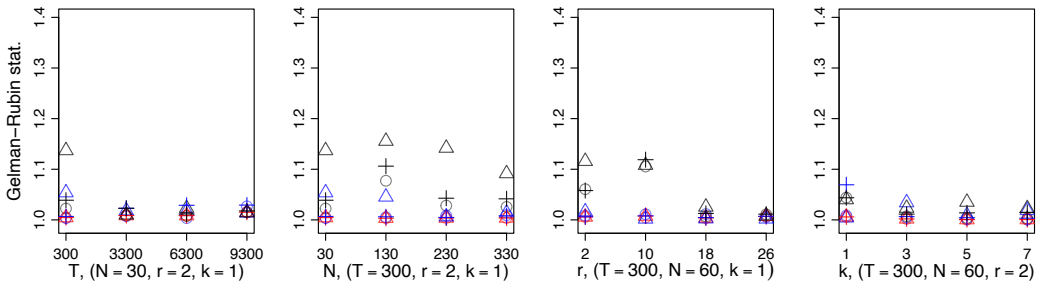


Figure A.1.: *Average Gelman-Rubin statistics for independent samples obtained via the proposed MCMC algorithm for a collection of scenarios:* $\lambda_s = 0$ *(Red);* $\lambda_s = 0.2$ *(Blue);* $\lambda_s = 0.6$ *(Black); zero window (Cross); fixed window (Circle); variable window (Triangle).*

Figure A.1 displays the Gelman-Rubin test statistics for a collection of scenarios considered in the experiment, which is discussed in Section 3.4.3. It is apparent the Gelman-Rubin test statistics are less than 1.2, suggesting that the sampler has converged.

A.1.2. Detection of clusters of changepoints on a graph of time series

Section 3.4.4.1 demonstrates the roles of the graph G and the changepoint parameters p and λ using simulations corresponding to scenarios with $v = 0$ for synchronous changepoints and with signal strength $\mu = 90$.

For other signal strengths considered in the experiment discussed in Section 3.4.4.1, corresponding to $\mu \in \{55, 120\}$, results are similar to results for the scenario where $\mu = 90$ when it comes to the role of changepoint prior parameters. Yet, by considering results for the dependence structure corresponding to the 2-chain displayed in Figure A.3, we note that, as the signal strength increases, the region of low MSE for k_i for time series $i \in C_1, C_2$ translates along the $\bar{p}$ axis on the $\bar{p}, \lambda$-plane.

The roles of G, p and λ are same when detecting asynchronous changepoints. It is apparent in Figure A.2 which displays results corresponding to scenarios with $v = 10$ for asynchronous changepoints and assuming the *fixed window* scenario for the graphical changepoint models. Results are similar for other levels of asynchrony v. The role of changepoint lags when detecting asynchronous changepoints is discussed in Section 3.4.4.2.

A.1.3. Extension for asynchronous changepoints

Section 3.4.4.2 demonstrates the role of changepoint lags when detecting asynchronous changepoints for the graphical changepoint models assuming $\bar{p} = -100$. As shown in Figure A.4, the role of changepoint lags is the same when assuming different $\bar{p}$. The roles of the graph G and the changepoint parameters $\bar{p}$ and λ are discussed in Section 3.4.4.1.

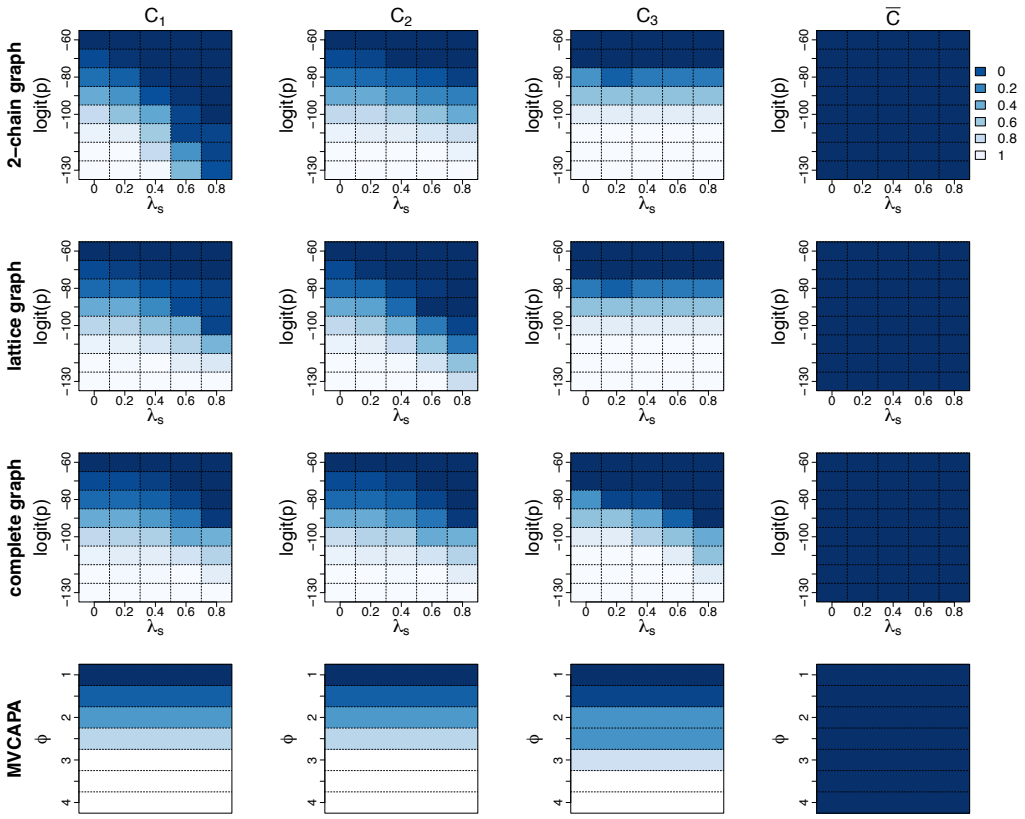


Figure A.2.: *Mean squared error for k_i for time series i in clusters $C_1, C_2, C_3, \bar{C}$. Top rows: estimates for the graphical changepoint model assuming the 2-chain graph, the lattice graph and the complete graph, as a function of $\bar{p} = logit(p)$ and λ_s. Bottom row: estimates for MVCAPA as function of ϕ. Results correspond to scenarios with $v = 10$ and $\mu = 90$.*

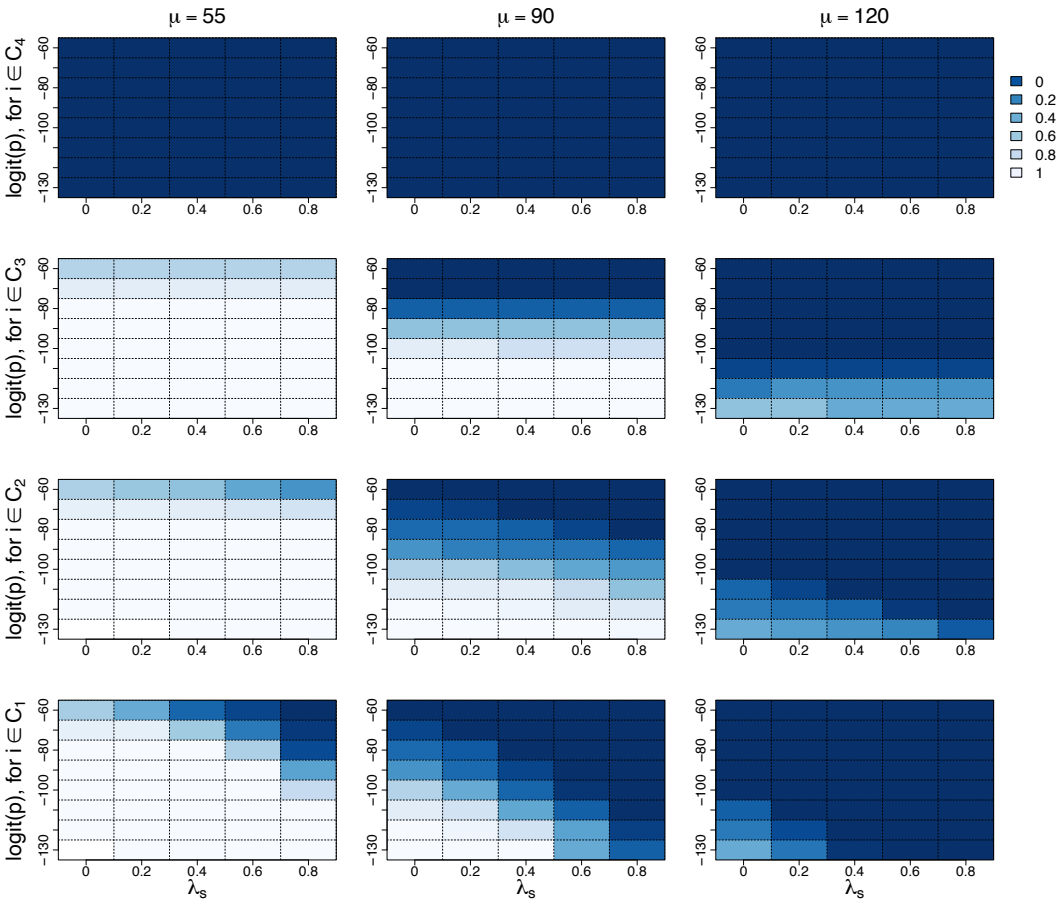


Figure A.3.: *Mean squared error for k_i for times series $i \in C_1, C_3$, given the 2-chain graph based dependence structure for changepoints, as a function of $\bar{p} = logit(p)$ and λ_s, and for different strength signals for changepoints $\mu \in \{55, 90, 120\}$.*

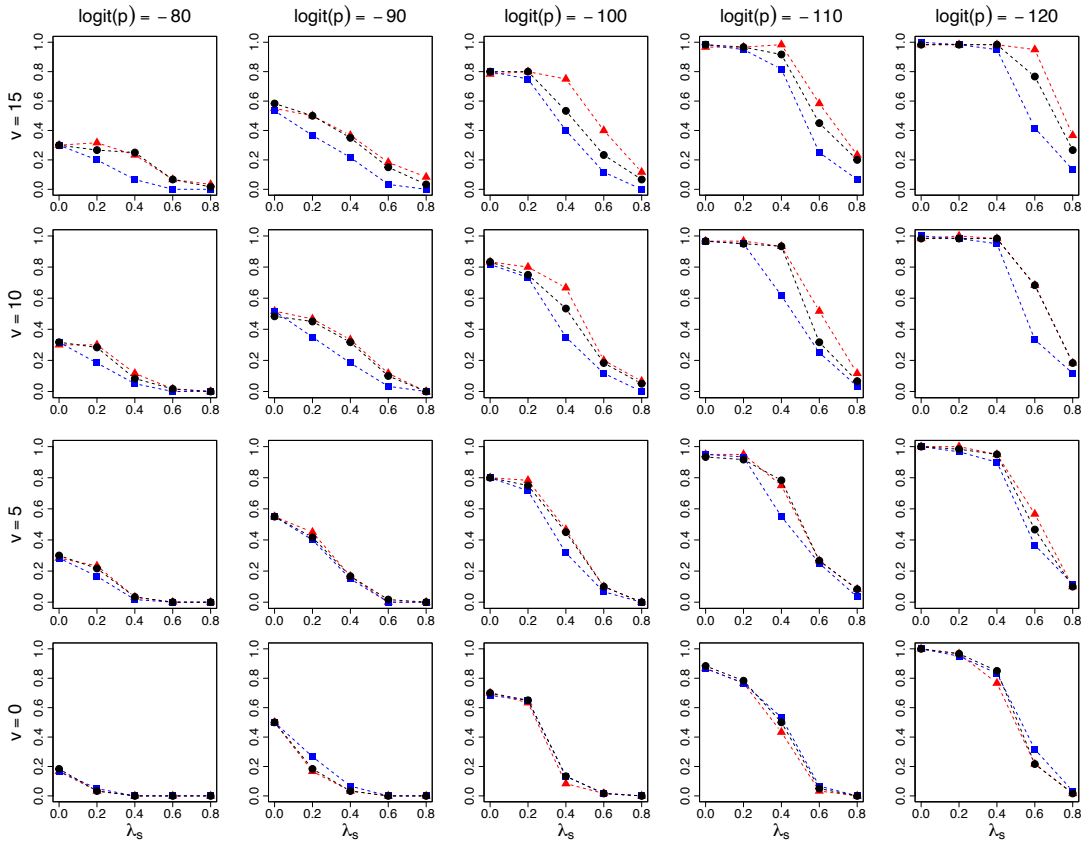


Figure A.4.: *Mean square error for k_i for time series $i \in C_1$, given the 2-chain dependence structure for changepoints, corresponding to different assumptions for the upper bounds for the lags* - fixed window *(blue squares),* variable window *(black circles), and* zero window *(red triangles) - for increasing levels of asynchrony of the changepoints v, and for a collection of changepoint prior parameters p, λ_s.*

A.2. Non-exchangeable data

Section 4.5 discusses a simulation study that demonstrates the benefits of the moving-sum changepoint model. Figure A.5 and Figure A.6 display the average precisions and recalls of the changepoint estimations, for each model of interest for a collection of scenarios, that correspond to the F1 scores displayed in Figure 4.9.

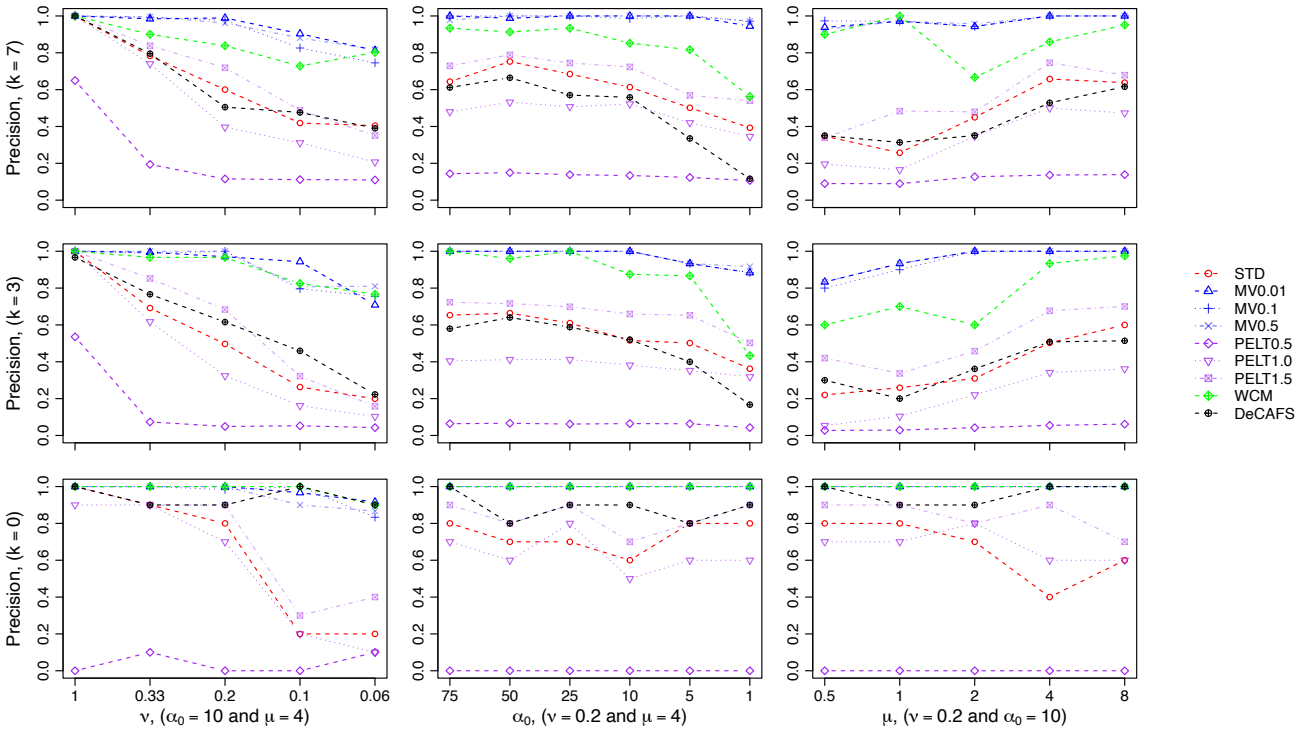


Figure A.5.: *Average precisions of changepoint estimates for the changepoint models of interest (STD for the standard changepoint model* (2.4)*; MVρ for the moving-sum changepoint model with hyperparameter ρ for orders of dependence; PELTϕ for PELT with MBIC penalty scaled by ϕ; WCM for WCMgSa; and DeCAFS) for a collection of scenarios considered in the study.*

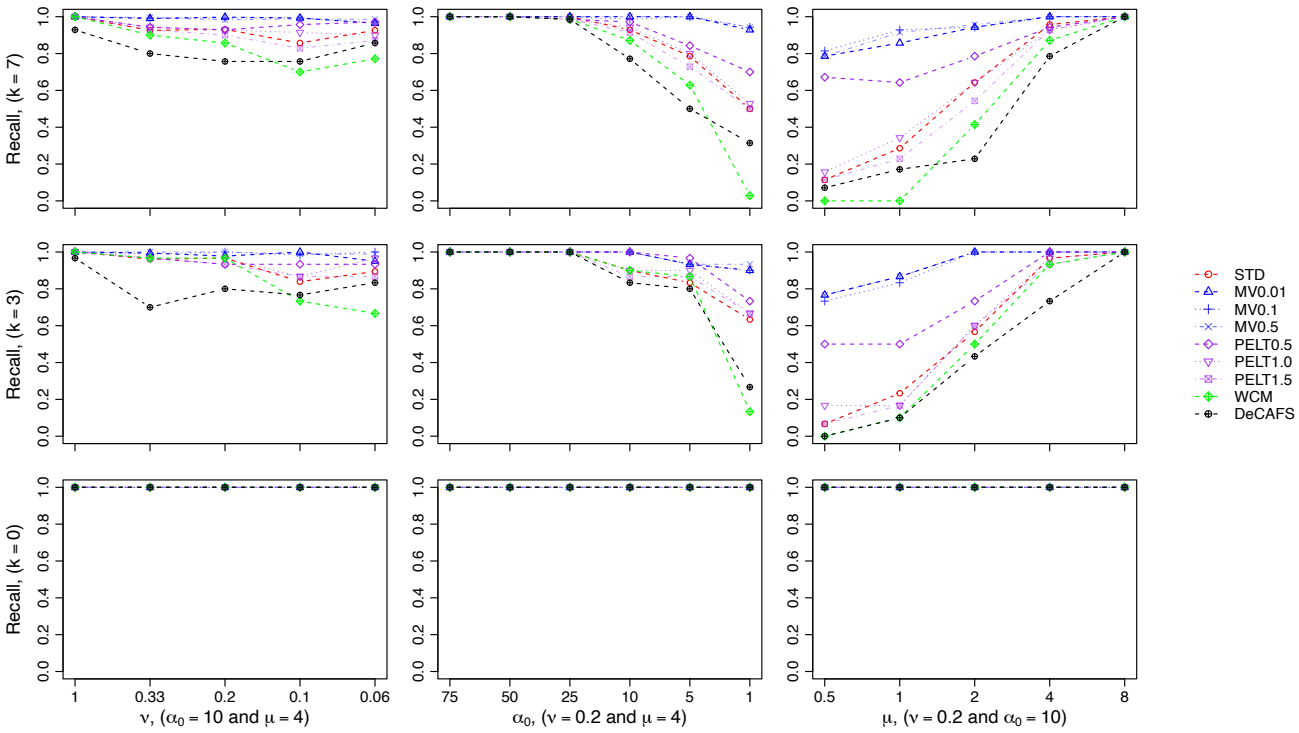


Figure A.6.: *Average recalls of changepoint estimates for the changepoint models of interest (STD for the standard changepoint model* (2.4)*; MVρ for the moving-sum changepoint model with hyperparameter ρ for orders of dependence; PELTϕ for PELT with MBIC penalty scaled by ϕ; WCM for WCMgSa; and DeCAFS) for a collection of scenarios considered in the study.*

A.3. Change intervals

For the simulation study discussed in Section 5.4, which demonstrates the benefits of the change interval model, Figure A.7 and Figure A.8 display the average precisions and recalls of the change interval border estimations, for each model of interest for a collection of scenarios, that correspond to the F1 scores displayed in Figure 5.6.

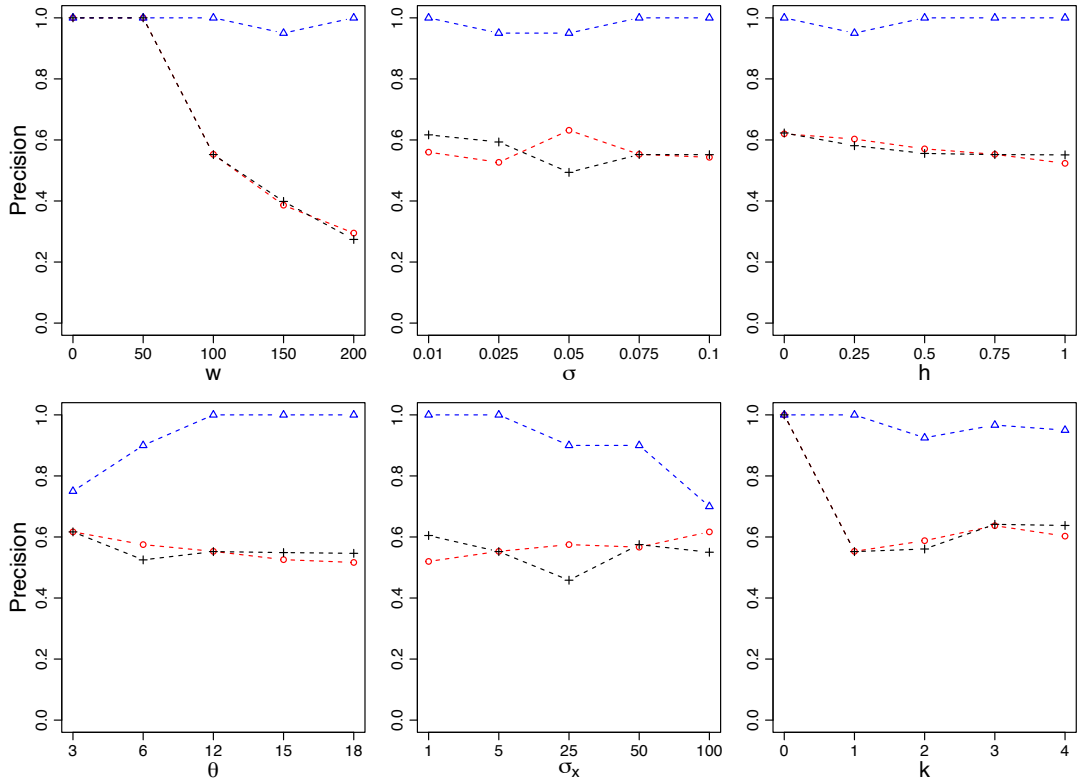


Figure A.7.: *Average precision for the three models of interest (change interval model: blue triangles. Standard changepoint model: red circles. DeCAFS: black crosses) for a collection of scenarios considered in the study.*

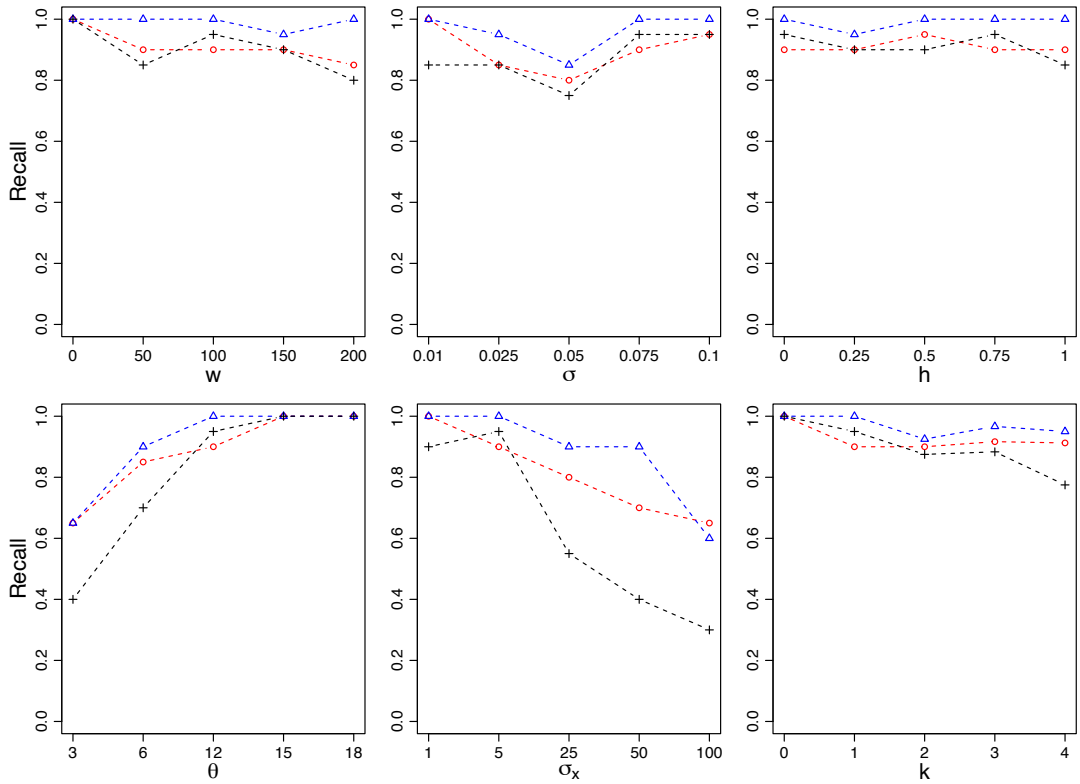


Figure A.8.: *Average recall for the three models of interest (change interval model: blue triangles. Standard changepoint model: red circles. DeCAFS: black crosses) for a collection of scenarios considered in the study.*

B

Proofs and results

This chapter provides proofs of some results that were stated in this thesis.

B.1. Proof of Proposition 1

(i) Let $w \geqslant 0$, $k \geqslant 0$ and $\tau_{1:k} = (\tau_1, \ldots, \tau_k) \in \mathcal{T}_k$. We introduce some notations to facilitate the proof. For all $0 \leqslant j \leqslant l \leqslant k$, let

$$\Omega_{j:l} = \Omega_{j:l}(w, k, \tau_{1:k}) = \{d_{j:l} = (d_j, \ldots, d_l) : \forall i,\ 0 \leqslant d_i < \rho(i, 0) = \min\{w + 1, T + 1 - \tau_i\}\}$$

and

$$\mathcal{Q}_{j:l} \equiv \mathcal{Q}_{j:l}(w, k, \tau_{1:k}) = \{d_{j:l} = (d_j, \ldots, d_l) \in \Omega_{j:l} : \tau_j + d_j \geqslant \tau_{j+1} + d_{j+1} \geqslant \cdots \geqslant \tau_l + d_l\}.$$

Note that $|\Omega_{j:l}| = \prod_{i=j}^{l} \rho(i, 0)$ and $|\mathcal{Q}_{j:l}| = Q(j, l - j)$ where Q is defined in (3.12).
It is immediate that if $k = 0$ then (3.13) holds. If $k = 1$ then $|\mathcal{D}(w, k, \tau_{1:k})| = \min\{w + 1, T + 1 - \tau_1\} = Z(1)$. We show by induction that (3.13) holds for $k > 1$. If $k = 2$ then

$$\begin{aligned} |\mathcal{D}(w, k, \tau_{1:k})| &= |\Omega_{1:2}| - |\mathcal{Q}_{1:2}| \\ &= \rho(1, 0)\rho(2, 0) - Q(1, 1) \\ &= Z(1)Q(2, 0) - Z(0)Q(1, 1) \\ &= Z(2). \end{aligned}$$

Assume that (3.13) holds for $k = 1, \ldots, k' - 1$ for some $k' > 2$. We show that (3.13) holds

for $k = k'$. We have

$$\begin{aligned}
|\mathcal{D}(w,k,\tau_{1:k})| &= \sum_{j=1}^{k}(-1)^{k-j}|\{d_{1:k} \in \Omega_{1:k} : d_{1:(j-1)} \in \mathcal{D}(w,j-1,\tau_{1:(j-1)}),\, d_{j:k} \in \mathcal{Q}_{j:k}\}| \\
&= \sum_{j=1}^{k}(-1)^{k-j}|\mathcal{D}(w,k-j,\tau_{1:(k-j)})|\,|\mathcal{Q}_{j:k}| \\
&= \sum_{j=1}^{k}(-1)^{k-j}Z(j-1)Q(j,k-j) \\
&= Z(k).
\end{aligned}$$

(ii) Follows immediately from (3.11).

B.2. Proof of Proposition 3

(i) For all n and for all reals $z_1, z_2, \ldots, z_n$,

$$\min\{z_1,\ldots,z_n\} = -\max\{-z_1,\ldots,-z_n\},$$

and therefore $U^m = x_1 - L^m_{m+1}$ for all m by definition (4.9).
Moreover, for all ℓ and for all reals (z_{ij}),

$$\sum_{i=1}^{\ell}\max\{0, z_{i1}, z_{i2}, z_{i3}, \ldots\} \geqslant \max\left\{0, z_{11}, z_{11}+z_{21}, z_{11}+z_{21}+z_{31}, \ldots, \sum_{j}\sum_{i=1}^{\ell} z_{ij}\right\},$$

and therefore, with ℓ such that $m+1 = \ell(m'+1)$ and using the notations introduced in (4.9), for all $r = 1, \ldots, m'+1$,

$$\begin{aligned}
\sum_{j=0}^{\ell-1} L^m_{j(m'+1)+r} &= \sum_{j=0}^{\ell-1}\max\left\{0, \triangle x_{j(m'+1)+r}, \ldots, \sum_{q=0}^{\kappa_{j\bar{m}'+r}} \triangle x_{(q\ell+j)(m'+1)+r}\right\} \\
&\geqslant L^{m'}_r.
\end{aligned}$$

As a result, for all $\gamma_{1:m}$ such that $U^m \geqslant \sum_r \gamma_r$ and $\gamma_r \geqslant L^m_r$ for all $r = 1, \ldots, m$, that is $\gamma_{1:m} \in \mathcal{Y}_m$, it follows that

$$\begin{aligned} U^{m'} = x_1 - L^{m'}_{m'+1} &\geqslant x_1 - \sum_{j=0}^{\ell-1} L^{m}_{j(m'+1)+(m'+1)} \\ &= U^m - \sum_{j=0}^{\ell-2} L^{m}_{(j+1)(m'+1)} \\ &\geqslant \left(\sum_{r=1}^{m} \gamma_r\right) - \sum_{j=0}^{\ell-2} \gamma_{(j+1)(m'+1)} = \sum_{r=1}^{m'} \gamma'_r \end{aligned}$$

and $\gamma'_r \geqslant \sum_j L^m_{j(m'+1)+r} \geqslant L^{m'}_r$ for all $r = 1, \ldots, m'$, that is $\gamma'_{1:m'} \in \mathcal{Y}_{m'}$.

(ii) Follows from (i).

B.3. Proof of Proposition 4

Note that (4.2) and (4.3) imply that $\triangle x_t = -\nabla^{(m+1)} y_{t+1}$ for all t. Hence, by definition of the bounds in (4.9), for all n,

$$L^{m'}_r(x_{1:n}) = \max\left\{0, -\nabla^{(m+1)} y_{r+1}, \ldots, -\sum_{q=0}^{\kappa_r(n)} \nabla^{(m+1)} y_{q(m'+1)+r+1}\right\}, \tag{B.1}$$

with $\kappa_r(n)$ defined to be the largest $q \in \mathbb{N}_0$ such that $q(m'+1)+r+1 \leqslant n$, for all r.

If $m' \in \mathcal{D}(m)$, that is there is some $\ell > 0$ such that $m+1 = \ell(m'+1)$, then for all q,

$$\nabla^{(m+1)} y_{q(m'+1)+r+1} = y_{q(m'+1)+r+1} - y_{(q-\ell)(m'+1)+r+1},$$

and we note that $q(m'+1)+r+1 = (q+\ell-\ell)(m'+1)+r+1$. Therefore

$$L^{m'}_r(x_{1:n}) = \left(\sum_{q=0}^{\ell-1} y_{(q-\ell)(m'+1)+r+1}\right) - B_n = \gamma'_r - B_n,$$

where

$$B_n = \min\left\{\sum_{q=j-\ell}^{j-1} y_{q(m'+1)+r+1}, \quad j = 0, \ldots, \kappa_r(n)\right\}$$

is a random variable that converges almost surely to 0, since it is non-increasing and converges in probability to the infimum of the support of $f_m(\cdot \mid \theta)$ which is 0. Hence $L^{m'}_r(x_{1:n})$ converges almost surely to γ'_r.

However if $m' \notin \mathcal{D}(m)$ then for all $q_1, q_2 \in \{0, \ldots, \kappa_r(n)\}$, we have that $q_1(m'+1)+r+1 = q_2(m'+1)+r+1$ if and only if $q_1 = q_2$, so that no cancellations occur in (B.1). Hence $L^{m'}_r(x_{1:n})$,

which is non-decreasing, converges almost surely to the supremum of the support of $f_m(\cdot \mid \theta)$ which is ∞.

Finally, the results on the upper bound then follow from the observation that

$$\begin{aligned} U^{m'}(x_{1:n}) &= x_1 - L^{m'}_{m'+1}(x_{1:n}) \\ &= \left(\sum_{r=1}^{m'} \gamma'_r\right) + \left(\sum_{q=0}^{\ell-1} y_{(q-\ell+1)(m'+1)+1}\right) - L^{m'}_{m'+1}(x_{1:n}). \end{aligned}$$

B.4. Order of dependence estimation

The sampling algorithm proposed in Section 4.4 can be used to estimate changepoint parameters as well as orders of dependence within segments. This section proposes an approach to estimate the posterior distribution of the order of dependence within a segment, which may be used to validate MCMC based estimations of orders of dependence conditional on MAP changepoint parameters.

Let $x_1, \ldots, x_n$ be some data observed within a single segment, assuming the moving-sum model defined in Section 4.1.1 for some $m \geqslant 0$ and θ. The posterior density of m, $\pi(m \mid x_{1:m}) \propto \pi(m) I_m \mathbf{1}_{\mathcal{M}}(m)$, where

$$I_m = \int_{\gamma_{1:m} \in \mathcal{Y}_m} L(x_{1:n}, \gamma_{1:m} \mid m) \mathrm{d}\gamma_{1:m}$$

is not tractable in general. However, an importance sampling estimate of I_m can be computed for each m, as described below, to obtain an approximation to the posterior density of m. Given N samples $\gamma^{(1)}_{1:m}, \ldots, \gamma^{(N)}_{1:m}$ from the distribution with density

$$q(\gamma_{1:m}) = \prod_{\ell=1}^{m} f_m(\gamma_\ell \mid \hat{\theta}),$$

where $\hat{\theta}$ is the MAP estimate of θ, which is tractable by conjugacy of θ,

$$\hat{I}_m = \frac{1}{N} \sum_{i=1}^{N} \frac{L(x_{1:n}, \gamma^{(i)}_{1:m} \mid m)}{q(\gamma^{(i)}_{1:m})}$$

is an estimator of I_m that converges almost surely to I_m as N tends to infinity. The posterior density of m can be estimated by the density

$$\hat{\pi}(m \mid x_{1:m}) \propto \pi(m) \hat{I}_m \mathbf{1}_{\mathcal{M}}(m).$$