Bayesian On-line Change-point Detection

Spatio-temporal point processes

Ioannis Zachos

Supervisor: Dr. Theo Damoulas

In partial fulfilment of
BSc Data Science

Department of Computer Science
University of Warwick
April 2018

To my father, Georgios (also a Warwick graduate), and my mother, Efterpi, for their invaluable love, support and encouragement.

Acknowledgements

I would like to thank Theodoros Damoulas for his valuable guidance and support throughout the duration of the project and Jeremias Knoblauch for giving me access to his codebase.

Abstract

This thesis details an approach known as *change-point detection* (CPD) that aims to detect changes in the mean, variance and covariance of a time series. The scope of CPD is limited to an on-line (real-time) Bayesian spatio-temporal setting. In this setting, the goal of CPD is to provide step-ahead predictions and partition the time series into disjoint segments every time a new datum is received using Bayesian inference. This is achieved by modelling each datum as a sample from a data-generating process which we are imitating using a probability distribution as a model. At each time step the most likely model is chosen among a universe of potential models. This leads to the development of the Bayesian on-line change-point detection and model selection (BOCDMS) algorithm which has a linear computational and storage complexity in the number of observations.

Model selection is narrowed by employing two conjugate point process models: the Poisson Gamma (PG) and Multinomial Dirichlet (MD) models. We study the properties of these models and assess their sensitivity and performance on four synthetic and three real-world datasets, the latter of which are related to crime in Chicago, property transactions in the UK and cryptocurrency transactions.

Keywords: Change-point detection, point processes, Bayesian inference, spatio-temporal statistics, machine learning

Contents

List of Figures

List of Tables

Nomenclature

In this thesis, different typeface is used to denoted different objects. Standard notation includes representing a scalar as x, a vector as $\boldsymbol{x}$, and a matrix as $\boldsymbol{X}$. The i-th element of a vector $\boldsymbol{x}$ is written as x_i while the (i,j)-th element of a matrix is indexed as X_{ij}. Moreover, we try to minimise notation collisions and announce such rare occasions using footnotes.

Regarding terminology, we use the words model and probability distribution interchangeably as the models we employ are probabilistic. Also, a data stream is used to denote input data in the form of a univariate time series. Additional nomenclature is listed below.

Nomenclature

T_i	Arrival time i
$\mathcal{O}(\cdot)$	Big Oh notation
C_i	Change-point i
$\boldsymbol{y}_{1:t}$	Collection of data $\boldsymbol{y}_1$ to $\boldsymbol{y}_t$ or $(\boldsymbol{y}_1, \dots, \boldsymbol{y}_t)$
$Cov(\cdot,\cdot)$	Covariance
$\sim$	Distributed as
$\overset{d}{=}$	Equal in distribution
$\mathbb{E}[\cdot]$	Expectation
$\forall$	For all
$\Gamma(\cdot)$	Gamma function
$\vert$	'Given'. Used to denote conditionality
$\widehat{\boldsymbol{y}}_{t+h}$	h-step ahead predictive outcome
$\widehat{\boldsymbol{y}}_t^h$	h-step ahead predictive interval
$H(\cdot)$	Hazard function

$\mathbb{1}\{\cdot\}$	Indicator function
∞	Infinity
$\int$	Integral
$\cap$	Intersection
S_i	Inter-arrival time i (clashes with segment S_t)
$l_{segment}$	Length of segment
m_t	Model at time t
$\mathcal{M}$	Model universe
$\boldsymbol{y}_t$	Multidimensional time series at time t
$\boldsymbol{\eta}$	Multinomial Dirichlet model hyper-parameter
$\mathbb{N}$	Natural numbers (not including zero)
n_{cps}	Number of change-points
$\boldsymbol{\Theta}_m$	Parameter space of model m
$\boldsymbol{\theta}_{m_t}$	Parameter vector of model m_t
ρ	Point process intensity
$\boldsymbol{\alpha}$	Poisson gamma model hyper-parameter (shape)
$\boldsymbol{\beta}$	Poisson gamma model hyper-parameter (rate)
$\mathbb{P}$	Probability or probability mass
$d\mathbb{P}$	Probability density
$\prod$	Product
$\widehat{\boldsymbol{Y}}_{t+h}$	Random variable denote h-step ahead prediction
$\mathbb{R}$	Real numbers
r_t	Run-length
S_t	Segment at time t
$S^{(i)}$	Segment i in terms of the i-th change-point
$\subseteq$	Subset
$\sum$	Sum
$supp(\cdot)$	Support of a random variable
$\exists$	There exists

t	Time

Acronyms

This thesis uses a variety of acronyms. We provide a list of acronyms used throughout the literature as well as those defined by the author.

Common acronyms

AR	Auto-regressive
BOCDMS	Bayesian on-line change-point detection and model selection
BTC	Bitcoin (cryptocurrency)
CP	Change-point
CPD	Change-point detection
CPE	Change-point estimation
CUMSUM	Cumulative sum
DGP	Data-generating process
ETH	Ethereum (cryptocurrency)
HMM	Hidden Markov model
iid	Independent and identically distributed
LR	Likelihood ratio
LTC	Litecoin (cryptocurrency)
LTP	law of total probability
MAP	Maximum a posteriori
MD	Multinomial-Dirichlet
MLE	Maximum likelihood estimation
MSBOCD	Model-specific Bayesian on-line change-point detection
PDF	Probability density function

PG	Poisson-Gamma
PMF	Probability mass function
PP	Point process
PPM	Product partition model
r.v.	Random variable
TS	Time series
TSA	Time series analysis

Chapter 1

Introduction

Objectives:

- ✓ Introducing and formulating change-point detection problems.
- ✓ Motivating the need for change-point detection algorithm.
- ✓ Conducting a comprehensive literature review of the change-point detection techniques used in a variety of settings.
- ✓ Limiting the scope of this thesis and outlining its contributions.

The study of sequences of data ordered in time, otherwise known as *time series* (TS), dates back to the work of G. U Yule in the 1920s (Yule, 1927). A natural way of improving the understanding of such sequences is to partition or segment them into smaller sub-sequences while maintaining their natural (time) ordering. The boundaries between successive partitions are known as structural breaks or *change-points* (CPs) and capture important pieces of information, such a natural interpretation in the context of a particular application.

The task of splitting a time series into time-ordered blocks is sometimes known as *segmentation*. Time series segmentation can be a fruitful activity as it can reveal the hidden properties of the source of the series. However, it often the case that these properties are discovered only in some regions of the TS. Another complication arises from the fact that the same properties do not appear periodically in different regions of the TS. Therefore, it is unrealistic to assume that the dynamics that govern a time series (and therefore its properties) do not change over time due to the stochastic or non-deterministic nature of these series. To avoid committing to naive assumptions about the behaviour of a time series as a whole, it is useful to identify segments where certain regimes govern the data. Hence, simplifying assumptions can be made locally without overestimating the applicability of these assumptions. A key assumption that is vital for modelling and inference is *stationarity*[1]. By modelling a time series as a piecewise stationary process, change-point detection (CPD) becomes more accurate in modelling real-world TS.

The changes that prevent us from accepting such assumptions are attributed to external factors and/or internal systematic changes of a system's dynamics (Aminikhanghahi and Cook, 2017). The latter cause of change-points is more interesting as internal changes are not often visible to the human eye and are therefore more difficult to detect. By taking these changes into account a change-point detection system aims to identify whether a change has occurred and to pinpoint its exact location in time. For the reasons outlined above, CPD constitutes a necessary tool for time series modelling as it boosts predictive performance significantly.

1.1 Problem statement

A formal introduction to the change-point problem tackled in this dissertation is provided below. Consider the following formulation of a one-dimensional change-point problem.

Let $\boldsymbol{y}_{1:t} := (y_1, y_2, \ldots, y_t)$ be a finite data stream $\forall y_i \in \mathbb{R}$, $i, t \in \mathbb{N}$. The sequence $\boldsymbol{y}_{1:t}$ is commonly referred to as a time series. Suppose that $\exists\ S^{(1)}, S^{(2)}, \ldots, S^{(l)}$ disjoint subsets of $\boldsymbol{y}_{1:t}$ known as segments, where the number of CPs $l \in \mathbb{N}$ is unknown. Assume that the elements in each one of $S^{(1)}, S^{(2)}, \ldots, S^{(l)}$ are arranged in increasing order of time. Define a change-point C_i to be the index of the last element of subset $S^{(i)}$, $\forall i \in \{1, \ldots, l\}$. The goal is to develop a CPD that determines each set $S^{(i)}$ and therefore the values of each C_i subject to the segmentation $S^{(1)}, S^{(2)}, \ldots, S^{(l)}$ being optimal according to some optimality metric. This optimality constraint is abstractly defined here because it varies across different approaches

[1] Stationarity is rigorously defined in Chapter 2.

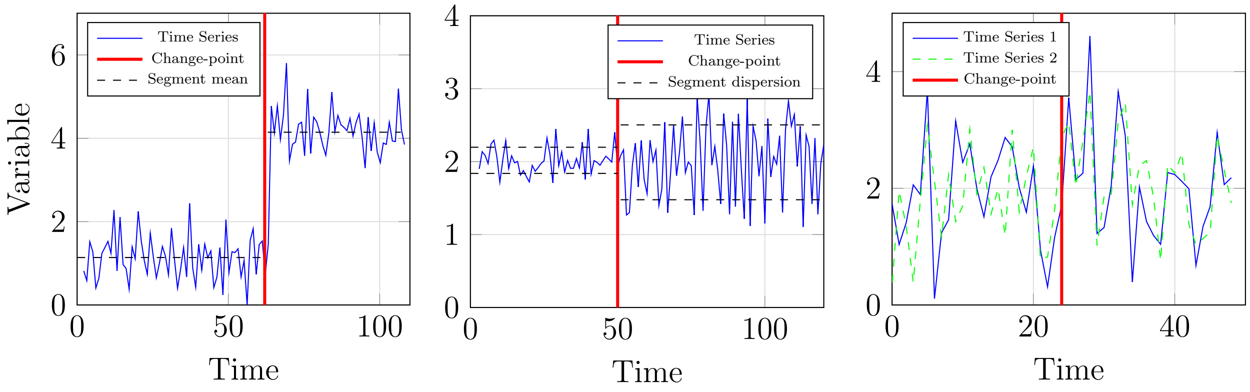


Figure 1.1: Change-points in mean (**left**), variance (**middle**) and covariance (**right**) of time series generated by samples from a Gaussian distribution.

adopted to solve the problem. A precise definition of this metric is provided in later chapters.

Apart from the optimality constraint mentioned above, one might define a "feasibility constraint" about each C_i. Specifically, each C_i corresponds to a point in time characterised by an abrupt change in a parameter and/or property of the time series (Aminikhanghahi and Cook, 2017; Page, 1954). At every change-point C_i at least one of the following properties of a collection time series must change:

- Mean.
- Variance.
- Covariance between any two time series.

Figure 1.1 illustrates changes in each of these properties on three separate plots. In real-world applications it is possible that a combination of properties of a time series changes. For instance, it is possible that both the mean and variance of a time series changes at the same time. It is therefore an essential property of a CPD to be able to detect multiple change-points at any instance of time.

1.2 Applications

To illustrate the usefulness of change-point detection we list two common applications of CPD in medicine and climate change (Aminikhanghahi and Cook, 2017).

In medical applications CPD appears useful in autonomous patient monitoring. By identifying changes in physiological variables such as heart rate and electrocardiogram, patients can be supervised in real-time (Staudacher et al., 2005). The understanding of brain activity as well as the study of sleeping patterns, epilepsy and other conditions can be aided by the use of CPD systems. An example of CPD in real-time heart rate monitoring is shown in Figure 1.2.

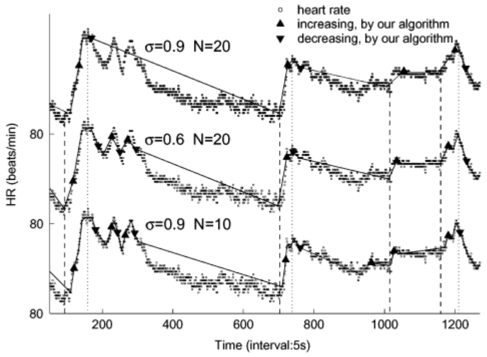


Figure 1.2: Change point detection in heart rate monitoring with different sensitivity control σ and different window size N (Staudacher et al., 2005).

Another field in which CPD gained popularity is climate change. CPD techniques focus on observing possible changes in the climate by examining increases in greenhouse gases (Aminikhanghahi and Cook, 2017) changes in the concentration of CO_2 levels. In a case study about the atmospheric carbon dioxide concentrations at Mauna Loa (Beaulieu et al., 2012), CPD techniques identified a shift of model parameters in 1991, the same year an in Mount Pinatubo occurred.

The above two showcase examples demonstrate the immense benefits CPD can have on a variety of unrelated scientific fields. A more detailed presentation of real-world applications can be found in Chapter 5, where the current CPD system is applied to crime, real estate and finance.

1.3 Literature

Due to the pertinence and applicability of change-point detection in various fields, change-point problems have been considered in a variety of settings. One area of research (Xuan, 2007; Xuan and Murphy, 2007) has been devoted to the *off-line* or *retrospective* or *batch-mode* version of the problem, where the whole dataset is available for processing. There are also *on-line* approaches (Fearnhead and Liu, 2007; Adams and MacKay, 2007) that process data in batches of size one, i.e. sequentially. In such approaches every time a new datum is received the algorithm performs updates based on previously stored information.

The first attempt to conceptualise and tackle change-point problems was made by (Page, 1954) using a frequentist framework. In his paper Page motivated the need to detect change-points in time series and proposed an on-line framework that identifies changes in a parameter θ of a time series. In particular, Page suggested the use of moving averages to model the time series' mean and then employed hypothesis testing to examine whether a CP has occurred or not. He also utilised control charts and the cumulative sum (CUMSUM) statistic to identify change-points along with many authors that adapted his approach (Jr. and Samuel, 2001; JR, 2007; Khoo, 2004; Taylor, 2000). Other frequentist techniques that emerged in the early years of CPD focused on estimating the abundance of CPs in a time series. These include the works of (Yao, 1988) and (Lee, 1995) who proposed the use of a penalised least-squares estimator for the number of CPs in a time series.

Once probabilistic approaches become more established in change-point problems, research was driven towards Maximum Likelihood Estimation (MLE) (Samuel et al., 1998; Fahmy and Elsayed, 2006) of CPs and Likelihood Ratio (LR) tests (Pignatiello and Simpson, 2002; Mahmoud et al., 2006) applied on CUMSUM statistics. Another promising approach that reduced CPD to time series outlier detection was the *Change Finder* method (Yamanishi and Takeuchi, 2002), which modelled the time series as an Autoregressive (AR) process. An adaptation to the Change Finder method was developed by (Liu et al., 2013) and (Kawahara and Sugiyama, 2011) and was inspired by LR tests. Their work proposed a non-parametric technique of computing the ratio of densities, which is a measure of dissimilarity between successive segments of a time series. A high dissimilarity measure implied that the existence of a CP between the two segments was likely.

On the Bayesian front, a first approach based on Page's work was introduced by (Smith, 1975). Smith applied Bayesian inference on the location of CPs and exemplified his framework using the cases of Binomial and Gaussian distributions. Despite the fact that this framework was suitable for retrospective segmentation, he also informally illustrated a way of extending his Bayesian analysis to sequential data. Another Bayesian off-line framework was developed by Xuan and Murphy in 2007 (Xuan and Murphy, 2007) which introduced inference on dependent multi-dimensional time series.

It wasn't until the works of (Fearnhead and Liu, 2007) and (Adams and MacKay, 2007) in 2007 that CPD was more rigorously defined in an on-line setting. The latter authors improved upon the ideas of the former, which involved deriving recursions that can be implemented efficiently using a dynamic programming paradigm. Another notable mention is the work of (Barry and Hartigan, 1992) on Product Partition Models (PPMs) that constituted the building blocks of many pioneering improvements in CPD including Adams and MacKay's and Xuan and Murphy's work. The work of on-line change-point detection culminates in Knoblauch and Damoulas' paper on Bayesian CPD in a spatio-temporal setting (Knoblauch and Damoulas, 2017). Notable competitive methods include the one by (Saatçi et al., 2010) which employs Gaussian Processes to create a non-parametric time series model for CPD.

For completeness, it is worth mentioning that kernel (Harchaoui et al., 2009), graph-based (Chen and Zhang, 2012), and clustering methods (Keogh et al., 2001) were also employed in a CPD context, but received comparatively less attention.

Regarding point processes (PP) the use of which is discussed in Chapter 3, relevant work was done by (Byrd et al., 2017). They proposed an l-lag exact on-line CPD algorithm that utilises point process models. The actual PP models were derived in a 1985 paper by (Nelson, 1985b), who extended univariate models to a multivariate framework.

1.4 Scope

The current CPD system tackles change-point problems in a broader framework than the one described in the problem definition. Specifically, CPD is applied on multi-dimensional data instead of uni-dimensional. In terms of notation, this translates to defining each y_i as a k-dimensional vector $\boldsymbol{y}_i$ and $\boldsymbol{y}_{1:t}$ as a $(t \times k)$ matrix $\boldsymbol{Y}_{1:t}$. Multi-dimensional data can be either dependent or independent collections of time series. In addition to the ordering of data in time, there is also a spatial arrangement of data in an effort to encode dependencies between different data-generating processes. This type of data is known as *spatio-temporal*.

Furthermore, the CPD algorithm is executed sequentially on data (on-line). As a result, there is an attempt to make effective use of models and data structures to upper bound the computational and storage complexities. Another implicit consequence of operating in an on-line setting is that the range of possible models is limited[2]. Also, Bayesian approaches are highly compatible with on-line CPD frameworks. For that reason, Bayesian inference techniques are employed to provide a solution to this class of change-point problems. In a Bayesian context, CPs are *inferred* or *estimated* based on updated beliefs about the location of the CPs and the models describing the data-generating process. Hence, in exact terms the algorithm developed performs *change-point estimation* (CPE) instead of CPD. Finally, there is an attempt to consider a set of potential models for modelling the data instead of only one model. This gives rise to the need for selecting the most appropriate model at any time instance. The proposed algorithm exploits the properties of the Bayesian framework to incorporate model selection in the algorithm, as inspired by (Fearnhead and Liu, 2007).

Given the scope specified above, the algorithm developed performs a Bayesian On-line Change-point Detection and Model Selection, abbreviated as BOCDMS.

1.4.1 Contributions

The work by Knoblauch and Damoulas has so far been restricted to continuous data. This dissertation implements the following extensions, details of which are provided in later chapters:

[2] We discuss this thoroughly in Chapter 2.

- Extending the current framework to point processes.
- Implementing two models for count and categorical data on both a univariate and multivariate setting.
- Assessing model sensitivity and stability in a formal analysis.
- Testing each model on three real-world datasets.

1.5 Synopsis

The second chapter of the dissertation outlines the building blocks of the CPD algorithm with reference to the assumptions and concepts that led to its development. It proceeds by defining important quantities of the algorithm and deriving key equations. Then, the pseudocode of the algorithm is provided and its computational complexity is analysed. In chapter 3, the two models for point processes are defined and are tested on synthetic datasets. Following each model description, there is a discussion of the model's sensitivity to prior specification and an evaluation of the model's performance in pathological cases. The Chapter ends by drawing comparisons between the two models and listing a models-specific version of the BOCDMS algorithm. Chapter 4 is devoted to applying each model to three real-world datasets and evaluating its performance. Finally, Chapter 5 summarises the conclusions of the dissertation and proposes future extensions to the current algorithm. This chapter also attempts to evaluate the project and consider legal, ethical and professional issues that emerged.

Chapter 2

Change-point detection framework

Objectives:

✓ Introducing basic concepts in TSA.

✓ Delineating the building blocks of the current CPD framework.

✓ Describing and deriving important quantities in CPD.

✓ Citing the CPD algorithm and analysing its computational and space complexities.

2.1 Preliminaries

We start this chapter by defining some fundamental concepts in time series analysis.

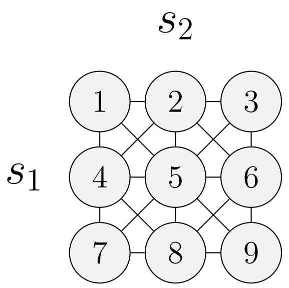


Figure 2.1: Graphical illustration of a collection nine TS models with spatial dimensions $s_1 = 3$, $s_2 = 3$.

Let $\boldsymbol{y}_{1:t} := (y_1, \ldots, y_t)$ be a one-dimensional time series $\forall y_i \in \mathbb{R}, 1 \leq i \leq t < \infty$. Then, let $\boldsymbol{y}_t$ be a $(s \times 1)$ vector denoting a collection of one-dimensional time series at time t for some $s \in \mathbb{N}$, or in other words a s-dimensional (multivariate) TS. In a spatio-temporal series each one of the $s := s_1 \times s_2$ TS can assume a spatial representation, such as the $s_1 \times s_2$ fully-connected spatial grid shown in Figure 2.1. The notation is then adjusted to account for the spatial structure by denoting the s-dimensional TS vector as a

$(s_1 \times s_2)$ matrix $\boldsymbol{Y}_t$, where Y_{ij} is the one-dimensional TS located in the (i, j)-th position of the grid at time t for $i \in \{1, \ldots, s_1\}, j \in \{1, \ldots, s_2\}$. Note that there is a direct mapping of the associations between TS to the spatial grid in Figure 2.1. In that mapping, edges appearing in the grid represent an association between two vertex TS. According to the structure of the graph in Figure 2.1, each TS is only correlated with TS that are in its close vicinity. This graphical structure can be exploited when studying real-world applications of spatio-temporal time series.

A time series is said to be stationary if it is *strictly* and/or *weakly* stationary.

A time series is weakly stationary (Hamilton, 1994) if its mean and auto-covariance are independent of time t:

$$\begin{aligned} \mathbb{E}(y_t) &= \mu \quad \forall t \in \mathbb{N} \\ Cov(y_t, y_{t+j}) &= \gamma_j \quad \forall t, j \in \mathbb{N}. \end{aligned}$$

A time series is strictly stationary (Hamilton, 1994) if the joint distribution of $(y_{i+1}, \ldots, y_{i+t})$ depends only on i $\forall t \in \mathbb{N}$:

$$(y_1, \ldots, y_t) \overset{d}{=} (y_{i+1}, \ldots, y_{i+t}) \quad \forall i \in \mathbb{N}.$$

Stationarity can be extended to spatio-temporal series: A spatio-temporal series is stationary if and only if all its component time series are stationary. The importance of a stationary process lies on the fact that its future can be modelled in the same way as its past, which simplifies prediction significantly.

Furthermore, it is important to define some additional tools in probability that are going to be useful for the rest of this chapter: law of total probability, Bayes' theorem, independence and conjugacy. The law of total probability (LTP) states that for $\{B_n : n \in \mathbb{N}\}$ a finite or countably infinite partition of a sample space Ω and an event $A \subseteq \Omega$, it follows that

$$\mathbb{P}(A) = \sum_{j=1}^{n} \mathbb{P}(A \cap B_j).$$

LTP decomposes the probability of an event into a sum of mutually exclusive conjunctions of events, which is a convenient property exploited throughout the derivations of this chapter.

Additionally, Bayes' theorem states that for any two events A and B with $\mathbb{P}(B) > 0$,

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} \propto \mathbb{P}(B|A)\mathbb{P}(A).$$

Based on this theorem we can define the notion of *independence* between two events A and B. Two events A and B with $A, B \subseteq \Omega$ are independent if and only if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ or equivalently, $\mathbb{P}(A|B) = \mathbb{P}(A)$ or $\mathbb{P}(B|A) = \mathbb{P}(B)$. Also, a corollary to the law of total probability that uses Bayes' theorem is the following:

$$\mathbb{P}(A|C) = \sum_{j=1} \mathbb{P}(A|C \cap B_j)\mathbb{P}(B_j|C) \tag{2.1}$$

Bayes' theorem can also be extended to model random variables (RVs) instead of events. For two random variables X and Y, we can interpret the theorem in the following way. Let f,g,h be well-defined distributions. Suppose that $g(X)$ expresses a *prior* belief over the possible values of X and that $h(Y|X)$ denotes the *likelihood* of Y under regime X. Then, $f(X|Y) \propto h(Y|X)g(X)$ is the *posterior* belief about X in the light of new information Y. Alternatively, posterior belief $\propto$ likelihood × prior belief. This formulation of Bayes' theorem is rather intuitive and reflects a thought process many humans follow when presented with new information about the world.

If the prior distribution g and posterior distribution f belong to the same family of distributions, then g and f are said to be *conjugate* distributions. We say that g is a conjugate prior to the likelihood function h. Also, conjugate models are said to be *closed under sampling* or in closed form because when integrated over their posterior distribution has an analytic expression. Conjugacy has many consequences in the CPD developed, all of which are discussed in relevant sections of this chapter.

2.2 Main assumptions

The arguments posed in the remaining of this chapter apply for a generalised form of a time series and can be trivially extended to a spatio-temporal series.

The construction of a CPD framework is based upon two fundamental assumptions about the properties of TS segmentation. These assumptions give rise to a Bayesian on-line CPD framework of spatio-temporal series, as specified in the Scope of this dissertation in Chapter 1. One instrumental assumption that encompasses the approach adopted in this dissertation is that there is an underlying *data-generating process* (DGP) generating the TS data. In other words, data points are assumed to be samples from a probability distribution that governs the behaviour of the DGP. In Machine Learning terminology this is known as a *generative* approach of modelling data. According to the generative model, the "hidden" distribution of the DGP is known but the parameters of the model are unknown. Formally, we say that each datum $\boldsymbol{y}_t$ at time t is described by a model m_t of some distribution $p(\boldsymbol{y}_t|\boldsymbol{\theta}_{m_t})$ with some set of parameters $\theta \in \mathbb{R}^d$, where d is the dimension of the parameter space. Notationally, this is articulated as $\boldsymbol{y}_t \sim p(\boldsymbol{y}_t|\boldsymbol{\theta}_{m_t})$.

The second fundamental assertion made is that CPs are modelled as a *product partition model* (PPM), which was developed by (Barry and Hartigan, 1992) and used in (Adams and MacKay, 2007) on-line CPD paper. This assertion is predicated on the previously adopted premise that a generative view of the world reflects reality more accurately. According to the one-dimensional case of the PPM model, each CP C_i is randomly drawn from a product partition distribution. Conditionally on known the location of the CPs and therefore the segments of the TS, data from two different segments $S^{(i)}$ and $S^{(j)}$ are independent $\forall i \neq j = 1, \ldots, l$. Mathematically, the last condition can be expressed as

$$\mathbb{P}(\boldsymbol{y}_{1:t}|C_1, \ldots, C_m) = \prod_{i=1}^{l} \mathbb{P}(S^{(i)}). \tag{2.2}$$

An implicit consequence of this condition is that any two segments $S^{(i)}$ and $S^{(j)}$ are described by a different model $\forall i \neq j \in \mathbb{N}$. A different model may either imply a different distribution or the same distribution with different parameters. There is a disparity between this assumption and the one made in (Fearnhead and Liu, 2007), which states that the models describing any two segments and their parameters are both independent. Our set of assumptions is more similar to the ones in (Adams and MacKay, 2007), where the PPM model is assumed.

The PPM model can also be compared to a Hidden Markov Model (HMM) (Chib, 1998), that is a model specified by an finite space of hidden states and the transition probabilities between them. A PPM can be treated as a HMM by modelling the partition variables C_i's as hidden states (Xuan and Murphy, 2007). However, the PPM model has an unbounded number of states, which is why it is more similar to a HMM with infinite number of states. Also, in a PPM it is impossible to revisit a state whereas in a HMM model this transition often receives non-zero probability. Therefore, a HMM model is more suitable to applications where the number of regimes is known and therefore it is useful to model transitions between known regimes. In contrast, the PPM model is more flexible in that it does not assume that the number of segments is known a priori.

In addition to the two central assumptions above, we assume that the data of any given segment S_i are independent and identically distributed (iid) draws from the model governing the segment, $\forall i \in \{1, \ldots, l\}$. Moreover, the time series is assumed to be piecewise stationary, where each segment's data corresponds to a stationary sub-series.

2.3 Prior beliefs

The assumptions declared in the previous section are compatible with a Bayesian framework of CPD, as demonstrated in (Adams and MacKay, 2007). The most important of these assumptions rests on the idea that conditionally on knowing the location of the CPs, the modelling of the time series can be simplified. One such way of modelling the location C_i of a CP for some $i \in \{1, \ldots, l\}$ is to introduce a random variable r_t, called the *run-length*, denoting the length of the current segment $S^{(i)}$. For instance, $r_t = j$ implies that there is a CP at time $t - j$. Therefore, the run-length encodes information about the location of CPs and can be used to find the boundaries of segments. It is defined recursively as:

$$r_t = \begin{cases} 0 & \text{If there is a CP at time } t \\ r_{t-1}+1 & \text{If there is no CP at time } t. \end{cases} \tag{2.3}$$

The run-length is illustrated using a dummy TS example shown in Figure 2.2. This Figure depicts the "true" run-length whereas in practise the distribution over the run-length is maintained.

Another quantity of interest that is monitored in every time instance is the model m_t, which is also modelled as a r.v.. We define m_t such that it is the model describing $\boldsymbol{y}_{(t-r_t):t}$. Since every segment is governed by only one model, it follows that $r_t = j$ implies $m_{t-j} = \cdots = m_{t-1} = m_t$. A model consists of a conditional probability density $d\mathbb{P}(\boldsymbol{Y}_t|\boldsymbol{\theta}_m)$ on $\mathbb{R}^s$ and a parameter prior $d\mathbb{P}(\boldsymbol{\theta}_m)$ on $\boldsymbol{\Theta}_m \subseteq \mathbb{R}^d$ with fixed *hyper-parameters* $\boldsymbol{\theta}_m^0$ that can be optimized depending on the application. Hyper-parameters are used to calibrate the model by expressing a prior belief about the model parameters. Also, one model may not suffice to model the different types of data satisfactorily and therefore it is beneficial to define a *model universe* $\mathcal{M}$, that is a set of potential models with associated probability. At every iteration, the most probable model is chosen to model the DGP.

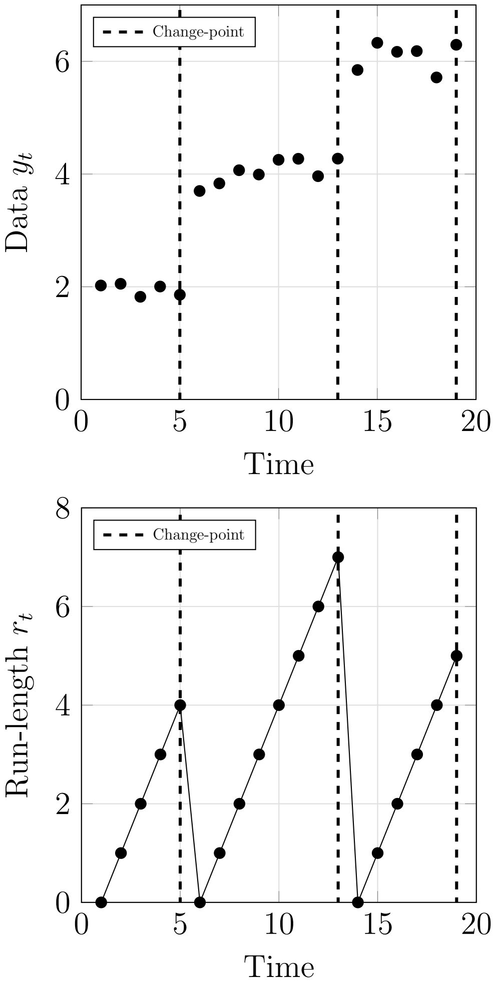


Figure 2.2: Plots of TS data with three change-points versus time (see top) and plots or run-length versus time (bottom).

Furthermore, there is no *a priori* knowledge about the relative abundance and location CPs as well as the models governing each segment. Since we are developing a Bayesian CPD framework, it is common to remedy this lack of knowledge by assigning prior beliefs to the random variables encoding that information. This idea was conceived independently by (Adams and MacKay, 2007) and (Fearnhead and Liu, 2007). Specifically, let g, h

be probability mass functions (PMFs) and π_{m_t}, f_{m_t} be probability density functions $\forall m_t \in \mathcal{M}$ and suppose that

$$
\begin{aligned}
r_t &\sim g(r_t) && \forall\ r_t \in \mathbb{N}, && (2.4)\\
m_t|r_t, m_{t-1} &\sim q(m_t|r_t, m_{t-1}) && \forall\ m_t \in \mathcal{M}, && (2.5)\\
\boldsymbol{\theta}_{m_t}|m_t &\sim \pi_{m_t}(\boldsymbol{\theta}_{m_t}|m_t) && \forall\ \boldsymbol{\theta}_{m_t} \in \boldsymbol{\Theta}_m \subseteq \mathbb{R}^d, && (2.6)\\
\boldsymbol{y}_t|m_t, \boldsymbol{\theta}_{m_t} &\sim f_{m_t}(\boldsymbol{y}_t|m_t, \boldsymbol{\theta}_{m_t}) && \forall\ \boldsymbol{y}_t \in \mathbb{R}^s. && (2.7)
\end{aligned}
$$

In the above prior specification $\boldsymbol{y}_t$ denotes the $(s \times 1)$ vector of one-dimensional TS at time $t \in \mathbb{N}$.

Next we define the transition probabilities for r_t and m_t for $|\mathcal{M}|= 1$ as defined in Adams and MacKay (2007). Given a Hazard function $H : \mathbb{N} \mapsto [0, 1]$, and model prior $q : \mathcal{M} \mapsto [0, 1]$, the transition probabilities are

$$
\mathbb{P}(r_t|r_{t-1}) = \begin{cases} 1 - H(r_{t-1} + 1) & \text{if } r_t > 0 \\ H(r_{t-1} + 1) & \text{if } r_t = 0 \\ 0 & \text{otherwise} \end{cases} \qquad (2.8)
$$

$$
\mathbb{P}(m_t|m_{t-1}, r_t) = \begin{cases} \mathbb{1}_{m_{t-1}}(m_t) & \text{if } r_t > 0 \\ q(m_t) & \text{if } r_t = 0 \\ 0 & \text{otherwise} \end{cases} \qquad (2.9)
$$

For a probability density function $P(x)$ and its distribution function $F(x)$ of a random variable X, the Hazard function H is defined (Evans et al., 2011) as follows:

$$
H(x) = \frac{P(x)}{1 - F(x)}
$$

$\forall x \in supp(X)$. In the case where $P(x)$ belongs in the exponential family of distributions, the process is memoryless and the Hazard function is constant for all $x \in supp(X)$ (Adams and MacKay, 2007). We choose $P(x)$ to be a Geometric distribution with constant intensity λ, where λ is a hyper-parameter. Overall, the Hazard function allows us to express a prior belief about the probability of CP occurring at a given time t. By default, we set $\lambda = 30$ in all applications considered as this belief has little effect on posterior beliefs given a sufficiently large number of observations.

Moreover, for a framework with more than one potential models ($|\mathcal{M}| > 1$) equation 2.10 is modified as follows:

$$\mathbb{P}(m_t|m_{t-1}, r_t) = \begin{cases} \mathbb{1}_{m_{t-1}}(m_t)\mathbb{P}(m_{t-1}|\boldsymbol{y}_{1:(t-1)}, r_t) & \text{if } r_t > 0 \\ q(m_t) & \text{if } r_t = 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.10}$$

The above equations indicate that a change of model occurs only in the presence of a CP, in which case a model is sampled from the model prior q.

2.4 Useful recursions

The CPD algorithm uses the quantities defined in the previous section to compute predictions and estimate the most probable segmentation. We proceed by listing supplementary quantities required to achieve these two tasks and deriving recursive equations for their computation.

At every time step t, the algorithm computes for all models and run-lengths the posterior predictive density of $\boldsymbol{y}_t$ conditional on the model, run-length and previous data:

$$d\mathbb{P}(\boldsymbol{y}_t|\boldsymbol{y}_{1:(t-1)}, m_t, r_t) = \int_{\Theta_{m_t}} \underbrace{d\mathbb{P}(\boldsymbol{y}_t|\boldsymbol{\theta}_{m_t})}_{\text{likelihood}} \underbrace{d\mathbb{P}(\boldsymbol{\theta}_{m_t}|\boldsymbol{y}_{(t-r_t):(t-1)})}_{\text{parameter posterior}} d\boldsymbol{\theta}_{m_t} \tag{2.11}$$

The above equation follows directly by the LTP in 2.1. Under the PPM assumption, we can all ignore data from previous segments conditional on the run-length in the parameter posterior. Hence, the interval $(t - r_t) : (t - 1)$ corresponds to the current segment before receiving the new datum $\boldsymbol{y}_t$ at time t. Since the parameter space $\boldsymbol{\Theta}_{m_t}$ is continuous we integrate over the parameters instead of summing over. In Bayesian analysis this is known as marginalising out parameter uncertainty since the parameter does not appear in the LHS of the equation.

This is achieved by "summing" over the possible parameter values and computing the product shown in the RHS. Unless conjugacy is exploited (Xuan and Murphy, 2007), the calculation of the integral is non-trivial. This dissertation focuses on obtaining exact expressions for 2.11 in a computationally efficient way and therefore utilises conjugate parametric models. If a conjugate model is used, then the product in 2.11 will have a closed form and the integral can be computed efficiently. Conjugacy is also exploited in (Adams and MacKay, 2007) but is only limited to conjugate exponential models. We remove the restriction of the models belonging in the exponential family of distributions. However, it is worth mentioning that there are approximations to the integral in 2.11 for non-conjugate models using sequential Monte Carlo techniques explored in Turcotte and Heard (2015).

Another recursion that can be efficiently computed using conjugate models is the joint distribution of the data, run-length and model:

$$\begin{aligned} d\mathbb{P}(\boldsymbol{y}_{1:t}, r_t, m_t) = \sum_{m_{t-1}} \sum_{r_{t-1}} \{ & \underbrace{d\mathbb{P}(\boldsymbol{y}_t | \boldsymbol{y}_{1:(t-1)}, r_t, m_t)}_{\text{equation 2.11}} \underbrace{\mathbb{P}(m_t | \boldsymbol{y}_{1:(t-1)}, r_t, m_{t-1})}_{\text{Model transition in 2.10}} \\ & \underbrace{\mathbb{P}(r_t | r_{t-1})}_{\text{equation 2.8}} \underbrace{d\mathbb{P}(\boldsymbol{y}_{1:(t-1)}, r_{t-1}, m_{t-1})}_{\text{Recursive term}} \} \end{aligned} \tag{2.12}$$

The above equation is used to assess whether a CP has occurred or not. In an on-line Bayesian setting this hypothesis is tested by computing the CP probability (run-length becomes zero) and its complement (run-length grows by a unit). These are referred to as CP and growth probabilities respectively and their calculations are shown below:

$$\begin{aligned} d\mathbb{P}(\boldsymbol{y}_{1:t}, r_t = 0, m_t) &= d\mathbb{P}(\boldsymbol{y}_t | \boldsymbol{y}_{1:(t-1)}, r_t, m_t) q(m_t) \\ &\quad \times \sum_{m_{t-1}} \sum_{r_{t-1}} \{ H(r_{t-1} + 1) d\mathbb{P}(\boldsymbol{y}_{1:(t-1)}, r_{t-1}, m_{t-1}) \} \qquad (2.13) \\ d\mathbb{P}(\boldsymbol{y}_{1:t}, r_t = r_{t-1} + 1, m_t) &= d\mathbb{P}(\boldsymbol{y}_t | \boldsymbol{y}_{1:(t-1)}, r_t, m_t) \mathbb{P}(m_{t-1} | \boldsymbol{y}_{1:(t-1)}, r_t) \\ &\quad \times (1 - H(r_t)) d\mathbb{P}(\boldsymbol{y}_{1:(t-1)}, r_{t-1}, m_{t-1}) \qquad (2.14) \end{aligned}$$

These expressions follow by substituting the appropriate run-length case from 2.8, 2.10 into 2.12. The rest of the calculations can be performed using the equations presented above. The marginal likelihood of the data known as *evidence* is equal to

$$d\mathbb{P}(\boldsymbol{y}_{1:t}) = \sum_{m_t}\sum_{r_t}\underbrace{d\mathbb{P}(\boldsymbol{y}_{1:t}, m_t, r_t)}_{\text{equation 2.12}} \tag{2.15}$$

Based on the evidence, we can compute the following mass functions:

$$\mathbb{P}(r_t, m_t|\boldsymbol{y}_{1:t}) = \frac{\overbrace{d\mathbb{P}(\boldsymbol{y}_{1:t}, r_t, m_t)}^{\text{equation 2.12}}}{\underbrace{d\mathbb{P}(\boldsymbol{y}_{1:t})}_{\text{equation 2.15}}} \tag{2.16}$$

$$\mathbb{P}(m_t|\boldsymbol{y}_{1:t}) = \sum_{r_t}\underbrace{\mathbb{P}(r_t, m_t|\boldsymbol{y}_{1:t})}_{\text{equation 2.16}} \tag{2.17}$$

$$\mathbb{P}(r_t|m_t, \boldsymbol{y}_{1:t}) = \frac{\overbrace{d\mathbb{P}(r_t, m_t|\boldsymbol{y}_{1:t})}^{\text{equation 2.16}}}{\underbrace{d\mathbb{P}(m_t|\boldsymbol{y}_{1:t})}_{\text{equation 2.17}}} \tag{2.18}$$

$$\mathbb{P}(r_t|\boldsymbol{y}_{1:t}) = \sum_{m_t}\underbrace{\mathbb{P}(r_t, m_t|\boldsymbol{y}_{1:t})}_{\text{equation 2.16}} \tag{2.19}$$

$$\mathbb{P}(m_{t-1}|\boldsymbol{y}_{1:(t-1)}, r_t) = \frac{\overbrace{\mathbb{P}(r_{t-1}, m_{t-1}|\boldsymbol{y}_{1:(t-1)})}^{\text{equation 2.16}}}{\underbrace{\mathbb{P}(r_{t-1}|\boldsymbol{y}_{1:(t-1)})}_{\text{equation 2.19}}} \tag{2.20}$$

Equations 2.16, 2.18 and 2.20 follow immediately by applying Bayes' theorem while 2.17 and 2.19 follow by law of total probability. The names of the quantities above order from top to bottom are the joint model and run-length distribution, the model posterior, the model-specific run-length posterior, the run-length posterior, and the conditional model posterior used in equation 2.10. It is important to note that 2.20 was utilised for the calculation in equation 2.10.

2.5 Prediction and Segmentation

The purpose of computing and storing the quantities above is to implement prediction and segmentation at the end of each time step t. The task of prediction or forecasting is also performed in a Bayesian setting for an arbitrary number of steps h ahead of the current time. For example, a two-step ahead prediction corresponds to forecasting $\boldsymbol{y}_{t+2}$. We denote the h-step ahead prediction as $\widehat{\boldsymbol{Y}}_{t+h}$, which is not to be confused with the spatial matrix of the multivariate time series $\boldsymbol{Y}_t$. We also define an additional quantity to simplify the forecasting expression. Let $\widehat{\boldsymbol{y}}_t^h$ be the predictive interval, that is the vector of predictions for a specified period period of time $t+1$ to $t+h-1$. Then,

$$\widehat{\boldsymbol{y}}_t^h = \begin{cases} \emptyset & \text{if } h = 1 \\ \widehat{\boldsymbol{y}}_{(t+1):(t+h-1)} & \text{otherwise} \end{cases}$$

The outcome $\widehat{\boldsymbol{y}}_{t+h}$ of the h-step ahead forecast is equal to $\mathbb{E}(\widehat{\boldsymbol{Y}}_{t+h}|\boldsymbol{y}_{1:t}, \widehat{\boldsymbol{y}}_t^h)$, the posterior expectation of the h-step ahead prediction. This expectation is computed with respect to the posterior predictive distribution of $\widehat{\boldsymbol{Y}}_{t+h}$

$$\mathbb{P}(\widehat{\boldsymbol{Y}}_{t+h}|\boldsymbol{y}_{1:t}) = \sum_{r_t}\sum_{m_t}\{\underbrace{d\mathbb{P}(\widehat{\boldsymbol{Y}}_{t+h}|\boldsymbol{y}_{1:t}, \widehat{\boldsymbol{y}}_t^h, r_t, m_t)}_{\text{equation 2.11}}\underbrace{d\mathbb{P}(r_t, m_t|\boldsymbol{y}_{1:t})}_{\text{equation 2.16}}\} \tag{2.21}$$

We expect prediction that are far ahead in the future to be less accurate than near-future predictions because the dynamics governing the data-generating process are more likely to change over a longer period of time. For that reason the most accurate forecast is of a TS is its posterior expectation.

The task of segmentation also follows the Bayesian paradigm. In the introduction of the change-point problem in Chapter 1, we claimed that the segmentation must be optimal in some way. Optimality is judged by considering all possible segmentations and choosing the most likely conditional on the current data. This is the analogue of maximum likelihood estimation in a Bayesian setting and is called *Maximum A Posteriori* (MAP) segmentation.

MAP segmentation was inspired by (Fearnhead and Liu, 2007) and is defined as follows:

$$MAP_t = \max_{r,m} \{d\mathbb{P}(\boldsymbol{y}_{1:t}, r_t = r, m_t = m) MAP_{t-r-1}\}, \tag{2.22}$$

with $MAP_0 = 1$. The MAP estimators for the run-length and model are

$$(r_t^*, m_t^*) = \operatorname*{argmax}_{r,m} \{d\mathbb{P}(\boldsymbol{y}_{1:t}, r_t = r, m_t = m) MAP_{t-r-1}\},$$

with $(r_0^*, m_0^*) = (0, m_0)$, where m_0 is a randomly chosen model from $\mathcal{M}$. Let S_t be the segment at time t containing the elements of the current segment, which is notationally different than the definition of $S^{(k)}$ for some $k \in \{1, \dots, l\}$. Then, the segmentation is defined as $S_t = S_{t-r_t^*-1} \cup \{(t - r_t^*, m_t^*)\}$ with $S_0 = \emptyset$, where $(t', m_{t'}) \in S_t$ denotes a CP at time $t' \leq t$ with $m_{t'} \in \mathcal{M}$ the model for $\boldsymbol{y}_{\boldsymbol{t'}\,:\,\boldsymbol{t}}$.

The MAP segmentation algorithm by (Fearnhead and Liu, 2007) is very similar to a forwards-filtering backwards-sampling algorithm for HMMs (Scott, 2002), where the “hidden variable” is a time index encoding where the location of the last CP.

2.6 Algorithm

The computations of the quantities described in the previous sections are compiled into Algorithm 1 shown below.

Every time a new datum $\boldsymbol{y}_t$ is received, the algorithm initialises or updates the growth and CP probabilities for every possible model in $\mathcal{O}(1)$, contributing $\mathcal{O}(|\mathcal{M}|)$ to the overall complexity of the algorithm. This is because both $d\mathbb{P}(\boldsymbol{y}_{1:t}, r_t = 0, m_t = m)$ and $d\mathbb{P}(\boldsymbol{y}_{1:t}, r_t, m_t = m)$ are based on equation 2.12. By storing this joint density for every possible run-length allows its computation to be done recursively and efficiently using a dynamic programming paradigm. The overhead of this computation is a $\mathcal{O}(|\mathcal{M}|t)$ space complexity, where t is the time of the last observation. The joint density in 2.12 also requires by definition the computation of 2.8, 2.10 and 2.11. Quantities 2.8 and 2.10 are both computed in $\mathcal{O}(1)$ at each iteration using the Hazard function and the model prior, respectively. Equation 2.11 can also be computed in $\mathcal{O}(1)$ under the assumption that only conjugate models are employed, as explained before.

Algorithm 1 Bayesian On-line Change-point Detection with Model Selection (BOCDMS)

```
Input at time 0: model universe $\mathcal{M}$, Hazard $H$, prior $q$.
Input at time $t$: next observation $\boldsymbol{y}_t$.
Output at time $t$: prediction $\widehat{\boldsymbol{y}}_{(t+1):(t+h_{max})}$, segment $S_t$, model posterior $\mathbb{P}(m_t|\boldsymbol{y}_{1:t})$.
for next observation $\boldsymbol{y}_t$ at time $t$ do
    # Step 1: Calculate model-specific quantities
    for $m \in \mathcal{M}$ do
        if $t = 1$ then
            Initialise $d\mathbb{P}(\boldsymbol{y}_{1:t}, r_t = 0, m_t = m)$ with prior
        else
            Update $d\mathbb{P}(\boldsymbol{y}_{1:t}, r_t, m_t = m)$ using 2.13 and 2.14
            Prune model-specific run-length distribution
        end if
    end for
    # Step 2: Aggregate over models
    Obtain joint distribution over $\mathcal{M}$ via 2.15-2.20
    Perform prediction and segmentation using 2.21, 2.22
    Update model parameters $\boldsymbol{\theta}_{m_t}$ for all models $m_t \in \mathcal{M}$
    Output: $\widehat{\boldsymbol{y}}_{(t+1):(t+h_{max})}$, $S_t$, $\mathbb{P}(m_t|\boldsymbol{y}_{1:t})$.
end for
```

Furthermore, the computation of the mass functions 2.15-2.20 can be achieved by summing over all possible run-lengths and models giving a complexity of $\mathcal{O}(|\mathcal{M}|t)$. On the other hand, 2.16-2.20 can be achieved in $\mathcal{O}(1)$ by calculating these quantities in the order they are listed. This is because each one of 2.16 - 2.20 is a multiplicative factor of the quantities listed above it or of quantities stored in the previous step of the algorithm, such as the joint distribution of the model, run-length and data in 2.12. This type of calculation entails minimal storage overheads and is therefore computationally efficient.

Moreover, prediction can also be achieved in $\mathcal{O}(1)$ as it requires the calculation of the posterior predictive in 2.11 and the joint model and run-length distribution in 2.16. The former quantity requires $\mathcal{O}(1)$ time due to use of conjugacy while the latter is available from the previous step of the algorithm. However, segmentation is based on a more burdensome calculation shown in 2.22. Maximising over the run-length r_t and model m_t takes $\mathcal{O}(\max(|\mathcal{M}|, t)) = \mathcal{O}(t)$ for large t. Finally, model parameter updates can be performed in $\mathcal{O}(1)$ using conjugate models. Therefore, in the worst-case scenario, each iteration of the algorithm would take $\mathcal{O}(|\mathcal{M}|t)$ resulting in an overall time and space complexities of $\mathcal{O}(|\mathcal{M}|t^2)$ and $\mathcal{O}(|\mathcal{M}|t)$.

2.6.1 Extensions

To avoid the heavy computational burden induced by the segmentation computation, it was proposed that the run-length distribution is trimmed. This would result in maximising over the pruned run-length instead of all possible run-lengths, which are in the order of t. The effect of pruning the run-length distribution would be a constant in time calculation of the MAP estimate instead of $\mathcal{O}(t)$. Reductions in complexity can also be found in equations such as 2.15 and 2.21, where we sum over all possible run-lengths.

An intuitive proposal is to discard all run-lengths whose probability is less than some threshold $1/R_{max}$ for some $R_{max} \in \mathbb{R}$ or to keep only the R_{max} most probable run-lengths, as suggested in (Adams and MacKay, 2007). The latter suggestion guarantees an upper bound of R_{max} on the number of run-lengths, where $\mathcal{O}(R_{max})$ is constant in time. Another proposal that was utilised in (Fearnhead and Liu, 2007) is called *Stratified Rejection Control*, which had comparatively similar performance to the two previous approaches. The BOCDMS algorithm prunes on the model-specific run-length distribution $d\mathbb{P}(r_t|m_t, \boldsymbol{y}_t)$ instead of $d\mathbb{P}(r_t|\boldsymbol{y}_t)$ in an attempt to encode knowledge about the models in the pruning process.

As a result of this pruning, the MAP estimate's computational complexity was reduced to $\mathcal{O}(|\mathcal{M}|R_{max})$ and therefore the running-time complexity of each iteration of Algorithm 1 is $\mathcal{O}(|\mathcal{M}|R_{max})$. For small model universes this complexity yields very fast solutions to online CPD problems. Overall, the time complexity is linear with time $\mathcal{O}(|\mathcal{M}|R_{max}t)$ with a storage cost of $\mathcal{O}(|\mathcal{M}|R_{max})$. In comparison, on-line approaches involving Gaussian Processes (GP), such as the one by (Saatçi et al., 2010), have a complexity of $\mathcal{O}(R_{max}^3 t)$. The on-line approaches by (Adams and MacKay, 2007) and (Fearnhead and Liu, 2007) have a performance linear in the number of time or number of observations. Therefore, the current BOCDMS algorithm has a faster performance compared to competitive on-line CPD algorithms.

Chapter 3

Spatio-temporal point processes

Objectives:

✓ Introducing basic concepts in point processes.

✓ Outlining the Poisson Gamma and Multinomial-Dirichlet models.

✓ Conducting Bayesian analysis for each of the two models.

✓ Performing a sensitivity analysis of each model.

✓ Addressing the limitations of each model.

Although CPD is frequently applied on continuous data, as in (Adams and MacKay, 2007) and (Fearnhead and Liu, 2007), there is a number of applications in areas such as ecology, seismology and material science (Møller and Waagepetersen, 2007) that use event-based data. This chapter explores two models that are applied on event-based data using point processes (PP).

3.1 Preliminaries

We devote this introductory section to providing details about important concepts in point processes.

Contrary to continuous data, a point process is a model used to study the occurrence of *events* in time (Weil, 2007). To avoid introducing measure theoretic results in this dissertation, the precise definition of a PP is omitted. For the purposes of this chapter, a PP can be thought of as a mapping between two well-defined finite measurable spaces. A simple one-dimensional point process is shown in Figure 3.1, which depicts the occurrence of events (dots) along time (line). In the context of a real-world application, each dot in Figure 3.1 may represent the event that a call was received in a call centre while the location of the dots encodes the time a call was received. In this type of modelling of PPs, these events or dots are modelled as *arrival times*. Each dot corresponds to a time T_i for $1 \leq t \leq t$, where $T_i < T_j \; \forall i < j$. Also, each T_i is modelled as a random variable. However, the restriction imposed by $T_i < T_j$ implies that there is a strong dependence between T_i's and therefore this type of mathematical modelling may become convoluted.

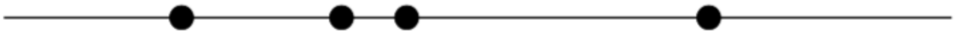


Figure 3.1: One-dimensional point process (Weil, 2007).

An alternative way of modelling point processes is to study *inter-arrival times*, that is the intervals between successive arrival times $S_i = T_{i+1} - T_i$ [1].These have important theoretical guarantees, such as the fact that $S_1, \ldots, S_t$ are independent and identically distributed Poisson random variables.

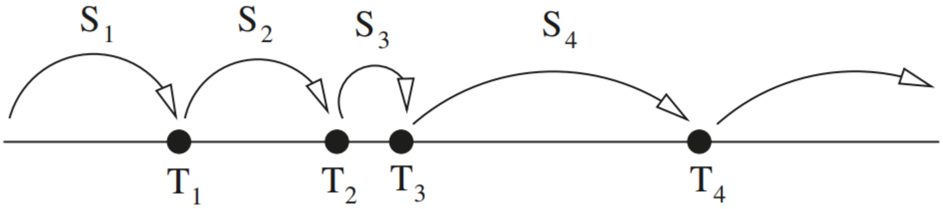


Figure 3.2: Arrival and inter-arrival times model of one-dimensional point process (Weil, 2007).

A third way of modelling PPs is to "formulate a point process in terms of the cumulative *counting process*" (Weil, 2007). The cumulative counting process N_t is defined as the number

[1] There is a notation overlap here. The definition of inter-arrival time S_i is different from the definition of segment S_i defined in the previous chapters.

of points arriving up to time t:

$$N_t = \sum_{i=1}^{\infty} \mathbb{1}\{T_i \leq t\},$$

$\forall t \geq 1$. A simplification of the cumulative counting process is to use interval counts $N(a, b] := N_b - N_a$ for $0 \leq a \leq b$. Instead of counting number of events cumulatively, the number of events is counted on given intervals of time. Interval counts also have useful properties if they are applied on disjoint intervals of time. An illustration of that type of counting is shown in Figure 3.3. In the call centre example, the analogue to interval counts may be the number of calls received by a centre every five minutes.

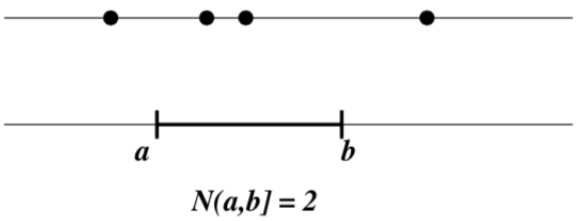


Figure 3.3: Interval counts on one-dimensional point process (Weil, 2007).

Since the focus of CPD detection is to model spatio-temporal processes, it is important to define a point process in a spatio-temporal setting. A *spatial point process* $\boldsymbol{X}$ is "a finite random subset of a given bounded region $S \subset \mathbb{R}^2$, and a realization of such a process is a spatial point pattern $\boldsymbol{x} = \{x_1, \dots, x_n\}$ of $n \geq 0$ points contained in S", as defined in (Møller and Waagepetersen, 2007). The number of points of a compact (bounded and closed) region or set $B \subseteq S$ is denoted by the random variable $N(B)$. A spatial PP is shown in Figure 3.4. The advantage of using interval or region counts over arrival and inter-arrival times is that there is no analogue of the latter in higher dimensions. Also, there is a clear distinction between a spatial PP and a spatio-temporal PP. A spatio-temporal point process is a collection of spatial PPs ordered in time. In terms of a spatio-temporal PP, Figure 3.4 constitutes a snapshot in time of that PP.

Additionally, it might be the case that more than one types of events occur within the same process. Therefore, it is intuitive to encode this extra information by labelling events using marks or ticks. These type of point processes are known as *marked* point processes. Mathematically, this translates to modelling each point in the realisation of the process as a tuple (x_i, l_i), where m_i is the label or mark of point i. In the example of the call centre, the

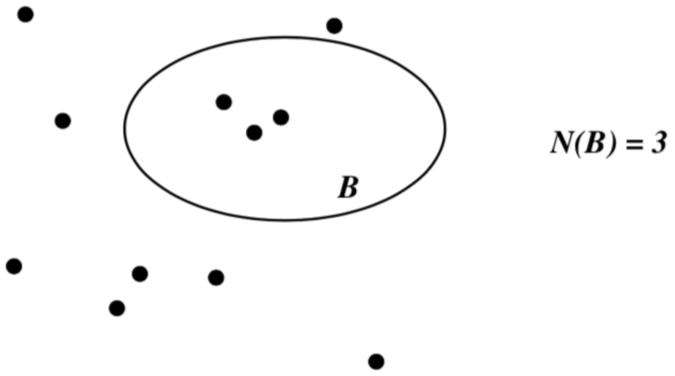


Figure 3.4: Example of a spatial point process in $\mathbb{R}^2$ with counting r.v. (Weil, 2007).

incoming calls may be split into domestic and international calls. An illustration of a marked spatial point process is shown in Figure 3.5.

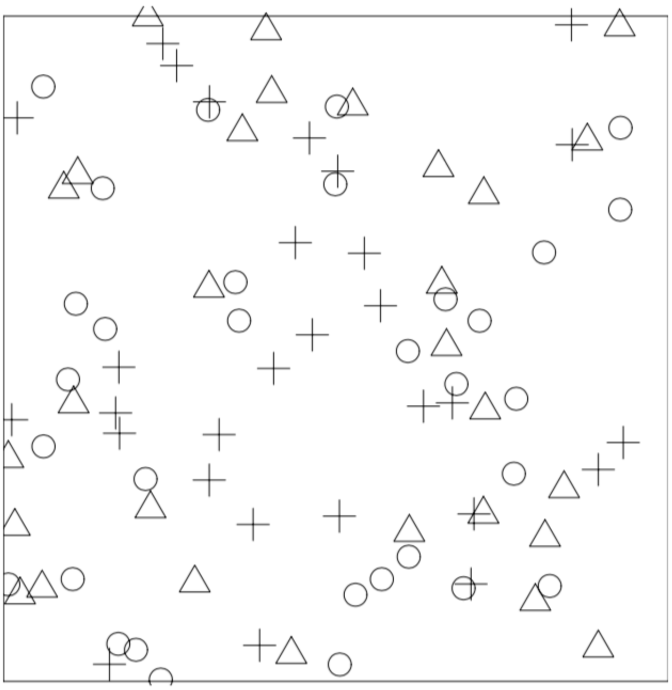


Figure 3.5: Example of a marked spatial point process in $\mathbb{R}^2$ with three marks (triangle,cross,circle). (Weil, 2007).

A special type of a point process is the *Poisson point process*. According to (Møller and Waagepetersen, 2007), a Poisson process $\boldsymbol{X}$ defined on S with intensity measure μ and intensity function ρ satisfies for any bounded region $B \subseteq S$ with $\mu(B) > 0$,

- $N(B)$ is Poisson distributed with mean $\mu(B)$,
- Conditional on $N(B)$, the points $\boldsymbol{X}_B$ are i.i.d. with density proportional to $\rho(u)$, $u \in B$.

In the case where $\rho(u)$ is constant $\forall u \in S$, the Poisson process is called *homogeneous*. Homogeneity in PPs implies that the dispersion of points in a space S (usually $\mathbb{R}^2$) is governed by the intensity function ρ. However, homogeneity is irrelevant to the shape realisations of homogeneous processes take, which may chaotic with large number of points concentrated in specific regions. Non-homogeneity implies that the number of points appearing in a region depends on the location of that region. It is therefore possible to model non-homogeneous processes as piecewise homogeneous, but this is outside the scope of this dissertation.

An entirely equivalent concept to a homogeneous point process is the stationary Poisson process. A point process is said to be stationary on $S \subseteq \mathbb{R}^2$ (Weil, 2007) if

$$\mathbb{P}(\{N(a, a+b] = k\})$$

depends on the length b but not on location b, $\forall\ b > 0$, $k \in \{0, 1, \ldots, \}$. We also define $\boldsymbol{X}$ to be stationary respective *isotropic* if its distribution has rotational and translational invariance in the space that $\boldsymbol{X}$ is defined.

Based on the above definitions, for the remainder of this chapter we will assume that the data generating process $\boldsymbol{Y}_{1:t}$ is a stationary and isotropic Poisson process with some fixed intensity ρ.

3.2 Poisson Gamma model

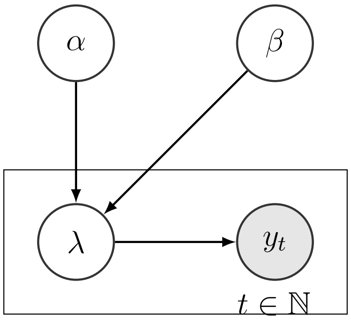


Figure 3.6: Plate diagram for the univariate Poisson Gamma model.

A conjugate model that effectively identifies change-points in count data is the Poisson Gamma (PG) model. We therefore assume that the DGP is a homogeneous Poisson process. The underlying rationale is that each point of the DGP is assumed to be drawn from a Poisson distribution with discrete intensity λ, where λ itself is drawn from a Gamma distribution. The plate diagram of this model is shown in Figure 3.6. The PG model is by nature an one-dimensional model of count data but can be extended to an arbitrary number of dimensions, say k, by defining a k-dimensional vector of independent univariate Poisson Gamma processes. However, a limitation emerges from this type of modelling; covariance between processes is ignored. As a result, the PG model identifies changes in the mean and variance of the only parameter $\boldsymbol{\theta}_{PG} = (\lambda)$.

3.2.1 Bayesian Analysis

Suppose that $\boldsymbol{y}_t = (y_{1t}, \ldots y_{kt})$ is a k-dimensional vector. Let $y_{1t}, \ldots y_{kt}$ be *iid* Poisson random variables with $y_{ij} \sim Poisson(\lambda_i)\ \forall\ 1 \leq i \leq k$, where λ_i is itself a random variable with $\lambda_i \sim Gamma(\alpha_{i0}, \beta_{i0})$. The probability mass functions of the Poisson and Gamma distributions are outlined in Appendix B. Also, note that α_{i0} and β_{i0} are fixed hyper-parameters that control the shape and rate of the Gamma distribution which itself control the intensity λ_i of the Poisson likelihood.

Consider the one-dimensional PG model (Hitchcock, 2014). Denote the distributions of $\lambda_i|y_{ij}$, $y_{ij}|\lambda_i$, and $\lambda_i|\alpha_0, \beta_0$ by π, g and p, respectively $\forall\ 1 \leq i \leq k$, $1 \leq j \leq t$. By Bayes' theorem $\forall\ 1 \leq i \leq k$, $1 \leq j \leq t$ it follows that

$$\begin{aligned}
\pi(\lambda_i|y_{ij}) &\propto g(y_{ij}|\lambda_i)p(\lambda_i|\alpha_{it}, \beta_{it}) \\
&= e^{-\lambda_i}\lambda_i^{y_{ij}}\lambda_i^{\alpha_{it}-1}e^{-\beta_{it}\lambda_i} \\
&= \lambda_i^{y_{ij}+\alpha_{it}-1}e^{-\lambda_i(\beta_{it}+1)}.
\end{aligned}$$

The parameter posterior distribution π at time t is also a Gamma distribution with shape and rate parameters equal to $\alpha + y_{ij}$ and $\beta + 1$, which implies that the PG model is conjugate. The parameter posterior derived above is used in equation 2.11 for computing the posterior predictive density of $\boldsymbol{y}_t$, as explained in the previous Chapter. This posterior update is performed independently for each one of the k data streams.

By performing the above update iteratively, we can obtain the following parameter updates at the end of the t-th iteration:

$$\begin{aligned} \alpha_{it} &= \alpha_{i0} + \sum_{j=1}^{t} y_{ij} \\ \beta_{it} &= \beta_{i0} + t, \end{aligned} \tag{3.1}$$

where α_{i0} and β_{i0} are the fixed hyper-parameters of the model for each $i = 1, \ldots, k$. Computationally, the parameter update is implemented by storing $\sum_{j=1}^{t} y_{ij}$ for every i and possible run-length. This quantity is known as a *sufficient statistic* because it is a function of the data y sufficient to fully describe the parameter update.

At each iteration this update can be performed in $\mathcal{O}(1)$, given that relevant quantities are stored throughout the execution ofthe algorithm. This is computationally very efficient and facilitates CPD.

Moreover, $\forall\ 1 \leq i \leq k$, $1 \leq j \leq t$ the posterior mean is equal to

$$\begin{aligned} \frac{\alpha_{ij} + \sum_{j=1}^{t} y_{ij}}{\beta_{ij} + t} &= \frac{\sum_{j=1}^{t} y_{ij}}{\beta_{ij} + t} + \frac{\alpha_{ij}}{\beta_{ij} + t} \\ &= \frac{t}{\beta_{ij} + t}\left(\frac{1}{t}\sum_{j=1}^{t} y_{ij}\right) + \frac{\beta_{ij}}{\beta_{ij} + t}\left(\frac{\alpha_{ij}}{\beta_{ij}}\right), \end{aligned}$$

where $\frac{\alpha_{ij}}{\beta_{ij}}$ is the prior mean, and $\frac{1}{t}\sum_{i=1}^{t} y_{ij}$ is the empirical mean. The expression above shows that the posterior mean is a weighted average of the prior and empirical means. As $t \to \infty$ or $\beta_{ij} \to 0$, the empirical mean has a larger effect on the update and the prior influence fades. Therefore, an unrepresentative prior mean would eventually not affect significantly the informed or updated estimate of the mean. Analogously, the posterior parameter variance is equal to

$$\frac{\alpha_{ij} + \sum_{i=1}^{t} y_{ij}}{(\beta_{ij} + t)^2}.$$

Parameter variance controls the uncertainty in the estimate of the parameter, which is modelled as an r.v.. A smaller variance conveys more certainty about the parameter estimate and vice versa.

Finally, the univariate ($k = 1$) posterior predictive (Hitchcock, 2014) $\mathbb{P}(y_{t+1}|y_{1:t})$ is equal to

$$\begin{aligned}
&\int_0^\infty \mathbb{P}(y_{t+1}|\lambda)\mathbb{P}(\lambda|y_{1:t})d\lambda && (3.2)\\
&= \int_0^\infty \frac{e^{-\lambda}\lambda^{y_{t+1}}}{(y_{t+1})!} \frac{(\beta+t)^{\alpha+\sum_{i=1} ty_i}\lambda^{\alpha+\sum_{i=1}^t y_i-1}e^{-\lambda(\beta+t)}}{\Gamma(\alpha+\sum_{i=1}^t y_i)}d\lambda && (3.3)\\
&= \frac{(\beta+t)^{\alpha+\sum_{i=1} ty_i}}{\Gamma(y_{t+1}+1)\Gamma(\alpha+\sum_{i=1}^t y_i)} \int_0^\infty e^{-\lambda(\beta+t+1)}\lambda^{\alpha-1+\sum_{i=1}^{t+1} y_i}d\lambda && (3.4)\\
&= \frac{(\beta+t)^{\alpha+\sum_{i=1} ty_i}}{\Gamma(y_{t+1}+1)\Gamma(\alpha+\sum_{i=1}^t y_i)} \frac{1}{Z}\int_0^\infty Ze^{-\lambda(\beta+t+1)}\lambda^{(\alpha+\sum_{i=1}^{t+1} y_i)-1}d\lambda, && (3.5)
\end{aligned}$$

where $Z \triangleq \frac{(\beta+t+1)^{\alpha+\sum_{i=1}^t y_i}}{\Gamma(\alpha+\sum_{i=1}^{t+1} y_i)}$. Therefore, $Ze^{-\lambda(\beta+t+1)}\lambda^{(\alpha+\sum_{i=1}^{t+1} y_i)-1}$ is a Gamma density and therefore the integral evaluates to 1 by axioms of probability. As a result, the posterior predictive becomes

$$\begin{aligned}
&\frac{(\beta+t)^{\alpha+\sum_{i=1}^t y_i}}{\Gamma(y_{t+1}+1)\Gamma(\alpha+\sum_{i=1}^t y_i)} \frac{\Gamma(\alpha+\sum_{i=1}^{t+1} y_i)}{(\beta+t+1)^{\alpha+\sum_{i=1}^t y_i}}\\
&= \left(\frac{\beta+t}{\beta+t+1}\right)^{\alpha+\sum_{i=1}^t y_i}\left(\frac{1}{\beta+t+1}\right)^{y_{t+1}} \frac{\Gamma(\alpha+\sum_{i=1}^{t+1} y_i)}{\Gamma(y_{t+1}+1)\Gamma(\alpha+\sum_{i=1}^t y_i)}, && (3.6)
\end{aligned}$$

which is a Negative Binomial distribution function with $p = \frac{\beta+t}{\beta+t+1}$ and $r = \alpha + \sum_{i=1}^t y_i$. The posterior predictive is used in equation 2.21 for prediction, as shown in the previous Chapter. Analogously to the posterior update, the posterior predictive can be extended to k dimensions by performing the update above independently for all k data streams.

3.2.2 Sensitivity Analysis

Based on the Bayesian analysis in the previous section, we can only infer the PG model's ability to model the data generating process and estimate CPs to a limited extent. For instance, the properties of the PG model reveal that it is impossible to identify changes in the covariance between the intensities generating two different DGPs. In order to formally assess the PG model's flexibility to identifying CPs in various settings, we conduct a sensitivity analysis on synthetic (artificially constructed) datasets.

Before considering instances of CPD in synthetic datasets, it is important to identify the factors affecting a model's detection precision. One factor arises from the nature of the CPD framework itself. On-line CPD incurs an important trade-off between detection latency and accuracy. In simple terms, a CPD algorithm is more likely to accurately estimate CPs long after they have occurred and vice versa. The reason accounting for that is the fact that the algorithm has more evidence about the intensity of the process after the CP and can therefore test the hypothesis that a CP occurred in the past more accurately. Another factor affecting detection accuracy is the length of the segment preceding the CP. A longer stationary and homogeneous segment allows the BOCDMS algorithm to establish a strong and stable belief about the intensity of a DGP. Any systematic change to the trend governing that segment will provide sufficient evidence for the presence of a CP and will thus increase the likelihood of declaring a CP. In addition, in volatile data streams the BOCDMS algorithm is more likely to declare a larger number of CPs, increasing the false positives rate. This is a result of the frequent fluctuations in the data, which do not allow the algorithm to identify locally stationary environments. Furthermore, the magnitude of the change in the model parameters directly affects CP estimation. Small in magnitude changes are less likely to be classified as CPs due to the fact that they are likely under the evidence density and can therefore be explained by the variance of the model. On the contrary, large and abrupt changes are more likely to be declared as CPs, as the on-line algorithm cannot assess whether they are outliers or part of a new trend. Finally, prior specification of the hyper-parameters has a significant impact on CPD. A strong unrepresentative prior may pollute the belief about the true value of the model parameters. Under the presence of new data the effect of the prior information on parameter mean and variance may gradually disappear. In the case of the PG model the rate at which priors seize to effect posterior beliefs about the mean is exponential, as shown in the derivation of the posterior parameter mean in the previous section.

Among all the factors mentioned, prior hyper-parameter specification appears to have the largest impact on CPD. Depending on the model, strong prior beliefs about model parameters that are close enough to their 'true' values will yield more accurate CPD results that strong beliefs that are far away from the truth. To avoid introducing variance in performance and accuracy when assessing models in the synthetic datasets below, we assign weak uninformative to model hyper-parameters. For the PG model, we set $\alpha = \beta = 1$ for the remainder of this

Chapter.

We now consider a showcase example of CPD using the PG on a simple synthetic dataset shown in Figure 3.7. Each dimension of the dataset is a piece-wise constant function. In this example, all CPs were identified by the PG model independently for each of the three dimensions. This is evident at time $t = 120$, where the run-length distribution is bimodal indicating the existence of two likely hypotheses. The one hypothesis is that there is no CP, which is the case for the green and red data streams, while the second hypothesis supports the existence of a CP, which occurs in the blue stream. Also, the lack of volatility allows the PG model to accurately estimate CPs without increasing the false positive or false negative rate.

However, under a different regime the PG model fails to identify CPs on-line, as shown in Figure 3.8. Denote the true segment length by $l_{segment}$ and the true number of CPs by n_{cps}. It is clear that the small segment length does not enable the algorithm to establish a baseline behaviour of the process and therefore fails to identify any CP. Equally important are the small in magnitude changes that occur. Under a Poisson likelihood with intensity λ, the mean and variance are both equal to λ. The mean is of the two streams vary from 11 to 15 units while the changes vary from 2 to 4 units. Therefore, under the assumption that the DGP is generated by a Poisson model these changes can be explained by the variance of the model, which also varies from 11 to 15 units. Despite that limitation, the algorithm has gathered evidence for the existence of a CP at times $t = 5$ to $t = 20$, $t = 35$ to $t = 60$ and $t = 65$ to $t = 100$ without declaring a CP at any of those times.

Based on the conclusions drawn from Figure 3.8, we extended the segment length $l_{segment}$ and increased the absolute magnitude of the changes appearing that dataset. The results are summarised in Table 3.1. By tuning the segment length and the magnitude of the changes, one can determine when the PG model 'breaks'.

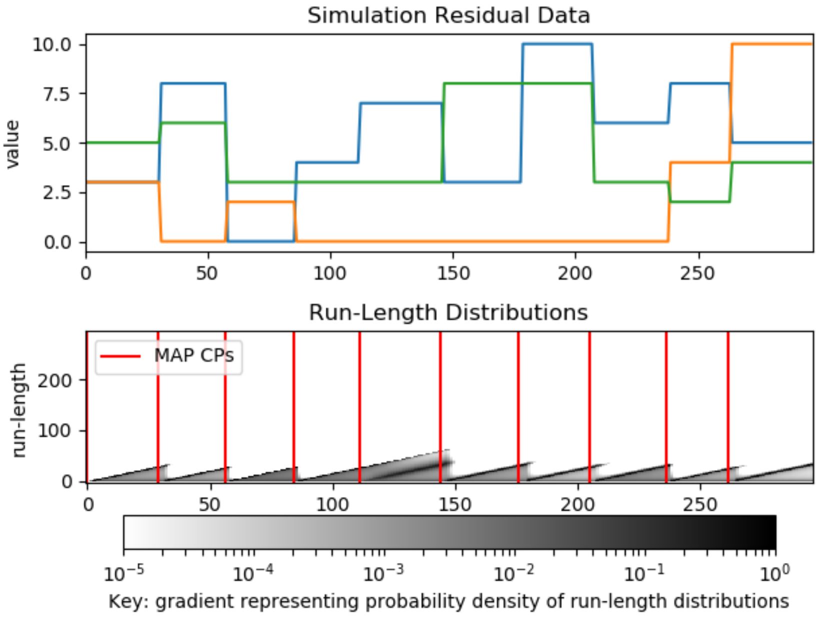


Figure 3.7: CPD using the PG model on a three-dimensional synthetic dataset. The dataset is shown on the top while the run-length distribution and CPs are shown in the bottom part of the Figure

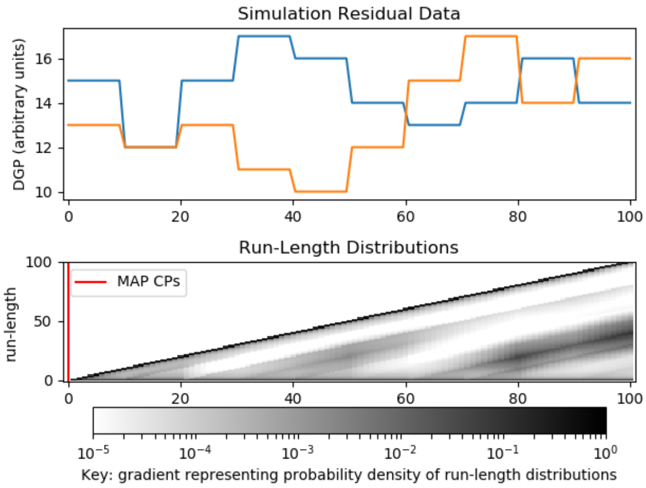
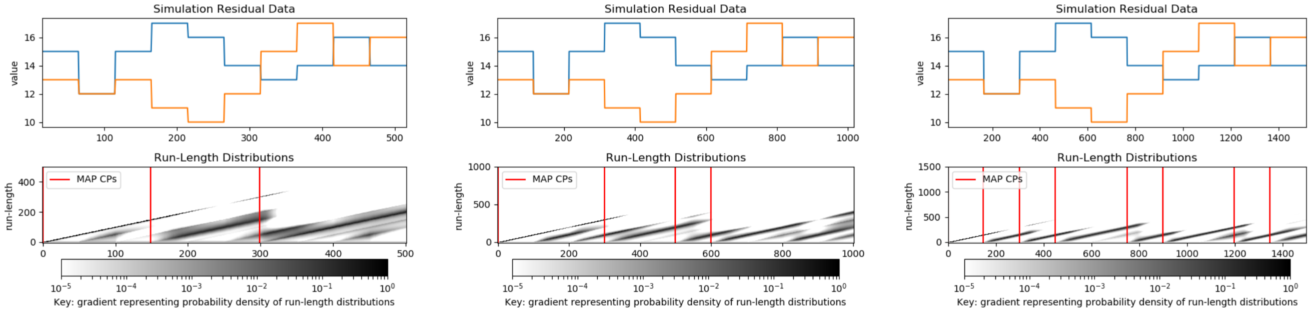


Figure 3.8: CPD using the PG model on a synthetic dataset with $l_{segment} = 10$, $n_{cps} = 10$.

Figure Name	$l_{segment}$	Change stretching factor	n_{cps}	% of CPs identified
3.9a	50	1	9	22%
3.9b	100	1	9	33%
3.9c	150	1	9	78%
3.10a	10	2	8	38%
3.10b	10	3	8	63%
3.10c	10	4	8	88%

Table 3.1: Table summarising segment extensions, change increases and the corresponding detection accuracy.

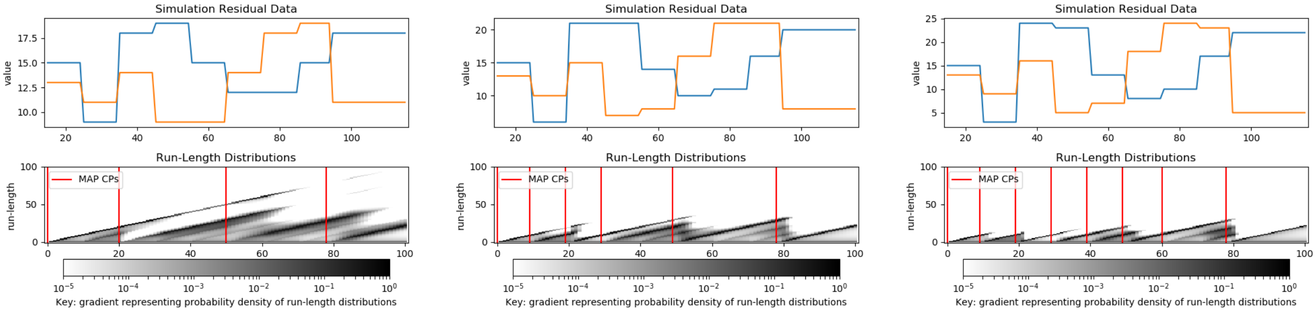
According to Table 3.1, increasing the segment length has a small effect on the detection accuracy. The evidence for the existence of a CP at the time referred to above is stronger in the extended datasets shown in Figures 3.9a to 3.9c but not sufficient for declaring a CP using the MAP estimator. On the contrary, increasing the magnitude of changes significantly boosts detection accuracy. This is attributed to the fact that the absolute changes are proportional to the variance of the process, as modelled by the Poisson distribution. We can therefore infer that the Poisson process is more likely to perform satisfactorily in volatile regimes where the absolute changes in the process are proportional to its variance. The run-length distribution in Figures 3.10a to 3.10c indicates that the abundance of possible hypotheses decreases as the magnitude of changes increases. This is particularly evident in Figure 3.10c, where the algorithm rejects almost any alternative hypothesis, which makes CP detection more clear-cut. In contrast, the run-length distribution in Figure 3.10a does not reject the hypothesis that there is no CP since $t = 0$ until $t = 70$, which renders the MAP segmentation more uncertain.



(a) CPD using the PG model on dataset 3.8 extended to $l_{segment} = 50$.

(b) CPD using the PG model on dataset 3.8 extended to $l_{segment} = 100$.

(c) CPD using the PG model on dataset 3.8 extended to $l_{segment} = 150$.



(a) CPD using the PG model on dataset 3.8 stretched by a factor of 2.

(b) CPD using the PG model on dataset 3.8 stretched by a factor of 3.

(c) CPD using the PG model on dataset 3.8 stretched by a factor of 4.

3.3 Multinomial Dirichlet model

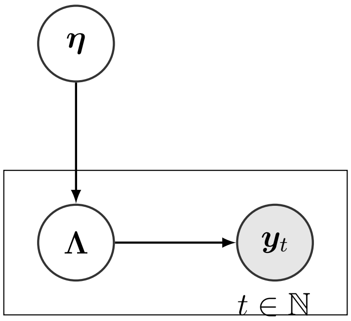


Figure 3.11: Plate diagram for the multivariate Multinomial Dirichlet model.

A multivariate extension to the PG model that also belongs in the conjugate family of distributions is the Multinomial Dirichlet (MD) model (Nelson, 1985a). The MD model addresses the limitation of the PG model and takes into account covariance when assessing the likelihood of a CP. The MD model differs from the PG model in the way it models intensity ρ. In particular, MD treats each one of the k data-streams as fractions of one cumulative data stream with intensity Λ. Therefore, each data stream has intensity $\Lambda_i := \frac{\lambda_i}{\Lambda}$, where λ_i and Λ_i are the absolute and relative or normalised intensities of data stream $i \leq k$. By definition, $\Lambda := \sum_{i=1}^{k} \lambda_i$ and as a result $\sum_{i=1}^{k} \Lambda_i = 1$. Despite this normalisation, the Multinomial component of the MD model identifies changes in the absolute intensities that generate the point process. The structure of the MD model indicates that it is particularly useful in categorical data applications. This model is depicted in a plate diagram shown in Figure 3.11.

3.3.1 Bayesian Analysis

Suppose that $\boldsymbol{y}_t = (y_{1t}, \ldots y_{kt})$, $\boldsymbol{\Lambda}_t = (\Lambda_{1t}, \ldots, \Lambda_{kt})$ and $\boldsymbol{\eta}_t = (\eta_{t1}, \ldots, \eta_{tk})$ are all k-dimensional vectors. Let

$$\boldsymbol{y}_t \sim Multinomial(\boldsymbol{\Lambda}_t, n),$$

where $n := \sum_{i=1}^{k} y_{it}$, $0 < \Lambda_{it} < 1 \; \forall \; i = 1, \ldots, k$, $\sum_{i=1}^{k} \Lambda_{it} = 1$ and

$$\boldsymbol{\Lambda}_t \sim \; Dirichlet(\boldsymbol{\eta}_t),$$

with $\eta_{it} > 0 \; \forall \; i = 1, \ldots, k,$. The probability density functions of the Multinomial and Dirichlet distributions are outlined in Appendix B.

Denote the distributions of $\boldsymbol{\Lambda}_t | \boldsymbol{y}_t$, $\boldsymbol{y}_t | \boldsymbol{\Lambda}_t, n$, and $\boldsymbol{\Lambda}_t | \boldsymbol{\eta}_t$ be π, g and p, respectively. Therefore, by Bayes' theorem the posterior distribution after one update (Tu, 2014) is equal to

$$\begin{aligned}\pi(\boldsymbol{\Lambda}_t|\boldsymbol{y}_t) &\propto p(\boldsymbol{\Lambda}_t|\boldsymbol{\eta}_t)g(\boldsymbol{y}_t|\boldsymbol{\Lambda}_t,n)\\ &\propto \prod_{i=1}^{k}\Lambda_{it}^{\eta_{it}-1}\prod_{i=1}^{k}\Lambda_{it}^{y_{it}}\\ &= \prod_{i=1}^{k}\Lambda_{it}^{\eta_{it}+y_{it}-1}\end{aligned}$$

Therefore, π is exactly the density of the Dirichlet distribution with updated parameters $\eta_{it}+y_{it}$ $\forall 1\leq i\leq k$ and, which implies that the MD model is conjugate. Extending this argument to argument to t updates yields the following update on $\boldsymbol{\eta}$:

$$\boldsymbol{\eta}_t = \boldsymbol{\eta}_0 + \sum_{\boldsymbol{y}_\tau\in\{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_t\}}\boldsymbol{y}_\tau \tag{3.7}$$

where $\boldsymbol{\eta}_0$ is the fixed prior hyper-parameter. The sufficient statistic of this update is $\sum_{\boldsymbol{y}_\tau\in\{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_t\}}\boldsymbol{y}_\tau$, which is stored for every dimension and possible run-length.

The posterior mean of $\boldsymbol{\Lambda}_t$ after one update is equal to

$$\frac{\eta_{it}+y_{it}}{\sum_{j=1}^{k}\eta_{jt}+y_{jt}},$$

$\forall i=1,\ldots,k$, while its posterior variance is equal to

$$\frac{(\eta_{it}+y_{it})(\eta_0-(\eta_{it}+y_{it}))}{\eta_0^2(\eta_0+1)},$$

where $\eta_0 := \sum_{i=1}^{k}\eta_{it}$ $\forall i=1,\ldots,k$. Finally, the posterior covariance between two normalised intensities Λ_{it} and Λ_{jt} is equal to

$$\frac{-\eta_{it}\eta_{jt}}{\eta_0^2(\eta_0+1)},$$

$\forall i\neq j\in\{1,\ldots,k\}$.

The posterior predictive distribution (Tu, 2014) at time t in equation 2.21 is equal to

$$\frac{\Gamma(n+1)}{\prod_{i=1}^{k}\Gamma(y_{it}+1)}\frac{\Gamma(\sum_{i=1}^{k}\eta'_{it})}{\prod_{i=1}^{k}\Gamma(\eta'_{it})}\frac{\prod_{i=1}^{k}\Gamma(y_{it}+\eta'_{it})}{\Gamma(n+\sum_{i=1}^{k}\eta'_{it})},$$

where η'_{it} is the updated posterior parameter of data stream i and Γ is the Gamma function. The derivation of the posterior predictive is omitted because it requires the use of tools that are outside the scope of this dissertation. The kernel of its posterior predictive is similar to the kernel of a Multinomial Dirichlet distribution, which is not directly available in a Python library. This implies that its computation was achieved by calculating the products of gamma functions, which incurred delays in the computation. For that reason we provide a simplified efficient calculation by taking the logarithm of the posterior predictive. Hence, the expression above is simplified to:

$$\begin{aligned}\log\left(\Gamma(n+1)\Gamma\Big(\sum_{i\,=1}^{k}\eta'_{it}\Big)\right)+\sum_{i\,=1}^{k}\log\left(\Gamma(y_{it}+\eta'_{it})\right)\\ -\sum_{i\,=1}^{k}\Big(\log\left(\Gamma(y_{it}+1)\Gamma(\eta'_{it})\right)\Big)-\log\left(\Gamma\Big(n+\sum_{i\,=1}^{k}\eta'_{it}\Big)\right)\end{aligned} \tag{3.8}$$

3.3.2 Sensitivity Analysis

The MD model addresses two limitations of the PG model: the problem of relative sizes of the changes in a DGP with respect to its variance and the inability to identify CPs in covariance between DGPs. For the remainder of this section we fix $\boldsymbol{\eta}$ to be equal to the k-dimensional unit vector, where k is the number of dimensions in the data.

A showcase example of an application of the MD model on a synthetic dataset is shown in Figure 3.12. The run-length distribution provides strong evidence for a correct MAP estimate, which is partially attributed to the sufficiently large segment length.

Regarding pathological cases, we applied the MD model on the dataset shown in Figure 3.8. The improvement in performance was marginally better than the performance of the PG model as only one CP was identified. However, the MD model provided a stronger evidence for the existence of other undeclared CPs compared to the evidence provided by the PG model.

For a direct comparison, the MD model was applied to the same synthetic datasets as the PG model. The effect of extending segment lengths and increasing the magnitudes of changes in the data is summarised in Table 3.2. According to that Table and Figures 3.14a to

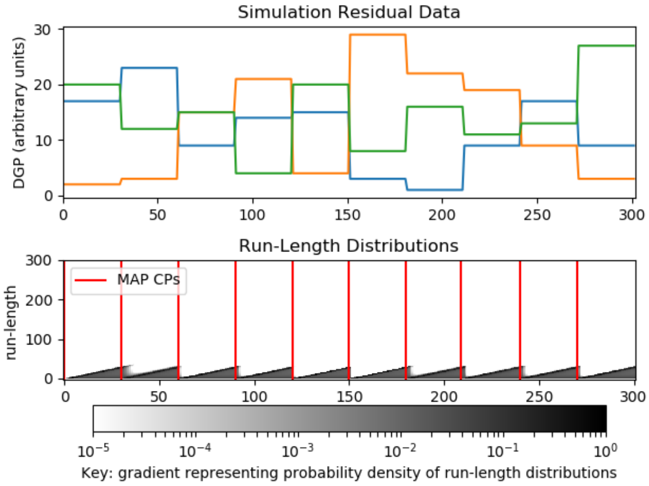


Figure 3.12: CPD using the MD model on a three-dimensional synthetic dataset with $l_{segment} = 50$. The dataset is shown on the top while the run-length distribution and CPs are shown in the bottom part of the Figure.

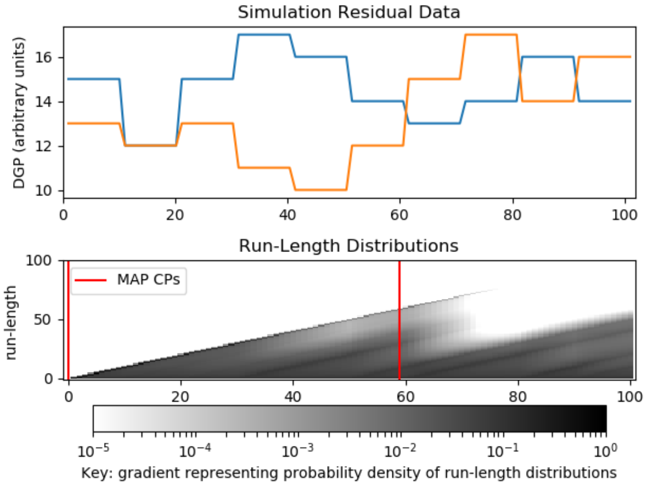
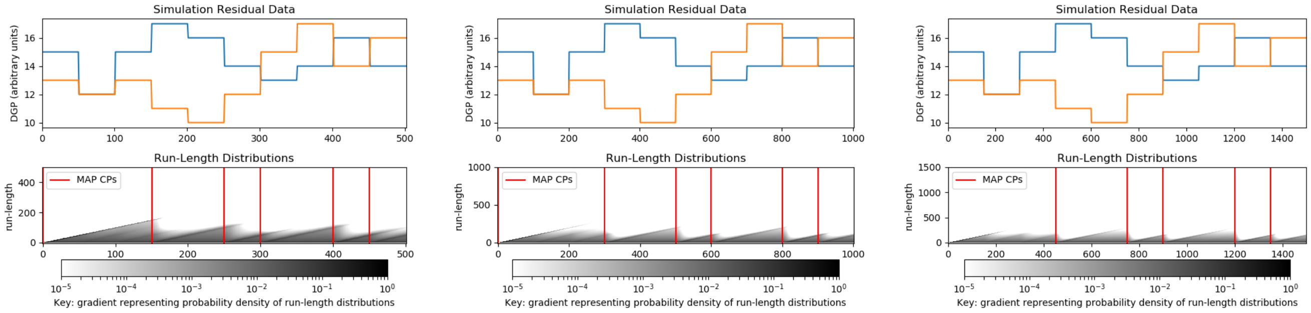


Figure 3.13: CPD using the MD model on a synthetic dataset with $l_{segment} = 50$, $n_{cps} = 9$.

Figure Name	$l_{segment}$	Change stretching factor	n_{cps}	MD model accuracy	PG model accuracy
3.14a	50	1	9	56%	22%
3.14b	100	1	9	56%	33%
3.14c	150	1	9	56%	78%
3.15a	10	2	8	38%	38%
3.15b	10	3	8	63%	63%
3.15c	10	4	8	88%	88%

Table 3.2: Table summarising increases in the magnitudes of the changes of DGPs and the corresponding detection accuracy.

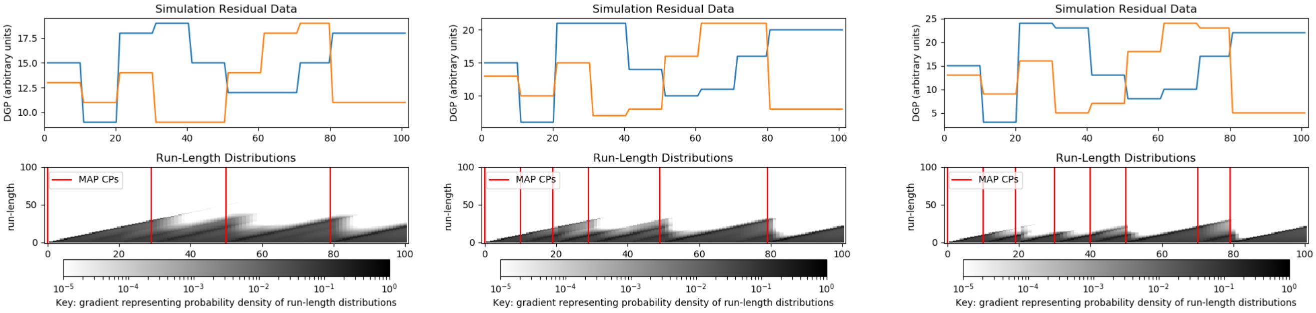
3.14c, the segment length has in general very little effect in the number of CPs identified by the MD model. Therefore, CPD under the MD model had segment-length invariance, which guaranteed consistent performance of the MD model in different settings. On the other hand, increasing the magnitudes of the changes, as shown in Figures 3.15a to 3.15c, increased the detection accuracy. In fact, the increase in change magnitude had the same effect on CPD under both the PG and MD models. As the stretching factor of changes increased, CP declarations became more distinct. The fact that both models' capacity to detect CPs is directly affected by the magnitudes of the changes themselves indicates that both models are more likely to be susceptible to declare outliers as CPs.



(a) CPD using the MD model on dataset 3.13 extended to $l_{segment} = 50$.

(b) CPD using the MD model on dataset 3.13 extended to $l_{segment} = 100$.

(c) CPD using the MD model on dataset 3.13 extended to $l_{segment} = 150$.



(a) CPD using the MD model on dataset 3.13 stretched by a factor of 2.

(b) CPD using the MD model on dataset 3.13 stretched by a factor of 3.

(c) CPD using the MD model on dataset 3.13 stretched by a factor of 4.

3.4 Model comparison

Overall, we can argue that the MD model is an extension of the PG model in the multivariate setting. in essence, the MD model encapsulates the capabilities of the PG model. In terms of application, the PG model is more appropriately used in count data while the MD model performs best in categorical data. However, the MD model can identify more subtle changes in covariance of the intensities governing the DGPs. Moreover, the PG model is more prone to not identifying CPs when the length of the segments is small whereas the MD model's ability to detect CPs appears to be independent of the segment length. Finally, both models are less effective in identifying CPs when the magnitudes of the changes are insignificant. However, given the current on-line CPD framework the models maintain a good performance in a variety of different data regimes. However, there is ample opportunity for improving the CPD in a discrete setting by enriching the model universe with other point processes models. Suggestions about such models are proposed in Chapter 5.

3.5 Model-specific algorithm

We end this chapter by summarising the use of the quantities derived in the Bayesian Analyses sections of the PG and MD models in Algorithm 2. We call this Algorithm Model-specific Bayesian On-line Change-point Detection (MSBOCD). MSBOCD illustrates the execution of BOCDMS for a given choice of model.

Regarding the time complexity of MSBOCD, it is linear in the number of observations, as is BOCDMS. Lines 6-9 take $\mathcal{O}(1)$ as they correspond to initialising appropriate data structures. The second step of the algorithm (lines 13-16) is computed in $\mathcal{O}(R_{max})$, where R_{max} is the number of retained run-lengths. Specifically, line 13 utilises the posterior predictive distributions 3.6 and 3.8 of the PG and MD models, respectively. Overall, the model-specific BOCPD has computational complexity equal to $\mathcal{O}(R_{max}t)$ after t observations.

As far as space complexity is concerned, the possible run-lengths and their corresponding probability, which together specify the run-length distribution, are stored in `numpy` arrays.

After run-length pruning, this translates to a space complexity of $\mathcal{O}(R_{max})$. In addition to that array, the PG and MD models keep track of their corresponding sufficient statistics 3.1 and 3.7, which for a k-dimensional model incurs a storage cost of $\mathcal{O}(R_{max}k)$. This translates to an overall space complexity of $\mathcal{O}(R_{max}k)$ for the MSBOCD algorithm.

Algorithm 2 Model-specific Bayesian On-line Change-point Detection (MSBOCD)

```
Input at time 0: Hazard $H$, prior $q$.
Input at time $t$: next observation $\boldsymbol{y}_t$.
Output at time $t$: Posterior predictive distribution of $\widehat{\boldsymbol{Y}}_{(t+1):(t+h_{max})}$, Run-length $r_t$ distribution.
for next observation $\boldsymbol{y}_t$ at time $t$ do
  if $t = 1$ then
    # Step 1: Initialisation.
    Compute model log-evidence according to 2.15.
    Initialise run-length distribution using 2.8.
    Initialise relevant sufficient statistics for all possible run-lengths.
  else
    # Step 2: Posterior computations.
    for $r_t = 0, 1 \ldots, t$ do
      Compute the log densities of $\boldsymbol{y}_t$ using the predictive posteriors shown in 2.11.
      Update sufficient statistics.
      Compute posterior predicted expectation from the current posteriors.
      Compute posterior predicted variance from the current posteriors.
    end for
    Output: Posterior predictive distribution of $\widehat{\boldsymbol{Y}}_{(t+1):(t+h_{max})}$, Run-length $r_t$ distribution.
  end if
end for
```

Chapter 4

Real-data applications

Objectives:

✓ Applying the MSBOCD algorithm to Chicago crime data.

✓ Applying the MSBOCD algorithm to UK Property transactions data.

✓ Applying the MSBOCD algorithm to cryptocurrency transaction frequency data.

✓ Fine-tuning model hyper-parameters and interpreting CPs where possible.

✓ Assessing individual model performance and drawing comparisons between models.

So far, we have examined the performances of the Poisson Gamma and Multinomial Dirichlet models on synthetic simulations of data. In order to obtain a holistic view of the flexibility and robustness of each model, we apply them on three real-world datasets. The first case study we consider is the Chicago crime data[1] obtained from the Chicago Data Portal. The second case study includes a dataset about the property transactions in the UK[2], as registered by the Office for National Statistics. Finally, we apply the CPD algorithm in cryptocurrency transaction volume data[3] obtained from Kaggle. The BOCDMS algorithm is executed using both the Poisson Gamma and Multinomial Dirichlet model separately on each dataset. Since these

[1] Data accessible here

[2] Data accessible here

[3] Data accessible here

datasets have not been used in previous approaches in literature, that is they are unsupervised, we attempt to establish a benchmark for assessing their performance by mapping their declared CPs to real-world events.

4.1 Chicago crime

A common application of point processes is found in crime data (Albertetti et al., 2016). The location and abundance of crimes renders crime data ideal for spatio-temporal modelling. This is because the number of crimes across a fixed geographical region S can be modelled as a marked point process, where each mark corresponds to a different location or sub-region of S. Also, the fact that crime data is seasonal enables us to assess model performance under the presence of seasonality without any deseasonalisation preprocessing. We focus of identifying CPs in the number of crimes committed on a daily basis in Chicago, Illinois from 2001 to 2017. The locations of crimes during that period are depicted in Figure 4.1. This Figure indicates that it is possible to model the DGP as a marked PP with each label representing a different region of the Chicago area. However, in practise it is difficult to establish boundaries and define local regions within the observation window shown in Figure 4.1. A naive spatial segmentation where the Chicago area is divided into equal in area regions would result in areas where very few and possibly no crimes are recorded. This phenomenon is commonly referred to in PP theory as *edge effects*. More advanced spatial segmentation techniques are outside the scope of this dissertation and are therefore not considered. For that reason, we model all crimes as an one-dimensional PP. The consequence of aggregating over all Chicago regions is that the flexibility of the model becomes limited. CPs that occur in specific areas of Chicago may not be detected if aggregation occurs due to a phenomenon known in Statistics as *Simpson's paradox*.

4.1.1 Hyper-parameter tuning

CP declaration is heavily dependent on hyper-parameter prior specification, as argued in Chapter 3. Different priors will yield different CPD results. Therefore, it is vital to fine-tune each model's prior hyper-parameters. For the PG model, these are α and β. These are used in the Gamma distribution to control the intensity of the Poisson likelihood. An uninformative

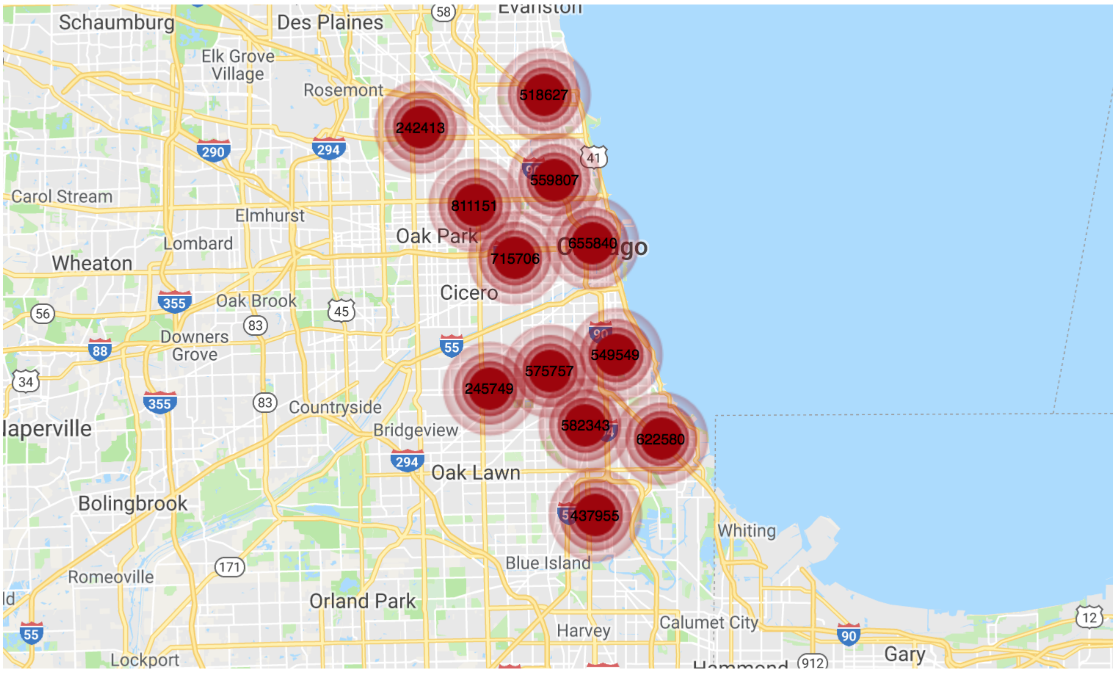


Figure 4.1: Location and number of crimes committed in Chicago during 2001 to 2017.

prior sets both α and β to equal to one and the prior Hazard equal to 30. As there is no expertise knowledge about the abundance of crimes we choose an uninformative prior. For comparison, we set $\alpha = 8000$ and $\beta = 20$, which translate to a prior mean and variance equal to 400 and 20, respectively.

4.1.2 CPD results

Since the data is univariate, the MD and PG models produce almost identical CPD results as the PG model is the univariate version of the MD model. Hence, we only illustrate the results of running the PG model in Figure 4.2. According to that Figure, the data has a strong seasonal factor that the CPD algorithm effectively identifies. Hence, a trend-blind model manages to detect seasonality in a an on-line setting. However, this has consequences on the detection of the trend of the data. It is evident that there is a decreasing linear trend which is concealed by the seasonality and is thus not detected.

By manually fine-tuning the prior specification, the results of the CPD differ significantly, as depicted in Figure 4.3. The effect of the informed prior is evident on the run-length distribution, which is piecewise linear. Therefore, there are no alternative hypotheses for the MAP

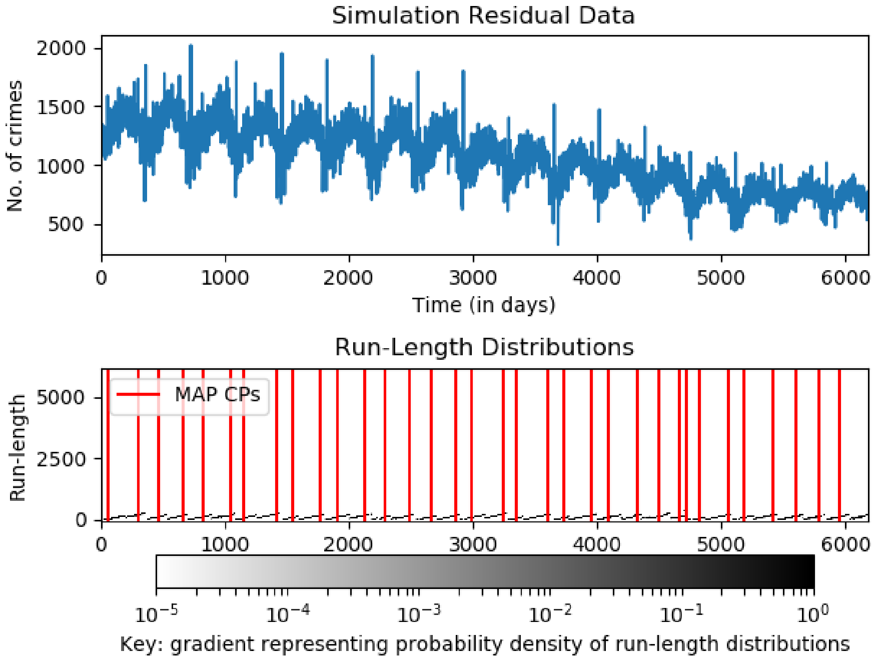


Figure 4.2: CPD using the PG model on the Chicago dataset using $\alpha = \beta = 1$ as prior specification.

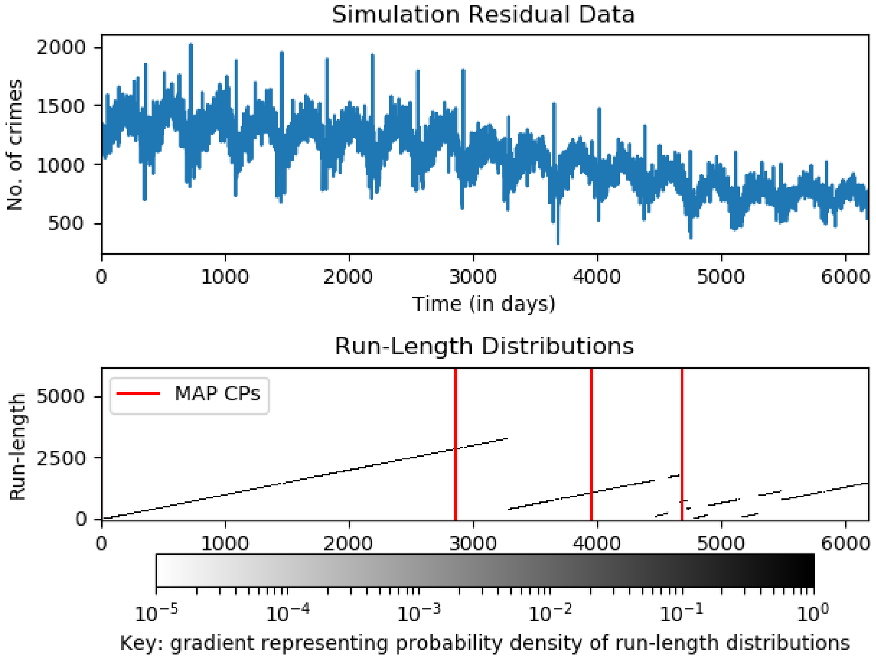


Figure 4.3: CPD using the PG model on the Chicago dataset using $\alpha = 8000$, $\beta = 20$ as prior specification.

estimator to assess. Moreover, the cause of each CP identified is cited in Table 4.1. It is evident that the CPs identified can be attributed to real events and therefore were correctly identified. For instance, at $t = 3956$ (April 2004) the Chicago Police Department reported that it had the lowest crime rate since 1965. As a result, we can conclude that the performance of the two models is significantly boosted by fine-tuning the hyper-parameters.

Time t	Date	Event
2861	January 2003	Chicago Police Department installs POD's (Police Observation Devices) in high crime areas.
3956	April 2004	Chicago Police Department adopts IT-based crime-fighting techniques recommended by the Los Angeles and New York City Police Departments.
4687	March 2005	14 Chicago Mafia members are indicted near the end of April 2005.

Table 4.1: Mapping CPs to events from 2001 to 2017 related to crime in the vicinity of Chicago, Illinois.

4.2 UK property transactions

Another area where CPD can provide useful insights is real estate. The ability to identify the time when the housing market undergoes changes enables a real estate agent to invest accordingly. For instance, being able to detect a housing bubble while its happening can help agents make lucrative investments. That is why we consider the monthly number of property transactions in the United Kingdom from April 2005 to February 2018 with value £40000 and above. We model the PP as marked, where the labels are England, Scotland, Wales and Northern Ireland.

4.2.1 Hyper-parameter tuning

After fine-tuning the hyper-parameters of the two models, we found that an uninformative prior would cause the MD model to over-fit the data while an uninformative prior would result in the PG model under-fitting the data. For that reason, we omitted these cases from consideration. For the MD model, we set $\boldsymbol{\eta} = (1150, 1150, 1150, 1150)^T$ and the prior Hazard to be equal to 30.This results in a prior mean intensity of the process equal to $(0.25, 0.25, 0.25, 0.25)^T$ and a prior variance in the intensity equal to $(0.0.0000408, 0.0.0000408, 0.0.0000408, 0.0.0000408)^T$, according to the Dirichlet prior.

4.2.2 CPD results

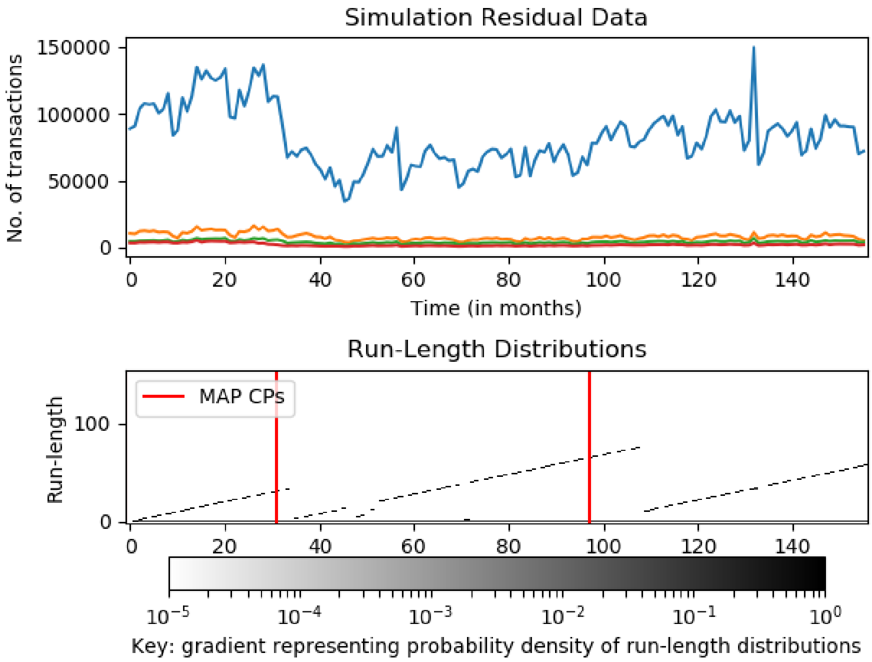


Figure 4.4: CPD using the PG model on the property transactions dataset using $\boldsymbol{\alpha} = \boldsymbol{\beta} = (1, 1, 1, 1)^T$ as prior specification.

The results of applying the two models in the four-dimensional property transactions dataset are shown in Figures 4.4 and 4.5. In these Figures, England, Scotland, Wales and Northern Ireland are drawn in blue, orange, green and red respectively. We replicate results of CPD using both models in Figure 4.6 in order to avoid issues of relative scale. Also, we interpret some of the CPs identified by the two models in Table 4.2.

As far as the PG model is concerned, it detects two clear shifts in the mean of the data at times $t = 31$ and $t = 97$ and does not misinterpret the outlier at $t = 131$ as a CP. These shifts can be identified by closely examining both plots in Figure 4.6. The fact that property transactions among different areas of the UK are strongly correlated implies that CPs are often common for all dimensions of the data and is therefore difficult to attribute the cause of a CP to the changing pattern of a specific dimension of the data. According to Table 4.2, the CP at $t = 31$ coincides with the beginning of the global financial crisis. Therefore, we can argue the declared CPs by the PG model are to a certain extent indicative of the structural changes the

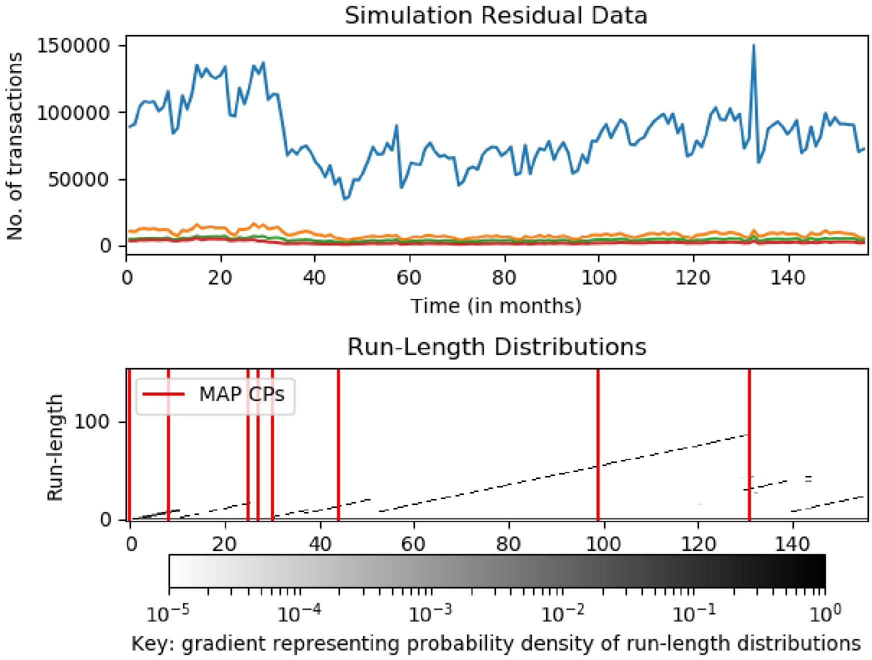


Figure 4.5: CPD using the MD model on the property transactions dataset using $\boldsymbol{\eta} = (1150, 1150, 1150, 1150)^T$ as prior specification.

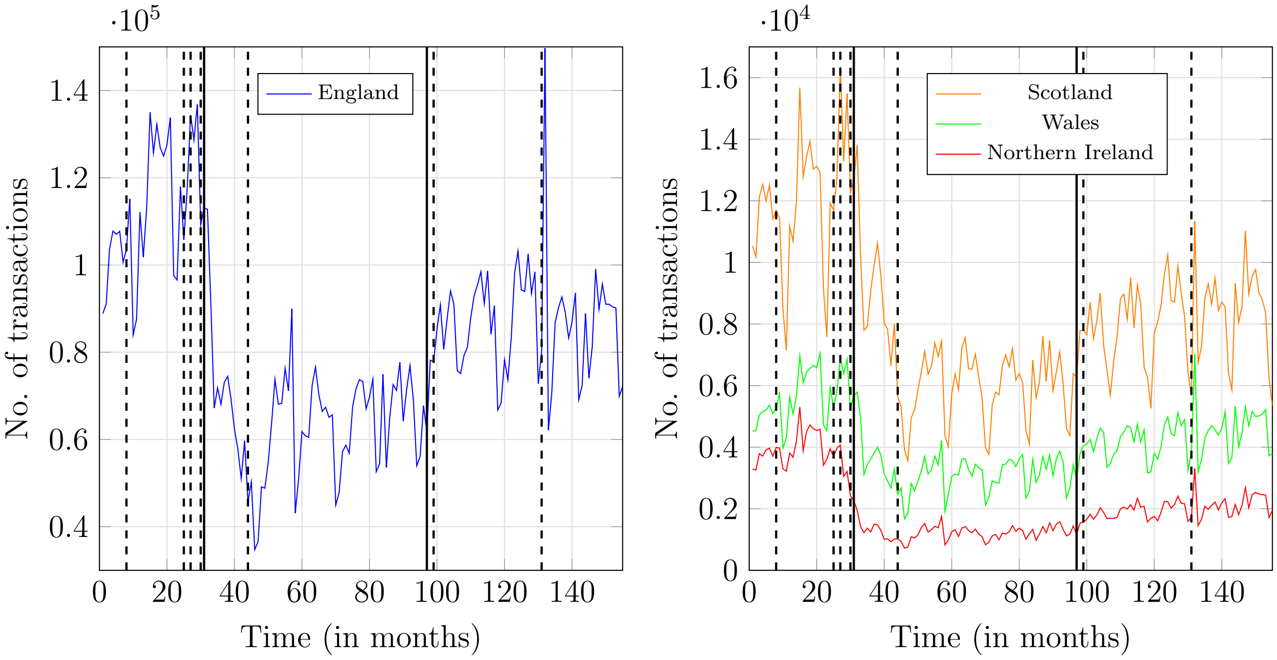


Figure 4.6: Monthly property transactions in England (blue), Scotland (orange), Wales (green) and Northern Ireland (red) from April 2005 to February 2018. CPs detected by the PG and MD models are shown in black dashed and solid lines, respectively.

UK property market experienced.

Regarding the MD model, it identified seven CPs two of which were also declared under the PG model. After the first detected CP at $t = 27$, the variance of transactions increases abruptly in all UK regions. In addition, during the three CPs identified between April ($t = 25$) and September ($t = 30$) 2017 the pairwise covariance between property transactions in Scotland, Wales and Northern Ireland appears to be changing. The detection of change in covariance illustrates the high flexibility of the MD model compared to the naive multivariate extension of the PG model. Moreover, the CP at November 2008 ($t = 44$) appears to pinpoint a temporary drop in the mean transactions across all regions. According to Table 4.2, the house prices witnessed the largest percentage drop in that November. Finally, the last CP detected at $t = 131$ corresponds an outlier in the data most likely caused by the announcement of the Brexit referendum, as alleged in Table 4.2.

Time t	Date	Event
31	October 2007	Global financial crisis begins with a crisis in the subprime mortgage market in the United States.
44	November 2008	Second largest percentage decrease in UK house prices in the post-financial crisis era according to Nationwide.
97	April 2013	Mortgage costs start to fall due to the Government and Bank of England's Funding for Lending scheme.
131	February 2016	Former prime minister David Cameron announces Brexit referendum on the following summer.

Table 4.2: Mapping CPs identified by the PG and MD models to events from April 2005 to February 2018 related to real estate in the United Kingdom.

4.3 Cryptocurrency transactions

The last application we consider is based on transaction counts of three prominent cryptocurrencies. CPs in cryptocurrency transactions are important because they can reveal shifts in demand of these currencies and can assist market analysts in predictive tasks. We model the transaction counts of Bitcoin (BTC), Ethereum (ETH) and Litecoint (LTC) jointly using a three-dimensional model. Also, we narrow the time frame to daily transactions between 20

February 2017 and 20 February 2018.

4.3.1 Hyper-parameter tuning

In regards to hyper-parameters, an informed prior specification on the both models results in data over-fitting. For the PG model, we set $\boldsymbol{\alpha} = (150, 150, 150)^T$ and $\boldsymbol{\beta} = (6, 6, 6)^T$, which suggests a prior intensity mean and variance equal to 25 and 4.17, respectively, for every dimension of the data. The MD model receives a prior $\boldsymbol{\eta} = (4.000.000.000, 4.000.000.000, 4.000.000.000)^T$ based on the expected number of daily transactions of cryptocurrencies. The prior Hazard is set by default to 30 for both models.

4.3.2 CPD results

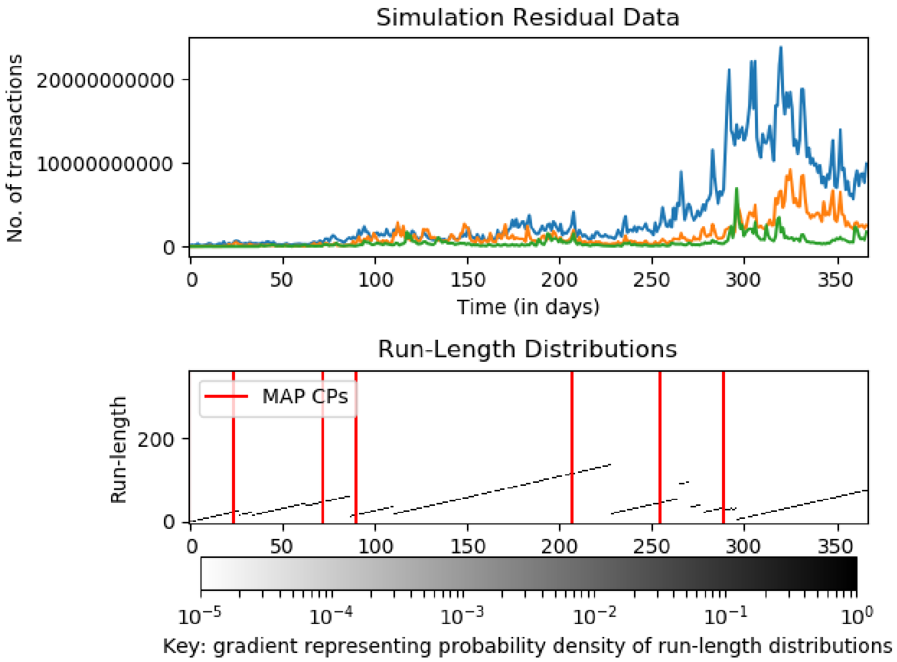


Figure 4.7: CPD using the PG model on the cryptocurrency transaction volumes dataset using $\boldsymbol{\alpha} = (150, 150, 150)^T$ and $\boldsymbol{\beta} = (6, 6, 6)^T$ as prior specification.

The results of applying the BOCDMS algorithm on the cryptocurrency data are shown in Figures 4.7 and 4.8 for the PG and MD models, respectively. Figures 4.9 and 4.10 replicate the CPD results by depicting the LTC data on a separate plot for clarity. Table 4.3 attempts to identify the possible cause of a some the identified CPs.

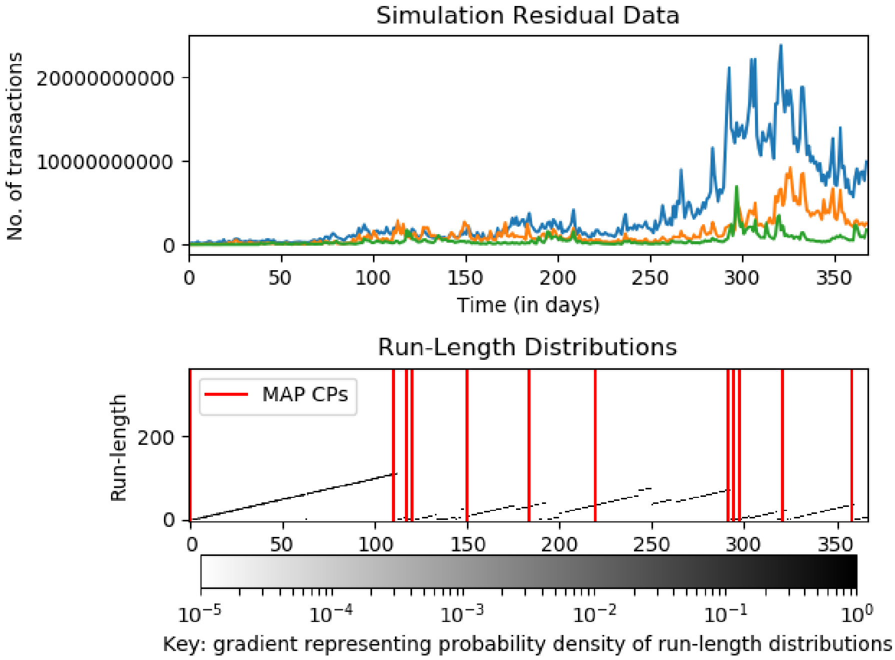


Figure 4.8: CPD using the MD model on the cryptocurrency transaction volumes dataset using $\boldsymbol{\eta} = (4.000.000.000, 4.000.000.000, 4.000.000.000)^T$ as prior specification.

The PG model identifies CPs in the variance of ETH & LTC at $t = 72, 90, 207$ and BTC, ETH & LTC at $t = 289$. The CP in the latter case can also be attributed to a shift in the mean of the data. However, due to the high volatility in the data we can argue that these CPs are also caused by a shift in the variance. Shifts in the mean are observed at $t = 23$ and $t = 254$ but it is equally possible that these CPs are caused by a change in variance. Furthermore, an important feature of the PG model is that it effectively discovers the major structural changes in the data by analysing it macroscopically.

On the other hand, the MD model is more sensitive to the fluctuations in the data and therefore detects a larger number of CPs. Contrary to the PG model, it examines the data on a microscopic level and identifies subtle changes in it. The sensitivity of the model can be adjusted by optimising the prior hyper-parameter specification. Moreover, the CPs detected at $t = 110, 117, 120, 150, 219$ are most likely caused by a change in mean or variance while the CP $t = 183$ seems to be cause by a change in covariance between ETC and LTC. The three CPs occurring between $t = 291$ and $t = 297$ identify changes in the mean variance and covariance of all three cryptocurrencies. Finally, the CP at $t = 321$ pinpoints a shift in the mean of ETH and BTC while the CP at $t = 358$ can associated with a change in variance of the transactions of all three cryptocurrencies.

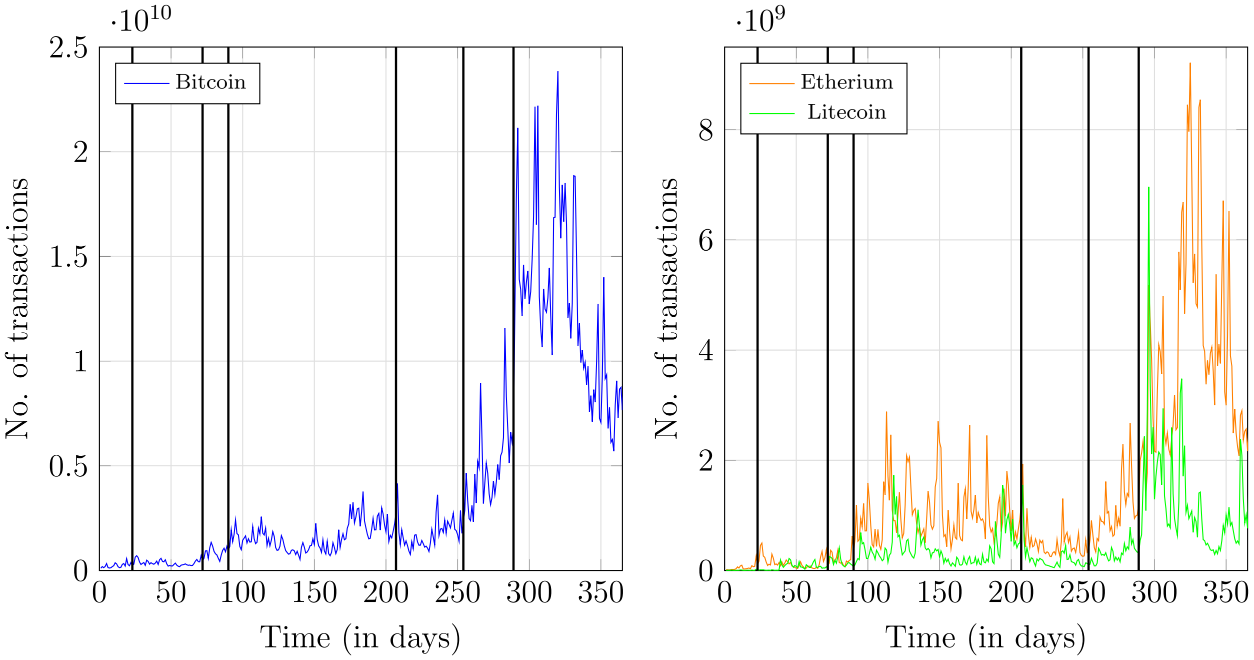


Figure 4.9: Daily cryptocurrency transactions in Bitcoin (blue), Etherium (orange) and Litecoin (green) from 20 February 2017 to 20 February 2018. CPs detected by the PG model are shown in black solid lines.

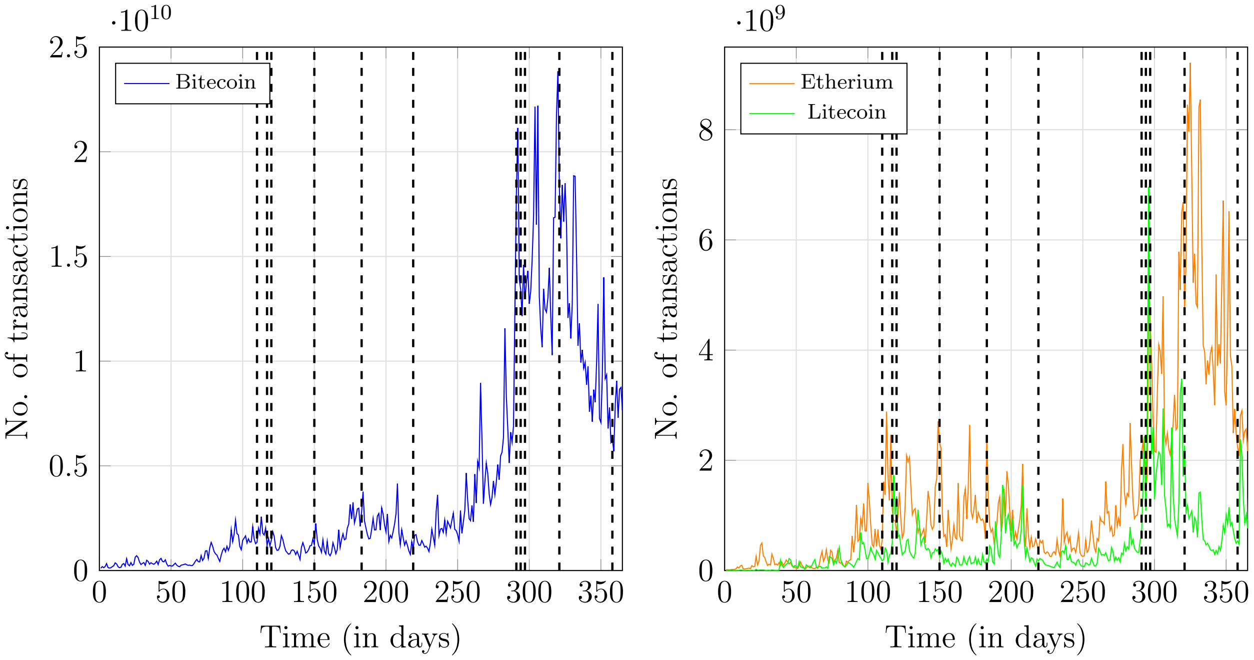


Figure 4.10: Daily cryptocurrency transactions in Bitcoin (blue), Etherium (orange) and Litecoin (green) from 20 February 2017 to 20 February 2018. CPs detected by the MD model are shown in black solid lines.

Time t	Date	Event
23	14 March 2017	The U.S. Securities and Exchange Commission on Friday denied a request to list Winklevoss twins' Bitcoin Ethereum COIN.
90	20 May 2017	Bitcoin and Ethereum are two of Google's most popular searches this week.
207	14 September 2017	Beijing orders cryptocurrency exchanges to stop trading and blocks new registrations.
254	31 October 2017	CME Group, the world's largest exchange operator by market value, is readying plans to offer futures on bitcoin.
289	05 December 2017	Bitcoin jumps above $12,000 to record high.

Table 4.3: Mapping CPs to events from 20 February 2017 to 20 February 2018 related to Bitcoin, Etherium, and Litecoin transactions.

Chapter 5

Conclusions

Objectives:

✓ Summarising thesis work and personal contributions.

✓ Discussing achievements and limitations with reference to the project's objectives.

✓ Proposing directions for future work.

✓ Addressing legal, social, ethical and professional issues.

In this thesis we have introduced CPD problems in the context of time series and provided a formal mathematical formulation of a general CPD problem. Chapter 1 demonstrates the notion of change-point with the aid of graphs illustrating changes in the mean, variance and covariance of time series. The need for CPD algorithms is motivated by the usefulness of their application in medicine and climate change. Moreover, the motivation of CPD is complemented by a comprehensive literature review of CPD techniques in off-line, frequentist and temporal settings. By addressing the limitations of existing methods, we propose a Bayesian on-line CPD framework in a spatio-temporal setting based on the works of (Knoblauch and Damoulas, 2017; Adams and MacKay, 2007; Fearnhead and Liu, 2007).

We proceed by defining important tools required for the construction of the BOCDMS algorithm, such as stationarity and Bayes' theorem. Then, we outline the building blocks of the CPD system by summarising the assumptions required for its development. Under the PPM model, we define the run-length and model random variables that facilitate the computation of the MAP segmentation estimators and the h-step ahead predictions. After specifying the prior distributions of these r.v.'s we cite key recursive equations that are computed in the system using a dynamic programming paradigm. These equations are compiled into the BOCDMS algorithm and its extensions: model selection and run-length pruning. Chapter 2 ends by analysing the computational and storage complexities of the BOCDMS algorithm, which are $\mathcal{O}(|\mathcal{M}|R_{max}T)$ and $\mathcal{O}(|\mathcal{M}|R_{max})$ for T data observations.

Furthermore, Chapter 3 extends the algorithm to point processes by introducing fundamental concepts in point process modelling, such as homogeneity and isotropy. Then, we enrich the model universe $\mathcal{M}$ by incorporating two parametric Bayesian conjugate models: the Poisson Gamma and Multinomial Dirichlet models. For each model, the parameter posterior and posterior predictive distributions are derived. After carrying out a Bayesian analysis, we examine the sensitivity of each model in synthetic datasets and address its limitations by applying it on datasets in which it underperforms. What follows is a comparative analysis of the two models in reference to their attributes and capabilities.

Finally, we apply the two models separately on three real-world datasets: crime in Chicago, property transactions in the UK and cryptocurrency transactions. For each dataset, the model's priors are fine-tuned and its performance is evaluated. There is also an attempt to map the detects CPs to real-world events in the news in order to establish a benchmark for the performance of the models.

We end this dissertation by discussing project achievements and limitations, suggesting future points of action and addressing legal, social, ethical and professional issues.

5.1 Project evaluation

For the discussion of the project's achievements and limitations we provide account of the project's objectives in Appendix A, as outlined in the Progress Report.

According to these requirements, Chapters 1 and 2 meet **OBJ1** as outlined in the thesis summary above. As far as **OBJ2** is concerned, *OBJ2.1* and *OBJ2.2* are met in Chapter 3.

However, we modified objective *OBJ2.3* and instead of implementing Gaussian Cox processes, we implemented the Multinomial Dirichlet model. This decision was based on the fact that the MD model is the multivariate extension of the PG model and therefore the MD model's implementation was a more natural progression after the achievement of *OBJ2.1* and *OBJ2.2*. Finally, **OBJ3** was also completed and extended. In an effort to introduce variety in the real-world applications and provide a holistic evaluations of both models, we fed a third type of data (UK property transactions) to the algorithm. Consequently, *OBJ3.1* was modified to include three sources data instead of two. Moreover, hyper-parameter tuning (*OBJ3.2*) and performance metrics (*OBJ3.3*) can be found in Chapters 2 and 3. Regarding visual representations of CPD results (*OBJ3.4*), they are illustrated in Figures included in Chapters 2, 3 and 4. Finally, we extended **OBJ3** and achieved an additional objective: "Establish performance benchmarks for unsupervised datasets to assess model-specific CPD results".

Overall, this project manages to go beyond the initial requirements set in the Progress report. The personal contributions made in Chapters 3 and 4 assist the work of Theodoros Damoulas and Jeremias Knoblauch and implements extensions proposed by the latter in Knoblauch and Damoulas (2017).

Furthermore, the project faced a number of obstacles during its completion, all of which were overcome. The implementation of the MD model was delayed due to the inefficient calculation of the posterior predictive and the existence of a bug in the code. However, after consulting the supervisor, Theodoros Damoulas, these issues were quickly resolved. In addition, the fact that CPD is an active research area meant that the requirements of the project had to be changed relatively frequently. In response to this agile project environment, we had frequent meetings with Jeremias Knoblauch and Theodoros Damoulas to discuss future directions of the project. Finally, near the end of the project key components of the code were updated by Jeremias Knoblauch to resolve bugs in model selection. However, the code written for the PG and MD models could not be updated in order to achieve compatibility with other key components of the codebase. As a result, model selection using a model universe with both the PG and MD models was not implemented. However, the fact that the MD model is the multivariate version of the PG model eliminated the need for model selection between univariate and multivariate models.

In addition, the limitations faced in this project can be divided into algorithm and model specific. Regarding the algorithm, hyper-parameter tuning was manually done and therefore it is not guaranteed that the hyper-parameters chosen are optimized. Also, the lack of a loss function did not provide us with a measure of model fit and error.

The second type of limitations originated from the models chosen to equip the model universe. While the choice of spatio-temporal point process models yielded high performances in the CPD task, they were based on assumptions about the structure and properties of the data they modelled. First of all, the model universe was restricted to conjugate models because of their high efficiency in computing MAP estimates in an on-line setting. However, non-conjugate models may have been able to provide additional capabilities for detecting CPs. Moreover, the assumptions about space homogeneity and isotropy of the PP may not be applicable to a number of real-world datasets. Due to the stochastic nature of most real-world datasets, it is often unrealistic to assume that the data meets some stability conditions. As a result, the models used to mimic the DGP may often be simplifications of reality and therefore may not capture the signal in the data.

5.2 Legal, Social, Ethical and Professional Issues

This dissertation did not face any legal, social, ethical or professional issues. Regarding the use of data, the datasets in Chapter 3 were obtained from open licence-free sources and therefore there were no legal requirements for their use. Also, there were no professional issues raised during the collaboration with Jeremias Knoblauch and Theodoros Damoulas. Overall, the project was completed smoothly and met all the required legal, social, ethical and professional obligations.

5.3 Future work

This thesis provides ample opportunities for future work based on the limitations addressed in the previous section.

First, we propose the development of a new component in the BOCDMS algorithm that optimizes the value of the hyper-parameters based on a fitness function such as the one suggested by (Khan et al., 2016). Alternative approaches may involve placing another prior on the hyper-parameters and sampling its posterior using sequential Monte Carlo techniques as proposed by (Svensson et al., 2015). This also motivates the computation of a loss function or its Bayesian analogue: Bayes factors. As a result, we can establish a mathematical benchmark for model performance.

Next, we propose the use of non-conjugate models in the models universe. Despite the fact that these models tend to be more computationally costly, there have been efficient on-line approaches (Turcotte and Heard, 2015) that do not directly compute the parameter posterior but sample from it. Finally, we suggest the implementation of inhomogeneous point process models that minimize the number of assumptions made about the data's properties. These models include Cox processes (Cressie, 1993), which are inhomogeneous generalisations of Poisson processes where the intensity $\rho(t)$ is itself a time-dependent stochastic process.

Appendix A

Project objectives

Objective	Priority	Description
OBJ1	High	Conduct a comprehensive literature review of the current methods used in various settings, i.e. CP detection in offline (batch) multivariate setting.
OBJ1.1	High	Understand how the PPM model works in a Bayesian setting.
OBJ1.2	High	Understand the use of MAP technique to estimate and predict CPs.
OBJ1.3	High	Study the use of sparse GMs to model relationships between variables.
OBJ1.4	High	Study the main algorithm combining these techniques to detect CPs.
OBJ2	High	Extend the software's functionality to CP detection of point processes.
OBJ2.1	High	Study the Poisson Gamma model and the effect of conjugacy on its posterior distributions.
OBJ2.2	High	Implement the PG model in Python while adhering to the principles of dynamic programming.
OBJ2.3	Medium-Low	Incorporate more sophisticated models for Gaussian Cox processes.
OBJ3	High	Feed one or two types of data to the system to evaluate its performance in action.
OBJ3.1	High	Include crime and/or financial data as two different sources of data.

OBJ3.2	Medium	Experiment with online algorithm by running simulations with different initial parameters and assess their impact on CP detection and prediction.
OBJ3.3	Medium	Attempt to quantify misdetection and false alarm rates by running sufficient number of simulations.
OBJ3.4	Medium	Visually represent CPs by including relevant graphs, such as run-length versus time.

Table A.1: List of objectives and their respective priorities ordered by expected time of completion.

Appendix B

Probability distributions

We list the mass and density functions, means and variances of the distributions of models used in Chapter 3 as evidenced in (Evans et al., 2011).

B.1 Poisson

The Poisson distribution is a univariate discrete distribution with:

Support (domain)	$0 \leq x < \infty$
Parameters	intensity $\lambda > 0$
Probability function	$\frac{\lambda^x e^{(-\lambda)}}{x!}$
Mean	λ
Variance	λ

Table B.1: Parameters, mass function, mean and variance of Poisson distribution.

B.2 Gamma

The Gamma distribution is a univariate continuous distribution with:

Support (domain)	$0 \leq x < \infty$
Parameters	scale $b > 0$, shape $c > 0$
Probability function	$(\frac{x}{b})^{c-1}\frac{e^{(-\frac{x}{b})}}{b\Gamma(c)}$, where Γ is the Gamma function
Mean	bc
Variance	b^2c

Table B.2: Parameters, mass function, mean and variance of Gamma distribution.

B.3 Multinomial

The Multinomial distribution is the multivariate generalisation of the Binomial distribution. The k-dimensional Multinomial distribution has:

Support (domain)	$\boldsymbol{x} = (x_1, \ldots, x_k)^T > 0$
Parameters	number of trials $n := \sum_{i=1}^k x_i$, probability $\boldsymbol{p} = (p_1, \ldots, p_k)$ with $0 < p_i < 1 \ \forall i = 1, \ldots, k$ and $\sum_{i=1}^k p_i = 1$
Probability function	$n! \prod_{i=1}^k (\frac{p_i^{x_i}}{x_i!})$
Mean	np_i for dimension $i = 1, \ldots, k$
Variance	$np_i(1 - p_i)$ for dimension $i = 1, \ldots, k$
Covariance	$-np_ip_j$ between dimensions i and j with $i \neq j$

Table B.3: Parameters, density function, mean and variance of Multinomial distribution.

B.4 Dirichlet

The Dirichlet distribution is the multivariate generalisation of the Beta distribution. The k-dimensional Multinomial distribution has:

Support (domain)	$\boldsymbol{x} = (x_1, \ldots, x_k)^T > 0$ with $\sum_{i=1}^k x_i \leq 1$
Parameters	$\boldsymbol{c} = (c_1, \ldots, c_k)^T > 0$ and c_0
Probability function	$\frac{\Gamma\left(\sum_{i=0}^k c_i\right)}{\prod_{i=1}^k \Gamma(c_i)} \prod_{i=1}^k x_i^{c_i - 1}(1 - \sum_{i=1}^k x_i)^{c_0 - 1}$
Mean	$\frac{c_i}{c}$ for dimension $i \ \forall i = 1, \ldots, k$, where $c = \sum_{i=0}^k c_i$
Variance	$\frac{c_i(c-c_i)}{c^2(c+1)}$ for dimension $i \ \forall i = 1, \ldots, k$, where $c = \sum_{i=0}^k c_i$
Covariance	$\frac{-c_ic_j}{c^2(c+1)}$ between dimensions i and j with $i \neq j \ \forall i, j = 1, \ldots, k$, where $c = \sum_{i=0}^k c_i$

Table B.4: Parameters, density function, mean and variance of Dirichlet distribution.

Appendix C

Code

As part of this dissertation, code was written in Python 3 for the implementation of the PG and MD models in Chapter 3. The scripts are found under the names `poisson_gamma.py` and `multinomial_dirichlet_model.py` for the PG and MD models, respectively. They include the following routines:

- `initialization`: Initialises key data structures for storing parameters and probability distributions.

- `evaluate_predictive_log_distribution`: Returns the log densities of $\boldsymbol{y}_t$ using the predictive posteriors for all possible run-lengths $r_t = 0, 1, \ldots, t-1$.

- `evaluate_log_prior_predictive`: Returns the prior log density of the predictive distribution for all possible run-lengths $r_t = 0, 1, \ldots, t-1$.

- `update_predictive_distributions`: Takes next observations $\boldsymbol{y}_t$ and updates sufficient statistics for all possible run-lengths $r_t = 0, 1, \ldots, t-1$.

- `get_posterior_expectation`: Returns the predicted value/expectation from the current posteriors at time point t, for all possible run-lengths.

- `get_posterior_variance`: Returns the predicted variance from the current posteriors at time point t, for all possible run-lengths.

- `prior_update`: Updates the prior expectation & variance to be the posterior expectation and variances weighted by the run-length distribution.

- `prior_log_density`: Computes the log-density of $\boldsymbol{y}_t$ under the prior.

- `trimmer`: Prunes key quantities based on the k most probable run-lengths.

Each of these code scripts extends the `ProbabilityModel` class, which is an abstract class developed by Jeremias Knoblauch. This class is called by the `Detector` object upon receiving a new datum $\boldsymbol{y}_t$.

Bibliography

R. P. Adams and D. J. MacKay. Bayesian online changepoint detection. *University of Cambridge*, 2007. doi: arXiv:0710.3742.

Fabrizio Albertetti, Lionel Grossrieder, Olivier Ribaux, and Kilian Stoffel. Change points detection in crime-related time series: An on-line fuzzy approach based on a shape space representation. *Applied Soft Computing*, 40:441 – 454, 2016. ISSN 1568-4946. doi: https://doi.org/10.1016/j.asoc.2015.12.004. URL http://www.sciencedirect.com/science/article/pii/S1568494615007838.

Samaneh Aminikhanghahi and Diane J. Cook. A survey of methods for time series change point detection. *Knowledge and information systems*, 51(2):339—-367, 2017. doi: 10.1007/s10115-016-0987-z.

Daniel Barry and J. A. Hartigan. Product partition models for change point problems. *The Annals of Statistics*, 20(1):260–279, 1992. doi: 10.1214/aos/1176348521.

Claudie Beaulieu, Jie Chen, and Jorge L. Sarmiento. Change-point analysis as a tool to detect abrupt climate variations. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 370(1962):1228–1249, 2012. ISSN 1364-503X. doi: 10.1098/rsta.2011.0383. URL http://rsta.royalsocietypublishing.org/content/370/1962/1228.

Michael Byrd, Linh Nghiem, and Jing Cao. Lagged exact bayesian online changepoint detection. 2017. doi: 1710.03276.

Hao Chen and Nancy Zhang. Graph-based change-point detection. 2012.

Siddhartha Chib. Estimation and comparison of multiple change-point models. *Journal of econometrics*, 86(2):221–241, 1998.

Noel AC Cressie. *Statistics for spatial data*. Wiley Online Library, 1993.

M. Evans, N. Hastings, and B. Peacock. *Statistical Distributions*. Wiley, 4th edition, 2011.

H. M. Fahmy and E. A. Elsayed. Drift time detection and adjustment procedures for processes subject to linear trend. *International Journal of Production Research*, 44(16):3257–3278, 2006. doi: 10.1080/00207540500410242.

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

James Douglas Hamilton. *Time series analysis*. Princeton Univ. Press, Princeton, NJ, 1994. ISBN 0691042896.

Zaid Harchaoui, Felicien Vallet, Alexandre Lung-Yut-Fong, and Olivier Cappe. A regularized kernel-based approach to unsupervised audio segmentation. pages 1665–1668, 04 2009.

David Hitchcock. The gamma/poisson bayesian model, 2014. URL http://people.stat.sc.edu/Hitchcock/slides535day5spr2014.pdf.

Joseph J. Pignatiello Jr. and Thomas R. Samuel. Estimation of the change point of a normal process mean in spc applications. *Journal of Quality Technology*, 33(1):82–95, 2001. doi: 10.1080/00224065.2001.11980049. URL https://doi.org/10.1080/00224065.2001.11980049.

Perry MB Pignatiello JJ Simpson JR. Estimating the change point of the process fraction nonconforming with a monotonic change disturbance in spc. *Quality and Reliability Engineering International*, 23(3):327—339, 2007.

Yoshinobu Kawahara and Masashi Sugiyama. Sequential change-point detection based on direct density-ratio estimation. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 5(2):114–127, 2011. doi: 10.1002/sam.10124. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sam.10124.

E. Keogh, S. Chu, D. Hart, and M. Pazzani. An online algorithm for segmenting time series. In *Proceedings 2001 IEEE International Conference on Data Mining*, pages 289–296, 2001. doi: 10.1109/ICDM.2001.989531.

Naveed Khan, Sally McClean, Shuai Zhang, and Chris Nugent. Optimal parameter exploration for online change-point detection in activity monitoring using genetic algorithms. *Sensors*, 16(11):1784, 2016.

Michael B. C. Khoo. Determining the time of a permanent shift in the process mean of cusum control charts. *Quality Engineering*, 17(1):87–93, 2004. doi: 10.1081/QEN-200028712. URL https://doi.org/10.1081/QEN-200028712.

J. Knoblauch and T. Damoulas. Bocmds: Bayesian on-line change-point detection & model selection report on the first oxwasp mini-project 2017. Awaiting Approval, 2017.

Chung-Bow Lee. Estimating the number of change points in a sequence of independent normal random variables. *Statistics and Probability Letters*, 25(3):241 – 248, 1995. ISSN 0167-7152. doi: https://doi.org/10.1016/0167-7152(94)00227-Y. URL http://www.sciencedirect.com/science/article/pii/016771529400227Y.

Song Liu, Makoto Yamada, Nigel Collier, and Masashi Sugiyama. Change-point detection in time-series data by relative density-ratio estimation. *Neural Networks*, 43:72 – 83, 2013. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2013.01.012. URL http://www.sciencedirect.com/science/article/pii/S0893608013000270.

Mahmoud A. Mahmoud, Peter A. Parker, William H. Woodall, and Douglas M. Hawkins. A change point method for linear profile data. *Quality and Reliability Engineering International*, 23(2):247–268, 2006. doi: 10.1002/qre.788.

Jesper Møller and Rasmus P Waagepetersen. Modern statistics for spatial point processes. *Scandinavian Journal of Statistics*, 34(4):643–684, 2007.

James F. Nelson. Multivariate gamma-poisson models. *Journal of the American Statistical Association*, 80(392):828–834, 1985a. ISSN 01621459. URL http://www.jstor.org/stable/2288540.

James F. Nelson. Multivariate gamma-poisson models. *Journal of the American Statistical Association*, 80(392):828–834, 1985b. ISSN 01621459. URL http://www.jstor.org/stable/2288540.

E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1):100–115, 1954. doi: 10.2307/2333009.

Joseph J. Pignatiello and James R. Simpson. A magnitude-robust control chart for monitoring and estimating step changes for normal process means. *Quality and Reliability Engineering International*, 18(6):429–441, 2002. doi: 10.1002/qre.487.

Yunus Saatçi, Ryan Turner, and Carl Edward Rasmussen. Gaussian process change point models. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pages 927–934, USA, 2010. Omnipress. ISBN 978-1-60558-907-7. URL http://dl.acm.org/citation.cfm?id=3104322.3104440.

Thomas R. Samuel, Joseph J. Pignatiello Jr., and James A. Calvin. Identifying the time of a step change with x control charts. *Quality Engineering*, 10(3):521–527, 1998. doi: 10.1080/08982119808919166.

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

A. F. M. Smith. A bayesian approach to inference about a change-point in a sequence of random variables. *Biometrika*, 62(2):407–416, 1975. ISSN 00063444. URL http://www.jstor.org/stable/2335381.

M. Staudacher, S. Telser, A. Amann, H. Hinterhuber, and M. Ritsch-Marte. A new method for change-point detection developed for on-line analysis of the heart beat variability during sleep. *Physica A: Statistical Mechanics and its Applications*, 349(3):582 – 596, 2005. ISSN 0378-4371. doi: https://doi.org/10.1016/j.physa.2004.10.026.

Andreas Svensson, Johan Dahlin, and Thomas B Schön. Marginalizing gaussian process hyperparameters using sequential monte carlo. In *Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2015 IEEE 6th International Workshop on*, pages 477–480. IEEE, 2015.

Wayne A. Taylor. Change-point analysis: A powerful new tool for detecting changes, 2000. URL http://www.variation.com/cpa/tech/changepoint.html.

Stephen Tu. The dirichlet-multinomial and dirichlet-categorical models for bayesian inference. *Computer Science Division, UC Berkeley*, 2014.

Melissa J. M. Turcotte and Nicholas A. Heard. Adaptive sequential monte carlo for multiple changepoint analysis, 2015.

Wolfgang Weil. *Spatial Point Processes and their Applications*, pages 1–75. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007. ISBN 978-3-540-38175-4. doi: 10.1007/978-3-540-38175-4_1. URL https://doi.org/10.1007/978-3-540-38175-4_1.

Xiang Xuan. Bayesian inference on change point problems, 2007. URL https://www.cs.ubc.ca/~murphyk/Students/Xuan_MSc07.pdf.

Xiang Xuan and K. Murphy. Modeling changing dependency structure in multivariate time series. *In Proceedings of the 24th international conference on machine learning*, pages 1055–1062, 2007.

Kenji Yamanishi and Jun-ichi Takeuchi. A unifying framework for detecting outliers and change points from non-stationary time series data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 676–681, New York, NY, USA, 2002. ACM. ISBN 1-58113-567-X. doi: 10.1145/775047.775148. URL http://doi.acm.org/10.1145/775047.775148.

Yi-Ching Yao. Estimating the number of change-points via schwarz' criterion. *Statistics and Probability Letters*, 6(3):181 – 189, 1988. ISSN 0167-7152. doi: https://doi.org/10.1016/0167-7152(88)90118-6. URL http://www.sciencedirect.com/science/article/pii/0167715288901186.

G. Udny Yule. Vii. on a method of investigating periodicities disturbed series, with special reference to wolfer's sunspot numbers. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 226(636-646):267–298, 1927. ISSN 0264-3952. doi: 10.1098/rsta.1927.0007. URL http://rsta.royalsocietypublishing.org/content/226/636-646/267.