Novel Methods for the Detection of Emergent Phenomena in Streaming Data

Edward Austin, B.Sc.(Hons.), M.Res

Submitted for the degree of Doctor of Philosophy at Lancaster University.

September 2022

STOR-i
excellence with impact

Abstract

In the fast paced and data rich world of today there is an increased demand for methods that analyse a stream of data in real time. In particular, there is a desire for methods that can identify phenomena in the data stream as they are emerging. These emergent phenomena can be viewed as observations being received that are surprising when compared to the history of the data. Motivated by challenges in the telecommunications sector, we develop methods that operate when the stream does not follow classical assumptions. This includes when the data are not independent or identically distributed, or when the phenomena occur gradually over time.

This thesis makes three contributions to the field of anomaly detection for streaming data. The first, Non-Parametric Unbounded Change (NUNC), provides a non-parametric method for identifying changes in the distribution of a data stream. The second, Functional Anomaly Sequential Test (FAST), provides a method for identifying deviations from an expected shape in a stream of partially observed functional data. The third, mvFAST, extends FAST to the multivariate functional data setting.

Acknowledgements

Firstly I would like to thank my friend and supervisor Idris Eckley for his tireless patience and assistance throughout my PhD. I am most grateful for the independence he has allowed me, and for his support turning my mis-shapen clay into the pottery it is today. Furthermore, I must also express my appreciation for the support he has given to me, and my family, throughout my PhD (in particular during my challenging first half of 2022!) I would also like to thank Lawrence Bardwell for his supervision during the first 15 months of my PhD; without his guidance my transition to independent research would have been a far more bumpy ride. Finally, I would like to thank the wider STOR-i, NG-CDI, and StatScale communities for the stimulating working environment that they have provided to me.

The work in this thesis would not have been possible without my collaborators: Gaetano Romano, Paul Fearnhead, and Priyanga Talagala. Their efforts helped shape my research, and their time and support helped enhance my contributions. I also cannot thank Gaetano enough for the time he spent over Teams throughout lockdown being the sounding board for all of my new ideas (and helping me to think of some of the acronyms, although Darth Fader never did make it...)

I am thankful to the STOR-i CDT, EPSRC, and British Telecommunications PLC (BT) for their funding of this thesis. I am grateful for BT and their employees, David Yearling, Adam Broadbent, and Trevor Burbridge for their helpful discussions and interesting applications that have inspired much of my research. Furthermore I would like to thank Peter Willis, formerly of BT, for getting the ball rolling with the anomaly detection challenge that kicked off this thesis.

On a personal level, I would like to thank my partner Lily for supporting me in

everything that I do. You have helped me reach far beyond where I would have gotten to alone. Lastly - despite, perhaps ironically, now always coming first - I would like to thank my newborn daughter Bonnie, born February 2022, for providing me with the perfect playmate with whom to break up the monotony of finishing off my PhD.

This thesis is dedicated to my grandma, Jean Wolstenholme.

Keep dancing with the daffodils.

Declaration

This thesis has not been submitted, either in whole or in part, for a degree at this, or any other, university. I declare that the work presented in this thesis is, to the best of my knowledge and belief, original and my own work. Listed below is a description of the status of each paper forming the main corpus of this thesis, and the external contributions to the sections I did not directly contribute to that have been left in place for completeness.

The work in Chapter 3 has been accepted for publication in the journal Computational Statistics and Data Analysis (Austin et al., 2023). The manuscript is a collaboration with my fellow PhD student, Gaetano Romano, his supervisor Paul Fearnhead, and our supervisor Idris Eckley. Our contributions to the paper are equal with myself leading the theoretical and application-focused aspects of the work, whilst G. Romano focused on the more computational aspects of the paper.

Chapter 4 is currently under review with Technometrics. The manuscript is joint work with Idris Eckley and Lawrence Bardwell. For the first 15 months of my PhD Lawrence was my supervisor, before he moved onto a new role in industry.

The work in Chapter 5 is joint work with Idris Eckley and Priyanga Talagala. My contribution was the development of the method, obtaining the theoretical results, and performing both the simulation and real data studies.

This thesis contains $28,572$ words.

Edward Austin

Contents

List of Figures

List of Tables

List of Abbreviations

AIC	Akaike Information Criterion
AR	Autoregressive
ARMA	Autoregressive Moving Average
BIC	Bayesian Information Criterion
CDF	Cumulative Distribution Function
CUSUM	Cumulative Sum
EWMA	Exponentially Weighted Moving Average
FAST	Functional Anomaly Sequential Test
FDA	Functional Data Analysis
FO	Functional Outlyingness
FOCuS	Functional Online Cumulative Sum
FPCA	Functional Principal Component Analysis
FPC	Functional Principal Component
FWER	Family-Wise-Error-Rate
i.i.d.	Independent and Identically Distributed
ISE	Integrated Square Error
KS	Kolmogorov-Smirnov
MA	Moving Average
MBD	Modified Band Depth
MCMC	Markov Chain Monte Carlo
MEI	Modified Epigraph Index
MLE	Maximum Likelihood Estimator
MO	Mean Outlyingness

MOSUM	Moving Sum
MS	Magnitude-Shape
NUNC	Non-Parametric Unbounded Change
ODE	Ordinary Differential Equation
OU	Ornstein-Uhlenbeck
PDA	Principal Differential Analysis
PFA	Probability of a False Alarm
ROC	Receiver Operating Characteristic
SLR	Sequential Likelihood Ratio
STR	Sequential Transform and Directional Outlyingness
VAR	Vector Autoregression
VO	Variational Outlyingness

Chapter 1

Introduction

Advances in computational power, and the growing emphasis on timely data driven decision making, mean that the analysis of a data stream in real time is a challenge that has become increasingly important in recent years. Motivated by the need to know what is happening *right now*, this real-time analysis has been applied in a variety of applications including healthcare (Aziz et al., 2016), wireless networks (Onibonoje and Olowu, 2017), urban infrastructure (Wan et al., 2022), air quality (Megantoro et al., 2021), and the Covid-19 pandemic (Leon et al., 2020). The work in this thesis focuses on a particular aspect of the streaming data analysis challenge: the detection of emergent anomalous phenomena. These are observations that can, in some way, be considered to be surprising when compared to previously observed data within the stream.

The detection of emergent anomalous phenomena is of critical importance in many sectors. One such sector is the digital infrastructure within today's telecommunications industry. The emergence of anomalous phenomena in this setting might, for example, indicate the occurrence of a significant event such as a service outage or hardware failure. It can be critical to identify these events before they impact on an end user. This is not only to preserve user connectivity, but also to help assure the reputation of a leading national infrastructure service provider.

This thesis is motivated by several challenges drawn from the telecommunications sector, and presents three novel methods for the detection of emergent phenomena

in a range of settings. These settings include when the data stream does not follow classical assumptions such as being independent or identically distributed, or when the phenomena emerge gradually over time. The remainder of this thesis is organised as follows: in Chapter 2 we review the literature pertinent to the work presented in this thesis. Specifically the focus of this chapter will be on sequential change and anomaly detection, and functional data analysis.

In Chapter 3 we introduce NUNC, a novel non-parametric window based algorithm for detection of changes in the distribution of a data stream. This method has been developed in response to an applied challenge of monitoring an operational metric on a given line with a view to taking pre-emptive remediation ahead of the customer reporting an issue. We present such an application along with empirical evaluations of the method, and a discussion on the computational properties of the algorithm. Additionally, we establish theoretical results for the choice of points to search in the window, and for the choice of the test threshold in order to control the false alarm rate at a desired level.

Chapter 4 introduces a method for monitoring partially observed functional data for emergent anomalies, and Chapter 5 extends this to the multivariate setting. We name this method the Functional Anomaly Sequential Test (FAST) and demonstrate the performance of our method in a range of settings, including monitoring points on an internet network. We also establish several theoretical results concerning threshold selection, bounds for the detection power, and a consistency result for mvFAST.

Finally, we close the thesis with a discussion on our contribution to the literature, and potential avenues for future work, in Chapter 6.

Chapter 2

Literature Review

This chapter seeks to set the foundations for the research introduced in Chapters 3 to 6. To achieve this, we draw upon various areas of statistics including sequential changepoint detection, functional data analysis, and anomaly detection. In Section 2.1 we begin by reviewing sequential changepoint methods including control charts, and window based approaches. We then turn to introduce functional data analysis (FDA) techniques in Section 2.4, exploring two pertinent FDA approaches: Functional Principal Components Analysis (FPCA) and Principal Differential Analysis (PDA). The chapter concludes with an overview of the literature on FDA applied to statistical anomaly detection.

2.1 Sequential Changepoint Detection

Before introducing the sequential changepoint detection setting, it is perhaps first helpful to consider the offline version of the changepoint problem. In the classical univariate changepoint setting we consider a sequence of observations $x_1, \ldots, x_n$, which are realisations of the random variables $X_1, \ldots, X_n$ respectively, containing a set of changepoints $\{\nu_0, \ldots, \nu_K\} \in \mathbb{N}$ with $\nu_0 = x_1$ and $\nu_K = x_n$. These $K + 1$ values divide the sequence of observations into K segments such that $X_i \sim F_k$ for $\nu_{k-1} < i \leq \nu_k$, and each $1 \leq k \leq K$. Here F_k is the (possibly unknown) data generating process for the kth segment of the data. The goal of a changepoint analysis is to accurately

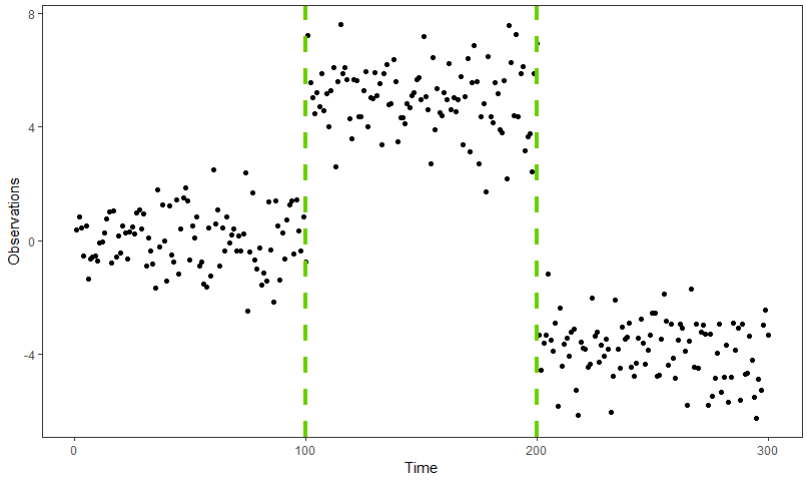


Figure 2.1.1: An example time series containing $K = 3$ segments. The dashed green lines denote the changepoint locations.

estimate the values of the $\{\nu_k\}_{k=1}^{K}$ based on the observed time series. An example sequence containing $K = 3$ segments is presented in Figure 2.1.1.

Many methods have been proposed for the detection of changepoints in an offline setting. See Truong et al. (2020) for an excellent overview of the area. In the applications addressed in this thesis, however, we make use of sequential changepoint detection tools. In this setting the data are collected individually, in sequence. As each new observation is made a test is performed for the occurrence of a change (Tartakovsky et al., 2014). Consider, by way of an example, a univariate sequence of observations $x_1, x_2 \ldots,$. A sequential approach amounts to testing, at each $t > 1$, whether there exists a $1 \leq \nu \leq t$ such that $x_1, \ldots, x_\nu$ are drawn from a data generating process F, and $x_{\nu+1}, \ldots, x_t$ are drawn from a different data generating process $G \neq F$. Here F and G denote the (potentially unknown) pre- and post-change data generating processes respectively.

An example time series that is being observed sequentially, and that contains a change, is depicted in Figure 2.1.2. In this figure we see a sequence of observations following the same data generating process (a), before a change in the mean of the process occurs (b). After a sufficient number of post-change observations have been seen an alarm is declared and a change is identified (c).

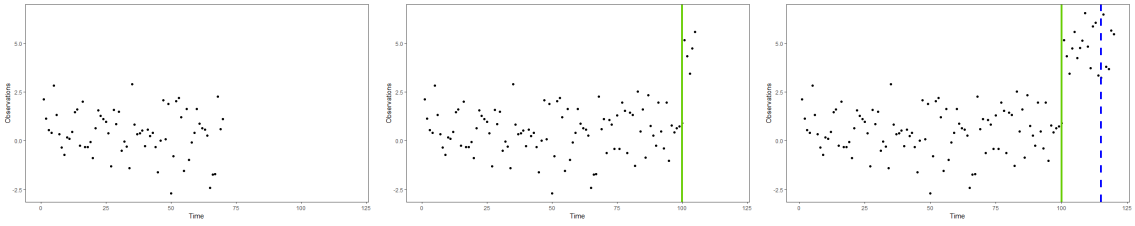
In the sequential setting the goal is not to identify the precise location of a change but instead to detect the occurrence of a change, as quickly as possible after it has happened, without declaring an excessive number of false alarms (Tartakovsky et al., 2014). This goal gives rise to the following two key metrics: (i) the probability of false alarm, and (ii) the detection delay. Letting ξ denote the stopping time for the sequential change detection test, with the stopping time indicating when the method declares a change has been detected, the probability of false alarm is defined as (Tartakovsky et al., 2014)

$$\mathrm{P}\left(\text{False Alarm}\right) = \mathrm{P}\left(\xi < \infty | \text{ No Changepoint }\right). \tag{2.1.1}$$

That is, it is the probability that the test incorrectly declares a change has occurred. The second key metric is the detection delay, $\mathcal{D}$. This is defined as (Johnson et al., 2017)

$$\mathcal{D} = \xi - \nu, \tag{2.1.2}$$

where ν is the true changepoint location.



(a) (b) (c)

Figure 2.1.2: In this figure we present an example of sequential change detection. In (a) we show an example time series containing no change. In (b) we show that the change has occurred (solid green line) but has not yet been detected. In (c) the change has been detected (dashed blue line).

Several different types of sequential changepoint detection methods have been proposed in the literature. For a broad review of the area, see Tartakovsky (2019). Two of the most common approaches are control charts and moving windows. For an excellent introduction to control charts and moving window based methods see Qiu (2013) and Aminikhanghahi and Cook (2017) respectively.

Both control charts and window based approaches are utilised in the work of this thesis. In Chapter 3 we will make use of a window-based approach as part of a nonparametric sequential changepoint detection method. Conversely in Chapters 4 and 5 we shall make use of a control chart approach to detect anomalies within sequentially observed functional data.

In the next section we will review the relevant literature in both these areas. However, we take this opportunity to note that numerous other approaches for sequential change detection have been introduced. These include Bayesian (Shiryaev, 1963; Adams and MacKay, 2007; Byrd et al., 2018; Agudelo-España et al., 2020), dynamic programming (Fisch et al., 2022a), and convex optimisation (Cao et al., 2018) based techniques. Each of these tools require parametric assumptions about the distribution of the data being observed and so, as we shall see, are unsuitable for use in the settings we consider in this thesis. As such, in this review we focus solely on the control chart and moving window methods as they are the most pertinent to the later chapters of this work.

2.2 Control Charts

First introduced by Shewhart (1930), a control chart is a method to monitor for a change to a process over time. See Qiu (2013) for a comprehensive review.

Formally, under the null hypothesis of no change, the process will generally be independent and identically distributed (i.i.d.). To monitor for a change to the process, a control chart computes a sequence of test statistics, $S_1, S_2, \ldots$, based on the data observed so far. Specifically, at each time t the test statistic, S_t, is computed from the observed data $x_1, \ldots, x_t$. This test statistic is then updated sequentially as each new observation arrives. Thus when x_{t+1} is observed S_{t+1} can be computed. As we shall see later, various forms for the test statistic have been considered in order to capture different properties of the data stream sequence.

To detect a change, a control chart uses the cumulative sum of the test statistic, $\Delta_t = \sum_{i=0}^{t} S_i$. Should Δ_t exceed a specified threshold then the control chart declares

that a change has occurred at observation t. This is represented by the stopping time, ξ, where

$$\xi = \inf \{t : \Delta_t > U \text{ or } \Delta_t < L\}. \tag{2.2.1}$$

Here U and L represent the upper and lower thresholds for the chart. Often, these thresholds are selected so that the probability of false alarm is controlled at some pre-determined level α (Polunchenko and Tartakovsky, 2012).

An early, and perhaps best known example of the control chart methods, is the approach introduced in Page (1954). The so-called CUSUM approach monitors cumulative sums of some statistic based on the observed data stream. Over the years various CUSUM schemes have been proposed. We will review some of the most common approaches below.

2.2.1 CUSUM Charts

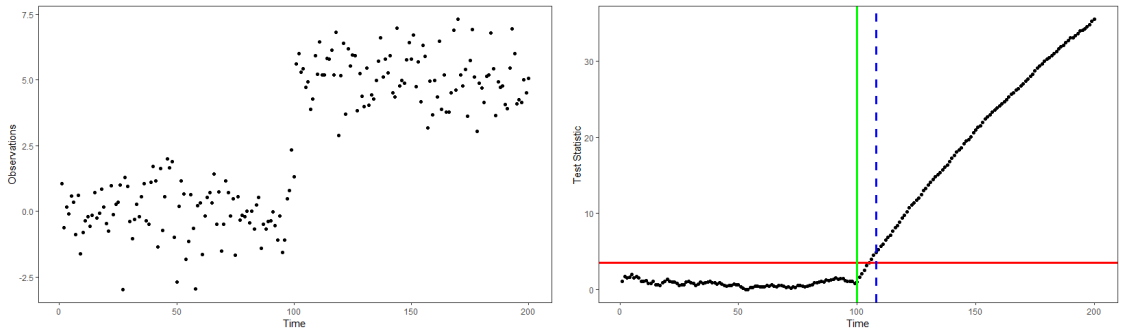
The original parametric CUSUM chart of Page (1954) tests for a change in the mean of a mean-zero data stream using the test statistic $S_t = \max(0, S_{t-1} + x_t)$, $S_0 = 0$. We then test to see whether the partial sum

$$S_t = \frac{1}{\sqrt{t}} \sum_{r=0}^{t} S_r, \tag{2.2.2}$$

exceeds the control chart threshold. As discussed above, this threshold will typically have been selected to control the probability of a false alarm at a pre-specified level. Note that the factor of $\sqrt{t}$ is introduced in equation (2.2.2) to control the variance of the partial sum. An example trace of the Page CUSUM test statistic is presented in Figure 2.2.1. Observe how, once the changepoint has occurred, the test statistic begins to increase rapidly leading to the CUSUM scheme declaring an alarm.

Kirch and Weber (2018) discuss an extension to the CUSUM test, the so-called Page-CUSUM. This test allows for any parameter, not just the mean, to be monitored by replacing the statistic, S_t, in equation (2.2.2) with a score statistic:

$$S(x_t; \theta) = \frac{d}{d\theta} \log \left(f(x_t, \theta) \right).$$



(a) (b)

Figure 2.2.1: In (a) we present an example time series that is tested for a change using the CUSUM test in equation (2.2.2). In (b) we plot the trace of the CUSUM test statistic, with the solid vertical green line the true change location, the horizontal line the test threshold, and the dashed blue line the time of detection.

Here θ is a parameter for the distribution of the data under the null hypothesis and $f(\cdot, \theta)$ is the corresponding density. The Page-CUSUM test takes the following form

$$K_t = \max_{0 \leq w \leq t} \frac{1}{\sqrt{w}} \sum_{r=t-w}^{t} |S(x_r; \theta)|, \tag{2.2.3}$$

The maximisation over w is implemented to increase test power by searching over a range of different segments, but at the price of quadratic computational cost (Yu et al., 2020).

An alternative parametric approach to the CUSUM-based approaches based on the chart of Page (1954) is the sequential likelihood ratio (SLR) test (Lorden, 1971). The SLR test is designed to detect changes in the distribution of the data, rather than changes in a set of parameters. To this end let $f(\cdot, \boldsymbol{\theta})$ and $g(\cdot, \boldsymbol{\theta})$ be the pre- and post-change densities, where $\boldsymbol{\theta}$ is a vector of parameters. Lorden (1971) then defines the sequential test as follows:

$$L_t = \max_{2 \leq \nu \leq t} \sum_{j=\nu}^{t} \log \frac{g(x_j; \boldsymbol{\theta})}{f(x_j; \boldsymbol{\theta})}. \tag{2.2.4}$$

One important challenge with Lorden's approach is that the post-change distribution is often unknown. To overcome this, the generalized likelihood ratio (GLR) test

was proposed by Siegmund and Venkatraman (1995). This replaces the unknown post change density, g, with the maximum likelihood estimate (MLE). Consequently only the family of the post-change distribution, rather than the full distribution, needs to be specified. In practice it can be challenging to specify this post-change distribution. This could occur, for example, because the data does not follow a parametric model. Motivated by this challenge, in Chapter 3 we present a novel non-parametric method for sequential changepoint detection where neither the pre-, nor post-, change distribution needs to be known.

The selection of the post-change distribution is not the only issue with the generalized likelihood ratio test. A second disadvantage is the computational cost of computing the post-change MLE, which is quadratic in the length of the data stream. One solution to this increased computational cost is the use of the sliding window approach that we describe in Section 2.3. Recently, however, Romano et al. (2021) have sought to address this issue from a control chart perspective. In particular, they provide an exact test for the Gaussian change in mean problem, FOCuS (Fast Online Cumulative Sum), with a logarithmic computational cost on average.

Over the years various alternative approaches to the parametric CUSUM test have been detailed in the literature. One such example is the Bayesian formulation of the CUSUM test known as the Shiryaev-Roberts procedure (Shiryaev, 1963). The test has also been adapted to determine deviations from a particular form of underlying model. Examples include the CUSUM test for changes in the parameters of a linear model (Ploberger and Kramer, 1992), the structure of hidden markov models (Fuh, 2003), sets of panel data (Wu, 2019), and the parameters of time series models (Song and Kang, 2020).

Each of the parametric CUSUM tests presented so far make the assumption that the observations are independent and identically drawn from some null distribution. Whilst many of the aforementioned methods allow for the parameters of the distribution to be estimated from the data, they at least require that the family of the distribution be specified up front. Such assumptions are inappropriate for the data structures considered in Chapters 4 and 5 of this work, where the data need not be

independent, identically distributed, or drawn from a known distribution. Motivated by the need to relax some of these assumptions, Tartakovsky et al. (2005) propose a non-parametric CUSUM test, and it is this approach that we shall build upon in Chapter 4 and 5. We briefly review the non-parametric CUSUM literature in the next section.

2.2.2 Non-Parametric CUSUM Chart

The essential ingredients of the non-parametric CUSUM control chart mirror that of the parametric one: the observations are used to compute a test statistic, and if the cumulative sum of this test statistic exceeds a threshold then an alarm is declared. The difference between the two approaches, however, is that in the non-parametric approach the test statistics do not rely upon there being an underlying distribution for the observed data (Hušková and Hlávka, 2012).

The non-parametric CUSUM control chart was first introduced by Tartakovsky et al. (2005). Under their approach the test statistic is formulated as

$$H_t = \sum_{j=1}^{t} h(x_j), \tag{2.2.5}$$

where $h(\cdot)$ is a user-defined function of the observed data. Tartakovsky et al. (2005) note that an appropriate choice for $h(\cdot)$ can overcome the need to assume the observed data are i.i.d. Furthermore, $h(\cdot)$ can be selected so that a particular property of the data stream can be monitored for a change. Tartakovsky et al. (2013) propose several choices for $h(\cdot)$, for example $h(\cdot) = x_t - m_t$, where m_t is the median estimated from the first t points; or $h(\cdot) = C_1 x_t + C_2 x_t^2 - C_3$, for constants C_i. The former choice is designed for monitoring for changes in the location of a data stream where the distribution for the data is heavy-tailed, whereas the latter is designed for monitoring for changes in both the mean and variance simultaneously.

One disadvantage of the non-parametric CUSUM chart, however, is that it may be time consuming to identify a suitable $h(\cdot)$ function. Furthermore, there are no theoretical guarantees that the selected function is the best possible, or even near optimal. That said, this approach does allow for the CUSUM test to be extended to

a broad number of settings, and in Chapter 4 we will incorporate this method into our anomaly detection test.

Whilst the classical CUSUM test, and variants of it, are some of the most common control charts in the literature various other monitoring schemes have been proposed. In the next section we shall present a brief overview of some other relevant recently proposed approaches.

2.2.3 Other Related Approaches

In a similar vein to the non-parametric CUSUM test of Tartakovsky et al. (2005), other control charts have been suggested that do not assume any knowledge of the distribution for the data. These methods use U-statistics to detect changes (Kirch and Stoehr, 2019). Example U-statistics include the Mann-Whitney (Mann and Whitney, 1947) or Mood (Mood and Graybill, 1963) statistics, used for changes in scale and location respectively (Kowalski and Tu, 2008). Several authors implement the U-statistic approach, for example Ross and Adams (2012); Zhou et al. (2014), and Xiang et al. (2019). In more recent literature, the U-Statistic approach has also been utilised to detect anomalies in autoregressive (AR) models, see Lee (2020) and Gamage and Ning (2020).

One issue with each of the control chart methods presented thus far is that, under the null hypothesis, they require the data to be drawn from a distribution that is weakly stationary - that is, the mean and variance are constant over time. Unfortunately such an assumption is not always suitable for the data generating process at hand. Work has been done to address this challenge by Qiu and Xiang (2014), who allow for a time varying mean structure under the null using a first order AR model. This is extended by Mathieu et al. (2020), who define the following model for the data

$$x_t = \begin{cases} \eta_t + v_t & 1 \leq t \leq \nu \\ \eta_t + v_t + f_{t,\delta} & t > \nu. \end{cases} \tag{2.2.6}$$

Here η_t is an Moving Average (MA) process, v_t is a slowly varying function, and $f_{t,\delta}$ represents a change in mean by the level δ. The use of an MA process and slowly

varying function in equation (2.2.6) allows for non-stationarity, and serial correlation, to be modelled under the null. A shortcoming of this method, however, is the requirement that the change be abrupt and occur as a step change at time ν. In Chapter 4 of this work we consider an application where changes from our non-stationary mean emerge gradually, rather than as a step change at a single point in time. As a result of this, the works of Qiu and Xiang (2014) and Mathieu et al. (2020) are unsuitable for use in our setting.

Little exists in the control chart literature regarding gradually emerging changes, however Shu et al. (2008) offers one approach by modelling a change from a constant mean under the null, to a mean that follows an ARMA time series model under the alternative. The effect of this is that the change can emerge gradually and increase in magnitude as more points are observed. As a result, the work of Shu et al. (2008) offers some progress towards the detection of gradual changes, albeit with the caveats that the post-change data follows an ARMA structure, and that the pre-change distribution is stationary. In Chapter 4 we will build upon Shu et al. (2008) by modelling both non-stationarity under the null and the emergence of gradual changes.

Another well studied alternative to the CUSUM chart is the exponentially weighted moving average (EWMA) chart. This chart, introduced by Hunter (1986) for use in sequential changepoint detection, features a test statistic given by

$$E_t = \lambda E_{t-1} + (1-\lambda) j(x_t, \ldots, x_{t-w+1}), \tag{2.2.7}$$

where $j(x_t, \ldots, x_{t-w+1})$ is a function of the previous w data points and $0 \leq \lambda \leq 1$. The advantage of this method is that it can place emphasis on changes in the most recent observations by suitably weighting equation (2.2.7) towards the contribution of the second term. This is in contrast to the CUSUM, where all points contribute equally to the test statistic. The EWMA chart has been applied by other authors in a variety of settings, including Dong et al. (2008) and Raza et al. (2015) for changes in non-stationary data, Ross (2014) and Keriven et al. (2020) for changes in distribution, and Plasse et al. (2021) for changes in transition matrices.

2.3 Window Based Sequential Methods

In contrast to the control chart approach, window based changepoint detection seeks to identify changes in a stream of data occurring within the points contained within a window. For a window of size w, this therefore corresponds to analysis of the data stream $x_{t-w+1}, \ldots, x_t$.

We begin our review of moving window based methods with the MOSUM, before proceeding to explore some alternative approaches.

2.3.1 The MOSUM Test

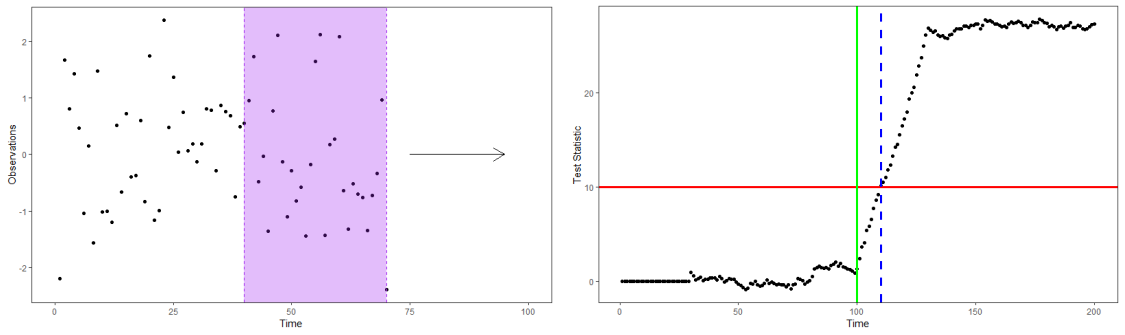
The MOSUM test of Bauer and Hackl (1978) monitors the cumulative sum of data within a sliding window in order to detect a change in mean. The test uses the statistic

$$R_{w,t} = \frac{1}{\sigma\sqrt{w}} \sum_{i=t-w+1}^{t} x_i - \mu, \tag{2.3.1}$$

where μ is the mean under the null, and σ is the standard deviation of the data. Under the MOSUM test, an alarm is declared if $|R_{w,t}| > \gamma$, where γ is the test threshold. The work of Bauer and Hackl (1978) was subsequently extended by Hušková (1990), and Chu et al. (1995), where asymptotic distributions for γ are proposed.

An example of a window based analysis is given in Figure 2.3.1. Figure 2.3.1(a) depicts a stream of data, and a window denoted by the shaded purple region. Any changepoint analysis that takes place is carried out using only the data contained within the window, and this window moves along the data stream as new observations are received. In Figure 2.3.1(b) we show the trace of the MOSUM test statistic, $R_{w,t}$, for the sample of data contained in Figure 2.2.1(a). Observe how, unlike the CUSUM test statistic plotted in Figure 2.2.1(b) the size of the MOSUM test statistic only increases significantly when the window contains the changepoint.

A more recent work by Eichinger and Kirch (2018) explores the utility of the MOSUM test to detect changes in the mean of an i.i.d. data stream following any



(a) (b)

Figure 2.3.1: In (a) we present an example of a window based change analysis. The window is of size $w = 30$ and depicted by the shaded region between the dashed purple lines. In (b) we plot the trace of the MOSUM test statistic on the data from Figure 2.2.1(a), with the solid vertical green line the true change location, the horizontal line the test threshold, and the dashed blue line the time of detection.

distribution. They define their MOSUM statistic as

$$M_{w,t} = \frac{1}{\sigma\sqrt{2w}} \left| \sum_{j=t-w+1}^{t} x_j - \sum_{j=t-2w+1}^{t-w} x_j \right| \tag{2.3.2}$$

and, similarly to the original MOSUM, declare a change if $|M_{w,t}| > \gamma$. As can be seen from equation (2.3.2), their test differs from the classical MOSUM test by searching not a window of size w, but by comparing two windows of size w against each other in order to identify a change in mean. This is proposed because it transpires that the test power is maximised in the middle of the window. This comes at the cost, however, of a potential increase in detection delay due to the wait for the changepoint to be located at the centre of the window.

There are several advantages to using the MOSUM over a control chart method for sequential changepoint detection. In particular, using a moving window provides the advantage of a bounded computational cost, overcoming the issue control charts face as the length of the data stream increases. It also introduces issues, however, such as the need to select an appropriate window size. That said, recently Xie et al. (2022) have provided a method for choosing the optimal window size under the assumption that the data are i.i.d.

Whilst the MOSUM test has generally been designed to detect a change in the mean of a data stream, Hsu (2007) have also applied the method to detect changes in variance. More recently, several authors have proposed sliding window based methods that seek to detect a broader range of changes than those in mean or variance. We overview some of these alternative approaches in the next section.

2.3.2 Alternative Sliding Window Based Methods

An example of a more generalized variant of a window based change detection method is outlined by Kirch and Weber (2018). Their approach utilises windowed sums given by

$$H_{w,t} = \sum_{j=t-w+1}^{t} h(x_j, \hat{\theta}), \tag{2.3.3}$$

where $h(\cdot, \cdot)$ is an estimating function and θ is a parameter estimated such that $\sum_{j=1}^{m} h(x_j, \hat{\theta}) = 0$, for a set of changepoint free historical data $x_1, \ldots, x_m$. Many existing window based methods, for example the MOSUM given in equation (2.3.1), are contained in the definition of $H_{w,t}$. Kirch and Weber (2018) demonstrate that location changes, changes in the parameters of non-linear models, and changes in integer-valued time series can all be detected by a suitable choice of estimating function. The complication here, however, is identifying such a function. Furthermore, there is a need to identify a set of training data that is changepoint free to set the parameter θ.

As Tout et al. (2018) discuss, a window based algorithm is also a useful tool for analysing a non-stationary stream of data. The authors propose a method similar to the MOSUM of Eichinger and Kirch (2018) to identify changes in the mean of a non-stationary data stream. A drawback of the method, however, is that it requires the data follow a Gaussian distribution and has constant variance across all observations. Talagala et al. (2020) present a windowed algorithm, OddStream, that relaxes these assumptions and detects changes in non-stationary and non-identically distributed data. To do so their method monitors a matrix of time series features generated from the windowed data. By monitoring several features at once, a broader range of changes

can be detected. A disadvantage of the method, however, is that the detection step uses a principal component projection of the feature matrix and so which features have changed cannot be identified.

Whilst each of the window based techniques presented so far search for a change inside the window of data, recently an alternative approach has been proposed where the window of data is compared to the out of window observations. Gösmann et al. (2020), for example, compare the estimate of the parameter of interest, θ, on the first $t - w$ points with the estimate obtained from the most recent w points:

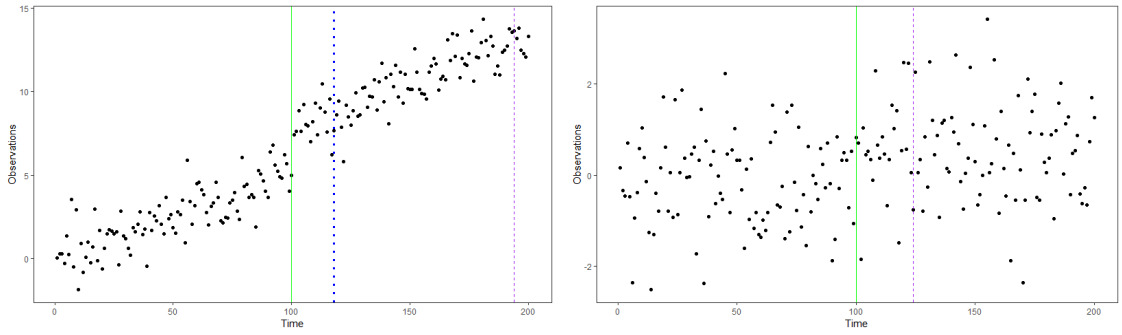
$$D_{t,w}(\theta) = |\hat{\theta}_{1:t-w} - \hat{\theta}_{t-w+1:t}|. \tag{2.3.4}$$

Here $\hat{\theta}_{s:t}$ is the estimate of the parameter of interest using observations $x_s, \ldots, x_t$. A similar approach is taken by Dette and Gösmann (2020) who instead use a likelihood ratio test to compare $\hat{\theta}_{1:t-w}$ and$\hat{\theta}_{t-w+1:t}$. The advantage of these approaches is that using the first $t - w$ observation to estimate θ offers a more accurate estimate than simply using the windowed data should the parameter of interest be unchanged over the whole data stream. A drawback of both Gösmann et al. (2020) and Dette and Gösmann (2020), however, is that they require parametric models for the data.

An interesting variation on the traditional window based method has been recently proposed by Padilla et al. (2019b), who introduce a test for a change in the distribution of a data stream. In contrast to the other works discussed so far, their approach does not use a single fixed window size, but instead computes a set of test statistics using multiple window sizes and then takes the maximum of these statistics. The test statistic used by Padilla et al. (2019a) is a Kolmogorov-Smirnov (KS) test statistic:

$$K_{s:t} = \sqrt{t - s + 1} \sup_y \left| F_0(y) - \hat{F}_{s:t}(y) \right|, \tag{2.3.5}$$

Here $\hat{F}_{s:t}(y)$ is the CDF estimated from the points $x_s, \ldots, x_t$ and $F_0(\cdot)$ is either the true CDF, or the CDF estimated from a sample of data drawn from the null distribution. The window statistic is then defined as $W_t = \max_{t-L \leq s \leq t} \hat{K}_{s:t}$. Here L is a cutoff point for the backwards movement of the window. The benefit of this approach is that considering various window sizes can improve the power of the test. On the



(a) (b)

Figure 2.3.2: In (a) we present an example time series that is non-stationary but contains a change in mean at time $\nu = 100$. The solid green line is the true changepoint, the dotted blue line the MOSUM detection time, and the dashed purple line the CUSUM detection time. In (b) we show an example where the change at time $\nu = 100$ is of a much smaller magnitude (equal to 0.5). Observe how the CUSUM test detects the change (dashed purple line) but the MOSUM test does not.

other hand, it also increases the computational cost by a factor of L as $K_{s:t}$ must be computed for each $t - L \leq s \leq t$.

Now that we have provided an overview of some of the key literature in the fields of both control charts and sliding windows it can be instructive to contrast these two schools of thought. In particular, in Figure 2.3.2 we present two examples for a change in mean comparing the CUSUM control chart with the MOSUM test. Figure 2.3.2 (a) shows how a sliding window method can provide favourable results compared to a control chart when the data stream is non-stationary. The reason for this is that a large value is needed for the control chart threshold in order to control the false alarm rate, and this in turn reduces test power and increases detection delay. On the other hand, we see that in Figure 2.3.2(b) the control chart offers better detection power when the change magnitude is smaller. This is because the control chart considers the whole of the data stream. As such, it is more sensitive to changes that are smaller in magnitude, whereas the MOSUM only considers the points in the window.

Now that we have laid the foundations for sequential changepoint detection, we can turn our attention to the other area of statistics that is frequently used in this

thesis: Functional Data Analysis. The next part of this literature review will briefly overview the essential ingredients needed to work in a functional data setting, before then moving on to an analysis of existing functional anomaly detection techniques.

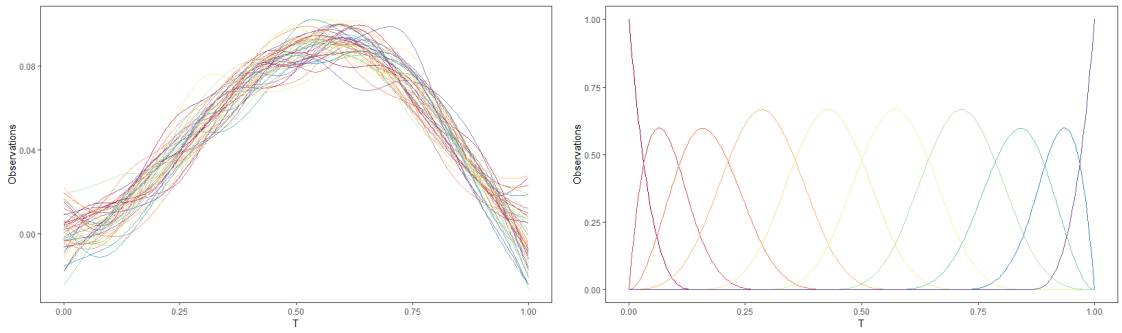
2.4 Functional Data

Functional data analysis is a growing field of statistics that focuses on the analysis of curves of data, rather than individual points. These curves can represent any observation that varies over a continuum, and it is typical for this continuum to be time. A wide variety of functional data methods exist in the literature, and for overviews Ramsay (2005) or Ferraty and Romain (2011) can be consulted. Hsing and Eubank (2015) also offer a theoretical perspective on the field. The main focus of this section is anomaly detection in functional data. In order to adequately describe the state of the art in this area we first introduce the foundations of FDA to the reader, and in particular highlight the approach of Ramsay (2005).

2.4.1 Foundations of Functional Data Analysis

Observations in functional data analysis consist of curves of data, $\{X_1(t), \ldots, X_n(t)\}$, where $t \in \mathcal{T}$. Here $\mathcal{T}$ is the compact interval the curves are observed over, and $X_i(t) \in L^2(\mathcal{T})$. In general, $\mathcal{T} = [a, b] \in \mathbb{R}$, however this is not a requirement. In Figure 2.4.1(a) we show an example of functional data observed over the interval $[0, 1]$. Notice how each observation consists of a smooth curve, and each curve has been observed over the same time interval. We also remark that in this example each of the curves has a similar shape, however this is not a requirement of functional data in general. The computations in Figure 2.4.1(a), along with all other functional data computations in this thesis, have been performed using the R package fda (Ramsay et al., 2021) unless stated otherwise.

As Ramsay (2005) note, in practice it is difficult to observe data in a continuous manner. To obtain a sample of functional data, then, an approximation for the functional observations, $\{X_i(t)\}$, is sought. This is obtained by using basis functions



(a) (b)

Figure 2.4.1: Figure (a) depicts 50 functional observations over the interval [0,1]. Figure (b) shows an example system of $K = 10$ cubic B-splines over the same interval.

to smooth a sample of point data observed over $\mathcal{T}$ into a set of curves. To be concrete, first consider $\mathcal{T}$ discretised into a fine grid of T points: $\{1, \ldots, T\}$. Then, if $x_{i1}, \ldots, x_{iT}$ represent the i^{th} discrete sample observed over the grid, and $\{b_k(t)\}_{k=1}^{K}$ are the K basis functions, a curve can be approximated from these points as

$$\hat{X}_i(t) = \sum_{k=1}^{K} c_{ik} b_k(t). \tag{2.4.1}$$

The coefficients $\boldsymbol{c_i} = \{c_{ik}\}_{k=1}^{K}$ are estimated such that

$$\hat{\boldsymbol{c_i}} = \arg\min_{\boldsymbol{c_i}} \sum_{t=1}^{T} ||x_{it} - \hat{X}_i(t)||_2^2 = \arg\min_{\boldsymbol{c_i}} \sum_{t=1}^{T} ||x_{it} - \sum_{k=1}^{K} c_{ik} b_k(t)||_2^2, \tag{2.4.2}$$

where $||\cdot||_2$ represents the L^2 norm.

A wide range of basis systems, $\{b_k(t)\}_{k=1}^{K}$, can be used including Fourier basis, wavelets, power series, and splines. An example set of 10 cubic b-spline basis functions is presented in Figure 2.4.1(b), and these are used to smooth a univariate time series in Figure 2.4.2. Alternatives to the basis function system approach have also been proposed in the literature, including the use of k-nearest neighbours algorithms (Ferraty and Vieu, 2006), kernel smoothers (Zhang and Chen, 2007), or Gaussian processes (Hall et al., 2008).

Systems of splines, and most notably B-Splines, see extensive use in the functional data literature. One reason for this is that for any K basis functions the system that

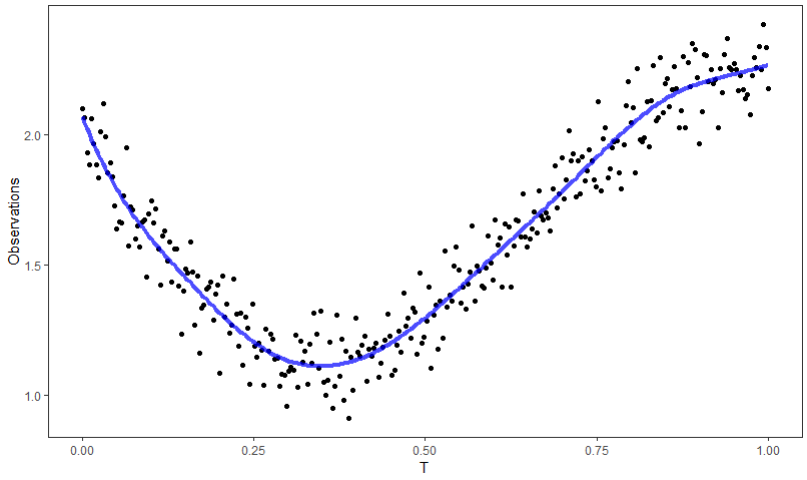


Figure 2.4.2: This figure shows an example time series that has been smoothed with 10 cubic b-spline basis functions. The spline fit is given by the solid blue line and has been obtained using equation (2.4.2).

provides the minimum mean square error when fit to a sample of data is a set of cubic splines (DeBoor, 2000). Furthermore, B-Splines have a set of properties that make them convenient to use computationally, for example closed forms for the basis functions, $\{b_k(t)\}_{k=1}^{K}$, and their derivatives. Further details on smoothing splines can be found in (DeBoor, 2000).

A well known challenge when smoothing point data is the tradeoff between bias and variance (Wasserman, 2006). An example is shown in Figure 2.4.3 highlighting both underfitting (high bias), see Figure 2.4.3(a); and overfitting (high variance), see Figure 2.4.3(b); of a set of basis functions. In order to balance the bias and variance, Kokoszka and Reimherr (2017) suggest the use of penalised basis function smoothing. This is where the c_{ik} in equation (2.4.1) are chosen as the minimisers of

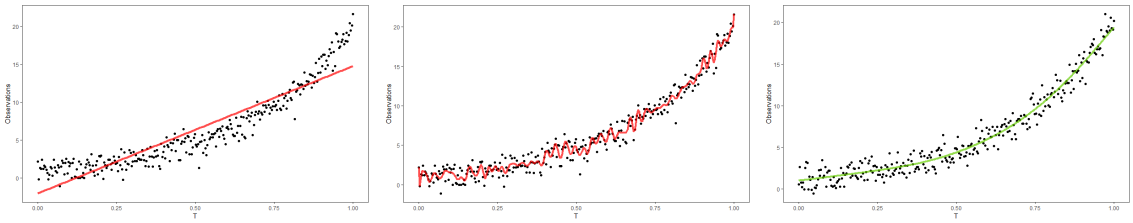
$$\hat{\boldsymbol{c_i}} = \arg\min_{\boldsymbol{c_i}} \sum_{t=1}^{T} \left[||x_{it} - \hat{X}_i(t)||_2^2 + \lambda \int_{\mathcal{T}} D^k \left(\hat{X}_i(t) \right) \, \mathrm{d}t \right], \tag{2.4.3}$$

where λ is the smoothing parameter and D^k represents the k^{th} derivative. Observe how in equation (2.4.3) as $\lambda \to 0$ we revert to unpenalised smoothing, and a high variance fit and low bias; whereas when $\lambda \to \infty$ we instead obtain low variance and a high bias. Ramsay (2005) propose leave-one-out-cross-validation to select the

choice of λ, however plug-in-estimators have also been proposed by Yoshida (2014). In Figure 2.4.3(c) we show an example of spline fitting that has used equation (2.4.3) in conjunction with cross-validation to select the value of λ.

Ensuring a balance between the bias and variance of the fit to the data is not the only advantage of using the penalised smoothing method described in equation (2.4.3). The second advantage is that it transforms the problem of fitting a set of K basis functions using least squares, as in equation (2.4.2), to a problem of selecting the optimal value of λ (Ramsay, 2005). This means that the fit obtained using equation (2.4.3) can be very similar for a wide range of basis functions, and different choices of K. The benefit of this is that it removes the problem of selecting the optimal basis system, and choice of K, for a set of data.

Now that the foundations of FDA have been introduced, we are in a position to present two important tools in the area. The first, Functional Principal Component Analysis, forms part of several existing anomaly detection methods. The second, Principal Differential Analysis, lies at the heart of the FAST and mvFAST methods developed in Chapters 4 and 5 of this thesis.



(a) (b) (c)

Figure 2.4.3: This figure presents three different instances of smoothing a univariate time series. In (a) we see high bias in the fit, in (b) we see high variance in the fit, and in (c) we see an appropriate fit. This demonstrates the utility of selecting the smoothing parameter in equation (2.4.3) using cross-validation.

2.4.2 Functional Principal Component Analysis

Functional Principal Component Analysis (FPCA), first proposed by Dauxois et al. (1982), is a functional extension of classical PCA (Pearson, 1901). FPCA uses a finite orthogonal basis system to represent functional observations, and this representation captures the dominant modes of variation within the data. A number of functional data techniques utilise FPCA in order to make computations tractable, or allow for existing statistical techniques to be adapted for the functional data analysis domain.

To compute the principal components for a sample of functional observations the covariance operator for the data, under the assumption that the covariance is stationary, is decomposed into a summation of its eigenelements. A covariance operator for functional data can be defined as

$$K(t,s) = \int_{\mathcal{T}} \int_{\mathcal{T}} \left(X_i(t) - \mu(t)\right) \left(X_i(s) - \mu(s)\right) \mathrm{d}t \, \mathrm{d}s,$$

where $\mu(t) = \mathrm{E}\left(X(t)\right)$ is the mean function for the data.

The covariance operator can be decomposed using Mercer's Theorem (Mercer and Forsyth, 1909) into the form

$$K(t,s) = \sum_{j=1}^{\infty} \lambda_j \phi_j(t) \phi_j(s).$$

Here λ_j and $\phi_j(t)$ are the eigenvalues and eigenfunctions of the covariance operator respectively. It is these eigenelements that are used to evaluate the functional principal component (FPC) scores, given by

$$\eta_{ij} = \int_{\mathcal{T}} \left(X_i(t) - \mu(t)\right) \phi_j(t) \, \mathrm{d}t. \tag{2.4.4}$$

These scores can be used to provide a Karhunen-Loeve expansion of the observations:

$$X_i(t) = \mu(t) + \sum_{j=1}^{\infty} \eta_{ij} \phi_j(t). \tag{2.4.5}$$

Example eigenfunctions and FPC scores for the data in Figure 2.4.1(a) are depicted in Figure 2.4.4. We remark that the FPC scores depicted in Figure 2.4.4(b) are not functional observations. The ability to represent functional data as multivariate non-functional observations has been exploited for various purposes including clustering

(Luz López García et al., 2015), forecasting (Foutz and Jank, 2010), and anomaly detection (Hyndman and Shang, 2010).

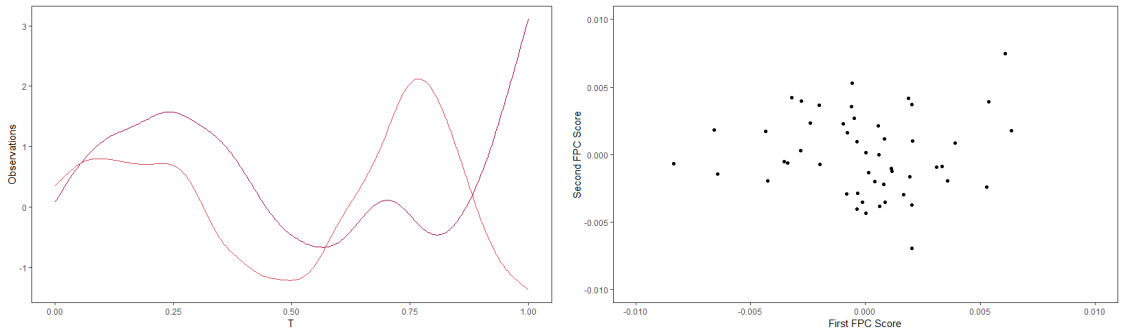
In practice the covariance operator, $K(s,t)$, and its eigensystem will need to be estimated from the data. The covariance operator can be estimated from a sample of data as follows:

$$\hat{K}(s,t) = \frac{1}{n}\sum_{i=1}^{n}\left(X_i(t) - \hat{\mu}(t)\right)\left(X_i(s) - \hat{\mu}(s)\right)$$
$$\hat{\mu}(t) = \frac{1}{n}\sum_{i=1}^{n} X_i(t).$$

The eigenfunctions, and eigenvalues, can then be estimated from $\hat{K}(s,t)$. To perform the estimation, one stated approach first proposed by Rao (1987) first discretises $\hat{K}(s,t)$ into a matrix of $T \times T$ values and then uses standard numerical linear algebra techniques to estimate the eigenelements, see e.g. Trefethen and Bau (1997) for details. Further computational efficiencies can be achieved by exploiting the basis function representation for the observations Ramsay (2005), or by first smoothing the estimate of the covariance operator Dubin and Müller (2005).

Once the eigenfunctions, $\hat{\phi}_j(t)$, and eigenvalues, $\hat{\lambda}_j$, have been estimated they can then be used in place of their true values in equations (2.4.4) and (2.4.5) to perform FPCA. The properties of these estimates is well studied, with Hall and Hosseini-Nasab (2006) demonstrating that they can be consistently estimated using the discretisation process described above subject to mild assumptions on the data.

The Karhunen-Loeve expansion for an observation that is given in equation (2.4.5) forms the foundation for one of the main benefits of using FPCA. Indeed, a d-dimensional representation for an observation can be obtained by truncating the sum in equation (2.4.5) to d components. Such a representation can be considered a projection of an observation into the d-dimensional subspace spanned by the first d eigenfunctions for the data. The benefit of this representation is that, for any choice of d, there exists no other basis of d functions that captures a greater amount of variation than the FPCA eigenbasis. As such, FPCA provides the most efficient d-dimensional orthogonal basis representation of a sample of curves possible (Kokoszka



(a) (b)

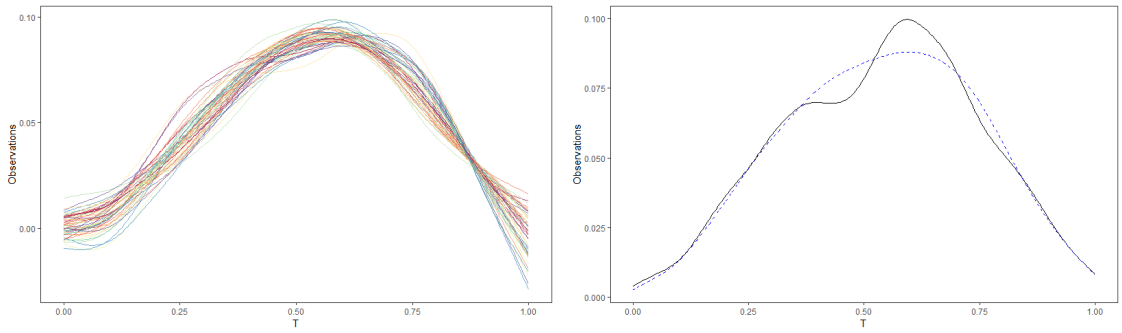
Figure 2.4.4: In this figure we present examples of eigenfunctions and FPC scores computed using the data in Figure 2.4.1(a). In figure (a) we show the first two eigenfunctions, and in figure (b) we plot the first two FPC scores for each observation, $\{(\eta_{i1}, \eta_{i2})\}_{i=1}^{n}$, against each other.

and Reimherr, 2017). As we shall see in Section 2.5.1, this low dimensional representation for a sample of functional data has been used in several functional anomaly detection methods.

In Figure 2.4.5 we present a reconstruction of the functional data from Figure 2.4.1(a) using $d = 2$ components. Observe how the projected data captures the overall shape of the observations, but arguably misses some of the smaller fluctuations particular to each curve.

One practical question when using FPCA is how to choose the number of principal components, d, to use. Ramsay (2005) advise that d should be chosen so that at least 85% of the variation is explained by the representation. A drawback of this, however, is that features of interest may only appear in a small number of curves, and so may only be contained in the higher components. As a result, Aue et al. (2018) argue that the low dimensional representation obtained using FPCA may omit key aspects of the data.

Two additional criticisms can be levelled at FPCA. The first is that it fails to take into account the temporal dependence between the functional observations, although the dynamic FPCA approach of Hörmann et al. (2015) has addressed this to an extent. The second criticism of FPCA is that it requires a stationary covariance operator,



(a) (b)

Figure 2.4.5: In this figure we use FPCA with $d = 2$ to provide a low dimensional representation of the data in Figure 2.4.1(a). In figure (a) we show the projection of the full data set, whereas in figure (b) we plot a true observation (solid black line) against its projection (dashed blue line).

which may be an unrealistic assumption to make in some settings. Recently, however, Elayouty et al. (2022) have shown how estimates for the FPCs can be allowed to vary over time to try to offset this. Their method requires the FPCs to be recomputed as each new observation is received, and so this approach may be too computationally expensive to use in a streaming setting. As an alternative to FPCA in the next section we present an approach for obtaining a finite dimensional form for the data that does not require any assumptions to be made regarding the covariance operator.

2.4.3 Principal Differential Analysis

An alternative model for functional data views the observations as a set of noisy realisations of the solution to an underlying order m linear ordinary differential equation (ODE) (Ramsay, 1996). Principal Differential Analysis can be used to estimate the coefficient for the ODE, and this ODE can be solved to obtain m linearly independent solutions for the differential equation. These solutions admit a low dimensional representation of the data, without the requirements of orthogonality of the basis functions, or stationarity of the covariance operator, required by FPCA.

The linear differential equation of order m to be fitted to the data takes the form

$$\mathcal{L} = \beta_0(t)\mathrm{D}^0 + \beta_1(t)\mathrm{D} + \ldots + \beta_{m-1}(t)\mathrm{D}^{m-1} + \mathrm{D}^m, \tag{2.4.6}$$

where D^k denotes the k^{th} derivative of a function, and the $\{\beta_j(t)\}_{j=0}^{m-1}$ are coefficient functions that are estimated using PDA. The coefficients are estimated by minimising $\sum_{i=1}^{n} ||\mathcal{L}\left(X_i(t)\right)||_2^2$, and Ramsay (1996) propose performing this minimisation on a discretised grid of T points. Using this approach the coefficients are estimated by

$$\boldsymbol{\beta}(t) = \left(\boldsymbol{Y}(t)^T\boldsymbol{Y}(t) + \lambda\boldsymbol{I}_m\right)\boldsymbol{Y}(t)^T\boldsymbol{z}(t).$$

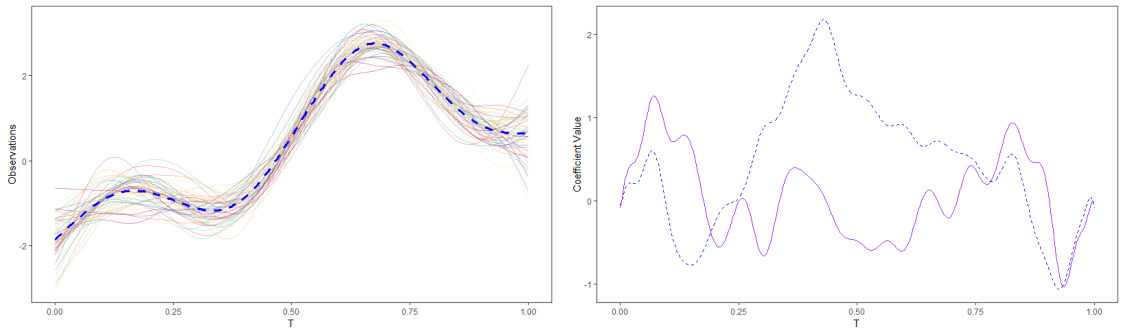
Here $\boldsymbol{\beta}(t) = \{\beta_j(t)\}_{j=0}^{m-1}$, $\boldsymbol{Y}(t)$ is an $n \times m$ matrix with entries $Y_{ij}(t) = \mathrm{D}^{j-1}(X_i(t))$, $\boldsymbol{z}(t) = \{D^m(X_i)(t)\}_{i=1}^{n}$, $\boldsymbol{I}_m$ is the $m \times m$ identity matrix, and λ is a parameter that controls the smoothness of the estimated coefficient functions. Ramsay (2005) remarks that this estimation can instead by performed using the a basis function representation for the coefficients in order to make the computation tractable. If K basis functions are used to represent the coefficients, then the computational cost of PDA is reduced from $\mathcal{O}(mT^3)$ to $\mathcal{O}(mK^3)$.

In Figure 2.4.6 we present an example in which an order two linear differential operator is fit to a set of data. We show both the solution to the estimated ODE in Figure 2.4.6(a), and the two estimated the coefficient functions in Figure 2.4.6(b). Notice how the fitted solution captures the overall shape of the observed data in a single curve.

Given that an order m ODE has m linearly independent solutions, denoted by $\{u_j(t)\}_{j=1}^{m}$, a functional observation can be represented as follows

$$X_i(t) = \sum_{j=1}^{m} c_j u_j(t) + f_i(t). \tag{2.4.7}$$

Here $f_i(t)$ is a residual function. Note that, as in FPCA, in general this residual function will not be mean zero but will decrease in size as the value of m increases. The contrast with FPCA, however, is that there is no theoretical guarantee that the expansion $\sum_{j=1}^{m} c_j u_j(t)$ will capture the optimal portion of variation for a basis system of size m. That said, a benefit of the model in equation (2.4.7) is that the $\{u_j(t)\}$ are



(a) (b)

Figure 2.4.6: In figure (a) we see an example set of functional observations and the solution (dashed blue line) to the linear differential operator fit to the data using PDA with $m = 2$. In figure (b) we provide the estimated coefficient functions.

allowed to be linearly independent, a weaker condition than the orthogonality imposed by FPCA. Furthermore, there is no requirement that the covariance operator for the data is stationary. This therefore allows for a larger range of finite dimensional models to be captured by PDA.

Several extensions to traditional PDA have been proposed, including extending PDA so that the ODE parameters depend on a set of covariates (Jin et al., 2013; Staniswalis et al., 2017). PDA itself has been applied in several settings, including the analysis of auction pricing, (Wang et al., 2008); Kinematics, (Reimer and Rudzicz, 2010); speech analysis, (Sattar and Rudzicz, 2016); and microbiology, (Chung et al., 2017).

One aspect of PDA that has received little attention to date, however, is the selection of the value of m. Ramsay (1996) suggest that it should be known beforehand, and Ramsay (2005) only discuss how to compare the operator in equation (2.4.6) with the operator given by $\tilde{\mathcal{L}} = D^m$ for a value of m that has been chosen beforehand. The lack of methods for selecting m is a key drawback of PDA. That said, choosing the ODE order to minimise the size of the residuals between the solution to the fitted ODE and the data is an approach discussed in both Hooker and Ellner (2015) and Ramsay and Hooker (2017) for ODE fitting methods unrelated to PDA. In Chapters 4 and 5 we will build upon this notion of using the residuals and propose methods for

the selection of the ODE order.

Now that the necessary foundations for FDA have been introduced, and two key techniques discussed, we conclude our review of background literature by providing an overview of recent contributions to anomaly detection in functional data.

2.5 Anomaly Detection in Functional Data

Functional anomaly detection is an area of functional data analysis concerned with identifying anomalous observations within a sample of functional data. Broadly speaking, two primary forms of anomaly are discussed in the literature: (i) magnitude anomalies, where the values of the curve differ from those in the sample by a substantial margin; and (ii) shape anomalies, where the appearance of the curve differs from the other shapes within the sample (Dai et al., 2020). A specific form of shape anomaly, phase anomalies, are considered by Tucker et al. (2013). This is where the curve is a similar shape but has been shifted in time, and detection of these anomalies is discussed in Harris et al. (2020).

Figure 2.5.1 displays examples of both shape, and magnitude, anomalies. Observe how the majority of curves in the sample take similar values. On the other hand, the magnitude anomaly (dashed orange line) takes greater values than the other curves in the sample, and the shape anomaly (solid red line) follows a different shape. Finally, the figure shows how an anomaly can be both a shape and magnitude anomaly, as depicted by the dotted black line.

In this section we will first explore the detection of anomalies in univariate functional data, a well studied area of research, before then outlining the detection of multivariate functional anomalies. This is an area that has received less attention, in part due to the high computational cost of working with multivariate functional data.

2.5.1 Univariate Functional Anomaly Detection Using FPCA

A wide range of different approaches for detecting anomalies in univariate functional data have been proposed and in this section we will outline some of the most im-

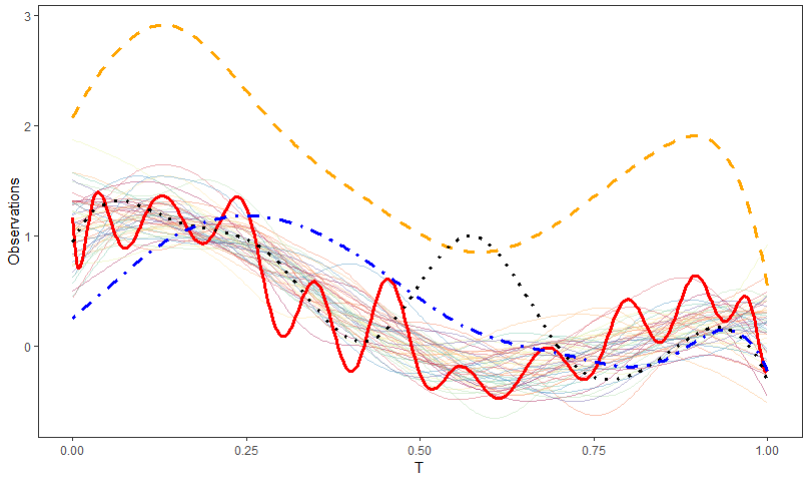


Figure 2.5.1: Figure showing example functional anomalies. The solid red line is a shape anomaly, the dashed orange line is a magnitude anomaly, and the dot and dash blue line is a phase anomaly. Finally, the dotted black line can be considered both a shape and magnitude anomaly, showing that the two are not mutually exclusive.

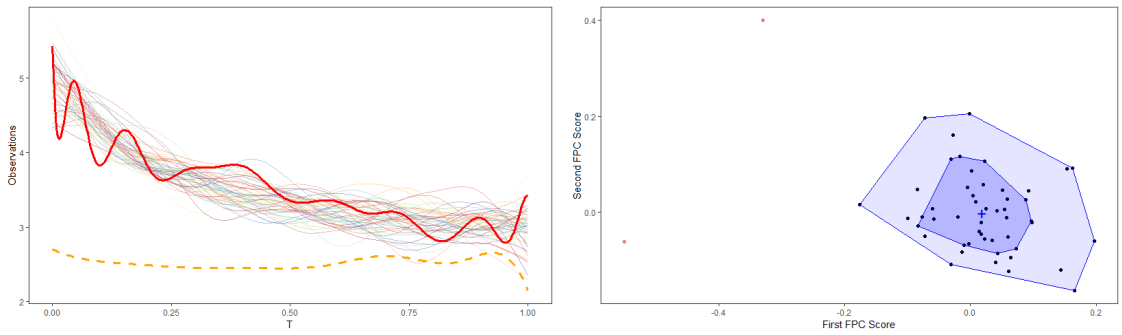
portant. To begin with we will present anomaly detection methods that rely upon FPCA, before then moving onto other methods such as those that build upon classical anomaly detection tools from multivariate statistics such as depth measures.

The motivation for using FPCA for anomaly detection is that the projection of each observation into the subspace spanned by the principal components can be said to "borrow information from across the curves" Wang et al. (2016). This means that observations with similar behaviours will have similar projections, and also similar FPC scores. Each of the methods discussed in this section are therefore seeking to identify when a projected observation is dissimilar to the majority of the projections in the sample. Such dissimilarity suggests that the observation is an anomaly because it has features that are not present within the majority of the observed data.

A leading approach for FPCA based functional anomaly detection, relying upon the FPC scores, is proposed by Hyndman and Shang (2010). Here two approaches are considered, both of which are founded upon taking the sample of principal component scores from equation (2.4.4) and applying multivariate outlier detection methods to them. In the first method, the Tukey depth (Tukey, 1975) of the first two scores is

calculated, and a bagplot (Rousseeuw et al., 1999) is used to identify anomalies. In the second method, kernel density estimation is applied to the first two scores, and then a highest density region (Hyndman, 1996) is computed for a desired level of probability coverage (eg, 99%). Points lying outside the highest density region are then identified as outliers.

An example using a bagplot in conjunction with the first two FPC scores is presented in Figure 2.5.2. The bagplot has been created using the R package applpack (Wolf, 2019). In the figure the inner region of the bag contains the median (plus symbol) and the most central 50% of the data, whereas the outer shaded region is the convex hull of observations contained within the fence (by convention the fence is not plotted). The fence is generated by inflating the inner bag by a factor of 3. Points outside the fence are identified as anomalies



(a) (b)

Figure 2.5.2: Figure (a) shows a sample of functional observations with a shape anomaly (solid red line) and a magnitude anomaly (orange dashed line). In figure (b) we follow the approach of Hyndman and Shang (2010) and present a bagplot of the first two FPC scores. The plus symbol is the median for the data, and the red points outside the fence have been identified as anomalies.

More recently, several other authors have considered using FPCA in conjunction with methods from multivariate statistics to detect anomalies. For example, Sawant et al. (2012) use the Mahalanobis distance on the FPC scores, Barreyre et al. (2020) apply a two-sample hypothesis test to two sets of functional data to determine if one set of FPC's differ from another, and Jarry et al. (2020) apply a hierarchical

clustering method to the FPC scores. Each of these methods, however, suffer from the usual drawbacks of FPCA insofar as they require stationary covariance between the observations, and assume the data are independent.

Combining FPCA with non-functional anomaly detection approaches from multivariate statistics is not the only way FPCA has been used to detect anomalies. Indeed both Hyndman and Shahid Ullah (2007) and Vilar et al. (2016) compare the observations with their low dimensional projections given by equation (2.4.5). Specifically they use the integrated square error, e_i, of the projections:

$$e_i = ||X_i(t) - \hat{X}_i(t)||_2^2. \tag{2.5.1}$$

Curves with a large integrated square error are then labelled as anomalies. One advantage of such an approach is that it uses the full information contained in the observations, rather than detecting anomalies based upon the projected data. Although we do not utilise FPCA, in Chapter 6 we shall briefly describe a method that makes use of integrated square error to detect anomalies.

Whilst FPCA can be seen as one of the key tools in the functional anomaly detection literature, many other approaches use measures of depth for functional data instead. We review these in Section 2.5.2. Although depth measures were originally used for outlier detection in multivariate statistics (Liu et al., 1999), the methods presented here use depth measures that have been designed for a functional data setting.

2.5.2 Univariate Functional Anomaly Detection Using Depth Measures

An early example of the use of depth measures is provided by Lopez-Pintado and Romo (2007), who propose a depth measure known as band depth to detect anomalies. The key idea here is that a band is delimited by two curves, $Y_1(t)$ and $Y_2(t)$, and a curve is contained within these bands if $Y_1(t) \leq X_i(t) \leq Y_2(t)$ for all $t \in \mathcal{T}$. The band depth of an observation can be calculated by assessing how many of the different bands generated using the sample $\{X_1(t), \ldots, X_n(t)\}$ the observation lies between.

The upper and lower bands for the sample are then set such that for any t a set proportion of the curves lie in the region between them. If a curve does not lie entirely inside the region defined by these bands, then it can be declared an anomaly.

In Figure 2.5.3 we illustrate an example of anomaly detection using band depth. The upper and lower bands have been calculated such that the area between the bands contains 99% of the data. We see that a magnitude anomaly (dashed orange line) lies entirely outside the two bands and so has been identified as anomalous.

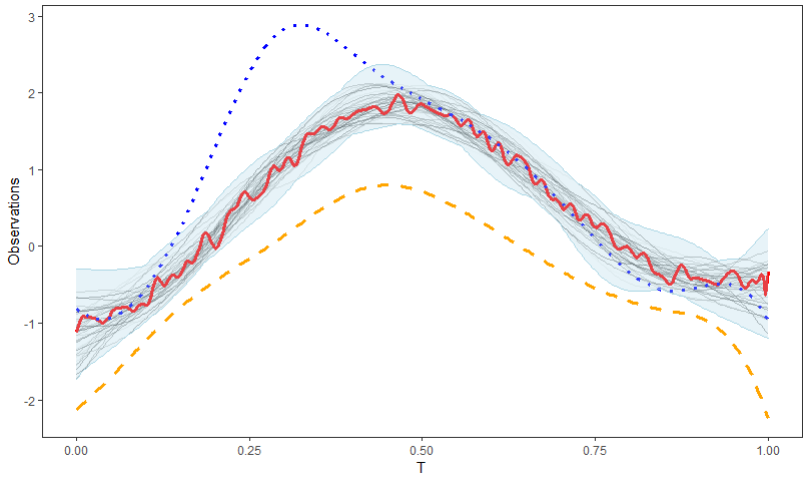


Figure 2.5.3: In this figure we illustrate how band depth (Lopez-Pintado and Romo, 2007) can be used for anomaly detection. The light blue shaded region denotes the area between the upper and lower bands, with observations entirely outside this region (eg, the dashed orange line) being declared as anomalous.

Unfortunately a key criticism of using the band depth of Lopez-Pintado and Romo (2007) for anomaly detection is that it does not consider the proportion of the time the observation spends outside the band. This is addressed by the modified band depth (MBD) approach of López-Pintado and Romo (2009), where the fraction of the interval where an observation is contained within the bands is considered; and also by the modified epigraph index (MEI) approach presented in Pintado and Romo (2011). The latter identifies the proportion of time a curve lies below the other curves in the sample. Should a curve have a large value of MBD or MEI, then this would indicate that it is an anomaly. In Figure 2.5.3, for example, both the

MBD and MEI are able to detect the dotted blue shape anomaly as well as the dashed orange magnitude anomaly. The disadvantage of these band depth methods, however, is that a shape anomaly may lie within the centre of the data and thus remain undetected (see, for example, the solid red line depicting a shape anomaly in Figure 2.5.3). As a consequence, these band-based approaches are best used to detect magnitude anomalies.

In multivariate statistics depth measures are combined with anomaly detection methods such as the boxplot or bagplot to detect anomalies (Liu et al., 1999). A similar approach has been taken in the functional anomaly detection literature when using depth measures, with several visual anomaly detection tools being proposed. Two of the most important approaches are the functional boxplot (Sun and Genton, 2011), which extends the concept of a boxplot to functional data (see Figure 2.5.4); and the outliergram (Arribas-Gil and Romo, 2014), which provides a way of combining the MBD and MEI depth measures to improve anomaly detection performance.

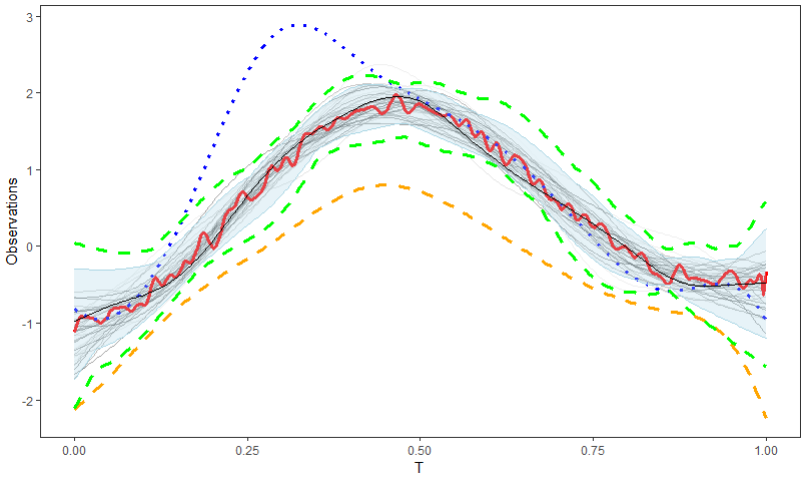


Figure 2.5.4: This figure presents an example of a functional boxplot. The black curve is the median, the shaded light blue region is the box, and the dashed green lines are the whiskers. Observations lying outside the whiskers are identified as anomalies. In this example this corresponds to the dashed orange, and the dotted blue, curves.

A range of other depth based approaches have been proposed in the literature. Narisetty and Nair (2016), for instance, propose the extremal band depth measure in

a bid to add robustness to the existing depth methods. Huang and Sun (2019), on the other hand, present the total variation depth measure. Unlike the measures discussed so far, the total variation depth performs well when detecting both magnitude and shape anomalies.

A challenge when detecting functional anomalies using depth measures is the selection of the depth measure to use. In principle, the choice of depth measure can be motivated by the anomalies that one wishes to detect. Unfortunately, this presupposes some knowledge of the anomalous structures present within the data. Such knowledge will not always be available. Dai and Genton (2018a) propose to address this challenge through the use of Directional Outlyingness. As we shall see in the next section, this quantity can detect both shape and magnitude anomalies and is compatible with a wide range of depth measures.

2.5.3 Univariate Functional Anomaly Detection Using Directional Outlyingness

In a bid to unify the range of existing depth measures for functional anomaly detection, Dai and Genton (2018a) propose the concept of Directional Outlyingness. This measure is compatible with any depth measure and can be used to detect both shape and magnitude anomalies. Directional outlyingness is defined by Dai and Genton (2018a) as

$$O(X_i(t)) = \left[\frac{1}{D(X_i(t))} - 1\right] \cdot v(t).$$

Here $D(\cdot)$ is a depth measure and $v(t)$ is defined as $\frac{X_i(t)-m(t)}{||X_i(t)-m(t)||_2}$, where $m(t)$ is the median of the sample of curves with respect to the chosen depth measure.

To use directional outlyingness to detect anomalies two different quantities must be computed. The first, MO, is the mean outlyingness and is designed to detect magnitude anomalies; and the second, VO, is the variational outlyingness and seeks

to identify shape anomalies. The MO and VO are defined as

$$MO(X_i(t)) = \int_{\mathcal{T}} O(X_i(t))w(t)\,\mathrm{d}t$$
$$VO(X_i(t)) = \int_{\mathcal{T}} ||O(X_i(t)) - MO(X_i(t))||_2^2 w(t)\,\mathrm{d}t,$$

where $w(t)$ is a weight function. In order to detect anomalies using these quantities Dai and Genton (2018a) propose plotting the MO and VO against each other. They term this the Magnitude-Shape (MS) plot and show how critical regions for the plot can be calculated and used for anomaly detection. Alternatively, Dai and Genton (2018b) discuss how directional outlyingness can be combined with the functional boxplot of Sun and Genton (2011).

In Figure 2.5.5 we present an example MS plot using the same data as in Figure 2.5.3. The depth measure used to calculate the directional outlyingness value is the random projection depth, as suggested in Dai et al. (2020), and it has been computed using the R package fdaoutlier (Ojo et al., 2021b). The threshold for anomaly detection is given by the dashed black ellipsis and has been computed using the approach laid out in Dai and Genton (2018b). As can be seen from Figure 2.5.5, unlike in the band depth (Figure 2.5.3) and functional boxplot (Figure 2.5.4) examples all three of the anomalies, including the shape anomaly contained in the centre of the observed curves (see Figure 2.5.3), have been detected.

Whilst directional outlyingness allows any depth measure to be incorporated into a single anomaly detection method, one issue with it is that the choice of both the depth measure, $D(\cdot)$; and weight function, $w(t)$; can lead to different observations being identified as anomalies. As such, the anomaly detection results can be difficult to interpret in practice.

As with the FPCA based methods, both depth based and directional outlyingness based approaches suffer from the issue that they assume the functional observations are independent. Outside of the anomaly detection literature, however, there are many examples of dependency structures in functional data. These include functional autoregressive processes, and other time series models, or sequences of curves that are strongly mixing (Hörmann et al., 2010). Whilst some of the depth based mea-

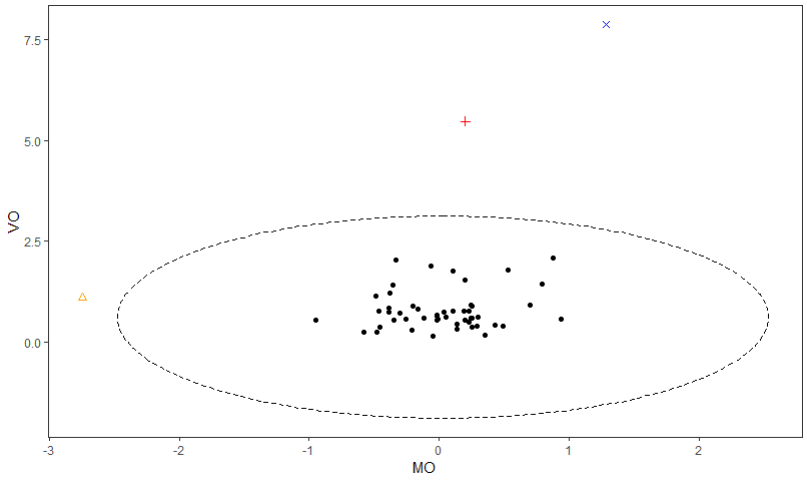


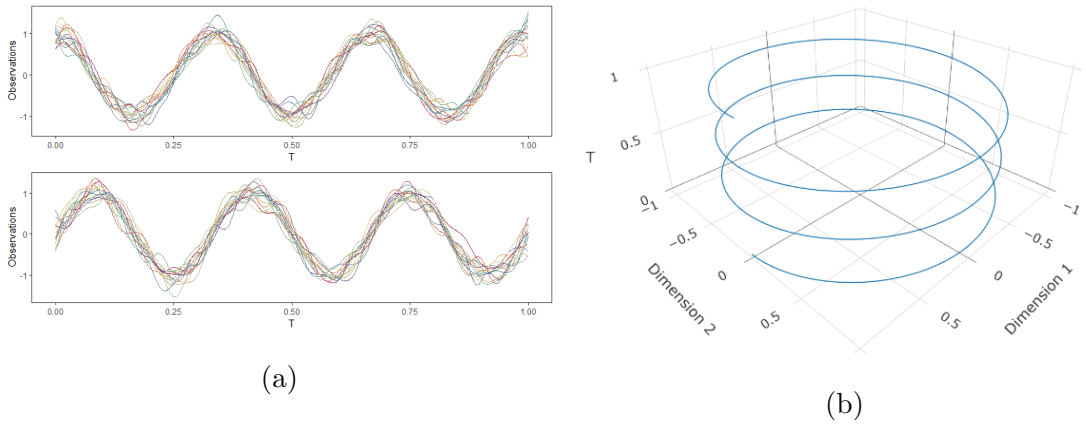
Figure 2.5.5: In this figure we present an MS plot for the same data as depicted in Figures 2.5.3. The dashed black ellipsis depicts the threshold for anomaly detection, with the three observations outside the elliptical region being identified as anomalies.

sures may allow for the assumption of independence to be relaxed, their results can be confounded by dependence structures that many functional time series possess. The FPCA methods, on the other hand, rely on the decomposition of the covariance operator for the data. When dependence between the observations is present, this covariance operator fails to model the relationships between the functional observations and this leads to incorrect anomaly detection results Hörmann et al. (2010).

Whilst the lion's share of functional anomaly detection tools rely upon either FPCA or depth measures, a small number of entirely different approaches are discussed in the literature. Both Vinue and Epifanio (2020) and Ojo et al. (2021a), for example, propose anomaly detection approaches that apply non-FPCA dimension reduction tools. Harris et al. (2020), on the other hand, apply a range of transformations to the observations in order to highlight anomalous shape structures within the data. None of these aforementioned approaches make use of tools relevant to the techniques presented in this thesis, however, and so we do not discuss them in more detail and instead direct the interested reader to these works.

Both the depth based approaches, and to a lesser extent the FPCA based techniques, have also been extended into the multivariate functional anomaly detection

domain. Although the literature in this area is sparser than in the univariate case, we present a review of it in the next section.



(a) (b)

Figure 2.5.6: In (a) we present an example of two-dimensional functional data with each dimension shown separately. In (b) we present a three-dimensional plot of the mean function for the bivariate functional data with $\mathcal{T}$ represented on the z-axis.

2.5.4 Multivariate Functional Anomaly Detection

Multivariate functional data are denoted by $\{X_{ij}(t)\}$ for $1 \leq i \leq n$ and $1 \leq j \leq p$, where p is the number of dimensions of the observation. An example of 50 multivariate functional observations of dimension $p = 2$ is shown in Figure 2.5.6. In this section we provide a brief overview of some important multivariate functional anomaly detection techniques. For a wider survey of the field of multivariate FDA, Górecki et al. (2018) can be consulted.

As in the univariate setting, depth measures form a cornerstone of the anomaly detection literature for multivariate functional data. Early attempts to extend depth measures to the multivariate domain resulted in approaches that first compute the depth of an observation in each of the p dimensions, and then either aggregate these depths (Ieva and Paganoni, 2013), or use multivariate non-functional anomaly detection tools such as bagplots on the p-dimensional depth values (Hubert et al., 2015). Rather than compute the depth for each dimension separately, the directional outly-

ingness measure (Dai and Genton, 2018a) can be used to consider all the dimensions of an observation together. Dai and Genton (2019), and Qu and Genton (2022) demonstrate how this can be combined with a univariate tools such as the MS-plot and functional boxplot in order to detect anomalies. More recently, Mozaffari et al. (2021) demonstrate how the functional boxplot can be used to identify anomalies within functional data representing two and three dimensional geometric shapes.

The key issue with extending depth based methods into the multivariate setting, however, is the computational cost of the method as the number of dimensions increase. Several authors have considered this issue, including the works of Wang and Yao (2015) and Ojo et al. (2022). Both these methods seek to address the problem of computational cost by first reducing the dimensionality of the observed data. As an alternative to the use of dimension reduction techniques, Lejeune et al. (2020a) compute various descriptive quantities for each dimension of the multivariate functional observations. Examples of these quantities include curvature, and arc length. Once the descriptors have been computed, they are aggregated across the dimensions before existing non-functional anomaly detection tools are used to identify outlying observations. A drawback of this, however, is noted by Lejeune et al. (2020b) who observe that the approach is best used for detecting shape anomalies only.

Chapter 3

An Online Non-Parametric Method for Detecting Changes in Distribution using NUNC

3.1 Introduction

The challenge of sequential nonparametric changepoint detection has seen significant development in recent years. See, for example, Tartakovsky et al. (2014) for an excellent introduction to the area. Contributions include the work of Gordon and Pollak (1994), Ross and Adams (2012), and Padilla et al. (2019a) who introduce novel approaches to detect changes in an unknown distribution. Others, including Chakraborti and van de Wiel (2008); Hawkins and Deng (2010); Murakami and Matsuki (2010); Ross et al. (2011); Mukherjee and Chakraborti (2012); Liu et al. (2013); Wang et al. (2017) and Coelho et al. (2017) seek to address a different non-parametric challenge: the sequential detection of changes in the mean, scale, or the location of the data. Such methods have also found application in a range of fields including monitoring financial systems (Pepelyshev and Polunchenko, 2017), monitoring viral intrusion in computer networks (Tartakovsky et al., 2005), detecting changes in social networks (Chen, 2019), genome sequencing (Siegmund, 2013), and radiological data (Padilla et al., 2019a).

Our work is motivated by a different challenge, increasingly encountered within many contemporary digital settings, such as those found in the telecommunications sector. In such environments it is important to perform device-side analyses on units with limited computational power and data storage capability. Existing methods, such as those mentioned above, are unsuitable for use in such cases as they require the entire data stream to be stored and analysed *a posteriori*, whereas our memory-constrained setting makes this impossible. As a consequence, an online approach that uses a lighter data footprint is required.

One example of such data, encountered by an industrial collaborator, can be seen in Figure 3.1.1. Here we display two sample data sets of a key operational metric that are routinely monitored to identify problems with networking devices. Figure 3.1.1(a) displays data from a healthy device, whereas the data in Figure 3.1.1(c) contains an event that triggers a user intervention. Note in particular how the structure of the data changes in the region when the event occurs. This can perhaps be more clearly seen in Figures 3.1.1(b,d). The ideal, therefore, is to be able to (i) identify the start of changing structure in advance of the user being required to start an intervention – we call the correct detection of such an event an 'anticipation'; (ii) using an approach that does not necessarily require the same underlying distribution pre- and post-change and (iii) can still permit (more subtle) non-anomalous changes in structure that occur over time due to typical operational issues (e.g. electrical interference, line optimisation, etc).

Many existing methods, such as those mentioned above, are unsuitable for use in this setting as they typically require the entire data stream to be stored. Unfortunately in our problem setting, such memory constraints are no longer possible. To overcome this one might, for example, consider adopting an online non-parametric changepoint detection approach using a sliding window, such as in the MOSUM test (Chu et al., 1995; Eichinger and Kirch, 2018; Kirch and Weber, 2018; Meier et al., 2021). Alternatively a control chart based approach, e.g. Ross and Adams (2012), might be considered. Unfortunately, as described above, our telecoms operational metric can exhibit non-anomalous shifts in mean and variance. Such structure, while

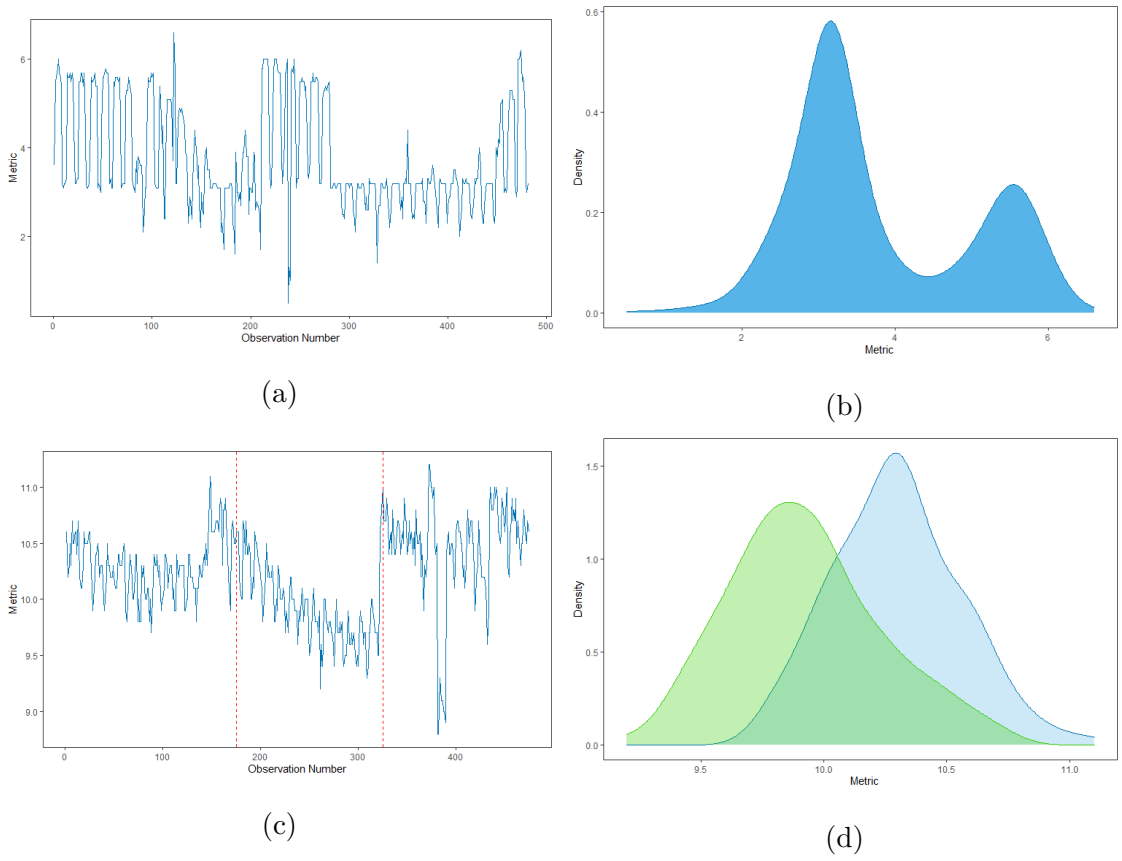


Figure 3.1.1: Example of telecoms operational data: (a) a series without an event, and (c) a series with an event taking place between the two red lines. The corresponding kernel density estimates for the time series in figures (a) and (c) are presented in (b) and (d) respectively. Note that in (d) the green represents the kernel density estimate of the series when the event is taking place.

operationally acceptable, may cause a control chart to return excessive false alarms due to the cumulative nature of the test statistic. Further, existing nonparametric sequential changepoint detection methods prove unsuitable as they do not expect the null distribution to change over time. While there are some methods that attempt to model and account for non-anomalous drift (Gama et al., 2014), these can be challenging to tune when the form of drift is unknown. We thus take a different approach and use a sliding window to deal with drift. To this end, we introduce a new windowed, non-parametric procedure to detect sequential changes in an online setting. Taken from the Latin, *nunc* ('now'), our approach provides a **N**on-Parametric **UN**bounded

Changepoint (NUNC) detection.

We propose two variants of NUNC: NUNC Local, and NUNC Global. The first of these algorithms, NUNC Local, performs the detection in a sliding window, considering only the points inside this window. This allows for the implementation to work in an online setting. The second, NUNC Global, uses an efficient updating step to compare the distribution of the historic data seen with the distribution of the data inside the sliding window; if these differ significantly, a change is identified.

The rest of this chapter is organised as follows: In Section 3.2 we outline the methodology behind our new sequential tests. In particular, we detail the existing non-parametric changepoint methods our work is based upon, and in Section 3.2.1 we provide details of our two new window-based changepoint detection tests. The remainder of the section then explores the properties of the test, including the choice of quantiles and threshold. In particular, a theoretical result is presented concerning the selection of the threshold in order to control the false alarm probability, and this result can also be utilised in an offline changepoint setting. The chapter concludes by exploring the performance of NUNC Local and NUNC Global using both simulated scenarios (Section 3.3) and data arising from the previously described telecommunications setting (Section 3.4).

3.2 Background and Methodology

Our approach builds on the recent work of Zou et al. (2014) and Haynes et al. (2017), utilising a non-parametric likelihood ratio test as the basis for the proposed sequential Non-Parametric test. In so doing, the method permits a range of data distributions to be modelled, without the need for restrictive parametric assumptions. Below we introduce both NUNC approaches, and provide a discussion of their various features including computational performance and the choice of quantiles. However, prior to doing so, we review the pertinent literature on non-parametric changepoint methods.

We begin by outlining some notation. Assume that we observe a data stream of real valued independent observations $x_1, x_2, \ldots, x_t$, and that the data stream can

contain a changepoint, at some (unknown) time point τ. Further, assume that x_1 is the start of this data stream, x_t is the most recently observed point, and that for $j > i$, $x_{i:j} = x_i \ldots, x_j$ denotes a segment of the data. If a change is present in the data at τ_1, then we refer to $x_1, \ldots, x_{\tau_1}$ as the pre-change segment, drawn from a distribution $F_1(\cdot)$. Similarly, $x_{\tau_1+1}, \ldots x_{\tau_2}$ are considered to be drawn from post-change distribution, $F_2(\cdot)$, with $F_1 \neq F_2$.

Following Zou et al. (2014), let $F_{i:t}(q)$ denote the (unknown) cumulative distribution function (CDF) for the segment $x_i, \ldots, x_t$, and $\hat{F}_{1:t}(q)$ as its associated empirical CDF. I.e.

$$\hat{F}_{1:t}(q) = \frac{1}{t} \sum_{j=1}^{t} \left\{ \mathbb{I}(x_j < q) + 0.5 \times \mathbb{I}(x_j = q) \right\}. \tag{3.2.1}$$

Under the assumption that the data are independent, then the empirical CDF will follow a Binomial distribution. That is,

$$t\hat{F}_{1:t}(q) \sim \text{Binom}(t, F_{1:t}(q)). \tag{3.2.2}$$

Using the Binomial distribution, the log-likelihood of the segment $x_{\tau_1+1}, \ldots, x_{\tau_2}$ is

$$\mathcal{L}(x_{\tau_1+1:\tau_2}; q) = (\tau_2 - \tau_1) \left[\hat{F}_{\tau_1+1:\tau_2}(q) \log(\hat{F}_{\tau_1+1:\tau_2}(q)) - (1 - \hat{F}_{\tau_1+1:\tau_2}(q)) \log(1 - \hat{F}_{\tau_1+1:\tau_2}(q)) \right].$$

This log-likelihood can be used to form a likelihood ratio test statistic for the detection of a change at a single quantile of the distribution as follows:

$$\max_{1 \leq \tau \leq t} 2 \left[\mathcal{L}(x_{1:\tau}; q) + \mathcal{L}(x_{\tau+1:t}; q) - \mathcal{L}(x_{1:t}; q) \right].$$

Following Zou et al. (2014) and Haynes et al. (2017), this test statistic can be averaged over multiple quantiles, $q_1, \ldots, q_K$, in order to search for a change in distribution. The statistic for such a test can be formulated as follows:

$$C_K(x_{1:t}) = \max_{1 \leq \tau \leq t} \frac{1}{K} \sum_{k=1}^{K} 2 \left[\mathcal{L}(x_{1:\tau}; q_k) + \mathcal{L}(x_{\tau+1:t}; q_k) - \mathcal{L}(x_{1:t}; q_k) \right]. \tag{3.2.3}$$

Here K is the fixed number of quantiles to be averaged over. Haynes et al. (2017) propose that a value of K is chosen that is proportionate to $\log(t)$. The choice of quantiles at which the empirical CDF can be evaluated will be discussed later in Section 3.2.2.

Using this test, a changepoint is declared when $C_K(x_{1:t}) - \beta \geq 0$. Thus the stopping time for our test becomes

$$\max_{1 \leq \tau \leq t} \sum_{k=1}^{K} 2 \left[\mathcal{L}(x_{1:\tau}; q_k) + \mathcal{L}(x_{\tau+1:t}; q_k) - \mathcal{L}(x_{1:t}; q_k) \right] \geq K\beta, \tag{3.2.4}$$

where β is the threshold for the test.

Having outlined how the existing (offline) non-parametric tests work, and how the cost function from this work can be used to devise a stopping rule for a sequential changepoint detection test, we are now in a position to introduce our two variant nonparametric approaches.

3.2.1 Two Sequential Changepoint Detection Algorithms

We now introduce two different, yet related, approaches that can be adopted within this nonparametric framework: NUNC Local and NUNC Global. Common to both is the use of a sliding window, and the test statistic given in equation (3.2.3). Where the two approaches differ, however, is the manner in which the data observed outside the window are handled. In NUNC Local, a simplistic perspective is adopted, taking the data contained within the sliding window into account – i.e., previously seen points that fall outside this window are forgotten. The advantage of this approach is that the sequential test is immune to false alarms that might be caused, for example, by a natural drift in the underlying distribution of the data. The drawback, however, is that the empirical CDF must be estimated only from the data in the window and so any historic information is lost.

NUNC Global seeks to overcome the short-comings of NUNC Local. Specifically, NUNC Global stores the empirical CDF that has been estimated using all data observed so far, and tests whether the data from such empirical distribution differ from the data observed in the current window. In Section 3.4 we will seek to contrast the differences between these two variants. However, prior to this, we describe both search methods more carefully, whilst also describing various properties and recommendations.

NUNC Local

Our first method takes a sliding window of size W and performs the test on the data within this sliding window. In the sliding window of points we have that

$$\mathcal{Q}_t^{local} = \max_{t-W+1\leq\tau\leq t} \sum_{k=1}^{K} 2\left[\mathcal{L}(x_{t-W+1:\tau}; q_k) + \mathcal{L}(x_{\tau+1:t}; q_k) - \mathcal{L}(x_{t-W+1:t}; q_k)\right], \quad (3.2.5)$$

where K is the number of quantiles. Letting β be the test threshold, when $\mathcal{Q}_t^{local} \geq K\beta$ then the algorithm stops at time t and declares that a change has occurred at time τ. A description of NUNC Local pseudocode can be found in Algorithm 1.

The choice of the parameters for NUNC Local, including the window size, quantiles, and threshold; will be discussed in Section 3.2.2 and in simulations in Section 3.3. We remark here, however, that the choice of K and the size of the window is related to the computational cost of NUNC Local. In particular, the naive cost of computing NUNC Local directly is $\mathcal{O}(KW^2)$. However, by using an Ordered Search Tree (Cormen et al., 2001) to calculate the empirical CDF this can be reduced to $\mathcal{O}(KW \log W)$.

Algorithm 1: NUNC Local Algorithm

Data: $\{x_{t-W+1}, ..., x_{t-1}, x_t\}$, the last W realizations from a data generating process X.

Input: $\beta > 0$, $K < W$, $q_1, \ldots, q_k$ quantiles

$\mathcal{Q} \leftarrow \max_{t-W+1\leq\tau\leq t} \left[\sum_{k=1}^{K} 2\left(\mathcal{L}(x_{(t-W):\tau}; q_k) + \mathcal{L}(x_{\tau+1:t}; q_k) - \mathcal{L}(x_{(t-W):t}; q_k)\right)\right]$;

$\tau^* \leftarrow \arg\max_{t-W+1\leq\tau\leq t} \left[\sum_{k=1}^{K} 2\left(\mathcal{L}(x_{(t-W):\tau}; q_k) + \mathcal{L}(x_{\tau+1:t}; q_k) - \mathcal{L}(x_{(t-W):t}; q_k)\right)\right]$;

if $\mathcal{Q} \geq K\beta$ **then**

  Return τ^* as a changepoint

In order to reduce the computational requirements of NUNC Local, which is quadratic in window size, it is possible to instead perform the search on a subset of the points in the sliding window. In this setting, we obtain the stopping condition:

$$\max_{\tau\in B_J} \sum_{k=1}^{K} 2\left[\mathcal{L}(x_{t-W+1:\tau}; q_k) + \mathcal{L}(x_{\tau+1:t}; q_k) - \mathcal{L}(x_{t-W+1:t}; q_k)\right] \geq K\beta,$$

where $B_J \subset \{t - W + 1, \ldots, t\}$. This corresponds to changing the maximisation in Algorithm 1 to taking place over the set B_J rather than the entire window. Using a subset of size $J << W$, the computational cost of NUNC Approximate is reduced to $\mathcal{O}(JKW)$. To find a suitable subset of values to search inside the sliding window, we first note that intuitively it only makes sense to search for a change in the right hand half of the window. This is because the data in the left of the window has already been scanned for a change several times. Moreover, we can also establish the following: that there exists a point on the right hand side of the window such that a changepoint cannot be detected to the right of this point.

Proposition 1. *For any quantile q the test statistic is bounded such that*

$$\mathcal{L}(x_{t-W+1:\tau}; q) + \mathcal{L}(x_{\tau+1:t}; q) - \mathcal{L}(x_{t-W+1:t}; q) \leq -\frac{\tau}{W} \log \frac{\tau}{W} - (W - \tau) \log \left(\frac{W - \tau}{W} \right).$$

Furthermore, for fixed W, this equation is decreasing as τ increases, and so if τ^* is the point such that

$$-\frac{\tau^*}{W} \log \frac{\tau^*}{W} - (W - \tau^*) \log \left(\frac{W - \tau^*}{W} \right) \leq \frac{\beta}{2}$$

then for $\tau > \tau^*$ detection of a change is impossible.

Proof. See Appendix. □

As a consequence of Proposition 1, only a portion of the right hand side of the window needs to checked, and the value of this cutoff can be found, with the value for τ^* being calculated numerically. Further computational efficiencies can be realised for NUNC Local if it is only performed on a spaced out grid of points. This is due to the value of the test statistic being correlated at nearby points. Consequently, if segmenting the data at t does not return a change, then it is unlikely that a change will be detected at $t + 1$. As a result, we propose to use an equally spaced grid of J points starting from the centre of the window, after it has been trimmed using the value of τ^*. However, one drawback of using the grid method is that there will be a higher detection delay for smaller values of J. This is due to it taking longer for the change to reach a point that we are checking. As such, we conclude that there

is a trade-off between computational efficiency and detection delay when using the approximated algorithm.

NUNC Global

NUNC Global differs from the NUNC Local. Specifically it tests whether or not the data in the window comes from a different distribution to all the data seen so far. To store the information in a memory efficient manner, we again fix K quantiles and update the long-run empirical CDF, denoted by $z_W^{(t)}(\cdot)$, each time a point leaves the sliding window. The recursive equations for this update step are as follows:

$$\begin{aligned} z_W^{(t)}(q) &= \hat{F}_{1:(t-W)}(q), \\ z_W^{(t+1)}(q) &= \frac{1}{t-W+1}\left[(t-W)z_W^{(t)}(q) + \hat{F}_{(t-W+1):(t-W+1)}(q)\right], \quad t \geq W. \end{aligned} \tag{3.2.6}$$

I.e. the long-run empirical CDF is updated to take into account the point that will leave the sliding window at the next iteration. The Global algorithm then compares the distribution for the long-run empirical CDF to the distribution of the data in the sliding window, denoted by $\hat{F}_{t-W+1:m}(\cdot)$.

To implement this approach, we need to obtain a CDF estimate of the full data. This is given by a weighted mixture of the long-run empirical CDF and the current segment empirical CDF estimate. Assuming we are at time t, and have a sliding window of size, W, we write this as

$$\hat{F}_{\text{full}}(q) = \hat{F}_{1:t}(q) = \frac{t-W}{t} z_W^{(t)}(q) + \frac{W}{t}\hat{F}_{t-W+1:t}(q).$$

With these distributions in place, we can obtain the equivalent likelihoods, given respectively by

$$\begin{aligned} \mathcal{L}(x_{1:t-W}; q) &= (t-W)\left[z_W^{(t)}(q)\log(z_W^{(t)}(q)) - (1-z_W^{(t)}(q))\log(1-z_W^{(t)}(q))\right] \\ \mathcal{L}(x_{t-W+1:t}; q) &= W\left[\hat{F}_{t-W+1:t}(q)\log(\hat{F}_{t-W+1:t}(q)) - (1-F_{t-W+1:t}(q))\log(1-\hat{F}_{t-W+1:t}(q))\right] \\ \mathcal{L}(x_{1:t}; q) &= t\left[\hat{F}_{\text{full}}(q)\log(\hat{F}_{\text{full}}(q)) - (1-\hat{F}_{\text{full}}(q))\log(1-\hat{F}_{\text{full}}(q))\right]. \end{aligned} \tag{3.2.7}$$

The test statistic is then given by:

$$\mathcal{Q}_t^{global} = \sum_{k=1}^{K} 2\left[\mathcal{L}(x_{1:t-W}; q_k) + \mathcal{L}(x_{t-W+1:t}; q_k) - \mathcal{L}(x_{1:t}; q_k)\right]. \tag{3.2.8}$$

When $\mathcal{Q}_t^{global} \geq K\beta$ we stop and declare a change at time t. Pseudocode outlining the Global algorithm is provided in Algorithm 2. We note that the computational cost of NUNC Global is $\mathcal{O}(KW)$. As with NUNC Local, this can be reduced to $\mathcal{O}(K \log W)$ by implementing an Ordered Search Tree for the empirical CDF (Cormen et al., 2001).

Algorithm 2: NUNC Global Algorithm

Data: $x_{(t-W+1):W}$, the last W realizations from a data generating process X;
$z_W^{(t)}(q_k)$ for $q_1, \ldots, q_K$.

Input: $\beta > 0$; $K < W$; $q_{1:K}$ the fixed quantiles.

$\mathcal{Q} \leftarrow \sum_{k=1}^{K} 2\left[\mathcal{L}(x_{1:t-W}; q_k) + \mathcal{L}(x_{t-W+1:t}; q_k) - \mathcal{L}(x_{1:t}; q_k)\right]$;

if $\mathcal{Q} \geq K\beta$ **then**
  Return $t - W$ as a changepoint.

else
  $z_W^{(t+1)}(q_k) \leftarrow \frac{1}{t-W+1}\left[(t-W)z_W^{(t)}(q_k) + F_{(t-W+1):(t-W+1)}(q_k)\right]$ for $q_k \in q_{1:K}$.

The advantage of this approach, over NUNC Local, is that only K pieces of information are required to store information about the estimate of the CDF of the null distribution, irrespective of the number of points observed so far or the size of the sliding window, satisfying the memory constraint requirement of our application.

3.2.2 Parameter selection

The execution of both NUNC Local and Global requires the selection of various parameters, including the K quantiles $q_1, \ldots, q_k$ and threshold β. Additionally, the size of the sliding window W must be chosen with care. In practice, W be chosen based on specific knowledge of the application and data generating process at hand. We defer further discussion of this until Section 3.3, where we consider the impact of W on different simulation scenarios, and Section 3.4 where we explore selecting W in a practical application by considering a range of different window sizes and selecting the one that offers the best detection power.

Next, we turn to the challenge of choosing the K quantiles $q_1, \ldots, q_k$. The value of K itself should be chosen to be proportionate to $\log(W)$, in line with the method

proposed by Haynes et al. (2017). In particular, the value $K = \lceil 4\log(n) \rceil$ was proposed, see Haynes et al. (2017, Section 4.3) for details. Given K, one approach to choosing the $\{q_k\}$ would be to evaluate evenly spaced empirical quantiles. However, an alternative approach is motivated by Haynes et al. (2017, Section 3.1). That is, we select q_k such that

$$q_k = \hat{F}^{-1}\left(1 + (2W+1)\exp\left[\frac{c}{K}(2k-1)\right]\right)^{-1}, \tag{3.2.9}$$

where $c = -\log(2W-1)$. The reason for making such a choice is that this gives a higher weight to values in the tail of the distribution (Haynes et al., 2017), allowing for more effective change detection. In the Local algorithm, the q_k will be updated as the window changes; in the Global algorithm, however, these K points are fixed in time. As such, the values of q_k must be obtained using the first W points of data the algorithm analyses. In some situations, however, this issue can be avoided because there is prior knowledge of the underlying distribution of the data. In this case known quantiles can be utilised rather than estimating them from the data.

Another important requirement for the two algorithms presented here, as in other sequential changepoint methods, is the ability to control the false alarm rate (Tartakovsky et al., 2014). In general, the value of β will be tuned so that the probability of a false alarm for data under the null hypothesis is set to some level α. This will be the case, for instance, in the telecommunications application where the threshold value will be tuned on devices where no event is detected. That said, we can follow a similar approach to that of Eichinger and Kirch (2018) to obtain an idea of how β relates to the probability of a false alarm. Indeed, we can (asymptotically) approximate the distribution of each term in the sum of equation (3.2.5) by a chi-square-1 distribution (Wilks, 1938). This is the asymptotic distribution of the likelihood-ratio test for a fixed t, τ, and q, assuming independent identically distributed (i.i.d.) data. With this approximation, it can be shown that:

Proposition 2. *If β is chosen such that $\beta = \max\{\beta_1, \beta_2\}$, where*

$$\beta_1 = 1 - 8K^{-1}\log\left(\frac{\alpha}{W(t-W+1)}\right)$$
$$\beta_2 = 1 + 2\sqrt{2\log\left(\frac{W(t-W+1)}{\alpha}\right)},$$

then the probability of a false detection by time t using NUNC Local is bounded above by α.

Proof. This result follows from bounds on the tail of sums of chi-square distributions and a Bonferonni correction – see the Appendix for details. □

It should be noted, that in situations where the window size is large this bound may be conservative due to it being an asymptotic bound. Furthermore, when the assumptions of data independence and identical distribution are not met, this bound may not hold, as illustrated in simulations. In such settings we suggest selecting a threshold by tuning on a data stream that does not contain any changes.

If using an approximate grid of size $J < W$, then it is necessary to replace the values of W in the above proposition with the value J. Furthermore, as a corollary to Proposition 2, a bound can be obtained for use in NUNC Global.

Corollary 1. *If β is chosen such that $\beta = \max\{\beta_1, \beta_2\}$, where*

$$\beta_1 = 1 - 8K^{-1}\log\left(\frac{\alpha}{(t-W+1)}\right)$$
$$\beta_2 = 1 + 2\sqrt{2\log\left(\frac{(t-W+1)}{\alpha}\right)},$$

then the probability of a false detection by time t using NUNC Global is bounded above by α.

Proof. The proofs follow similarly to Proposition 2, however we perform only one test, rather than W tests, per window. □

Due to the connection between our online methodology and the test of Zou et al. (2014) and Haynes et al. (2017) these results can also be extended to the offline setting.

In particular selecting β, as per Proposition 2, as the penalty in these changepoint detection algorithms would control the probability of a false alarm in a segment of length t.

Now that the methodology behind NUNC has been presented, and methods for quantile and threshold selection discussed, we consider its performance within various simulation settings.

3.3 Simulation Study

In this section we examine the properties of both NUNC approaches in various simulation settings. These can be seen in Figure 3.3.1. The first (Figure 3.3.1(a)) is a change in the mixture proportions of a bi-modal Gaussian distribution, whilst the second example considers a change in the scale of a Cauchy Distribution. The third setting is a change in the amplitude of a sinusoidal process, and the final setting is that of a change in the drift parameter of an Ornstein-Uhlenbeck (OU) process. Both the sinusoidal and OU examples are included to highlight how NUNC performs when the independence assumption is not met. These latter two scenarios are also motivated by our telecommunications application as both exhibit drift. Realisations of each of these data generating processes can be seen in Figure 3.3.1.

In what follows, we explore the performance of each method across the four given scenarios. In particular we consider the influence of window size on the power, and the detection delay, of the test. In each setting we will also compare the NUNC-based tests against a MOSUM test, as implemented by Meier et al. (2021), a competitor non-parametric online changepoint algorithm that has a lightweight data footprint.

3.3.1 False Alarm Probability

We begin by considering the false alarm rates returned by the three methods (NUNC Local, NUNC Global and MOSUM) for 100 replicates of each of our four data generating scenarios, without a change being present. In each case, the series generated was of length 1000, with $K = 20$ and $W = 150$ for NUNC Local and NUNC Global.

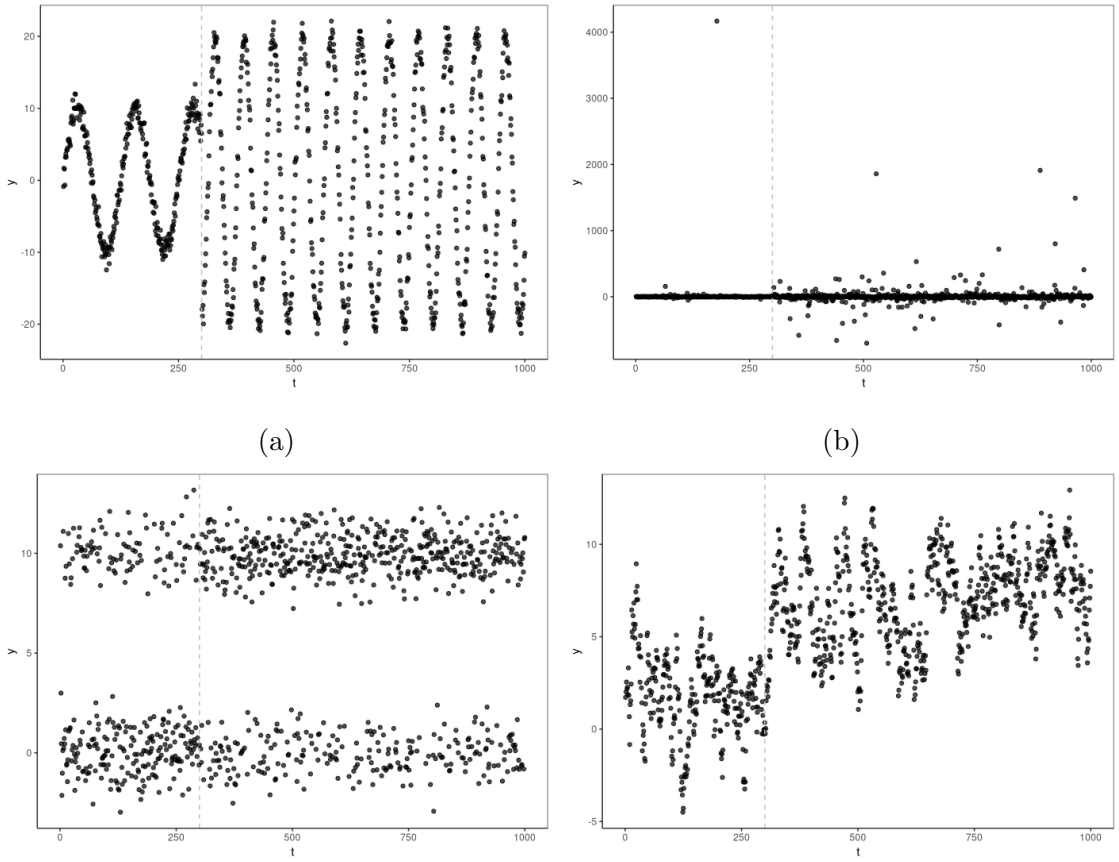
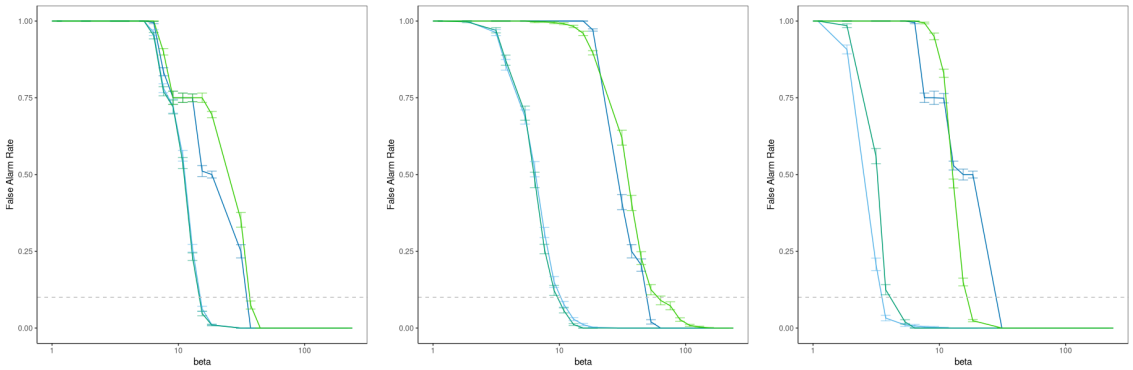


Figure 3.3.1: Four different simulations scenarios: (a) change in the amplitude of a sinusoidal process; (b) change in scale of a Cauchy distribution; (c) change in mixture proportions of a bi-modal Gaussian distribution; and (d) change in the drift parameter of an Ornstein–Uhlenbeck process.

For comparison, we also compared against the equivalent MOSUM procedure (i.e. $W = 150$, and other settings set to default). The resulting false alarm rates for a range of thresholds can be seen in Figure 3.3.2.

To explore the practical utility of Proposition 2, we compare the thresholds required for the i.i.d. multi-modal Gaussian and Cauchy change-in-scale false alarm rate when seeking to achieve a 10% false alarm rate. Our study highlights that penalty values of 9 and 10 are required by the NUNC Global algorithm for the multi-modal Gaussian and Cauchy scenarios respectively. In these two settings, penalties of 11.6 and 12.6 respectively were required for NUNC Local. This compares favourably with the approximate penalty values selected using Proposition 2 (9.51 and 12.30 for the Global and Local cases respectively). Unsurprisingly, in the case of the (non-i.i.d., temporally dependent) sinusoidal and OU scenarios, the penalties required for a 10% false alarm rate differ from those provided by Proposition 2.



(a) (b) (c)

Figure 3.3.2: The False Alarm rate for increasing threshold values for the four different simulation scenarios analysed with (a) NUNC Local, (b) NUNC Global and (c) MOSUM. The dotted line indicates a false alarm rate of 0.10, and the error bars indicate two standard deviations. In each plot the simulation scenarios are represented by sinusoidal change-in-amplitude (dark blue); the Cauchy change-in-scale (light blue); in emerald green, the change-in-mixture proportions (emerald green); and the change-in-drift in a OU process (light green).

3.3.2 Detection Power and Detection Delay

We now turn to consider the detection power and detection delay of the NUNC algorithms in a variety of settings. Following Tartakovsky et al. (2014), we define the detection power as the probability that a changepoint is detected after it has occurred, and the detection delay as the difference between the stopping time of the test and the time the changepoint is known to have emerged. Again we focus on 100 replicates of each of the four scenarios displayed in Figure 3.3.1, where each series is of length 1000 and the change occurs at time $t = 300$. In each case we seek to estimate the detection power and detection delay, controlling the false alarm rate at 10% and allowing the window size W of the algorithms to vary.

Results for the detection power, and detection delay, are summarised in Tables 3.3.1 and 3.3.2 respectively. It is notable that NUNC is able to detect changes in a variety of settings, including those where the data has time-dependent structure. We also note that NUNC Global outperforms NUNC Local in most cases, except when the underlying distribution is sinusoidal. This is perhaps to be expected since NUNC Global incorporates the long-run empirical CDF which stores the historical data. This allows for better identification of departures from the null when the data is stationary. When the data is non-stationary, however, this is not so beneficial and so the performance of NUNC Local is comparable.

Turning to consider the results obtained for the detection delay, displayed in Table 3.3.2, it is evident that NUNC Local demonstrates stronger performance than that of NUNC Global. This is as expected, because NUNC Global checks if the distribution of the data in the window differs from the long run empirical CDF, whereas NUNC Local checks each point in the window (after pruning as per Proposition 1) for a changepoint within the window.

In comparison to the MOSUM, the detection power of NUNC typically exceeds it except in the specific case of multi-modal data being analysed with a large window. The reason MOSUM performs so well in this case is due to the fact that the change in mixture proportions can also be cast as a change in mean. For the sinusoidal process,

however, the non-stationarity of the data means that the threshold that is required is too high for detection to take place.

The results in Table 3.3.1 also illustrate how, as one might expect, the performance of NUNC Local improves for stationary data as the size of the window increases. Specifically, the larger window provides a better estimate of the CDF of the (stationary) data stream, which in turn makes it easier to identify when a change has occurred. The price for this increased power, however, comes in the form of an increase in computational cost due to the larger window size. As such, there is a trade off between detection power and the computational burden of NUNC Local. For NUNC Global, on the other hand, in many situations the use of the long run CDF provides a better estimate of the distribution under the null. This somewhat reduces the need to increase the window size.

3.4 Application

We now revisit the telecommunications example, briefly introduced in Section 3.1, to explore the utility of NUNC in this setting. Recall that the data consists of historic records of a key operational metric routinely monitored on devices that have limited computational power and data storage capability. We have records for 473 such devices, of which 133 were known to contain a (series specific) event that triggered a user intervention. Due to the specifics of the application, engineers believed that it is possible to identify the start of the event in the operational data *before* a user identifies and makes an intervention. If this is true, then it would be desirable to identify the start of changing structure in advance of the user identifying and making an intervention. We call the correct detection of such a change in advance of user identification, an 'anticipation'. Conversely, the detection of such an event before it is resolved is called a 'detection'. The aim of this exploratory analysis, therefore, is to identify to what extent NUNC can (a) identify the correct (event-containing) series and (b) to what extent it can be used to 'anticipate' or 'detect'.

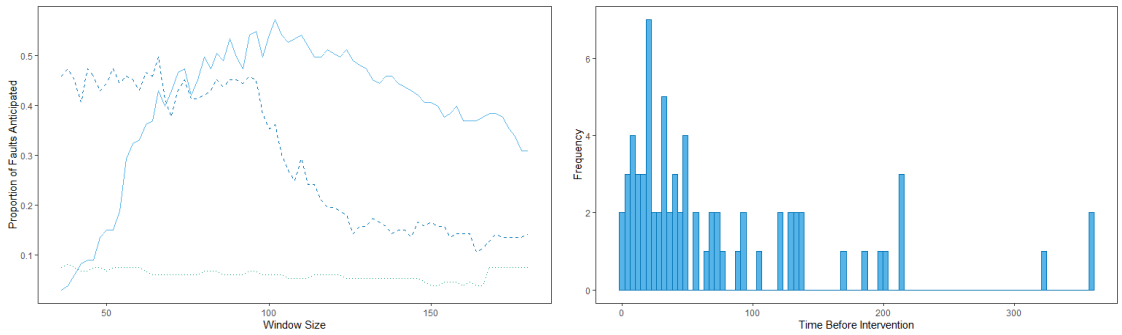
Before summarising the results, we briefly discuss the various parameter choices

Window	50			100			150			200		
Process	Local	Global	MOSUM	Local	Global	MOSUM	Local	Global	MOSUM	Local	Global	MOSUM
Cauchy	0.56	**0.9**	0.02	0.71	**0.89**	0.05	0.81	**0.91**	0.02	0.85	**0.91**	0.01
Multimodal	0.45	**0.87**	0.43	0.31	**0.79**	0.71	0.58	0.8	**0.92**	0.58	0.83	**0.92**
Sinusoidal	**0.95**	0	0.93	**1**	0.92	0	0.17	**0.91**	0	**0.98**	0.92	0
OU	0.25	**0.92**	0.42	0.47	**0.91**	0.68	0.36	**0.9**	0.86	0.79	**0.93**	**0.93**

Table 3.3.1: Table showing the detection power of NUNC Local, NUNC Global, and the MOSUM for various window sizes, with the best performance for each scenario highlighted in bold.

Window	50			100			150			200		
Process	Local	Global	MOSUM	Local	Global	MOSUM	Local	Global	MOSUM	Local	Global	MOSUM
Cauchy	150.16	**67.13**	345	**51.67**	109.39	211	**72.99**	133.64	494	**48.96**	143.69	285
Multimodal	285.29	**131.46**	169.21	285.82	238.63	**49.23**	242.59	266.96	**48.03**	206.09	297.05	**59.07**
Sinusoidal	**19.59**	Inf	411.69	**5**	102.03	Inf	**9.16**	159.3	Inf	**7.58**	238.61	Inf
OU	284.32	**115.88**	173.88	246.71	136.81	**77.65**	242.85	173.01	**78.83**	155.7	221.06	**81.34**

Table 3.3.2: Table showing the average detection delay of NUNC Local, NUNC Global, and the MOSUM for various window sizes, with the best performance for each scenario highlighted in bold. Note that in line with Tartakovsky et al. (2014), if no detection takes place then this corresponds to a detection delay of infinity.



(a) (b)

Figure 3.4.1: Comparison of anticipation rates achieved by the Local (Solid Line) and Global (Dashed Line) variants of NUNC, and the MOSUM (Dotted Line), for varying window sizes, for a window size of 100 and a false alarm rate of 15% in (a). A histogram illustrating the distribution of the time between the detection of an event by NUNC Local and the report by a customer, for a window size of 100 and a false alarm rate of 15% in (b).

made: specifically, the threshold β, the window size W, and the choice of K. In line with Haynes et al. (2017), we choose $K = \lceil 4 \log(W) \rceil$. The choice of β was made to control the false alarm rate at a desired level after discussion with domain experts. In this particular setting, false alarms can be tolerated if this results in improved identification of real events. Consequently a false alarm rate of 15% was selected. In order to identify the appropriate value of β to achieve this, we first fix a window size and then perform NUNC on the 340 data series without the event, choosing a value of β that gives the desired false alarm rate. NUNC is then applied to the 133 series known to contain an event using this β, for the chosen window size, to explore the power of the approach for different window sizes. A similar process is also used to implement the MOSUM test; again, the threshold is chosen to control the false alarm rate at 15% for a given window size.

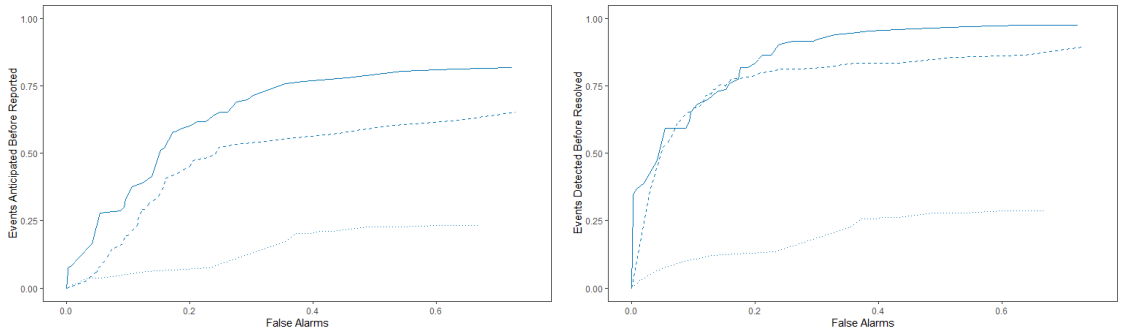
In Figure 3.4.1(a) we present the anticipation rate for a range of window sizes. As can be seen, a window of size between 80 and 120 performs the best, with NUNC Local correctly identifying (i.e. anticipating) $> 50\%$ of events in advance of user intervention. We also note that NUNC outperforms the MOSUM for various choices

of W. One reason for this is because the MOSUM test threshold is set to avoid detecting the non-anomalous changes in mean that many of the series exhibit, and this reduces detection power.

False Alarms	1%	5%	10%	15%
Local	0.08 (0.37)	0.28 (0.59)	0.38 (0.68)	0.51 (0.77)
Global	0.03 (0.36)	0.06 (0.51)	0.20 (0.67)	0.35 (0.76)
MOSUM	0.02 (0.02)	0.04 (0.08)	0.05 (0.10)	0.06 (0.12)

Table 3.4.1: Table illustrating proportion of events anticipated (and detected) for varying rates of false alarms for both NUNC Local and NUNC Global.

Finally, for a fixed window size ($W = 100$) we explore the anticipation and detection rate as the false alarm rate (or equivalently β) varies. The results are summarised in Table 3.4.1, and Figure 3.4.2.



(a) (b)

Figure 3.4.2: Comparison of event anticipation (a) and event detection (b) rates achieved by both the Local (Solid Line) and Global (Dashed Line) variants of NUNC, and the MOSUM (dotted line), for a window of size $W = 100$ and a varying false alarm rate.

From the results presented, one remark that can be made is that the detection power for NUNC Local and NUNC Global is similar, with both achieving over 75% for a false alarm rate of 15% and window of $W = 100$, but the anticipation power of NUNC Local is significantly better for a range of false alarm rates and window sizes.

As in the simulations on detection delay, this is due to the way that NUNC Local checks every point within the window for a change, allowing for a shorter detection delay. A second observation is that NUNC achieves better results than the MOSUM in terms of both anticipation, and power, as the false alarm rate varies.

3.5 Concluding Remarks

This chapter has introduced NUNC Local and NUNC Global: two related, non-parametric changepoint methods with a lightweight data footprint. NUNC Global offers greater power in instances where there is a stationary underlying null distribution for the data (*cf.* Section 3.3.2). Conversely NUNC Local shows greater resilience, and is able to outperform NUNC Global, in settings where the process contains time-dependent or other non-independent structure, such as the telecommunications example that we consider. Finally, we have explored how both forms of NUNC compare against MOSUM, an existing non-parametric window based changepoint detection test with lightweight footprint, in both simulations and the telecommunications application.

To support the implementation of the algorithms, we also consider their computational properties and key theoretical results. The first of these provides a cutoff value for the sliding window in NUNC Local, so that only relevant parts of the sliding window are checked for a change. This provides a theoretical justification for reducing the computational cost of the algorithm. A method for (approximately) selecting the threshold was also provided, enabling the control of the false alarm rate at a fixed level.

As with any approach, NUNC has various weaknesses that might be identified for criticism. For example, the estimation of the empirical CDF from windowed data means that gradual changes are likely to go undetected. In addition, as identified by our simulation study, NUNC Global struggles with changing structure in time-dependent series. The investigation of potential alternatives to the NUNC framework, that can resolve such weaknesses, are left as avenues for future research.

The reason for developing NUNC was to enable to data-driven identification of an event that would anticipate a user intervention related to a telecommunications device, rather than simply waiting for direct communication from end-users. As seen in Section 3.4 this can be achieved with some success, with NUNC Local allowing for detection in advance of the event in a meaningful number of cases. This provides a powerful diagnostic tool, and demonstrates the ability of NUNC to successfully detect changes that anticipate user intervention in a large number of cases.

Chapter 4

Detection of Emergent Anomalous Structure Phenomena in Functional Data

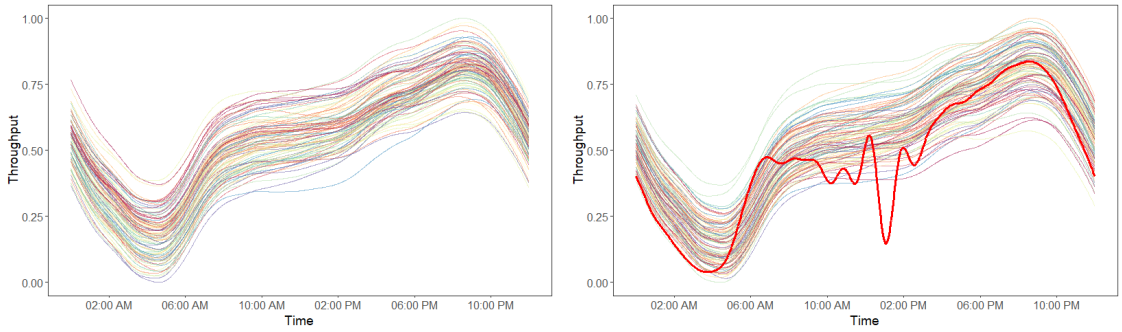
4.1 Introduction

In recent years, anomaly detection for functional data has received substantial attention. Particular attention has been focused on identifying a completely observed anomalous curve that deviates in shape or magnitude from the other observations in a sample of functional data. Recent contributions to this area include the works of Pintado and Romo (2011); Arribas-Gil and Romo (2014); Rousseeuw et al. (2018); Dai and Genton (2018a); Huang and Sun (2019); Harris et al. (2020); Rieser and Filzmoser (2022).

An extension of the classical functional anomaly detection problem is that of sequentially monitoring for anomalies in functional data. In the classical non-sequential setting, a given approach analyses an entire sample of functional data to identify curves that are anomalous. In one version of the sequential setting, however, *complete* functional observations are received sequentially and compared to some baseline sample to determine whether the new curve appears anomalous. As such, a sequential method does not analyse the whole sample of data and instead makes a decision based

solely on the observations as they are recorded. Existing sequential methods include the functional profile monitoring techniques of Hall et al. (2001); Jeong et al. (2006); Colosimo and Pacella (2010); Paynabar et al. (2016); Yan et al. (2018).

This chapter is motivated by a different, yet related, anomaly detection challenge arising from a telecommunications application monitoring throughput data at a point on a digital network. In this setting the observed data exhibits a similar shape over any given day, and this shape takes the form of a smooth curve. It is this smooth curve structure that motivates our functional data approach. Figure 4.1.1 shows an example of this throughput functional data. In some instances, an anomaly is observed where there is a deviation from the daily shape, as shown for example in Figure 4.1.1(b). Such behaviour might indicate the occurrence of a significant event on the network that requires an operational intervention.



(a) (b)

Figure 4.1.1: Observations of throughput data recorded at one minute intervals over 100 days at a point on a telecommunications network and represented as a functional time series. Each curve denotes a single day of data (a). An instance of an anomalous day is overlaid in red in (b).

To perform the operational interventions as soon as possible, we cannot wait for a new curve to be completely observed and instead need to identify the anomalies as they emerge. As such, we ideally need a method that monitors a *partial* curve sequentially, as the data is collected over time. Existing functional anomaly detection methods such as those proposed by Hyndman and Shang (2010) or Harris et al. (2020) require the whole curve to be observed, and so are not applicable to this

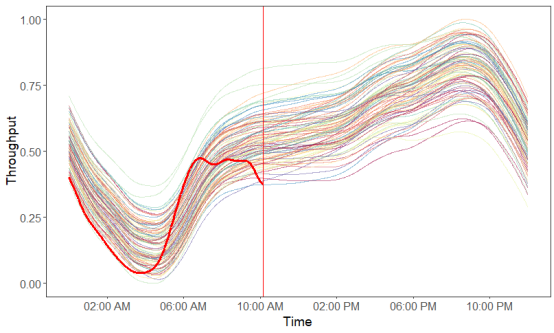


Figure 4.1.2: Sequential detection of emergent behaviour observed in Figure 1(b).

problem. Furthermore, it would not be computationally feasible to perform these existing methods iteratively, as each new point on the curve is received. On the other hand, non-functional methodologies for monitoring non-stationary data, e.g. Zhang et al. (2018), Wang et al. (2018), or Talagala et al. (2020) may not handle the temporal dependence contained in functional data, or require the anomalies to consist of an abrupt change in mean. To address this, we propose detecting smooth deviations from an underlying functional data shape sequentially, as the partially observed functional data emerges over time. To be specific, we seek to identify both the anomalous curves as they arise, and identify where in time on the curve the anomaly takes place. (See Figure 4.1.2). To this end we introduce a Functional Anomaly Sequential Test (FAST).

The essence of our proposed approach can be summarised as follows: FAST first estimates the best fitting linear differential operator from a set of training data, before then applying this operator to a new observation to obtain a new estimated residual function. This residual function is used to calculate the test statistic for a non-parametric CUSUM test (Tartakovsky et al., 2005), where if the cumulative value of this test statistic exceeds a threshold then an anomaly is declared. By performing this CUSUM test sequentially on a partially observed curve anomalies can be detected as they emerge.

The remainder of the chapter is organised as follows: Section 4.2 introduces the functional data framework and methodology that underpins the FAST procedure. Various theoretical properties of the anomaly detection procedure are established

in Section 4.3, including a threshold that controls the false alarm rate, and results relating to the power of the test. A simulation study is then described in Section 4.4, illustrating the performance of the method across a range of settings. We conclude in Section 4.5 with an analysis of the throughput telecommunications data.

4.2 Background and Methodology

We start by outlining our approach to detecting emergent anomalies in functional data. In Sections 4.2.1 and 4.2.2 we introduce the model which we assume describes the data observed and our estimation approach, prior to introducing the FAST algorithm in Section 4.2.3.

4.2.1 Model

In what follows we adopt the functional data model of Ramsay (2005). Consider a sequence of functional observations $\{X_1(t), X_2(t), \ldots, X_n(t)\} \in L^2(\mathcal{T})$, none of which are anomalous, and where the subscript denotes the i-th observation over the interval $\mathcal{T}$. In this work the interval is a day, so it is convenient to view $\{X_i(t)\}$ as denoting the observed value for the i-th day at time t.

As described in Section 4.1, each of our non-anomalous observations follows an underlying shape. In order to represent this underlying structure we utilise the approach of Ramsay (2005) and model the observations as noisy realisations of the solution to an order m linear ODE, with the solution to the ODE describing the underlying shape structure. The linear operator for this ODE takes the following form:

$$\mathcal{L} = \beta_0(t)\mathrm{D}^0 + \beta_1(t)\mathrm{D} + \ldots + \beta_{m-1}(t)\mathrm{D}^{m-1} + \mathrm{D}^m, \tag{4.2.1}$$

where D^k denotes the k^{th} derivative of a function, and the $\{\beta_j(t)\}_{j=0}^{m-1}$ are coefficient functions that are estimated from the data. The estimation of the operator, and the choice of m, will be discussed in the next section.

Under the assumption that each observation is a noisy realisation of the solution to the ODE underpinned by the operator in equation (4.2.1), an observation is given

by

$$X_i(t) = \sum_{j=1}^{m} c_j u_j(t) + f_i(t), \quad c_j \in \mathbb{R}, 1 \leq i \leq n, t \in \mathcal{T}. \tag{4.2.2}$$

Here $\{u_j(t)\}_{j=1}^m$ are the solutions to the ODE and $\sum_{j=1}^m c_j u_j(t)$ represents the underlying shape for the data. In equation (4.2.2) it is assumed that the $\{u_j(t)\}$ are linearly independent m-times differentiable functions, and the $\{f_i(t)\}$ are m-times differentiable mean-zero functional random variables representing noise.

We make the following assumption about the noise functions:

Assumption 1. *The noise functions, $\{f_i(t)\}$, are independent mean zero Gaussian processes with stationary covariance functions.*

Whilst FAST does not require this assumption to work in practice, it is used for model selection in Section 4.2.2, and to develop an asymptotic distribution for the test statistic in Section 4.3.

The aim of FAST is to detect when an observation has been contaminated by an anomaly. We model an anomaly as a function, $\{g_i(t)\}$, causing a deviation from the underlying shape for the data. We make the following assumptions regarding these anomaly functions:

Assumption 2. *The anomaly functions $\{g_i(t)\}$ are (i) m-times differentiable, and (ii) not equal to a linear combination of the $\{u_j(t)\}$.*

Assumption 2(i) is required so that the differentiability condition on the $\{X_i(t)\}$ is still satisfied when an anomaly is present. Additionally Assumption 2(ii) ensures that the anomaly does not follow the underlying shape for the data. We also remark that it is not necessary for an anomaly function to contaminate the entire observation period, and that it can be present on some sub-region, $\mathcal{S} \subset \mathcal{T}$.

Building on equation (4.2.2), a model for a contaminated observation can be defined as:

$$X_i(t) = \begin{cases} \sum_{j=1}^m c_j u_j(t) + f_i(t) & t \in \mathcal{T} \setminus \mathcal{S}_i \\ \sum_{j=1}^m c_j u_j(t) + f_i(t) + g_i(t) & t \in \mathcal{S}_i. \end{cases} \tag{4.2.3}$$

Whilst the above functional data model for the observations assumes continous time, this may not be possible in practice. In such circumstances we follow the approach taken by other authors such as Nagy et al. (2017), and model the observations on a fine grid. We let $\{1, \ldots, T\}$ be the fine grid that discretizes the compact set $\mathcal{T}$, and let the discretized observation be denoted by $\{X_i(\tau)\}_{\tau=1}^{T}$ for $1 \leq i \leq n$.

With our model for the data generating process outlined, we now turn to consider the estimation scheme that will be used. As we shall see in Section 4.2.3, the FAST algorithm uses an estimate of the linear differential operator in equation (4.2.1), rather than estimates of the solution functions in equation (4.2.2). To estimate the operator, we follow the approach of Ramsay (1996) and use Principal Differential Analysis, as we describe in the next section.

4.2.2 Estimating the Model Using Principal Differential Analysis

We begin by recalling that the approach of Ramsay (1996) to estimating an ODE to model functional data uses Principal Differential Analysis (PDA). PDA estimates the best fitting order m linear differential operator for a sample of functional data. As described in the previous section, in this work the linear differential operator takes the form of equation (4.2.1).

The linear ordinary differential operator is estimated by selecting the values of $\boldsymbol{\beta(t)} = \{\beta_j(t)\}_{j=0}^{m-1}$ that minimise $||\mathcal{L}\left(\boldsymbol{X}(t)\right)||_2^2$, where $||\cdot||_2$ represents the usual L^2 norm. As described in Section 2.4.3, Ramsay (1996) propose performing this minimisation on a discretised grid of T points using the follow estimation scheme:

$$\boldsymbol{\beta}(t) = \left(\boldsymbol{Y}(t)^T\boldsymbol{Y}(t) + \lambda\boldsymbol{I}_m\right)^{-1}\boldsymbol{Y}(t)^T\boldsymbol{z}(t).$$

Here $\boldsymbol{\beta}(t) = \{\beta_j(t)\}_{j=0}^{m-1}$, $\boldsymbol{Y}(t)$ is an $n \times m$ matrix with entries $Y_{ij}(t) = \mathrm{D}^{j-1}(X_i(t))$, $\boldsymbol{z}(t) = \{D^m(X_i)(t)\}_{i=1}^{n}$, $\boldsymbol{I}_m$ is the $m \times m$ identity matrix, and λ is a parameter that controls the smoothness of the estimated coefficient functions. The estimated ODE, $\hat{\mathcal{L}} = 0$, will have m linearly independent solutions, and these estimated solution functions are estimates of $\{u_j(t)\}$ in equation (4.2.2). Further computational efficiencies

can be achieved when estimating the coefficients through use of basis function smoothing. In the appendix we present simulations that explore the effect this has on the estimated differential operator.

To detect anomalies the FAST algorithm applies the estimated ODE to the observed data. Applying the operator $\mathcal{L}$ to an observation produces a residual function, $\{\epsilon_i(t)\}_{i=1}^n$, and it is these residual functions that form a key part of our test. We denote the estimated residual obtained by applying the estimated ODE by $\{\hat{\epsilon}_i(t)\}_{i=1}^n$. Note that under Assumption 1, and using the well known properties of Gaussian processes, the residual functions are also themselves mean-zero Gaussian processes (Rasmussen and Williams, 2005).

An explicit connection between the residual functions and the model for the data presented in equation (4.2.2) is given by the following:

$$\begin{aligned}\hat{\mathcal{L}}(X_i(t)) &= \hat{\mathcal{L}}\left(\sum_{j=1}^{m} c_j u_j(t) + f_i(t)\right) \\ &= \hat{\mathcal{L}}\left(f_i(t)\right) \\ &= \hat{\epsilon}_i(t), \quad 1 \leq i \leq n. \end{aligned} \tag{4.2.4}$$

Although it will not be known *a-priori* if an observation is contaminated with an anomaly, it can also be instructive to consider what happens when the differential operator is applied to such an observation. Applying $\hat{\mathcal{L}}$ to the model in equation (4.2.3) gives the following:

$$\hat{\mathcal{L}}(X_i(t)) = \begin{cases} \hat{\epsilon}_i(t) & t \in \mathcal{T} \setminus \mathcal{S}_i \\ \hat{\mathcal{L}}(g_i(t)) + \hat{\epsilon}_i(t) & t \in \mathcal{S}_i. \end{cases} \tag{4.2.5}$$

It is this property that motivates the FAST test: observations that deviate from the underlying shape for the data will have contaminated residual functions.

One aspect of the model estimation that is yet to be addressed is the selection of m, the order of the ODE to fit to the data. This is a subject that has received little attention in the literature to date. It is suggested in Ramsay (1996) that the m should be specified by discussion with experts or through use of physical modelling.

Unfortunately in many applications, for example the telecommunications example in Section 4.5, the order will not be known a priori.

As an alternative we propose selecting the order of the differential operator in a data driven manner, motivated by the selection of regression models. In particular, it is proposed that the Bayesian Information Criterion (BIC) (Schwarz, 1978) is used to select the order of the ODE to fit to the data (Teräsvirta and Mellin, 1986). BIC is given by

$$\text{BIC}(m) = m\log(n) - 2\hat{\ell} \tag{4.2.6}$$

were $\hat{\ell}$ is the log-likelihood of the fitted model at the MLE, m is the order for the ODE and n is the number of observed curves. Using Assumption 1 the residual functions are mean zero Gaussian processes, and so on a discretised grid $\{\epsilon_i(\tau)\}_{\tau=1}^{T}$ will follow a mean zero multivariate Gaussian distribution. As such, the log-likelihood in equation (4.2.6) is that of a multivariate normal distribution. In the case where the noise functions are independent the log-likelihood reduces to (Johnson and Wichern, 1988):

$$\hat{\ell} = -\frac{n}{2}\log(\frac{\sum_{i=1}^{n}\sum_{\tau=1}^{T}\hat{\epsilon}_i^2(\tau)}{n}), \tag{4.2.7}$$

where $\sum_{i=1}^{n}\sum_{\tau=1}^{T}\hat{\epsilon}_i^2(\tau)$ is the sum of square error between the observations and the solution to the estimated ODE.

4.2.3 FAST: A Test for Intra-Functional Anomaly Detection

Having established our approach for estimating the ODE for non-anomalous data, we are in a position to introduce FAST. The aim of FAST is to sequentially monitor an emerging functional observation for anomalous behaviour. To achieve this it combines the ODE estimation introduced in the previous section with a non-parametric CUSUM test to identify when an anomaly is emerging. The pseudocode for FAST can be found in Algorithm 3.

The first stage of the FAST algorithm takes as input a set of anomaly-free training observations, $\{X_1, \ldots, X_n\} \in L^2(\mathcal{T})$, and uses PDA to estimate $\hat{\mathcal{L}}$. This best fitting linear differential operator is then applied to a new observation $\{X_{n+1}(t)\}$ to obtain

a residual function, $\{\hat{\epsilon}_{n+1}(t)\}$. The residual function can then be used to define the test function, $Z_i(\tau)$, used by FAST. This is given by

$$Z_{n+1}(\tau) = \frac{S_{n+1}(\tau) - \mu(\tau)}{\sigma(\tau)}, \quad 2 \leq \tau \leq T \tag{4.2.8}$$

where

$$S_{n+1}(\tau) = \left(\hat{\epsilon}_{n+1}(\tau) - \hat{\epsilon}_{n+1}(\tau - 1)\right)^2, \quad 2 \leq \tau \leq T.$$

Note that $\mu(\tau)$ and $\sigma(\tau)$ are the mean and standard deviation of S_{n+1} at time τ respectively. In some settings, the structure for the noise function will be known and so $\mu(\tau)$ and $\sigma(\tau)$ can be obtained analytically. In others, however, these values will be unknown. In this case they can be estimated from the training data that has been used to estimate $\hat{\mathcal{L}}$. We then replace the $\mu(\tau)$ and $\sigma(\tau)$ in equation (4.2.8) with their estimators $\hat{\mu}(\tau)$ and $\hat{\sigma}(\tau)$.

We standardise the test function in equation (4.2.8) so that at each time step it is a mean-zero, unit variance random variable. This is necessary so that a threshold for the test can be derived that controls the false alarm rate when the observations are dependent, as is expected in the functional data setting. The cumulative sum of equation (4.2.8) can then be treated as the test statistic, $\Delta_{n+1}(\tau)$; this is given by

$$\Delta_{n+1}(\tau) = \sum_{r=2}^{\tau} Z_{n+1}(r), \quad 2 \leq \tau \leq T. \tag{4.2.9}$$

When performing FAST, the absolute value of $\Delta_{n+1}(\tau)$ is taken so that the test is one sided. This is convenient since it allows for a single threshold for the test to be selected by a user; and also because we do not make the distinction between large positive, or negative, deviations of $\Delta_{n+1}(\tau)$. Furthermore, in line with the existing CUSUM literature, the test statistic is scaled by the square root of the length of the cumulative sum. The purpose of this is to control the increasing variation of the test statistic under the null as the value of T increases.

An anomaly is detected should the scaled test statistic cross a threshold denoted by $\gamma \geq 0$. This gives rise to the following stopping rule for FAST:

$$\xi = \min\left\{2 \leq \tau \leq T \ : \ \frac{|\Delta_{n+1}(\tau)|}{\sqrt{\tau - 1}} \geq \gamma\right\}, \tag{4.2.10}$$

where $\min\{\emptyset\} = \infty$ indicates no alarm being sounded. The threshold, γ, is chosen so that the false alarm probability for FAST is controlled. Whilst in practice this could be obtained through simulation, it is also possible to derive a theoretical choice for the threshold for settings where Assumption 1 holds. This choice is discussed in Section 4.3.

Algorithm 3: FAST Algorithm

Compute $\hat{\mathcal{L}}$ from the observed data $X_1, \ldots, X_n$. ;

for τ *in* $2 : T$ **do**

  Compute $S_1(\tau), \ldots, S_n(\tau)$;

  $\mu(\tau) = n^{-1} \sum_{i=1}^{n} S_i(\tau)$;

  $\sigma(\tau) = (n-1)^{-1} \sum_{i=1}^{n} (S_i(\tau) - \mu(\tau))^2$;

end

Initialisation ;

Set $\Delta(1) = 0$;

Choose the threshold γ ;

for τ *in* $2 : T$ **do**

  Compute $Z_{n+1}(\tau)$;

  $\Delta_{n+1}(\tau) = \Delta_{n+1}(\tau - 1) + Z_{n+1}(\tau)$;

  if $|\Delta_{n+1}(\tau)| \geq \gamma\sqrt{\tau - 1}$ **then**

    BREAK;

  end

end

4.3 Theoretical Properties of FAST

We now derive various theoretical properties for FAST, including the selection of a threshold for the test that controls the false alarm rate at a suitable level, and results for the power of the test for both a finite sample setting and as $T \to \infty$.

In the classical (non-functional) online changepoint setting, it is common to consider the average run length to false alarm (ARL). In practice selection of the threshold for the changepoint test is made based on controlling the ARL at some pre-specified level. This is done using training data in many applications due to the complexity of the run length distribution (see Tartakovsky et al. (2014) for details). Within the functional data generating scenario that we consider, however, we only observe each curve at T points and so the test terminates at time T. Consequently, the following is a more appropriate quantity to consider in this setting: the probability of false alarm by time T. Below we propose a threshold that controls this quantity at a desired level.

The probability of a false alarm by time T is defined as

$$\mathrm{P}(\xi \leq T \,|\, \text{No Change}) = \sum_{\tau=2}^{T} \mathrm{P}\left(\xi = \tau \,|\, \text{No Change}\right), \tag{4.3.1}$$

where ξ is the stopping time for the test.

To establish a suitable threshold that controls the false alarm probability, we follow Eichinger and Kirch (2018) and use a central limit argument to approximate the asymptotic distribution of the test statistic in equation (4.2.10). We use this approximation to select the threshold as per the following proposition. As we shall see later (Section 4.4.1), whilst this is an asymptotic bound, it also performs favourably in the finite sample setting.

Proposition 3. *Let*

$$\gamma = \Phi^{-1}\left(1 - \frac{\alpha}{2(T-1)}\right) \tag{4.3.2}$$

be the threshold used in the detection test. Then the probability of a false alarm by time T is bounded above by α.

Proof. See appendix. □

Whilst control of the false alarm rate is an important component for any sequential test, theoretical guarantees for the power of the test are also desirable. In the next result we consider an anomaly function, $\{g_{n+1}(\tau)\}$, emerging at time ν, for a period of

length s; and a test threshold set as in Proposition 3. Using a Central Limit Theorem approximation we can establish the following bounds on the probability of detecting this anomaly.

Proposition 4. *Let $\{g_{n+1}(\tau)\}$ be an anomaly for a new observation that emerges at time ν, until time $\nu + s$; and let γ be the test threshold set as in Proposition 3. The probability of detecting an anomaly by time T is bounded above and below by*

$$\mathrm{P}\left(|U_T| \geq \gamma\right) \leq \mathrm{P}(\xi \leq T) \leq \sum_{\tau=2}^{T} \mathrm{P}\left(|U_\tau| \geq \gamma\right). \tag{4.3.3}$$

Here $U_\tau \sim \mathrm{N}\left(\frac{1}{\sqrt{\tau-1}} \sum_{r=2}^{\tau} \frac{\lambda(r)}{\sqrt{2}}, 1 + \frac{1}{\tau-1} \sum_{r=2}^{\tau} 2\lambda(r)\right)$, $\lambda(r) = \frac{(\hat{\mathcal{L}}(g_{n+1})(r) - \hat{\mathcal{L}}(g_{n+1})(r-1))^2}{\mu(r)}$, and $\hat{\mathcal{L}}(g_{n+1})(r) = 0$ outside of $\nu \leq r \leq \nu + s$.

Proof. See appendix. □

Proposition 4 illustrates how, as the size of the anomaly increases, the probability of detection increases. The connection between the result and the duration of time the anomaly emerges for, s, can also be obtained by rewriting U_τ as follows:

$$U_\tau \sim \begin{cases} \mathrm{N}(0,1) & \tau < \nu \\ \mathrm{N}\left(\frac{1}{\sqrt{\tau-1}} \sum_{r=\nu}^{\tau} \frac{\lambda(r)}{\sqrt{2}}, 1 + \frac{1}{\tau-1} \sum_{r=\nu}^{\tau} 2\lambda(r)\right) & \nu \leq \tau \leq \nu + s \\ \mathrm{N}\left(\frac{1}{\sqrt{\tau-1}} \sum_{r=\nu}^{\nu+s} \frac{\lambda(r)}{\sqrt{2}}, 1 + \frac{1}{\tau-1} \sum_{r=\nu}^{\nu+s} 2\lambda(r)\right) & \tau > \nu + s. \end{cases}$$

In other words, as the size of s increases, both the mean and variance of U_τ increase. Consequently the probability of detection also increases, much as one might expect.

Although the previous proposition considers the detection performance of FAST in a finite sample setting, the detection power of FAST as $T \to \infty$ can also be considered. The reason such a setting is useful to consider is that the threshold, γ, proposed in Proposition 3 grows larger as T increases. In light of this one may ask if, for large values of T, choosing the threshold using Proposition 3 will reduce the power of the test. The next proposition establishes that, given certain conditions on the anomaly, as $T \to \infty$ the detection of an anomaly occurs with probability equal to one.

Proposition 5. *Let* $\{g_{n+1}(\tau)\}$ *be an anomaly for a new observation that emerges for all* $\tau \geq \nu$, *where* $\nu = \lfloor \theta T \rfloor$ *and* $0 < \theta < 1$, *and let* A *be the event that* $\left\{\hat{\mathcal{L}}(g_{n+1})(\tau) - \hat{\mathcal{L}}(g_{n+1})(\tau - 1) = 0\right\}$ *at most finitely many times. Then*

$$\lim_{T\to\infty} \mathrm{P}\left[\xi < \infty \middle| A\,,\, \gamma = \Phi^{-1}\left(1 - \frac{\alpha}{2(T-1)}\right)\right] = 1. \tag{4.3.4}$$

Proof. See appendix. □

Having developed theoretical results for both the control of the false alarm probability under the null, and the power of detection under the alternative, we now turn our attention to the finite sample performance of FAST.

4.4 Simulation Study

In this section we will begin by considering the finite sample behaviour of the asymptotic threshold derived in Proposition 3, before assessing the performance of the detection test on anomalous functions in a variety of cases. As we will see, FAST is able to detect different forms of anomalies in a range of settings, including scenarios where the test is mis-specified (i.e. where the assumptions underpinning Section 4.2 are not fully satisfied).

Three different scenarios will be used in the simulation studies in this section. The first scenario is one where the assumptions laid out in our functional data framework are satisfied: a sinusoidal functional perturbed by a Gaussian noise:

$$X_i(t) = \sin\left(\frac{1}{100}\pi t\right) + \cos\left(\frac{1}{100}\pi t\right) + f_i(t), \quad 1 \leq i \leq n, t \in [1, 500]. \tag{4.4.1}$$

Here $f_i(t)$ is a Gaussian process with covariance operator, $K(s,t)$, given by $K(s,t) = 0.3\exp\left(\frac{1}{2}\left(\frac{s-t}{100}\right)^2\right)$. Simulations where the choice of covariance function is a Matérn function have also been completed, leading to similar results, and these are contained in the appendix.

The second scenario we consider consists of replacing the Gaussian process innovations, $\{f_i(t)\}$, in equation (4.4.1) with innovations drawn from a Student's t-process. We explore the scenario where the t-process has two degrees of freedom and scale

operator given by the same squared exponential covariance operator used for the innovations in equation (4.4.1).

Finally, the third scenario we explore corresponds to one where there is a constant underlying shape for the data plus some innovations $\{f_i(t)\}$. This data generating process is defined as

$$X_i(t) = c + f_i(t), \quad 1 \leq i \leq n,\, t \in [1, 500],\, c \in \mathbb{R}.$$

where the $\{f_i(t)\}$ are innovations that follows the same distribution as those in equation (4.4.1). When $c \neq 0$ this system will not be the solution to any linear operator of the form in equation (4.2.1) with non-zero coefficients, and consequently PDA should not be able to estimate the shape of the data.

4.4.1 Finite Sample Threshold Performance

Our first simulations explore the Type 1 error properties of FAST. To this end, the threshold derived in Proposition 3 was used to monitor observations from a model that contained no anomalies. We consider this for each of the three scenarios introduced above, performing 1000 repetitions of each simulation. For each repetition we generate 100 observations from the data generating process over an interval of 500 time points.

Scenario Alpha	Scenario 1	Scenario 2	Scenario 3
0.01	0.01	0.05	0.04
0.05	0.04	0.06	0.06
0.1	0.05	0.07	0.07
0.2	0.06	0.08	0.08

Table 4.4.1: Comparison of empirical false alarm rate for different data generating processes when the threshold has been set using Proposition 3 to control the false alarm rate at α.

As can be seen from the results in Table 4.4.1, for scenario one - where the assumptions of our setting are satisfied - the threshold controls the false alarm probability at

no more than the desired α. The results in the table also show that, in particular for higher levels of α, the empirical false alarm rate is lower than the desired value. This is a common feature of many sequential testing approaches, and is due to the use of a Bonferroni Correction to control the type one error in a series of dependent hypothesis tests (Tartakovsky et al., 2014). In particular, the correction leads to Proposition 3 providing an overinflated threshold.

On the other hand, in scenarios two and three the false alarm probability is not always controlled at the desired level. This is, perhaps, to be expected, as the assumptions required for Proposition 3 do not hold. That said, similar to in Scenario one, for $\alpha = 0.05$ and above the empirical false alarm rate does not exceed the level α. The cause is again related to the effect of the Bonferroni Correction and the fact that the test threshold is overinflated. The implication of these simulations results are that Proposition 3 can be effective at controlling the false alarm probability even in settings where our assumptions are not met.

Although control of the false alarm rate is important, the main purpose of FAST is to detect emergent anomalies in a sample of functional data. To this end, in the next section we will explore the detection performance of FAST in a range of settings.

4.4.2 Anomaly Detection Performance

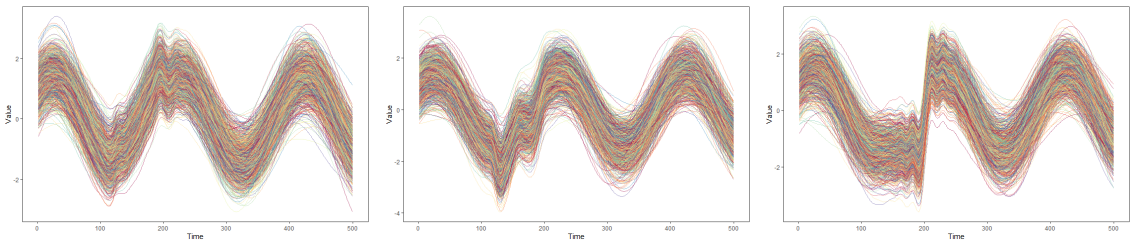
Having shown that the theoretical result for the penalty is able to control the false alarm rate in a finite sample, attention can now be given to the finite sample anomaly detection performance of FAST.

Three different anomaly functions will be inserted into the data generating processes. These anomalies are given by:

1. **Polynomial Anomaly** $g(t) = a_0 + a_1(t-100) + a_2(t-100)^2 + a_3(t-100)^3 + a_4(t-100)^4, a_0 \sim \mathrm{U}[0, 0.5]$, $a_i \sim \mathrm{U}[0.5, 1.5]$, $t \in [100, 200]$.

2. **Sinusoidal Anomaly** $g(t) = b_0 \sin\left(\frac{2\pi t}{100}\right) + b_1 \cos\left(\frac{2\pi t}{100}\right)$, $b_i \sim \mathrm{U}[0.01, 0.2]$, $t \in [100, 200]$.

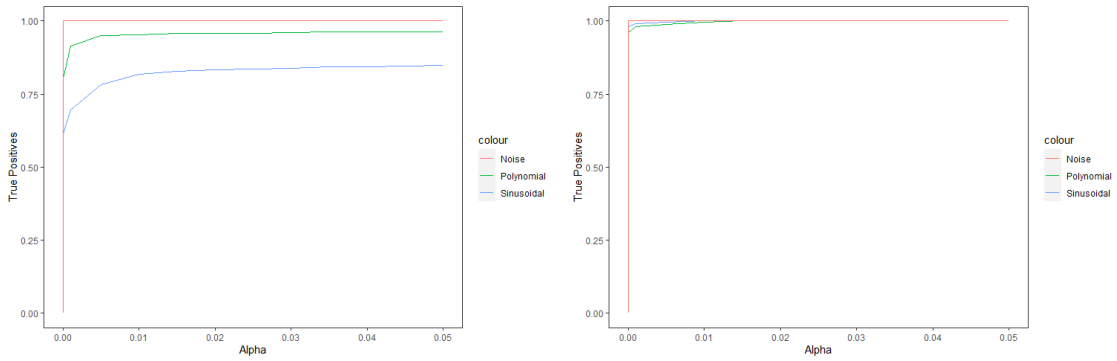
3. **Loss Of Shape Anomaly** $g(t) = 0 + f_i(t)$, $t \in [100, 200]$.



(a) (b) (c)

Figure 4.4.1: 1000 instances of the three anomaly functions; polynomial, sinusoidal, and loss of underlying shape; inserted into the Gaussian process residual function data generating process.

The polynomial and sinusoidal anomalies represent magnitude and shape anomalies respectively, whereas the loss of shape anomaly represents the underlying shape for the data disappearing. The result of this is that the observations consist only of noise. This is motivated by a behaviour seen in our application, where the throughput at a point on the network becomes zero for a period of time.



(a) Order One Model (b) Order Two Model

Figure 4.4.2: ROC curves for the Gaussian process residual data generating process. The sinusoidal anomaly is the blue line, the polynomial anomaly is the green line, and the noise only anomaly is the red line.

For each data generating process, and anomaly type, both the number of detections, and the average detection delay, will be recorded. In line with Tartakovsky et al. (2014) we define detection power as the number of anomalies that are detected

before the end of the anomalous region $\mathcal{S}$, and the detection delay as the time taken between the emergence of anomaly at time $\nu = 100$, and the detection by FAST. This is also referred to as the conditional detection delay, as we record the delay conditional on a successful detection (Tartakovsky et al., 2014).

The metrics used in this simulation study can be connected to the two anomaly detection problems that FAST has been designed to consider. The first problem is the detection of anomalous curves, and the second is the sequential identification of where in time the anomaly emerges. The detection power indicates the ability of FAST to identify the anomalies, whilst the detection delay indicates how quickly the emergence of the anomaly is detected. A further property of FAST assessed in these simulations is the robustness of the method to model mis-specification. In particular, we fit both the correct order ODE model (an order two model), and an incorrect order one model.

To carry out the study, a sample of 100 functions is drawn from a data generating process and used as a training set for detection of an anomaly. Each of the data generating processes in this section consist of a single underlying shape, and the anomalies deviate from this underlying shape. As in the previous section, we perform 1000 repetitions of each scenario, and generate the data over an interval of 500 time points. Furthermore for each test we set the threshold using Propositon 3 with $\alpha = 0.05$.

The performance of FAST with respect to these two metrics is presented in the tables below, and ROC curves are provided in Figure 4.4.2 for the Gaussian process residual data generating process. In particular these results indicate that in almost all scenarios FAST has high power when the data generating process satisfies our assumptions, even when the order of the ODE is mis-specified as an order one model. Detection delay results are also encouraging. Note, in particular, that the average detection delay in each scenario is less than the size of the anomalous region. The results in the table also show that the average detection delay is lower when the order for the ODE is specified correctly.

There are, however, two instances where the performance of FAST is reduced.

Order One Model	Gaussian process Residuals		t-process Residuals		No Underlying Shape	
Anomaly Type	Power	ADD	Power	ADD	Power	ADD
Polynomial	0.96	35.30 (37.04)	0.37	62.03 (50.02)	0.99	6.92 (13.46)
Sinusoidal	0.83	25.69 (23.02)	0.21	31.45 (33.46)	0.99	12.93 (2.06)
Loss of Shape	1.00	45.71 (15.50)	0.97	86.76 (18.44)	0.03	41.88 (92.04)

Order Two Model	Gaussian process Residuals		t-process Residuals		No Underlying Shape	
Anomaly Type	Power	ADD	Power	ADD	Power	ADD
Polynomial	1.00	1.49 (12.72)	0.99	13.53 (23.53)	1.00	22.46 (6.67)
Sinusoidal	1.00	2.74 (4.74)	0.95	14.51 (15.20)	1.00	1.14 (0.75)
Loss of Shape	1.00	2.67 (10.44)	0.99	14.13 (13.93)	0.05	38.21 (95.84)

Table 4.4.2: Tables showing the detection power, and average detection delay (and standard deviation of detection delay) for each anomaly.

The first is where there is a polynomial or sinusoidal anomaly, and the residuals follow a Students's t-process. The second is the loss of shape anomaly when there is no underlying shape for the data. In the former case, the reason for this is due to the magnitude of the Student's t-process residuals obscuring the anomalous behaviour. That said, for a second order model these anomalies are detected, showing that FAST still works well for this data generating process if the model is correctly specified. In the latter case, the loss of shape anomaly is undetected because the underlying process itself only consists of noise, and so this anomaly does not represent a change in the underlying structure for the data.

4.5 Application to Internet Network Data

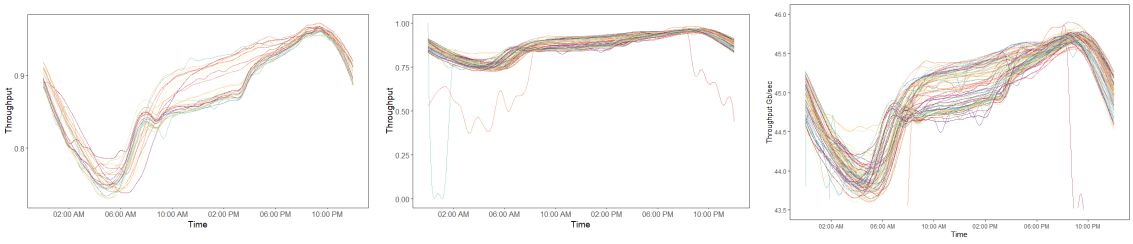
We now finally turn to consider the application that originally motivated this research. Here we use data taken from a testbed node somewhere on an internet network in the United Kingdom. The data measures the volume of traffic passing through the node, called throughput, each minute for the first four months of a given year. Fol-

lowing Ramsay (2005) we prepare the minute by minute data by smoothing it using a roughness penalty and cross-validation in conjunction with a system of 100 order 6 B-Splines. In Figure 4.5.1 the data for both January, and the remaining months, can be seen to follow an underlying shape. This shape can be attributed as follows:

(i) Lower demand for internet data during the early hours of the morning where many customers are asleep.

(ii) An increase in demand during the day as more consumers utilise the internet for both personal, and commercial, purposes.

Such behaviour is what causes the data to have a strongly repeated shape, and it is this feature that makes the dataset a candidate for our anomaly detection method.



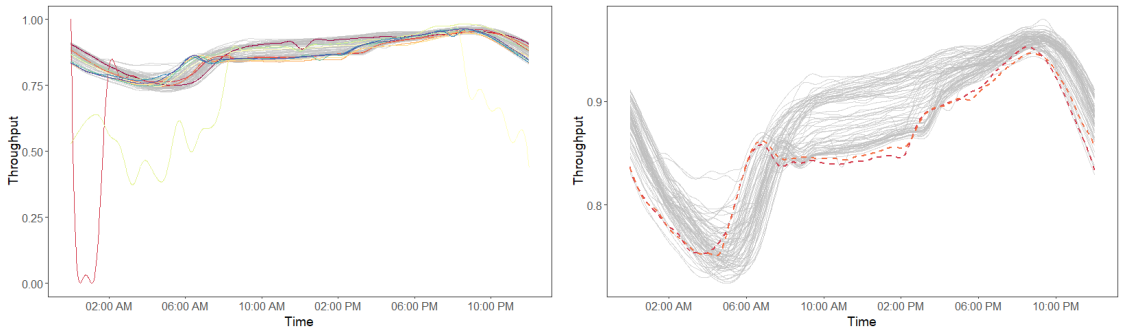
(a) (b) (c)

Figure 4.5.1: Training data taken from the first 30 days of observations (a); test data that the anomaly detection test will be performed on (b); and the test data with axis scaled to show a similar underlying shape to the training data (c).

We use FAST in this setting to detect curves that are statistically atypical and warrant further investigation with domain experts. In some instances these curves may well be within the range of normal operation, whilst in others may suggest anomalous behaviour.

To perform FAST, a training dataset consisting of the first 30 days of the dataset is used. To select the order of the differential operator to fit to the training data we use BIC, as described in Section 4.2.2, and this leads to an order four differential operator being fitted to the training data. FAST is carried out with the test threshold set to control the probability of false alarm at the 5% level. 11 alarms are declared,

and these instances are depicted in Figure 4.5.2(a). One interesting aspect of the results is that two of the days on which a detection takes place are on a public holiday weekend. Whilst not in themselves anomalous, these detection events may indicate different consumer behaviour over these days, and are depicted in Figure 4.5.2(b).



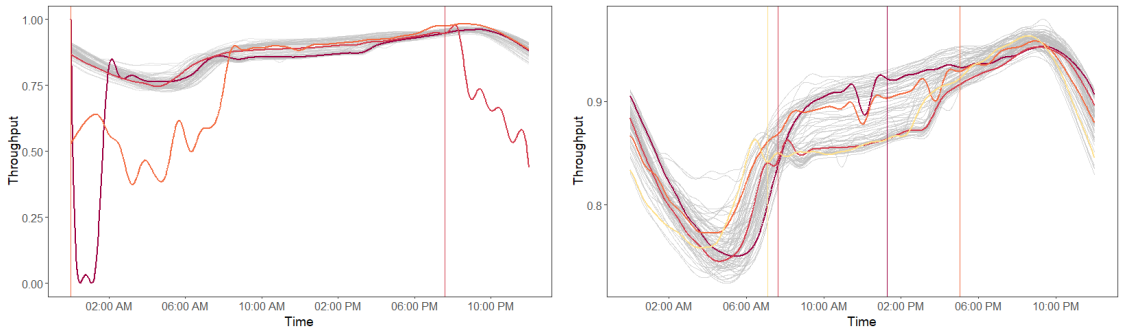
(a) (b)

Figure 4.5.2: Figure (a) illustrates the 11 detections returned by FAST, and (b) illustrates the two public holidays on which FAST identifies atypical behaviour. In both figures the grey lines represent the days where FAST does not detect anything.

FAST detects statistically atypical behaviour that reflects both a change in magnitude, and a change in shape, compared to the core underlying shape of the data. We show examples of such instances in Figures 4.5.3(a) and 4.5.3(b) respectively. Finally, we remark that the detection delay is less than than 40 minutes from when the atypical behaviour appears to start. This illustrates how FAST is able to respond quickly to these events as they emerge in this dataset.

4.6 Discussion

In this chapter we have introduced a new method for detecting the emergence of anomalous structures within functional data. This is done using FAST, a non-parametric CUSUM test with a threshold set to control the false alarm rate that takes into account the intra-functional dependence. The main advantage of this method is that it detects the emergence of smooth anomalies, rather than changes in a parameter as in other sequential detection tests. A second advantage is that, unlike other



(a) (b)

Figure 4.5.3: This figure illustrates a selection of the days on which FAST identifies the emergence of atypical behaviour. The vertical lines indicate the point where FAST first identifies this behaviour.

functional anomaly detection procedures, the sequential nature of our test suggests what part of the functional observation is anomalous; this increases the inferential power of the new method.

In Functional Data Analysis the typical dimension reduction approach uses Functional Principal Components Analysis (FPCA). This has been used in anomaly detection methods such as Hyndman and Shang (2010) and Aue et al. (2018). In our approach we instead estimate the low dimensional structure using PDA. One advantage of using PDA is that it allows for the underlying shape to consist of linearly independent functions, relaxing the condition of orthogonality in FPCA. A second advantage is that PDA does not require the covariance operator for the functional data to be stationary.

In addition to the methodological contributions made by this work, several theoretical results have been derived for FAST. One of these allows for a user to set a threshold that controls the rate of false alarms observed when no anomaly is present, and another provides bounds for the finite sample detection power of the test.

Simulation studies have been performed to further examine the finite sample performance of FAST. These studies have demonstrated that control of the probability of false alarms is possible using the threshold set out in Proposition 3. Furthermore, we have illustrated how FAST exhibits both high detection power, and low detection

delay, for a range of different settings including those where the assumptions of our model are not satisfied, or there are multiple underlying shapes for the data. Finally, FAST has attained good results in situations where the choice of ODE order is mis-specified.

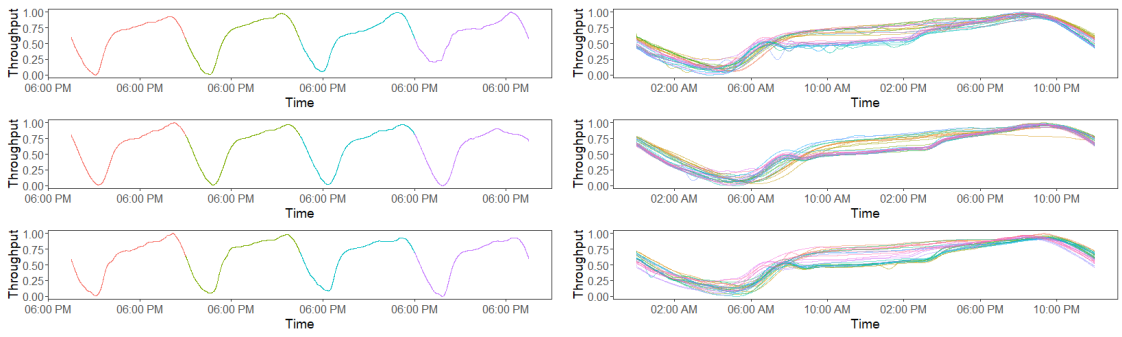
Although the performance of FAST on simulated data is encouraging, the motivation for this paper was a telecommunications anomaly detection challenge. We have demonstrated that FAST can detect atypical behaviours as they emerge in throughput data taken from an internet network. The fact FAST can detect such events in this setting is a primary contribution of the paper.

Chapter 5

Extending FAST to the Multivariate Setting

5.1 Introduction

In Chapter 4 we considered the problem of detecting emergent anomalies in univariate functional data. It is natural, perhaps, to consider the challenges involved in extending our approach to the multivariate setting. This is the focus of the work we now present.



(a) (b)

Figure 5.1.1: Four days of throughput data (a), and then four weeks of throughput data smoothed into a curve for each day (b), taken from three locations on the network.

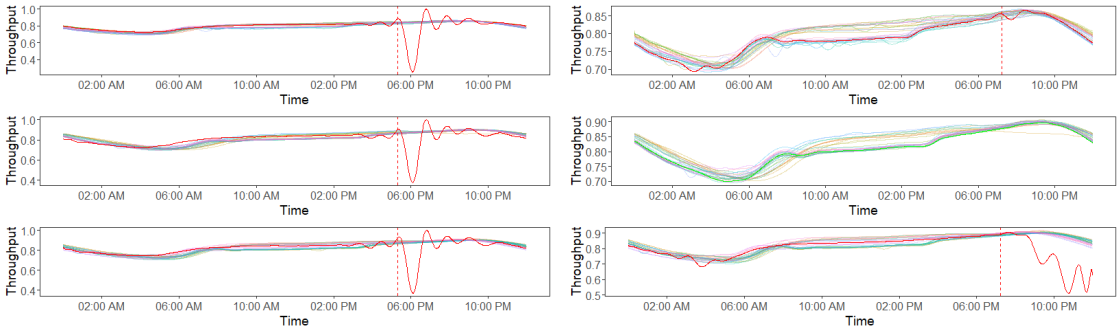
The motivation for our work stems from a challenge in the telecommunications sector where multiple points on a data network are monitored for potentially atypical

behaviour. As can be seen in Figure 5.1.1, the data at all locations has a repeated shape over the course of each day, and is recorded at a high frequency. In some instances, however, the observations deviate from this expected shape in either all, or a subset of, the dimensions of the data. In Figure 5.1.2 we present examples of such deviations. Identified atypical events are subsequently referred to a network specialist for further consideration.

Data of the form seen in Figure 5.1.1 are most naturally modelled as multivariate functional data. In recent years, several authors have considered the challenge of detecting anomalies in this setting. One approach to the problem utilises depth measures. See for example López-Pintado et al. (2014); Hubert et al. (2015); Rousseeuw et al. (2018); Dai and Genton (2019); Dai et al. (2020); and Qu and Genton (2022). Alternatives to the use of depth measures also exist, including the approaches proposed by Wang and Yao (2015); Lejeune et al. (2020a) and Archimbaud et al. (2022). Looking at the example in Figure 5.1.2, we see that it would be desirable for an anomaly detection method to incorporate the following features: (i) to detect the anomaly before the curve has been observed in full, (ii) to locate where on the curve the anomaly can first be identified and, (iii) to identify which of the dimensions can be considered anomalous. This means the existing methods are unsuitable for use in our setting, and a new approach is required.

From a modelling perspective, there is a further challenge that we might consider. As the data is drawn from a network, there is a dependency structure between the different dimensions of the data and it is desirable to model this structure. This is because it allows for instances where the behaviour of one location is unexpected with reference to the behaviour of other locations. Whilst current multivariate functional anomaly detection methods do not all require independence, they do not explicitly seek to capture the structure within the data and so do not provide the insight that we seek.

To incorporate features (i), (ii), and (iii) required by our data challenge we propose the multivariate Functional Anomaly Sequential Test (mvFAST). Central to our method is the modelling of the underlying shape for the data using the solution to



(a) (b)

Figure 5.1.2: Examples of atypical levels of throughput data in which: (a) all three locations experience the phenomena; or (b) where only a subset of the locations experience it. Note that in this example the atypical behaviour appears to occur at the same time.

an ODE system, and it is this system that allows us to capture relationships between different dimensions of the data. The differential operator defining this ODE system is then applied to an observation to obtain a residual function. This residual function is tested for anomalies using a CUSUM test with a two stage test threshold. As we shall see, the purpose of the two stage test threshold is to distinguish between anomalies affecting every dimension of an observation, and anomalies only affecting a subset of dimensions. Crucially for our particular application, the CUSUM test can be applied sequentially to a partially observed curve in order to detect anomalies as they emerge.

The remainder of this chapter is organised as follows: in Section 5.2 we explore the model for the functional data, and the methodology underpinning mvFAST. In Section 5.3 we turn to consider theoretical results for the test, establishing bounds on the detection power and proposing an approach for selecting test thresholds. Subsequently, in Section 5.4 the finite sample power of the test is examined in a range of settings. In particular, we consider (a) detection of both anomalies that contaminate every dimension, and (b) anomalies that contaminate a subset of the dimensions. In addition, we explore detection of anomalies in situations where there is trend in the underlying shape for the data, and illustrate the performance of the test when the order of the ODE has been mis-specified. Finally, we apply our method to detect

anomalies at key locations on a telecommunications network, before concluding with a discussion.

5.2 Methodology

In this section we outline the model that underpins our analysis of multivariate functional data, and discuss how the model parameters can be estimated from the data. We then present our anomaly detection test, mvFAST, demonstrating how it can be applied offline to detect anomalous observations within a sample of fully observed functional data. Additionally we consider its potential to sequentially detect anomalies within a new, partially observed, curve.

5.2.1 Model

Before presenting our model for the data, we first describe the functional data setting used in this chapter. We consider a time series of n p-dimensional functional observations, $\{\mathbf{X}_i(t)\}_{i=1}^n$, with $\mathbf{X}_i(t) = \{X_{ij}(t)\}_{j=1}^p$ and $t \in \mathcal{T}$. We assume that each $X_{ij}(t) \in L^2(\mathcal{T})$, and that $\mathcal{T}$ represents an interval of time. For our application the time interval is one day, and the p dimensions locations on a network. Hence it is convenient to consider $X_{ij}(t)$ as the observed value of the jth location at time t on the ith day.

As described in Section 1, each of the non-anomalous observations follow a repeated shape in each dimension. In order to model the p-dimensional set of repeated structures we utilise the approach of Ramsay (2005), representing the data as noisy realisations of the set of solutions to a p-dimensional, order m linear ODE. The order m linear differential operator for the ODE system is represented by $\boldsymbol{\mathcal{L}} = \{\mathcal{L}_i\}_{i=1}^p$. This operator describes a relationship between the mth derivative of the ith dimension of an observation with the derivatives up to order $m-1$ of every observed dimension. This gives rise to the following representation for each equation in the system:

$$\mathcal{L}_i = \sum_{j=1}^{p} \sum_{k=0}^{m-1} \beta_{i,j,k}(t) D_j^k + D_i^m, \quad 1 \leq i \leq p. \tag{5.2.1}$$

Here D_j^k represents the kth derivative of the jth variable of the observation, and $\{\beta_{i,j,k}(t)\}$ represents the functional coefficient for the kth derivative of the jth dimension for the equation for the ith dimension. The choice of m, and estimation of the coefficient functions $\{\beta_{i,j,k}(t)\}$, will be discussed in the next section.

Ramsay (1996) observes that if $\mathbf{S}(t)$ is a solution to the system, $\boldsymbol{\mathcal{L}}$, then each of the p dimensions can be represented in the following form:

$$S_j(t) = \sum_{k=1}^{m} c_{jk} u_{jk}(t), \quad 1 \leq j \leq p. \tag{5.2.2}$$

Here $\{c_{jk}\} \in \mathbb{R}$ and the $\{u_{jk}(t)\} \in L^2(\mathcal{T})$ are the linearly independent solution functions for the system. We can use the representation in (5.2.2) as the foundation of a model for our data. Indeed, under the assumption that each observation is a noisy realisation of the solution to a p-dimensional order m ODE we have the following:

$$X_{ij}(t) = \sum_{k=1}^{m} c_{jk} u_{jk}(t) + f_{ij}(t), \quad 1 \leq j \leq p. \tag{5.2.3}$$

Here $\{c_{jk}\}$, and $\{u_{jk}(t)\} \in L^2(\mathcal{T})$ are as above, and the $\{f_{ij}(t)\} \in L^2(\mathcal{T})$ are innovation functions.

To ensure that our model can be consistently estimated we make the following assumption:

Assumption 3. *(i) For each $1 \leq i \leq n$ the functions $\left\{D^k(\boldsymbol{X}_i)(t)\right\}_{k=0}^{m-1}$ are linearly independent and (ii) the innovation functions, $\{f_{ij}(t)\}$ are independent and identically distributed m-times differentiable mean-zero functions.*

The objective of this work is to detect anomalies that manifest as deviations from a common shape structure for the data. As described in the introduction, these anomalies can occur in either a subset of the dimensions of an observation (i.e. a sparse anomaly); or in all of the dimensions (i.e. a dense anomaly). We also allow for the anomalies to emerge in different dimensions of $\boldsymbol{X}_i(t)$ at different points in time. Letting $\{g_{ij}(t)\}$ be the contaminating function that drives the deviation from the underlying shape structure, we have the following model for the anomalies:

$$X_{ij}(t) = \begin{cases} \sum_{k=1}^{m} c_{jk} u_{jk}(t) + f_{ij}(t) & t \in \mathcal{T} \setminus \mathcal{S}_{ij} \\ \sum_{k=1}^{m} c_{jk} u_{jk}(t) + f_{ij}(t) + g_{ij}(t) & t \in \mathcal{S}_{ij}. \end{cases} \tag{5.2.4}$$

Here $\mathcal{S}_{ij}$ is the period in which the anomaly present in the jth dimension of the ith function is present. We make two assumptions regarding the anomaly functions to ensure (i) that the observations remain differentiable when contaminated by an anomaly, and (ii) that the anomaly functions are not solutions to $\boldsymbol{\mathcal{L}}$:

Assumption 4. *The anomaly functions $\{g_{ij}(t)\}$ are (i) m-times differentiable, and (ii) not equal to a linear combination of the $\{u_{jk}(t)\}$.*

A practical issue when working with functional data is that the observations will not be available in continuous time. To address this, many authors (e.g. Nagy et al. (2017)) have worked with the observations on a discretised grid. We let $\{1, \ldots, T\}$ be the grid that discretises the compact set $\mathcal{T}$, and denote the discretised observation for the jth location on the ith day by $\{X_{ij}(\tau)\}_{\tau=1}^{T}$.

We next turn attention to the estimation of our model for the data. Our anomaly detection test requires an estimate of the ODE operator, $\mathcal{L}$, for it to be applied to the observations. As we describe in the next section, our approach will make use of Principal Differential Analysis (PDA) (Ramsay, 1996).

5.2.2 Model Estimation Using Principal Differential Analysis

To detect anomalies, mvFAST will apply the p-dimensional order m ODE operator, $\boldsymbol{\mathcal{L}}$, to a set of observations. In practice the operator underpinning the ODE system will be unknown, and so needs to be estimated from the data. To do so, we use Principal Differential Analysis (Ramsay, 1996). PDA estimates the coefficients $\boldsymbol{\beta}(t) = \{\beta_{l,j,k}(t)\}$ as

$$\boldsymbol{\beta}(t) = \arg\min_{\beta_{l,j,k}} \sum_{i=1}^{n} ||\mathcal{L}(X_i)(t)||_2^2 + \lambda \sum_{l=1}^{p} \sum_{j=1}^{p} \sum_{k=1}^{m} ||\beta_{ljk}(t)||_2^2, \tag{5.2.5}$$

where $||\cdot||_2$ is the L^2 norm. Here $\lambda \geq 0$ is a parameter that controls the amount of penalisation that takes place and controls the smoothness of the estimates of the coefficients. Furthermore, Ramsay and Hooker (2017) discuss how increasing the value of the parameter prevents overfitting.

Following the approach of Ramsay (1996) the coefficients for each of the p equations of $\boldsymbol{\mathcal{L}}$; denoted by $\left\{\boldsymbol{\beta^{(l)}}\right\}_{l=1}^{p}$, $\boldsymbol{\beta^{(l)}} = \{\beta_{l,1,0}(t), \ldots, \beta_{l,1,m-1}(t), \beta_{l,2,0}(t), \ldots, \beta_{l,p,m-1}(t)\}$;

can be estimated on a discretised grid as follows

$$\boldsymbol{\beta}^{(\boldsymbol{l})}(t) = \left(\boldsymbol{Y}(t)^T\boldsymbol{Y}(t) + \lambda\boldsymbol{I}_{mp}\right)^{-1}\boldsymbol{Y}(t)^T\boldsymbol{z}^{(\boldsymbol{l})}(t).$$

In the above equation $\boldsymbol{Y}(t)$ is a matrix of dimension $n \times mp$ and each row of this matrix is given by $Y_{i,\cdot}(t) = \{\mathrm{D}^0(X_{i1}(t)), \ldots, \mathrm{D}^{m-1}(X_{i1}(t)), \mathrm{D}^0(X_{i2}(t)), \ldots, \mathrm{D}^{m-1}(X_{ip}(t))\}$, $\boldsymbol{z}^{(\boldsymbol{l})}(t) = \{D^m(X_{il})(t)\}_{i=1}^n$ is a vector of length n, and $\boldsymbol{I}_{mp}$ is the $mp \times mp$ identity matrix.

The purpose behind estimating the ODE system is not so that the solution functions, $\{u_{jk}(t)\}$, can be estimated but instead so that anomalies can be detected by applying the estimate of the differential operator for the ODE system to the observations. When $\mathcal{L}$ is applied to an observation we obtain residuals functions, $\mathcal{L}(X_{ij})(t) = \epsilon_{ij}(t)$. An explicit connection between the model for the data in equation (5.2.3) can also be derived as follows:

$$\begin{aligned}\mathcal{L}(X_{ij})(t) &= \mathcal{L}\left(\sum_{k=1}^{m} c_{jk}u_{jk}(t)\right) + \mathcal{L}(f_{ij})(t) \\ &= 0 + \epsilon_{ij}(t).\end{aligned}$$

On the other hand if an anomaly is present then we have

$$\begin{aligned}\mathcal{L}(X_{ij})(t) &= \mathcal{L}\left(\sum_{k=1}^{m} c_{jk}u_{jk}(t)\right) + \mathcal{L}(f_{ij})(t) + \mathcal{L}(g_{ij})(t) \\ &= 0 + \epsilon_{ij}(t) + \mathcal{L}(g_{ij})(t).\end{aligned}$$

In practice, we will apply the estimated operator obtained from equation (5.2.5) and in doing so obtain estimated residual functions, $\{\hat{\epsilon}_{ij}(t)\}$. It is these estimated residual functions that will be monitored by mvFAST.

Aside: ODE Order Selection Before we present our anomaly detection test, we provide a brief discussion on how to select the order, m, of the ODE system. This aspect of PDA has received limited attention in the literature, in particular for multivariate functional data. Ramsay and Hooker (2017) suggest that, for an unrelated ODE selection method, the structure of the ODE should be based upon minimising the residuals of the fitted model. Motivated by this argument, and the similar

challenge of selecting the order of a Vector Autoregressive (VAR) model (Lütkepohl, 1993), we propose to use the Bayesian Information Criterion (BIC) (Schwarz, 1978) to select the order of the ODE. Following Lütkepohl (1993) we select the order of the ODE using BIC as follows:

$$\text{BIC}(m) = \log\left(n^{-1}\text{SSE}_{\text{m}}\right) + \frac{\log(n)\, mp^2}{n}, \tag{5.2.6}$$

where $\text{SSE}_{\text{m}} = \sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{\tau=1}^{T}\epsilon_{ij}^2(\tau)$ is the sum of square error for the fitted ODE of order m. We remark that under the assumption that the innovation functions are mean-zero Gaussian processes then this will lead to consistent estimation of the true order for the data (Lütkepohl, 1993).

The reason we draw on the literature for VAR models is because the coefficient structure is identical in both the VAR and PDA settings, with the distinction between the two models being the fact that VAR uses a difference operator, not a differential operator.

With our model for the data in place, and estimation of the model from a sample of data discussed, we next introduce our anomaly detection test.

5.2.3 Anomaly Detection Test

Recall that the primary goal of this work is to detect both observations that deviate from a common shape structure, and to identify the location on the curve where this deviation first occurs. To do this, our proposed anomaly detection test estimates the shape structure using a system of differential equations and then monitors the residuals of the fitted system to identify deviations. Below we present our anomaly detection test in both the offline, where the curves have already been observed in full; and online setting, where a new curve is monitored as it emerges.

Offline Anomaly Detection

For both offline anomaly detection, and the testing of a new observation for an anomaly, the steps of mvFAST are similar. We begin with the offline testing problem and then describe how it can be adapted to the sequential testing environment.

Formally, the goal in the offline anomaly detection setting is to test the hypothesis

$$H_0 : \boldsymbol{\mathcal{L}}(g_{ij})(\tau) = 0$$
$$H_1 : \boldsymbol{\mathcal{L}}((g_{ij})(\tau) \neq 0$$

for at least one $1 \leq j \leq p$ and $1 \leq \tau \leq T$, for each $1 \leq i \leq n$.

Motivated by the fact that under the null $\mathcal{L}(X_{ij})(t) = \epsilon_{ij}(t)$, whereas under the alternative $\mathcal{L}(X_{ij})(t) = \epsilon_{ij}(t) + \mathcal{L}(g_{ij})(t)$, we propose to monitor the residual functions in order to detect the presence of anomalous behaviour. To do so we must obtain estimates of the residual functions, and so to this end we obtain an estimate of $\mathcal{L}$ using PDA. We then apply $\hat{\mathcal{L}}$ to each observation to generate estimates of the residual functions: $\{\hat{\epsilon_{ij}}(\tau)\}_{\tau=1}^{T}$.

To monitor these residual functions mvFAST uses a CUSUM statistic of the form

$$\Delta_{ij}(\tau) = \sum_{r=1}^{\tau} \hat{\epsilon}_{ij}(r), \qquad 1 \leq i \leq n,\, 1 \leq j \leq p,\, 1 \leq \tau \leq T. \tag{5.2.7}$$

A test statistic based on the CUSUM test is adopted, in contrast to other multivariate functional anomaly detection techniques, since it allows for the location on the curve of the anomaly's emergence to be identified.

Recalling that we seek to detect both dense and sparse anomalies we present a threshold regime that differentiates between both forms. To achieve this we follow an approach similar to that proposed by Tickle et al. (2021), utilising a two stage threshold for the test. In particular we consider a sparse threshold, γ to identify if, given the structure of the system, a single dimension of a curve is anomalous; and a dense threshold, Γ, to identify when every dimension of a curve contains an anomaly. Should a monitoring test statistic exceed γ, then we declare that individual series is anomalous. That is, the jth dimension of the ith observation is anomalous at time τ onwards if

$$|\Delta_{ij}(\tau)| > \gamma.$$

On the other hand, all p dimensions of the ith observation are declared as anomalous at time τ onwards if

$$\left| \sum_{j=1}^{p} \Delta_{ij}(\tau) \right| - p\gamma > \Gamma.$$

Using the two detection rules presented above, we obtain the following stopping time, ξ_{ij}, for the jth dimension of the ith location of the data:

$$\xi_{ij} = \min\left\{1 \leq \tau \leq T \ : \ |\Delta_{ij}(\tau)| > \gamma, \left|\sum_{j=1}^{p} \Delta_{ij}(\tau)\right| - p\gamma > \Gamma\right\}, 1 \leq i \leq n, 1 \leq j \leq p. \tag{5.2.8}$$

If neither threshold has been crossed then no anomaly has been identified. In this particular case, following Baron and Tartakovsky (2006) and Tartakovsky et al. (2014), we declare $\min\{\emptyset\} = \infty$.

The motivation behind the use of a two stage threshold for the detection test is that it allows for determination of whether a subset of the dimensions, or every dimension, is anomalous. Whilst the sparse threshold, γ, identifies anomalous behaviour in any particular dimension, the dense threshold, Γ, is used to signal that, when taken as a whole, the data is sufficiently anomalous to suggest that every dimension is contaminated with an anomaly function.

In some scenarios, the detection of only one form of anomaly will be desired. In these cases, only one of the two test statistics should be utilised. In the sparse case, this will result in a stopping time of

$$\xi_{ij} = \min\{1 \leq \tau \leq T \ : \ |\Delta_{ij}(\tau)| > \gamma\}, 1 \leq i \leq n, 1 \leq j \leq p, \tag{5.2.9}$$

whilst in the dense case this stopping time will be

$$\xi_{ij} = \min\left\{1 \leq \tau \leq T \ : \ \left|\sum_{j=1}^{p} \Delta_{ij}(\tau)\right| - p\gamma > \Gamma\right\}, \ 1 \leq i \leq n, 1 \leq j \leq p. \tag{5.2.10}$$

In Section 5.3, results will be presented for choosing a threshold in each of the settings in order to control the family-wise error rate.

Algorithm 4: mvFAST Online Algorithm

Inputs: Observed data $\boldsymbol{X_1}(t), \ldots, \boldsymbol{X_n}(t)$, a new observations $\boldsymbol{X_{n+1}}(t)$, the ODE order m, and the thresholds γ and Γ. ;

Outputs: $\mathcal{C}$, a $1 \times p$ matrix to record the anomalous dimensions of the new observation; and $\mathcal{D}$, a $1 \times p$ matrix to record the detection times. ;

Initialise each entry of $\mathcal{C}$ to be $FALSE$ and each entry of $\mathcal{D}$ to be ∞. ;

Compute $\hat{\boldsymbol{\mathcal{L}}}$ from the training data $\boldsymbol{X_1}(t), \ldots, \boldsymbol{X_n}(t)$. ;

Initialise $\boldsymbol{\Delta_{n+1}}(0) = 0$;

for r *in* $1 : T$ **do**

  Apply $\hat{\boldsymbol{\mathcal{L}}}$ to the new observation $\boldsymbol{X_{n+1}}(r)$ to obtain $\boldsymbol{\epsilon_{n+1}}(r)$. ;

  $\boldsymbol{\Delta_{n+1}}(r) = \boldsymbol{\Delta_{n+1}}(r-1) + \boldsymbol{\epsilon_i}(r)$;

  if $|\sum_{j=1}^{p} \Delta_{n+1,j}(\tau)| - p\gamma > \Gamma$ **then**

    $\mathcal{C}[1, 1:p] = TRUE$;

    $\mathcal{D}[1, 1:p] = r$;

    break;

  end

  else if $\max_j |\Delta_{n+1,j}(r)| > \gamma$ **then**

    for j *in* $1 : p$ **do**

      if $|\Delta_{n+1,j}(r)| > \gamma$ **then**

        $\mathcal{C}[1, j] = TRUE$;

        $\mathcal{D}[1, j] = r$;

      end

    end

    break;

  end

end

if *No Anomaly Detected* **then**

  Update $\boldsymbol{X_1}(t), \ldots, \boldsymbol{X_n}(t)$ to $\boldsymbol{X_2}(t), \ldots, \boldsymbol{X_{n+1}}(t)$;

end

return $\mathcal{C}$ and $\mathcal{D}$

Online Anomaly Detection

We next turn our attention to the detection of anomalies sequentially within a new, partially observed, curve as it emerges. That is, given that we have observed n curves so far, we observe the $(n+1)$th curve $\boldsymbol{X}_{n+1}(1), \boldsymbol{X}_{n+1}(2), \ldots, \boldsymbol{X}_{n+1}(T)$ sequentially. In the online setting this corresponds to testing, at each time $1 \leq \tau \leq T$, the hypothesis

$$H_0 \ : \ \boldsymbol{\mathcal{L}}(g_{n+1,j})(\tau) = 0$$
$$H_1 \ : \ \boldsymbol{\mathcal{L}}(g_{n+1,j})(\tau) \neq 0$$

for at least one $1 \leq j \leq p$.

Similar to in the offline setting, $\hat{\mathcal{L}}$ is applied to the new observation as each new point on the curve is received. That is, for each $1 \leq \tau \leq T$ we apply $\hat{\boldsymbol{\mathcal{L}}}$ to $\boldsymbol{X}_{n+1}(\tau)$ and compute the test statistic $\boldsymbol{\Delta}_{n+1}(\tau)$ using equation (5.2.7). The decision as to whether the new observation contains an anomaly can then be made using a similar stopping time to the offline setting:

$$\xi_{n+1,j} = \min \left\{ 1 \leq r \leq \tau \ : \ |\Delta_{n+1,j}(r)| > \gamma, \left| \sum_{j=1}^{p} \Delta_{n+1,j}(r) \right| - p\gamma > \Gamma \right\}, \ 1 \leq j \leq p.$$

To sequentially monitor the partially observed curve for anomalies as it emerges over time we repeat the process iteratively. This approach is used to analyse the telecommunications data in Section 5.5.

One distinction between the offline and online problem is that an estimate of $\mathcal{L}$ must be obtained before the online analysis can begin. To facilitate this the estimate is obtained from some anomaly-free training data $\{\boldsymbol{X}_i(t)\}_{i=1}^{n}$. Using this training data, an estimate of operator $\hat{\mathcal{L}}$ is obtained using PDA as in the offline setting.

Another distinction between the offline and online detection challenge is that, in some situations, the underlying shape for the data may change as more complete curves are observed. In this case, the sequential version of mvFAST should be adapted such that, if the most recent observation $\boldsymbol{X}_{n+1}(t)$ does not contain an anomaly, then it is substituted into the training data in place of $\boldsymbol{X}_1(t)$. An updated estimate for the differential operator for the system can then be computed using PDA. In Algorithm 4 we present pseudocode for the sequential form of mvFAST.

5.3 Theoretical Results

We next turn to consider some key theoretical properties of mvFAST. These include thresholds that can be employed to control the false alarm rate at a given level, bounds for the detection power of the test, and a proof of the asymptotic consistency of the location where the anomaly is detected on the curve. Results are provided for both the dense, and sparse, anomaly scenarios and apply to both offline, and online, mvFAST.

We begin by exploring how the threshold for mvFAST can be chosen to control the family-wise error rate (FWER). The FWER is defined as

$$\mathrm{P}\left(\bigcup_{i=1}^{n}\bigcup_{j=1}^{p}\xi_{ij} < \infty \,\middle|\, \text{No Anomaly}\right),$$

i.e. the probability of incorrectly identifying one or more of the p dimensions of an observation as anomalous. Typically we seek to control this at some given level α. That is, the proportion of the n observations where one or more of the p series is incorrectly identified as anomalous does not exceed α. The following result describes the selection of the threshold so that the FWER is controlled for each of the three stopping times (equations (5.2.8) - (5.2.10)) discussed in Section 5.3.

Proposition 6. *Let Assumption 3 hold. Then the following choices for γ and Γ bound the FWER above by α:*

1. *Sparse Anomalies Only (Equation* (5.2.9)*)*

$$\gamma = \left(\frac{np\left(\sum_{\tau=1}^{T} V_\tau\right)}{\alpha}\right)^{\frac{1}{2}}$$

2. *Dense Anomalies Only (Equation* (5.2.10)*)*

$$\Gamma = 2\left(\left[\frac{np\left(\sum_{\tau=1}^{T} V_\tau\right)}{\alpha}\right]^{\frac{1}{2}} - p\gamma\right), \text{ for any } \gamma < \left(\frac{n\sum_{\tau=1}^{T} V_\tau}{p\alpha}\right)^{\frac{1}{2}}.$$

3. Both Dense and Sparse Anomalies Simultaneously (Equation (5.2.8)*)*

$$\gamma = \left(\left[\frac{np\left(\sum_{\tau=1}^{T} V_\tau\right)}{\alpha} \right] \left[\frac{(p+\log(p))^2 + 1}{(p+\log(p))^2} \right] \right)^{\frac{1}{2}}$$
$$\Gamma = \gamma \log(p)$$

where $V_\tau = \mathrm{Var}\left(\Delta_{ij}(\tau)\right)$.

Proof. See appendix. □

The next result concerns the power of the detection test to detect a change by time T. To this end let $D = \{\xi \leq T \,|\, g_{ij}(\tau) \neq 0, \forall 1 \leq i \leq n, 1 \leq j \leq p, 1 \leq \tau \leq T\}$ be the event that a detection has taken place by time T given an anomaly is present. Equipped with this we are in a position to state our result for the bounds on the probability of detecting an anomaly:

Proposition 7. *Assume that the ith observation is contaminated with a p-dimemsional anomaly function* $\{g_{ij}(t)\}_{j=1}^{p}$. *Furthermore, assume that for any* $1 \leq \tau \leq T$, $\mathrm{E}(\epsilon_{ij}(\tau)^3) < \infty$. *Then, under Assumption 3, the probability of detecting the anomaly by time* T *is bounded by*

$$\frac{-2\mu_1}{K} + \frac{11\mu_2}{4K^2} - \frac{3\mu_3}{4K^3} \leq \mathrm{P}(D)$$
$$\leq \sum_{\tau=1}^{T} \left[\frac{pV_\tau}{\left(\Gamma + p\gamma - |\sum_{j=1}^{p} \mathcal{L}(G_{ij})(\tau)|\right)^2} + \frac{V_\tau^p}{\min_{1\leq j\leq p}\left(\gamma - |\mathcal{L}(G_{ij})(\tau)|\right)^2} \right].$$

Here $K = \Gamma + p\gamma + |\sum_{j=1}^{p} \mathcal{L}(G_{ij})(T)|$, $\mu_1 = \mathrm{E}\left(|\sum_{j=1}^{p} \Delta_{ij}(T)|\right)$, $\mu_2 = \mathrm{E}\left[\left(|\sum_{j=1}^{p} \Delta_{ij}(T)|\right)^2\right]$, $\mu_3 = \mathrm{E}\left[\left(|\sum_{j=1}^{p} \Delta_{ij}(T)|\right)^3\right]$, V_τ *is defined as in Proposition 6, and* $\mathcal{L}(G_{ij})(\tau) = \sum_{r=1}^{\tau} \hat{\mathcal{L}}(g_{n+1})(r)$.

Proof. See appendix for details. □

The previous results are concerned with the performance of mvFAST on a finite grid. It is perhaps also natural to question to ask how the test performs as the fineness of the grid increases. The next pair of results seeks to show the conditions under which

an anomaly is guaranteed to be detected as $T \to \infty$. Further, we provide a result that connects this guarantee to the detection delay of the test.

Our first result is concerned with the probability of detecting an anomaly as $T \to \infty$. In particular, we continue on from a step in the proof of the lower bound in Proposition 7 and show that if $\lim_{T\to\infty} \frac{|\sum_{j=1}^{p} \mathcal{L}(G_{ij})(T)|}{T}$ is bounded away from zero then we are guaranteed to detect the anomaly:

Corollary 2. *Let the p-dimensional anomaly function* $\{g_{ij}(t)\}_{j=1}^{p}$ *emerge at a time* $\nu = \lfloor \theta T \rfloor$, $0 < \theta < 1$, *and assume that* $\lim_{T\to\infty} \frac{|\sum_{j=1}^{p} \mathcal{L}(G_{ij})(T)|}{T} = M$ *for some* $M > 0$. *Then, if Assumption 3 holds, and the residual functions,* $f_{ij}(t)$, *are independent in* t, $\lim_{T\to\infty} \mathrm{P}(D) = 1$

Proof. See Appendix. □

This corollary can also be used to provide a proof of the consistency of mvFAST subject to the same assumptions on the anomaly and innovation functions. In the changepoint literature the consistency of a changepoint estimator refers to the performance of the estimator as the size of the sample, n, tends to infinity. In particular, if $\hat{\tau}$ is an estimator of the changepoint τ then $\hat{\tau}$ is a consistent estimator if $\frac{|\hat{\tau}-\tau|}{n} \to 0$ as $n \to \infty$ (Truong et al., 2020). We apply this concept to our setting by seeking to show that if ξ is the stopping time for the mvFAST test, and ν is the true changepoint, then as $T \to \infty$, $\frac{\xi-\nu}{T} \to 0$. This is a desirable property, indicating that the detection delay of the test does not increase at the same rate as the fineness of the grid. The following proposition illustrates when this property holds.

Proposition 8. *Consider observations over the interval* $\mathcal{T} = [0, 1]$, *and with this interval discretised into* T *points with the sequence* $\{t_i\}_{i=0}^{T}$, $t_0 = 0$, *and* $t_T = 1$. *Let the p-dimensional anomaly function* $\{g_{ij}(t)\}_{j=1}^{p}$ *emerge at a time* $\nu = t_{\lfloor \theta T \rfloor}$, $0 < \theta < 1$. *Furthermore, let* $A_\epsilon = \left\{t_{\lfloor \theta T \rfloor}, t_{\lfloor \theta T \rfloor} + a_1 \ldots, t_{\lfloor \theta T \rfloor} + a_m\right\}$, *where* $a_m < \epsilon$ *for some* $0 < \epsilon < 1 - \theta$ *and any* $m \in \mathbb{N}$.

Then if $\lim_{m\to\infty} \frac{|\sum_{j=1}^{p} \mathcal{L}(G_{ij})(t_{\lfloor \theta T \rfloor} + a_m)|}{m} > 0$, *Assumption 3 holds, and the residuals functions are independent in* t, *as* $T \to \infty$, $\frac{\xi-\nu}{T} \to 0$.

Proof. See appendix. □

In addition to the dense setting, each of the results presented in this section can be trivially extended to the sparse setting where the anomaly function only emerges in a subset, $\mathcal{J}$, of the p dimensions. This simply requires a modification of the anomaly functions by setting $g_{ij}(t) = 0$ for each $j \notin \mathcal{J}$ and every $1 \leq t \leq T$. Doing so ensures that the conclusions of Propositions 7 and 7, and Corollary 2 still hold, so long as the appropriate assumptions for each result are satisfied.

5.4 Simulation Studies

With key theoretical properties established, we next consider a suite of simulations to study the performance of mvFAST in a finite sample setting. We begin by examining the power and detection delay when detecting several different anomalies for a range of data generating processes. Then, we will explore how mvFAST performs when there is a mis-specification of the ODE order, or the underlying shape for the data is non-stationary.

In Sections 5.4.1 and 5.4.2, five different forms of data generating processes are considered over the interval $\mathcal{T} = [1, 300]$. Namely:

(i) **Single Underlying Shape:** $X_{ij}(t) = \sin\left(\frac{2\pi t}{100}\right) + \cos\left(\frac{2\pi t}{200}\right) + f_{ij}(t)$.

(ii) **Multiple Underlying Shapes:** $X_{ij}(t) = a_{ij}\sin\left(\frac{2\pi t}{100}\right) + b_{ij}\cos\left(\frac{2\pi t}{200}\right) + f_{ij}(t)$, $a_{ij}, b_{ij} \sim \text{U}[0.8, 1.2]$.

(iii) **Heavy Tailed Innovation Functions:** $X_{ij}(t) = a_{ij}\sin\left(\frac{2\pi t}{100}\right) + b_{ij}\cos\left(\frac{2\pi t}{200}\right) + \tilde{f}_{ij}(t)$, $a_{ij}, b_{ij} \sim \text{U}[0.8, 1.2]$,

(iv) **Correlated Innovation Functions:** $X_{ij} = \sin\left(\frac{2\pi t}{100}\right) + \cos\left(\frac{2\pi t}{200}\right) + f_{ij}(t)$, with $f_{ij}(t)$ correlated between different observations.

(v) **Bimodal Underlying Shape:**

$$X_{ij}(t) = \begin{cases} a_{ij}\sin\left(\frac{2\pi t}{100}\right) + b_{ij}\cos\left(\frac{2\pi t}{200}\right) + f_{ij}(t) & \text{if } p_{ij} = 1 \\ 3 + a_{ij}\sin\left(\frac{2\pi t}{100}\right) + b_{ij}\cos\left(\frac{2\pi t}{200}\right) + f_{ij}(t) & \text{if } p_{ij} = 0. \end{cases}$$

$a_{ij}, b_{ij} \sim \text{U}[0.8, 1.2]$ and $p_{ij} = \text{Bernoulli}\left(\frac{1}{2}\right)$.

In the above settings $\{f_{ij}(t)\}$ are Gaussian processes with squared exponential covariance kernel, $K(s,t) = 0.2\exp\left(\frac{1}{2}\left(\frac{s-t}{20}\right)^2\right)$, and $\left\{\tilde{f}_{ij}(t)\right\}$ are Student's t-processes with 5 degrees of freedom and scale matrix parameterised by the same covariance kernel, $K(s,t)$.

The first two data generating processes enable analysis of mvFAST when the assumptions of our model are met. On the other hand the heavy tailed, and correlated, innovation function settings allow us to examine performance when the model approach is somewhat mis-specified. Finally, the bimodal underlying shape corresponds to the situation where no single model for the ODE exists, and there is more than one underlying shape for the data.

As part of our study we will seek to assess mvFAST on shape, magnitude, and phase anomalies. We thus consider seven different anomaly functions. Furthermore, we consider both dense, and sparse, settings. The seven instances we consider are as follows, with each anomaly function emerging over the interval $100 \leq t \leq 200$:

(i) **Single Shape Anomaly:** $g_{ij}(t) = c_i \cos\left(\frac{2\pi t}{30}\right) + d_i \cos^2\left(\frac{2\pi t}{200}\right)$, $c_i, d_i \sim \text{U}[1,3]$.

(ii) **Multiple Shape Anomaly:** $g_{ij}(t) = c_{ij} \cos\left(\frac{2\pi t}{30}\right) + d_{ij} \cos^2\left(\frac{2\pi t}{200}\right)$, $c_{ij}, d_{ij} \sim \text{U}[1,3]$.

(iii) **Polynomial Shape Anomaly:** $g_{ij}(t) = c_{ij}\left(\frac{t}{100}\right)^{\frac{3}{2}}\left(1 - \frac{t}{100}\right)$, $c_{ij} \sim \text{U}[1,3]$.

(iv) **Constant Magnitude Anomaly:** $g_{ij}(t) = c_{ij}$, $c_{ij} \sim \text{U}[1,5]$.

(v) **Exponential Magnitude Anomaly:** $g_{ij}(t) = c_{ij}\exp\left(\frac{t}{100}\right)$, $c_{ij} \sim \text{U}(0.8, 1.2)$.

(vi) **Phase Anomaly:** $g_{ij}(t) = -a_{ij}\sin\left(\frac{2\pi t}{100}\right) - b_{ij}\cos\left(\frac{2\pi t}{200}\right) + \sin\left(\frac{2\pi t}{c_{ij}}\right) + \cos\left(\frac{2\pi t}{d_{ij}}\right)$, $c_{ij}, d_{ij} \sim \text{Discrete U}[20, 80]$.

(vii) **Consistency Anomaly:** $g_{ij}(t) = 3f_{ij}(t)$ (Dai and Genton, 2019).

We remark that the constant magnitude anomaly, unlike the other scenarios, also represents a setting where the anomaly function does not satisfy the required differentiability conditions. We also remark that the consistency anomaly, proposed by Dai

and Genton (2019), is the situation where the anomaly function is a multiple of the innovation functions. Such a scenario corresponds to when additional noise is added to the underlying shape for the data. mvFAST should not detect this as an anomaly as the underlying shape structure is unchanged. As such, this study examines if mvFAST is susceptible to declaring a false alarm when there has been an increase in the noisiness of different data generating processes.

To assess the comparative strengths of our new method, this study proposes a comparison of the detection power of mvFAST with the detection power of the method of Dai and Genton (2019). We implement Dai and Genton (2019) using the R package fdaOutlier (Ojo et al., 2021b). Further, in order to improve power, we first apply the pointwise outlier transformation discussed in Dai et al. (2020) to the observations. Although the method of Dai and Genton (2019) does not aim to detect the location of an anomaly in partially observed multivariate functional data, or identify which dimension is anomalous, it is still possible to compare the two methods. To ensure a fair comparison, we allow the competitor method to see all the points on the curve from the start, rather than observing the curve sequentially as mvFAST does. Furthermore, we only require the competitor to identify the anomalous observation, and not the particular dimension containing the anomaly. This is different to with mvFAST, where a successful detection requires that the correct dimensions are identified.

5.4.1 Anomaly Detection Power

In this section, we consider the power of the mvFAST detection test on each of the data generating processes and anomalies defined above. In all simulations, we have 50 observations with number of dimensions $p = 10$. We perform 500 repetitions of each setting, control the FWER to be 5%, and for mvFAST fit the order $m = 2$ differential operator.

Table 5.4.1 presents the results for the dense setting. In general, mvFAST performs well even in cases where the underlying assumptions of the model are not satisfied (columns 3, 4, and 5). However, it is clear that mvFAST is unable to detect phase anomalies (row 6). One reason for this may be because the ODE structure that

would be estimated from a sample of anomalous observations is similar to the ODE structure that is estimated from non-anomalous data. This makes detection of the anomaly using mvFAST difficult.

The results for the sparse scenarios are presented in Tables 5.4.3. The results are similar to the dense case, although power is weaker in some cases owing to the smaller number of dimensions that have been contaminated. Note that, as in the dense case, the performance with respect to phase anomalies is poor (row 6). This may be for similar reasons as described in the previous praragraph.

In terms of comparison with the method of Dai et al. (2020), our method performs competitively in the majority of settings. This is particularly encouraging given that the method of Dai et al. (2020) has access to all of the observation for the detection test from the start, whereas mvFAST does not. In situations where depth measures are less suitable, for instance when the observations are heavy-tailed (column 3), we also see that mvFAST outperforms the method of Dai et al. (2020). A final remark that can be made regarding the differences between mvFAST and the competitor method is that, unlike mvFAST, the competitor identifies consistency anomalies. This is as expected, however, because the method of Dai et al. (2020) seeks to detect such anomalies whereas our method considers them a false alarm because there is no change to the underlying shape for the data.

Whilst the detection of the anomalies is the key aim of mvFAST, a further performance metric that can be considered is the time taken to detect the anomalies given they emerge at time 100. The detection delay results are presented in the next section.

5.4.2 Average Detection Delay

We next explore the detection delay for each of the data generating processes and anomalies, for the three different sparsity settings. In order to assess the detection delay performance, we use the definition of the conditional average detection delay

Process	Single Shape		Multiple Shape		Heavy Tailed		Mixing Innovations		Bimodal Shape	
Anomaly	mvFAST	Dai	mvFAST	Dai	mvFAST	Dai	mvFAST	Dai	mvFAST	Dai
Single	**1**	**1**	**1**	**1**	0.99	0.11	0.86	**1**	**1**	0.97
Multiple	**1**	**1**	**1**	**1**	**0.99**	0.1	0.89	**1**	**1**	0.98
Polynomial	**1**	**1**	**1**	**1**	**0.91**	0.56	0.98	**1**	**1**	0.99
Phase	**0.41**	0.39	0.31	**0.33**	**0.1**	0.07	0.48	**1**	**1**	**1**
Constant	**1**	**1**	**1**	**1**	**0.98**	0.83	0.91	**1**	**1**	**1**
Exponential	**1**	**1**	**1**	**1**	**0.99**	**0.99**	0.99	**1**	**1**	**1**
Consistent	**0.06**	1	**0.08**	1	**0.11**	0.48	**0.39**	1	**0.05**	0.99

Table 5.4.1: Comparison of detection power of mvFAST and the sequential transform and directional outlyingness approach (Dai) of Dai et al. (2020) in the dense setting. The better result in each scenario is highlighted in **bold**.

Process Anomaly	Single Shape	Multiple Shape	Heavy Tail	Correlated	Bimodal Shape
Single Shape	5.70 (3.78)	6.41 (4.02)	13.58 (13.82)	29.93 (46.20)	6.04 (3.40)
Multiple Shape	7.42 (1.28)	7.65 (1.36)	14.09 (13.95)	46.91 (58.92)	5.41 (2.25)
Polynomial	38.79 (12.53)	43.51 (14.97)	55.99 (27.41)	77.31 (50.54)	105.70 (10.74)
Phase	29.40 (44.65)	24.93 (44.16)	9.37 (31.23)	33.76 (64.27)	28.40 (48.42)
Constant Magnitude	60.37 (3.34)	0.36 (3.60)	23.94 (28.98)	52.15 (57.92)	32.74 (15.20)
Exponential Magnitude	70.33 (3.07)	0.49 (3.88)	23.21 (27.98)	41.37 (51.93)	27.99 (11.91)
Consistent	8.33 (33.97)	8.95 (33.52)	12.55 (39.82)	38.12 (78.13)	9.14 (40.09)

Table 5.4.2: Average detection delay (and standard deviation) of mvFAST in the dense setting.

Process Anomaly	Single Shape		Multiple Shape		Heavy Tailed		Mixing Innovations		Bimodal Shape	
	mvFAST	Dai	mvFAST	Dai	mvFAST	Dai	mvFAST	Dai	mvFAST	Dai
Single	**0.95**	0.79	**0.95**	0.8	**0.81**	0.13	0.94	**1**	**0.88**	0.12
Multiple	**0.95**	0.78	**0.95**	0.78	**0.84**	0.15	0.95	**1**	**0.85**	0.1
Polynomial	0.93	**0.94**	0.93	**0.94**	**0.64**	0.25	0.99	**1**	**0.93**	0.61
Phase	0.01	**0.06**	0.02	**0.05**	0.01	**0.04**	0.43	**0.99**	0.01	**0.05**
Constant	0.85	**0.88**	**0.88**	**0.88**	**0.7**	0.6	0.67	**1**	0.25	**0.37**
Exponential	**1**	**1**	**1**	**1**	**0.99**	0.05	0.71	**1**	0.67	**0.98**
Consistent	**0.23**	0.62	**0.21**	0.64	**0.13**	0.24	**0.83**	0.98	**0.03**	0.24

Table 5.4.3: Comparison of detection power of mvFAST and the sequential transform and directional outlyingness approach (Dai) of Dai et al. (2020) in the sparse setting. The better result in each scenario is highlighted in **bold**.

Process Anomaly	Single Shape	Multiple Shape	Heavy Tail	Correlated	Bimodal Shape
Single	42.00 (16.67)	44.66 (17.55)	53.59 (37.37)	59.46 (61.32)	22.26 (22.44)
Multiple	51.48 (19.04)	53.15 (19.67)	78.71 (38.94)	51.32 (79.76)	54.33 (42.48)
Polynomial	28.97 (18.01)	33.96 (23.00)	45.49 (26.26)	1.78 (16.80)	26.88 (34.27)
Phase	2.35 (20.63)	0 (0)	0 (0)	23.96 (59.05)	0 (0)
Constant	133.06 (35.38)	114.50 (40.43)	1.58 (15.51)	78.73 (85.52)	0 (0)
Exponential	111.36 (4.50)	113.61 (12.39)	1.67 (13.67)	133.19 (87.42)	119.91 (21.95)
Consistent	0 (0)	0 (0)	0 (0)	9.21 (41.55)	0 (0)

Table 5.4.4: Average detection delay (and standard deviation) of mvFAST in the sparse setting.

given by Tartakovsky et al. (2014):

$$\text{ADD} = \text{E}\left(\xi - \nu | \xi \geq \nu\right),$$

where ξ is the stopping time for the test. That is, we report the average detection delay given that the detection occurs after the anomaly has emerged at time $\nu = 100$. The results for the detection delay in the dense case are presented in Table 5.4.2, and the sparse case in Table 5.4.4. No comparison with Dai (2020) is given as their method is not designed to work with partially observed curves.

The results for the study indicate that mvFAST is quick to identify shape anomalies in each of the settings considered. The magnitude anomalies, however, are detected more slowly. The reason for this is that the shape anomalies incorporate a change in each of the derivatives of an observations, whereas the magnitude anomalies only change the first term, with the higher order derivatives unchanged. As such, the contribution of the anomaly to the residual function is smaller than with a shape anomaly, and as such the time taken to exceed the threshold is greater.

A final point to make about the simulations in this section and Section 5.4.1 is that they have considered when either all the dimensions, or one dimension, of an observation contain an anomaly. Further results have also been obtained for when 5 of the 10 dimensions are contaminated. These results are similar to those for the sparse setting presented in Tables 5.4.3 and 5.4.4 and so are included in Appendix C.2.

5.4.3 Performance Under Model Mis-Specification

A primary assumption of mvFAST is that applying the estimated differential operator to an observation removes the underlying shape, and so the test statistic under the null hypothesis contains only an innovation function. For this to occur, however, it is necessary to select the correct order for the ODE system. If the model for the data is underfit, then not all of the shape structure will be captured and the test statistic may contain more than just noise; if the model is overfit then anomalous behaviour may be captured by the model and so remain undetected.

Process	Underfit		Overfit	
Anomaly	Dense	Sparse	Dense	Sparse
Single	0.77	0.13	1	1
Multiple	0.92	0.2	1	1
Polynomial	1	1	1	1
Phase	0.09	0.01	1	0.95
Constant	0.96	0.45	1	1
Exponential	1	1	1	1
Consistent	0.31	0.07	0.2	0.93

Table 5.4.5: Detection power of mvFAST when mis-specifying the order of the ODE model for the data.

To enable the finite sample performance of mvFAST to be studied in this setting, we perform mvFAST on the multiple underlying shape data generating process (scenario (ii) in Sections 5.4.1-5.4.2). We test each of the seven anomalies, for both the dense and sparse case. In each of the simulations, we have 50 observations with number of dimensions $p = 10$. We perform 500 repetitions of each setting; fit both an order $m = 1$, and $m = 3$, differential operator; and control the FWER to be 5%. The results for these simulations are presented in Table 5.4.5. The results show that power is only lost when underfitting and that this loss is only substantial in the sparse case. The reason the dense case is only marginally affected is attributed to the fact that combining the test statistic across the dimensions enhances the power of the test, allowing for detection even in this challenging situation.

So far, we have made the assumption that the underlying shape for the data is stationary over time. Below, however, we will consider a setting where this is not the case. We shall first discuss how, where necessary, mvFAST can be adapted to handle such non-stationarity before then presenting simulations to explore the efficacy of our proposed approach.

5.4.4 Modelling Trend in the Data

Whilst in some applications the underlying shape for each dimension of $\{\boldsymbol{X}_i(t)\}$ will remain constant, in others the shape will evolve over time. This is the case for example in our telecommunications setting, where evolving user behaviours lead to a small upward trend in the data over time. In this section we consider two different cases: the first is where there is additive trend, and the second is where there is a scalar multiplicative trend. As we shall see, additive trend affects the estimation of the differential operator and so needs removing before this estimation takes place, whereas a scalar multiplicative trend does not.

We are not the first authors to consider the problem of estimating a differential operator when there is a underlying trend in the observed data stream. Indeed, both Hooker (2009) and Hooker and Ellner (2015) discuss estimation of an unknown ODE structure that may also contain an additional unknown function, $h(t)$, termed a forcing function, and this forcing function may be subject to a trend. Unfortunately, the estimate of this trend would be contaminated by the presence of the innovation functions, $f_{ij}(t)$, that are present in our model for the data but not in the ODE models estimated in these works. As such, we seek an alternative approach by drawing on the existing literature for estimating a trend contained in a functional time series.

To begin we first consider additive trend. We follow the approach of Kokoszka and Young (2017) and define the additive trend function, $\{d_j(t,i)\}$, as follows: $d_j(t,i) = iy_j(t)$, for some function $\{y_j(t) \in L^2(\mathcal{T})\}_{j=1}^p$. Note that the i, j, and t in the additive drift function index the same quantities as in the observations $\{X_{ij}(t)\}$. Using the additive trend function we have the following model for our data:

$$X_{ij}(t) = d_j(t,i) + \sum_{k=1}^{m} c_{jk}u_{jk}(t) + f_{ij}(t), \quad 1 \leq j \leq p. \tag{5.4.1}$$

Using the model in equation (5.4.1) we see that the differences, $d_j(t,i) - d_j(t,i-1)$, are stationary. The estimation of the additive trend function can therefore be performed using differencing (Martínez-Hernández and Genton, 2021):

$$\hat{d}_j(t,i) = \frac{i}{n}\sum_{i=2}^{n} X_{ij}(t) - X_{i-1,j}(t).$$

Once estimates for $\{d_j(t,i)\}$ have been obtained, the trend can be removed from the data so that the underlying shape becomes stationary. This then allows for the differential operator for the data to be estimated using PDA in line with the approach presented in Section 5.2.2.

The second form of trend we consider is a scalar multiplicative trend. This is where a sample of data evolves over time by scaling. The function for a scalar multiplicative trend is given by $e_j(i) = ci$ for some $c \in \mathbb{R}$. The model for data affected by a scalar multiplicative trend is thus

$$X_{ij}(t) = e_j(i) \sum_{k=1}^{m} c_{jk} u_{jk}(t) + f_{ij}(t), \quad 1 \leq j \leq p. \tag{5.4.2}$$

Unlike additive trend, a scalar multiplicative trend has no affect on the estimation of the ODE for the data. This is due to the fact $\mathcal{L}$ is a linear operator and so the solutions to the ordinary differential equation $\mathcal{L}$ are scale invariant. As a result, if $\mathcal{L}(\sum_{k=1}^{m} c_{jk} u_{jk}(t)) = 0$ then $\mathcal{L}\left(e_j(i) \sum_{k=1}^{m} c_{jk} u_{jk}(t)\right) = 0$ also, since

$$\mathcal{L}\left(e_j(i) \sum_{k=1}^{m} c_{jk} u_{jk}(t)\right) = e_j(i) \mathcal{L}\left(\sum_{k=1}^{m} c_{jk} u_{jk}(t)\right) = 0.$$

As such, mvFAST can be performed without modification when a scalar multiplicative trend is present.

To highlight the performance of mvFAST when both additive, and multiplicative, trend affect the model for the data we shall perform simulations using two additional forms of data generating process. These are:

(vi) **Additive Trend:** $X_{ij}(t) = d_j(t,i) + a_{ij} \sin\left(\frac{2\pi t}{100}\right) + b_{ij} \cos\left(\frac{2\pi t}{200}\right) + f_{ij}(t)$, $a_{ij}, b_{ij} \sim$ U$[0.8, 1.2]$, $d_j(t,i) = iy(t)$. Here $y(t)$ is an uncorrelated Gaussian process such that, at any t, $y(t) \sim$ N$(1, 0.5)$.

(vii) **Multiplicative Trend:** $X_{ij}(t) = e_j(i)\left[a_{ij} \sin\left(\frac{2\pi t}{100}\right) + b_{ij} \cos\left(\frac{2\pi t}{200}\right)\right] + f_{ij}(t)$, $a_{ij}, b_{ij} \sim$ U$[0.8, 1.2]$, $e_j(i) \sim$ N$(1, 0.5)$.

In both the scenarios above $\{f_{ij}(t)\}$ is defined as in the other simulation settings. That is, a mean-zero Gaussian process with covariance kernel $K(s,t) = 0.2 \exp\left(\frac{1}{2}\left(\frac{s-t}{20}\right)^2\right)$.

As in the simulations in Sections 5.4.1- 5.4.2 we shall consider the power, and detection delay, of mvFAST with respect to each of the seven anomaly functions. Furthermore, we consider both the dense and sparse settings, and provide a comparison with the method of Dai and Genton (2019). In each of the simulations we have 50 complete observations with number of dimensions $p = 10$. We perform 500 repetitions of each setting, control the FWER to be 5%, remove the additive trend of the observations using differencing as described above, and for mvFAST fit an order $m = 2$ differential operator. The results for the detection power and detection delay are presented in Tables 5.4.6 and 5.4.7 respectively.

Trend	Additive		Multiplicative	
Anomaly	mvFAST	Dai	mvFAST	Dai
Single	**1**	0.76	**1**	**1**
Multiple	**1**	0.79	**1**	**1**
Polynomial	0.97	**0.98**	**1**	**1**
Phase	0.06	**0.41**	**0.29**	0.21
Constant	**0.99**	0.92	**1**	**1**
Exponential	**1**	**1**	**1**	**1**
Consistent	**0.03**	0.22	**0.08**	1

(a)

Trend	Additive		Multiplicative	
Anomaly	mvFAST	Dai	mvFAST	Dai
Single	**0.69**	0.31	**0.94**	0.69
Multiple	**0.75**	0.3	**0.96**	0.66
Polynomial	**0.58**	0.56	**0.84**	0.8
Phase	0.03	**0.41**	0.01	**0.04**
Constant	**0.14**	0.13	**0.82**	0.81
Exponential	**0.43**	0.39	**1**	**1**
Consistent	**0.01**	0.15	**0.16**	0.51

(b)

Table 5.4.6: Detection power for mvFAST in the dense (a) and sparse (b) settings when there is a trend in the underlying shape. The better result in each scenario is highlighted in **bold**.

As can be seen from the results in Tables 5.4.6(a) and (b), in general mvFAST performs well in both settings. Furthermore, the performance of mvFAST in the non-stationary setting is similar to that of the stationary setting for the multiple shape data generating process. This highlights the value of our proposed detrending step when the trend is additive. It also reinforces the mathematical argument that mvFAST is unaffected by a multiplicative trend.

Trend Anomaly	Additive	Multiplicative
Single	1.98 (5.98)	6.17 (4.04)
Multiple	7.40 (4.43)	21.52 (41.52)
Polynomial	105.29 (19.60)	42.70 (15.62)
Phase	4.99 (25.61)	21.52 (41.52)
Constant	28.98 (44.99)	2.09 (1.88)
Exponential	36.04 (42.44)	2.37 (3.40)
Consistent	3.55 (21.021574)	9.74 (34.46)

(a)

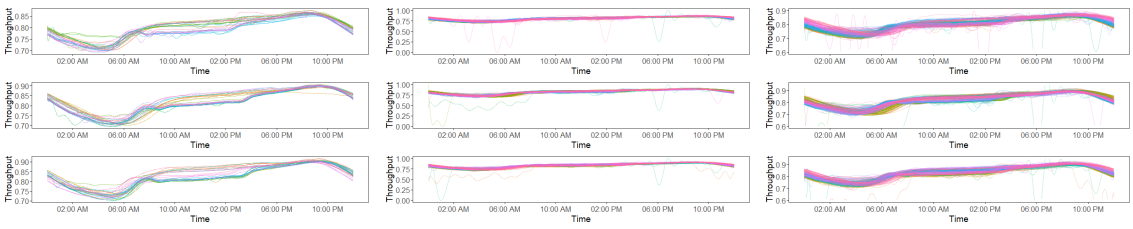
Trend Anomaly	Additive	Multiplicative
Single	45.66 (22.08)	42.52 (18.21)
Multiple	41.87 (39.44)	52.89 (19.12)
Polynomial	6.37 (29.61)	33.58 (22.10)
Phase	3.58 (19.33)	7.66 (3.21)
Constant	10.02 (37.33)	116.00 (62.94)
Exponential	43.73 (61.64)	113.24 (10.61)
Consistent	3.07 (23.50)	19.18 (4.32)

(b)

Table 5.4.7: Average detection delay (and standard deviation) for mvFAST in the dense (a) and sparse (b) settings when there is a trend in the underlying shape.

Now that the finite sample performance of mvFAST has been explored in a variety of settings, attention can be directed towards the primary purpose of mvFAST: monitoring for anomalies in telecommunications data streams. We present this application in the next section.

5.5 Telecommunications Application



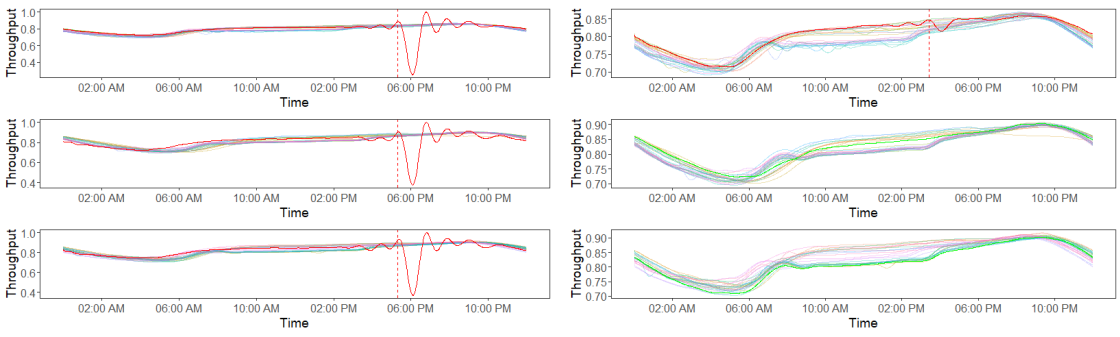
(a) (b) (c)

Figure 5.5.1: Training (a) and test (b) throughput data drawn from three locations on the network and used for the application. In figure (c) we present the data over a smaller range of throughput values.

Recall from Section 1 that our dataset is a record of throughput data at three key geographically dispersed points on a telecommunications network over 242 days. In order to implement a functional data approach, the data has been partitioned

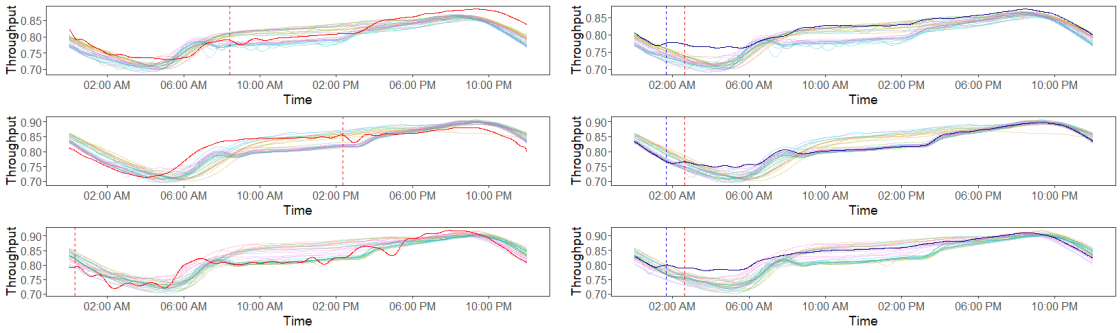
into daily measurements and smoothed into a single curve for each day following the approach of Ramsay (2005). In particular, we fit a system of 50 order 4 B-spline functions using penalised smoothing, and use cross-validation to select the penalty. The aim of the application is to analyse new data sequentially, in order to detect the onset of atypical behaviour as soon as practicable. To demonstrate this, we form a training dataset from the first 28 days of data, and then use mvFAST to test the remaining days sequentially as time progresses. The training, and test, data are depicted in Figure 5.5.1.

Our analysis indicates that several (statistically) interesting features are identified within the dataset. These not only include the two instances of atypical behaviour presented in Figure 5.1.2, but also several other potential events (see Figures 5.5.2 and 5.5.3). For example Figure 5.5.2(a) shows a deviation from the expected shape for the data that affects all three variates, whereas in Figure 5.5.2(b) the behaviour affects only one. Over the three locations and 207 days tested, mvFAST detects 35 days requiring further expert analysis. In each case, the false alarm rate of the training sample is controlled at the 5% level.



(a) (b)

Figure 5.5.2: Further examples of dense (a) and sparse (b) detections by mvFAST in throughput data drawn from three locations on the network.



(a) (b)

Figure 5.5.3: Deviations from the expected shape recorded on the same day, but at different times (a), and a deviation identified in the dataset linked with atypical morning demand (b). Observe that the blue line is when the observation appears to deviate from the underlying shape, and the red line is when detection takes place using mvFAST.

5.6 Discussion

In this chapter we have presented a novel method for the identification of anomalous curves in a sample of multivariate functional data. Our proposed approach allows for a subset of the dimensions of an observation to be labelled anomalous, and can be used to detect the time point on the curve the anomaly can be said to have taken place. The ability to detect anomalies and their location within a subset of the dimensions of multivariate functional data is a key contribution of this work. A second advantage is that this detection test can be extended to a sequential setting, as illustrated by the application in Section 5.5.

At the heart of mvFAST lies the estimation of an underlying model for the data using PDA. Several methods exist in the literature for estimating parameters of an ODE model, however there are several reasons for our choice of PDA. Foremost amongst these is that, unlike other methods such as those discussed in Ramsay and Hooker (2017) or Carey and Ramsay (2020), PDA does not require a computationally expensive two stage optimisation process to estimate the coefficients for the differential operator $\boldsymbol{\mathcal{L}}$. Furthermore, the method is non-parametric and so allows for the ODE to

be estimated without the need for assumptions of Gaussianity as made in Paul et al. (2011). The use of PDA also allows for the dependency between different dimensions to be captured in the representation for the data. This would not be the case if using multivariate FPCA (Chiou et al., 2014).

The use of PDA, however, is not without drawbacks. One drawback is that it requires each ODE equation to be of the same order, and so requires mp^2 parameters to be estimated. Further, every linking term in the system is included in the fitted model. This may lead to overfitting, however to guard against this problem we utilise the penalty term in equation (5.2.5). An alternative approach such as that of Carey and Ramsay (2020) could be considered, where the structure of the system can be specified and so fewer parameters may be needed. In many cases no prior model for the system will be known, however, and so specifying a system with specific linkage relationships between the p different dimensions will be challenging.

A second drawback of PDA is that parameter estimation may be affected by the dependence structures in the data. In particular, we have no guarantee that the parameter estimators are minimum variance when the innovation functions are dependent. However, simulations have shown that successful anomaly detection can still be carried out in these situations. Due to the fact that the primary concern of this chapter is anomaly detection rather than ODE parameter estimation, we argue that our simulations have shown that the limitations of PDA for model fitting do not have a significant effect on anomaly detection.

Finally, recall that the main aim of this work was to detect anomalies in a telecommunications network as quickly as possible so that appropriate further action can be taken. We have demonstrated that mvFAST is able to achieve this goal in Section 5.5, where we have analysed critical operational data drawn from such a network.

Chapter 6

Conclusions and Future Directions

This thesis has introduced three new anomaly detection methods: NUNC, FAST, and mvFAST. In this closing chapter we shall discuss the key contributions of each of the three proposed approaches, and suggest some avenues for future research.

The first of our contributions, NUNC, was introduced in Chapter 3 and offers a computationally efficient method for sequentially detecting changes in the distribution of nonparametric data. The proposed approach was motivated by a telecommunications monitoring application, and has been shown to outperform a competitor sequential method in this setting. A theoretical result for selecting an appropriate threshold for controlling the false alarm rate of the test has also been presented, and this result could be extended to other binary segmentation approaches in the at most one change, or sequential, settings.

Several extensions to NUNC could be considered, with the most obvious being to extend the method to multivariate data. One avenue for doing this would be to consider using copulas to model the multivariate empirical distribution for the data. This would have the added advantage of capturing the dependence structure between the different dimensions of the data. Additionally, in a similar vein to mvFAST, a multivariate form of NUNC could incorporate a test for changes affecting both every dimension, or a subset of the dimensions. This could be implemented, for example, by using a combination of the copula and the marginal empirical CDFs.

In Chapters 4 and 5 we introduce FAST, and the multivariate extension mvFAST.

The key novelty of FAST is that it permits the detection of anomalies within partially observed functional data, with mvFAST extending this to the multivariate functional data setting. This extension is non-trivial, as in the multivariate setting anomalies can affect either every dimension, or a subset of the dimensions of the data. Furthermore, the anomalies may occur at different times in the different dimensions of a single observation. Through the use of a two-stage threshold for the CUSUM test, mvFAST is able to distinguish between the different forms of anomaly. Various theoretical results have been obtained for both FAST and mvFAST, including results concerning the choice of the threshold for the tests, and bounds on the test detection power.

One intriguing aspect of FAST that could provide a source of future research is the estimation of the ODE function. Whilst the estimation of the residual functions using PDA is unaffected by dependence in the data, the individual coefficient estimates may be either biased or have high variance. This means that inference regarding a single model coefficient cannot be performed. Whilst FAST is not directly affected by this issue as it tests the residuals of the fitted ODE, the ability to test individual coefficients for anomalous structures could provide additional insights. As an alternative, a fitting step that uses the method of Carey and Ramsay (2020) could be considered. Their method provides guarantees on the fit of the individual ODE coefficients, and so inference on these coefficients could be incorporated into the anomaly detection process.

Finally, we turn to the problem outlined in mvFAST. This, and closely related, settings are ripe for further development. Consider for example a scenario where each dimension of a multivariate functional observation follows one of a finite set of p underlying shapes (see Figure 6.1.1). Over time, it may be the case that a particular dimension experiences a change from one shape to another (see Figure 6.1.2). Alternatively, there may be instances when an observation follows none of the shapes available. We term the former of these events as a change, and the latter an anomaly. Whilst functional anomaly detection methods such as mvFAST would be able to sequentially identify when both changes and anomalies have occurred, they would not distinguish between the two. Furthermore, they would not provide any

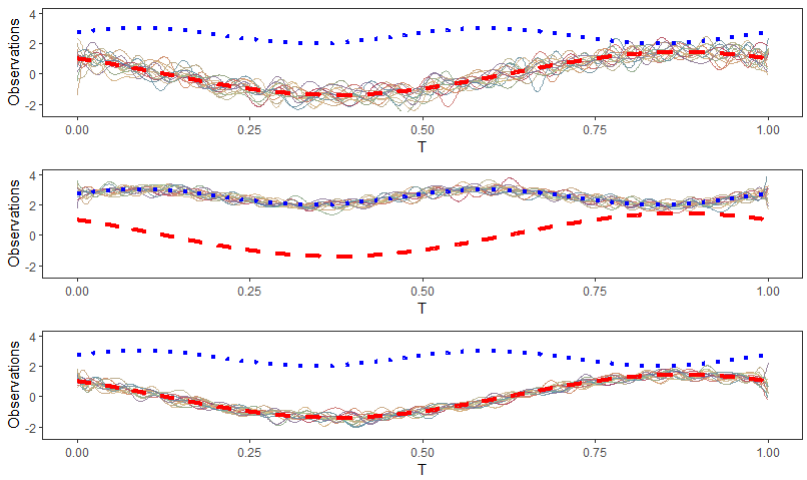


Figure 6.1.1: Example of three-dimensional multivariate data containing $p = 2$ underlying shapes (denoted by the dashed red and dotted blue lines respectively).

information as to what shape is being followed post-change. Consequently it would be interesting to develop a method that can (i) sequentially detect when a dimension of the data deviates from its underlying shape and, should a detection occur, (ii) classify which, if any, of the other p shapes for the data are now being followed. Below, we introduce a potential model for this setting and outline a suggested approach to address this sequential anomaly detection and classification challenge.

In our setting we consider n observations of m different dimensions of a multivariate functional dataset. That is, $\{X_{ij}(t) : 1 \leq i \leq m, 1 \leq j \leq n, t \in \mathcal{T}\} \in L^2(\mathcal{T})$, with $\mathcal{T}$ representing the domain of time the functions are observed over. In this setting $X_{ij}(t)$ represents the jth observation of the ith dimension at time t. As described above, each of the observed dimensions of an observation follows one of p different shapes, $\{C_k(t)\}_{k=1}^{p}$. If dimension i follows the kth shape at time j then we have the following model for the observations:

$$X_{ij}(t) = C_k(t) + \epsilon_{ij}(t), \tag{6.1.1}$$

where $\epsilon_{ij}(t) \in L^2(\mathcal{T})$ represents an independent and identically distributed mean-zero innovation function.

Over time, it may be the case that the shape followed by a dimension changes, and transitions to either one of the other $p - 1$ shapes or becomes anomalous. Below

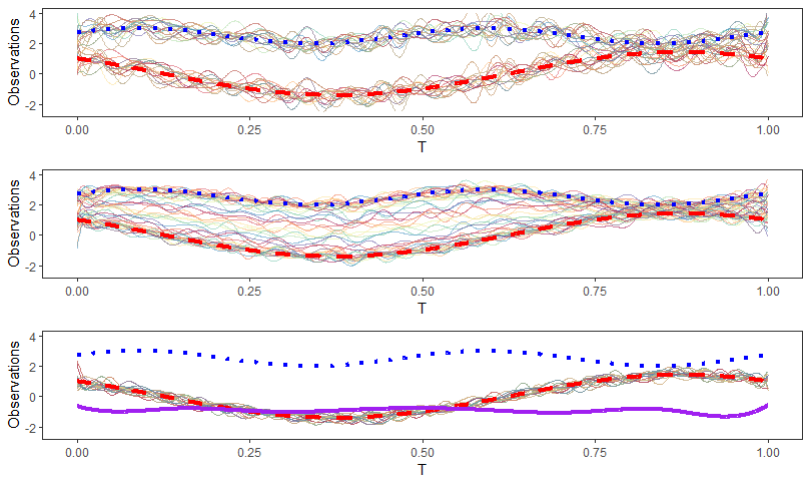


Figure 6.1.2: This figure shows examples of changes in multivariate functional data with $p = 2$ underlying shapes (denoted by the dashed red and dotted blue lines respectively). In this example the first dimension experiences an abrupt change (equation (6.1.2)), the second dimension experiences a gradual change (equation (6.1.3)), and in the third dimension an anomaly (equation (6.1.4) is recorded (solid purple line).

three models are provided to represent this. The first model (6.1.2) represents an abrupt change in shape at time ν, whereas the second (6.1.3) represents a gradual change over time. The third (6.1.4) contains an anomaly function, $\{g_{ij}(t)\} \in L^2(\mathcal{T})$, that causes a deviation from the underlying shape. Examples of each of these three occurrences are presented in Figure 6.1.2.

$$X_{ij}(t) = C_k(t)\mathbb{I}_{j\leq\nu} + C_l(t)\mathbb{I}_{j>\nu} + \epsilon_{ij}(t) \tag{6.1.2}$$

$$X_{ij}(t) = C_k(t) + f(j)\left(C_l(t) - C_k(t)\right) + \epsilon_{ij}(t) \tag{6.1.3}$$

$$X_{ij}(t) = C_k(t) + g_{ij}(t) + \epsilon_{ij}(t) \tag{6.1.4}$$

In equation (6.1.3) $f : \mathbb{Z} \to [0, 1]$ represents a function controlling the drift, with $f(1) = 0$ and $f(n) = 1$, so that after n steps the shape has changed from the kth to the lth.

In order to detect these changes, a test to compare an observation to its assigned shape is required. If dimension i follows shape k at time j then we can define a

distance, Z_{ij}, as follows:

$$Z_{ij} = \int_{\mathcal{T}} \left(X_{ij}(t) - C_k(t)\right)^2 \, \mathrm{d}t. \tag{6.1.5}$$

Intuitively, a large value of Z_{ij} suggests that an observation differs substantially from its assigned shape. With this in mind, one could use Z_{ij} as the basis for a CUSUM test statistic:

$$\Delta_{ij} = \sum_{r=1}^{j} \left(Z_{ir} - \mu_{ij}\right). \tag{6.1.6}$$

Here μ represents the expected value of the distance under the null hypothesis of no change. Should $|\Delta_{ij}| > \gamma$, where γ is a threshold for the test, then an alarm would be declared. We remark that by using the cumulative sum of the distance, Z_{ij}, it is hoped that it is possible to detect both abrupt (equation (6.1.2)), and gradual (equation (6.1.3)), changes in the data.

Once the detection test declares an alarm, one interesting possibility would be to consider classifying the nature of the change. Using a similar measure to equation (6.1.5) one might compare the distance of the post-change observation with the other $p-1$ shapes for the data. Should the observation be sufficiently close to one of these shapes, then the dimension could be reclassified as following this shape. On the other hand, should no shape offer a suitable fit then an anomaly would be declared.

The approach outlined above offers a glimpse into the challenge of combining anomaly detection and classification for multivariate functional data. Many interesting avenues could be explored to build upon our proposed approach, including incorporating non-stationarity into the shapes for the data, or modelling the dependency structure of the different dimensions.

Appendix A

Chapter 3 - NUNC

A.1 Proof of Results

A.1.1 Proof of Proposition 1

Proof. In order to prove the desired bound, we focus on the extreme case where either $x_{1:\tau} < q$ and $x_{\tau+1:W} > q$, or vice versa, and note that this case maximises the expression

$$\mathcal{L}(x_{t-W+1:\tau}; q) + \mathcal{L}(x_{\tau+1:t}; q) - \mathcal{L}(x_{t-W+1:t}; q), \tag{A.1.1}$$

for any quantile q.

By writing out the likelihoods in the above equation, it can be observed that $\mathcal{L}(x_{t-W+1:\tau}; q) = 0$, $\mathcal{L}(x_{\tau+1:t}; q) = 0$, and

$$\mathcal{L}(x_{t-W+1:t}; q) = \frac{\tau}{W} \log \frac{\tau}{W} - (W - \tau) \log \left(\frac{W - \tau}{W} \right) \tag{A.1.2}$$

in the case where each $x_{1:\tau} > q$ and $x_{\tau+1:W} < q$. We also have

$$\mathcal{L}(x_{t-W+1:t}; q) = -\frac{\tau}{W} \log \frac{\tau}{W} + (W - \tau) \log \left(\frac{W - \tau}{W} \right) \tag{A.1.3}$$

when $x_{1:\tau} < q$ and $x_{\tau+1:W} > q$. Given both $\frac{\tau}{W}$ and $\frac{W-\tau}{W}$ are less than one, the log terms in equations (A.1.2) and (A.1.3) are both negative. As such, we can bound both equations (A.1.2) and (A.1.3) above by the following bound:

$$\mathcal{L}(x_{t-W+1:t}; q) \leq -\frac{\tau}{W} \log \frac{\tau}{W} - (W - \tau) \log \left(\frac{W - \tau}{W} \right).$$

As this case dealt with the maximum of equation (A.1.1), this then means that for any quantile q and window of data $x_1, \ldots, x_W$ we have that

$$\mathcal{L}(x_{t-W+1:\tau}; q) + \mathcal{L}(x_{\tau+1:t}; q) - \mathcal{L}(x_{t-W+1:t}; q) \leq -\frac{\tau}{W} \log \frac{\tau}{W} - (W - \tau) \log \left(\frac{W - \tau}{W} \right), \tag{A.1.4}$$

as required. Additionally, we note that this equation is decreasing in τ for fixed W.

We then consider the test statistic, given by equation (3.2.4). As a result of the bound in equation (A.1.4) if

$$2K \left[-\frac{\tau}{W} \log \frac{\tau}{W} - (W - \tau) \log \left(\frac{W - \tau}{W} \right) \right] \leq K\beta$$

then detection is impossible, as the bound for the test statistic does not exceed the threshold for the test. As a result, we conclude that if

$$-\frac{\tau^*}{W} \log \frac{\tau^*}{W} - (W - \tau^*) \log \left(\frac{W - \tau^*}{W} \right) \leq \frac{\beta}{2}$$

then due to the fact the expression decreases as τ increases then for $\tau > \tau^*$ detection is impossible. This completes the proof. □

A.1.2 Proof of Proposition 2

Proof. A false alarm by the time t under NUNC Local can be written as

$$\begin{aligned} &= \mathrm{P}\left(\bigcup_{s=W}^{t} \max_{s-W+1 \leq \tau \leq s} \sum_{k=1}^{K} 2 \left[\mathcal{L}(x_{s-W+1:\tau}; q_k) + \mathcal{L}(x_{\tau+1:s}; q_k) - \mathcal{L}(x_{s-W+1:s}; q_k) \right] \geq K\beta \right) \\ &\leq \sum_{s=W}^{t} \sum_{\tau=s-W+1}^{s} \mathrm{P}\left(\sum_{k=1}^{K} 2 \left[\mathcal{L}(x_{s-W+1:\tau}; q_k) + \mathcal{L}(x_{\tau+1:s}; q_k) - \mathcal{L}(x_{s-W+1:s}; q_k) \right] \geq K\beta \right) \end{aligned} \tag{A.1.5}$$

The next part of the proof uses the fact that, under the i.i.d. assumption, asymptotically for any quantile the following holds

$$2 \left[\mathcal{L}(x_{s-W+1:\tau}; q_k) + \mathcal{L}(x_{\tau+1:s}; q_k) - \mathcal{L}(x_{s-W+1:s}; q_k) \right] \sim \chi^2_1.$$

As in Wainwright (2019), it can be shown that if a random variable X_i follows a χ^2_1 distribution then it is Sub-Exponential with parameters 4 and 4. That is, $X_i \sim \mathrm{SE}(4, 4)$.

Under dependence, as is the case between different quantiles, if $X_i \sim \text{SE}(\nu_i^2, b_i)$ then $\sum_{i=1}^n X_i - \text{E}(X_i) \sim \text{SE}\left(\left(\sum_{i=1}^n \nu_i\right)^2, \max_i b_i\right)$, and so

$$\sum_{k=1}^{K} [\mathcal{L}(x_{s-W+1:\tau}; q_k) + \mathcal{L}(x_{\tau+1:s}; q_k) - \mathcal{L}(x_{s-W+1:s}; q_k)] - K \sim \text{SE}(4K^2, 4).$$

We then use a well known bound on the subexponential tail (Vershynin, 2018) to obtain

$$\begin{aligned}
&\sum_{s=W}^{t} \sum_{\tau=s-W+1}^{s} \text{P}\left(\sum_{k=1}^{K} 2\left[\mathcal{L}(x_{s-W+1:\tau}; q_k) + \mathcal{L}(x_{\tau+1:s}; q_k) - \mathcal{L}(x_{s-W+1:s}; q_k)\right] \geq K\beta\right) \\
&= \sum_{s=W}^{t} \sum_{\tau=s-W+1}^{s} \text{P}\left(\sum_{k=1}^{K} 2\left[\mathcal{L}(x_{s-W+1:\tau}; q_k) + \mathcal{L}(x_{\tau+1:s}; q_k) - \mathcal{L}(x_{s-W+1:s}; q_k)\right] - K \geq K\beta - K\right) \\
&\leq W(t-W+1)\exp\left(-\frac{1}{2}\min\left\{\frac{K\beta-K}{4}, \frac{(K\beta-K)^2}{4K^2}\right\}\right). \qquad \text{(A.1.6)}
\end{aligned}$$

We can set this final line equal to α to control our desired false alarm rate. We have two cases in equation (A.1.6) and must bound above by the largest of these. The first case is that

$$W(t-W+1)\exp\left(-\frac{K\beta_1 - K}{8}\right) = \alpha$$

in which case we choose

$$\beta_1 = 1 - 8K^{-1}\log\left(\frac{\alpha}{W(t-W+1)}\right).$$

On the other hand we have the case where

$$W(t-W+1)\exp\left(-\frac{(K\beta_2 - K)^2}{8K^2}\right) = \alpha$$

and solving this gives

$$\beta_2 = 1 + 2\sqrt{2\log\left(\frac{W(t-W+1)}{\alpha}\right)}.$$

We then choose the larger of β_1 and β_2, completing the proof. □

Appendix B

Chapter 4 - FAST

B.1 Proof of Results

In this section we provide proofs of the results in Section 4.3.

B.1.1 Proof of Proposition 3

Proof. From definition of the probability of false alarm it is the case that

$$\mathrm{P}(\xi \leq T) = \mathrm{P}\left(\bigcup_{\tau=2}^{T}\{\xi = \tau\}\right) \leq \sum_{\tau=2}^{T}\mathrm{P}(\xi = \tau),$$

where any stopping time at τ can be bounded above by

$$\mathrm{P}(\xi = \tau) \leq \mathrm{P}\left(\frac{|\Delta_{n+1}(\tau)|}{\sqrt{\tau - 1}} \geq \gamma\right)$$

because $\{\xi = \tau\} \subset \{\frac{|\Delta_{n+1}(\tau)|}{\sqrt{\tau-1}} \geq \gamma\}$. Putting these two facts together gives

$$\mathrm{P}(\xi \leq T) \leq \sum_{\tau=2}^{T}\mathrm{P}\left(\frac{|\Delta_{n+1}(\tau)|}{\sqrt{\tau - 1}} \geq \gamma\right).$$

The quantity $\Delta_{n+1}(\tau) = \sum_{r=2}^{\tau} Z_{n+1}(r)$ may contain dependent terms in the summation, however a Central Limit Theorem result for dependent sequences (Billingsley, 1995) can be applied as each term in the sum is drawn from a Gaussian Process with a stationary covariance matrix (Bradley, 2005). Under the Central Limit Theorem approximation, we have that $\frac{\Delta_{n+1}(\tau)}{\sqrt{\tau-1}} \sim \mathrm{N}(0, 1)$.

Using this approximate distribution, and the symmetry of the Gaussian distribution, we obtain the following bound for the test statistic:

$$\sum_{\tau=2}^{T} \mathrm{P}\left(\frac{|\Delta_{n+1}(\tau)|}{\sqrt{\tau-1}} \geq \gamma\right) = 2\sum_{\tau=2}^{T} 1 - \Phi\left(\gamma\right) = 2\left(T - 1 - \sum_{\tau=2}^{T} \Phi\left(\gamma\right)\right).$$

To complete the proof we now need to find suitable γ so that $\mathrm{P}(\xi \leq T) \leq \alpha$, and to do this we equate

$$\alpha = 2\left(T - 1 - \sum_{\tau=2}^{T} \Phi\left(\gamma\right)\right).$$

Hence we take

$$\gamma = \Phi^{-1}\left(1 - \frac{\alpha}{2(T-1)}\right),$$

to complete the proof.

□

B.1.2 Proof of Proposition 4

Proof. The initial steps of this proof follows similar steps to those taken in the proof of Proposition 2 in Fisch et al. (2022b). In particular, we follow their approach to establish equation (B.1.1). After this, our proof differs from theirs as the nature, and distribution, of their test statistic is different to the FAST test statistic.

The probability of detecting a contaminated observation by time T, is given by the expression

$$\mathrm{P}\left(\bigcup_{\tau=2}^{T} \frac{|\Delta_{n+1}(\tau)|}{\sqrt{\tau-1}} \geq \gamma\right),$$

where γ is the test threshold. To see why observe that the probability of an anomaly not being detected by time T is given by $\mathrm{P}(\xi > T) = 1 - \mathrm{P}(\xi \leq T)$ and that

$$\mathrm{P}(\xi > T) = \mathrm{P}\left(\bigcap_{\tau=2}^{T} \frac{|\Delta_{n+1}(\tau)|}{\sqrt{\tau-1}} < \gamma\right)$$

Hence taking the complementary event

$$\begin{aligned}\mathrm{P}(\xi \leq T) &= \mathrm{P}\left(\left[\bigcap_{\tau=2}^{T} \frac{|\Delta_{n+1}(\tau)|}{\sqrt{\tau-1}} < \gamma\right]^{c}\right) \\ &= \mathrm{P}\left(\bigcup_{\tau=2}^{T} \frac{|\Delta_{n+1}(\tau)|}{\sqrt{\tau-1}} \geq \gamma\right).\end{aligned}$$

A bound on the probability of an anomaly being detected by time T can be calculated by bounding this expression as follows:

$$\begin{aligned} \mathrm{P}\left(\frac{|\Delta_{n+1}(T)|}{\sqrt{T-1}} \geq \gamma\right) &\leq \mathrm{P}\left(\bigcup_{\tau=2}^{T}\left(\frac{|\Delta_{n+1}(\tau)|}{\sqrt{\tau-1}} \geq \gamma\right)\right) \\ &\leq \sum_{\tau=2}^{T} \mathrm{P}\left(\frac{|\Delta_{n+1}(\tau)|}{\sqrt{\tau-1}} \geq \gamma\right). \end{aligned} \tag{B.1.1}$$

In order to proceed we require expressions for the test statistic, $\Delta_{n+1}(\tau)$, when it has been contaminated by the anomaly function. Using the equation for the test statistic, (4.2.9), for $\tau > \nu + s$ the contaminated test statistic is given by

$$\Delta_{n+1}(\tau) = \sum_{r=2}^{\tau} Z^c_{n+1}(r) \tag{B.1.2}$$

with $Z^c_{n+1}(\tau)$ given by

$$Z^c_{n+1}(r) = \frac{(\hat{\epsilon}_{n+1}(r) + \hat{\mathcal{L}}(g_{n+1})(r) - \hat{\epsilon}_{n+1}(r-1) - \hat{\mathcal{L}}(g_{n+1})(r-1))^2 - \mu(r)}{\sigma(r)}. \tag{B.1.3}$$

We remark that although the anomaly only occurs on $\nu \leq r \leq \nu + s$, we set $g(r) = 0$ outside this range to avoid the need to write the contaminated test statistic in equation (B.1.3) using three summations.

For the lower bound in equation (B.1.1), we proceed as in the proof of Proposition 3 and use the Central Limit Theorem Approximation (Billingsley, 1995) for the sum in equation (B.1.3). The difference between this proof and that of Proposition 3, however, is that the contaminated values $Z^c_{n+1}(\tau)$ will not be mean-zero or have unit variance.

In order to use the central limit theorem approximation we therefore first derive the mean and variance of $(\hat{\epsilon}_{n+1}(r) + \hat{\mathcal{L}}(g_{n+1})(r) - \hat{\epsilon}_{n+1}(r-1) - \hat{\mathcal{L}}(g_{n+1})(r-1))^2$. As this quantity is the square of a Gaussian random variable with non-zero mean and non-unit variance it is the case that

$$(\hat{\epsilon}_{n+1}(r) + \hat{\mathcal{L}}(g_{n+1})(r) - \hat{\epsilon}_{n+1}(r-1) - \hat{\mathcal{L}}(g_{n+1})(r-1))^2 \sim \psi(r) V_r \tag{B.1.4}$$

where $\psi(r) = \mathrm{Var}(\hat{\epsilon}_{n+1}(r) - \hat{\epsilon}_{n+1}(r-1))$, and V_r is a non-central chi-square random variable with one degree of freedom and non-centrality parameter $\lambda(r) = \frac{(\hat{\mathcal{L}}(g_{n+1})(r) - \hat{\mathcal{L}}(g_{n+1})(r-1))^2}{\psi(r)}$ (Johnson et al., 1994).

We then combine equation (B.1.4) with equation (B.1.3) to see that

$$Z^c_{n+1}(r) \sim \frac{\psi(r)V_r - \mu(r)}{\sigma(r)}. \tag{B.1.5}$$

This can then be further simplified by considering the relationships between $\mu(r)$, $\sigma(r)$, and $\psi(r)$. To see the connection between $\mu(r)$ and $\psi(r)$ we note that $\mathrm{E}(\hat{\epsilon}_{n+1}(r) - \hat{\epsilon}_{n+1}(r-1)) = 0$ and so

$$\begin{aligned}\mu(r) &= \mathrm{E}(\hat{\epsilon}_{n+1}(r) - \hat{\epsilon}_{n+1}(r-1))^2 \\ &= \mathrm{Var}(\hat{\epsilon}_{n+1}(r) - \hat{\epsilon}_{n+1}(r-1)) \\ &= \psi(r).\end{aligned}$$

The relationship between $\psi(r)$ and $\sigma(r)$ can also be derived by observing that

$$(\hat{\epsilon}_{n+1}(r) - \hat{\epsilon}_{n+1}(r-1))^2 \sim \psi(r)W,$$

where W is a χ^2_1 random variable, because the LHS of the above is the square of a zero mean non-unit variance Gaussian (Johnson et al., 1994). Now by definition

$$1 = \mathrm{Var}\left(\frac{(\hat{\epsilon}_{n+1}(r) - \hat{\epsilon}_{n+1}(r-1))^2}{\sigma(r)}\right) = \mathrm{Var}\left(\frac{\psi(r)W}{\sigma(r)}\right),$$

because $\sigma^2(r) = \mathrm{Var}(\hat{\epsilon}_{n+1}(r) - \hat{\epsilon}_{n+1}(r-1))^2$, and so because $\mathrm{Var}(W) = 2$ it is the case that $2\psi^2(r) = \sigma^2(r)$. Hence we conclude that $\sigma(r) = \sqrt{2}\psi(r)$.

Combining the expressions for $\mu(r)$ and $\sigma(r)$ with equations (B.1.3) and (B.1.5), we have therefore shown that

$$\left(\frac{(\hat{\epsilon}_{n+1}(r) + \hat{\mathcal{L}}(g_{n+1})(r) - \hat{\epsilon}_{n+1}(r-1) - \hat{\mathcal{L}}(g_{n+1})(r-1))}{\sqrt{\sigma(r)}}\right)^2 \sim \frac{V_r - 1}{\sqrt{2}},$$

where V_r is a non-central Chi-square distribution with one degree of freedom and non-centrality parameter $\lambda(r) = \frac{(\hat{\mathcal{L}}(g_{n+1})(r) - \hat{\mathcal{L}}(g_{n+1})(r-1))^2}{\psi(r)}$.

Using results for the mean and variance of a non-central Chi-square distribution (Johnson et al., 1994), the mean of each summand is therefore

$$\mathrm{E}(Z^c_{n+1}(r)) = \frac{\lambda(r)}{\sqrt{2}},$$

and the variance is

$$\mathrm{Var}(Z^c_{n+1}(r)) = 1 + 2\lambda(r).$$

Using the mean and variance in conjunction with the Central Limit Theorem, we can conclude the test statistic is distributed as

$$\frac{\Delta_{n+1}(\tau)}{\sqrt{\tau-1}} \sim \mathrm{N}\left(\frac{1}{\sqrt{\tau-1}}\sum_{r=2}^{\tau}\frac{\lambda(r)}{\sqrt{2}}, \frac{1}{\tau-1}\sum_{r=2}^{\tau} 1+2\lambda(r)\right). \tag{B.1.6}$$

Putting equation (B.1.6) into the lower bound in equation (B.1.1), we can conclude that

$$\mathrm{P}\left(\frac{|\Delta_{n+1}(T)|}{\sqrt{T-1}} \geq \gamma\right) = \mathrm{P}\left(|U_T| \geq \gamma\right),$$

where $U_T \sim \mathrm{N}\left(\frac{1}{\sqrt{T-1}}\sum_{r=2}^{T}\frac{\lambda(r)}{\sqrt{2}}, \frac{1}{T-1}\sum_{r=2}^{T} 1+2\lambda(r)\right)$, and where $\lambda(r) = \frac{(\hat{\mathcal{L}}(g_{n+1})(r)-\hat{\mathcal{L}}(g_{n+1})(r-1))^2}{\mu(r)}$, and $\hat{\mathcal{L}}(g_{n+1})(r) = 0$ outside of $\nu \leq r \leq \nu+s$.

The upper bound in equation (B.1.1) can be deduced similarly to the lower bound, because by equation (B.1.6) we have

$$\sum_{\tau=2}^{T}\mathrm{P}\left(\frac{|\Delta_{n+1}(\tau)|}{\sqrt{\tau-1}} \geq \gamma\right) = \sum_{\tau=2}^{T}\mathrm{P}\left(|U_\tau| \geq \gamma\right),$$

with $U_\tau \sim \mathrm{N}\left(\frac{1}{\sqrt{\tau-1}}\sum_{r=2}^{\tau}\frac{\lambda(r)}{\sqrt{2}}, \frac{1}{\tau-1}\sum_{r=2}^{\tau} 1+2\lambda(r)\right)$.

Finally, to complete the result, we move the plus one term in the expression for the variance outside the summation, and thus conclude that $U_\tau \sim \mathrm{N}\left(\frac{1}{\sqrt{\tau-1}}\sum_{r=2}^{\tau}\frac{\lambda(r)}{\sqrt{2}}, 1+\frac{1}{\tau-1}\sum_{r=2}^{\tau} 2\lambda(r)\right)$, as required. □

B.1.3 Proof of Proposition 5

Proof. This result requires proof that

$$\lim_{T\to\infty}\mathrm{P}\left(|\Delta_{n+1}(T)| \geq \gamma\sqrt{T}\right) = 1,$$

and in order to demonstrate this we first observe that

$$\lim_{T\to\infty}\mathrm{P}\left(|\Delta_{n+1}(T)| \geq \gamma\sqrt{T}\right) = \lim_{T\to\infty}\mathrm{P}\left(|\sum_{\tau=2}^{\nu-1} Z_{n+1}(\tau) + \sum_{\tau=\nu}^{T} Z^c_{n+1}(\tau)| \geq \gamma\sqrt{T}\right),$$

and that this is equivalent to

$$\lim_{T\to\infty}\mathrm{P}\left(|\Delta_{n+1}(T)| \geq \gamma\sqrt{T}\right) = \lim_{T\to\infty}\mathrm{P}\left(\left|\frac{\sum_{\tau=2}^{\nu-1} Z_{n+1}(\tau)}{T-\nu+1} + \frac{\sum_{\tau=\nu}^{T} Z^c_{n+1}(\tau)}{T-\nu+1}\right| \geq \frac{\gamma\sqrt{T}}{T-\nu+1}\right).$$

As ν is fixed it is the case that $\frac{\sum_{\tau=2}^{\nu} Z_{n+1}(\tau)}{T-\nu+1} \to 0$ as $T \to \infty$. Hence we can state that

$$\lim_{T\to\infty} \mathrm{P}\left(\frac{\sum_{\tau=2}^{\nu} Z_{n+1}(\tau)}{T-\nu+1} \geq -\frac{\delta_1}{3}\right) = 1$$

for any $\delta_1 > 0$. Turning our attention to the second term in the LHS, we recall equation (B.1.3):

$$Z^c_{n+1}(\tau) = \left[\frac{(\hat{\epsilon}_{n+1}(\tau) + \hat{\mathcal{L}}(g_{n+1})(\tau) - \hat{\epsilon}_{n+1}(\tau-1) - \hat{\mathcal{L}}(g_{n+1})(\tau-1))^2 - \mu(\tau)}{\sigma(\tau)}\right],$$

and expand this to see that we have

$$\begin{aligned} Z^c_{n+1}(\tau) = &\frac{\left(\hat{\epsilon}_{n+1}(\tau) - \hat{\epsilon}_{n+1}(\tau-1)\right)^2 - \mu(\tau)}{\sigma(\tau)(T-\nu+1)} \\ &+ \frac{2\Big(\hat{\epsilon}_{n+1}(\tau) - \hat{\epsilon}_{n+1}(\tau-1)\Big)\Big(\hat{\mathcal{L}}(g_{n+1})(\tau) - \hat{\mathcal{L}}(g_{n+1})(\tau-1)\Big)}{\sigma(\tau)(T-\nu+1)} \\ &+ \frac{\Big(\hat{\mathcal{L}}(g_{n+1})(\tau) - \hat{\mathcal{L}}(g_{n+1})(\tau-1)\Big)^2}{\sigma(\tau)(T-\nu+1)}, \end{aligned}$$

with

$$\frac{\left(\hat{\epsilon}_{n+1}(\tau) - \hat{\epsilon}_{n+1}(\tau-1)\right)^2 - \mu(\tau)}{\sigma(\tau)(T-\nu+1)} = \frac{Z_{n+1}(\tau)}{T-\nu+1}.$$

We then work with the following three pieces from the above expression:

$$\sum_{\tau=\nu}^{T} \frac{Z_{n+1}(\tau)}{T-\nu+1} \tag{B.1.7}$$

$$\sum_{\tau=\nu}^{T} \frac{2\Big(\hat{\epsilon}_{n+1}(\tau) - \hat{\epsilon}_{n+1}(\tau-1)\Big)\Big(\hat{\mathcal{L}}(g_{n+1})(\tau) - \hat{\mathcal{L}}(g_{n+1})(\tau-1)\Big)}{\sigma(\tau)(T-\nu+1)} \tag{B.1.8}$$

$$\sum_{\tau=\nu}^{T} \frac{\Big(\hat{\mathcal{L}}(g_{n+1})(\tau) - \hat{\mathcal{L}}(g_{n+1})(\tau-1)\Big)^2}{\sigma(\tau)(T-\nu+1)}. \tag{B.1.9}$$

Using Assumption 1 we see that by the weak law of large numbers (Karlin and Taylor, 2012), and because both equation (B.1.7) and equation (B.1.8) are the sum of mean-zero random variables, these two pieces of the sum converge in probability to zero. As a result we can state that

$$\lim_{T\to\infty} \mathrm{P}\left(\sum_{\tau=\nu}^{T} \frac{Z_{n+1}(\tau)}{T-\nu+1} > -\frac{\delta_1}{3}\right) = 1,$$

and

$$\lim_{T\to\infty} \mathrm{P}\left(\sum_{\tau=\nu}^{T} \frac{2\left(\hat{\epsilon}_{n+1}(\tau) - \hat{\epsilon}_{n+1}(\tau-1)\right)\left(\hat{\mathcal{L}}(g_{n+1})(\tau) - \hat{\mathcal{L}}(g_{n+1})(\tau-1)\right)}{\sigma(\tau)(T-\nu+1)} > -\frac{\delta_1}{3} \right) = 1$$

for any $\delta_1 > 0$. On the other hand, due to the fact that the summand of equation (B.1.9) is equal to zero only finitely many times the sum will take a positive value. As such, we have that for some $\delta_2 > 0$:

$$\lim_{T\to\infty} \mathrm{P}\left(\sum_{\tau=\nu}^{T} \frac{\left(\hat{\mathcal{L}}(g_{n+1})(\tau) - \hat{\mathcal{L}}(g_{n+1})(\tau-1)\right)^2}{\sigma(\tau)(T-\nu+1)} \geq \delta_2 \right) = 1.$$

Thus if we choose $\delta_1 < \delta_2$ and set $\delta = \delta_2 - \delta_1 > 0$ we can put these four inequalities together to show that

$$\lim_{T\to\infty} \mathrm{P}\left(\frac{\sum_{\tau=2}^{\nu-1} Z_{n+1}(\tau)}{T-\nu+1} + \frac{\sum_{\tau=\nu}^{T} Z^c_{n+1}(\tau)}{T-\nu+1} \geq -\frac{\delta_1}{3} - \frac{\delta_1}{3} - \frac{\delta_1}{3} + \delta_2 \right) = 1.$$

As $\delta = \delta_2 - \delta_1 > 0$ both sides are positive, and so taking absolute values gives

$$\lim_{T\to\infty} \mathrm{P}\left(\left| \frac{\sum_{\tau=2}^{\nu-1} Z_{n+1}(\tau)}{T-\nu+1} + \frac{\sum_{\tau=\nu}^{T} Z^c_{n+1}(\tau)}{T-\nu+1} \right| \geq \delta \right) = 1.$$

Hence we have shown that for some $\delta > 0$

$$\lim_{T\to\infty} \mathrm{P}\left(\frac{|\Delta_{n+1}(T)|}{T-\nu+1} \geq \delta \right) = 1.$$

We now turn to consider the threshold the asymptotic behaviour of the threshold, $\frac{\gamma\sqrt{T}}{T-\nu+1}$. As γ has been chosen as in Proposition 3 we find that

$$\frac{\gamma\sqrt{T}}{T-\nu+1} = \frac{\Phi^{-1}\left(1 - \frac{\alpha}{2(T-1)}\right)\sqrt{T}}{T-\nu+1} = \mathcal{O}\left(\frac{\Phi^{-1}(1-\frac{\alpha}{2T})}{\sqrt{T}} \right),$$

It is stated in Blair et al. (1976) that

$$\Phi(1-T^{-1}) = \mathcal{O}\left(\sqrt{-2\log(T^{-1})}\right)$$

and it is well known that

$$\sqrt{\frac{2\log(T)}{T}} = o(1).$$

Putting this together, we have that

$$\frac{\gamma\sqrt{T}}{T-\nu+1} = o(1),$$

and so for any $\epsilon > 0$ and large enough T

$$\frac{\gamma\sqrt{T}}{T-\nu+1} \leq \epsilon.$$

Due to the fact that there exists some $\delta > 0$ such that

$$\lim_{T\to\infty} \mathrm{P}\left(\frac{|\Delta_{n+1}(T)|}{T-\nu+1} \geq \delta\right) = 1.$$

we can therefore conclude that, by letting $\epsilon = \delta$ and choosing T large enough, we have

$$\begin{aligned}
1 &= \lim_{T\to\infty} \mathrm{P}\left(\frac{|\Delta_{n+1}(T)|}{T-\nu+1} \geq \delta\right) \\
&= \lim_{T\to\infty} \mathrm{P}\left(\frac{|\Delta_{n+1}(T)|}{T-\nu+1} \geq \epsilon\right) \\
&\leq \lim_{T\to\infty} \mathrm{P}\left(\frac{|\Delta_{n+1}(T)|}{T-\nu+1} \geq \frac{\gamma\sqrt{T}}{T-\nu+1}\right).
\end{aligned}$$

We then use this to conclude that

$$\lim_{T\to\infty} \mathrm{P}\left(\left|\frac{\sum_{\tau=2}^{\nu-1} Z_{n+1}(\tau)}{T-\nu+1} + \frac{\sum_{\tau=\nu}^{T} Z^c_{n+1}(\tau)}{T-\nu+1}\right| \geq \frac{\gamma\sqrt{T}}{T-\nu+1}\right) = 1,$$

as required. □

B.2 Additional Simulation Results

In this section we shall present additional simulation results concerning both the estimation of $\mathcal{L}$ and the detection performance of FAST. First, we shall consider whether performing PDA with penalised basis smoothing has a significant impact on the residuals obtained when estimating the differential operator for the observations using PDA. Following on from this, we shall present a repeat of the simulations contained in Section 4.4.1 but for when the covariance operator for the data has a Matérn kernel.

B.2.1 Choice of Basis in PDA

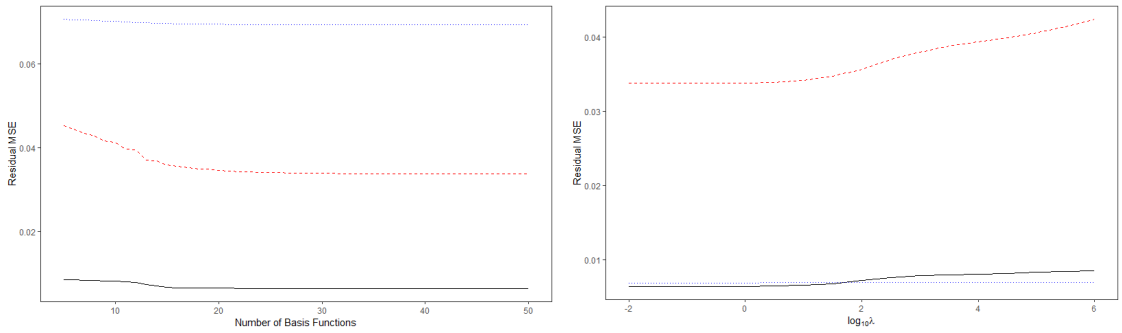
As described in Section 4.2.2, it is possible to improve the computational performance of PDA by estimating the coefficient functions, $\{\beta_j(t)\}$ using basis functions and penalised smoothing. A natural question to ask is whether or not the number of basis functions, or choice of penalty, influences the residual functions to a significant extent. If this is the case, then this would influence the output of FAST. Conversely, if this was not the case then penalised basis smoothing could be implemented to improve the computational performance of FAST without hindering detection power.

To assess the impact of the basis system and penalty term in the fitting of the differential operator to a sample of functional data, we again consider the scenarios of Section 4.4.1. We generate 100 non-anomalous functions from each of the data generating processes and perform PDA on this data for a varying number, K, of order six B-Spline basis functions and a varying smoothing parameter, λ. In each iteration of the study we record the sum of square error of the residuals, given by $\sum_{i=1}^{100}\sum_{\tau=1}^{500}\epsilon_i^2(\tau)$, and for each K and λ perform 1000 replications of each scenario.

The results for the study are presented in Figure B.2.1. As can be seen, in general neither the choice of K, nor λ, have a significant impact on the mean square error of the fit. This implies that the computational performance of FAST can be enhanced by using penalised smoothing without having a serious effect on anomaly detection results. One comment can be made, however, that for very large values of λ the error does increase somewhat. This can be attributed to the fact that as the penalty term increases in size the model begins to underfit. This is similar to what has been observed in the related challenge of fitting smooth functions using penalised basis systems, as discussed in Section 2.4.1.

B.2.2 Simulation Studies with a Matérn Covariance Kernel

In this section we carry out simulations identical to those in Section 4.4.1 with the only difference being the covariance operator is now specified by a Matérn structure



(a) (b)

Figure B.2.1: Mean square error of the model fit using PDA for varying K (a), and λ (b). In both figures the solid black line denotes the Gaussian process residual setting, the dotted red line the T-process residual setting, and the dashed blue line the no underlying shape setting.

given by:

$$K(s,t) = \frac{3}{40\Gamma(3)} \left(\frac{\sqrt{6}\,(s-t)}{50} \right)^3 K_3 \left(\frac{\sqrt{6}\,(s-t)}{50} \right).$$

Here $\Gamma(\cdot)$ is the Gamma function and $K_\nu(\cdot)$ is the modified Bessel function of the second kind. The difference between this operator and the squared exponential operator used in Section 4.4.1 is that it is not infinitely differentiable (the above model is only twice differentiable), and so the observed functions are far less smooth (Rasmussen and Williams, 2005).

Scenario Alpha	Scenario 1	Scenario 2	Scenario 3
0.01	0.01	0.02	0.01
0.05	0.02	0.02	0.01
0.1	0.04	0.03	0.02
0.2	0.05	0.03	0.03

Table B.2.1: Comparison of empirical false alarm rate for different data generating processes when the threshold has been set using Proposition 3 to control the false alarm rate at α.

The results for the simulations for the threshold are given in Table B.2.1, and the results for anomaly detection performance in Table B.2.2. As can be seen, in general the results are similar in terms of both false alarm control and the power of the detection test, however a small increase in detection delay is also observed. One reason for this may be because the Matérn covariance operator gives rise to rougher functional observations than the squared exponential operator used in Section 4.4.1, and this makes the onset of anomalous behaviour harder to identify.

Order One Model	Gaussian process Residuals		t-process Residuals		No Underlying Shape	
Anomaly Type	Power	ADD	Power	ADD	Power	ADD
Polynomial	1.00	10.01 (12.33)	0.88	5.53 (42.53)	1.00	26.99 (13.46)
Sinusoidal	1.00	2.56 (15.06)	0.88	14.45 (54.37)	1.00	79.00 (19.02)
Loss of Shape	1.00	32.20 (12.00)	0.78	91.45 (24.09)	0.20	169.04 (194.70)

Order Two Model	Gaussian process Residuals		t-process Residuals		No Underlying Shape	
Anomaly Type	Power	ADD	Power	ADD	Power	ADD
Polynomial	1.00	4.95 (39.55)	0.89	23.36 (92.17)	1.00	96.02 14.11
Sinusoidal	1.00	4.34 (19.14)	0.90	31.99 (107.31)	1.00	97.58 (19.03)
Loss of Shape	1.00	4.72 (21.47)	0.81	15.77 (75.77)	0.26	169.04 (200.12)

Table B.2.2: Tables showing the detection power, and average detection delay (and standard deviation of detection delay) for each anomaly.

Appendix C

Chapter 5 - mvFAST

C.1 Proof of Results

C.1.1 Proof of Proposition 6

Proof. To establish this result we will consider each of the three anomaly detection settings in turn. Recall that these settings are: sparse anomalies only, dense anomalies only, and both forms of anomaly. In each of the three cases we will draw on Chebychev's inequality. To make use of this inequality, we will need expressions for the expectation and variance of the test statistic, $\Delta_{ij}(\tau)$, which has been defined in equation (5.2.7).

We observe that by the mean-zero assumption for the residual functions,

$$\mathrm{E}\left(\Delta_{ij}(\tau)\right) = \mathrm{E}\left(\sum_{r=1}^{\tau} \epsilon_{ij}(r)\right) = 0.$$

Further, we note that the variance is given by

$$\mathrm{Var}\left(\Delta_{ij}(\tau)\right) = \sum_{r=1}^{\tau} \mathrm{Var}\left(\epsilon_{ij}(r)\right) + 2 \sum_{1 \le a < b \le \tau} \mathrm{Cov}\left(\epsilon_{ij}(a)\epsilon_{ij}(b)\right).$$

As the innovation functions are identically distributed, we denote the variance as $\mathrm{Var}\left(\Delta_{ij}(\tau)\right) = V_\tau$.

A further result, required for the dense setting, is the mean and variance of $\sum_{j=1}^{p} \Delta_{ij}(\tau)$. By Assumption 3 each innovation function is independent and iden-

tically distributed. Hence the residual functions will be uncorrelated across the p dimensions. Thus

$$\begin{aligned}
\mathrm{E}\left(\sum_{j=1}^{p}\Delta_{ij}(\tau)\right) &= 0, \\
\mathrm{Var}\left(\sum_{j=1}^{p}\Delta_{ij}(\tau)\right) &= \sum_{j=1}^{p}\left[\sum_{r=1}^{\tau}\mathrm{Var}\left(\epsilon_{ij}(r)\right) + 2\sum_{1\le a<b\le\tau}\mathrm{Cov}\left(\epsilon_{ij}(a)\epsilon_{ij}(b)\right)\right] \\
&= pV_\tau.
\end{aligned}$$

Using these results, we are in a position to establish the choices of threshold presented in the proposition.

Sparse Anomalies Only For the sparse setting a false alarm is the event

$$\begin{aligned}
\mathrm{P}\left(\text{False}\right) &= \mathrm{P}\left(\bigcup_{i=1}^{n}\bigcup_{j=1}^{p}\bigcup_{\tau=1}^{T}|\Delta_{ij}(\tau)| \ge \gamma\right) \\
&\le \sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{\tau=1}^{T}\mathrm{P}\left(|\Delta_{ij}(\tau)| \ge \gamma\right), \qquad \text{(C.1.1)}
\end{aligned}$$

using the union bound. Using Chebychev's inequality, and the above results for the expectation and variance of $|\Delta_{ij}(\tau)|$, it can be seen that

$$\mathrm{P}\left(|\Delta_{ij}(\tau)| \ge \gamma\right) \le \frac{V_\tau}{\gamma^2}.$$

Combining this with equation (C.1.1) we see that

$$\begin{aligned}
\sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{\tau=1}^{T}\mathrm{P}\left(|\Delta_{ij}(\tau)| \ge \gamma\right) &\le \sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{\tau=1}^{T}\frac{V_\tau}{\gamma^2} \\
&= \frac{np\sum_{\tau=1}^{T}V_\tau}{\gamma^2}. \qquad \text{(C.1.2)}
\end{aligned}$$

Setting equation (C.1.2) equal to α and rearragning demonstrates that

$$\gamma = \left(\frac{np\left(\sum_{\tau=1}^{T}V_\tau\right)}{\alpha}\right)^{\frac{1}{2}},$$

as required.

Dense Anomalies Only For the dense case we note that a false alarm is given by the event

$$\begin{aligned}\mathrm{P}\left(\text{False}\right) &= \mathrm{P}\left(\bigcup_{i=1}^{n}\bigcup_{\tau=1}^{T}|\sum_{j=1}^{p}\Delta_{ij}(\tau)| - p\gamma \geq \Gamma\right) \\ &\leq \sum_{i=1}^{n}\sum_{\tau=1}^{T}\mathrm{P}\left(|\sum_{j=1}^{p}\Delta_{ij}(\tau)| \geq p\gamma + \Gamma\right).\end{aligned}$$

Again, by a similar application of Chebychev's Inequality we obtain

$$\mathrm{P}\left(\text{False}\right) \leq \frac{np\sum_{\tau=1}^{T}V_\tau}{\left(\Gamma + p\gamma\right)^2}. \tag{C.1.3}$$

Setting the RHS of equation (C.1.3) equal to α and rearranging the subsequent quadratic equation provides us with the expression

$$\Gamma^2 + 2p\gamma\Gamma + \left(p^2\gamma^2 - \frac{np\sum_{\tau=1}^{T}V_\tau}{\alpha}\right) = 0.$$

This can be solved to show that

$$\Gamma = 2\left(\left[\frac{np\left(\sum_{\tau=1}^{T}V_\tau\right)}{\alpha}\right]^{\frac{1}{2}} - p\gamma\right).$$

Here the positive part of the root has been taken to ensure that $\Gamma > 0$. Further, since Γ must be positive, we also require

$$\gamma < \left(\frac{n\sum_{\tau=1}^{T}V_\tau}{p\alpha}\right)^{\frac{1}{2}}.$$

Both Dense and Sparse Anomalies In this final case a false alarm is given by either the sparse, or the dense, thresholds being exceeded. Using the union bound we see that this can be written as

$$\mathrm{P}\left(\text{False}\right) = \mathrm{P}\left(\text{Dense} \cup \text{Sparse}\right) \leq \mathrm{P}\left(\text{Dense}\right) + \mathrm{P}\left(\text{Sparse}\right).$$

Using the same result as in the sparse setting, we see that

$$\mathrm{P}\left(\text{Sparse}\right) \leq \frac{np\sum_{\tau=1}^{T}V_\tau}{\gamma^2}, \tag{C.1.4}$$

and using a similar series of steps as in the dense setting but with $\Gamma = \gamma \log(p)$, we see that

$$\mathrm{P}\left(\text{Dense}\right) \leq \frac{np \sum_{\tau=1}^{T} V_\tau}{\gamma^2 \left(p + \log(p)\right)^2}. \tag{C.1.5}$$

As such, we set the sum of equations (C.1.4) and (C.1.5) equal to α and rearrange to obtain

$$\gamma = \left(\left[\frac{np \left(\sum_{\tau=1}^{T} V_\tau \right)}{\alpha} \right] \left[\frac{(p + \log(p))^2 + 1}{(p + \log(p))^2} \right] \right)^{\frac{1}{2}},$$

as required. □

C.1.2 Proof of Proposition 7

Proof. In the dense setting we split the detection probability into two cases:

$$\mathrm{P(D)} = \mathrm{P}\left(\left[\bigcup_{\tau=1}^{T} |\sum_{j=1}^{p} \Delta_{ij}^{c}(\tau)| > \Gamma + p\gamma \right] \cup \left[\bigcup_{\tau=1}^{T} \bigcap_{j=1}^{p} |\Delta_{ij}^{c}(\tau)| > \gamma \right] \right). \tag{C.1.6}$$

The first term on the RHS of (C.1.6) represents an anomaly with respect to the dense threshold, and the second term an anomaly detected in each dimension with respect to the sparse threshold.

We begin with the upper bound. By the union bound

$$\mathrm{P(D)} \leq \mathrm{P}\left(\bigcup_{\tau=1}^{T} |\sum_{j=1}^{p} \Delta_{ij}^{c}(\tau)| > \Gamma + p\gamma \right) + \mathrm{P}\left(\bigcup_{\tau=1}^{T} \bigcap_{j=1}^{p} |\Delta_{ij}^{c}(\tau)| > \gamma \right).$$

We therefore consider the contributions from the dense and sparse terms separately.

For the dense case we note that

$$\mathrm{P}\left(\bigcup_{\tau=1}^{T} |\sum_{j=1}^{p} \Delta_{ij}^{c}(\tau)| > \Gamma + p\gamma \right) \leq \sum_{\tau=1}^{T} \mathrm{P}\left(|\sum_{j=1}^{p} \Delta_{ij}^{c}(\tau)| > \Gamma + p\gamma \right),$$

and consider the $\Delta_{ij}(\tau)$ term when it has been contaminated by an anomaly. By equation (5.2.7) when the ith observation contains an anomaly the test statistic is $\Delta_{ij}(\tau) = \sum_{r=1}^{\tau} \epsilon_{ij}(r) + \hat{\mathcal{L}}(g_{n+1})(r)$. Using the triangle inequality together with the

fact that $\sum_{r=1}^{\tau} \hat{\mathcal{L}}(g_{n+1})(r) = \mathcal{L}(G_{ij})(\tau)$, we note that

$$\begin{aligned}\sum_{\tau=1}^{T} \mathrm{P}\left(|\sum_{j=1}^{p} \Delta_{ij}^{c}(\tau)| > \Gamma + p\gamma \right) &\leq \sum_{\tau=1}^{T} \mathrm{P}\left(|\sum_{j=1}^{p} \Delta_{ij}(\tau)| + |\sum_{j=1}^{p} \mathcal{L}(G_{ij})(\tau)| > \Gamma + p\gamma \right) \\ &\leq \sum_{\tau=1}^{T} \mathrm{P}\left(|\sum_{j=1}^{p} \Delta_{ij}(\tau)| > \Gamma + p\gamma - |\sum_{j=1}^{p} \mathcal{L}(G_{ij})(\tau)| \right).\end{aligned} \tag{C.1.7}$$

Using a similar approach as to that of Proposition 6, we utilise Chebychev's inequality to place an upper bound on the probability in equation (C.1.6). As such, we find

$$\sum_{\tau=1}^{T} \mathrm{P}\left(|\sum_{j=1}^{p} \Delta_{ij}(\tau)| > \Gamma + p\gamma - |\sum_{j=1}^{p} \mathcal{L}(G_{ij})(\tau)| \right) \leq \sum_{\tau=1}^{T} \left(\frac{pV_{\tau}}{\left(\Gamma + p\gamma - |\sum_{j=1}^{p} \mathcal{L}(G_{ij})(\tau)| \right)^{2}} \right).$$

This is the bound for the dense term in equation (C.1.6).

For the sparse term in equation (C.1.6) we bound the intersection probability using Fréchet's bound:

$$\mathrm{P}\left(\bigcup_{\tau=1}^{T} \bigcap_{j=1}^{p} |\Delta_{ij}^{c}(\tau)| > \gamma \right) \leq \sum_{\tau=1}^{T} \min_{j} \mathrm{P}\left(|\Delta_{ij}(\tau)| > \gamma \right).$$

When then proceed in a similar manner to the dense term, finding that

$$\sum_{\tau=1}^{T} \min_{j} \mathrm{P}\left(|\Delta_{ij}^{c}(\tau)| > \gamma \right) \leq \sum_{\tau=1}^{T} \min_{j} \mathrm{P}\left(|\Delta_{ij}(\tau)| > \gamma - |\mathcal{L}(G_{ij})(\tau)| \right).$$

Again, using Chebychev's inequality results in

$$\begin{aligned}\sum_{\tau=1}^{T} \min_{j} \mathrm{P}\left(|\Delta_{ij}(\tau)| > \gamma - |\mathcal{L}(G_{ij})(\tau)| \right) &\leq \sum_{\tau=1}^{T} \min_{j} \left(\frac{V_{\tau}}{\left(\gamma - |\mathcal{L}(G_{ij})(\tau)| \right)^{2}} \right) \\ &\leq \sum_{\tau=1}^{T} \left(\frac{V_{\tau}}{\max_{j} \left(\gamma - |\mathcal{L}(G_{ij})(\tau)| \right)^{2}} \right).\end{aligned}$$

Putting this together with the bound for the dense case completes the proof for the upper bound.

For the lower bound we begin with equation (C.1.6) and proceed as follows:

$$\begin{aligned}
\mathrm{P(D)} &= \mathrm{P}\left(\left[\bigcup_{\tau=1}^{T} |\sum_{j=1}^{p} \Delta_{ij}^{c}(\tau)| > \Gamma + p\gamma\right] \cup \left[\bigcup_{\tau=1}^{T} \bigcap_{j=1}^{p} |\Delta_{ij}^{c}(\tau)| > \gamma\right]\right) \\
&\geq \mathrm{P}\left(\bigcup_{\tau=1}^{T} |\sum_{j=1}^{p} \Delta_{ij}^{c}(\tau)| > \Gamma + p\gamma\right) \\
&\geq \mathrm{P}\left(|\sum_{j=1}^{p} \Delta_{ij}^{c}(T)| > \Gamma + p\gamma\right) \\
&= \mathrm{P}\left(|\sum_{j=1}^{p} \Delta_{ij}(T) + G_{ij}(T)| > \Gamma + p\gamma\right) \qquad \text{(C.1.8)} \\
&\geq \mathrm{P}\left(|\sum_{j=1}^{p} \Delta_{ij}(T)| > \Gamma + p\gamma + |\sum_{j=1}^{p} G_{ij}(T)|\right),
\end{aligned}$$

where the final line follows from the fact that for any $x, y, \in \mathbb{R}$ we have $|x+y| \geq |x|-|y|$. Labelling $K = \Gamma + p\gamma + |\sum_{j=1}^{p} \mathcal{L}(G_{ij})(T)|$, $\mathrm{E}\left(|\sum_{j=1}^{p} \Delta_{ij}(T)|\right) = \mu_1$, $\mathrm{E}\left[\left(|\sum_{j=1}^{p} \Delta_{ij}(T)|\right)^2\right] = \mu_2$, and $\mathrm{E}\left[\left(|\sum_{j=1}^{p} \Delta_{ij}(T)|\right)^3\right] = \mu_3$, we can utilise the following bound due to Rohatgi and Székely (1992):

$$\mathrm{P}\left(|\sum_{j=1}^{p} \Delta_{ij}(T)| > \Gamma + p\gamma + |\sum_{j=1}^{p} G_{ij}(T)|\right) \geq \frac{-2\mu_1}{K} + \frac{11\mu_2}{4K^2} - \frac{3\mu_3}{4K^3},$$

as required. □

C.1.3 Proof of Corollary 2

Proof. In order to prove this result we need to show that $\lim_{T\to\infty} \mathrm{P}(D) = 1$. We begin from equation (C.1.8) in the proof of Proposition 7. That is,

$$\mathrm{P(D)} \geq \mathrm{P}\left(|\sum_{j=1}^{p} \Delta_{ij}(T) + \mathcal{L}(G_{ij})(T)| > \Gamma + p\gamma\right).$$

The next step is to by divide the inequality in the RHS above by T, giving:

$$\mathrm{P\,(D)} = \mathrm{P}\left(\frac{|\sum_{j=1}^{p} \Delta_{ij}(T) + \mathcal{L}(G_{ij})(T)|}{T} > \frac{\Gamma + p\gamma}{T}\right). \qquad \text{(C.1.9)}$$

Our strategy is now to apply a form of Kolmogorov's Strong Law of Large Numbers (Sen and Singer, 1994) to the RHS of the above equation. That is, we will apply the

Kolmogorov's Strong Law to:

$$\frac{|\sum_{j=1}^{p} \Delta_{ij}(T) + \mathcal{L}(G_{ij})(T)|}{T}.$$

By Kolmogorov's Strong Law

$$\mathrm{P}\left(\lim_{T\to\infty} \frac{|\sum_{j=1}^{p} \Delta_{ij}(T) + \mathcal{L}(G_{ij})(T)|}{T} = \lim_{T\to\infty} \mathrm{E}\left(\frac{|\sum_{j=1}^{p} \Delta_{ij}(T) + \mathcal{L}(G_{ij})(T)|}{T}\right)\right) = 1,$$

and so we will need to compute the expectation in the above expression. Using Jensen's inequality we obtain

$$\mathrm{E}\left(\frac{|\sum_{j=1}^{p} \Delta_{ij}(T) + \mathcal{L}(G_{ij})(T)|}{T}\right) \geq \left|\mathrm{E}\left(\frac{\sum_{j=1}^{p} \Delta_{ij}(T) + \mathcal{L}(G_{ij})(T)}{T}\right)\right|.$$

Further, because $\Delta_{ij}(T)$ is the sum of mean-zero random variables, we find

$$\left|\mathrm{E}\left(\frac{\sum_{j=1}^{p} \Delta_{ij}(T) + \mathcal{L}(G_{ij})(T)}{T}\right)\right| = \left|\frac{\sum_{j=1}^{p} \mathcal{L}(G_{ij})(T)}{T}\right|.$$

By assumption, $\lim_{T\to\infty} \frac{|\sum_{j=1}^{p} \mathcal{L}(G_{ij})(T)|}{T} = M > 0$, and so

$$\lim_{\mathrm{T}\to\infty} \mathrm{E}\left(\frac{|\sum_{j=1}^{p} \Delta_{ij}(T) + \mathcal{L}(G_{ij})(T)|}{T}\right) \geq M > 0. \tag{C.1.10}$$

Using the fact that the residual functions are independent in t, the Strong Law of Large Numbers, and the lower bound on the expectation in equation (C.1.10) on equation (C.1.9) we see that

$$\mathrm{P}\left(\lim_{T\to\infty} \frac{|\sum_{j=1}^{p} \Delta_{ij}(T) + \mathcal{L}(G_{ij})(T)|}{T} > \frac{\Gamma + p\gamma}{T}\right) = 1.$$

This is because the LHS of the inequality converges almost surely to at least $M > 0$ and the RHS converges to zero. Now we recall from the proof of Proposition 7, and in particular equation (C.1.8), that for any T, if $|\sum_{j=1}^{p} \Delta_{ij}(T) + \mathcal{L}(G_{ij})(T)| > \Gamma + p\gamma$, then a detection must have occurred. As a result of this, we obtain

$$1 = \mathrm{P}\left(\lim_{T\to\infty} \frac{|\sum_{j=1}^{p} \Delta_{ij}(T) + \mathcal{L}(G_{ij})(T)|}{T} > \frac{\Gamma + p\gamma}{T}\right) \leq \mathrm{P}\left(\lim_{T\to\infty} D\right).$$

To complete the proof we observe that the probability of detecting an anomaly by time T forms an increasing sequence of events in T and so we can exchange the limit and the probability (Billingsley, 1995). As such, we have that

$$1 = \mathrm{P}\left(\lim_{T\to\infty} D\right) = \lim_{T\to\infty} \mathrm{P}\left(D\right),$$

as required. □

C.1.4 Proof of Proposition 8

Proof. This proof follows similar logic to the proof of Proposition 3.1 in Tickle et al. (2021).

We consider the subset of $\mathcal{T}$, $A_\epsilon = \left\{t_{\lfloor \theta T \rfloor}, t_{\lfloor \theta T \rfloor} + a_1 \ldots, t_{\lfloor \theta T \rfloor} + a_m\right\}$, where $a_m < \epsilon$ for some $0 < \epsilon < 1 - \theta$. This can be visualised as a fine grid of $m + 1$ points over the interval $[\nu, \nu + \epsilon)$. We also let D_{A_ϵ} be the event that a detection takes place on the grid A_ϵ, i.e $D_{A_\epsilon} = \mathrm{P}(\xi < \nu + \epsilon)$.

As $m \to \infty$ the grid spacing becomes vanishingly small. Using the assumption that the residual functions are independent in t, $\lim_{m\to\infty} \frac{|\sum_{j=1}^{p} \mathcal{L}(G_{ij})(t_{\lfloor \theta T \rfloor} + a_m)|}{m} > 0$, and Corollary 2 it is therefore the case that $\mathrm{P}(D_{A_\epsilon}) = 1$. Hence $\xi \leq \nu + a_m$.

As a result of this, we can conclude that as m, $T \to \infty$ it is the case that $\frac{\xi - \nu}{T} \leq \frac{a_m}{T} < \epsilon$, as required. □

C.2 Additional Simulation Results

C.2.1 Further Results on Detection Power and Delay

In this study we perform an identical set of simulations to those carried out in Section 5.4.1, with the exception that $p = 5$ dimensions contain an anomaly. The results are presented in Table C.2.2, and are similar to the case examined in Section 5.4.1 where $p = 1$ dimensions are contaminated with an anomaly.

Process	Single Shape		Multiple Shape		Heavy Tailed		Mixing Innovations		Bimodal Shape	
Anomaly	mvFAST	Dai	mvFAST	Dai	mvFAST	Dai	mvFAST	Dai	mvFAST	Dai
Single	**1**	0.99	**1**	1	**1**	0.94	**1**	**1**	**0.98**	0.81
Multiple	**1**	**1**	**1**	**1**	**1**	0.99	0.91	**1**	**1**	0.91
Polynomial	**1**	**1**	**1**	**1**	0.98	**1**	0.94	**1**	**1**	**1**
Phase	0.14	**1**	0.15	**1**	**0.12**	0.05	0.8	**1**	**0.11**	0.05
Constant	**1**	**1**	**1**	**1**	0.98	**1**	0.97	**1**	0.59	**0.99**
Exponential	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
Consistent	**0.11**	1	**0.06**	0.24	**0.08**	0.92	**0.64**	1	**0.09**	0.75

Table C.2.1: Comparison of detection power of mvFAST and the sequential transform and directional outlyingness approach (Dai) of Dai et al. (2020) in the sparse setting with 50% anomaly contamination. The better result in each scenario is highlighted in **bold**.

Process	Single Shape	Multiple Shape	Heavy Tail	Mixing Innovations	Bimodal Shape
Single	13.56 (8.63)	13.43 (8.62)	17.33 (12.74)	17.60 (45.13)	11.21 (10.10)
Multiple	12.32 (7.11)	10.21 (2.93)	13.69 (8.75)	44.42 (69.81)	8.49 (2.45)
Polynomial	80.56 (23.91)	88.68 (24.65)	93.79 (24.26)	65.66 (78.48)	49.28 (56.89)
Phase	14.55 (38.48)	15.45 (39.42)	13.44 (40.03)	33.90 (65.26)	16.02 (47.76)
Constant	43.35 (26.79)	38.60 (37.58)	58.31 (32.96)	41.53 (73.33)	14.59 (47.60)
Exponential	45.43 (25.73)	38.32 (32.62)	63.43 (35.92)	30.86 (64.80)	74.35 (50.37)
Consistent	11.70 (35.69)	6.30 (26.42)	7.25 (32.96)	59.74 (81.71)	14.32 (47.50)

Table C.2.2: Average detection delay (and standard deviation) of mvFAST in the sparse setting with 50% anomaly contamination.

Bibliography

Ryan Prescott Adams and David JC MacKay. Bayesian online changepoint detection, 2007. URL https://doi.org/10.48550/arXiv.0710.3742.

Diego Agudelo-España, Sebastian Gomez-Gonzalez, Stefan Bauer, Bernhard Schölkopf, and Jan Peters. Bayesian online prediction of change points. In *Conference on Uncertainty in Artificial Intelligence*, pages 320–329. PMLR, 2020. URL https://doi.org/10.48550/arXiv.1902.04524.

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. URL https://doi.org/10.1007/s10115-016-0987-z.

Aurore Archimbaud, Feriel Boulfani, Xavier Gendre, Klaus Nordhausen, Anne Ruiz-Gazen, and Joni Virta. Ics for multivariate functional anomaly detection with applications to predictive maintenance and quality control. *Econometrics and Statistics*, 2022. ISSN 2452-3062. doi: https://doi.org/10.1016/j.ecosta.2022.03.003. URL https://www.sciencedirect.com/science/article/pii/S2452306222000247.

Ana Arribas-Gil and Juan Romo. Shape outlier detection and visualization for functional data: the outliergram. *Biostatistics*, 15(4):603–619, 03 2014. ISSN 1465-4644. doi: 10.1093/biostatistics/kxu006. URL https://doi.org/10.1093/biostatistics/kxu006.

Alexander Aue, Gregory Rice, and Ozan Sönmez. Detecting and dating structural breaks in functional data without dimension reduction. *Journal of the Royal*

Statistical Society: Series B (Statistical Methodology), 80(3):509–529, 2018. doi: 10.1111/rssb.12257. URL https://doi.org/10.1111/rssb.12257.

Edward Austin, Gaetano Romano, Idris A. Eckley, and Paul Fearnhead. Online non-parametric changepoint detection with application to monitoring operational performance of network devices. *Computational Statistics & Data Analysis*, 177:107551, 2023. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2022.107551. URL https://www.sciencedirect.com/science/article/pii/S0167947322001311.

Kahtan Aziz, Saed Tarapiah, Salah Haj Ismail, and Shadi Atalla. Smart real-time healthcare monitoring and tracking system using gsm/gps technologies. In *2016 3rd MEC International Conference on Big Data and Smart City (ICBDSC)*, pages 1–7. IEEE, 2016. doi: 10.1109/ICBDSC.2016.7460394. URL https://doi.org/10.1109/ICBDSC.2016.7460394.

M. Baron and A. G. Tartakovsky. Asymptotic optimality of change-point detection schemes in general continuous-time models. *Sequential Analysis*, 25(3):257–296, 2006. doi: 10.1080/07474940600609597. URL https://doi.org/10.1080/07474940600609597.

C. Barreyre, B. Laurent, J. M. Loubes, L. Boussouf, and B. Cabon. Multiple testing for outlier detection in space telemetries. *IEEE Transactions on Big Data*, 6(3):443–451, 2020. doi: 10.1109/TBDATA.2019.2954831. URL https://doi.org/10.1109/TBDATA.2019.2954831.

Peter Bauer and Peter Hackl. The use of mosums for quality control. *Technometrics*, 20(4):431–436, 1978. ISSN 00401706. doi: 10.1080/00401706.1978.10489697. URL https://doi.org/10.1080/00401706.1978.10489697.

P. Billingsley. *Probability and Measure*. Wiley Series in Probability and Statistics. Wiley, 1995. ISBN 9780471007104. URL https://books.google.co.uk/books?id=z39jQgAACAAJ.

J. M. Blair, C. A. Edwards, and J. H. Johnson. Rational chebyshev approximations for the inverse of the error function. *Mathematics of Computation*, 30(136):827–830, 1976. ISSN 00255718, 10886842. doi: 10.2307/2005402. URL https://doi.org/10.2307/2005402.

C. Bradley. Basic properties of strong mixing conditions. a survey and some open questions. *Probability Surveys*, 2:107–144, 11 2005. doi: 10.1214/154957805100000104.

Michael Byrd, Linh Nghiem, and Jing Cao. Lagged exact bayesian online changepoint detection with parameter estimation, 2018. URL https://doi.org/10.48550/arXiv.1710.03276.

Yang Cao, Liyan Xie, Yao Xie, and Huan Xu. Sequential change-point detection via online convex optimization. *Entropy*, 20(2), 2018. ISSN 1099-4300. doi: 10.3390/e20020108. URL https://www.mdpi.com/1099-4300/20/2/108.

Michelle Carey and James O Ramsay. Fast stable parameter estimation for linear dynamical systems. *Computational Statistics & Data Analysis*, 156:107124, 2020. doi: 10.1016/j.csda.2020.107124. URL https://doi.org/10.1016/j.csda.2020.107124.

Subhabrata Chakraborti and Mark A. van de Wiel. A nonparametric control chart based on the mann-whitney statistic. *Institute of Mathematical Statistics Collections*, page 156–172, 2008. doi: 10.1214/193940307000000112. URL http://dx.doi.org/10.1214/193940307000000112.

Hao Chen. Sequential change-point detection based on nearest neighbors. *The Annals of Statistics*, 47(3):1381–1407, 06 2019. doi: 10.1214/18-AOS1718. URL https://doi.org/10.1214/18-AOS1718.

Jeng-Min Chiou, Yu-Ting Chen, and Ya-Fang Yang. Multivariate functional principal component analysis: A normalization approach. *Statistica Sinica*, 24(4):1571–1596, 2014. ISSN 10170405, 19968507. doi: 10.5705/ss.2013.305. URL http://dx.doi.org/10.5705/ss.2013.305.

Chia-Shang J. Chu, Kurt Hornik, and Chung-Ming Kuan. Mosum tests for parameter constancy. *Biometrika*, 82(3):603–617, 1995. ISSN 00063444. doi: 10.1093/biomet/82.3.603. URL https://doi.org/10.1093/biomet/82.3.603.

Matthias Chung, Justin Krueger, and Mihai Pop. Identification of microbiota dynamics using robust parameter estimation methods. *Mathematical biosciences*, 294: 71–84, 2017. doi: 10.1016/j.mbs.2017.09.009. URL https://doi.org/10.1016/j.mbs.2017.09.009.

M.L.I. Coelho, M.A. Graham, and S. Chakraborti. Nonparametric signed-rank control charts with variable sampling intervals. *Quality and Reliability Engineering International*, 33(8):2181–2192, 2017. doi: 10.1002/qre.2177. URL https://doi.org/10.1002/qre.2177.

Bianca M. Colosimo and Massimo Pacella. A comparison study of control charts for statistical monitoring of functional data. *International Journal of Production Research*, 48(6):1575–1601, 2010. doi: 10.1080/00207540802662888. URL https://doi.org/10.1080/00207540802662888.

T.H. Cormen, T.H. Cormen, C.E. Leiserson, Inc Books24x7, Massachusetts Institute of Technology, MIT Press, R.L. Rivest, C. Stein, and McGraw-Hill Publishing Company. *Introduction To Algorithms*. Introduction to Algorithms. MIT Press, 2001. ISBN 9780262032933. URL https://books.google.co.uk/books?id=NLngYyWFl_YC.

Wenlin Dai and Marc Genton. Multivariate functional data visualization and outlier detection. *Journal of Computational and Graphical Statistics*, page 30, 03 2018a. doi: 10.1080/10618600.2018.1473781. URL https://doi.org/10618600.2018.1473781.

Wenlin Dai and Marc Genton. Functional boxplots for multivariate curves: Multivariate curves. *Stat*, 7:e190, 08 2018b. doi: 10.1002/sta4.190. URL https://doi.org/10.1002/sta4.190.

Wenlin Dai and Marc G. Genton. Directional outlyingness for multivariate functional data. *Computational Statistics & Data Analysis*, 131:50 – 65, 2019. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2018.03.017. URL http://www.sciencedirect.com/science/article/pii/S016794731830077X.

Wenlin Dai, Tomáš Mrkvička, Ying Sun, and Marc G. Genton. Functional outlier detection and taxonomy by sequential transformations. *Computational Statistics & Data Analysis*, 149:106960, 2020. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2020.106960. URL https://www.sciencedirect.com/science/article/pii/S0167947320300517.

J. Dauxois, A. Pousse, and Y. Romain. Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference. *Journal of Multivariate Analysis*, 12(1):136–154, 1982. ISSN 0047-259X. doi: https://doi.org/10.1016/0047-259X(82)90088-4. URL https://www.sciencedirect.com/science/article/pii/0047259X82900884.

C. DeBoor. *A Practical Guide to Splines*. Springer New York, 2000. ISBN 9780387989228. URL https://books.google.co.uk/books?id=a4yHkQEACAAJ.

Holger Dette and Josua Gösmann. A likelihood ratio approach to sequential change point detection for a general class of parameters. *Journal of the American Statistical Association*, 115(531):1361–1377, 2020. doi: 10.1080/01621459.2019.1630562. URL https://doi.org/10.1080/01621459.2019.1630562.

Yuping Dong, A. S. Hedayat, and B. K. Sinha. Surveillance strategies for detecting changepoint in incidence rate based on exponentially weighted moving average methods. *Journal of the American Statistical Association*, 103(482):843–853, 2008. ISSN 01621459. doi: 10.1198/016214508000000166. URL http://www.jstor.org/stable/27640106.

Joel A Dubin and Hans-Georg Müller. Dynamical correlation for multivariate longitudinal data. *Journal of the American Statistical Association*, 100(471):872–

881, 2005. doi: 10.1198/016214504000001989. URL https://doi.org/10.1198/016214504000001989.

B. Eichinger and C. Kirch. A mosum procedure for the estimation of multiple random change points. *Bernoulli*, 24:526–564, 02 2018. doi: 10.3150/16-BEJ887. URL https://doi.org/10.3150/16-BEJ887.

Amira Elayouty, Marian Scott, and Claire Miller. Time-varying functional principal components for non-stationary epCO_2 in freshwater systems. *Journal of Agricultural, Biological and Environmental Statistics*, pages 1–17, 2022. doi: 10.1007/s13253-022-00494-2. URL https://doi.org/10.1007/s13253-022-00494-2.

F. Ferraty and P. Vieu. *Nonparametric Functional Data Analysis: Theory and Practice*. Springer Series in Statistics. Springer New York, 2006. ISBN 9780387366203. URL https://books.google.co.uk/books?id=lMy6WPFZYFcC.

Frederic Ferraty and Yves Romain. *The Oxford Handbook of Functional Data Analysis*. Oxford University Press, 01 2011.

Alexander Fisch, Lawrence Bardwell, and Idris A Eckley. Real time anomaly detection and categorisation. *Statistics and Computing*, 32(4):1–15, 2022a. doi: 10.1007/s11222-022-10112-3. URL https://doi.org/10.1007/s11222-022-10112-3.

Alexander T. M. Fisch, Idris A. Eckley, and Paul Fearnhead. A linear time method for the detection of collective and point anomalies. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 15(4):494–508, 2022b. doi: https://doi.org/10.1002/sam.11586. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sam.11586.

Natasha Zhang Foutz and Wolfgang Jank. Research note—prerelease demand forecasting for motion pictures using functional shape analysis of virtual stock markets. *Marketing Science*, 29(3):568–579, 2010. doi: 10.1287/mksc.1090.0542. URL https://doi.org/10.1287/mksc.1090.0542.

Cheng-Der Fuh. Sprt and cusum in hidden markov models. *The Annals of Statistics*, 31(3):942–977, 2003. ISSN 00905364. doi: 10.1214/aos/1056562468. URL http://www.jstor.org/stable/3448426.

João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM computing surveys (CSUR)*, 46(4):1–37, 2014. doi: 10.1145/2523813. URL https://doi.org/10.1145/2523813.

Ramadha D Piyadi Gamage and Wei Ning. Empirical likelihood for change point detection in autoregressive models. *Journal of the Korean Statistical Society*, pages 1–29, 2020. doi: 10.1007/s42952-020-00061-w. URL https://doi.org/10.1007/s42952-020-00061-w.

Louis Gordon and Moshe Pollak. An Efficient Sequential Nonparametric Scheme for Detecting a Change of Distribution. *The Annals of Statistics*, 22(2):763 – 804, 1994. doi: 10.1214/aos/1176325495. URL https://doi.org/10.1214/aos/1176325495.

Tomasz Górecki, Mirosław Krzyśko, Łukasz Waszak, and Waldemar Wołyński. Selected statistical methods of data analysis for multivariate functional data. *Statistical Papers*, 59(1):153–182, 2018. doi: 10.1007/s00362-016-0757-8. URL https://doi.org/10.1007/s00362-016-0757-8.

Josua Gösmann, Tobias Kley, and Holger Dette. A new approach for open-end sequential change point monitoring. *Journal of Time Series Analysis*, 42(1):63–84, 2020. doi: 10.1111/jtsa.12555. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/jtsa.12555.

Peter Hall and Mohammad Hosseini-Nasab. On properties of functional principal components analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):109–126, 2006. doi: 10.1111/j.1467-9868.2005.00535.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2005.00535.x.

Peter Hall, D S Poskitt, and Brett Presnell. A functional data—analytic approach to signal discrimination. *Technometrics*, 43(1):1–9, 2001. doi: 10.1198/00401700152404273. URL https://doi.org/10.1198/00401700152404273.

Peter Hall, Hans-Georg Müller, and Fang Yao. Modelling sparse generalized longitudinal observations with latent gaussian processes. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 70(4):703–723, 2008. ISSN 13697412, 14679868. doi: 10.1111/j.1467-9868.2008.00656.x. URL https://doi.org/10.1111/j.1467-9868.2008.00656.x.

Trevor Harris, J. Derek Tucker, Bo Li, and Lyndsay Shand. Elastic depths for detecting shape anomalies in functional data. *Technometrics*, 0(0):1–11, 2020. doi: 10.1080/00401706.2020.1811156. URL https://doi.org/10.1080/00401706.2020.1811156.

Douglas M. Hawkins and Qiqi Deng. A nonparametric change-point control chart. *Journal of Quality Technology*, 42(2):165–173, 2010. doi: 10.1080/00224065.2010.11917814. URL https://doi.org/10.1080/00224065.2010.11917814.

Kaylea Haynes, Paul Fearnhead, and Idris A. Eckley. A computationally efficient nonparametric approach for changepoint detection. *Statistics and Computing*, 27 (5):1293–1305, Sep 2017. doi: 10.1007/s11222-016-9687-5. URL https://doi.org/10.1007/s11222-016-9687-5.

Giles Hooker. Forcing function diagnostics for nonlinear dynamics. *Biometrics*, 65(3):928–936, 2009. doi: https://doi.org/10.1111/j.1541-0420.2008.01172.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1541-0420.2008.01172.x.

Giles Hooker and Stephen P. Ellner. Goodness of fit in nonlinear dynamics: Misspecified rates or misspecified states? *The Annals of Applied Statistics*, 9(2):754 – 776, 2015. doi: 10.1214/15-AOAS828. URL https://doi.org/10.1214/15-AOAS828.

Siegfried Hörmann, Piotr Kokoszka, et al. Weakly dependent functional data. *The Annals of Statistics*, 38(3):1845–1884, 2010. doi: 10.1214/09-AOS768. URL https://doi.org/10.1214/09-AOS768.

Tailen Hsing and Randall Eubank. *Theoretical foundations of functional data analysis, with an introduction to linear operators*, volume 997. John Wiley & Sons, 2015.

Chih-Chiang Hsu. The mosum of squares test for monitoring variance changes. *Finance Research Letters*, 4(4):254–260, 2007. doi: 10.1016/j.frl.2007.09.003. URL https://doi.org/10.1016/j.frl.2007.09.003.

Huang Huang and Ying Sun. A decomposition of total variation depth for understanding functional outliers. *Technometrics*, 61(4):445–458, 2019. doi: 10.1080/00401706.2019.1574241. URL https://doi.org/10.1080/00401706.2019.1574241.

Mia Hubert, Peter J Rousseeuw, and Pieter Segaert. Multivariate functional outlier detection. *Statistical Methods & Applications*, 24(2):177–202, 2015. doi: 10.1007/s10260-015-0297-8. URL https://doi.org/10.1007/s10260-015-0297-8.

J. Stuart Hunter. The exponentially weighted moving average. *Journal of Quality Technology*, 18(4):203–210, 1986. doi: 10.1080/00224065.1986.11979014. URL https://doi.org/10.1080/00224065.1986.11979014.

Marie Hušková. Asymptotics for robust mosum. *Commentationes Mathematicae Universitatis Carolinae*, 31(2):345–356, 1990. URL https://dml.cz/handle/10338.dmlcz/106864.

Marie Hušková and Zdeněk Hlávka. Nonparametric sequential monitoring. *Sequential Analysis*, 31(3):278–296, 2012. doi: 10.1080/07474946.2012.694345. URL https://www.tandfonline.com/doi/abs/10.1080/07474946.2012.694345.

Rob J. Hyndman. Computing and graphing highest density regions. *The American Statistician*, 50(2):120–126, 1996. ISSN 00031305. doi: 10.2307/2684423. URL https://doi.org/10.2307/2684423.

Rob J. Hyndman and Md. Shahid Ullah. Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics & Data Analysis*, 51 (10):4942–4956, June 2007. doi: 10.1016/j.csda.2006.07.028. URL `https://doi.org/10.1016/j.csda.2006.07.028`.

Rob J. Hyndman and Han Lin Shang. Rainbow plots, bagplots, and boxplots for functional data. *Journal of Computational and Graphical Statistics*, 19(1):29–45, 2010. doi: 10.1198/jcgs.2009.08158. URL `https://doi.org/10.1198/jcgs.2009.08158`.

Siegfried Hörmann, Łukasz Kidziński, and Marc Hallin. Dynamic functional principal components. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(2):319–348, 2015. doi: 10.1111/rssb.12076. URL `https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12076`.

Francesca Ieva and Anna M. Paganoni. Depth measures for multivariate functional data. *Communications in Statistics - Theory and Methods*, 42(7):1265–1276, 2013. doi: 10.1080/03610926.2012.746368. URL `https://doi.org/10.1080/03610926.2012.746368`.

Gabriel Jarry, Daniel Delahaye, Florence Nicol, and Eric Feron. Aircraft atypical approach detection using functional principal component analysis. *Journal of Air Transport Management*, 84:101787, 2020. ISSN 0969-6997. doi: https://doi.org/10.1016/j.jairtraman.2020.101787. URL `http://www.sciencedirect.com/science/article/pii/S0969699719303266`.

M. K. Jeong, J.-C. Lu, and N. Wang. Wavelet-based spc procedure for complicated functional data. *International Journal of Production Research*, 44(4):729–744, 2006. doi: 10.1080/00207540500222647. URL `https://doi.org/10.1080/00207540500222647`.

Seoweon Jin, Joan G Staniswalis, and Indika Mallawaarachchi. Principal differential analysis with a continuous covariate: low-dimensional approximations for functional

data. *Journal of statistical computation and simulation*, 83(10):1964–1980, 2013. doi: 10.1080/00949655.2012.675575. URL https://doi.org/10.1080/00949655.2012.675575.

N.L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions*. Number v. 2 in Continuous Univariate Distributions. Wiley & Sons, 1994. ISBN 9780471584940. URL https://books.google.co.uk/books?id=0QzvAAAAMAAJ.

P Johnson, J Moriarty, and G Peskir. Detecting changes in real-time data: a user's guide to optimal detection. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 375(2100):20160298, 2017. doi: 10.1098/rsta.2016.0298. URL https://doi.org/10.1098/rsta.2016.0298.

R.A. Johnson and D.W. Wichern. *Applied Multivariate Statistical Analysis*. Prentice-Hall series in statistics. Prentice-Hall, 1988. ISBN 9780130411464. URL https://books.google.co.uk/books?id=2lYZAQAAIAAJ.

S. Karlin and H.E. Taylor. *A First Course in Stochastic Processes*. Elsevier Science, 2012. ISBN 9780080570419. URL https://books.google.co.uk/books?id=dSDxjX9nmmMC.

Nicolas Keriven, Damien Garreau, and Iacopo Poli. Newma: A new method for scalable model-free online change-point detection. *IEEE Transactions on Signal Processing*, 68:3515–3528, 2020. doi: 10.1109/TSP.2020.2990597. URL https://ieeexplore.ieee.org/document/9078835.

Claudia Kirch and Christina Stoehr. Sequential change point tests based on u-statistics, 2019. URL https://doi.org/10.48550/arXiv.1912.08580.

Claudia Kirch and Silke Weber. Modified sequential change point procedures based on estimating functions. *Electron. J. Statist.*, 12(1):1579–1613, 2018. doi: 10.1214/18-EJS1431. URL https://doi.org/10.1214/18-EJS1431.

Piotr Kokoszka and Matthew Reimherr. *Introduction to functional data analysis*. Chapman and Hall/CRC, 2017.

Piotr Kokoszka and Gabriel Young. Testing trend stationarity of functional time series with application to yield and daily price curves. *Statistics and its interface*, 10:81–92, 01 2017. doi: 10.4310/SII.2017.v10.n1.a8. URL https://doi.org/10.4310/SII.2017.v10.n1.a8.

J. Kowalski and X.M. Tu. *Modern Applied U-Statistics*. Wiley Series in Probability and Statistics. Wiley, 2008. ISBN 9780470186459. URL https://books.google.co.uk/books?id=tWPRBaQXCbAC.

Sangyeol Lee. Location and scale-based cusum test with application to autoregressive models. *Journal of Statistical Computation and Simulation*, 90(13):2309–2328, 2020. doi: 10.1080/00949655.2020.1775833. URL https://doi.org/10.1080/00949655.2020.1775833.

Clément Lejeune, Josiane Mothe, Adil Soubki, and Olivier Teste. Shape-based outlier detection in multivariate functional data. *Knowledge-Based Systems*, 198:105960, 2020a. doi: 10.1016/j.knosys.2020.105960. URL https://doi.org/10.1016/j.knosys.2020.105960.

Clément Lejeune, Josiane Mothe, and Olivier Teste. Outlier detection in multivariate functional data based on a geometric aggregation. In *EDBT - International Conference on Extending Database TechnologyAt: Copenhagen*, pages 1–4, 04 2020b. doi: 10.5441/002/edbt.2020.38. URL https://doi.org/10.5441/002/edbt.2020.38.

David A Leon, Vladimir M Shkolnikov, Liam Smeeth, Per Magnus, Markéta Pechholdová, and Christopher I Jarvis. Covid-19: a need for real-time monitoring of weekly excess deaths. *The Lancet*, 395(10234):e81, 2020. doi: 10.1016/S0140-6736(20)30933-8. URL https://doi.org/10.1016/S0140-6736(20)30933-8.

Liu Liu, Xuemin Zi, Jian Zhang, and Zhaojun Wang. A sequential rank-based nonparametric adaptive ewma control chart. *Communications in Statistics - Simulation and Computation*, 42(4):841–859, 2013. doi: 10.1080/03610918.2012.655829. URL https://doi.org/10.1080/03610918.2012.655829.

Regina Y. Liu, Jesse M. Parelius, and Kesar Singh. Multivariate analysis by data depth: Descriptive statistics, graphics and inference. *The Annals of Statistics*, 27 (3):783–840, 1999. ISSN 00905364. doi: 10.1214/aos/1018031260. URL `http://www.jstor.org/stable/120138`.

Sara Lopez-Pintado and Juan Romo. Depth-based inference for functional data. *Computational Statistics & Data Analysis*, 51(10):4957–4968, 2007. doi: 10.1016/j.csda.2006.10.029. URL `https://doi.org/10.1016/j.csda.2006.10.029`.

Sara López-Pintado, Ying Sun, Juan K Lin, and Marc G Genton. Simplicial band depth for multivariate functional data. *Advances in Data Analysis and Classification*, 8(3):321–338, 2014. doi: 10.1007/s11634-014-0166-6. URL `https://doi.org/10.1007/s11634-014-0166-6`.

G. Lorden. Procedures for reacting to a change in distribution. *Ann. Math. Statist.*, 42(6):1897–1908, 12 1971. doi: 10.1214/aoms/1177693055. URL `https://doi.org/10.1214/aoms/1177693055`.

H. Lütkepohl. *Introduction to Multiple Time Series Analysis*. Springer Berlin Heidelberg, 1993. ISBN 9783540569404. URL `https://books.google.co.uk/books?id=ANB-8BdjbsYC`.

María Luz López García, Ricardo García-Ródenas, and Antonia González Gómez. K-means algorithms for functional data. *Neurocomputing*, 151:231–245, 2015. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2014.09.048. URL `https://www.sciencedirect.com/science/article/pii/S0925231214012521`.

Sara López-Pintado and Juan Romo. On the concept of depth for functional data. *Journal of the American Statistical Association*, 104(486):718–734, 2009. ISSN 01621459. doi: 10.1198/jasa.2009.0108. URL `https://doi.org/10.1198/jasa.2009.0108`.

H. B. Mann and D. R. Whitney. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*,

18(1):50 – 60, 1947. doi: 10.1214/aoms/1177730491. URL https://doi.org/10.1214/aoms/1177730491.

Israel Martínez-Hernández and Marc G. Genton. Nonparametric trend estimation in functional time series with application to annual mortality rates. *Biometrics*, 77(3):866–878, 2021. doi: https://doi.org/10.1111/biom.13353. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.13353.

Sophie Mathieu, Rainer von Sachs, Véronique Delouille, Laure Lefèvre, and Christian Ritter. Nonparametric robust monitoring of time series panel data, 2020. URL https://doi.org/10.48550/arXiv.2010.11826.

Prisma Megantoro, Brahmantya Aji Pramudita, P Vigneshwaran, Abdufattah Yurianta, and Hendra Ari Winarno. Real-time monitoring system for weather and air pollutant measurement with html-based ui application. *Bulletin of Electrical Engineering and Informatics*, 10(3):1669–1677, 2021. doi: 10.11591/eei.v10i3.3030. URL https://doi.org/10.11591/eei.v10i3.3030.

Alexander Meier, Claudia Kirch, and Haeran Cho. mosum: A package for moving sums in change-point analysis. *Journal of Statistical Software*, 97(8):1–42, 2021. ISSN 1548-7660. doi: 10.18637/jss.v097.i08. URL https://www.jstatsoft.org/v097/i08.

James Mercer and Andrew Russell Forsyth. Xvi. functions of positive and negative type, and their connection the theory of integral equations. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 209(441-458):415–446, 1909. doi: 10.1098/rsta.1909.0016. URL https://royalsocietypublishing.org/doi/abs/10.1098/rsta.1909.0016.

A.M. Mood and F.A. Graybill. *Introduction to the Theory of Statistics*. McGraw-Hill series in probability and statistics. McGraw-Hill, 1963. URL https://books.google.co.uk/books?id=FPRQAAAAMAAJ.

Ahmad Mozaffari, Shojaeddin Chenouri, Ehsan Toyserkani, and Usman Ali. Functional boxplots for outlier detection in additive manufacturing, 2021. URL https://doi.org/10.48550/arXiv.2110.10867.

A. Mukherjee and S. Chakraborti. A distribution-free control chart for the joint monitoring of location and scale. *Quality and Reliability Engineering International*, 28(3):335–352, 2012. doi: 10.1002/qre.1249. URL https://doi.org/10.1002/qre.1249.

Hidetoshi Murakami and Takashi Matsuki. A nonparametric control chart based on the mood statistic for dispersion. *The International Journal of Advanced Manufacturing Technology*, 49:757–763, 2010. doi: 10.1007/s00170-009-2439-3. URL https://doi.org/10.1007/s00170-009-2439-3.

Stanislav Nagy, Irène Gijbels, and Daniel Hlubinka. Depth-based recognition of shape outlying functions. *Journal of Computational and Graphical Statistics*, 26(4):883–893, 2017. doi: 10.1080/10618600.2017.1336445. URL https://doi.org/10.1080/10618600.2017.1336445.

Naveen N. Narisetty and Vijayan N. Nair. Extremal depth for functional data and applications. *Journal of the American Statistical Association*, 111(516):1705–1714, 2016. doi: 10.1080/01621459.2015.1110033. URL https://doi.org/10.1080/01621459.2015.1110033.

Oluwasegun Ojo, Antonio Fernández Anta, Rosa Lillo, and Carlo Sguera. Detecting and classifying outliers in big functional data. *Advances in Data Analysis and Classification*, pages 1–36, 08 2021a. doi: 10.1007/s11634-021-00460-9. URL https://doi.org/10.1007/s11634-021-00460-9.

Oluwasegun Ojo, Rosa E. Lillo, and Antonio Fernández Anta. Outlier detection for functional data with r package fdaoutlier, 2021b. URL https://doi.org/10.48550/arXiv.2105.05213.

Oluwasegun Taiwo Ojo, Antonio Fernández Anta, Marc G. Genton, and Rosa E. Lillo. Multivariate functional outlier detection using the fastmuod indices, 2022. URL `https://arxiv.org/abs/2207.12803`.

Moses Oluwafemi Onibonoje and Temitayo Olayemi Olowu. Real-time remote monitoring and automated control of granary environmental factors using wireless sensor network. In *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*, pages 113–118. IEEE, 2017. doi: 10.1109/ICPCSI.2017.8391925. URL `https://doi.org/10.1109/ICPCSI.2017.8391925`.

Oscar Hernan Madrid Padilla, Alex Athey, Alex Reinhart, and James G. Scott. Sequential nonparametric tests for a change in distribution: An application to detecting radiological anomalies. *Journal of the American Statistical Association*, 114(526):514–528, 2019a. doi: 10.1080/01621459.2018.1476245. URL `https://doi.org/10.1080/01621459.2018.1476245`.

Oscar Hernan Madrid Padilla, Alex Athey, Alex Reinhart, and James G. Scott. Sequential nonparametric tests for a change in distribution: An application to detecting radiological anomalies. *Journal of the American Statistical Association*, 114(526):514–528, 2019b. doi: 10.1080/01621459.2018.1476245. URL `https://doi.org/10.1080/01621459.2018.1476245`.

Ewan S Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954. doi: 10.1093/biomet/41.1-2.100. URL `http://biomet.oxfordjournals.org/cgi/doi/10.1093/biomet/41.1-2.100`.

Debashis Paul, Jie Peng, and Prabir Burman. Semiparametric modeling of autonomous nonlinear dynamical systems with application to plant growth. *Ann. Appl. Stat.*, 5(3):2078–2108, 09 2011. doi: 10.1214/11-AOAS459. URL `https://doi.org/10.1214/11-AOAS459`.

Kamran Paynabar, Changliang Zou, and Peihua Qiu. A change-point approach for phase-i analysis in multivariate profile monitoring and diagnosis. *Technometrics*, 58

(2):191–204, 2016. doi: 10.1080/00401706.2015.1042168. URL https://doi.org/10.1080/00401706.2015.1042168.

K. Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2 (11):559–572, 1901. doi: 10.1080/14786440109462720. URL https://doi.org/10.1080/14786440109462720.

Andrey Pepelyshev and Aleksey Polunchenko. Real-time financial surveillance via quickest change-point detection methods. *Statistics and its interface*, 10:93–106, 01 2017. doi: 10.4310/SII.2017.v10.n1.a9. URL https://doi.org/10.4310/SII.2017.v10.n1.a9.

Sara Pintado and Juan Romo. A half-region depth for functional data. *Computational Statistics & Data Analysis*, 55:1679–1695, 04 2011. doi: 10.1016/j.csda.2010.10.024. URL https://doi.org/10.1016/j.csda.2010.10.024.

Joshua Plasse, Henrique Hoeltgebaum, and Niall M Adams. Streaming changepoint detection for transition matrices. *Data Mining and Knowledge Discovery*, pages 1–30, 2021. doi: 10.1007/s10618-021-00747-7. URL https://doi.org/10.1007/s10618-021-00747-7.

Werner Ploberger and Walter Kramer. The CUSUM Test with OLS Residuals. *Econometrica*, 60(2):271–285, March 1992. doi: 10.2307/2951597. URL https://www.jstor.org/stable/2951597.

Aleksey S Polunchenko and Alexander G Tartakovsky. State-of-the-art in sequential change-point detection. *Methodology and computing in applied probability*, 14(3): 649–684, 2012. doi: 10.1007/s11009-011-9256-5. URL https://doi.org/10.1007/s11009-011-9256-5.

P. Qiu. *Introduction to Statistical Process Control*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2013. ISBN 9781439847992. URL https://books.google.co.uk/books?id=DWRmAQAAQBAJ.

Peihua Qiu and Dongdong Xiang. Univariate dynamic screening system: An approach for identifying individuals with irregular longitudinal behavior. *Technometrics*, 56: 248–260, 05 2014. doi: 10.1080/00401706.2013.822423. URL https://doi.org/10.1080/00401706.2013.822423.

Zhuo Qu and Marc G. Genton. Sparse functional boxplots for multivariate curves. *Journal of Computational and Graphical Statistics*, 0(0):1–14, 2022. doi: 10.1080/10618600.2022.2066680. URL https://doi.org/10.1080/10618600.2022.2066680.

J. Ramsay and G. Hooker. *Dynamic Data Analysis: Modeling Data with Differential Equations*. Springer Series in Statistics. Springer New York, 2017. ISBN 9781493971909. URL https://books.google.co.uk/books?id=FgUqDwAAQBAJ.

J. O. Ramsay. Principal differential analysis: Data reduction by differential operators. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(3):495–508, 1996. ISSN 00359246. doi: 10.1111/j.2517-6161.1996.tb02096.x. URL https://doi.org/10.1111/j.2517-6161.1996.tb02096.x.

J. O. Ramsay, Spencer Graves, and Giles Hooker. *fda: Functional Data Analysis*, 2021. URL https://CRAN.R-project.org/package=fda. R package version 5.5.1.

James Ramsay. *Functional Data Analysis*. Springer, Dordrecht, 2nd ed. edition, 2005. ISBN 9780387227511.

C. Radhakrishna Rao. Prediction of future observations in growth curve models. *Statistical Science*, 2(4):434–447, 1987. ISSN 08834237. doi: 10.1214/ss/1177013119. URL http://www.jstor.org/stable/2245538.

C.E. Rasmussen and C.K.I. Williams. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning series. MIT Press, 2005. ISBN 9780262182539. URL https://books.google.co.uk/books?id=GhoSngEACAAJ.

Haider Raza, G. Prasad, and Yuhua Li. Ewma model based shift-detection methods for detecting covariate shifts in non-stationary environments. *Pattern Recognit.*,

48:659–669, 2015. doi: 10.1016/j.patcog.2014.07.028. URL https://doi.org/10.1016/j.patcog.2014.07.028.

Michael Reimer and Frank Rudzicz. Identifying articulatory goals from kinematic data using principal differential analysis. In *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

Christopher Rieser and Peter Filzmoser. Outlier detection for pandemic-related data using compositional functional data analysis. In *Pandemics: Insurance and Social Protection*, pages 251–266. Springer, Cham, 2022. doi: 10.1007/978-3-030-78334-1_12. URL https://doi.org/10.1007/978-3-030-78334-1_12.

V. K. Rohatgi and Gábor J. Székely. An inverse markov-chebychev inequality. *Periodica Polytechnica Civil Engineering*, 36(4):455–458, 1992. doi: N/A. URL https://pp.bme.hu/ci/article/view/3862.

Gaetano Romano, Idris Eckley, Paul Fearnhead, and Guillem Rigaill. Fast online changepoint detection via functional pruning cusum statistics, 2021. URL https://doi.org/10.48550/arXiv.2110.08205.

Gordon J. Ross. Sequential change detection in the presence of unknown parameters. *Statistics and Computing*, 24(6):1017–1030, November 2014. ISSN 0960-3174. doi: 10.1007/s11222-013-9417-1. URL https://doi.org/10.1007/s11222-013-9417-1.

Gordon J. Ross and Niall M. Adams. Two nonparametric control charts for detecting arbitrary distribution changes. *Journal of Quality Technology*, 44(2):102–116, 2012. doi: 10.1080/00224065.2012.11917887. URL https://doi.org/10.1080/00224065.2012.11917887.

Gordon J. Ross, Dimitris K. Tasoulis, and Niall M. Adams. Nonparametric monitoring of data streams for changes in location and scale. *Technometrics*, 53(4):379–389, 2011. doi: 10.1198/TECH.2011.10069. URL https://doi.org/10.1198/TECH.2011.10069.

Peter J Rousseeuw, Ida Ruts, and John W Tukey. The bagplot: a bivariate boxplot. *The American Statistician*, 53(4):382–387, 1999. doi: 10.1080/00031305.1999.10474494. URL https://doi.org/10.1080/00031305.1999.10474494.

Peter J. Rousseeuw, Jakob Raymaekers, and Mia Hubert. A measure of directional outlyingness with applications to image data and video. *Journal of Computational and Graphical Statistics*, 27(2):345–359, 2018. doi: 10.1080/10618600.2017.1366912. URL https://doi.org/10.1080/10618600.2017.1366912.

Farook Sattar and Frank Rudzicz. Principal differential analysis for detection of bilabial closure gestures from articulatory data. *Computer Speech & Language*, 36: 294–306, 2016. doi: 10.1016/j.csl.2015.07.002. URL https://doi.org/10.1016/j.csl.2015.07.002.

Pallavi Sawant, Nedret Billor, and Hyejin Shin. Functional outlier detection with robust functional principal component analysis. *Computational Statistics*, 27(1): 83–102, March 2012. doi: 10.1007/s00180-011-0239-3. URL https://doi.org/10.1007/s00180-011-0239-3.

Gideon Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6 (2):461 – 464, 1978. doi: 10.1214/aos/1176344136. URL https://doi.org/10.1214/aos/1176344136.

P.K. Sen and J.M. Singer. *Large Sample Methods in Statistics: An Introduction with Applications*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 1994. ISBN 9780412042218. URL https://books.google.co.uk/books?id=Q-8Tp201fGMC.

W. A. Shewhart. Economic quality control of manufactured product1. *Bell System Technical Journal*, 9(2):364–389, 1930. doi: https://doi.org/10.1002/j.1538-7305.1930.tb00373.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/j.1538-7305.1930.tb00373.x.

A. N. Shiryaev. On optimum methods in quickest detection problems. *Theory of Probability & Its Applications*, 8(1):22–46, 1963. doi: 10.1137/1108002. URL https://epubs.siam.org/doi/10.1137/1108002.

Lianjie Shu, Wei Jiang, and Kwok-Leung Tsui. A weighted cusum chart for detecting patterned mean shifts. *Journal of Quality Technology*, 40(2):194–213, 2008. doi: 10.1080/00224065.2008.11917725. URL https://www.tandfonline.com/doi/abs/10.1080/00224065.2008.11917725.

D. Siegmund and E. S. Venkatraman. Using the generalized likelihood ratio statistic for sequential detection of a change-point. *Ann. Statist.*, 23(1):255–271, 02 1995. doi: 10.1214/aos/1176324466. URL https://doi.org/10.1214/aos/1176324466.

David Siegmund. Change-points: From sequential detection to biology and back. *Sequential Analysis*, 32(1):2–14, 2013. doi: 10.1080/07474946.2013.751834. URL https//:doi.org/10.1080/07474946.2013.751834.

Junmo Song and Jiwon Kang. Sequential change point detection in arma-garch models. *Journal of Statistical Computation and Simulation*, 90(8):1520–1538, 2020. doi: 10.1080/00949655.2020.1734807. URL https://doi.org/10.1080/00949655.2020.1734807.

Joan G Staniswalis, Christopher Dodoo, and Anu Sharma. Local principal differential analysis: Graphical methods for functional data with covariates. *Communications in Statistics-Simulation and Computation*, 46(3):2346–2359, 2017. doi: 10.1080/03610918.2015.1043387. URL https://doi.org/10.1080/03610918.2015.1043387.

Ying Sun and Marc G. Genton. Functional boxplots. *Journal of Computational and Graphical Statistics*, 20(2):316–334, 2011. doi: 10.1198/jcgs.2011.09224. URL https://doi.org/10.1198/jcgs.2011.09224.

Priyanga Dilini Talagala, Rob J. Hyndman, Kate Smith-Miles, Sevvandi Kandanaarachchi, and Mario A. Muñoz. Anomaly detection in streaming nonstation-

ary temporal data. *Journal of Computational and Graphical Statistics*, 29(1):13–27, 2020. doi: 10.1080/10618600.2019.1617160. URL https://doi.org/10.1080/10618600.2019.1617160.

A. Tartakovsky. *Sequential Change Detection and Hypothesis Testing: General Non-i.i.d. Stochastic Models and Asymptotically Optimal Rules*. Chapman & Hall/CRC Monographs on Statistics and Applied Probability. CRC Press, 2019. ISBN 9781498757591. URL https://books.google.co.uk/books?id=WvrDDwAAQBAJ.

A. Tartakovsky, I. Nikiforov, and M. Basseville. *Sequential Analysis: Hypothesis Testing and Changepoint Detection*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press, 2014. ISBN 9781439838211. URL https://books.google.co.uk/books?id=th3OBQAAQBAJ.

A. G. Tartakovsky, A. S. Polunchenko, and G. Sokolov. Efficient computer network anomaly detection by changepoint detection methods. *IEEE Journal of Selected Topics in Signal Processing*, 7(1):4–11, 2013. doi: 10.1109/JSTSP.2012.2233713. URL https://ieeexplore.ieee.org/abstract/document/6380529.

Alexander G Tartakovsky, Boris L Rozovskii, and Khushboo Shah. A nonparametric multichart cusum test for rapid intrusion detection. In *Proceedings of Joint Statistical Meetings*, volume 7, page 11, 2005. doi: 10.1.1.348.5320. URL https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.348.5320.

Timo Teräsvirta and Ilkka Mellin. Model selection criteria and model selection tests in regression models. *Scandinavian Journal of Statistics*, pages 159–171, 1986. URL https://www.jstor.org/stable/4616022.

S. O. Tickle, I. A. Eckley, and P. Fearnhead. A computationally efficient, high-dimensional multiple changepoint procedure with application to global terrorism incidence. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184 (4):1303–1325, 2021. doi: 10.1111/rssa.12695. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssa.12695.

Karim Tout, Florent Retraint, and Rémi Cogranne. Non-stationary process monitoring for change-point detection with known accuracy: Application to wheels coating inspection. *IEEE Access*, 6:6709–6721, 2018. doi: 10.1109/ACCESS.2018.2792838. URL https://ieeexplore.ieee.org/document/8255557.

L.N. Trefethen and D. Bau. *Numerical Linear Algebra*. Other Titles in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104), 1997. ISBN 9780898719574. URL https://books.google.co.uk/books?id=JaPtxOytY7kC.

Charles Truong, Laurent Oudre, and Nicolas Vayatis. Selective review of offline change point detection methods. *Signal Processing*, 167:107299, 2020. ISSN 0165-1684. doi: 10.1016/j.sigpro.2019.107299. URL https://www.sciencedirect.com/science/article/pii/S0165168419303494.

J. Derek Tucker, Wei Wu, and Anuj Srivastava. Generative models for functional data using phase and amplitude separation. *Computational Statistics & Data Analysis*, 61:50–66, 2013. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2012.12.001. URL https://www.sciencedirect.com/science/article/pii/S0167947312004227.

J. W. Tukey. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians, Vancouver, 1975*, pages 523–531, 1975. URL https://cir.nii.ac.jp/crid/1573950399770196096.

R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, page 37. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. ISBN 9781108415194. URL https://books.google.co.uk/books?id=NDdqDwAAQBAJ.

Juan Vilar, P. Raña, and Germán Aneiros-Pérez. Using robust fpca to identify outliers in functional time series, with applications to the electricity market. *SORT*, 40:321–348, 01 2016. URL https://raco.cat/index.php/SORT/article/view/316148.

Guillermo Vinue and Irene Epifanio. Robust archetypoids for anomaly detection in big functional data. *Advances in Data Analysis and Classification*, 08 2020. doi: 10.1007/s11634-020-00412-9. URL https://doi.org/10.1007/s11634-020-00412-9.

M.J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. ISBN 9781108498029. URL https://books.google.co.uk/books?id=IluHDwAAQBAJ.

Shaohua Wan, Songtao Ding, and Chen Chen. Edge computing enabled video segmentation for real-time traffic monitoring in internet of vehicles. *Pattern Recognition*, 121:108146, 2022. doi: 10.1016/j.patcog.2021.108146. URL https://doi.org/10.1016/j.patcog.2021.108146.

Dabuxilatu Wang, Liang Zhang, and Qiang Xiong. A non parametric cusum control chart based on the mann–whitney statistic. *Communications in Statistics - Theory and Methods*, 46(10):4713–4725, 2017. doi: 10.1080/03610926.2015.1073314. URL https://doi.org/10.1080/03610926.2015.1073314.

Huangang Wang and Ma Yao. Fault detection of batch processes based on multivariate functional kernel principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 149:78–89, 2015. doi: 10.1016/j.chemolab.2015.09.018. URL https://doi.org/10.1016/j.chemolab.2015.09.018.

Jane-Ling Wang, Jeng-Min Chiou, and Hans-Georg Müller. Functional data analysis. *Annual Review of Statistics and Its Application*, 3(1):257–295, 2016. doi: 10.1146/annurev-statistics-041715-033624. URL https://doi.org/10.1146/annurev-statistics-041715-033624.

Shanshan Wang, Wolfgang Jank, Galit Shmueli, and Paul Smith. Modeling price dynamics in ebay auctions using differential equations. *Journal of the American Statistical Association*, 103(483):1100–1118, 2008. doi: 10.1198/016214508000000670. URL https://doi.org/10.1198/016214508000000670.

Xing Wang, Jessica Lin, Nital Patel, and Martin Braun. Exact variable-length anomaly detection algorithm for univariate and multivariate time series. *Data Mining and Knowledge Discovery*, 32:1806–1844, 07 2018. doi: 10.1007/s10618-018-0569-7. URL https://doi.org/10.1007/s10618-018-0569-7.

L. Wasserman. *All of Nonparametric Statistics*. Springer Texts in Statistics. Springer New York, 2006. ISBN 9780387306230. URL https://books.google.co.uk/books?id=MRFlzQfRg7UC.

S. S. Wilks. The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *The Annals of Mathematical Statistics*, 9(1):60 – 62, 1938. doi: 10.1214/aoms/1177732360. URL https://doi.org/10.1214/aoms/1177732360.

Hans Peter Wolf. *aplpack: Another Plot Package (version 190512)*, 2019. URL https://cran.r-project.org/package=aplpack.

Yanhong Wu. *Sequential Common Change Detection and Isolation of Changed Panels in Panel Data*, pages 391–409. Springer Singapore, Singapore, 2019. ISBN 978-981-15-0864-6. doi: 10.1007/978-981-15-0864-6_20. URL https://doi.org/10.1007/978-981-15-0864-6_20.

Dongdong Xiang, Shulin Gao, Wendong Li, Xiaolong Pu, and Wen Dou. A new nonparametric monitoring of data streams for changes in location and scale via cucconi statistic. *Journal of Nonparametric Statistics*, 31(3):743–760, 2019. doi: 10.1080/10485252.2019.1632307. URL https://doi.org/10.1080/10485252.2019.1632307.

Liyan Xie, George V. Moustakides, and Yao Xie. Window-limited cusum for sequential change detection, 2022. URL https://arxiv.org/abs/2206.06777.

Hao Yan, Kamran Paynabar, and Jianjun Shi. Real-time monitoring of high-dimensional functional data streams via spatio-temporal smooth sparse decomposition. *Technometrics*, 60(2):181–197, 2018. doi: 10.1080/00401706.2017.1346522. URL https://doi.org/10.1080/00401706.2017.1346522.

Takuma Yoshida. Direct determination of smoothing parameter for penalized spline regression. *Journal of Probability and Statistics*, 2014:1–11, 04 2014. doi: 10.1155/2014/203469. URL https://doi.org/10.1155/2014/203469.

Yi Yu, Oscar Hernan Madrid Padilla, Daren Wang, and Alessandro Rinaldo. A note on online change point detection, 2020. URL https://doi.org/10.48550/arXiv.2006.03283.

Jin-Ting Zhang and Jianwei Chen. Statistical inferences for functional data. *The Annals of Statistics*, 35(3):1052–1079, 2007. ISSN 00905364. doi: 10.1214/009053606000001505. URL http://www.jstor.org/stable/25463592.

Ruizhi Zhang, Yajun Mei, and Jianjun Shi. *Wavelet-Based Profile Monitoring Using Order-Thresholding Recursive CUSUM Schemes*, pages 141–159. Springer, 01 2018. ISBN 978-3-319-99388-1. doi: 10.1007/978-3-319-99389-8_6. URL https://doi.org/10.1007/978-3-319-99389-8_.

Maoyuan Zhou, Wei Geng, and Zhaojun Wang. Likelihood ratio-based distribution-free sequential change-point detection. *Journal of Statistical Computation and Simulation*, 84(12):2748–2758, 2014. doi: 10.1080/00949655.2014.899599. URL https://doi.org/10.1080/00949655.2014.899599.

Changliang Zou, Guosheng Yin, Long Feng, Zhaojun Wang, et al. Nonparametric maximum likelihood approach to multiple change-point problems. *The Annals of Statistics*, 42(3):970–1002, 2014. doi: 10.1214/14-AOS1210. URL https://doi.org/10.1214/14-AOS1210.