Change Point Analysis in Piecewise Polynomial Signals Using Trend Filtering

by

Reza Valiollahi Mehrizi

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 2021

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Richard Lockhart Professor, Dept. of Statistics and Actuarial Science, Simon Fraser University
Supervisor:	Shojaeddin Chenouri Professor, Dept. of Statistics and Actuarial Science, University of Waterloo
Internal Members:	Paul Marriott Professor, Dept. of Statistics and Actuarial Science, University of Waterloo
	Greg Rice Associate Professor, Dept. of Statistics and Actuarial Science, University of Waterloo
Internal-External Member:	Pascal Poupart Professor, Cheriton School of Computer Science, University of Waterloo

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Change point analysis, an active area of research in many fields, including statistics, has attracted a lot of attention in recent years. The focus of this thesis is change point detection, where the purpose is to estimate the number and locations of changes in the structure of a data sequence. Despite the recent attention, few papers have addressed change point analysis for piecewise polynomial signals. To address this gap, the work focuses on the mean change point problem for such signals. We approach this problem by applying trend filtering and introduce a method called *Pattern Recovery Using Trend Filtering (PRUTF)* to estimate change point locations.

We develop an extension of the trend filtering algorithm in order to construct a dual solution path that allows us to detect change points. We demonstrate that the dual solution path constitutes a Gaussian bridge process, which enables us to derive an exact and efficient stopping rule to terminate the search algorithm. Finally, we prove that the estimates produced by this algorithm are asymptotically consistent for piecewise polynomial signals. This result holds even in the presence of staircase patterns (consecutive change points of the same sign) in the signal, which to the best of our knowledge, previous works have been unable to address.

Additionally, we employ the post-selection framework to make statistical inference for change points once selected by the PRUTF method. The key development has been to represent the set of estimated change points and their signs as a polyhedron in the sample space. This representation gives us tools for exact statistical inference such as the construction of confidence intervals and testing the significance of the estimated change points. We also provide some truncation techniques in order to improve the confidence intervals and enhance the power of the tests.

Acknowledgements

I would like to thank many people who have helped me through the completion of this dissertation. First, I would like to express my sincere gratitude to my supervisor Professor Shojaeddin Chenouri, with whom I have had the luck to collaborate. I also wish to express my appreciation to Professor Paul Marriott and Professor Greg Rice as members of my supervisory committee. A lot of thanks are due to them for inspiring useful changes to my thesis. My appreciation also goes to all the members at the Department of Statistics and Actuarial Science, at the University of Waterloo, for their help and assistance during my PhD studies.

Special thanks to my parents, my brother and sisters for their constant emotional support. Finally, my thanks also go to those who, directly or indirectly, helped me to complete my dissertation.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Scientific research has focused on understanding the behaviour of processes for over a century. This understanding allows researchers to draw conclusions that impact the future behaviour of processes. The desire to discover the unobserved structure of evolutionary time series has inspired systematic methods for recording measurements in various fields such as economics, medical science and environmental studies. These systematic recording methods, along with technological advances in data storage and management, have enabled researchers to better understand the structure of underlying processes. The immense growth of digital storage and data processing capacities has recently moved the analysis of time series data to a new stage. The quantity and quality of data in various fields such as climate studies, speech analysis, finance, physiology, internet traffic and astronomy have increased. This improvement in data recording provides researchers with opportunities to develop new tools for analyzing, understanding and interpreting this data.

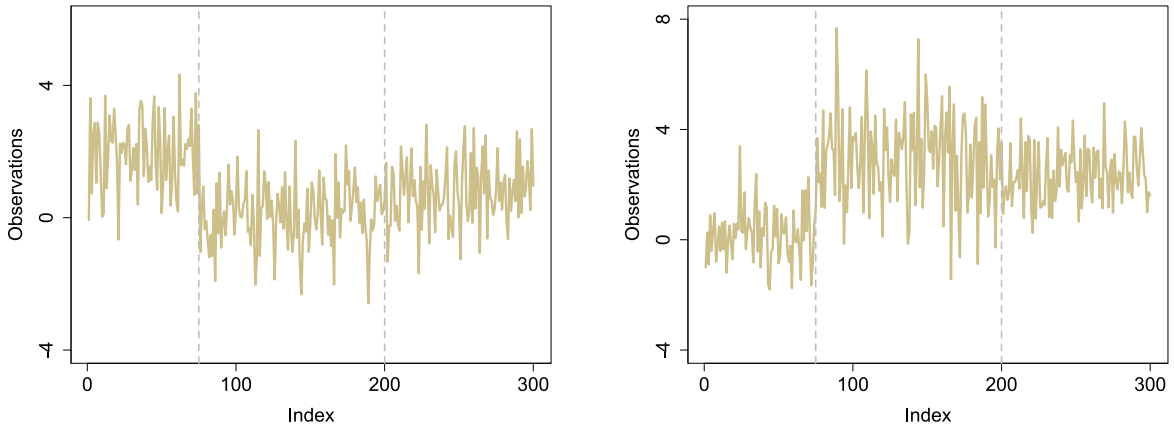
Evolutionary processes undergo abrupt transitions, allowing researchers to identify the point of change and its causes and consequences. For example, policymakers and economists divide the state of an economy into recession and recovery classes and seek to determine time points when a transition between these two states occurs. By identifying such transition points, economists can decide on appropriate policy measures accordingly [43]. In medicine, doctors monitor physiological measurements of patients over time to determine sudden abnormalities that may indicate a medical condition [20]. Similarly, in seismology, scientists deal with predicting earthquakes by tracking velocity changes in seismic waves [113].

An interesting and informative research field, referred to as change-point detection, is to estimate the number and location of points where a transition occurs. A change-

point refers to a location at which the observations before and after it follow two different models. Change-point analyses are extensively applied in medicine, health sciences, finance, geographics, environment, etc. In statistics, detecting all change points is crucial because the homogeneity of a statistical model is affected any time a change point exists. If there is a change point, performing a statistical analysis without considering its existence would usually lead to misleading and possibly invalid results.

Here, we present an introductory example for the statistical concept of a change point. This concept is visualized in Figure 1.1. Statistical properties of both datasets in Figure 1.1 change at locations 75 and 200. In panel (a), the mean of the observations changes at both locations. As can be seen, there is a noticeable mean shift at the location 75, whereas the mean shift at the location 200 may not be easily detectable. The observations in panel (b) undergo changes in both mean and variance. For example, the variation of the observations in the second segment is more considerable than that of the other two segments.

The main purpose of this thesis is to introduce a novel statistical approach for change point detection. Specifically, we employ trend filtering to identify change points in mean models with piecewise polynomial structure. This approach called *Pattern Recovery Using Trend Filtering* (PRUTF) can be used in a range of different applications, thereby allowing for more insightful analysis and meaningful interpretation of data. We also address the gap between detecting change points and performing inference for change points after detection.



(a) Changes in the mean.

(b) Changes in both mean and variance.

Figure 1.1: Simulated datasets with change points at locations 75 and 200.

In particular, we study conducting statistical inference on the significance of change points after detection in the framework of post-selection inference.

Prior to a review of existing literature, a few real-world applications are provided to emphasize the critical contributions that change point analysis makes to many fields. The following list of applications also expands the scope of data structures to time series, images and panel datasets.

1.1 Motivational Examples

The following paragraphs describe examples selected from neuroscience, genetics and finance. It is worth pointing out that these are fields in which change point analysis is frequently applied.

The first example is selected from the field of neuroscience. Epileptic seizures and episodes of abnormal brain activity have received much attention from researchers in efforts to better understand disruptions in normal brain functioning. Such analysis has helped neuroscientists develop more accurate diagnoses, improved therapy, and effective early warning systems for seizure activity. During a seizure, sudden and abrupt changes occur in the spectral profile of brain activity. To identify such changes, electroencephalographic (EEG) recordings have been studied for over a few decades. The EEG datasets are frequently modelled as piecewise stationary processes with potential change points [89]. As a direct implication, a change point detection approach allows neuroscientists to determine such change points.

The second example has been selected from genetics. A copy number variant (CNV) in DNA sequences is a type of structural variation that causes a genome to possess an abnormal number of copies of DNA. These variants have been established as sources of variation within a population. For a given cell or individual, CNV observations are often measured in the form of "log-R ratios" for a range of probes with different locations along the genome. These observations are measured as log base 2 of the ratio of the observed probe intensity to the reference intensity for a given probe. The mean of log-R ratios for normal regions of the genome is 0, whereas the mean of log-R ratios for CNVs is far from zero [81, 17]. Therefore, to locate CNVs, the goal would be to identify the regions of the genome whose means of log-R ratios are not zero.

Finally, the third example is selected from the field of finance. Corporate finance datasets contain the annual value of a range of different financial indicators, such as the amount of dividend paid out by firms or the values of the firms' assets. Corporate cash

holding, a phenomenon in corporate finance whereby firms hold considerable cash, has grown in recent years. For example, the cash holding by United States firms has doubled in the past decades. Now, policymakers and scientists are starting to pay a lot of attention to this corporate cash holding. Scientists believe that the evolution of the phenomenon can be explained by changes in the "cost of carry", that is, the net cost of financing one dollar of liquid assets, [11, 63]. Therefore, an important subject for a scientist would be to detect change points in cost of carry data.

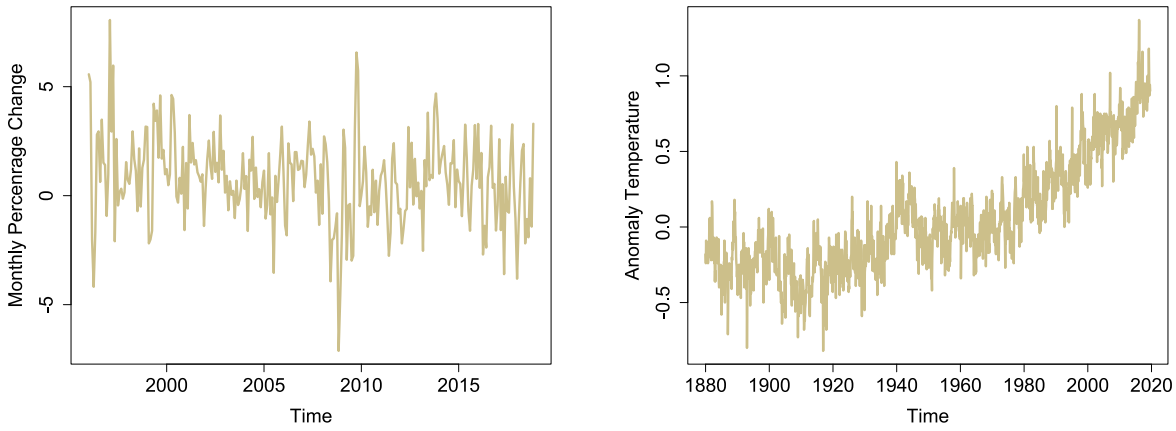
1.2 Real Datasets For Analysis

In this thesis, to explain how to apply our proposed method in practice, we will provide three real examples including the economic data (House Price Index), climate data (GISS Surface Temperature), as well as health science data (Covid-19). These examples are interesting in the context of change point analysis, as we will illustrate in Chapter 3.

Example 1.1 *The UK House Price Index (HPI) is a National Statistic that shows changes in the value of residential properties in England, Scotland, Wales and Northern Ireland. A house price index (HPI) measures the price changes of residential housing. The HPI measures the price change of completed house sale transactions as a percentage change from some specific start date. The UK index uses the hedonic regression model as a statistical approach to produce estimates of the change in house prices for each period. For more details, see* https://landregistry.data.gov.uk/app/ukhpi .

Many researchers including [4], [54], and [57] studied the UK HPI dataset for change point analysis. We consider monthly percentage change in the UK HPI at Tower Hamlets (an east borough of London) from January 1996 to November 2018. The dataset is presented in Figure 1.2.

Example 1.2 *Goddard Institute for Space Studies (GISS) monitors broad global changes around the world. The GISS Surface Temperature Analysis (GISTEMP) is an estimate of the global surface temperature change (see* https://www.giss.nasa.gov *). In the analysis of GISTEMP data, the temperature anomalies are used rather than the actual temperature. A temperature anomaly is a difference from an average or baseline temperature. The baseline temperature is typically computed by averaging 30 or more years of temperature data (1951 to 1980 in the current dataset). A positive anomaly indicates the*



(a) UK HPI dataset.

(b) GISTEMP dataset.

Figure 1.2: Plots of both Tower Hamlet HPI and GISTEMP datasets provided in examples 1.1 and 1.2.

observed temperature was warmer than the baseline. In contrast, a negative anomaly indicates the observed temperature was cooler than the baseline (for more details, see [67] and [98]).

The GISTEMP dataset has been frequently explored in change point literature, for example see [134], [80] and [16]. Figure 1.2 displays monthly land-ocean temperature anomalies recorded from January 1880 to August 2019 (see `https://data.giss.nasa.gov/gistemp` *). The plot reveals the presence of a linear trend with several potential change points in the dataset.*

Example 1.3 *Since the initial outbreak of Novel Coronavirus Disease 2019 (COVID-19) in Wuhan, China, in mid-November 2019, the virus has rapidly spread throughout the world. The pandemic has infected 167.25 million people and killed more than 3.46 millions, until April 30, 2021 (*`https://covid19.who.int/`*), greatly stressing public health systems and adversely influencing global society and economies. Therefore, every country has attempted to slow down the transmission rate by various regional and national policies such as the declaration of national emergencies, quarantines and mass testing. Of vital interest to governments is understanding the pattern of the epidemic growth and assessing the effectiveness of policies undertaken. A scientist can investigate these matters by analyzing the sequence of infection data for COVID-19. Change-point detection is one possible*

framework for studying the behaviour of COVID-19 infection curves. By detecting the locations of alterations in the curves, change point analysis gives us insights into changes in the transmission rate or efficiency of interventions. It also enables us to raise warning signals if the disease pattern changes.

For the analysis, we will consider the log-scale of the cumulative number of confirmed cases from March 10, 2020 to April 30, 2021, for Australia, Canada, the United Kingdom and the United States. Figure 1.3 displays these datasets as well as the number of confirmed

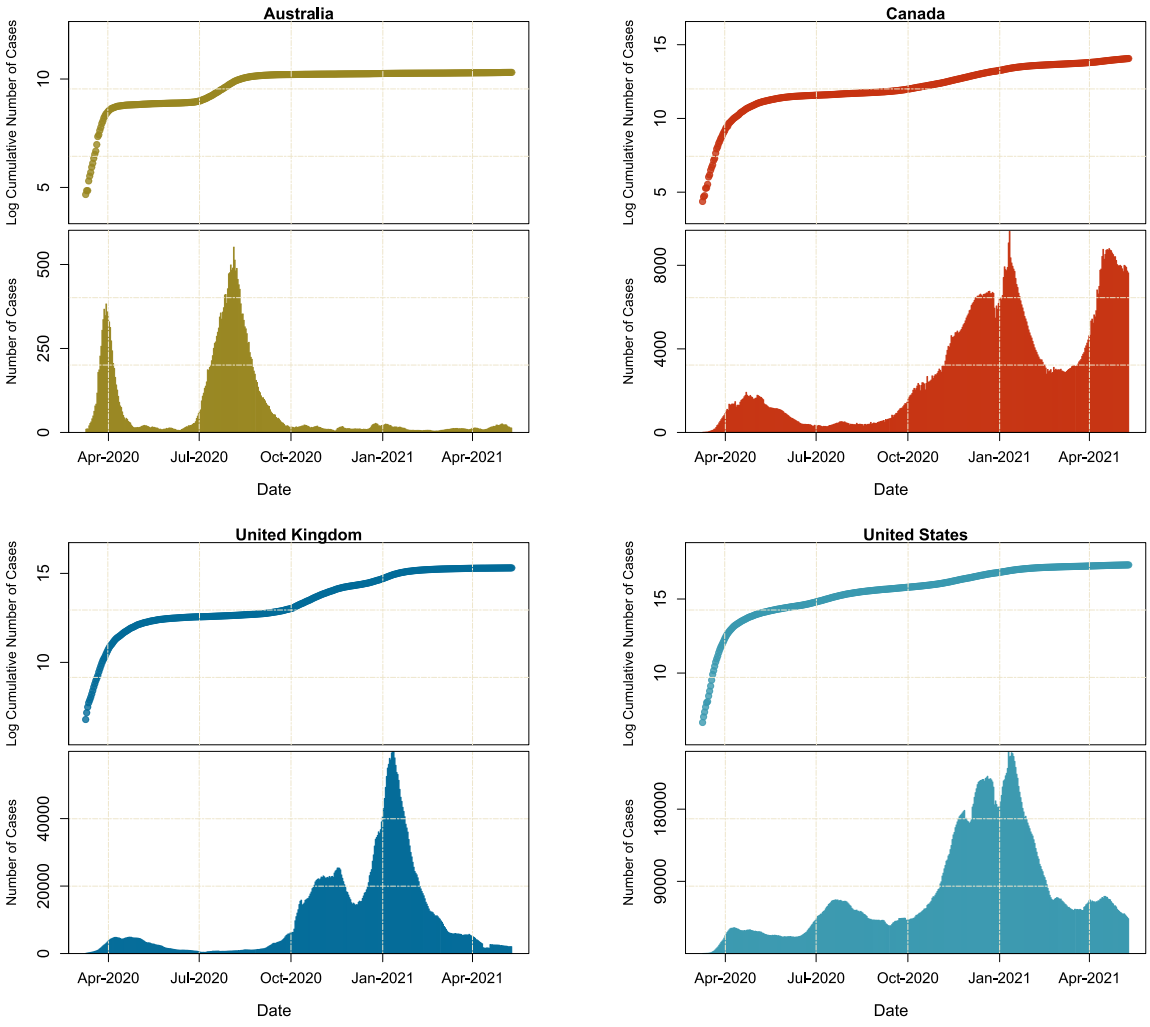


Figure 1.3: Number of COVID-19 cases as well as the log cumulative number of cases for Australia, Canada, United Kingdom and the United States.

cases for Australia, Canada, the United Kingdom and the United States.

1.3 Thesis Structure

The main objective of this thesis is to introduce a new approach for detecting the potential change points in a univariate data sequence. The goal is to use trend filtering to develop an algorithm that identifies change points in a mean model with a piecewise polynomial structure. Simulation studies are conducted to provide useful insights into the proposed approach and illustrate its performance. The proposed approach is also compared with several existing state-of-the-art change point approaches, focusing on accuracy and efficiency. Finally, we employ our proposed approach to analyze the real datasets in Examples 1.1 – 1.3.

Chapter 2 begins with a review of existing literature on change point detection setting, introducing the reader to the relevant basic foundations and concepts that appear in later chapters. Various problem formulations and detection approaches are also discussed, focusing on the state-of-the-art methods that we use as benchmark approaches later in this thesis.

In Chapter 3, we present our novel change point detection approach, referred to as *Pattern Recovery Using Trend Filtering* (PRUTF). The approach aims to split a data sequence into segments where the properties of the data change from one segment to another. We assume that the true mean model follows a piecewise polynomial model. This piecewise polynomial mean model is a more-general problem of estimating the number and locations of change points and has not received much attention in the literature. Additionally, the consistency and the rate of convergence for the detected change points are established. To illustrate the efficiency of PRUTF, simulation studies, as well as real data analyses, are provided.

In Chapter 4, we discuss a question related to that of Chapter 3: how significant are the change points estimated using PRUTF ? In other words, our goal is to perform statistical inference for change points that have been detected using PRUTF. A newly developed framework called *post-selection inference* has made it possible to develop inferential tools for change points after they have been estimated using a detection procedure. This inference includes computing p-values for the significance of the change points in magnitude and constructing confidence intervals. Specifically, we propose two approaches that lead to inferences with high-power tests and narrow confidence intervals.

In Chapter 5, we conclude this thesis and discuss several interesting avenues for future research.

Chapter 2

Literature Review

Change point detection has gained considerable attention in recent years and found applications in many fields such as finance and econometrics, [15, 66], bioinformatics and genomics, [59, 90], climatology, [100, 121], and technology, [136, 116, 104, 125, 60]. In statistical literature, change point analysis dates back to the 1950s [118, 117]. This type of analysis has since been a very active field of research due to its broad application, such as in time series analysis and signal processing, in which data is routinely collected over time. Since 2010, several factors, including the demand for computationally fast and statistically efficient segmentation approaches, have renewed statistical researchers' interest in the change point problem. Consequently, there is vast and rich literature on this subject.

The problem of change point detection arises when an ordered sequence of data undergoes abrupt and meaningful changes. These changes can occur in the mean, variance, slope of a trend, or distribution of the underlying data sequence. In general, change point detection approaches can be classified as offline (retrospective) or online (sequential). The offline change point framework includes approaches in which the full dataset for analysis is available. In contrast, in the online change point framework [18, 109, 87, 41, 141], the analysis is carried out sequentially as more data become available. More-recent papers address both offline and online change point detection problems in more-complicated situations, see [8] and [74]. The focus of the research described in this thesis is the offline change point framework.

Consider a data sequence of length n, denoted by $\mathbf{y} = (y_1, , \ldots, y_n)$, which is ordered by time, location, or some other attributes. The goal is to find the locations of sudden transition in the data's structure. We assume that the true number of change points is J_0, with exact locations given by the entries of the vector $\boldsymbol{\tau} = \{\tau_1, \ldots, \tau_{J_0}\}$. The 0-th and

(J_0+1)-th change points are conventionally defined as $\tau_0 = 0$ and $\tau_{J_0+1} = n$. Statistically, it is assumed that each data point has a cumulative distribution function F; that is, $y_i \sim F_i,\ i = 1, \ldots, n$. For the canonical change point problem, a change then occurs in the location τ if $F_\tau \neq F_{\tau+1}$. Formally, we have the following definition for change points.

Definition 2.1 *For an observation vector* $\mathbf{y} = (y_1, \ldots, y_n)$ *with a piecewise constant structure and the assumption* $Y_i \sim F_i$*, the locations* $\boldsymbol{\tau} = \{\tau_1, \ldots, \tau_{J_0}\} \subset \{1, \ldots, n\}$ *are called change points if*

$$F_1 = \ldots = F_{\tau_1} \neq F_{\tau_1+1} = \ldots = F_{\tau_2} \neq F_{\tau_2+1} = \ldots = F_{\tau_{J_0}} \neq F_{\tau_{J_0}+1} = \ldots = F_n\,, \tag{2.1}$$

where J_0 *is the number of change points.*

It is worth emphasizing that, in applications, the number of change points J_0 and their locations $\boldsymbol{\tau} = \{\tau_1, \ldots, \tau_{J_0}\}$ are usually unknown. The aim in a change point problem is to consistently estimate J_0 and $\boldsymbol{\tau} = \{\tau_1, \ldots, \tau_{J_0}\}$ from an observed dataset. We denote the estimates of J_0 and $\boldsymbol{\tau} = \{\tau_1, \ldots, \tau_{J_0}\}$ by $\widehat{J}_0$ and $\widehat{\boldsymbol{\tau}} = \{\widehat{\tau}_1, \ldots, \widehat{\tau}_{\widehat{J}_0}\}$, respectively.

A very common type of change point analysis is the one that looks for transitions in the mean of the data. In the piecewise constant mean shift change, the goal is to find locations $\boldsymbol{\tau} = \{\tau_1, \ldots, \tau_{J_0}\}$ such that

$$\mu_1 = \ldots = \mu_{\tau_1} \neq \mu_{\tau_1+1} = \ldots = \mu_{\tau_2} \neq \mu_{\tau_2+1} = \ldots = \mu_{\tau_{J_0}} \neq \mu_{\tau_{J_0}+1} = \ldots = \mu_n, \tag{2.2}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$ is the true mean of the underlying dataset. This piecewise constant mean change model is a fundamental problem in change point analysis, and most of the existing approaches, including the one presented in this thesis, have been developed specifically for this type of change. Note that in a mean shift problem with an r degree piecewise polynomial structure, at the location of a change point at least one of the coefficients in the polynomial function varies.

Section 2.1 briefly reviews some works of change point analysis, emphasizing its growing application in various fields. Since there are many approaches in change point analysis, we classify them into two broad categories and then review some state-of-the-art and popular ones. Our primary focus is on the penalized likelihood models with ℓ_1-norm for discovering change points. Section 2.2 is devoted to a general overview of research in this field. Particularly, trend filtering, which has recently attracted huge attention, is reviewed and its properties are outlined. A new framework called post-selection inference has been recently developed in order to make inferences for derived models through a selection procedure. In Section 2.3, we summarize some proposed techniques for post-selection inference, which will be later applied to test the significance of change point locations in Chapter 4.

2.1 Change Point Methodologies

This section reviews a body of literature regarding the offline change point framework closely related to our proposal. We refer the reader to [47, 29, 155] and [34] for more comprehensive reviews. Depending on various aspects of data, many different techniques and algorithms exist in the literature. Some techniques, such as likelihood-ratio-based methods, are very efficient in a single change point detection framework. The penalized likelihood is another approach that performs well when the number of change points is fixed and bounded. Search-based techniques like Binary Segmentation and its variants can identify multiple change points; however, they are computationally intensive, particularly in large datasets.

We begin with a straightforward problem of change point detection in which we assume that there is at most one change point in the underlying dataset. In this setting, we often attempt to test the hypothesis of no change versus one change for the entire data sequence:

$$H_0 : \ F_1 = \ldots = F_n \qquad \text{v.s.} \qquad H_1(\tau) : \ F_1 = \ldots = F_\tau \neq F_{\tau+1} = \ldots = F_n \,. \tag{2.3}$$

As a popular statistical approach for hypothesis testing, the likelihood ratio test can be applied to examine (2.3) for any given value of $\tau = 1, 2, \ldots, n-1$. If H_0 is rejected, the change point estimate is the maximizer of the likelihood ratio. In order to state the method in statistical terminology, let R_τ denote the likelihood ratio of the testing problem in (2.3) for any given τ. The test statistic then becomes, [71]

$$T = \max_{1 \leq \tau \leq n-1} \ -2 \, \log \left(R_\tau \right), \tag{2.4}$$

which rejects H_0 if its value exceeds a certain threshold. In such a case, the change point estimate is given by

$$\widehat{\tau} = \arg\max_{1 \leq \tau \leq n-1} \ -2 \, \log \left(R_\tau \right). \tag{2.5}$$

The mean change point problem (2.2) when observations follow a Gaussian distribution with a possible change in the mean and a fixed variance σ^2 was studied in [38].

Another popular approach, though closely related to likelihood ratio method, used to test the hypothesis in (2.2) is *cumulative sum* (CUSUM), which is constructed based on partial sums of data points. The CUSUM for a change in the mean of the data sequence $y_1, \ldots, y_n$ is defined by

$$U_n(s) = \frac{1}{\sigma \sqrt{n}} \sum_{i=1}^{\lfloor ns \rfloor} \left(y_i \, - \, \bar{y}_{1:n} \right), \qquad\qquad 0 \, \leq \, s \, \leq \, 1, \tag{2.6}$$

where $y_1, \ldots, y_n$ are independently and identically distributed with mean μ and variance σ^2, and $\bar{y}_{1:n}$ is their average value. If a change occurs at the location $\lfloor ns \rfloor$, then the value of $U_n(s)$ becomes large. On the other hand, if there is no change, data points on each side of the mean will cancel each other out and the value of $U_n(s)$ remains small. Under the null hypothesis of no change, $U_n(s)$ is shown to converge to a Brownian bridge [24].

In practice, though, a data sequence frequently undergoes multiple changes and needs special search algorithms for their change discovery. The multiple change point detection literature includes many techniques with their mechanism specialized for specific applications. Depending on whether a method searches for change points all at once or one at a time, the existing methods in the change point literature fall into two broad categories. The first category contains methods that aim to detect change points all at once, mainly by solving an optimization problem consisting of a loss function (often a negative log-likelihood function) coupled with a penalty function. The other category is closely related to the change point testing framework, in which the single change point test is locally conducted to estimate multiple change points. In the following, we review the literature on the univariate mean change point analysis according to both penalized likelihood and testing categories.

2.1.1 Penalized Likelihood Approaches

Under a more general distributional assumption on the observations $\mathbf{y}$, one may write the penalized likelihood function as

$$-\ell(\mathbf{y}, \mathbf{X}\boldsymbol{\beta}) + P_\lambda(\boldsymbol{\beta}), \tag{2.7}$$

where $\ell(\cdot)$ is the log-likelihood function and $P_\lambda(\cdot)$ is a penalty function with the regularization parameter $\lambda > 0$. The problem (2.7) has been extensively studied in the literature with various choices of penalties and design matrices $\mathbf{X}$.

Many researchers have studied the problem (2.7) with the penalty function linear in the number of change points (ℓ_0-norm). Assuming a piecewise constant signal, [168, 169] applied the Schwarz Information Criterion (SIC) as a penalty function with the square error loss to consistently estimate the bounded number of change points and their locations for Gaussian observations. Since then, under the same framework as [169], some works have focused on various forms of penalty function to improve the properties of change point estimations. For instance, [91] proposed a method using a slightly different version of SIC, which is still linear in the number of change points. A penalty function, called Modified Information Criterion (MIC), was proposed by [119], who also showed that its

estimates are consistent in terms of both the number and the location of change points. Several other penalties such as the modified Bayesian Information Criterion (mBIC), [171], Simultaneous Information Theoretic Criterion (SITC), [167] and modified SIC [35, 36] have been proposed, and their change point estimates have been proved to be consistent.

In a recent paper, [160] studied the change point problem for a piecewise constant mean model using an ℓ_0-penalized least square method. [160] proved that the method is nearly minimax optimal in terms of the localization error. [170] have since generalized that method to higher degree polynomials. Specifically, they considered change point localization in a piecewise polynomial mean model. Their paper proposes a two-step change point detection method using an ℓ_0-penalized least square problem. In the first step, the method identifies change points by applying the minimal partitioning algorithm. With the estimated change points and their associated segments provided in the first step, the method updates the location of change points by minimizing a squared error loss function over each segment. The authors have established that the method is nearly minimax optimal. Furthermore, [166] studied the problem of jump detection in piecewise smooth trends under complex temporal dynamics whose covariance and higher-order structures evolve both smoothly and abruptly over time. Their method, called multiscale jump point detection (MJPD), is based on a multiscale statistic that applies an optimal wavelet/filter to the underlying time series. MJPD is a two-step procedure in which the first step applies the multiscale statistic to identify the number and locations of jumps, and the second step updates these estimated jump locations using a CUSUM statistic. The method has been shown to efficiently and accurately identify all jump locations.

All of the aforementioned methods involve computationally intensive search algorithms, which become infeasible as the size of data increases. Consequently, many researchers have invented approaches that enable the fast and efficient running of these algorithms. Segment Neighborhood [79] and Optimal Partitioning [9] are examples of dynamic programming algorithms for solving optimization problems. Some pruning algorithms such as PELT, [85, 70], pDPA [127], FPOP, and SNIP [105, 50] have also developed, and allow for more efficient implementation of the Segment Neighborhood and Optimal Partitioning algorithms.

In the regression context, when an ℓ_1-norm is used as the penalty function, the problem (2.7) is known as lasso [147], and has been greatly studied over the past twenty-five years. In the context of change point analysis, the application of ℓ_1-penalized likelihood approaches has attracted a lot of attention recently, mainly due to the wealth of works on ℓ_1-regularized regression. The formulation of the change point problem as a penalized regression was considered in [76]. Dealing with a DNA copy number dataset, they used the optimization

problem

$$\min_{\boldsymbol{\mu}\in\mathbb{R}^n} \frac{1}{2}\left\|\mathbf{y}-\boldsymbol{\mu}\right\|_2^2 + \lambda \sum_{i=1}^{n-1} \left|\mu_{i+1}-\mu_i\right|, \tag{2.8}$$

to fit a piecewise constant model for the DNA copy number mean $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$. The problem (2.8) was originally proposed in the signal processing framework by [133] and was called *total variation denoising*. It was later reproposed as fused lasso in [148]. Motivated by the work of [76], [68] applied fused lasso to estimate the locations of change points. Specifically, the consistency of its estimates for change point locations was proved when the number of change points is known.

[129] proposed sparse fused lasso, composed of both ℓ_1-norm and total variation seminorm, which yields a sparse piecewise constant fit. [129] established the consistency of the fused lasso estimates when the variance of noise terms vanishes and the minimum magnitude of jumps is bounded from below. However, [132] argued that the consistency results achieved by [129] are incorrect when a frequently viewed pattern, called staircase, exists in the signal. The staircase phenomenon occurs in a piecewise-constant model when there are either two successive jumps down or jumps up in its mean structure. This concept will be discussed in more detail later in Chapter 3. In addition, the lasso problem derived by transforming the underlying fused lasso does not satisfy the so-called Irrepresentable Condition ([173]), which is necessary and sufficient for exact pattern recovery. [132] introduced an alternative property called approximate sign consistency, which is more practical, and showed that it is satisfied by the fused lasso estimates. The fused lasso estimates for change point detection were also discussed in [123]. In particular, its inconsistency is illustrated by the irrepresentable condition as in [132], and a new approach based on puffer transformation of [83] is then proposed. They named the method Preconditioned Fused Lasso and established that it can recover the exact pattern with a probability approaching one. The other work that addressed the impact of staircase patterns on fused lasso estimates is [138]. They established that, in the presence of a staircase pattern, the objective function of fused lasso fails to improve when adjacent blocks are merged. This fact arises because as the regularization parameter decreases, the bias remains zero and prevents staircase blocks from adjoining. Applying this result, a modified version of the algorithm, presented in [72], leads to consistent pattern recovery. Because of the pivotal importance of fused lasso and its generalization for the research conducted in this manuscript, we will review and discuss the topic in more detail in Section 2.2.

Several authors have studied the problem of change point detection for the canonical parameters in the exponential family. The consistency of estimates derived from minimization of the negative log-likelihood along with SIC as the penalty function has been studied

by [69] and [92]. A more popular work of change point detection for the exponential family has been provided by [53]. For change point detection in the canonical parameter of a one-dimensional exponential family of the form $f_\theta(x) = \exp\{\theta x - \psi(\theta)\}$, [53] proposed the Simultaneous Multiscale Change-point Estimator (SMUCE). This technique estimates the number of change points and their locations by minimizing the number of change points subject to a constraint on a multiscale test statistic. To be more specific, SMUCE solves the optimization problem

$$\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \qquad \text{subject to} \qquad T_n(\mathbf{Y}, \boldsymbol{\theta}) \leq c, \tag{2.9}$$

where $J(\boldsymbol{\theta})$ is the number of change points, and c is a threshold. The multiscale statistic $T_n(\mathbf{Y}, \boldsymbol{\theta})$ is given by the maximum of the local likelihood ratio statistic over all possible segments of the parameter $\boldsymbol{\theta}$ in which no change occurs. In other words, $T_n(\mathbf{Y}, \boldsymbol{\theta})$ is the maximum of the likelihood ratio statistic of the hypothesis $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$, over the interval $\big[i/n\,,\, j/n\big]$, for $i, j = 1, \ldots, n$. One advantage of this method is that it provides tools for building a confidence set for the canonical parameter $\boldsymbol{\theta}$. The computational complexity of SMUCE via dynamic programming is of the order $O(n^2)$ and can be reduced by some pruning techniques. Because SMUCE controls the family-wise error rate (FWER), the method is conservative and usually underestimates the number of change points, particularly in data sequences with many change points or a small signal-to-noise ratio. This shortcoming was addressed in [99], and the method FDRseg, based on False Detection Rate (FDR) control, is suggested. The problem of change point detection in canonical parameters of exponential families was studied in [120], and [40] for heteroscedastic and dependent noise variables, respectively.

In a more general case, [111] proposed an approach for change point detection using a class of non-convex penalty functions. They compared the performance of the penalized likelihood problem with the SCAD penalty [49], the bridge penalty [58], and the unbounded penalty [94] in the context of change point analysis. It has been shown that the penalized likelihood method using the above penalty functions does not guarantee consistent recovery of change points. By combining the lasso and the unbounded penalty, [111] introduced a new penalty, termed modified unbounded, which allows consistent estimation of the number of change points, their locations, and their magnitudes.

2.1.2 Test-Based Approaches

Methods that search for change points one at a time often use the likelihood ratio test or CUSUM as their basis for finding a single change at each step. Among the most popular

methods, we review Binary Segmentation [158] and its extensions: Circular Binary Segmentation (CBS) [114], Wild Binary Segmentation (WBS) [56] and Wild Binary Segmentation 2 and Steepest Drop to Low Levels (WBS2.SDLL) [54]. We also discuss more-recent change point detection methods in the literature including Tail Greedy Unbalanced Haar (TGUH) [57], Narrowest Over Threshold (NOT) [16], Isolate-Detect (ID) [4] and Narrowest Significance Pursuit (NSP) [55].

As an extensively used method in multiple change point detection, Binary Segmentation is a forward selection method that recursively searches for a single mean change within a certain regime of data points. It starts with testing the existence of a change in the entire original data sequence using a CUSUM statistic. After detecting a change, it divides the data sequence into two subsegments, one before and one after the change point. Then the search for a change continuous within each subsegment, and splitting is carried out. This procedure is repeated until no further division is possible. The following algorithm elaborates Binary Segmentation in steps.

Algorithm 2.2 Binary Segmentation Algorithm:

Step 1. *Perform the test* (2.3) *for mean change and, if the null hypothesis is not rejected, then stop the algorithm and return* $\boldsymbol{\tau} = \emptyset$ *as the change point set. Otherwise, set* $\boldsymbol{\tau} = \{\widehat{\tau}_1\}$ *where* $\widehat{\tau}_1$ *is computed as the maximizer of the CUSUM statistic in* (2.6)*, and split the data into two subsegments* $\mathbf{y}_{1:\widehat{\tau}_1} = \big(y_1, \ldots, y_{\widehat{\tau}_1}\big)$ *and* $\mathbf{y}_{\widehat{\tau}_1+1:n} = \big(y_{\widehat{\tau}_1+1}, \ldots, y_n\big)$.

Step 2. *Similar to step 1, test for the existence of a change for these new subsegments, and update the change point set* $\boldsymbol{\tau}$.

Step 3. *Repeat the procedure until no further segmentation is possible.*

Proposed by [158], Binary segmentation is a prevalent method in change point detection analysis and has been widely used by scientists in various disciplines. Simplicity, fast implementation with the computational complexity of order $O(n \log n)$, and relative accuracy account for its popularity. The consistency of Binary Segmentation in identifying both the number and location of change points has been established for cases in which the minimum distance among neighbouring segments is of an order greater than $O(n^{3/4})$, [157], [56]. However, Binary Segmentation is suboptimal in rates of convergence when the number of change points is fixed or is growing with the sample size n. On the other hand, lack of ability to identify a change point located within a long segment is a major limitation of Binary Segmentation because this method searches only for one change point

inside each subsegment. The limited nature of such searches leads to poor performance if there are frequent change points. Thus, Binary Segmentation is likely to estimate change point locations inaccurately. Some modified versions of this method, [114], [56], [54] and [57] have been developed to solve this shortcoming and improve the accuracy of Binary Segmentation.

The first adaptation of Binary Segmentation, Circular Binary Segmentation (CBS) [114], suggests making a circular sequence by tying both ends of the original signal together. The test (2.3) can then be performed for two segments $\mathbf{y}_{s+1:e}$ and $\left(\mathbf{y}_{1:s}\,,\,\mathbf{y}_{e+1:n}\right)$, for a fixed pair (s,e), $s<e$. Implementation of the test over all possible values of $(s\,,\,e)$ makes CBS computationally expensive; however, it is more powerful than Binary Segmentation in determining short segments.

To improve efficiency, [56] developed a method called Wild Binary Segmentation (WBS) for mean change point problems (2.2). The idea is to randomly select a user-specified number of subsegments, say M, from a segment of data and perform the single change test for these selected subsegments. The method is based on the hope that one of the selected subsegments will catch the change within short segments. More precisely, suppose that intervals $[s_j\,,\,e_j]$, $j=1,\ldots,M$, such that $s_j<e_j$ and $s_j,\,e_j\in\{s,\ldots,e\}$, are randomly selected from the original dataset with the start point s and endpoint e. The interest is then in testing

$$H_{0j}:\mu_{s_j}=\ldots=\mu_{e_j}\quad\text{versus}\quad H_{1j}(\tau):\mu_{s_j}=\ldots=\mu_{\tau}\neq\mu_{\tau+1}=\ldots=\mu_{e_j}, \tag{2.10}$$

for $\tau=s_j,\,\ldots\,,e_j-1$. The maximum value of the CUSUM statistic in (2.6) over the j-th random interval is given by $U_j=\max_{s_j\leq\tau\leq e_j-1}U_j(\tau)$. The change point estimate is the location in the random intervals $[s_j\,,\,e_j]$ that maximizes the U_j over all $j=1,\ldots,M$. In other words, let $\widehat{j}=\underset{1\leq j\leq M}{\arg\max}\,U_j$, then the change point location is

$$\widehat{\tau}_{\text{WBS}}\;=\;\underset{s_{\widehat{j}}\leq\tau\leq e_{\widehat{j}}-1}{\arg\max}\;U_{\widehat{j}}\,(\tau). \tag{2.11}$$

Note that, in addition to the threshold value for CUSUM tests, the number of random draw subsamples M is crucial for both the accuracy and efficiency of the WBS.

In recent years, some papers have attempted to improve the performance of Binary Segmentation and WBS. [57] argue that the weak performance of Binary Segmentation is in part due to its forward (top-down) nature, meaning that it starts with the entire signal and splits it into shorter segments as the process goes on. The authors suggest a backward (bottom-up) mechanism, labelled Tail Greedy Unbalanced Haar (TGUH), which operates

by fusing a number of successive neighbouring segments that most likely share the same structure. In light of the need to merge multiple consecutive segments and to facilitate the procedure, TGUH uses Unbalanced Haar transformation on the original data sequence. The computational cost of the method is $O(n \log^2(n))$ regardless of the number of change points. It also guarantees the consistency of the estimations in both the number of change points and their locations.

WBS is designed to handle cases in which more than one change point exists in each segment. It has been shown that it performs well for signals with a small or moderate number of change points. Nonetheless, this performance deteriorates substantially in datasets with frequent changes. [54] have offered a method called Wild Binary Segmentation 2 and Steepest Drop to Low Levels (WBS2.SDLL) to deal with the aforementioned challenges in scenarios with frequent change points. The method solves the issues in two separate steps. In the first step, called WBS2, an entire solution path for the problem is constructed with $0, 1, \ldots, n-1$ change points by executing WBS with a much smaller number of drawn subsamples M. The distinction between WBS and WBS2 is in the size of M. In WBS2, M is relatively tiny compared to WBS (100 in WBS2 compared to 20000 in WBS). The second step, named SDLL, focuses on choosing the next change points. To this end, SDLL computes the ratio of the two successive CUSUM statistics obtained in the WBS2 step. The statistic with the largest ratio that exceeds a provided threshold is regarded as the next change point. The method has been proven fast in run time and accurate in detection.

The method Narrowest Over Threshold (NOT), put forward by [16], provides a solution to the problem of change point detection in general models, particularly piecewise constant and piecewise linear models. In the first step, M random subsamples denoted by $[s_j\,,\, e_j],\ j = 1, \ldots, M$ are drawn from the entire dataset. A suitable cost function (a negative log-likelihood in most cases) is chosen and then applied to select potential change points within each drawn subsample. More specifically, let $C^{b}_{s\,,\,e}(\mathbf{y})$ denote the value of the cost function for all data points b inside the subsample $[s\,,\, e]$. Calculate the maximum value of the cost function over subsample $[s\,,\, e]$, i.e., $C^{b_{\max}}_{s\,,\,e}(\mathbf{y}) = \max\limits_{b\,\in\,\{s,\,\ldots\,,e\}} C^{b}_{s\,,\,e}(\mathbf{y})$. Repeat this calculation for all M randomly drawn subsamples.

In the second step, all derived values $C^{b_{\max}}_{s_j\,,\,e_j}$, for $j = 1, \ldots, M$, are compared with a pre-specified threshold to test which interval contains a significant change. The interval $[s_j\,,\, e_j]$ with the smallest length (narrowest) is chosen from all intervals with a significant change. Finally, the location that maximizes the cost function for the narrowest interval is identified as a change point estimation. A similar procedure is recursively carried out to the left and right sides of the estimated change point. The search for more change points stops when there is no interval with a cost function that exceeds the threshold.

The computational complexity of NOT is of the order $O(n\log(n))$ if we select the optimal choice for M. The order is almost linear in the size of the data. The method can be extended to cover more general models such as piecewise polynomials if a suitable cost function is defined. Moreover, NOT provides an asymptotically consistent estimator of the number and locations of change points. The benefit of NOT over methods like WBS is that it considers the smallest interval, thereby preventing subsamples from containing more than one change point.

In a recent work, [4] developed a new approach called Isolate-Detect (ID) to consistently estimate change points in a sequence of data. The ID method in change point analysis involves two stages: isolation and estimation. In the isolation stage, the aim is to identify subintervals of the entire domain that contain only one change point. This isolation is carried out by a process called interval expansion, which updates intervals upon detection. Having such subintervals at hand, the estimation stage seeks to locate the change point within each subinterval. For a suitably chosen loss function, the location of its maximum value is taken to be a change point if the maximum value exceeds a certain threshold. Once an appropriate loss function has been appointed, the ID technique can estimate changes in models other than piecewise constant ones.

ID localizes change points faster than the NOT and WBS methods, which is of great importance, particularly in data sequences with a large number of change points. Moreover, due to its interval expansion, it covers the entire domain of the dataset and enables searching for all possible change points. The authors have demonstrated that under certain mild conditions, the change point estimates derived using ID are consistent.

[55] has also studied the problem of change point detection in a linear model and has proposed a method called Narrowest Significance Pursuit (NSP). This method is a forward procedure and proceeds recursively to the right and the left side of a new change point until no further significant change point can be found. NSP uses a particular multiresolution sup-norm loss function at each step to find the shortest interval on which the slope of the underlying linear model is significantly changed. It is shown that NSP performs well with the assumptions beyond identical and independent Gaussian random noises.

Most existing change point detection methods in the literature focus on univariate data with possible changes in the mean only and assume that the other characteristics are unchanged. Other methods provide solutions to the change point problem in other characteristics such as variance [62, 161] and covariance [7, 10, 42]. In principle, more complex change point problems that study changes in characteristics other than the mean can be converted to the canonical mean change problem [28]. This conversion can be executed by applying a suitable transformation to the original dataset. See [31, 73, 32] and

[33].

Another important field of active research in change point analysis deals with high-dimensional datasets. For example, the performance of the least square estimator of a single change point in a high dimensional setting was studied in [14]. [172] and [48] applied CUSUM statistics to estimate change points in a high dimensional setting. Additionally, under the sparsity assumption for change points in a high dimensional setting, [84, 137, 163, 162] and [130] have proposed solutions to the change point problem.

There is no single best approach for all change point problems and applications. The existing approaches can be compared based on their properties, such as consistency and the rate of convergence to estimate the true number and exact locations of change points J, plus the approaches' computational complexity and scalability. In the change point literature, a suitable approach is frequently considered to be the one that would closely estimate the number of change points, and their locations as the number of data points grows. In statistical terminology, this point implies that the method must yield consistent estimations.

2.2 Trend Filtering

Over the past 30 years, an active line of statistical research has been devoted to pattern recovery – techniques for estimating unknown parameter vectors by imposing certain structures on them and then verifying the conditions under which the techniques perform well. Pattern recovery, in which the positions of nonzero elements of an unknown parameter vector are identified, appears in a wide variety of disciplines, including compressed sensing [44], signal processing [30], and model selection in regression [109]. One way of approaching pattern recovery problems is to use an optimization problem to compromise a loss function and a penalty criterion. A natural choice of penalty function is ℓ_0-norm, which regularizes the number of nonzero coordinates of a parameter vector. Unfortunately, minimization of the ℓ_0-regularization problem is intractable due to its non-convexity. As a computationally feasible surrogate, the ℓ_1-regularization problem has attracted much attention over the past three decades.

As a generalized form of ℓ_1-regularization and in order to cover a wide range of models with structural constraints [154] introduced *generalized lasso*. The generalized lasso objective function contains negative log-likelihood loss along with the ℓ_1-norm of a specific matrix times coefficients vector. More clearly, let $\mathbf{y} \in \mathbb{R}^n$ be the response vector, and

$\mathbf{X} \in \mathbb{R}^{n \times p}$ be the predictor matrix; then the least square generalized lasso is formulated as

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{2} \left\| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \right\|_2^2 + \lambda \left\| \mathbf{D}\boldsymbol{\beta} \right\|_1, \tag{2.12}$$

where $\mathbf{D} \in \mathbb{R}^{m \times p}$ is called the penalty matrix and reflects the structure of the model and $\lambda \geq 0$ is the regularization parameter.

The r-th order trend filtering is a special case of the generalized lasso if the predictor and penalty matrices are replaced by the identity and the r-th order discrete difference operator, respectively. Such difference operator imposes sparsity in the r-th difference of the model, and as a result, the r-th trend filtering fit is a piecewise polynomial of order r. In more concrete terms, define the first-order discrete difference operator matrix $\mathbf{D}^{(1)} \in \mathbb{R}^{n-1 \times n}$ as

$$\mathbf{D}^{(1)} = \begin{pmatrix} -1 & 1 & 0 & 0 & \dots & 0 \\ 0 & -1 & 1 & 0 & \dots & 0 \\ 0 & 0 & -1 & 1 & \dots & 0 \\ \vdots & & & & & \vdots \end{pmatrix},$$

then for $r \geq 1$, the $(r+1)$-th operator $\mathbf{D}^{(r+1)} \in \mathbb{R}^{n-r-1 \times n}$ can be recursively computed as $\mathbf{D}^{(r+1)} = \mathbf{D}^{(1)} \times \mathbf{D}^{(r)}$ where $\mathbf{D}^{(1)}$ is the $(n-r-1) \times (n-r)$ version of the first discrete difference matrix. Here are the examples of $\mathbf{D}^{(2)}$ and $\mathbf{D}^{(3)}$,

$$\mathbf{D}^{(2)} = \begin{pmatrix} 1 & -2 & 1 & 0 & \dots & 0 \\ 0 & 1 & -2 & 1 & \dots & 0 \\ 0 & 0 & 1 & -2 & \dots & 0 \\ \vdots & & & & & \vdots \end{pmatrix}, \qquad \mathbf{D}^{(3)} = \begin{pmatrix} -1 & 3 & -3 & 1 & \dots & 0 \\ 0 & -1 & 3 & -3 & \dots & 0 \\ 0 & 0 & -1 & 3 & \dots & 0 \\ \vdots & & & & & \vdots \end{pmatrix}.$$

Hence, the r-th order trend filtering estimate is the solution of

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^n} \frac{1}{2} \left\| \mathbf{y} - \boldsymbol{\beta} \right\|_2^2 + \lambda \left\| \mathbf{D}^{(r+1)} \boldsymbol{\beta} \right\|_1. \tag{2.13}$$

If $r = 0$, trend filtering reduces to one-dimensional fused lasso, [148], also well-known as one-dimensional total variation denoising [133], and produces piecewise constant estimates. The existence and uniqueness of the trend filtering estimates are discussed in Section 3.3.

The problem (2.13) was first studied by [140] in context of the image processing and called *higher order total variation regularization*. It was later rediscovered by [86] and

termed as the trend filtering. In their paper, [86] established the properties of linear trend filtering ($r = 1$) which fits piecewise linear models. An extensive study of the trend filtering is carried out by [151]. The paper investigates the properties of trend filtering and justifies it as a great tool in nonparametric regression. It points out that the trend filtering resembles smoothing splines [64] and locally adaptive regression splines [107] in the sense of being its continuous-time versions. More importantly, [151] establishes the convergence of trend filtering to the true underlying function with minimax rate. In a recent paper, [149] have also explored connections between trend filtering and discrete splines.

From a computational and algorithmic standpoint, [86] describe the Primal-Dual Interior-Point (PDIP) method to derive estimates of the linear trend filtering at a fixed λ which can be readily carried over to the trend filtering of any order. The idea relies on solving a constrained quadratic problem using the interior point method. The computational complexity of the method in practice is of order $O(n)$, which indicates its efficiency. Another well-suited approach for deriving estimates of the trend filtering is constructing a solution path using its dual problem. Such an algorithm is developed in [154] for the generalized lasso, which is simply applicable for trend filtering by setting $\mathbf{D} = \mathbf{D}^{(r+1)}$ and $\mathbf{X} = \mathbf{I}$.

In order to accelerate solving the trend filtering optimization, [164] suggest the falling factorial basis, which enables fast computation of the solution in order of $O(n)$. [124] derived a fast and efficient algorithm for solving (2.13) based on the Alternating Direction Method of Multipliers (ADMM) presented first in [25]. They then compared their approach with PDIP, empirically and theoretically and showed that the specialized ADMM algorithm converges faster and more accurately.

2.3 Post-Selection Inference

Classical statistical theory provides tools for performing inference about pre-specified statistical questions, determined before observing the data– such as hypothesis testings and constructing confidence intervals. However, in many applications, data are collected without such questions. An exploratory data analysis is usually used to generate interesting questions and then provides preliminary answers to them. Consequently, most statistical methods are composed of two stages [51]:

(i) *Model Selection:* A practitioner applies the underlying dataset to decide which models are worthy of attention. This stage allows us to determine statistical questions, such as formulating estimations and hypothesis testings.

(ii) *Target Inference:* The practitioner would like to evaluate the selected model using inferential methods. Essentially, this stage deals with statistically answering questions generated in the model selection stage.

In statistical terminology, a selected model is called adaptive if a dataset is used in its derivation. Adaptive models are stochastic due to their dependence on random observations. Two broad classes of such stochastic models obtained from certain selection procedures are [22]:

- Models derived from variable selection procedures such as Forward Stepwise or Backward Selection, or models derived from regularized regression, for instance, ridge regression and Lasso.
- Models derived in an ill-defined way such as those chosen by visual inspection or by some regression diagnostic methods.

A consequence of such adaptive models is having random hypothesis testings and confidence intervals. To be more explicit, we present the following example.

Example 2.3 *Consider the linear regression setup with the $n \times p$ predictor matrix $\mathbf{X}$ and response vector $\mathbf{y} \in \mathbb{R}^n$ drawn from $N(\boldsymbol{\mu}\,, \sigma^2\,\mathbf{I})$. Suppose that lasso is run at a fixed value of the regularization parameter λ, and a subset of variables $\widehat{\mathbf{M}}(\mathbf{y}) = \widehat{\mathbf{M}} \subset \{1, \ldots, p\}$ is chosen (model selection). One might be interested in computing p-value for the significance of a specific coefficient in the selected model, $\beta_{j\,,\,\widehat{M}}$, that is, $H_0 : \beta_{j\,,\,\widehat{M}} = 0$, for $j \in \widehat{\mathbf{M}}$ (target inference).*

Classical statistical theory treats the selected model as a fixed model and assumes that the hypothesis is known in advance. To test the hypothesis $H_0 : \boldsymbol{\eta}^T\boldsymbol{\mu} = 0$ for a pre-specified nonzero vector $\boldsymbol{\eta} \in \mathbb{R}^p$, one can use the usual Z-test (t-test for unknown σ^2). In our setting, these tests are invalidated because the data are used to choose which coefficients should be tested. The preceding hypothesis $H_0 : \beta_{j\,,\,\widehat{M}} = 0$ is equivalent to $H_0 : \boldsymbol{\eta}^T\boldsymbol{\mu} = 0$, where

$$\boldsymbol{\eta} = \mathbf{X}_{\widehat{M}}\Big(\mathbf{X}_{\widehat{M}}^T\,\mathbf{X}_{\widehat{M}}\Big)^{-1}\boldsymbol{e}_j.$$

In this representation, $\mathbf{X}_{\widehat{M}}$ is the matrix extracted from columns of the design matrix $\mathbf{X}$ indexed by the model $\widehat{\mathbf{M}}$, and $\mathbf{e}_j$ is the basis vector with the j-th element being 1, and remaining elements being 0. The classical statistical tools are misleading in this case since $\boldsymbol{\eta} = \boldsymbol{\eta}\big(\,\widehat{\mathbf{M}}(\mathbf{y})\,\big)$ depends on the dataset and thus random.

As explained above, classical statistical inference is misleading for models chosen by a selection procedure. The reason is that the selected models are stochastic and are not accounted for in the classical theory. For example, classical inference considers the same sampling distribution for post-selection inference. However, this distribution is no longer the same as the original sampling distribution, because of the stochastic nature of a selection procedure. Moreover, the randomness of data-driven hypotheses implies that the Type-I error differs from that of classical theory in which the hypotheses are pre-specified. We make this point in the following example.

Example 2.4 (File Drawer Effect [51]) *Suppose random variables* Y_i, $i = 1, \ldots, n$ *are drawn independently from* $N(\mu_i\,,\,1)$. *For selection, we consider the set* $\widehat{I} = \{i:\ |Y_i| > 1\}$, *which selects variables with large effects. Then, the goal is to test* $H_{0\,,i}:\ \mu_i = 0$ *at the nominal level* $\alpha = 0.05$. *For* $i \in \widehat{I}$, *classical inference ignores the selection of large effect variables and rejects* $H_{0\,,i}$ *when* $|Y_i| > 1.96$. *However, the error rate for hypotheses selected for testing is no longer* 0.05. *More specifically, let* n_0 *denote the number of true null effects and assume that* $n_0 \longrightarrow \infty$ *as* $n \longrightarrow \infty$. *Then, the fraction of errors among the true nulls we select is*

$$\frac{\#\textit{False Rejections}}{\#\textit{True Nulls Selected}} = \frac{\sum_{i:\,H_{0\,,i}\ true} \mathbb{1}\left\{i \in \widehat{I}\ ,\ \textit{Reject } H_{0\,,i}\right\} / n_0}{\sum_{i:\,H_{0\,,i}\ \ true} \mathbb{1}\left\{i \in \widehat{I}\right\} / n_0}$$

$$\rightarrow \frac{Pr_{H_{0\,,i}}\left(i \in \widehat{I}\ ,\ \textit{Reject } H_{0\,,i}\right)}{Pr_{H_{0\,,i}}\left(i \in \widehat{I}\right)}$$

$$= Pr_{H_{0\,,i}}\left(\textit{Reject } H_{0\,,i} \mid i \in \widehat{I}\right) = \frac{\Phi(-1.96)}{\Phi(-1)} = 0.16.$$

For years, a major gap existed between model selection procedures and inferential tools for selected models. Fortunately, a growing line of work has recently been dedicated to statistical inference after model selection. The framework to make inferences for an adaptive model, referred to as *post-selection inference*, also known as *selective inference*, attempts to perform statistical inference such as hypothesis testing and confidence interval construction.

The problem of post-selection inference was first discussed in [122], and later, in a sequence of articles, [95, 96, 97] explored how to estimate post-selection distributions. This

subject has attracted much attention more recently as applications for model selection have proliferated. Several researchers have attempted to address the post-selection inference problem by applying conditional statistical inference. In particular, in a hypothesis testing problem, if the model $\mathbf{M}$ and the null hypothesis H_0 are adaptively chosen, the goal is to control the Type-I error rate

$$\Pr\Big(\text{Reject } H_0 \,\big|\, (\mathbf{M}\,,\, H_0)\Big),$$

at the nominal level α.

Classical hypothesis testing assumes that the null hypothesis and the model are independent of the data used for inference and, therefore, conditioning on $(\mathbf{M}\,,\, H_0)$ is not required. This fact has inspired a class of inferential approaches called *data splitting*. The idea, originated by [37], states that the data can be divided into two independent parts. One part is applied to choose a model, and the other part is used to make inferences about that model. The inferential targets in the data splitting framework are fixed due to the independence of the two splits. Consequently, conditioning on the selected model is not required. However, carrying out this idea is not free. The cost is in the decreased size of data for both selection and inference phases. Moreover, it is not always possible to split the data into two independent parts, such as autocorrelated spatial and temporal data. For more details on data splitting, we refer the reader to [165, 110].

Apart from data splitting, one natural solution to post-selection inference is to condition the analysis on the selected model. More specifically, to perform hypothesis testing $H_{0,i}^{\mathbf{M}}$, for any i in the selected model $\mathbf{M}$, we are interested in tests that control the post-selection Type-I error at level α, i.e.,

$$\Pr\Big(\text{Reject } H_{0,i}^{\mathbf{M}} \;\big|\; (\mathbf{M}\,,\, H_{0,i}^{\mathbf{M}})\Big) \;\leq\; \alpha\,, \qquad\qquad i \in \mathbf{M}. \tag{2.14}$$

These types of tests are called *post-selection hypothesis tests*. In the same manner, a *post-selection confidence interval* forms a confidence interval for a parameter of interest in the selected model $\mathbf{M}$, namely $\theta_i^{\mathbf{M}}$, for $i \in \mathbf{M}$. More precisely, the interval $I_i^{\mathbf{M}}$ is a post-selection confidence interval at the level of $(1-\alpha)$ for $\theta_i^{\mathbf{M}}$, if

$$\Pr\Big(\theta_i^{\mathbf{M}} \in I_i^{\mathbf{M}} \;\big|\; i \in \mathbf{M}\Big) \;\geq\; 1-\alpha. \tag{2.15}$$

The interpretation for the post-selection confidence interval is as follows: if we were to repeatedly generate $\mathbf{y}$ from the underlying model and apply the corresponding selection procedure, and only focus on cases in which we selected model $\mathbf{M}$, then among such cases,

the constructed confidence intervals $I_i^{\mathbf{M}}$, would contain $\theta_i^{\mathbf{M}}$, $i \in \mathbf{M}$, with a frequency tending to $1-\alpha$. See [153] and [152]. Analogous to the classical case, there is a one-to-one correspondence between the post-selection tests and confidence intervals. Indeed, the post-selection confidence intervals can be computed by inverting post-selection tests, [51].

2.3.1 Post-Selection Inference in Regression Setting

Hereafter, we restrict our attention to the regression framework as one of the important applications of post-selection inference, especially when one is conducting hypothesis testing and constructing a confidence interval for model parameters. In a regression model, let $\mathbf{y} \in \mathbb{R}^n$ be a response vector and $\mathbf{X} = \left(X_1, \ldots, X_p\right) \in \mathbb{R}^{n \times p}$ be the full predictor matrix. Also, suppose that the model $\mathbf{M} \subset \{1, \ldots, p\}$ is chosen using a selection procedure. The interest is then to make inference for parameters associated with components of $\mathbf{M}$, denoted by $\boldsymbol{\beta}^{\mathbf{M}}$. In other words, targets of inference in post-selection inference are coefficients contained in the selected model, but not those that are excluded. This setting is called the *submodel view*, which assumes that the selected model has its own variables and that the excluded variables do not exist. In contrast, the *full model view* assumes that excluded variables are in the model but with zero coefficients [22].

In practice, a model is adaptively chosen through a selection procedure that uses random response $\mathbf{y}$. This estimated model, denoted by $\widehat{\mathbf{M}} = \widehat{\mathbf{M}}(\mathbf{y})$ is a result of the map $\widehat{\mathbf{M}} : \mathbf{y} \longmapsto \widehat{\mathbf{M}}(\mathbf{y})$ with some crucial natural properties. First, due to the involvement of random response $\mathbf{y}$ in model selection, $\widehat{\mathbf{M}} = \widehat{\mathbf{M}}(\mathbf{y})$ is a random variable and so are its corresponding parameters $\boldsymbol{\beta}^{\widehat{\mathbf{M}}}$. Second, the selected model $\widehat{\mathbf{M}}$ may or may not contain a fixed $i \in \{1, \ldots, p\}$. According to the submodel view, this fact implies that the inference of $\beta_i^{\widehat{\mathbf{M}}}$ is undefined when $i \notin \widehat{\mathbf{M}}$. Third, the interpretation and estimation of the parameter $\beta_i^{\widehat{\mathbf{M}}}$, given $i \in \widehat{\mathbf{M}}$, changes due to its dependence on other parameters contained in the model.

[22] described a method for deriving a confidence interval by controlling the family-wise error rate (FWER), regardless of the selection procedure. The method leads to universally valid confidence intervals. Universal validity is a strong property with applications in areas where model selection is not pre-specified. However, this property is very conservative and produces undesirably long confidence intervals.

[102] considered the significance test of a predictor variable that enters the active set of lasso. The authors applied the fitted values of lasso at a given value of regularization parameter to propose a test statistic, the *covariance test statistic*. They established that the asymptotic null distribution of the covariance test statistic is the standard exponential

under the assumptions that the entries of the predictor matrix are in general position and that the true model is linear. It is important to point out that this test only checks whether a predictor variable that enters the current model is significant.

[93] proposed a novel idea for statistical inference of the lasso estimates. The idea states that the selected model $\widehat{\mathbf{M}} = \widehat{\mathbf{M}}(\mathbf{y})$ corresponds to a polyhedron set over the sample space of random response $\mathbf{y}$. More specifically, they established that for a lasso selected model $\widehat{\mathbf{M}}$ with sign $\widehat{\mathbf{s}}_{_M}$ at a fixed value of regularization parameter, the event $\left(\widehat{\mathbf{M}}, \widehat{\mathbf{s}}_{_M}\right)$ can be characterized as a polyhedron of the form $\{\mathbf{y} \in \mathbb{R}^n : \ \mathbf{A}\,\mathbf{y} \geq \mathbf{q}\}$. Consequently, the procedure of making inference given the selected model boils down to the restriction of the sample space of $\mathbf{y}$ to a polyhedron. This characterization could be of great help for inference after model selection. In Particular, when the distribution of the response vector is Gaussian, an exact test based on a truncated Gaussian distribution has been derived [93, 153]. We will describe this idea in more detail in Chapter 4.

[153] concurrently demonstrated the same results as [93] for other model selection methods such as the forward stepwise and LARS as well as lasso. The extension of the truncated Gaussian test in this work is based on the models derived after a fixed number of steps. This characteristic distinguishes this work from [93], which considers models chosen at a fixed value of the regularization parameter. In a very recent paper, [77] study characterization of the polyhedron in the generalized lasso framework. Comparison of the exact truncated Gaussian among various selection techniques reveals that each technique creates a typical polyhedron. A number of extensions to different frameworks and applications are given in [145], [103], [144], [126] and [31].

Several authors have extended the post-selection inference setting to non-Gaussian distributions. [51] study the theoretical properties of post-selection inference and generalize the framework to the broad class of distributions: the exponential family. [145] establish an asymptotic framework for post-selection inference in a high-dimensional setting while removing the Gaussian assumption. This method is performed for a specific class of model selection covering the affine selection procedures. [152] proves the uniform convergence of the post-selection test statistics in the case of non-Gaussian observations. The same asymptotic framework has been proposed by [146] for randomized responses.

Chapter 3

Detection of Change Points in Piecewise Polynomial Signals Using Trend Filtering

While many approaches have been proposed for discovering abrupt changes in piecewise constant signals, few methods are available to capture these changes in piecewise polynomial signals. In this chapter, we propose a change point detection method, PRUTF, based on trend filtering. By providing a comprehensive dual solution path for trend filtering, PRUTF allows us to discover change points of the underlying signal for either a given value of the regularization parameter or a specific number of steps of the algorithm. We demonstrate that the dual solution path constitutes a Gaussian bridge process that enables us to derive an exact and efficient stopping rule for terminating the search algorithm. We also prove that the estimates produced by this algorithm are asymptotically consistent in pattern recovery. This result holds even in the presence of staircases (consecutive change points with the same signs) in the signal. Finally, we investigate the performance of our proposed method for various signals and then compare its performance against some state-of-the-art methods in the context of change point detection. We apply our method to three real-world datasets, including the UK House Price Index (HPI), the GISS Surface Temperature Analysis (GISTEMP) and the Coronavirus disease (COVID-19) pandemic.

3.1 Introduction

We consider the univariate signal plus noise model

$$y_i = f_i + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{3.1}$$

where $f_i = f(i/n)$ is a deterministic and unknown signal with equally spaced input points over the interval [0, 1]. The error terms $\varepsilon_1, \ldots, \varepsilon_n$ are assumed to be independently and identically distributed Gaussian random variables with mean zero and finite variance σ^2. We assume that $f(\cdot)$ undergoes J_0 unknown and distinct changes at point fractions $0 = \omega_0 < \omega_1 < \ldots < \omega_{J_0} < \omega_{J_0+1} = 1$, where the number of change point fractions, J_0 can grow with the sample size n. Additionally, we assume that $f(\cdot)$ is a piecewise polynomial function with any arbitrary but fixed order r. These assumptions imply that, associated with $\omega_0, \ldots, \omega_{J_0+1}$, there are change points locations $0 = \tau_0 < \tau_1 < \ldots < \tau_{J_0} < \tau_{J_0+1} = n$, which partition the entire signal $\mathbf{f} = (f_1, \ldots, f_n)$ into J_0+1 segments. More specifically, any subsignal of $\mathbf{f}$ within segments created by the change points follows an r-degree polynomial structure with or without a continuity constraint at the change points. In other words, for a piecewise polynomial signal of order r, at least one of the coefficients in the polynomial function alters at a change point location. See Figure 3.2 for visual schematic of such signals. Change in the level of a piecewise constant signal, known as the canonical multiple change point, and change in the slope of a piecewise linear signal are examples of the problem under consideration in this chapter. In change point analysis, the objective is to estimate the number of change points, J_0, as well as their locations $\boldsymbol{\tau} = \{\tau_1, \ldots, \tau_{J_0}\}$ based on the observations $\mathbf{y} = (y_1, \ldots, y_n)$.

The canonical multiple change point problem, where the signal f is modelled as a piecewise constant function, has been extensively studied in the literature. Beyond the canonical change point problem, signals in which f is modelled as a piecewise polynomial of order $r \geq 1$ have attracted less attention in the literature despite many applications. For instance, piecewise linear signals are applied in monitoring patient health ([3], [139]), climate change ([131]), and finance ([23]). In such a framework, [13] introduced a method based on Wald-type sequential tests, and [105] devised a dynamic programming applied to an ℓ_0-penalized least square model. In continuous piecewise linear models, [86] developed a methodology called ℓ_1-trend filtering. Furthermore, [16] put forward the method of Narrowest Over Threshold (NOT), and [4] developed an approach called Isolate-Detect (ID) which both estimate change points in more-general change point problems.

This chapter aims to introduce a unifying method covering the canonical change point problem and beyond. More precisely, the method can detect change points in piecewise

polynomial signals of order r ($r = 0, 1, 2, \ldots$) with and without continuity constraint at the locations of change points.

The canonical change point problem for a sequence of data can be formulated as a penalized regression fitting problem. According to our notation, the quantity $f_\tau - f_{\tau+1}$ is nonzero if the signal f undergoes a change at point τ, and is zero otherwise. Moreover, if we assume that change points are sparse, that is, the number of locations where f changes, J_0, is much smaller than the number of observations n, change points can be estimated using the one-dimensional fused lasso problem

$$\min_{\mathbf{f}\in\mathbb{R}^n} \frac{1}{2} \left\| \mathbf{y} - \mathbf{f} \right\|_2^2 + \lambda \sum_{i=1}^{n-1} \left| f_{i+1} - f_i \right|,$$

where $\mathbf{f} = (f_1, \ldots, f_n)$.

This formulation of the canonical change point problem was first considered in [76] and was applied to analyze a DNA copy number dataset. [68] considered the same formulation and proved the consistency of the respective change point estimates when the number of change points is bounded. Employing sparse fused lasso, which is composed of both the ℓ_1-norm and the total variation seminorm penalties, [129] proposed a sparse piecewise constant fit and established the consistency of the corresponding estimates when the variance of the noise terms vanishes and the minimum magnitude of jumps is bounded from below. However, [132] argued that the consistency results achieved by [129] are incorrect when a frequently viewed pattern, called *staircase*, exists in the signal. The staircase phenomenon occurs in a piecewise constant model when there are either two consecutive downward jumps or upward jumps in its mean structure. The staircase pattern will be discussed in more detail in Section 3.7. Additionally, [123] showed that the lasso problem, when derived by transforming fused lasso, does not satisfy the Irrepresentable Condition ([173]) that is necessary and sufficient for exact pattern recovery. In particular, [123] proposed an approach called preconditioned fused lasso based on the puffer transformation of [83] and established that it can recover the exact pattern with probability approaching one.

A similar approach to that of the piecewise constant signals can be considered for estimating change points in piecewise polynomial signals. In particular, a positive integer τ is a change location in an r-th degree piecewise polynomial signal f if τ-th element of the vector $\mathbf{D}^{(r+1)}\mathbf{f}$ is non-zero, denoted by $\left[\mathbf{D}^{(r+1)}\mathbf{f}\right]_\tau \neq 0$. Here $\mathbf{D}^{(r+1)}$ is a penalty matrix that was defined in Section 2.2. Hence, change points can be estimated from nonzero elements of $\mathbf{D}^{(r+1)}\widehat{\mathbf{f}}$, where $\widehat{\mathbf{f}}$ is the solution of

$$\min_{\mathbf{f}\in\mathbb{R}^n} \frac{1}{2} \left\| \mathbf{y} - \mathbf{f} \right\|_2^2 + \lambda \left\| \mathbf{D}^{(r+1)}\mathbf{f} \right\|_1. \tag{3.2}$$

The trend filtering problem (3.2) fits a piecewise polynomial of order r for the true signal f. First-order trend filtering was developed by [86] for piecewise linear fits. They provided a primal-dual interior-point algorithm to fit a model at a specified value of the regularization parameter $\lambda > 0$. For an arbitrary $r \in \mathbb{N}$, [154] also provided a solution path algorithm that yields the trend filtering estimates over all values of the regularization parameter λ. In a separate paper, [151] focused mainly on the statistical properties of the trend filtering estimates and compared various algorithms in terms of computational efficiency. For more details, see Section 2.2.

In this chapter, we develop a new methodology called *Pattern Recovery Using Trend Filtering* (PRUTF) for identifying unknown change points in piecewise polynomial signals with no continuity restriction at change point locations. Therefore, a change point is defined as a sudden jump in the signal and its all derivatives up to order r. Figure 3.2 displays such change points for various r. In this chapter, we make the following contributions.

- We propose a generic dual solution path algorithm along with the regularization parameter for trend filtering. This solution path, whose basic idea is borrowed from [154] enables us to determine change points at each level of the regularization parameter. Our algorithm, PRUTF, is different from that of [154] as we remove $(r+1)$ coordinates of dual variables after identifying each change point. This adjustment to the algorithm allows us to have independent dual variables between each pair of neighbouring change points. Besides, the elimination of $(r+1)$ coordinates at each step leads to faster implementation of the algorithm.

- We establish a stopping criterion that plays an essential role in the PRUTF algorithm used to find change points. Notably, we show that the dual variables of trend filtering between consecutive change points constitute a Gaussian bridge process. This finding allows us to introduce a threshold for terminating our proposed algorithm.

- If the signal contains a staircase pattern, we prove that the method is statistically inconsistent, making it unfavourable. Explaining the reason for this end, we modify the PRUTF algorithm to produce estimates consistent in terms of both the number and location of change points.

This chapter is organized as follows: we first describe how to characterize the dual optimization problem of trend filtering. In Section 3.4, we develop our main algorithm, PRUTF, to use in constructing the dual solution path of trend filtering and, in turn, identifying the locations of change points. Section 3.5 discusses the properties of this dual

solution path. We establish that the dual variables derived from the solution path form a Gaussian bridge process that makes them favourable for statistical inference. Applying these properties, we develop a stopping rule for the change point search algorithm in Section 3.6. The quality of the PRUTF algorithm is validated in terms of pattern recovery of the true signal in Section 3.7. It is established that the proposed technique in its naive form fails to consistently identify the true signal when a special pattern, called staircase, is present in the signal. Section 3.8 elaborates on how to modify PRUTF in order to consistently estimate the true pattern. Simulation results and real-world applications are presented in Section 3.9.

3.2 Notations

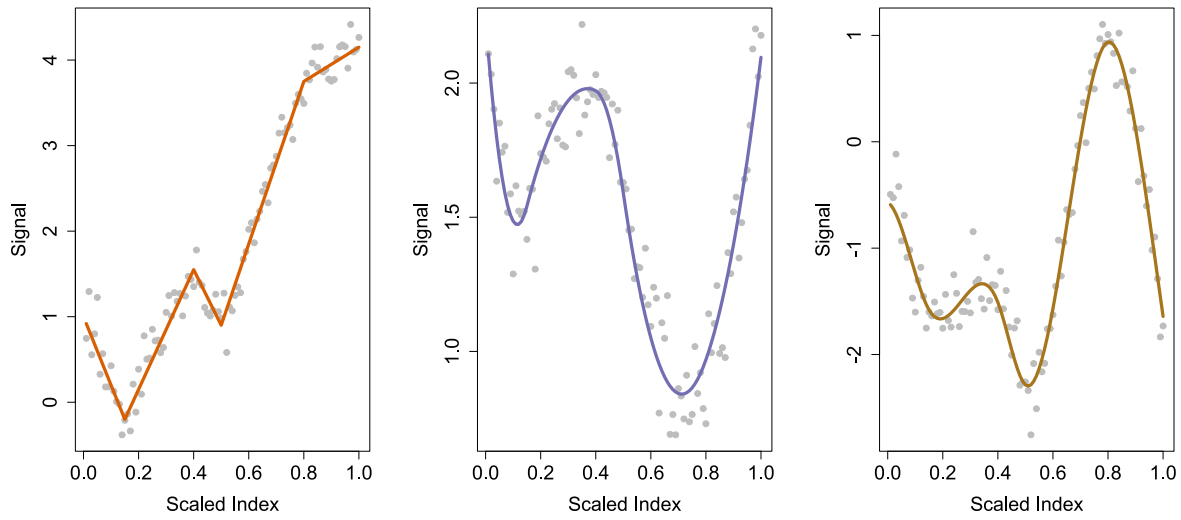
We begin this section with setting up notations that will be used throughout this thesis. For an $m \times n$ matrix $\mathbf{A}$, we denote its rows by $\mathbf{A}_1, \ldots, \mathbf{A}_m$ and express the matrix as $\mathbf{A} = (\mathbf{A}_1^T, \ldots, \mathbf{A}_m^T)^T$. Now for the set of indices $\mathcal{I} = \{i_1, \ldots, i_k\} \subseteq \{1, \ldots, m\}$, the notation $\mathbf{A}_{\mathcal{I}} = (\mathbf{A}_{i_1}^T, \ldots, \mathbf{A}_{i_k}^T)^T$ represents the submatrix of $\mathbf{A}$ whose row labels are in the set $\mathcal{I}$. In a similar manner, for a vector $\boldsymbol{a}$ of length m, we let $\boldsymbol{a}_{\mathcal{I}} = (a_{i_1}, \ldots, a_{i_k})^T$ denote a subvector of $\boldsymbol{a}$ whose coordinate labels are in $\mathcal{I}$. We write $\mathbf{A}_{-\mathcal{I}}$ and $\boldsymbol{a}_{-\mathcal{I}}$ to denote $\mathbf{A}_{\{1,\ldots,m\}\setminus\mathcal{I}}$ and $\boldsymbol{a}_{\{1,\ldots,m\}\setminus\mathcal{I}}$, respectively, where $\mathcal{J}\setminus\mathcal{I}$ is the set of indices in $\mathcal{J}$ but not in $\mathcal{I}$. Furthermore, for selecting i-th row of $\mathbf{A}$, the notation $[\mathbf{A}]_i$ and for its (i,j)-th element the notation $[\mathbf{A}]_{ij}$ are used. Also, $[\boldsymbol{a}]_i$ extracts the i-th elements of the vector $\boldsymbol{a}$. We write $\mathrm{diag}(\mathbf{A})$ to denote the vector of the main diagonal entries of the matrix $\mathbf{A}$. Moreover, for a real number x, $\lfloor x \rfloor$ denotes the greatest integer less than or equal x, and $\lceil x \rceil$ denotes the least integer greater or equal x. For a set B, the indicator function is denoted by $\mathbb{1}(B)$.

3.3 Dual Problem of Trend Filtering

Recall the trend filtering problem

$$\min_{\mathbf{f}\in\mathbb{R}^n} \frac{1}{2} \left\| \mathbf{y} - \mathbf{f} \right\|_2^2 + \lambda \left\| \mathbf{D}^{(r+1)} \mathbf{f} \right\|_1, \tag{3.3}$$

where $\lambda \geq 0$ is the regularization parameter for controlling the effect of smoothing, and the $(n-r-1)\times n$ penalty matrix $\mathbf{D}^{(r+1)}$ is the difference operator of order $(r+1)$. For



(a) Linear **(b)** Quadratic **(c)** Cubic

Figure 3.1: Trend filtering solutions for $r = 1, 2, 3$ producing (a) piecewise linear, (b) piecewise quadratic and (c) piecewise cubic fits, respectively.

$r = 0$, the first order difference matrix $\mathbf{D}^{(1)}$ is defined as

$$\mathbf{D}^{(1)} = \begin{pmatrix} -1 & 1 & 0 & 0 & \dots & 0 \\ 0 & -1 & 1 & 0 & \dots & 0 \\ 0 & 0 & -1 & 1 & \dots & 0 \\ \vdots & & & & & \vdots \end{pmatrix}, \tag{3.4}$$

and for $r \geq 1$, the difference operator of order $r + 1$ can be recursively computed by $\mathbf{D}^{(r+1)} = \mathbf{D}^{(1)} \times \mathbf{D}^{(r)}$. Notice that, in this matrix multiplication, $\mathbf{D}^{(1)}$ is the submatrix consisting of the first $n - r - 1$ rows and $n - r$ columns of the matrix in (3.4). Figure 3.1 displays the trend filtering fits for $r = 1, 2, 3$ for simulated data.

Although the objective function in the r-th order trend filtering (3.3) is strictly convex and thus the minimization has a guaranteed unique solution, the penalty term is not differentiable in $\mathbf{f}$, so solving the optimization in its current form is difficult. To overcome this difficulty, we follow the argument in [154] and convert this optimization problem into its dual form. Since the objective function in the primal problem is strictly convex with no constraint, the strong duality holds, meaning that the primal and the dual solutions coincide [26].

The trend filtering problem (3.3) can be rewritten as

$$\min_{\mathbf{f}\in\mathbb{R}^n} \ \frac{1}{2}\left\| \mathbf{y}-\mathbf{f} \right\|_2^2 \ + \ \lambda \left\| \mathbf{z} \right\|_1, \qquad\qquad \text{subject to} \qquad \mathbf{z}=\mathbf{D}\mathbf{f}\,,$$

where, for ease in the notation, we use $\mathbf{D}=\mathbf{D}^{(r+1)}$. For any given $\lambda>0$, the Lagrangian is

$$\mathcal{L}(\mathbf{f},\,\mathbf{z},\,\mathbf{u}) = \frac{1}{2}\left\|\mathbf{y}-\mathbf{f}\right\|_2^2 \ + \ \lambda \left\| \mathbf{z} \right\|_1 \ + \ \mathbf{u}^T\big(\mathbf{D}\mathbf{f}-\mathbf{z}\big)$$

and, thus the dual function is given by

$$g(\mathbf{u}) = \inf_{\mathbf{f}\in\mathbb{R}^n,\,\mathbf{z}\in\mathbb{R}^m} \ \mathcal{L}(\mathbf{f},\,\mathbf{z},\,\mathbf{u}),$$

which is a concave function defined on $\mathbb{R}^m$, where $m=n-r-1$ and takes values in the extended real line $\mathbb{R}\cup\{-\infty,\,\infty\}$. The vectors $\mathbf{f}$ and $\mathbf{u}$ are called the primal and dual variables, respectively. Taking the derivative of the Lagrangian $\mathcal{L}(\mathbf{f},\,\mathbf{z},\,\mathbf{u})$ with respect to $\mathbf{f}$ and setting it to be equal to zero, we obtain

$$\mathbf{f}=\mathbf{y}-\mathbf{D}^T\mathbf{u}. \tag{3.5}$$

Now substituting this back into the Lagrangian $\mathcal{L}(\mathbf{f},\,\mathbf{z},\,\mathbf{u})$, and performing certain algebraic manipulations, we obtain

$$\begin{aligned}\mathcal{L}^*(\mathbf{z},\,\mathbf{u}) = & \inf_{\mathbf{f}\in\mathbb{R}^n} \mathcal{L}(\mathbf{f},\,\mathbf{z},\,\mathbf{u}) \\ = & -\frac{1}{2}\left\|\mathbf{y}-\mathbf{D}^T\mathbf{u}\right\|_2^2+\frac{1}{2}\left\|\mathbf{y}\right\|_2^2+\lambda\left\|\mathbf{z}\right\|_1-\mathbf{u}^T\mathbf{z}\,.\end{aligned}$$

Minimizing $\mathcal{L}^*(\mathbf{z},\,\mathbf{u})$, or equivalently maximizing $\mathbf{u}^T\mathbf{z}-\lambda\|\mathbf{z}\|_1$, with respect to $\mathbf{z}\in\mathbb{R}^m$ leads us to the dual function $g(\mathbf{u})$. Notice that $\sup_{\mathbf{z}}\left\{\mathbf{u}^T\mathbf{z}-\lambda\left\|\mathbf{z}\right\|_1\right\}$ is the conjugate of the function $\lambda\left\|\mathbf{z}\right\|_1$ in the context of conjugate convex functions. See [27] and [26]. This conjugate function is given by

$$\sup_{\mathbf{z}}\left\{\mathbf{u}^T\mathbf{z}-\lambda\left\|\mathbf{z}\right\|_1\right\} = \begin{cases} 0 & \text{if } \left\|\mathbf{u}\right\|_\infty \le \lambda \\ \infty & \text{otherwise}\,. \end{cases}$$

From all these, the dual function is given as

$$g(\mathbf{u}) = -\frac{1}{2}\left\|\mathbf{y}-\mathbf{D}^T\mathbf{u}\right\|_2^2+\frac{1}{2}\left\|\mathbf{y}\right\|_2^2 \qquad \text{for} \quad \left\|\mathbf{u}\right\|_\infty \le \lambda\,,$$

and, thus the dual problem is to find the maximum of the dual function $g(\mathbf{u})$. This is equivalent to

$$\min_{\mathbf{u}\in\mathbb{R}^m} \frac{1}{2}\left\| \mathbf{y} - \mathbf{D}^T\mathbf{u} \right\|_2^2 \qquad \text{subject to} \qquad \|\mathbf{u}\|_\infty \leq \lambda\,. \tag{3.6}$$

The constraint in (3.6) is an ℓ_∞-ball or a hypercube centered at the origin with the boundaries given by the set $\{-\lambda,\, \lambda\}^m$. Since the matrix $\mathbf{D}$ is full row rank, the problem (3.6) is strictly convex and has a unique solution, see [150] and [2]. In addition, notice that the dimension of the dual vector $\mathbf{u}$ is m, which is smaller than that of the primal vector $\mathbf{f}$ and may lead to relatively faster computations. The connection between the primal and the dual solutions is given by the equations

$$\widehat{\mathbf{u}}_\lambda = \lambda\,\widehat{\boldsymbol{\gamma}}, \tag{3.7}$$

$$\widehat{\mathbf{f}}_\lambda = \mathbf{y} - \mathbf{D}^T\widehat{\mathbf{u}}_\lambda\,, \tag{3.8}$$

where $\widehat{\boldsymbol{\gamma}} \in \mathbb{R}^m$ is a subgradient of $\|\mathbf{x}\|_1$ computed at $\mathbf{x} = \mathbf{D}\widehat{\mathbf{f}}_\lambda$. This subgradient is given by

$$\widehat{\gamma}_i \in \begin{cases} \{+1\} & \text{if} \quad \left[\mathbf{D}\widehat{\mathbf{f}}_\lambda\right]_i > 0 \\ \{-1\} & \text{if} \quad \left[\mathbf{D}\widehat{\mathbf{f}}_\lambda\right]_i < 0 \\ [-1\,,\,+1] & \text{if} \quad \left[\mathbf{D}\widehat{\mathbf{f}}_\lambda\right]_i = 0\,. \end{cases} \tag{3.9}$$

The statements in Equations (3.7)–(3.9) are equivalent to the KKT optimality conditions of the primal problem (3.3). The dual problem (3.6) demonstrates that $\mathbf{D}^T\widehat{\mathbf{u}}_\lambda$ is the projection, $P_{\mathbb{C}}(\mathbf{y})$, of $\mathbf{y}$ onto the convex polyhedron $\mathbb{C} = \left\{ \mathbf{x} \in \mathbb{R}^m \,:\, \|\mathbf{x}\|_\infty \leq \lambda \right\}$. From this, the primal solution (3.8) can be rewritten in the form of $\widehat{\mathbf{f}}_\lambda = \left(\mathbf{I} - P_{\mathbb{C}}\right)(\mathbf{y})$, representing the residual projection map of $\mathbf{y}$ onto the polyhedron $\mathbb{C}$.

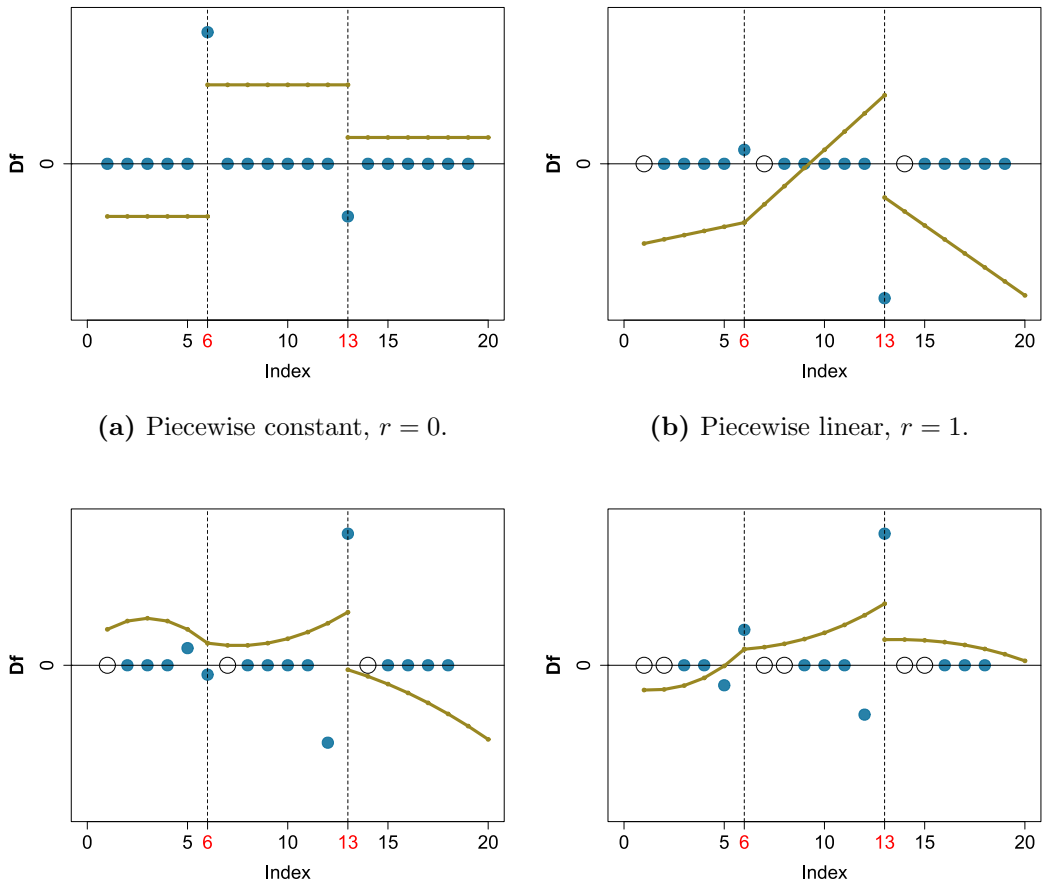
Our idea of applying trend filtering to discover change points in piecewise polynomial signals is inspired by [129] and its correction [128], in which change point detection is studied using fused lasso. Besides extending to piecewise polynomial signals, the novelty of our work is in providing an exact stopping criterion, which is based on the Gaussian bridge property of the trend filtering dual variables. In addition, we propose an algorithm which, unlike that proposed in [129], always produces consistent change points even in the presence of staircase patterns.

3.4 Solution Path of Trend Filtering and PRUTF Algorithm

In this section, we construct and study the solution path of dual variables $\widehat{\mathbf{u}}_\lambda$ as the regularization parameter decreases from $\lambda = \infty$ to $\lambda = 0$. In the following, the PRUTF algorithm is given to compute the entire dual solution path. This dual solution path identifies the corresponding primal solution using (3.8). For any given λ, we call any coordinate of $\widehat{\mathbf{u}}_\lambda$ a boundary coordinate if it is a vertex of the polyhedron $\mathbb{C} = \left\{\mathbf{x} \in \mathbb{R}^m : \|\mathbf{x}\|_\infty \leq \lambda\right\}$, meaning that its absolute value becomes λ. In the process of constructing the solution path, for any λ, we trace several sets, introduced below.

- The set $\mathcal{B} = \mathcal{B}(\lambda)$, called the boundary set, contains the boundary coordinates identified by $\widehat{\mathbf{u}}_\lambda$.
- The vector $\mathbf{s}_\mathcal{B} = \mathbf{s}_\mathcal{B}(\lambda)$, called the sign vector, represents collectively the signs of the boundary points in $\mathcal{B}(\lambda)$.
- The set $\mathcal{A} = \mathcal{A}(\lambda)$, called the augmented boundary set, contains the boundary coordinates in $\mathcal{B}(\lambda)$ as well as the first $r_a = \lfloor (r+1)/2 \rfloor$ coordinates immediately after.
- The vector $\mathbf{s}_\mathcal{A} = \mathbf{s}_\mathcal{A}(\lambda)$ represents collectively the signs of the augmented boundary points in $\mathcal{A}(\lambda)$.

In the following, we discuss the need for the augmented boundary set $\mathcal{A}$. We begin by studying the structure of the dual vector $\mathbf{u} = \mathbf{Df}$ in a piecewise polynomial signal of order r, where the signal is partitioned into a number of blocks defined by the position of the change points. Because the signal f is a piecewise polynomial of order r, to compute the i-th coordinate of the vector $\mathbf{u}$, we need $r_b = \lceil (r+1)/2 \rceil - 1$ points directly before the i-th element of $\mathbf{f}$ as well as $r_a = \lfloor (r+1)/2 \rfloor$ points immediately after that. Consequently, the first r_a elements of $\mathbf{Df}$ within each block cannot be computed. Moreover, within each block, the last $r_b + 1$ elements of $\mathbf{Df}$ are all nonzero due to the existence of a change point. This observation is depicted in Figure 3.2 for $r = 0, 1, 2, 3$. To explain this point clearly, consider the case of $r = 2$ in Figure 3.2 in which the structure of $\mathbf{Df}$ is shown, where the true change points are at 6 and 13. As can be seen, the points on the boundary – the nonzero coordinates of $\mathbf{Df}$ – are $\mathcal{B}(\lambda) = \{5, 6, 12, 13\}$ with their respective signs $\mathbf{s}_\mathcal{B}(\lambda) = \{1, 1, -1, -1\}$. Notice that $\mathbf{Df}$ does not exist at points 7 and 14. The augmented boundary set contains these points as well as the boundary points; that is $\mathcal{A}(\lambda) = \{5, 6, 7, 12, 13, 14\}$. The respective signs of the coordinates in the augmented



(a) Piecewise constant, $r = 0$.

(b) Piecewise linear, $r = 1$.

(c) Piecewise quadratic, $r = 2$.

(d) Piecewise cubic, $r = 3$.

Figure 3.2: Structure of **Df** for piecewise polynomial signals with various orders $r = 0, 1, 2, 3$. The olive lines display the true signals with two change points at the locations 6 and 13. Empty circles represent the indices at which **Df** does not exist.

boundary set $\mathcal{A}(\lambda)$ are given by $\mathbf{s}_{\mathcal{A}}(\lambda) = \{1, 1, 1, -1, -1, -1\}$. At each value of λ, we call the coordinates that belong to the augmented boundary set $\mathcal{A}(\lambda)$ the augmented boundary coordinates, and the rest, the interior coordinates.

At the j-th iteration with $\lambda = \lambda_j$, we assume that the boundary set and its corresponding sign vector are $\mathcal{B} = \mathcal{B}(\lambda)$ and $\mathbf{s}_{\mathcal{B}} = \mathbf{s}_{\mathcal{B}}(\lambda)$, respectively. Furthermore, we assume the

augmented boundary set and its sign vector are $\mathcal{A} = \mathcal{A}(\lambda)$ and $\mathbf{s}_{\mathcal{A}} = \mathbf{s}_{\mathcal{A}}(\lambda)$, respectively. Dual coordinates can be split into augmented boundary coordinates $\widehat{\mathbf{u}}_{\lambda_j, \mathcal{A}}$ and interior coordinates $\widehat{\mathbf{u}}_{\lambda_j, -\mathcal{A}}$. Recall from Section 3.2 that $\widehat{\mathbf{u}}_{\lambda_j, \mathcal{A}}$ represents the subvector of $\widehat{\mathbf{u}}_{\lambda_j}$ with the coordinate labels in the set $\mathcal{A}$ and $\widehat{\mathbf{u}}_{\lambda_j, -\mathcal{A}}$ represents the subvector of $\widehat{\mathbf{u}}_{\lambda_j}$ with the coordinate labels in the set $\{1, 2, \cdots, m\} \backslash \mathcal{A}$. It is apparent from the definition of the boundary coordinates that

$$\widehat{\mathbf{u}}_{\lambda_j, \mathcal{A}} = \lambda_j \, \mathbf{s}_{\mathcal{A}} \, . \tag{3.10}$$

Replacing the boundary coordinate with $\lambda_j \, \mathbf{s}_{\mathcal{A}}$ in (3.6) and solving the resulting quadratic problem with respect to the interior coordinates, lead to their least square estimates, given by

$$\widehat{\mathbf{u}}_{\lambda_j, -\mathcal{A}} = \Big(\mathbf{D}_{-\mathcal{A}} \mathbf{D}_{-\mathcal{A}}^T \Big)^{-1} \mathbf{D}_{-\mathcal{A}} \Big(\mathbf{y} - \lambda_j \mathbf{D}_{\mathcal{A}}^T \, \mathbf{s}_{\mathcal{A}} \Big). \tag{3.11}$$

It should be noted that for the purpose of simplicity, we denote $(\mathbf{D}_{\mathcal{A}})^T$ and $(\mathbf{D}_{-\mathcal{A}})^T$ with $\mathbf{D}_{\mathcal{A}}^T$ and $\mathbf{D}_{-\mathcal{A}}^T$, respectively. Notice that in (3.11), the first term $\big(\mathbf{D}_{-\mathcal{A}} \mathbf{D}_{-\mathcal{A}}^T \big)^{-1} \mathbf{D}_{-\mathcal{A}} \, \mathbf{y}$ simply yields the least square estimate of regressing the response vector $\mathbf{y}$ on the design matrix $\mathbf{D}_{-\mathcal{A}}$. The second term $-\lambda_j \big(\mathbf{D}_{-\mathcal{A}} \mathbf{D}_{-\mathcal{A}}^T \big)^{-1} \mathbf{D}_{-\mathcal{A}} \, \mathbf{D}_{\mathcal{A}}^T \, \mathbf{s}_{\mathcal{A}}$ can be interpreted as a shrinkage term due to the condition $\| \mathbf{u} \|_{\infty} \leq \lambda$. The expression (3.11) is true for $\lambda \leq \lambda_j$ until either an interior coordinate joins the boundary or a coordinate in the boundary set leaves the boundary. The following argument explains how to specify values of λ while the interior coordinates change.

We define the joining time associated with the interior coordinate $i \in \{1, 2, \cdots, m\} \backslash \mathcal{A}$ as the time at which this interior coordinate joins the boundary. To determine the next joining time, we reduce the value of λ in a linear direction starting from λ_j and solve $\widehat{\mathbf{u}}_{\lambda, -\mathcal{A}} = (\pm\lambda, \cdots, \pm\lambda)^T$. Note that the right-hand side of (3.11) can be expressed as $\boldsymbol{a} - \lambda_j \, \mathbf{b}$, where

$$\begin{aligned} \boldsymbol{a} &= \big(\mathbf{D}_{-\mathcal{A}} \mathbf{D}_{-\mathcal{A}}^T \big)^{-1} \mathbf{D}_{-\mathcal{A}} \, \mathbf{y} \, , \\ \mathbf{b} &= \big(\mathbf{D}_{-\mathcal{A}} \mathbf{D}_{-\mathcal{A}}^T \big)^{-1} \mathbf{D}_{-\mathcal{A}} \mathbf{D}_{\mathcal{A}}^T \, \mathbf{s}_{\mathcal{A}} \, . \end{aligned} \tag{3.12}$$

The joining time for every $i \in \{1, 2, \cdots, m\} \backslash \mathcal{A}$ is hence the solution of the equation $a_i - \lambda \, b_i = \pm\lambda$ with respect to λ, which is given by

$$\lambda_i^{\text{join}} = \frac{a_i}{b_i \pm 1} \, , \qquad\qquad i \in \{1, 2, \cdots, m\} \backslash \mathcal{A} \, .$$

Note that λ_i^{join} is uniquely defined because only one of the signs -1 or $+1$ yields $\lambda_i \in [0, \lambda_j]$.

Now we turn the attention to the characterization of a coordinate that leaves the boundary set $\mathcal{B}$. For $i \in \mathcal{B}$, the leaving time is defined as the time that the coordinate i leaves the boundary set $\mathcal{B}$. Since $\mathbf{s}_{\mathcal{B}}$ is the sign vector of changes captured by $\left[\mathbf{D}\,\widehat{\mathbf{f}}\right]_{\mathcal{B}}$, then $\operatorname{diag}(\mathbf{s}_{\mathcal{B}})\left[\mathbf{D}\,\widehat{\mathbf{f}}\right]_{\mathcal{B}} > \mathbf{0}$, which in turn, along with Equation (3.8), implies $\operatorname{diag}(\mathbf{s}_{\mathcal{B}})\big[\mathbf{D}\left(\mathbf{y} - \mathbf{D}^T\widehat{\mathbf{u}}_\lambda\right)\big]_{\mathcal{B}} > \mathbf{0}$. Here, for any vector $\boldsymbol{\eta}$, $\operatorname{diag}(\boldsymbol{\eta})$ denotes the diagonal matrix with the diagonal elements given by $\boldsymbol{\eta}$, and $\boldsymbol{\eta} > \mathbf{0}$ holds element-wise. Therefore, a coordinate $i \in \mathcal{B}$ leaves the boundary set $\mathcal{B}$ if $\operatorname{diag}(\mathbf{s}_{\mathcal{B}})\big[\mathbf{D}\left(\mathbf{y} - \mathbf{D}^T\widehat{\mathbf{u}}_\lambda\right)\big]_{\mathcal{B}} > \mathbf{0}$ is violated. Using the relation

$$\left[\mathbf{D}\left(\mathbf{y} - \mathbf{D}^T\widehat{\mathbf{u}}_\lambda\right)\right]_{\mathcal{B}} = \mathbf{D}_{\mathcal{B}}\left(\mathbf{y} - \mathbf{D}^T\widehat{\mathbf{u}}_\lambda\right),$$

and the decomposition $\mathbf{D}^T\widehat{\mathbf{u}}_\lambda = \mathbf{D}_{\mathcal{A}}^T\,\widehat{\mathbf{u}}_{\lambda,\mathcal{A}} + \mathbf{D}_{-\mathcal{A}}^T\,\widehat{\mathbf{u}}_{\lambda,-\mathcal{A}}$, we obtain

$$\operatorname{diag}(\mathbf{s}_{\mathcal{B}})\left[\mathbf{D}\left(\mathbf{y} - \mathbf{D}^T\widehat{\mathbf{u}}_\lambda\right)\right]_{\mathcal{B}} = \mathbf{c} - \lambda\,\mathbf{d}\,, \tag{3.13}$$

where

$$\begin{aligned}\mathbf{c} &= \operatorname{diag}(\mathbf{s}_{\mathcal{B}})\,\mathbf{D}_{\mathcal{B}}\big(\mathbf{y} - \mathbf{D}_{-\mathcal{A}}^T\,\boldsymbol{a}\big)\,,\\ \mathbf{d} &= \operatorname{diag}(\mathbf{s}_{\mathcal{B}})\,\mathbf{D}_{\mathcal{B}}\big(\mathbf{D}_{\mathcal{A}}^T\,\mathbf{s}_{\mathcal{A}} - \mathbf{D}_{-\mathcal{A}}^T\,\mathbf{b}\big).\end{aligned} \tag{3.14}$$

Hence, a leaving time is obtained from the equation $c_i - \lambda\,d_i > 0$ as

$$\lambda_i^{\text{leave}} = \begin{cases} \dfrac{c_i}{d_i}, & \text{if } c_i < 0 \;\text{ and }\; d_i < 0\,,\\ 0, & \text{otherwise}\,. \end{cases}$$

The conditions in the aforementioned equation is due to the fact that at the j-th iteration with $\lambda \le \lambda_j$, the expression $c_i - \lambda\,d_i > 0$ fails for $i \in \mathcal{B}$, if both c_i and d_i are negative. An alternative way to determine the next leaving time is to use the KKT optimality conditions of (3.6). We refer the reader to the supplementary materials of [154].

The following algorithm, PRUTF, describes the process of constructing the entire dual solution path of trend filtering.

Algorithm 3.1 (PRUTF Algorithm)

1. *Initialize the set of change points locations as* $\boldsymbol{\tau}_0 = \emptyset$*, the empty set.*

2. *At step $j = 1$, initialize the boundary set $\mathcal{B}_1 = \{\tau_1 - r_b, \tau_1 - r_b + 1, \ldots, \tau_1\}$ and its associated sign vector $\mathbf{s}_{\mathcal{B}_1} = \{s_1, \ldots, s_1\}$, both with cardinality of $r_b + 1$, where τ_1 is obtained by*

$$\tau_1 = \underset{i=1,\ldots,m}{\operatorname{argmax}} \mid \widehat{u}_i \mid, \tag{3.15}$$

and $s_1 = \operatorname{sign}(\widehat{u}_{\tau_1})$, where $\widehat{u}_i$ is the i-th element of the vector $\widehat{\mathbf{u}} = \left(\mathbf{D}\mathbf{D}^T\right)^{-1}\mathbf{D}\,\mathbf{y}$. The updated set of change points locations is now $\boldsymbol{\tau}_1 = \{\tau_1\}$. We also record the first joining time $\lambda_1 = \mid \widehat{u}_{\tau_1} \mid$ and keep track of the augmented boundary set $\mathcal{A}_1 = \{\tau_1 - r_b, \ldots, \tau_1 + r_a\}$ and its corresponding sign vector $\mathbf{s}_{\mathcal{A}_1} = \{s_1, \ldots, s_1\}$ of length $r + 1$. The dual solution is regarded as $\widehat{\mathbf{u}}(\lambda) = \left(\mathbf{D}\mathbf{D}^T\right)^{-1}\mathbf{D}\,\mathbf{y}$, for $\lambda \geq \lambda_1$.

3. *For step $j = 2, 3, \ldots,$*

   (a) *Obtain the pair $\left(\tau_j^{\text{join}}, s_j^{\text{join}}\right)$ from*

$$\left(\tau_j^{\text{join}}, s_j^{\text{join}}\right) = \underset{i \notin \mathcal{A}_{j-1},\, s \in \{-1,1\}}{\operatorname{argmax}} \ \frac{a_i}{s + b_i} \cdot \mathbb{1}\left\{0 \leq \frac{a_i}{s + b_i} \leq \lambda_{j-1}\right\}, \tag{3.16}$$

   and set the next joining time λ_j^{join} as the value of $\frac{a_i}{s+b_i}$, for $i = \tau_j^{\text{join}}$ and $s = s_j^{\text{join}}$.

   (b) *Obtain the pair $\left(\tau_j^{\text{leave}}, s_j^{\text{leave}}\right)$ from*

$$\left(\tau_j^{\text{leave}}, s_j^{\text{leave}}\right) = \underset{i \in \mathcal{B}_{j-1},\, s \in \{-1,1\}}{\operatorname{argmax}} \ \frac{c_i}{d_i} \cdot \mathbb{1}\Big\{c_i < 0 \ , \ d_i < 0\Big\}, \tag{3.17}$$

   and assign the next leaving time λ_j^{leave} as the value of $\dfrac{c_i}{d_i}$, for $i = \tau_j^{\text{leave}}$ and $s = s_j^{\text{leave}}$.

   (c) *Let $\lambda_j = \max\left\{\lambda_j^{\text{join}}, \lambda_j^{\text{leave}}\right\}$, then the boundary set $\mathcal{B}_j$ and its sign vector $\mathbf{s}_{\mathcal{B}_j}$ are updated in the following fashion:*

   - *Either append $\left\{\tau_j^{\text{join}} - r_b, \tau_j^{\text{join}} - r_b + 1, \ldots, \tau_j^{\text{join}}\right\}$ and the corresponding signs $\left\{s_j^{\text{join}}, \ldots, s_j^{\text{join}}\right\}$ to $\mathcal{B}_{j-1}$ and $\mathbf{s}_{\mathcal{B}_{j-1}}$, respectively, provided that $\lambda_j = \lambda_j^{\text{join}}$. Also, add τ_j^{join} to $\boldsymbol{\tau}_{j-1}$.*
   - *Or remove $\left\{\tau_j^{\text{leave}}, \tau_j^{\text{leave}} + 1, \ldots, \tau_j^{\text{leave}} + r_b\right\}$ and the corresponding signs $\left\{s_j^{\text{leave}}, \ldots, s_j^{\text{leave}}\right\}$ from $\mathcal{B}_{j-1}$ and $\mathbf{s}_{\mathcal{B}_{j-1}}$, respectively, provided that $\lambda_j = \lambda_j^{\text{leave}}$. Also, remove τ_j^{leave} from $\boldsymbol{\tau}_{j-1}$.*

In the same manner, the augmented boundary set, $\mathcal{A}_j$ and its sign, $\mathbf{s}_{\mathcal{A}_j}$ are formed by adding $\{\tau_j^{\text{join}} - r_b, \ldots, \tau_j^{\text{join}} + r_a\}$ and $\{s_j^{\text{join}}, \ldots, s_j^{\text{join}}\}$ to $\mathcal{A}_{j-1}$ and $\mathbf{s}_{\mathcal{A}_{j-1}}$, respectively, if $\lambda_j = \lambda_j^{\text{leave}}$ or, otherwise, by removing the associated set $\{\tau_j^{\text{leave}}, \ldots, \tau_j^{\text{leave}} + r\}$ and $\{s_j^{\text{leave}}, \ldots, s_j^{\text{leave}}\}$ from $\mathcal{A}_{j-1}$ and $\mathbf{s}_{\mathcal{A}_{j-1}}$. Thus, the dual solution is computed as $\widehat{\mathbf{u}}_{\mathcal{A}_j}(\lambda) = \boldsymbol{a} - \lambda\,\mathbf{b}$ for interior coordinates and $\widehat{\mathbf{u}}_{-\mathcal{A}_j}(\lambda) = \lambda\,\mathbf{s}_{\mathcal{A}_j}$ for boundary coordinates over $\lambda_j \leq \lambda \leq \lambda_{j-1}$.

4. Repeat step 3 until $\lambda > 0$.

The critical points $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$ indicate the values of the regularization parameter at which the boundary set changes.

Remark 3.2 *Notice that the vector $\boldsymbol{\tau}$ derived by the PRUTF algorithm represents the locations of change points for the dual variables. In order to obtain the locations of change points in primal variables, we must add r_a to any element of $\boldsymbol{\tau}$, that is, $\{\tau_1 + r_a, \tau_2 + r_a, \ldots\}$. This relationship between the primal and dual change point sets is visible from Figure 3.2.*

Remark 3.3 *For fused lasso, $r = 0$, Lemma 1 of [154], known as the boundary lemma, is satisfied since the matrix $\mathbf{D}\mathbf{D}^T$ is diagonally dominant, meaning that $\left|\left[\mathbf{D}\mathbf{D}^T\right]_{i,i}\right| \geq \sum_{j\neq i}\left|\left[\mathbf{D}\mathbf{D}^T\right]_{i,j}\right|$, for $i = 1, \ldots, m$. This lemma states that when a coordinate joins the boundary, it will stay on the boundary for the rest of the path. Consequently, part (b) of step 3 in Algorithm 3.1 is unnecessary, and hence the next leaving time in part (c) is set to zero, i.e., $\lambda_j^{\text{leave}} = 0$, for every step j. However, the boundary lemma is not satisfied for $r = 1, 2, 3, \ldots$.*

Remark 3.4 *There is a subtle and important distinction between our proposed algorithm, PRUTF, and the one presented in [154]. The latter work studies the generalized lasso problem for any arbitrary penalty matrix $\mathbf{D}$ (unlike $\mathbf{D}$ used in trend filtering, which must have a certain structure). The proposed algorithm in [154] relies on adding or removing only one coordinate to or from the boundary set at every step. The key attribute of our algorithm is to add or remove $r+1$ coordinates to or from the augmented boundary set, an approach inspired by the argument presented at the beginning of this section. Essentially, this attribute makes PRUTF, presented in Algorithm 3.1, well-suited for change point analysis. It is important to mention that PRUTF requires at least $r+1$ data points between neighbouring change points.*

Remark 3.5 *For a given λ, equations (3.10) and (3.11) give the values of the dual variables in $\widehat{\mathbf{u}}_\lambda$. The equations demonstrate that the dual solution path is a linear function of λ with change in the slopes at joining or leaving times $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$.*

Remark 3.6 *The number of iterations required for PRUTF, presented in Algorithm 3.1, is at most $(3^p+1)/2$, where $p = \lceil \frac{m}{r+1} \rceil$, see [150], Lemma 6. However, this upper bound for the number of iterations is usually very loose. The upper bound comes from the following realization discovered by [115] and later improved by [106]. Any pair $(\mathcal{A}, \mathbf{s}_\mathcal{A})$ appears at most once throughout the solution path. In other words, if $(\mathcal{A}, \mathbf{s}_\mathcal{A})$ is visited in one iteration of the algorithm, the pair $(\mathcal{A}, -\mathbf{s}_\mathcal{A})$ as well as $(\mathcal{A}, \mathbf{s}_\mathcal{A})$ cannot reappear again for the rest of the algorithm. Interestingly, this fact says that once a coordinate enters the boundary set, it cannot immediately leave the boundary set at the next step.*

Moreover, note that at one iteration of the PRUTF algorithm with the augmented boundary set $\mathcal{A}$, the dominant computation is in solving the least square problem

$$\min_{\mathbf{u} \in \mathbb{R}^m} \ \frac{1}{2} \left\| \mathbf{y} - \mathbf{D}_\mathcal{A}^T \mathbf{u} \right\|_2^2 . \tag{3.18}$$

One can apply QR decomposition of $\mathbf{D}_\mathcal{A}^T$ to solve the least square problem, and then update the decomposition as set $\mathcal{A}$ changes. However, since $\mathbf{D}_{-\mathcal{A}}\mathbf{D}_{-\mathcal{A}}^T$ is a banded Toeplitz matrix (see Section 3.5), a solution of (3.18) always exists and can be computed using a banded Cholesky decomposition. Thus, the computational complexity for the iteration is of order $O\big((m - |\mathcal{A}|)\, r^2\big)$, which is linear in the number of interior coordinates as r is fixed and usually small. Overall, if K is the total number of steps run by the PRUTF algorithm, then the total computational complexity is $O\big(K(m - |\mathcal{A}|)\, r^2\big)$. See [154] and [6].

3.5 Statistical Properties of the Solution Path

An important component of the methodology that we develop in this work involves computing algebraic expressions based on the matrix $\mathbf{D} = \mathbf{D}^{(r+1)}$. In this section, we describe the properties of such expressions. To begin with, let $\mathcal{A} = \{A_1, \ldots, A_J\}$ and $\mathbf{s}_\mathcal{A} = \{\mathbf{s}_1, \ldots, \mathbf{s}_J\}$ be the augmented boundary set and its corresponding sign vector, respectively, after a number of iterations of Algorithm 3.1, where $A_j = \{\tau_j - r_b, \tau_j - r_b + 1, \ldots, \tau_j + r_a\}$ and $\mathbf{s}_j = \{s_j, \ldots, s_j\}$ for $j = 1, \ldots, J$. This augmented boundary set corresponds to J change points $\{\tau_1, \ldots, \tau_J\}$ that partition all the dual variables into $J+1$ blocks $B_j = \{\tau_j + 1, \ldots, \tau_{j+1}\}$ for $j = 0, 1, \ldots, J$, with the conventions that $\tau_0 = 0$ and $\tau_{J+1} = m$. In the following, we list some properties of matrix multiplications involving $\mathbf{D}$.

$$\begin{pmatrix}
6 & -4 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
-4 & 6 & -4 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & -4 & 6 & -4 & 1 & 0 & 0 & 0 & 0 \\
0 & 1 & -4 & 6 & -4 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & -4 & 6 & -4 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & -4 & 6 & -4 & 1 & 0 \\
0 & 0 & 0 & 0 & 1 & -4 & 6 & -4 & 1 \\
0 & 0 & 0 & 0 & 0 & 1 & -4 & 6 & -4 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & -4 & 6
\end{pmatrix}$$

(a) Structure of $\mathbf{D}\mathbf{D}^T$.

$$\left(\begin{array}{ccc|cccc}
6 & -4 & 1 & 0 & 0 & 0 & 0 \\
-4 & 6 & -4 & 0 & 0 & 0 & 0 \\
1 & -4 & 6 & 0 & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & 6 & -4 & 1 & 0 \\
0 & 0 & 0 & -4 & 6 & -4 & 1 \\
0 & 0 & 0 & 1 & -4 & 6 & -4 \\
0 & 0 & 0 & 0 & 1 & -4 & 6
\end{array}\right)$$

(b) The structure of $\mathbf{D}_{-\mathcal{A}}\mathbf{D}_{-\mathcal{A}}^T$.

Figure 3.3: Structure of quadratic forms of matrix $\mathbf{D}$.

- It follows from the definition of the matrix $\mathbf{D}$ that it is a banded Toeplitz matrix with bandwidth $r+1$. It tuns out that the matrix $\mathbf{D}\mathbf{D}^T$ reveals the same property, meaning that it is a square banded Toeplitz matrix. Moreover, its $r+1$ nonzero row elements are consecutive binomial coefficients of order $2\,r+2$ with alternating signs. In other words, $(i\,,\,j)$-th element of $\mathbf{D}\mathbf{D}^T$ for $i \geq j$ is $(-1)^{i-j}\binom{2\,r+2}{r+1+i-j}$. An example, for $r=1$, is given in panel (a) of Figure 3.3. Note that $\mathbf{D}\mathbf{D}^T$ is a symmetric, nonsingular and positive definite matrix [39].

- The matrix $\mathbf{D}_{-\mathcal{A}}\mathbf{D}_{-\mathcal{A}}^T$ is a block diagonal matrix whose diagonal submatrices correspond to $J+1$ blocks. More precisely, the j-th submatrix on the diagonal of $\mathbf{D}_{-\mathcal{A}}\mathbf{D}_{-\mathcal{A}}^T$ is a matrix with the first $(\tau_{j+1} - \tau_j - r)$ rows and columns of $\mathbf{D}\mathbf{D}^T$, see panel (b) of Figure 3.3. Notice that, due to its non-singularity, $\mathbf{D}_{-\mathcal{A}}\mathbf{D}_{-\mathcal{A}}^T$ is always invertible. In fact, both $\left(\mathbf{D}_{-\mathcal{A}}\mathbf{D}_{-\mathcal{A}}^T\right)^{-1}$ and $\left(\mathbf{D}_{-\mathcal{A}}\mathbf{D}_{-\mathcal{A}}^T\right)^{-1}\mathbf{D}_{-\mathcal{A}}$ are block diagonal matrices. Another interesting result is that every row of the matrix $\left(\mathbf{D}_{-\mathcal{A}}\mathbf{D}_{-\mathcal{A}}^T\right)^{-1}\mathbf{D}_{-\mathcal{A}}$ is a contrast vector, meaning that for any $t = 1, \ldots, m$,

$$\sum_{i=1}^{n}\left[\left(\mathbf{D}_{-\mathcal{A}}\mathbf{D}_{-\mathcal{A}}^T\right)^{-1}\mathbf{D}_{-\mathcal{A}}\right]_{t,\,i} = 0\,.$$

- Another interesting term in analyzing the behaviour of the dual variables is $\mathbf{D}_{\mathcal{A}}^T\mathbf{s}_{\mathcal{A}}$. It can be shown that the vector $\mathbf{D}_{\mathcal{A}}^T\mathbf{s}_{\mathcal{A}}$ can be partitioned into $J+1$ subvectors associated with the change points τ_j, $j = 1, \ldots, J$. The subvector associated with

τ_j, $j = 2, \ldots, J-1$, is $\mathbf{D}^T_{A_j}\,\mathbf{s}_{A_j}$, whose elements are zero, except the first consecutive $r+1$ as well as the last consecutive $r+1$ elements. The first $r+1$ nonzero elements of $\mathbf{D}^T_{A_j}\,\mathbf{s}_{A_j}$ are the binomial coefficients in the expansion of $s_j\,(x-1)^r$, and its last $r+1$ elements are the binomial coefficients in the expansion of $-s_{j+1}\,(x-1)^r$. Furthermore, the first $r+1$ elements of the first subvector and the last $r+1$ elements of the last subvector are also equal to zero. For example, for a piecewise cubic signal, $r = 3$, with two change points $(\tau_1\,,\,\tau_2)$ and signs $(-1\,,\,1)$, the vector $\mathbf{D}^T_{\mathcal{A}}\,\mathbf{s}_{\mathcal{A}}$ becomes

$$\left(\underbrace{0,\ldots,\,0,\,1,\,-3,\,3,\,-1}_{1\,:\,(\tau_1+r_a)},\,\underbrace{-1,\,3,\,-3,\,1,\,0,\,\ldots,\,0,\,-1,\,3,\,-3,\,1}_{(\tau_1+r_a+1)\,:\,(\tau_2+r_a)},\,\underbrace{1,\,-3,\,3,\,-1,\,0,\,\ldots,\,0}_{(\tau_2+r_a+1)\,:\,m}\right).$$

Consequently, the structure of $\mathbf{D}^T_{A_j}\,\mathbf{s}_{A_j}$ allows us to write $\mathbf{D}^T_{\mathcal{A}}\,\mathbf{s}_{\mathcal{A}} = \sum_{j=0}^{J}\mathbf{D}^T_{A_j}\,\mathbf{s}_j$. Additionally, if the signs of two consecutive change points τ_j and τ_{j+1} are the same, then

$$\Big[\big(\mathbf{D}_{-\mathcal{A}}\mathbf{D}^T_{-\mathcal{A}}\big)^{-1}\mathbf{D}_{-\mathcal{A}}\Big]_t\,\Big(\mathbf{D}^T_{A_{j+1}}\mathbf{s}_{j+1} + \mathbf{D}^T_{A_j}\mathbf{s}_j\Big) = -s_j, \tag{3.19}$$

for $t = \tau_j + r_a\,,\,\ldots\,,\,\tau_{j+1} + r_b$.

- Let $\mathbf{P}_D = \mathbf{D}^T_{-\mathcal{A}}\big(\mathbf{D}_{-\mathcal{A}}\mathbf{D}^T_{-\mathcal{A}}\big)^{-1}\mathbf{D}_{-\mathcal{A}}$ be the projection matrix onto the row space of the matrix $\mathbf{D}_{-\mathcal{A}}$. Moreover, let $\mathbf{X}_j$ be the design matrix of the r-th polynomial regression on the indices of j-th segment $\{\tau_j + 1\,,\,\ldots\,,\,\tau_{j+1}\}$, that is,

$$\mathbf{X}_j = \begin{pmatrix} 1 & \frac{\tau_j+1}{n} & \left(\frac{\tau_j+1}{n}\right)^2 & \cdots & \left(\frac{\tau_j+1}{n}\right)^r \\ 1 & \frac{\tau_j+2}{n} & \left(\frac{\tau_j+2}{n}\right)^2 & \cdots & \left(\frac{\tau_j+2}{n}\right)^r \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \frac{\tau_{j+1}}{n} & \left(\frac{\tau_{j+1}}{n}\right)^2 & \cdots & \left(\frac{\tau_{j+1}}{n}\right)^r \end{pmatrix}.$$

The orthogonal projection matrix $\mathbf{I}-\mathbf{P}_D$ is a block diagonal matrix whose j-th block associated with the segment $\{\tau_j + 1\,,\,\ldots\,,\,\tau_{j+1}\}$ is equal to the projection map onto the column space of $\mathbf{X}_j$, i.e.,

$$\mathbf{I} - \mathbf{P}_D = \mathbf{X}_j\big(\mathbf{X}_j^T\,\mathbf{X}_j\big)^{-1}\,\mathbf{X}_j^T. \tag{3.20}$$

Equation (3.10) says that the absolute values of the boundary coordinates are λ, that is,

$$\widehat{u}(t;\lambda) = \lambda\, s_{_j} \qquad\qquad \text{for } t \in A_j. \tag{3.21}$$

On the other hand, the values of the interior coordinates are given by

$$\widehat{u}(t;\lambda) = \begin{cases} \left[\left(\mathbf{D}_{_{-\mathcal{A}}}\mathbf{D}_{_{-\mathcal{A}}}^T\right)^{-1}\mathbf{D}_{_{-\mathcal{A}}}\right]_{_t} \left(\mathbf{y} - \lambda\, \mathbf{D}_{_{A_1}}^T \mathbf{s}_{_1}\right), & 1 \le t < \tau_{_1} - r_{_b} \\ \left[\left(\mathbf{D}_{_{-\mathcal{A}}}\mathbf{D}_{_{-\mathcal{A}}}^T\right)^{-1}\mathbf{D}_{_{-\mathcal{A}}}\right]_{_t} \left(\mathbf{y} - \lambda \left(\mathbf{D}_{_{A_{j+1}}}^T \mathbf{s}_{_{j+1}} + \mathbf{D}_{_{A_j}}^T \mathbf{s}_{_j}\right)\right), & \tau_{_j} + r_{_a} < t < \tau_{_{j+1}} - r_{_b} \\ \left[\left(\mathbf{D}_{_{-\mathcal{A}}}\mathbf{D}_{_{-\mathcal{A}}}^T\right)^{-1}\mathbf{D}_{_{-\mathcal{A}}}\right]_{_t} \left(\mathbf{y} - \lambda\, \mathbf{D}_{_{A_J}}^T \mathbf{s}_{_J}\right), & \tau_{_J} + r_{_a} < t \le m\,. \end{cases} \tag{3.22}$$

For a given λ, the dual variables $\widehat{u}(t;\,\lambda)$ for $t = 0, \ldots, m$ can be collectively viewed as a random bridge, that is, a conditioned random walk with drift whose end points are set to zero. Moreover, $\widehat{u}(t;\,\lambda)$ is bounded between $-\lambda$ and λ. The quantity $\widehat{u}(t;\,\lambda)$ can also be decomposed into a sum of several smaller random bridges which are formed by blocks created from the change points. Recall that the last consecutive $r_{_b}+1$ elements of the block $B_{_j}$ are $\lambda\, s_{_j}$, for any $j = 0,\, 1,\, \cdots,\, J$. Hence, for $t = \tau_{_j} + r_{_a}, \ldots, \tau_{_{j+1}} - r_{_b}$, the random bridge associated with the j-th block is given by

$$\widehat{u}_{_j}(t;\,\lambda) = \left[\left(\mathbf{D}_{_{-\mathcal{A}}}\mathbf{D}_{_{-\mathcal{A}}}^T\right)^{-1}\mathbf{D}_{_{-\mathcal{A}}}\right]_{_t} \left(\mathbf{y} - \lambda \left(\mathbf{D}_{_{A_{j+1}}}^T \mathbf{s}_{_{j+1}} + \mathbf{D}_{_{A_j}}^T \mathbf{s}_{_j}\right)\right), \quad j = 0, \ldots, J\,, \tag{3.23}$$

with the conventions $\mathbf{s}_{_0} = \mathbf{s}_{_{J+1}} = \mathbf{0} \in \mathbb{R}^{r+1}$. It is important to note that similar to $\widehat{u}(t;\,\lambda)$, the process $\widehat{u}_{_j}(t;\,\lambda)$ satisfies the conditions $\widehat{u}_{_j}(\tau_{_j}+r_{_a};\,\lambda) = \lambda\, s_{_j}$ and $\widehat{u}_{_j}(\tau_{_{j+1}}-r_{_b};\lambda) = \lambda\, s_{_{j+1}}$. From (3.23), the process $\widehat{u}_{_j}(t;\,\lambda)$ is composed of the stochastic term

$$\widehat{u}_{_j}^{\,\mathrm{st}}(t) = \left[\left(\mathbf{D}_{_{-\mathcal{A}}}\mathbf{D}_{_{-\mathcal{A}}}^T\right)^{-1}\mathbf{D}_{_{-\mathcal{A}}}\right]_{_t} \mathbf{y}, \tag{3.24}$$

and the drift term

$$\widehat{u}_{_j}^{\,\mathrm{dr}}(t;\,\lambda) = -\lambda \left[\left(\mathbf{D}_{_{-\mathcal{A}}}\mathbf{D}_{_{-\mathcal{A}}}^T\right)^{-1}\mathbf{D}_{_{-\mathcal{A}}}\right]_{_t} \left(\mathbf{D}_{_{A_{j+1}}}^T \mathbf{s}_{_{j+1}} + \mathbf{D}_{_{A_j}}^T \mathbf{s}_{_j}\right). \tag{3.25}$$

According to model (3.1) with Gaussian noises, it turns out that the discrete time stochastic process term $\widehat{u}_{_j}^{\,\mathrm{st}}(t)$ can be embedded in a continuous time Gaussian bridge process. The following theorem describes the characteristics of this process.

Theorem 3.7 *Suppose the observation vector* $\mathbf{y}$ *is drawn from the model* (3.1)*, where the error vector* $\boldsymbol{\varepsilon}$ *has a Gaussian distribution with mean zero and covariance matrix* $\sigma^2\,\mathbf{I}$*. For given* $\mathbf{D}$ *and* $\mathcal{A}$*,*

(a) Define

$$W_j(t) = \left(\tau_{j+1} - \tau_j - r\right)^{-(2r+1)/2} \left[\left(\mathbf{D}_{-\mathcal{A}}\mathbf{D}_{-\mathcal{A}}^T\right)^{-1}\mathbf{D}_{-\mathcal{A}}\right]_{\lfloor mt\rfloor}\mathbf{y}, \tag{3.26}$$

for $(\tau_j + r_a)/m \;\le\; t \;\le\; (\tau_{j+1} - r_b)/m$*, where*

$$W_j\Big(\frac{\tau_j + r_a}{m}\Big) = W_j\Big(\frac{\tau_{j+1} - r_b}{m}\Big) = 0, \tag{3.27}$$

for $j = 0\,,\,\ldots\,,\,J$*. Then the stochastic process* $\mathbf{W}_j = \big\{W_j(t) :\; (\tau_j + r_a)/m \le t \le (\tau_{j+1} - r_b)/m\big\}$ *is a Gaussian bridge process with mean vector zero and covariance function*

$$\mathrm{Cov}\Big(W_j(t)\,,\,W_j(t')\Big) = \sigma^2\left[\left(\mathbf{D}_{-\mathcal{A}}\mathbf{D}_{-\mathcal{A}}^T\right)^{-1}\right]_{\lfloor mt\rfloor,\lfloor mt'\rfloor}, \tag{3.28}$$

for any $(\tau_j + r_a)/m \;\le\; t\,,\,t' \;\le\; (\tau_{j+1} - r_b)/m$*.*

(b) The processes $\mathbf{W}_j$ *and* $\mathbf{W}_{j'}$ *are independent, for* $j' \neq j$*.*

A proof is given in Appendix A.1.

This theorem could be extended to the case of non-Gaussian random variables and therefore establishes a Donsker type Central Limit Theorem for $\mathbf{W}_j$. Theorem 3.7 guarantees that the dual variable process associated with the j-th block, i.e.

$$\mathbf{u}_j = \Big\{\,\widehat{u}\big(\lfloor mt\rfloor;\,\lambda\big) :\; (\tau_j + r_a)/m \le t \le (\tau_{j+1} - r_b)/m\Big\}$$

is a Gaussian bridge process with the drift term

$$-\lambda\left[\left(\mathbf{D}_{-\mathcal{A}}\mathbf{D}_{-\mathcal{A}}^T\right)^{-1}\mathbf{D}_{-\mathcal{A}}\right]_{\lfloor mt\rfloor}\left(\mathbf{D}_{\mathcal{A}_{j+1}}^T\mathbf{s}_{j+1} + \mathbf{D}_{\mathcal{A}_j}^T\mathbf{s}_j\right), \tag{3.29}$$

and the covariance matrix stated in (3.28).

Recall that a standard Brownian bridge process defined on the interval $[a,\,b]$ is a standard Brownian motion $B(t)$ conditioned on the event $B(a) = B(b) = 0$. It is often characterized from a Brownian motion $B(t)$ with $B(a) = 0$, by setting

$$B_0(t) = B(t) - \frac{t-a}{b-a}\,B(b)\,.$$

The mean and covariance functions of the Brownian bridge $B_0(t)$ are given by $\mathrm{E}\big(B_0(t)\big) = 0$ and $\mathrm{Cov}\big(B_0(s),\, B_0(t)\big) = \min\{s-a,\, t-a\} - (b-a)^{-1}(s-a)(t-a)$ for any $s,\, t \in [a,\, b]$, respectively. A Gaussian bridge process is an extension of the Brownian bridge process when the Brownian motion $B(t)$, in the definition of the Brownian bridge $B_0(t)$, is replaced by a more general Gaussian process $G(t)$. See, for example, [61].

Remark 3.8 *The celebrated Donsker theorem [45] states that the partial sum process of a sequence of i.i.d. random variables, with mean zero and variance 1, converges weakly to a Brownian bridge process. See [156] or [24]. A version of Theorem 3.7 involving non-Gaussian random variables would extend this result to weighted partial sum processes and show that the limiting process is a Gaussian bridge with a certain covariance structure. So the Gaussian assumption in Theorem 3.7 is not restrictive. It is also interesting to show that for $r = 0,\, 1$, the process $\widehat{u}_j^{\,st}\big(\lfloor mt \rfloor\big)$ boils down to its respective CUSUM processes. To show this, consider the interval $\big[(\tau_j + r_a)/m\,, (\tau_{j+1} - r_b)/m\big]$,*

- *For the piecewise constant signals, $r = 0$, the quantity $\Big[\big(\mathbf{D}_{-\mathcal{A}}\mathbf{D}_{-\mathcal{A}}^T\big)^{-1}\mathbf{D}_{-\mathcal{A}}\Big]_{\lfloor mt \rfloor}\,\mathbf{y}$ can be written as*

$$\Big(0, \ldots, \underbrace{0}_{\tau_j}, 1 - \frac{\lfloor mt \rfloor}{\tau_{j+1} - \tau_j}, \ldots, \underbrace{1 - \frac{\lfloor mt \rfloor}{\tau_{j+1} - \tau_j}}_{\lfloor mt \rfloor}, -\frac{\lfloor mt \rfloor}{\tau_{j+1} - \tau_j}, \ldots, -\frac{\lfloor mt \rfloor}{\tau_{j+1} - \tau_j}, \underbrace{0}_{\tau_{j+1}}, \ldots, 0\Big)\,\mathbf{y}.$$

  Notice that the above statement is the CUSUM statistic for the j-th segment, that is

$$\sum_{k=\tau_j+1}^{\lfloor mt \rfloor} \Big(y_k - \overline{y}_{(\tau_j+1):\tau_{j+1}}\Big), \tag{3.30}$$

  where $\overline{y}_{(\tau_j+1):\tau_{j+1}}$ is the sample average of $\big(y_{\tau_j+1}, \ldots, y_{\tau_{j+1}}\big)$. It is well known that the CUSUM statistic (3.30) converges weakly to the Brownian bridge. In addition, for any $(\tau_j + r_a)/m \le t' \le t \le (\tau_{j+1} - r_b)/m$, the covariance function becomes

$$\Big[\big(\mathbf{D}_{-\mathcal{A}}\mathbf{D}_{-\mathcal{A}}^T\big)^{-1}\Big]_{(\lfloor mt' \rfloor, \lfloor mt \rfloor)} = (\lfloor mt' \rfloor - \tau_j) - \frac{(\lfloor mt' \rfloor - \tau_j)(\lfloor mt \rfloor - \tau_j)}{\tau_{j+1} - \tau_j},$$

  which is identical to the covariance function of the Brownian bridge.

- *For the piecewise linear signals $r = 1$, the quantity* $\left[\left(\mathbf{D}_{-\mathcal{A}}\mathbf{D}_{-\mathcal{A}}^T\right)^{-1}\mathbf{D}_{-\mathcal{A}}\right]_{\lfloor mt \rfloor}\mathbf{y}$ *reduces to*

$$\sum_{k=\tau_j+1}^{\lfloor mt \rfloor} k\left(y_k - \widehat{f}_k\right), \tag{3.31}$$

where $\widehat{f}$ is the least square fit of the simple linear regression of $\left(y_{\tau_j+1}, \ldots, y_{\tau_{j+1}}\right)$ onto $\left(\tau_j + 1, \ldots, \tau_{j+1}\right)$. As proved in Theorem 3.7, the preceding statistic (3.31) is also a Gaussian bridge process. Furthermore, using the results in [75], for any $(\tau_j + r_a)/m \leq t' \leq t \leq (\tau_{j+1} - r_b)/m$, the covariance function of this Gaussian bridge process is given by

$$\left[\left(\mathbf{D}_{-\mathcal{A}}\mathbf{D}_{-\mathcal{A}}^T\right)^{-1}\right]_{(\lfloor mt' \rfloor, \lfloor mt \rfloor)} = \frac{\left(\Delta_j - \lfloor mt \rfloor + \tau_j\right)\left(\Delta_j - \lfloor mt \rfloor + \tau_j + 1\right)}{3\,\Delta_j\left(\Delta_j + 1\right)\left(\Delta_j + 2\right)}$$

$$\times \left(\lfloor mt' \rfloor - \tau_j\right)\left(\lfloor mt' \rfloor - \tau_j + 1\right)$$

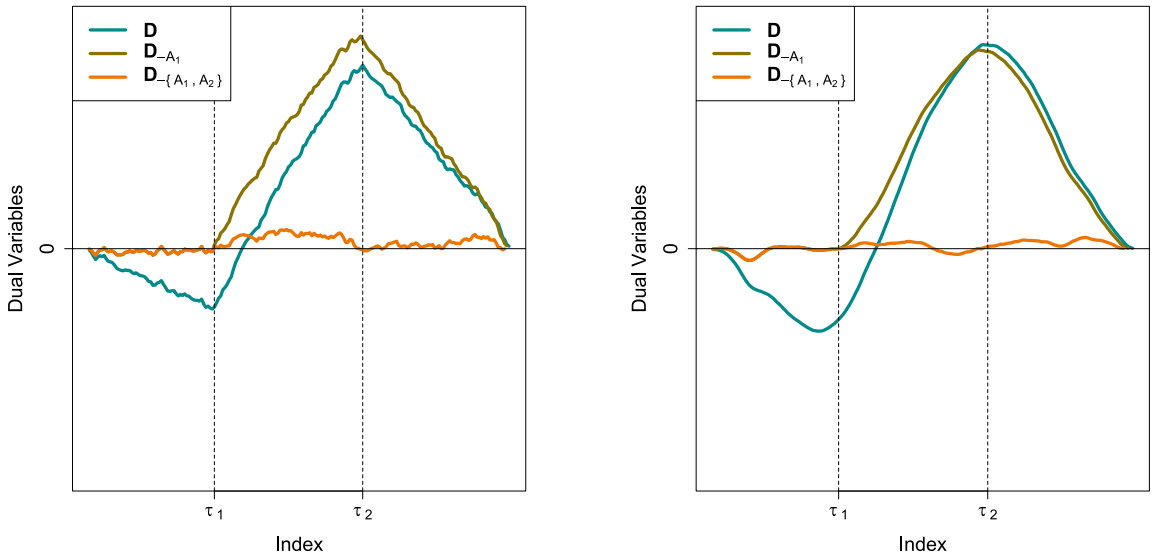
$$\times \left[\left(\lfloor mt \rfloor - \tau_j + 1\right)\left(\lfloor mt' \rfloor - \tau_j - 1\right)\left(\Delta_j + 2\right) - \left(\lfloor mt \rfloor - \tau_j\right)\left(\lfloor mt' \rfloor - \tau_j + 2\right)\Delta_j\right],$$

where $\Delta_j = \tau_{j+1} - \tau_j$.

3.6 Stopping Criterion

This section concerns developing a stopping criterion for the PRUTF algorithm. We provide tools for deriving a threshold value at which the PRUTF algorithm terminates the search if no values of dual variables exceed this threshold. Consider the dual variables at the first step of the algorithm, i.e. $\widehat{u}^{\text{st}}(t) = \left[\left(\mathbf{D}\mathbf{D}^T\right)^{-1}\mathbf{D}\right]_t\mathbf{y}$, for $t = 0, \ldots, m$, which correspond to $\widehat{u}^{\text{st}}(t)$ in (3.24) with $\mathcal{A} = \emptyset$. It turns out that $\widehat{u}^{\text{st}}(t)$ is a stochastic process with local minima and maxima attained at the change points. This structure is displayed with cyan-colored lines (—) in Figure 3.4 for both piecewise constant $r = 0$ and piecewise linear $r = 1$ signals. As the PRUTF algorithm detects more change points and forms the augmented boundary set $\mathcal{A}$, the local minima or maxima corresponding to these change points are removed from the stochastic process

$$\widehat{u}^{\text{st}}_{-\mathcal{A}}(t) = \left[\left(\mathbf{D}_{-\mathcal{A}}\mathbf{D}_{-\mathcal{A}}^T\right)^{-1}\mathbf{D}_{-\mathcal{A}}\right]_t\mathbf{y} = \sum_{j=0}^{J}\widehat{u}^{\text{st}}_j(t)\,\mathbb{1}\{t \in B_j\}, \tag{3.32}$$



(a) Piecewise constant with $r = 0$ **(b)** Piecewise linear with $r = 1$

Figure 3.4: The cyan-colored lines show the dual variables for the full matrix $\mathbf{D}$. Dual variables computed after removing rows of the matrix $\mathbf{D}$ associated with τ_1, that is $\mathbf{D}_{-\mathcal{A}_1}$, are displayed by the olive-colored lines. The augmented boundary set $\mathcal{A}_2$ corresponding to τ_1 and τ_2 results to the dual variables shown by orange-colored lines.

for $t = 1, \ldots, m - |\mathcal{A}|$. This fact is shown by olive-colored lines (—) in Figure 3.4. The last equality in (3.32) expresses that the $\widehat{u}^{\text{st}}_{-\mathcal{A}}(t)$ is the stochastic term of the dual variables for all the interior coordinates and is derived by stacking the stochastic terms of the dual variables associated with j-th block, $\widehat{u}^{\text{st}}_{j}(t)$, as defined in (3.24), for $j = 0, \ldots, J$. This behaviour suggests a way to introduce a stopping rule for the PRUTF algorithm. As can be viewed from the orange-colored lines (—) of Figure 3.4, if all true change points are captured by the algorithm and stored in the augmented set $\mathcal{A}_0$, the resulting process

$$\widehat{u}^{\text{st}}_{-\mathcal{A}_0}(t) = \left[\left(\mathbf{D}_{-\mathcal{A}_0} \mathbf{D}^T_{-\mathcal{A}_0} \right)^{-1} \mathbf{D}_{-\mathcal{A}_0} \right]_t \mathbf{y} \qquad \text{for} \quad t = 0, \ldots, m - |\mathcal{A}_0|,$$

contains no noticeable optimum points and tends to fluctuate close to the zero line (x-axis).

We terminate the search in Algorithm 3.1 at step j by checking whether the maximum of $\left| \widehat{u}^{\text{st}}_{-\mathcal{A}_j}(t) \right|$, for $t = 0, \ldots, m - |\mathcal{A}_j|$, is smaller than a certain threshold. To exactly specify this threshold, as suggested by Theorem 3.7, we need to calculate the *excursion*

probabilities of a Gaussian bridge process. As stated in [1], analytic formulas for the excursion probabilities are known to be available only for a small number of Gaussian processes. One of such Gaussian processes is the Brownian bridge process. It is well known that for the Brownian bridge process $B_0(t)$ defined on the interval $[a,\, b]$

$$\Pr\left(\sup_{a\le t\le b} \left| B_0(t) \right| \ge x \right) = 2 \sum_{i=1}^{\infty} (-1)^{i+1} \exp\left(\frac{-2\, i^2\, x^2}{b-a} \right) . \tag{3.33}$$

See, for example, [1], and [135]. Hence for the piecewise constant signals, the required threshold for stopping Algorithm 3.1 can be obtained from (3.33), for a suitably chosen interval $[a,\, b]$. That is, for a given value α, we choose x_α such that $\Pr\left(\sup_{a\le t\le b} | B_0(t) | \ge x_\alpha \right) = 1-\alpha$. Therefore, for $r=0$ and $a=0$, $b=1$, we stop Algorithm 3.1 at the iteration j_0 if

$$\max_{0\le t\le 1} \left| \widehat{\mathbf{u}}^{\,\mathrm{st}}_{-\mathcal{A}_{j_0}} \left(\lfloor kt \rfloor \right) \right| \;\le\; \sigma\, x_\alpha\, \sqrt{k}\,, \qquad\qquad \text{for} \quad t = 0,\, \ldots,\, m - |\mathcal{A}_{j_0}|\,,$$

and $k = m - |\mathcal{A}_{j_0}|$.

For $r \ge 1$, the threshold is obtained in a similar fashion. Although the excursion probabilities for the Gaussian bridge processes are not known, we notice that by adopting the steps for the proof of (3.33) in [19], we can establish a similar formula for the Gaussian bridge process $G_0(t)$ in Theorem 3.7 as

$$\Pr\left(\sup_{a\le t\le b} \left| G_0(t) \right| \ge x \right) = 2 \sum_{i=1}^{\infty} (-1)^{i+1} \exp\left(\frac{-2\, i^2\, x^2}{S_r^2(k)} \right) , \tag{3.34}$$

where $k = m - |\mathcal{A}_{j_0}|$, and the quantity $S_r^2(k)$ is the k-th diagonal element of the matrix

$$\left(\mathbf{D}_{-\mathcal{A}_{j_0}}\, \mathbf{D}^T_{-\mathcal{A}_{j_0}} \right)^{-1} .$$

Hence, we stop Algorithm 3.1 at the iteration j_0 if

$$\max_{0\le t\le 1} \left| \widehat{\mathbf{u}}^{\,\mathrm{st}}_{-\mathcal{A}_{j_0}} \left(\lfloor kt \rfloor \right) \right| \le \sigma x_\alpha\, (k-r)^{(2r+1)/2} , \qquad \text{for} \quad t = 0,\, \ldots,\, m - |\mathcal{A}_{j_0}|, \tag{3.35}$$

where x_α is derived from the equation

$$\sum_{i=1}^{\infty} (-1)^{i+1} \exp\left(\frac{-2\, i^2\, x_\alpha^2}{S_r^2(k)} \right) = \frac{\alpha}{2}\,. \tag{3.36}$$

3.7 Pattern Recovery and Theories

The main purpose of this section is to investigate whether the PRUTF algorithm can recover features of the true signal $\mathbf{f}$. We also demonstrate conditions under which the structure of the estimated signal $\widehat{\mathbf{f}}$ matches the true signal $\mathbf{f}$. To verify the performance of PRUTF in the discovery the true signal, we first define what we mean by pattern recovery.

Definition 3.9 *(Pattern Recovery): A trend filtering estimate $\widehat{\mathbf{f}}$ recovers the pattern of the true signal $\mathbf{f}$ if*

$$sign\big(\big[\,\mathbf{D}\widehat{\mathbf{f}}\,\big]_i\big) = sign\big(\big[\,\mathbf{D}\mathbf{f}\,\big]_i\big), \qquad for \quad i = 1, \ldots, m, \tag{3.37}$$

where $m = n - r - 1$ is the number of rows of matrix $\mathbf{D}$. We use the notation $\widehat{\mathbf{f}} \overset{pr}{=} \mathbf{f}$ to briefly denote the pattern recovery feature of $\widehat{\mathbf{f}}$.

In the asymptotic framework, a trend filtering estimate is called pattern consistent if

$$\Pr\big(\widehat{\mathbf{f}} \overset{pr}{=} \mathbf{f}\big) \longrightarrow 1 \qquad \text{as} \quad n \longrightarrow \infty, \tag{3.38}$$

where $\widehat{\mathbf{f}} = \widehat{\mathbf{f}}_n$, to denote its dependency to the sample size n. Pattern recovery is very similar to the concept of sign recovery in lasso [173, 159] as it deals with the specification of both locations of non-zero coefficients and their signs.

The problem of pattern recovery is studied for the special case of the fused lasso in several papers. [129] derived conditions under which fused lasso consistently identifies the true pattern. This was contradicted by [132], who argued that fused lasso does not always succeed in discovering the exact change points. [132] showed that fused lasso can be reformulated as the usual lasso, for which the necessary conditions for exact sign recovery have been established in the literature. Then, they proved that one such necessary condition, known as the irrepresentable condition, is not satisfied for the transformed lasso when there is a specific pattern called a staircase (Definition 3.11). Corrections to [129] appeared in [128]. Later on, [123] proposed a method called puffer transformation, which is shown to be consistent in specifying the exact change points, including in the presence of staircases.

In the remaining part of this section, we use the dual variables to demonstrate the situations in which PRUTF can correctly recover the pattern of the true signal. Exact pattern recovery implies that the dual variables are comprised of J_0+1 consecutive bounded processes whose endpoints correspond to the true change points. The following lemma describes the situations in which exact pattern recovery can be attained. A particular case of this result in the context of piecewise constant signals was established in [128].

Theorem 3.10 *Exact pattern recovery in PRUTF occurs when the discrete time processes* $\left\{\widehat{u}_j^{\mathrm{st}}(t),\ t = \tau_j + r_a, \ldots, \tau_{j+1} - r_b\right\}$, *for* $j = 0, \ldots, J_0$, *satisfy the following conditions simultaneously with probability one:*

(a) **First block constraint:** *for* $t = 1, \ldots, \tau_1 - r_b$,

$$-\lambda\left(1 - \left[(\mathbf{D}_{-\mathcal{A}}\mathbf{D}_{-\mathcal{A}}^T)^{-1}\mathbf{D}_{-\mathcal{A}}\right]_t \mathbf{D}_{A_1}^T\mathbf{1}\right) \le \widehat{u}_0^{\mathrm{st}}(t) \le \lambda\left(1 + \left[(\mathbf{D}_{-\mathcal{A}}\mathbf{D}_{-\mathcal{A}}^T)^{-1}\mathbf{D}_{-\mathcal{A}}\right]_t \mathbf{D}_{A_1}^T\mathbf{1}\right). \tag{3.39}$$

(b) **Last Block constraint:** *for* $t = \tau_{J_0} + r_a, \ldots, m$,

$$-\lambda\left(1 + \left[(\mathbf{D}_{-\mathcal{A}}\mathbf{D}_{-\mathcal{A}}^T)^{-1}\mathbf{D}_{-\mathcal{A}}\right]_t \mathbf{D}_{A_{J_0}}^T\mathbf{1}\right) \le \widehat{u}_{J_0}^{\mathrm{st}}(t) \le \lambda\left(1 - \left[(\mathbf{D}_{-\mathcal{A}}\mathbf{D}_{-\mathcal{A}}^T)^{-1}\mathbf{D}_{-\mathcal{A}}\right]_t \mathbf{D}_{A_{J_0}}^T\mathbf{1}\right). \tag{3.40}$$

(c) **Interior Block constraints:** *for* $t = \tau_j + r_a, \ldots, \tau_{j+1} - r_b$, *if* $s_j \neq s_{j+1}$

$$-\lambda\left(1 - \left[(\mathbf{D}_{-\mathcal{A}}\mathbf{D}_{-\mathcal{A}}^T)^{-1}\mathbf{D}_{-\mathcal{A}}\right]_t (\mathbf{D}_{A_{j+1}}^T\mathbf{1} - \mathbf{D}_{A_j}^T\mathbf{1})\right) \le \widehat{u}_j^{\mathrm{st}}(t)$$
$$\le \lambda\left(1 + \left[(\mathbf{D}_{-\mathcal{A}}\mathbf{D}_{-\mathcal{A}}^T)^{-1}\mathbf{D}_{-\mathcal{A}}\right]_t (\mathbf{D}_{A_{j+1}}^T\mathbf{1} - \mathbf{D}_{A_j}^T\mathbf{1})\right), \tag{3.41}$$

and if $s_j \neq s_{j+1}$, *which corresponds to a staircase block,* $\widehat{u}_j^{\mathrm{st}}(t) \le 0$ *or* $\widehat{u}_j^{\mathrm{st}}(t) \ge 0$.

In the foregoing equations, $\mathbf{1} \in \mathbb{R}^{r+1}$ is a vector of size $r+1$ whose elements are all 1. A proof of the theorem is given in Appendix A.2.

We analyze the performance of the PRUTF algorithm in pattern recovery in two different scenarios;

- signals with staircase patterns,
- signals without staircase patterns.

To our knowledge, [132] was the first paper to carefully investigate the staircase pattern for the piecewise constant signals in the change points analysis setting. In [132], a staircase pattern for a piecewise constant signal refers to the phenomenon of equal signs in two consecutive changes. We extend this concept to the general case, which covers any piecewise polynomial signals of order r, by applying the penalty matrix $\mathbf{D} = \mathbf{D}^{(r+1)}$.

Definition 3.11 *Suppose that the true signal* $\mathbf{f}$ *is a piecewise polynomial of order* r *with change points at the locations* $\boldsymbol{\tau} = \{\tau_1, \ldots, \tau_{J_0}\}$. *Moreover, let* $\mathbf{B} = \{B_0, \ldots, B_{J_0}\}$ *be blocks created by the change points in* $\boldsymbol{\tau}$. *A staircase occurs in block* B_j, $j = 1, \ldots, J_0 - 1$ *if*

$$\operatorname{sign}\big(\big[\,\mathbf{D}\mathbf{f}\,\big]_{\tau_j}\big) = \operatorname{sign}\big(\big[\,\mathbf{D}\mathbf{f}\,\big]_{\tau_{j+1}}\big). \tag{3.42}$$

The following theorem investigates the consistency of PRUTF in pattern recovery, in both with and without staircases. Specifically, it shows that for a signal without a staircase, the exact pattern recovery conditions are satisfied with probability one. On the other hand, in the presence of staircases in the signal, the probability of these conditions holding, which is equivalent to the probability of a Gaussian bridge process never crossing the zero line, converges to zero.

In the literature, the consistency of a change point method is usually characterized by the signal size n, the number of change points J_0, the noise variance σ_n^2, the minimal spacing between change points,

$$\underline{L}_n = \min_{j=0,\ldots,J_0} \big|L_{n,j}\big| = \min_{j=0,\ldots,J_0} \big|\tau_{j+1} - \tau_j\big|,$$

and the minimum magnitude of jumps between change points,

$$\delta_n = \min_{j=1,\ldots,J_0} \big|\mathbf{D}_{\tau_j}\mathbf{f}\big|.$$

All the above quantities are allowed to change as n grows.

In the following, we present our main theorem providing conditions under which the output of the PRUTF algorithm consistently recovers the pattern of the true signal $\mathbf{f}$.

Theorem 3.12 *Suppose that* $\mathbf{y}$ *follows the model in* (3.1). *Let* $\boldsymbol{\tau}$ *be the set of* J_0 *change points for the true signal* $\mathbf{f}$. *Additionally, assume that* $\widehat{\boldsymbol{\tau}}_n$ *and* $\widehat{\mathbf{f}}_n$ *are the set of change points estimates and the corresponding signal estimate obtained by the PRUTF algorithm, respectively. The followings hold for the PRUTF algorithm.*

(a) **Non-staircase Blocks:** *Suppose there is no staircase block in the true signal* $\mathbf{f}$. *For some* $\xi > 0$ *and with*

$$\lambda_n < \frac{\delta_n\, \underline{L}_n^{2r+1}}{n^{2r}\, 2^{\,r+2}},$$

if the conditions

- $$\frac{\delta_n \, \underline{L}_n^{r+1/2}}{n^r \, \sigma_n} \longrightarrow \infty \qquad \textit{and} \qquad \frac{\delta_n \, \underline{L}_n^{r+1/2}}{2^{\,r/2+2} \, n^r \, \sigma_n \, \sqrt{\log(J_0)}} \; > \; (1+\xi), \quad (3.43)$$

- $$\frac{\lambda_n \, \underline{L}_n^{r+1/2}}{n^r \, \sigma_n} \longrightarrow \infty \qquad \textit{and} \qquad \frac{2^{\,r/2+1} \, \lambda_n \, \underline{L}_n^{r+1/2}}{n^r \, \sigma_n \, \sqrt{\log(n - J_0)}} \; > \; (1+\xi), \quad (3.44)$$

hold, then the PRUTF algorithm guarantees exact pattern recovery with probability approaching one. That is,

$$\Pr\left(\widehat{\mathbf{f}}_n \overset{pr}{=} \mathbf{f}\right) \longrightarrow 1 \qquad \textit{as} \qquad n \longrightarrow \infty.$$

(b) **Staircase Blocks:** *On the other hand, if the true signal* $\mathbf{f}$ *contains at least one staircase block, then the probability of exact pattern recovery by the PRUTF algorithm converges to zero. That is,*

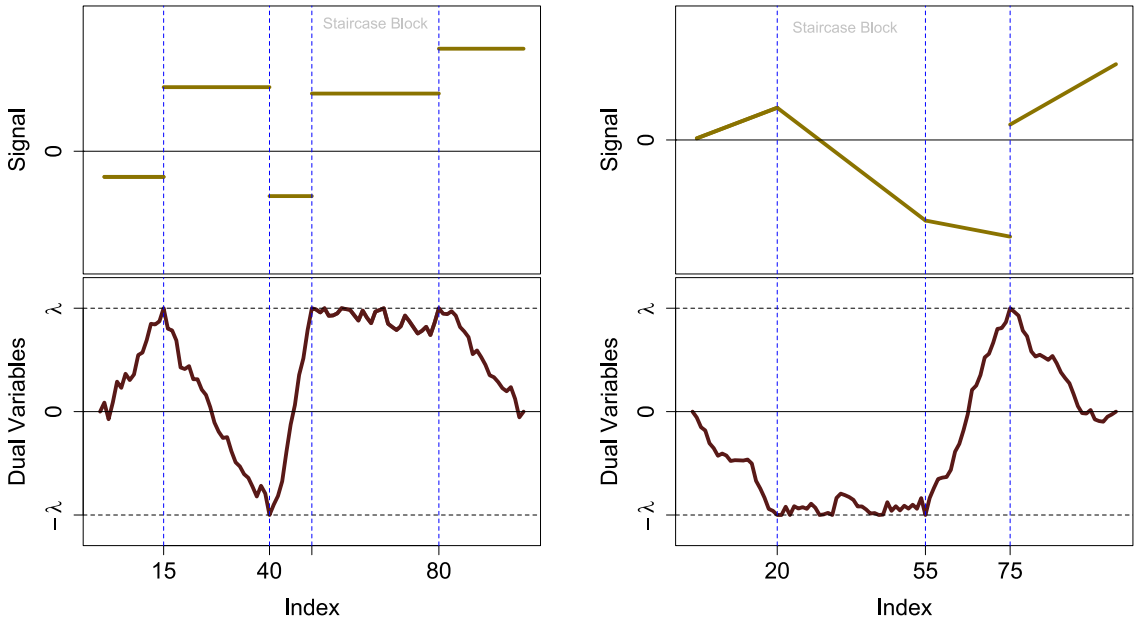
$$\Pr\left(\widehat{\mathbf{f}}_n \overset{pr}{=} \mathbf{f}\right) \longrightarrow 0 \qquad \textit{as} \qquad n \longrightarrow \infty.$$

A proof is given in Appendix A.3.

Remark 3.13 *The performance of PRUTF in terms of consistent pattern recovery relies on the quantity* $\delta_n \, \underline{L}_n^{r+1/2}/\sigma_n$ *and the choice of* λ_n*. In the piecewise constant case, the former quantity reduces to the well-known signal-to-noise-ratio quantity, which is crucial for a consistent change point estimation [56, 160]. The statements in 3.43 illustrate that the consistency of PRUTF in non-staircase blocks is achievable if the quantity* $\delta_n \, \underline{L}_n^{r+1/2}/\sigma_n$ *is of order* $O(n^{r+c})$*, for some* $c > 0$*. In addition, the number of the change points* J_0 *is allowed to diverge, provided*

$$\log\left(J_0\right) \; \lesssim \; \frac{\delta_n^2 \, \underline{L}_n^{2r+1}}{n^{2r} \, \sigma_n^2} \, .$$

The drift term (3.25) plays a key role in assessing the performance of PRUTF in pattern recovery. From (3.19), this drift for a staircase block B_j becomes $\lambda \, s_j$, which is constant in t for the entire block. Consequently, the interior dual variables $\widehat{u}_j(t;\lambda)$ for the staircase block B_j contain only the stochastic term $\widehat{u}_j^{\,\mathrm{st}}(t) = \left[\left(\mathbf{D}_{-\mathcal{A}}\mathbf{D}_{-\mathcal{A}}^T\right)^{-1}\mathbf{D}_{-\mathcal{A}}\right]_t \mathbf{y}$, which fluctuates around the line $\lambda \, s_j$. Recall that the KKT conditions for the dual problem of trend filtering require $\widehat{u}_j(t;\,\lambda)$ to stay within the lines $-\lambda$ and λ. Thus, for a signal with staircase



(a) Piecewise constant signal with staircase block (50, 80].

(b) Piecewise linear signal with staircase block (20, 55].

Figure 3.5: Piecewise constant and piecewise linear signals with staircase pattern at blocks (50, 80] and (20, 55] and their corresponding dual variables.

patterns, the PRUTF algorithm is sensitive to the variability of random noises and identifies change points once $\widehat{u}_j^{\mathrm{st}}(t)$ touches the $\pm\lambda$ boundaries. Examples of piecewise constant and piecewise linear signals, along with their corresponding dual variables, are depicted in Figure 3.5, in which the above argument can be clearly seen.

According to Theorem 3.12, if there is no staircase pattern in the underlying signal, the PRUTF algorithm consistently estimates the true signal, and fails to do so, otherwise. Given the results in Theorem 3.12, the natural question is whether Algorithm 3.1 could be modified to enjoy the consistent pattern recovery in any case. In the next section, we will present an effective remedy based on altering the sign of a change associated with a staircase block.

3.8 Modified PRUTF Algorithm

In this section, we attempt to modify the PRUTF algorithm in such a way that it produces consistent estimates of the number and locations of change points even in the presence of staircase patterns. As previously mentioned, for a staircase block, the drift term (3.25) is constant and leads to false discoveries in change points. This is shown in Figure 3.6 with a piecewise constant signal of size $n = 100$ and the true change points at $\boldsymbol{\tau} = \{15, 40, 50, 80\}$. The figure reveals that the staircase block (50, 80] leads to three false discoveries at the locations 52, 54 and 76.

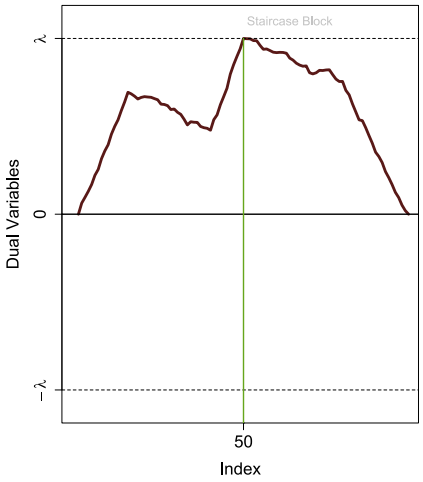
The inconsistency of PRUTF in the presence of a staircase as established in Theorem 3.12, stems from the fact that the change signs of the two consecutive change points at both ends of the staircase block are identical. That is, for the staircase block B_j, $\operatorname{sign}\big([\,\mathbf{Df}\,]_{\tau_j}\big) = \operatorname{sign}\big([\,\mathbf{Df}\,]_{\tau_{j+1}}\big)$. Therefore, a question arises: can we modify Algorithm 3.1 in such a way that the change signs of two neighbouring change points never become equal but still yield the solution path of trend filtering? We suggest a simple but very efficient solution to the above question.

Once a new change point is identified, we check whether its r-th order difference sign is the same as that of the change points right before and after. If these change signs are not identical, then the procedure continues to search for the next change point. Otherwise, we replace the sign of the neighbouring change point with zero. This replacement of the sign prevents the drift term (3.25) from becoming zero. This idea is implemented for the above signal, and the result is displayed in Figure 3.7. As shown in panel (b), the sign of the first change point at location 50 is set to zero since its sign is identical to the sign of the second change point at 15. This sign replacement vanishes false discoveries appeared in panel (b) of Figure 3.6.

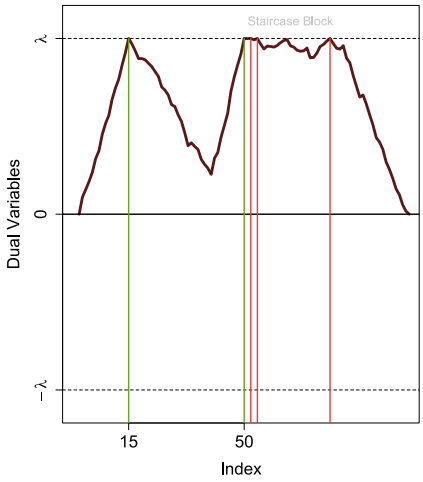
Based on the above argument, PRUTF presented in Algorithm 3.1 can be modified as follows to avoid false discovery and to produce consistent pattern recovery.

Algorithm 3.14 (mPRUTF)

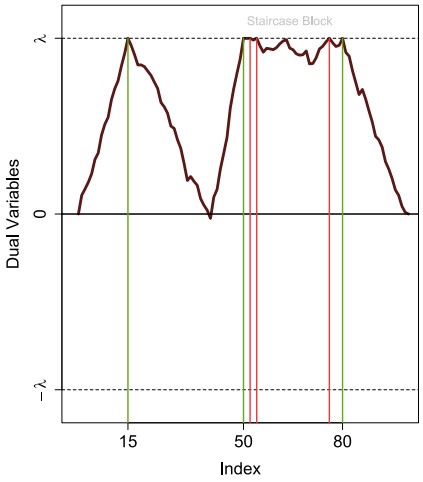
1. *Execute steps 1 and 2 of Algorithm 3.1.*
2. *(a) Execute part (a) of step 3 in Algorithm 3.1 to obtain* τ_j^{join} *and its sign* s_j^{join}. *At this point, the algorithm checks whether* s_j^{join} *is identical to the signs of the change points just before and after* τ_j^{join}. *If so, set the sign of change point which is identical to* s_j^{join} *to zero. Then, repeat part (a) of step 3 again to obtain new* τ_j^{join} *and* s_j^{join} *and update the sets* $\mathcal{A}_j$ *and* $\mathcal{B}_j$.



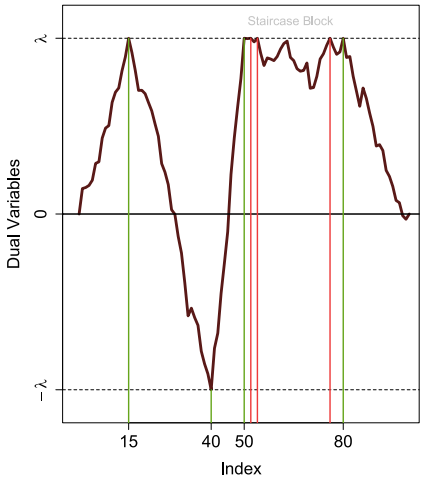
(a) First change point at $\tau = 50$.



(b) Second change point at $\tau = 15$.



(c) Third change point at $\tau = 80$.

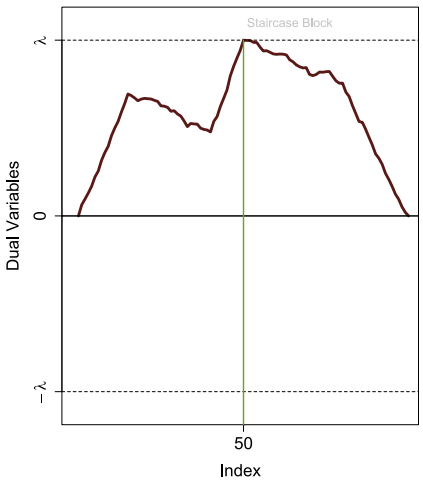


(d) Fourth change point at $\tau = 40$.

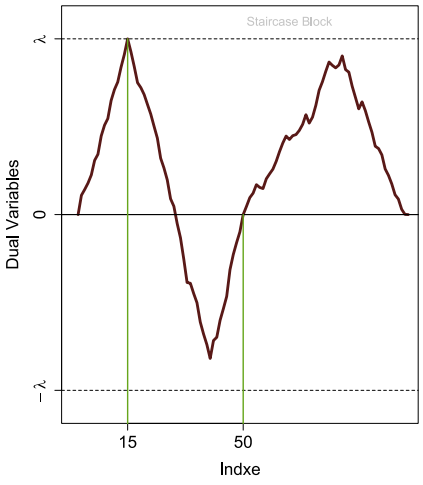
Figure 3.6: The process of detecting change points using PRUTF for a signal with a staircase pattern. In panel (b), there are three falsely detected change points {52, 54, 76} which is due to the staircase block (50, 80].

(b) Execute parts (b) and (c) of step 3 in Algorithm 3.1.

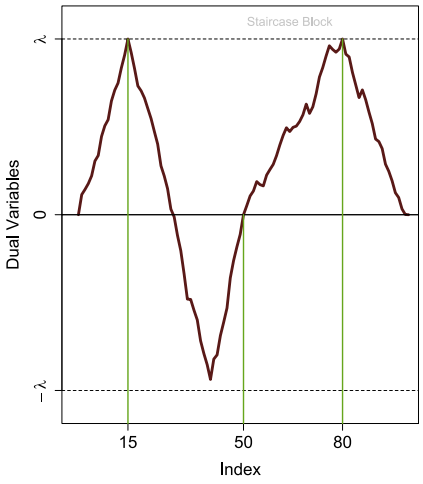
3. *Repeat step 3 until either* $\lambda_j > 0$ *or a stopping rule is met.*



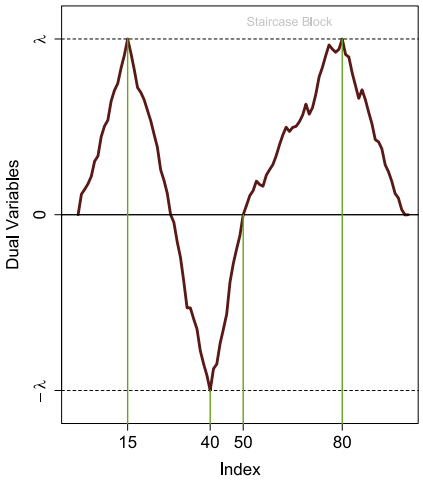
(a) First change point at $\tau = 50$.



(b) Second change point at $\tau = 15$.



(c) Third change point at $\tau = 80$.



(d) Fourth change point at $\tau = 40$.

Figure 3.7: Steps of the mPRUTF algorithm until all four true change points are identified.

The modified PRUTF (mPRUTF) algorithm produces consistent change point estimations, even in the presence of staircase patterns. This consistency has been achieved by converting the staircase patterns to non-staircase patterns that avoid false change point detection. In other words, running mPRUTF on an arbitrary signal (with or without staircases) is equivalent to running PRUTF on a signal without any staircase; see Figures 3.6 and 3.7. Thus, from part (a) of Theorem 3.12, the mPRUTF algorithm is consistent in pattern recovery.

Remark 3.15 *In step 2, part (a) of the mPRUTF algorithm, presented in Algorithm 3.14, it is impossible for the sign s_j^{join} of the new change point to be identical to the sign of both of its immediate neighbouring change points, because the algorithm has already checked the equality of signs at previous steps. If they are equal, the sign of the immediate neighbouring change point will be set to zero.*

Recall that the KKT optimality conditions for solutions of the trend filtering problem in (3.6) requires the dual variables $\widehat{\mathbf{u}}_\lambda$ to be less than or equal to λ in absolute values, i.e., $|\widehat{\mathbf{u}}_\lambda| \leq \lambda$. This condition still holds when we replace the sign values ($+1$ or -1) with 0. Consequently, we have the following theorem.

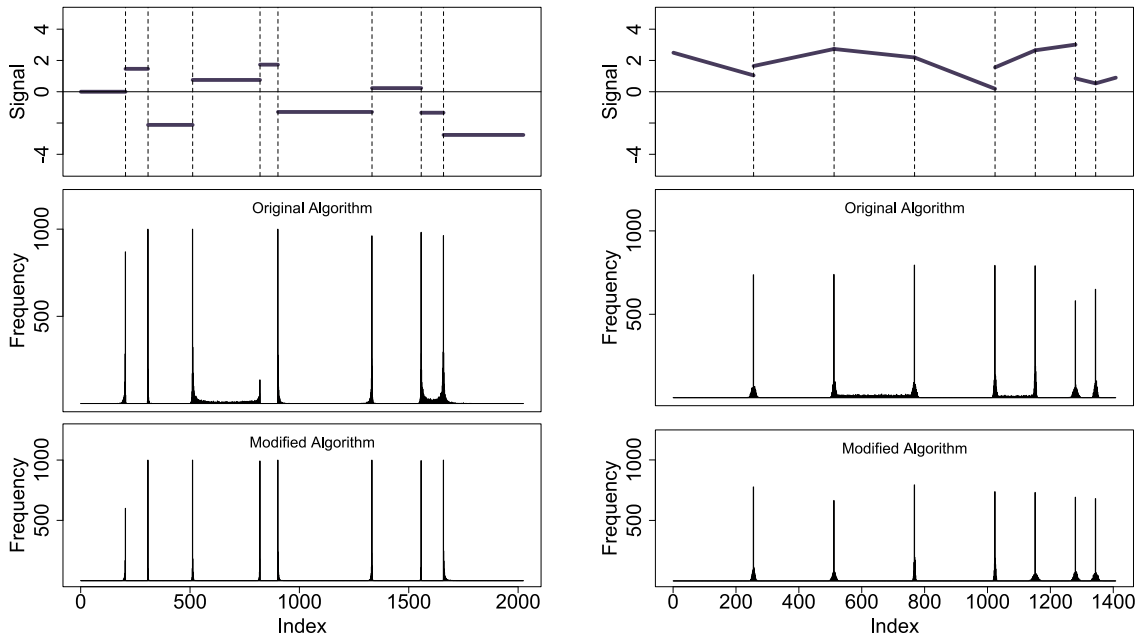
Theorem 3.16 *The mPRUTF algorithm presented in Algorithm 3.14 is a solution path of trend filtering.*

For brevity, we do not provide the proof of Theorem 3.16 here. We refer the reader to the similar arguments for the LARS algorithm of lasso in [150].

It is worth pointing out that the mPRUTF algorithm requires slightly more computation than the original PRUTF algorithm. The increase in computation time directly depends on the number of staircase blocks in the underlying signal. To show how mPRUTF resolves the problem of false discovery in signals with staircases, we ran both algorithms for 1000 generated datasets from a piecewise constant and piecewise linear signals. The frequency plot of the estimated change points for both algorithms are represented in Figure 3.8. The figure reveals that the original algorithm produces false discoveries within staircase blocks for both signals, whereas mPRUTF resolves this issue.

3.9 Numerical Studies

In this section, we provide numerical studies to demonstrate the effectiveness and performance of our proposed algorithm, mPRUTF . We begin with a simulation study and then provide real data analyses.



(a) A piecewise constant signal with blocks 4 and 8 as staircase blocks.

(b) A piecewise linear signal with blocks 3 and 5 as staircase blocks.

Figure 3.8: The frequency plots of estimated change points using the PRUTF and mPRUTF algorithms.

3.9.1 Simulation Study

In this section, we investigate the performance of our proposed method, mPRUTF, by a simulation study. We consider two scenarios, namely piecewise constant and piecewise linear signals with staircase patterns. We compare our method to some powerful state-of-the-art approaches in change point analysis. These methods, a list of their available packages on CRAN, and their applicability for different scenarios are listed in Table 3.1.

We have adopted the simulation setting of [16], and consider piecewise constant and piecewise linear signals as follows.

(i) A piecewise constant signal (PWC) of size $n = 2024$ with the number of change points $J_0 = 8$. The locations of the true change points are $\boldsymbol{\tau} = \{205, 308, 512,$

Method	Reference	R Package	Signal	
			PWC	PWL
PELT	[85]	**changepoint**	✓	×
WBS	[56]	**wbs**	✓	×
SMUCE	[53]	**stepR**	✓	×
NOT	[16]	**not**	✓	✓
ID	[4]	**IDetect**	✓	✓

Table 3.1: A list of change point detection and estimation methods with their packages in CRAN. The last two columns indicate which methods can be applied to piecewise constant or/and piecewise linear signals.

820, 902, 1332, 1557, 1659$\big\}$ with jump sizes 1.464, -0.656, 0.098, 1.830, 0.537, 0.768, -0.574, -3.335. We set the starting intercept to 0.

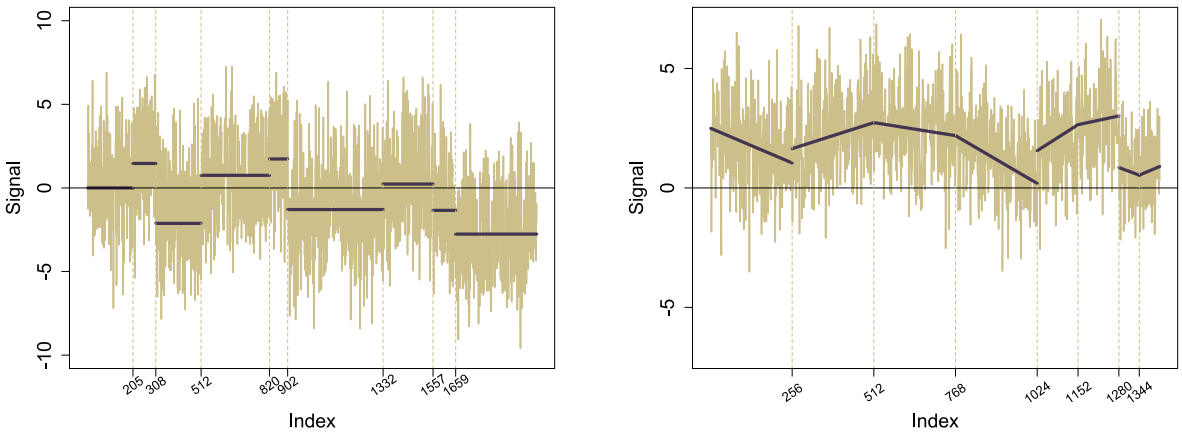
(ii) A piecewise linear signal (PWL) of size $n = 1408$ and the number of change points $J_0 = 7$. The true change points are located at $\boldsymbol{\tau} = \big\{256,\ 512,\ 768,\ 1024,\ 1152,\ 1280,\ 1344\big\}$. The corresponding intercepts and slopes for 8 created blocks by $\boldsymbol{\tau}$ are 0.111, 0.553, -0.481, 3.002, -7.169, -0.030, 7.217, -0.958 and -8, 6, -3, -11, 12, 4, -7, 8, respectively.

Figure 3.9 displays the true PWC and PWL signals, with their representative datasets generated using model (3.1). We note that both PWC and PWL signals contain two staircase blocks. These blocks for the PWC signal are $(512\,,\,820]$, $(1557\,,\,1659]$ and for PWL signal are $(512\,,\,768]$ and $(1024\,,\,1152]$.

We apply mPRUTF presented in Algorithm 3.14 to estimate the number and the locations of the change points for the PWC and PWL signals. In each iteration of the simulation study, we simulate a dataset according to model (3.1) under the assumption that the error terms are independently and identically distributed as $N(0\,,\,\sigma^2)$. Moreover, we set the significance level to $\alpha = 0.05$ for the stopping rule in (3.35).

In order to explore the impact of different noise levels on the change point methods, we run each simulation for various values of σ in $\big\{0.5,\ 1,\ 1.5,\ \ldots\ ,\ 4.5,\ 5\big\}$. We run the simulation $N = 5000$ times and report the results for each change point technique in terms of estimates of the number of change points, estimates of the mean square error given by $\text{MSE} = N^{-1}\sum_{i=1}^{N}\big(\widehat{f}_i - f_i\big)^2$, estimates of the scaled Hausdorff distance given by

$$d_H = \frac{1}{N}\,\max\left\{\max_{j=0,\,\ldots,\,J_0}\ \min_{i=0,\,\ldots,\,\widehat{J}_0}\big|\widehat{\tau}_i - \tau_j\big|\ \ ,\ \ \max_{i=0,\,\ldots,\,\widehat{J}_0}\ \min_{j=0,\,\ldots,\,J_0}\big|\widehat{\tau}_i - \tau_j\big|\right\},$$



(a) PWC signal with staircases at blocks (512 , 820] and (1557 , 1659].

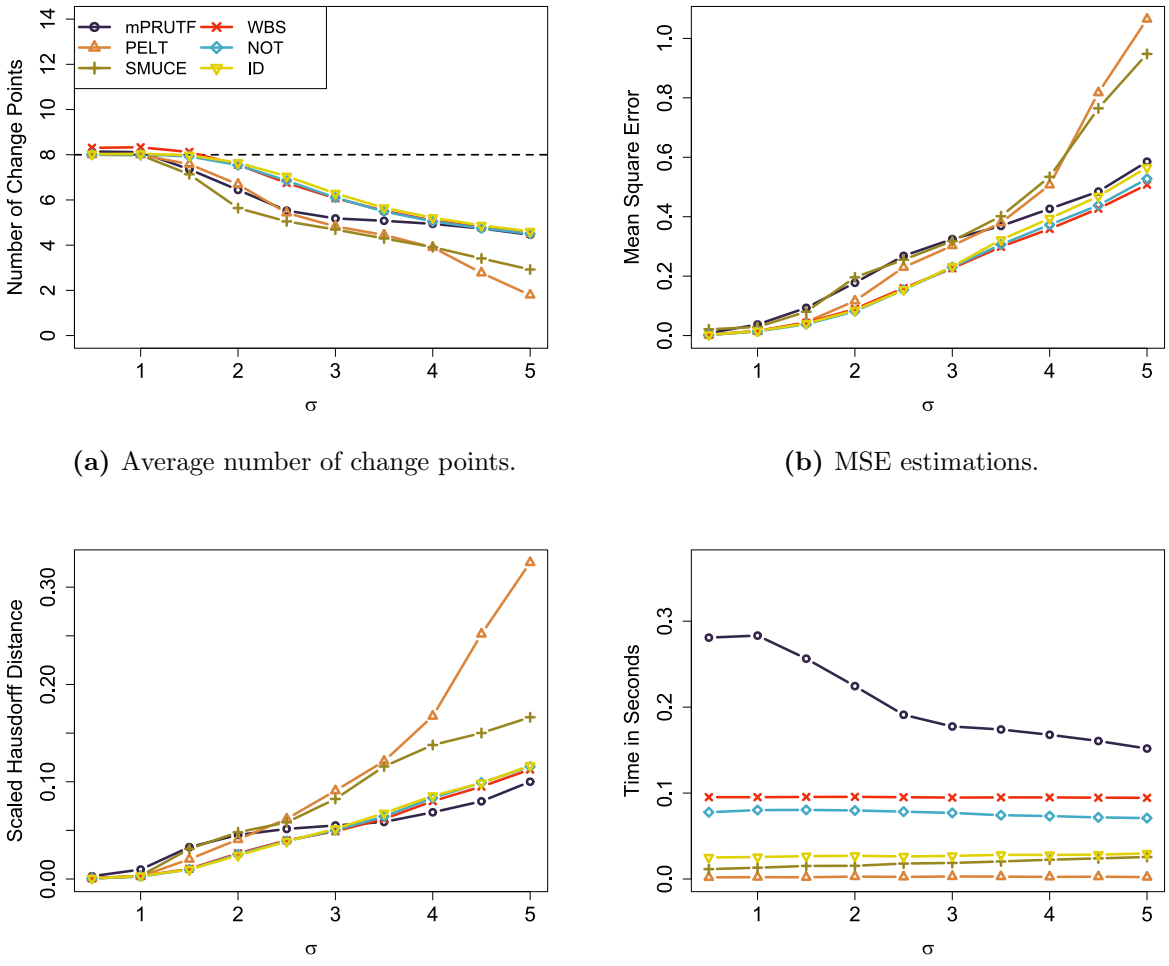
(b) PWL signal with staircases at blocks (512 , 768] and (1024 , 1152].

Figure 3.9: The piecewise constant (PWC) and piecewise linear (PWL) signals with the generated samples used in the simulation study.

and the computation time in seconds. These quantities are frequently used to assess the performance of a change point detection technique in the literature, for example, see [16], [4]. The signal estimate, $\widehat{f}$, is computed by the least square fit of a polynomial of order r to the observations within segments created by each change point method. We also remark that the tuning parameters and stopping criteria for the methods listed in Table 3.1 are set to the default values by the packages.

The results for the PWC and PWL signals are presented in Figures 3.10 and 3.11, respectively. In the case of piecewise constant signal, as in Figure 3.10, mPRUTF performs comparable to PELT and SMUCE in terms of the average number of change points, MSE and the scaled Hausdorff distance up to $\sigma = 3$, and outperforms them as σ increases. For $\sigma \geq 4$, similar performance to WBS, NOT and ID is viewed from these measurements. As indicated by the average number of change points, MSE and the scaled Hausdorff distance, WBS, NOT and ID outperform the other methods in almost all noise levels. From a computational point of view, mPRUTF takes a slightly longer time, mainly due to the matrix $\mathbf{D}$ multiplications, however, this computation time decreases as noise level σ increases. As in panel (d) of Figure 3.10, the methods PELT, SMUCE and ID are the fastest ones.

In the case of piecewise linear signal, mPRUTF is only compared to NOT and ID methods, which are applicable to the piecewise polynomials of order $r \geq 1$. As in Figure 3.11 mPRUTF outperforms both NOT and ID in terms of the average number of change



(a) Average number of change points.

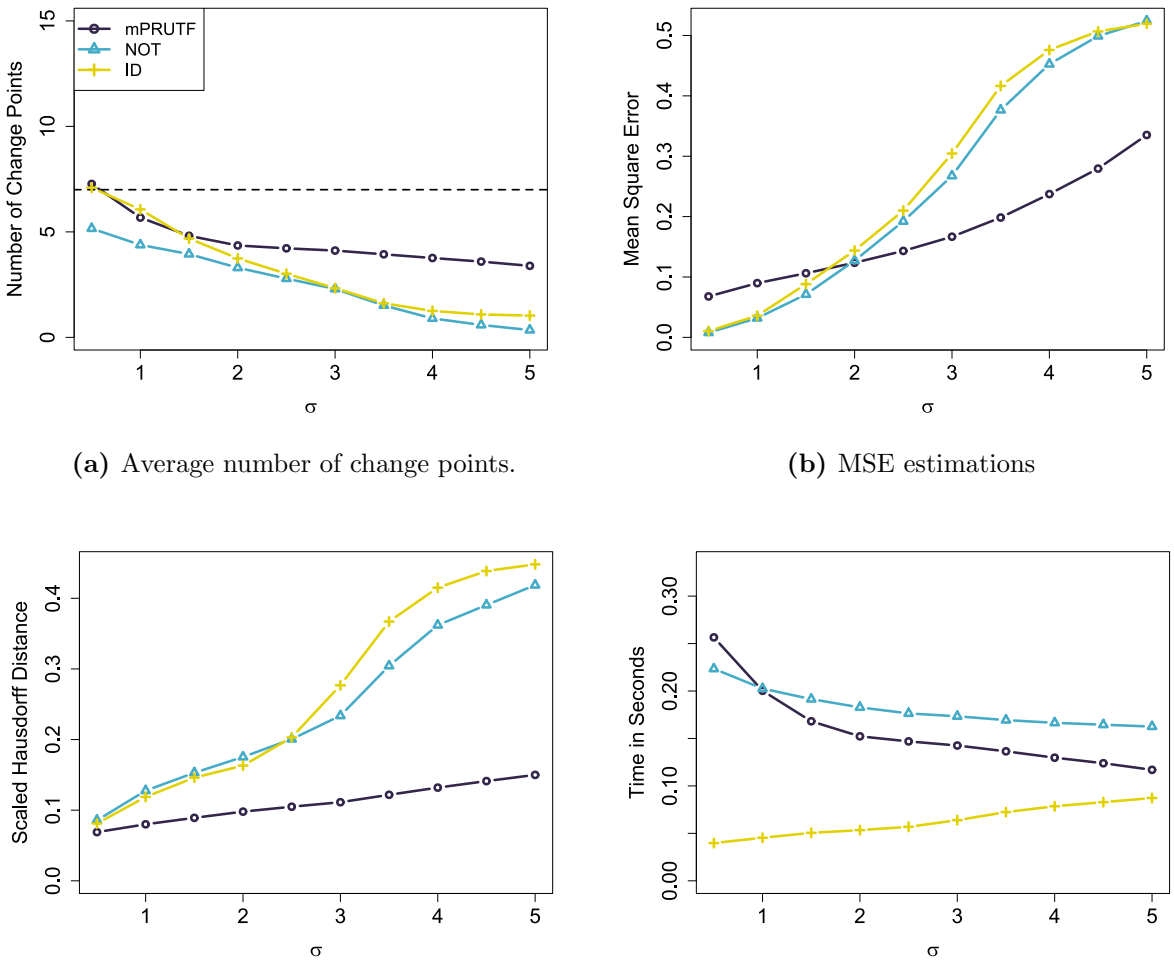
(b) MSE estimations.

(c) Scaled Hausdorff distance.

(d) Computation time.

Figure 3.10: The estimated average number of change points, MSE and Hausdorff distance, as well as the computation time of various methods for PWC signal. The results are provided for different values of the noise variability σ.

points and the scaled Hausdorff distance for all noise levels. In terms of MES, mPRUTF outperforms the other two for $\sigma \geq 2$. As shown in Panel (d) of Figure 3.11, the computation time of mPRUTF ranks second after ID.



(a) Average number of change points.

(b) MSE estimations

(c) Hausdorff distance.

(d) Computation time.

Figure 3.11: The estimated average number of change points, MSE and Hausdorff distance, as well as the computation time of various methods for PWL signal. The results are provided for different values of the noise variability σ.

The mPRUTF method performs well in terms of the estimation of the number of change points, their locations, as well as the true signals. In fact, simulation results for most of the scenarios indicate that mPRUTF is among the most competitive change point detection approaches in the literature.

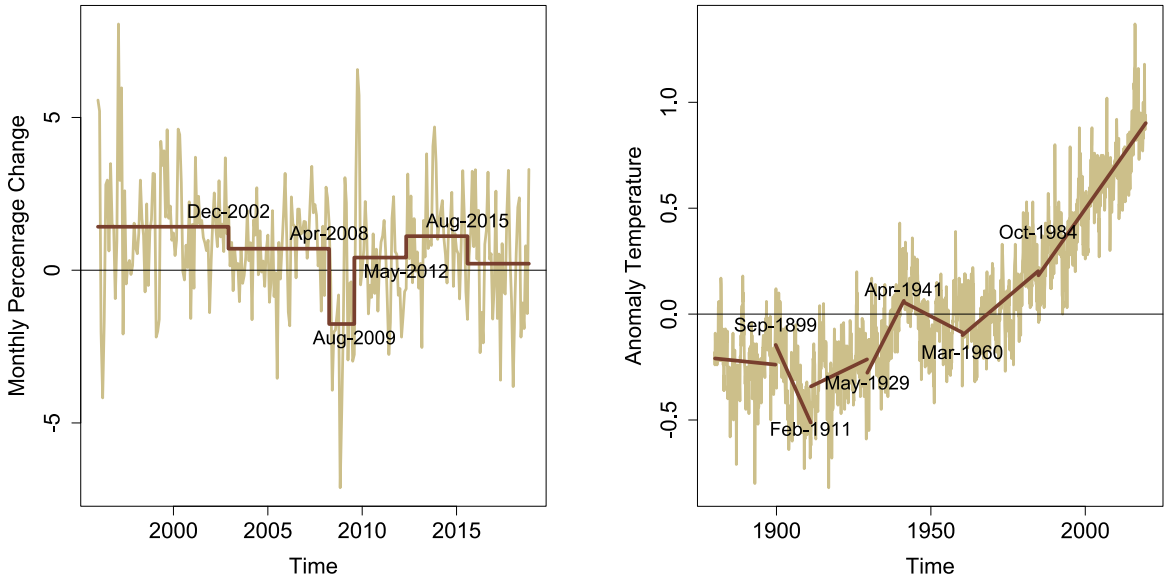
3.9.2 Real Data Analysis

In this section, we have analyzed UK HPI and GISTEMP and COVID-19 datasets, presented in Section 1.2, using our proposed algorithm. Because σ^2 is unknown for these real datasets, we applied median absolute deviation (MAD) proposed by [65], to robustly estimate σ^2. More specifically, a MAD estimate of σ for piecewise constant signals is given by $\widehat{\sigma} = \text{Median}\big(\mathbf{D}^{(1)}\,\mathbf{y}\big)/\big[\sqrt{2}\,\Phi^{-1}(0.75)\big]$ and for piecewise linear signals by $\widehat{\sigma} = \text{Median}\big(\mathbf{D}^{(2)}\,\mathbf{y}\big)/\big[\sqrt{6}\,\Phi^{-1}(0.75)\big]$, where $\Phi^{-1}(\cdot)$ represents the inverse cumulative density function of the standard normal distribution.

Example 3.17 (UK HPI Data) *Recall that the UK HPI dataset, discussed in Example 1.1, is monthly percentage changes in the UK HPI at Tower Hamlets from January 1996 to November 2018. We have applied the mPRUTF algorithm to the dataset. The algorithm have found five change points located at the dates December 2002, April 2008 and August 2009 (may be attributed to the Credit Crunch and Financial Crises), May 2012 (may be attributed to The London 2012 Summer Olympics) and August 2015 (may be attributed to regulatory and tax changes, and also by lower net migration from the EU). The dataset, the change points derived by mPRUTF and its piecewise constant fit are presented in panel (a) of Figure 3.12.*

Example 3.18 (GISTEMP Data) *The GISTEMP example, discussed in Example 1.2, considers the monthly land-ocean temperature anomalies recorded from January 1880 to August 2019. The plot reveals the presence of a linear trend with several potential change points in the dataset. For this dataset, we have identified six change points using mPRUTF located in September 1899, February 1911, May 1929, April 1941, March 1960, October 1984. The locations of change points and an estimate of the piecewise linear signal are presented in panel (b) of Figure 3.12.*

Example 3.19 (COVID-19 Data) *For the COVID-19 example, discussed in Example 1.3, we consider the log-scale of the cumulative number of confirmed cases for Australia,*



(a) UK HPI dataset and its piecewise constant fit. **(b)** GISTEMP dataset and its piecewise linear fit.

Figure 3.12: The time series and fitted signals for both Tower Hamlet HPI and GISTEMP datasets presented in examples

Canada, the United Kingdom and the United States, during the period March 10, 2020 through April 30, 2021. We have applied mPRUTF to detect change points that have occurred in the data for each country. We then fitted a piecewise linear model to the data using the selected change points, which provides a more direct perception of how the growth rate changes over time.

Figure 3.13 displays the locations of change points detected by the mPRUTF algorithm as well as the estimated linear trends for the four countries. For example, our algorithm has identified eight change points for Canada, on March 26, 2020; April 9, 2020; May 11, 2020; July 14, 2020; August 31, 2020; October 10, 2020; January 12, 2021 and March 18, 2021. The figure shows segments created by the estimated change points as well as their growth rate. The growth rate for the first segment (from March 10, 2020 to March 26, 2020) is remarkably high, but starts to slightly decline after the first change point on March 26, 2020. This mild decline may be linked to the declaration of the the state of emergency, quarantine and international travel ban declared by the Government of Canada. The third segment (from April 9, 2020 to May 11, 2020), the fourth segment (from May 11, 2020 to

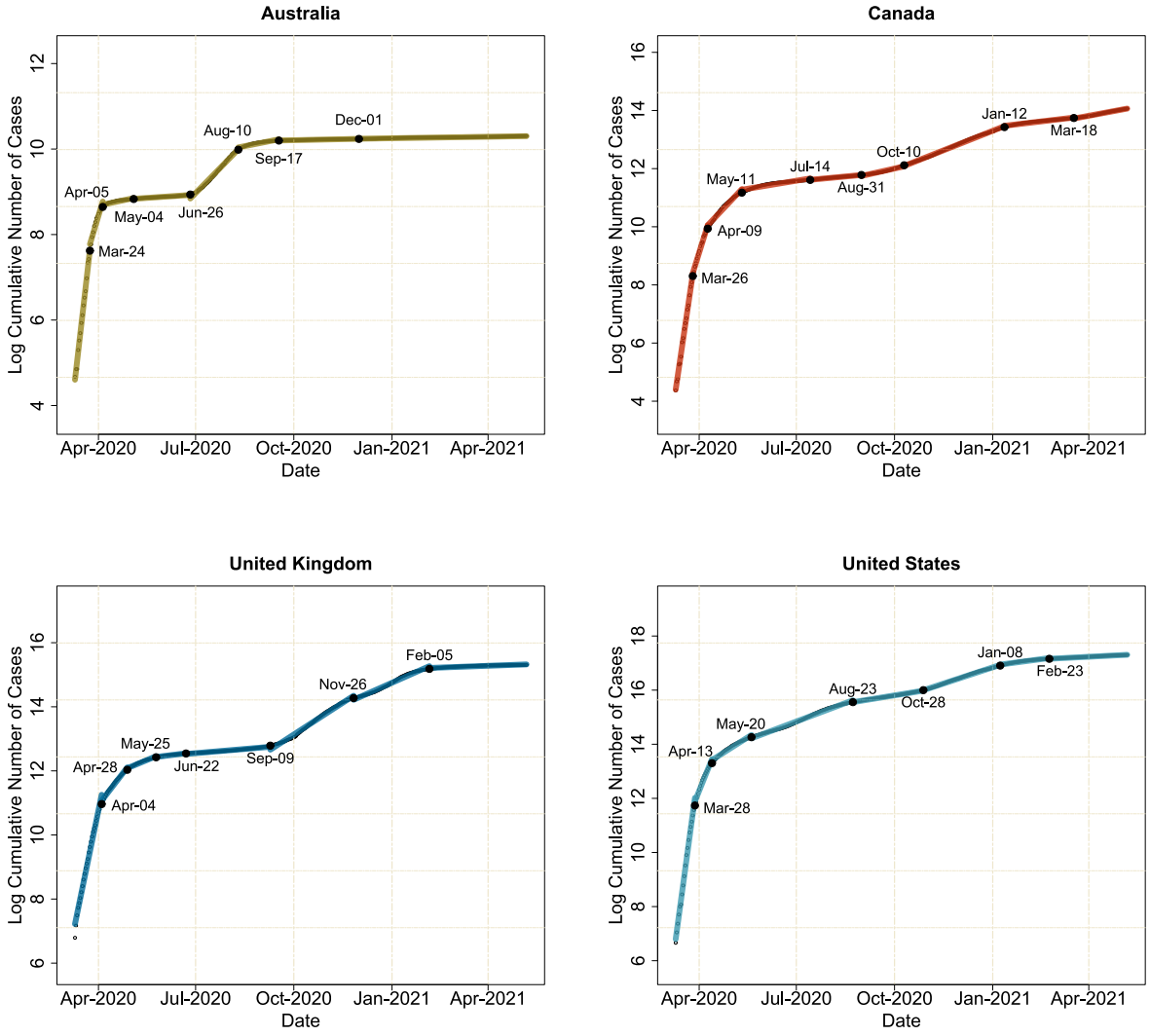


Figure 3.13: The change point locations and estimated linear trend for the transformed COVID-19 datasets in Example 3.19. The dates indicated on each plot show the detected change points.

July 14, 2020) and the fifth segment (from July 14, 2020 to August 31, 2020) have witnessed noticeable decreases in the growth rate. The decrease can perhaps/probably be explained by the mandatory use of face-coverings and the border closure with the United States for the third segment, and the use of COVID-19 serological tests and the national contact tracing for the fourth and fifth segments. An upward trend in the growth rate observed from August 31, 2020 to October 10, 2020 could have resulted from the opening of businesses and public

spaces. It seems that the second wave started on October 10, 2020, with a remarkable increase in the rate that continued until January 12, 2021. After this date, the rate again declined until March 18, 2021, which could be the result of provincial states of emergency and lockdowns. The last segment witnessed another surge in the rate, perhaps due to new variants of Coronavirus.

The mPRUTF algorithm has also detected seven change points for the United Kingdom on the following dates: April 4, 2020; April 28, 2020; May 25, 2020; June 22, 2020; September 9, 2020; November 26, 2020 and February 5, 2021. As can be viewed from the figure, there are remarkable declines in the growth rates for the second segment (perhaps due to the nationwide lockdown), the third segment (perhaps due to the international travel ban) and the segments from May 25, 2020 to September 9, 2020 (perhaps due to mandatory use of face masks and comprehensive contact tracing). The country witnessed a significant increase in the growth rate starting from September 9, 2020, which aligns with the reopening of businesses, schools and universities. The second national lockdown could be linked to the very small decrease in the slope of the segment from November 26, 2020 to February 5, 2021. Finally, the growth rate in the last segment seemed to be under control, which could be the result of COVID vaccinations.

3.10 More on Models With Frequent Change Points or With Dependent Errors

This section empirically investigates the performance of mPRUTF in models with frequent change points as well as models with dependent random errors.

3.10.1 mPRUTF in Signals With Frequent Change Points

In order to evaluate the detection power of mPRUTF in signals with frequent change points, we employ a teeth signal for the piecewise constant case and a wave signal for the piecewise linear case. For the teeth signal, we consider a signal with 29 change points and varying segment lengths defined as follows:

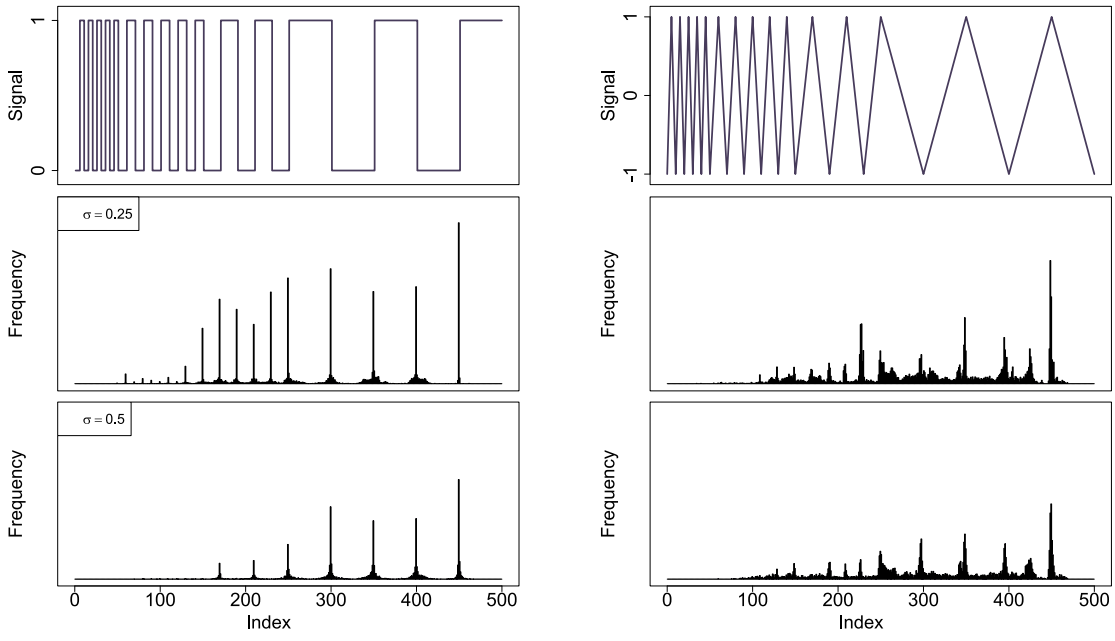
- for $1 \leq t \leq 50$, $f_t = 0$ if $(t \bmod 10) \in \{1, \ldots, 5\}$; $f_t = 1$, otherwise,
- for $51 \leq t \leq 150$, $f_t = 0$ if $(t \bmod 20) \in \{1, \ldots, 10\}$; $f_t = 1$, otherwise,
- for $151 \leq t \leq 250$, $f_t = 0$ if $(t \bmod 40) \in \{1, \ldots, 20\}$; $f_t = 1$, otherwise,

- for $251 \leq t \leq 500$, $f_t = 0$ if $(t \bmod 100) \in \{1, \ldots, 50\}$; $f_t = 1$, otherwise.

The signal is displayed in the top-left panel of Figure 3.14. The wave signal also has 29 change points with varying slopes which is defined as follows:

- for $1 \leq t \leq 50$, $f_t = -1 + 0.4\,t$ if $(t \bmod 10) \in \{1, \ldots, 5\}$; $f_t = 1 - 0.4\,t$, otherwise,
- for $51 \leq t \leq 150$, $f_t = -1 + 0.2\,t$ if $(t \bmod 20) \in \{1, \ldots, 10\}$; $f_t = 1 - 0.2\,t$, otherwise,
- for $151 \leq t \leq 250$, $f_t = -1 + 0.1\,t$ if $(t \bmod 40) \in \{1, \ldots, 20\}$; $f_t = 1 - 0.1\,t$, otherwise,
- for $251 \leq t \leq 500$, $f_t = -1 + 0.04\,t$ if $(t \bmod 100) \in \{1, \ldots, 50\}$; $f_t = 1 - 0.04\,t$, otherwise.

The top-right panel of Figure 3.14 shows this signal.



(a) Teeth signal **(b)** Wave signal

Figure 3.14: Histograms of the locations of change points for the teeth and wave signals. The histograms show the frequencies of the change points detected using mPRUTF in both signals. The result are displayed for two different σ values.

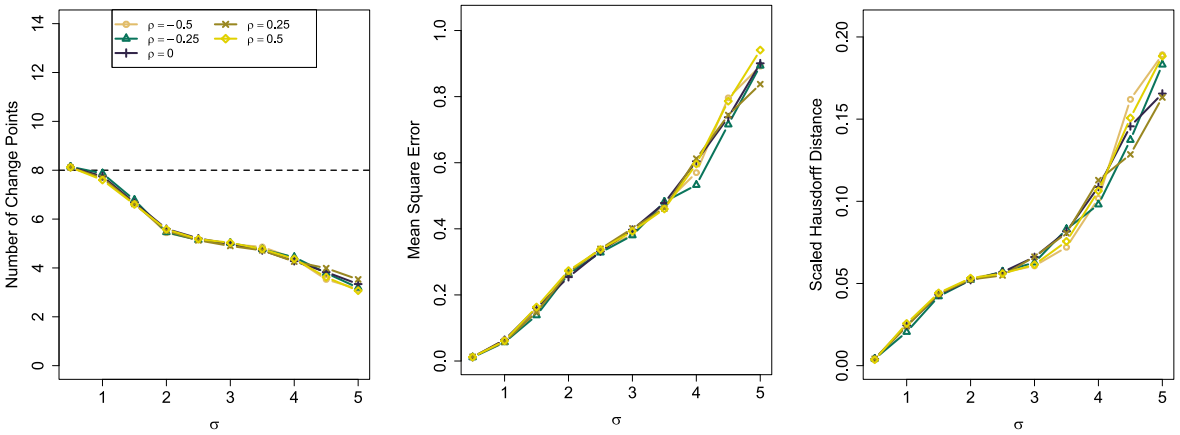
We generated 1000 independent samples of y_t in model (3.1) with $\varepsilon_t \overset{\text{i.i.d}}{\sim} N(0\,,\,\sigma^2)$ for both signals. The mPRUTF algorithm was then applied to these samples to estimate their change point locations. Figure 3.14 shows the histograms of the locations of these change points for the signals. The figure provides evidence that mPRUTF is unable to effectively detect change points in signals with frequent change points and short segments. It also shows that the results deteriorate when the noise variance σ^2 or the polynomial order r increase.

It turns out that the success of the mPRUTF algorithm critically relies on its stopping rule. Equation (3.35) verifies that estimating the noise variance σ^2 and specifying the threshold x_α from a Gaussian bridge process play crucial roles in the stopping rule. As discussed in [54], the two widely used robust estimators of σ, Mean Absolute Deviation (MAD) (used here) and Inter-Quartile Range (IQR), overestimate σ in frequent change point scenarios. In addition, determining the accurate value of the threshold x_α using (3.36) is affected in such scenarios. These two factors prevent the stopping rule from being effective in the mPRUTF algorithm and lead to the underestimation of change points for these scenarios. We must note that such poor performance in frequent change point scenarios is not specific to mPRUTF. As investigated in [54], state-of-the-art methods such as PELT, WBS, MOSUM, SMUCE and FDRSeg are among the approaches that fail in such scenarios.

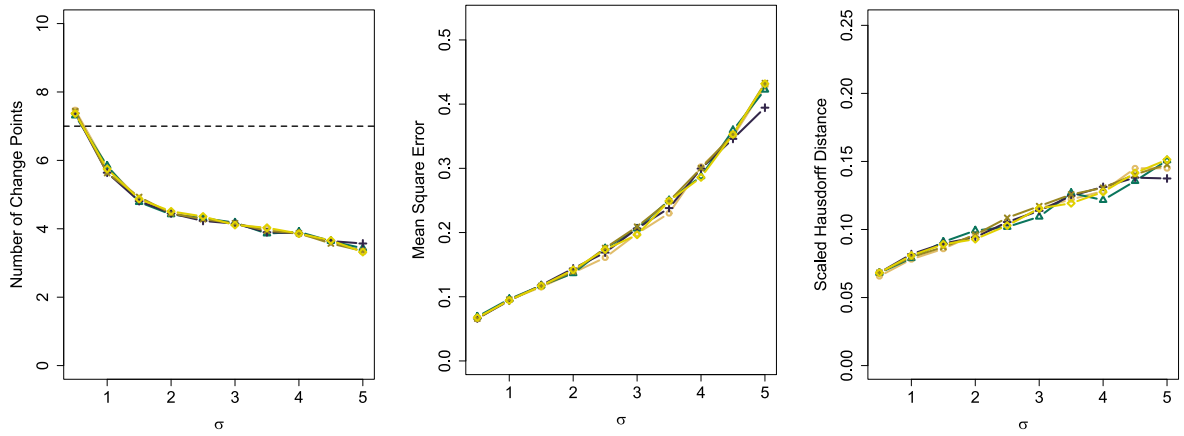
3.10.2 mPRUTF in Models With Dependent Error Terms

How can mPRUTF's performance be affected by various types of random errors, such as non-Gaussian or dependent errors? For example, having independent random error assumption for the three real datasets, analyzed in Section 3.9.2, may be violated. We explored this assumption by analyzing their residuals in Appendix A.4. This is of course an important question and will be the topic of future works. Notice that the dual solution path of trend filtering is not impacted by the type of random errors. However, the type of random errors plays a key role in the stopping rule of mPRUTF because the stopping rule is built based on Gaussian bridge processes established by Donsker's Theorem.

To empirically investigate the performance of mPRUTF for weakly dependent random errors, a simulation study is carried out here. To this end, we generate $N = 5000$ samples from model (3.1) with the PWC and PWL signals. We consider errors ε_i from an $AR(1)$ model with $\varepsilon_i = \rho\,\varepsilon_{i-1} + e_i$, for $i = 1, \ldots, n$. Note that for PWC signal $n = 2024$ and for PWL signal $n = 1408$. Here, e_i's are independent and identical random errors drawn from $N\big(0\,,\,(1-\rho^2)\,\sigma^2\big)$ with $\rho \in \{-0.5,\, -0.25,\, 0,\, 0.25,\, 0.5\}$ and $\sigma \in \big\{0.5,\, 1,\, 1.5,\, \ldots,\, 4.5,\, 5\big\}$.



(a) PWC signal



(b) PWL signal

Figure 3.15: The estimated average number of change points, MSEs and Hausdorff distances of various methods for both PWC and PWL signals. The results are based on weakly dependent observations and provided for various values of the error variability σ.

The results of mPRUTF for both PWC and PWL signals are provided in Figure 3.15. As can be seen, the results are very similar, in terms of the average number of change points, MSEs and the scaled Hausdorff distances, for various values of ρ. Therefore, it appears that the mPRUTF algorithm is quite robust against dependent error terms. Extensive

studies of mPRUTF for non-Gaussian and dependent random errors will be carried out in future research.

Chapter 4

Valid Post-Detection Inference for Change Points Identified Using PRUTF

There are many research works and methods about change point detection in the literature. However, there are only a few that provide inference for such change points after being estimated. This chapter mainly focuses on statistical analyses of change points estimated by the PRUTF algorithm, which incorporates trend filtering to determine change points in piecewise polynomial signals. We develop a methodology to perform statistical inference, such as computing p-values and constructing confidence intervals in the newly developed post-selection inference framework. Our work concerns both cases of known and unknown error variance. As pointed out in the post-selection inference literature, the length of such confidence intervals are undesirably long. To resolve this shortcoming, we also provide two novel methods, *global post-detection* and *local post-detection* inferences, which are based on the intrinsic properties of change points. We run our proposed methods on real-world as well as simulated datasets to evaluate their performances.

4.1 Introduction

As previously discussed, change point detection seeks the locations of changes in the distribution of a signal for which an ordered sequence of observations is available. There is a vast and rich literature on change-point detection in statistical research. Although there

is a huge body of work on canonical change point framework, a more general case in which signal f is modelled as a piecewise polynomial has attracted less attention. A few works, mainly focused on a piecewise linear model, exist in the literature; see [13], [105], [16], [4], and [170]. Recently, [108] have introduced the PRUTF method, which is designed for change point detection in piecewise polynomial signals. PRUTF exploits the trend filtering problem [151] and provides a new algorithm to estimate change points.

After analysts perform change point estimation, genuine interest is to make inferences about the uncertainty of the estimated points. Many research works such as [30], [47], and [74] have discussed conducting such inferences. However, they have ignored the data-driven nature of the detected change points and regarded them as fixed locations. It is important to note that this data-dependent property of detected change points invalidates the inferences mentioned earlier. There is a new and rapidly growing framework that provides tools for conducting inference after a selection procedure. This type of inference is called *post-selection inference* and has been mainly developed for inference after variable selection in high-dimensional regression. See for example [22, 93, 153, 51]. This chapter adopts this body of work for inference after change point detection in piecewise polynomial signals.

Interest in post-selection inference research ignited after the work of [122]. Later, in a sequence of articles, [95, 96, 97] explored the estimation of the post-selection distribution. This subject has attracted much attention more recently as applications for model selection approaches have proliferated. [22] performed valid and conservative post-selection inference by considering all possible procedures that produce a selected model. [102] suggested an asymptotic test procedure to test whether a nonzero estimated coefficient in the lasso regression coincides with the true nonzero coefficient. [93] developed an exact test based on truncated Gaussian distribution for the solution path of lasso at a fixed value of the regularization parameter λ. Concurrently, a similar test was proposed by [153] for the lasso, LARs and forward stepwise regression with a fixed number of steps in the solution path. [51] studied the theoretical properties of post-selection inference and generalized the framework to the broad class of the exponential family of distributions. A number of extensions to different frameworks and applications are given in [103], [145], [152], [101], and [12].

In a post-selection inference setting, since data have been used to fit a model, a conditional inference is required to restrict the sample space to unused data. This conditional approach prevents an inference procedure from using data twice (once for selection and once for inference). As a result of this conditioning, the post-detection distribution changes from Gaussian to truncated Gaussian [93, 153], reflecting the restriction on the sample space. However, it turns out that this truncation is often very severe and leads statistical tests

to lose power and their respective confidence intervals to become undesirably wide; see Section 4.6. These problems occur because most of the data are employed to select the model. Consequently, there is insufficient information to use for drawing conclusions after the selection procedure.

The application of post-selection inference in the change points context, referred to as post-detection inference, was first addressed in [77]. Their work, which is our main inspiration, applies a truncated Gaussian test for a given and selected number of steps in the solution path of the generalized lasso. Similarly, [78] studied post-detection inference for some popular change point detection methodologies such as Binary Segmentation, Circular Binary Segmentation, and Wild Binary Segmentation.

In this chapter, we study the problem of conducting valid statistical inferences for change points detected by the PRUTF algorithm introduced in Chapter 3. At the core of our framework, an important result states that the set of change points detected by PRUTF constitutes a polyhedron (a convex cone). We apply the post-selection inference framework to compute valid post-detection p-values as well as post-detection confidence intervals.

We make the following contributions in this chapter.

- One fundamental aspect of our contribution is establishing that the set of change points identified using the PRUTF algorithm characterizes a polyhedron set in $\mathbf{y}$. This characterization allows us to use the post-selection inference methodology for conducting statistical inference after change point detection. To the best of our knowledge, the inference procedures for the significance of detected change points developed in this chapter are the first of their kind for piecewise polynomial signals. Chapter 3 also developed a stopping criterion for selecting the number of required steps for the PRUTF algorithm, which uses the Gaussian bridge property of dual variables. We will show that this criterion also forms a polyhedron and, consequently, applies to the post-detection inference framework.

- The implication of post-detection inference for the Gaussian model in (3.1) leads us to propose two test statistics, one for known σ^2 and another for unknown σ^2. A significant feature of these test statistics is that their exact and finite sample distributions under the null hypothesis are $U\big(0,\,1\big)$. These test statistics allow us to conduct inference for the significance of change points estimated by PRUTF.

- We inspect the produced p-values and confidence intervals of detected change points using two test statistics for both known and unknown noise variance cases. We

show that the confidence intervals derived by these approaches are unacceptably wide. This behaviour also leads to a loss in the power of hypothesis tests. To resolve these shortcomings, we introduce two new methods of conditioning for post-detection inference. We call the first method *global post-detection* that focuses only on the target change point and makes inferences regardless of the other change points. We call the second method *local post-detection* which takes into account detected locations that are the most relevant to the target change point.

- We conduct a comprehensive simulation study to investigate the performance of the proposed procedures in terms of the power of post-detection tests and the length of their confidence intervals. We also demonstrate applications of our algorithms to several real datasets.

The rest of this chapter is structured as follows. We first review conducting post-selection inference based on the polyhedron characterization of a selection procedure. In Section 4.3, we explain how to capture the representation of the corresponding polyhedron of the PRUTF algorithm. This representation also determines cases where the number of steps for the algorithm is adaptively selected using the stopping rule. Next, two different approaches are established in Section 4.4 to conduct valid post-detection inference when the error variance is assumed both known and unknown. In Section 4.5, we discuss the shortcomings of the post-selection inference, conditional on the entire selected model and its signs, such as reduction in the power of tests and wide confidence intervals. We then propose two strategies in Section 4.6 to resolve these shortcomings. Section 4.7 provides some numerical investigation, including real-world data analyses and simulation studies.

4.2 Post-Selection Inference With Polyhedron Selection Procedures

In this section, we review some key concepts of post-selection inference in a linear regression setting. For details about a broader class of models, see [51]. Suppose that the observations $\mathbf{y} \in \mathbb{R}^n$ follow a Gaussian regression model with either known or unknown variance. Also, let $\mathcal{M}$ be a finite collection of all possible models $\mathbf{M}$ obtained from a model selection procedure, that is, $\mathcal{M} = \big\{\mathbf{M} : \mathbf{M} \subseteq \{1, \ldots, p\}\big\}$, for p to be the number of variables. The goal is, therefore, to carry out statistical inference for a selected model $\widehat{\mathbf{M}}(\mathbf{y}) = \mathbf{M}$. Since we adaptively choose $\widehat{\mathbf{M}}(\mathbf{y})$ using the data, for inference, it is natural to consider the conditional distribution given $\widehat{\mathbf{M}}(\mathbf{y})$. This conditional distribution is called post-selection

distribution. For testing null hypothesis H_0, we seek to control the post-selection type-I error, defined as $\Pr_{H_0}\left(\text{Reject } H_0 \,\middle|\, \widehat{\mathbf{M}}(\mathbf{y}) \in \mathcal{M}\right)$, at the nominal level α. By analogy to the classical theory, a post-selection confidence interval can then be built by inverting the associated post-selection hypothesis test.

Throughout this chapter, we restrict our attention to a specific class of selection procedures known as *polyhedron (affine) selection procedures*. Again, suppose a variable selection approach picks a model $\widehat{\mathbf{M}}(\mathbf{y})$ from a finite collection of models $\mathcal{M}$ using the data $\mathbf{y}$. This selection approach is called a polyhedron selection procedure if any $\widehat{\mathbf{M}}(\mathbf{y}) \in \mathcal{M}$ can be characterized as a polyhedron set in respect to $\mathbf{y}$. In other words, any selected model $\widehat{\mathbf{M}}(\mathbf{y})$ chosen by the polyhedron selection procedure can be written in the form of

$$\left\{\mathbf{y} : \ \mathbf{A}\mathbf{y} \geq \mathbf{q}\right\}, \tag{4.1}$$

where the matrix $\mathbf{A} \in \mathbb{R}^{p\times n}$ and the vector $\mathbf{q} \in \mathbb{R}^p$ are dependent on the model $\widehat{\mathbf{M}}(\mathbf{y})$. Observe that the inequality in (4.1) is interpreted componentwise. Selection approaches using the ℓ_1-penalized generalized linear model, including lasso [93, 153], LARS [142] and generalized lasso [77], are examples of such polyhedron selection procedures.

4.3 Construction of Polyhedron

In this section, we describe how the change points detection procedure using PRUTF can be represented as a polyhedron. This representation enables us to state our post-detection inference conditional on a polyhedron. First, consider the change point model

$$y_i = f_i + \varepsilon_i, \qquad\qquad i = 1, \ldots, n\,, \tag{4.2}$$

where f is a piecewise polynomial signal of order r with J_0 change points. We also assume that the error terms $\varepsilon_i,\ i = 1 \ldots, n$, have identical and independent Gaussian distribution with mean of zero and finite variance σ^2 (either known or unknown). Now, suppose $\left\{\widehat{\boldsymbol{\tau}}_j\,, \widehat{\mathbf{s}}_j\right\}$ is the set of detected change points and their signs at step j of the PRUTF algorithm over $\mathbf{y} = \left(y_1, \ldots, y_n\right)$. Below, we describe that $\left\{\widehat{\boldsymbol{\tau}}_j = \widehat{\boldsymbol{\tau}}_j(\mathbf{y})\,, \widehat{\mathbf{s}}_j = \widehat{\mathbf{s}}_j(\mathbf{y})\right\}$ as a function of $\mathbf{y}$ is indeed a polyhedron. Thus, formally, for a matrix $\mathbf{A}$ and vector $\mathbf{q}$,

$$\left\{\mathbf{y} : \ \widehat{\boldsymbol{\tau}}_j = \widehat{\boldsymbol{\tau}}_j(\mathbf{y})\,, \widehat{\mathbf{s}}_j = \widehat{\mathbf{s}}_j(\mathbf{y})\right\} = \left\{\mathbf{y} : \ \mathbf{A}\mathbf{y} \geq \mathbf{q}\right\}. \tag{4.3}$$

From now on, we call $\mathbf{A}$ the polyhedron matrix. Observe that, in general, $\widehat{\boldsymbol{\tau}}_j$ does not necessarily contain j entries since (except in the case of $r = 0$) the dual solution path of PRUTF can either add a change point to or remove it from the augmented boundary set at each step, see Remark 3.3.

4.3.1 Construction of Polyhedron Along the Solution Path

In the following, the construction process of the matrix $\mathbf{A}$ and vector $\mathbf{q}$ associated with the polyhedron along the dual solution path of PRUTF is provided. We present this construction in steps according to the steps of PRUTF in Algorithm 3.1.

1. For $j = 1$, the conditions for deriving $\left\{\widehat{\tau}_1\,,\,\widehat{s}_1\right\}$ can be rewritten as

$$\widehat{s}_1 \left[\left(\mathbf{D}\mathbf{D}^T\right)^{-1}\mathbf{D}\right]_{\widehat{\tau}_1}\mathbf{y} \;\geq\; \pm\left[\left(\mathbf{D}\mathbf{D}^T\right)^{-1}\mathbf{D}\right]_i\mathbf{y}, \tag{4.4}$$

   for any $i \neq \widehat{\tau}_1$. This implies that the polyhedron matrix after the first step, denoted by $\mathbf{A}_1$, has $2\,(m-1)$ rows, formed by $\widehat{s}_1 \left[\left(\mathbf{D}\mathbf{D}^T\right)^{-1}\mathbf{D}\right]_{\widehat{\tau}_1} \pm \left[\left(\mathbf{D}\mathbf{D}^T\right)^{-1}\mathbf{D}\right]_i$, for any $i \neq \widehat{\tau}_1$.

2. For $j = 1\,,\,2,\,\ldots$, assume that $\mathbf{A}_j$ is the polyhedron matrix associated with $\left\{\widehat{\boldsymbol{\tau}}_j\,,\,\widehat{\mathbf{s}}_j\right\}$ at step j of the PRUTF algorithm. In order to construct the polyhedron matrix for the set $\left\{\widehat{\boldsymbol{\tau}}_{j+1}\,,\,\widehat{\mathbf{s}}_{j+1}\right\}$, a number of rows will be appended to $\mathbf{A}_j$ according to Step 2 of Algorithm 3.1. Recall that part (a) of Step 2 corresponds to the specification of the joining coordinates and their signs. Therefore, specifying $\widehat{\boldsymbol{\tau}}_{j+1}$ is equivalent to satisfying the conditions

$$\operatorname{sign}(a_i)\left[\left(\mathbf{D}_{-\mathcal{A}_j}\mathbf{D}_{-\mathcal{A}_j}^T\right)^{-1}\mathbf{D}_{-\mathcal{A}_j}\right]_i\mathbf{y} \;\geq\; 0, \qquad\qquad \forall\, i \notin \mathcal{A}_j, \tag{4.5}$$

   and

$$\frac{\left[\left(\mathbf{D}_{-\mathcal{A}_j}\mathbf{D}_{-\mathcal{A}_j}^T\right)^{-1}\mathbf{D}_{-\mathcal{A}_j}\right]_{\widehat{\tau}_{j+1}^{\text{join}}}\mathbf{y}}{\widehat{s}_{j+1}^{\text{join}} + \left[\left(\mathbf{D}_{-\mathcal{A}_j}\mathbf{D}_{-\mathcal{A}_j}^T\right)^{-1}\mathbf{D}_{-\mathcal{A}_j}\right]_{\widehat{\tau}_{j+1}^{\text{join}}}\mathbf{D}_{\mathcal{A}_j}^T\widehat{\mathbf{s}}_{\mathcal{A}_j}} \;\geq\; \frac{\left[\left(\mathbf{D}_{-\mathcal{A}_j}\mathbf{D}_{-\mathcal{A}_j}^T\right)^{-1}\mathbf{D}_{-\mathcal{A}_j}\right]_i\mathbf{y}}{\operatorname{sign}(a_i) + \left[\left(\mathbf{D}_{-\mathcal{A}_j}\mathbf{D}_{-\mathcal{A}_j}^T\right)^{-1}\mathbf{D}_{-\mathcal{A}_j}\right]_i\mathbf{D}_{\mathcal{A}_j}^T\widehat{\mathbf{s}}_{\mathcal{A}_j}}, \tag{4.6}$$

   for any $i \notin \mathcal{A}_j \bigcup \{\,\widehat{\tau}_{j+1}^{\text{join}}\,\}$. The vector $\boldsymbol{a}$ in the aforementioned equations is given in (3.12). The above inequalities will add $2\left(m - |\mathcal{A}_j|\right) - 1$ rows to the matrix $\mathbf{A}_j$.

   Part (b) of Step 2 explains conditions for the determination of the leaving coordinates. First, rows corresponding to conditions $c_i < 0$ and $d_i < 0$ for $i \in \mathcal{B}_j$ are added to $\mathbf{A}_j$. To this end, we disregard those $i \in \mathcal{B}_j$ which $d_i \geq 0$ and partition the remaining entries into two groups. The first group called viable leaving coordinates includes those which $c_i < 0$ and the second group as its complement. Denoting $\mathbf{L}_{j+1}$ to be the collection of viable leaving coordinates, we consider

$$\mathbf{L}_{j+1} = \Big\{\, i \in \mathcal{B}_j :\; d_i < 0 \quad \text{and} \quad c_i < 0 \,\Big\}, \tag{4.7}$$

with the complementary set

$$\mathbf{L}_{j+1}^{c} = \Big\{ i \in \mathcal{B}_j : \; d_i < 0 \quad \text{and} \quad c_i \geq 0 \Big\}. \tag{4.8}$$

Since vector $\mathbf{d}$, defined in (3.14), does not depend on $\mathbf{y}$, therefore, given $d_i < 0$, conditions $c_i < 0$ and $c_i \geq 0$ can be expressed as

$$\Big[\mathrm{diag}\big(\widehat{\mathbf{s}}_{\mathcal{B}_j}\big)\, \mathbf{D}_{\mathcal{B}_j} \Big(\mathbf{I} - \mathbf{D}_{-\mathcal{A}_j}^T \big(\mathbf{D}_{-\mathcal{A}_j} \mathbf{D}_{-\mathcal{A}_j}^T\big)^{-1} \mathbf{D}_{-\mathcal{A}_j}\Big)\Big]_i \mathbf{y} \;\leq\; 0\,, \qquad \text{for} \quad i \in \mathbf{L}_{j+1},$$

$$\Big[\mathrm{diag}\big(\widehat{\mathbf{s}}_{\mathcal{B}_j}\big)\, \mathbf{D}_{\mathcal{B}_j} \Big(\mathbf{I} - \mathbf{D}_{-\mathcal{A}_j}^T \big(\mathbf{D}_{-\mathcal{A}_j} \mathbf{D}_{-\mathcal{A}_j}^T\big)^{-1} \mathbf{D}_{-\mathcal{A}_j}\Big)\Big]_i \mathbf{y} \;\geq\; 0\,, \qquad \text{for} \quad i \in \mathbf{L}_{j+1}^{c}, \tag{4.9}$$

respectively. The above equations correspond to $|\mathbf{L}_{j+1}| + |\mathbf{L}_{j+1}^{c}|$ rows of $\mathbf{A}_{j+1}$. Second, the condition associated with the specification of the pair $\big(\widehat{\tau}_{j+1}^{\,\text{leave}}, \widehat{s}_{j+1}^{\,\text{leave}}\big)$ must be added to the rows of $\mathbf{A}_j$. This condition can be captured for any $i \in \mathbf{L}_{j+1} \setminus \{\widehat{\tau}_{j+1}^{\,\text{leave}}\}$ by the inequality

$$\frac{\Big[\mathrm{diag}(\widehat{\mathbf{s}}_{\mathcal{B}_j})\mathbf{D}_{\mathcal{B}_j} \Big(\mathbf{I} - \mathbf{D}_{-\mathcal{A}_j}^T \big(\mathbf{D}_{-\mathcal{A}_j} \mathbf{D}_{-\mathcal{A}_j}^T\big)^{-1} \mathbf{D}_{-\mathcal{A}_j}\Big)\Big]_{\widehat{\tau}_{j+1}^{\,\text{leave}}} \mathbf{y}}{\Big[\mathrm{diag}(\widehat{\mathbf{s}}_{\mathcal{B}_j})\mathbf{D}_{\mathcal{B}_j} \Big(\mathbf{I} - \mathbf{D}_{-\mathcal{A}_j}^T \big(\mathbf{D}_{-\mathcal{A}_j} \mathbf{D}_{-\mathcal{A}_j}^T\big)^{-1} \mathbf{D}_{-\mathcal{A}_j}\Big)\Big]_{\widehat{\tau}_{j+1}^{\,\text{leave}}} \mathbf{D}_{\mathcal{A}_j}^T \widehat{\mathbf{s}}_{\mathcal{A}_j}} \;\geq$$

$$\frac{\Big[\mathrm{diag}(\widehat{\mathbf{s}}_{\mathcal{B}_j})\mathbf{D}_{\mathcal{B}_j} \Big(\mathbf{I} - \mathbf{D}_{-\mathcal{A}_j}^T \big(\mathbf{D}_{-\mathcal{A}_j} \mathbf{D}_{-\mathcal{A}_j}^T\big)^{-1} \mathbf{D}_{-\mathcal{A}_j}\Big)\Big]_i \mathbf{y}}{\Big[\mathrm{diag}(\widehat{\mathbf{s}}_{\mathcal{B}_j})\mathbf{D}_{\mathcal{B}_j} \Big(\mathbf{I} - \mathbf{D}_{-\mathcal{A}_j}^T \big(\mathbf{D}_{-\mathcal{A}_j} \mathbf{D}_{-\mathcal{A}_j}^T\big)^{-1} \mathbf{D}_{-\mathcal{A}_j}\Big)\Big]_i \mathbf{D}_{\mathcal{A}_j}^T \widehat{\mathbf{s}}_{\mathcal{A}_j}}, \tag{4.10}$$

which forms $|\mathbf{L}_{j+1}| - 1$ rows of $\mathbf{A}_{j+1}$.

Lastly, the decision of adding or removing a coordinate in part (c) is required. The condition for the joining time is $\lambda_{j+1}^{\text{join}} \geq \lambda_{j+1}^{\text{leave}}$, which can be expressed as

$$\frac{\Big[\big(\mathbf{D}_{-\mathcal{A}_j} \mathbf{D}_{-\mathcal{A}_j}^T\big)^{-1} \mathbf{D}_{-\mathcal{A}_j}\Big]_{\widehat{\tau}_{j+1}^{\,\text{join}}} \mathbf{y}}{\widehat{s}_{j+1}^{\,\text{join}} + \Big[\big(\mathbf{D}_{-\mathcal{A}_j} \mathbf{D}_{-\mathcal{A}_j}^T\big)^{-1} \mathbf{D}_{-\mathcal{A}_j}\Big]_{\widehat{\tau}_{j+1}^{\,\text{join}}} \mathbf{D}_{\mathcal{A}_j}^T \widehat{\mathbf{s}}_{\mathcal{A}_j}} \geq$$

$$\frac{\Big[\mathrm{diag}(\widehat{\mathbf{s}}_{\mathcal{B}_j})\mathbf{D}_{\mathcal{B}_j} \Big(\mathbf{I} - \mathbf{D}_{-\mathcal{A}_j}^T \big(\mathbf{D}_{-\mathcal{A}_j} \mathbf{D}_{-\mathcal{A}_j}^T\big)^{-1} \mathbf{D}_{-\mathcal{A}_j}\Big)\Big]_{\widehat{\tau}_{j+1}^{\,\text{leave}}} \mathbf{y}}{\Big[\mathrm{diag}(\widehat{\mathbf{s}}_{\mathcal{B}_j})\mathbf{D}_{\mathcal{B}_j} \Big(\mathbf{I} - \mathbf{D}_{-\mathcal{A}_j}^T \big(\mathbf{D}_{-\mathcal{A}_j} \mathbf{D}_{-\mathcal{A}_j}^T\big)^{-1} \mathbf{D}_{-\mathcal{A}_j}\Big)\Big]_{\widehat{\tau}_{j+1}^{\,\text{leave}}} \mathbf{D}_{\mathcal{A}_j}^T \widehat{\mathbf{s}}_{\mathcal{A}_j}}. \tag{4.11}$$

In the case of leaving time, which corresponds to $\lambda_{j+1}^{\text{join}} < \lambda_{j+1}^{\text{leave}}$, the sign of inequality in (4.11) will flip. Observe that the decision on whether to add or remove a coordinate will only add one row to $\mathbf{A}_j$. Also, it is important to note that $\mathbf{q} = \mathbf{0}$ in all the above steps.

4.3.2 Construction of Polyhedron After Stopping Rule

In the preceding section, we described how to construct the polyhedron matrix $\mathbf{A}$ and vector $\mathbf{q}$ after running the PRUTF algorithm for a fixed number of steps. However, the PRUTF algorithm, as discussed in Section 3.6, terminates using a stopping rule. This stopping rule is developed based on the stochastic term of the dual variables, $\widehat{u}^{\,\mathrm{st}}_{-\mathcal{A}}(t) = \Big[\big(\mathbf{D}_{-\mathcal{A}}\,\mathbf{D}^T_{-\mathcal{A}}\big)^{-1}\,\mathbf{D}_{-\mathcal{A}}\Big]_t\,\mathbf{y}$, for $t \in \{1,\ldots,m\}\backslash\mathcal{A}$. In the following, we will show that this stopping criterion can also be expressed as a polyhedron in $\mathbf{y}$.

The stopping rule in Section 3.6 states that the PRUTF algorithm stops upon the time the inequality

$$\max_{0\le t\le 1}\,\Big|\,\widehat{\mathbf{u}}^{\,\mathrm{st}}_{-\mathcal{A}}\big(\lfloor\, kt\,\rfloor\big)\,\Big| \;\le\; \sigma\, x_\alpha\big(k-r\big)^{(2\,r+1)/2}, \qquad \text{for} \quad k = m-|\mathcal{A}|\,, \tag{4.12}$$

is satisfied. In order to show that this stopping rule creates a polyhedron, note that, from (3.32), the condition (4.12) can be rewritten as

$$\begin{aligned}
\Big[\big(\mathbf{D}_{-\mathcal{A}}\,\mathbf{D}^T_{-\mathcal{A}}\big)^{-1}\,\mathbf{D}_{-\mathcal{A}}\Big]_t\,\mathbf{y} \;&\ge\; -\sigma\, x_\alpha\big(k-r\big)^{(2\,r+1)/2}, \qquad && \text{for} \quad t\notin\mathcal{A}\\
-\Big[\big(\mathbf{D}_{-\mathcal{A}}\,\mathbf{D}^T_{-\mathcal{A}}\big)^{-1}\,\mathbf{D}_{-\mathcal{A}}\Big]_t\,\mathbf{y} \;&\ge\; -\sigma\, x_\alpha\big(k-r\big)^{(2\,r+1)/2}, \qquad && \text{for} \quad t\notin\mathcal{A}.
\end{aligned} \tag{4.13}$$

These conditions append $2\,k$ rows to the matrix $\mathbf{A}$. Additionally, the above conditions add $2\,k$ non-zero values $-\sigma\, x_\alpha\big(k-r\big)^{(2\,r+1)/2}$ to $\mathbf{q}$.

Remark 4.1 *We have just shown that the event* $\big\{\widehat{\boldsymbol{\tau}} = \boldsymbol{\tau}\,,\,\widehat{\mathbf{s}} = \mathbf{s}\big\}$*, for all fixed* $\boldsymbol{\tau}$ *and* $\mathbf{s}$*, constitutes a polyhedron of the form* $\big\{\,\mathbf{y}:\;\;\mathbf{A}\,\mathbf{y}\,\ge\,\mathbf{q}\big\}$*. It would also be interesting to only characterize change point set* $\big\{\widehat{\boldsymbol{\tau}} = \boldsymbol{\tau}\big\}$ *in the form of a polyhedron. It turns out that* $\big\{\widehat{\boldsymbol{\tau}} = \boldsymbol{\tau}\big\}$ *can be represented as the union of such polyhedra. More precisely,*

$$\Big\{\mathbf{y}:\;\;\widehat{\boldsymbol{\tau}}(\mathbf{y}) = \boldsymbol{\tau}\Big\} = \bigcup_{\mathbf{s}}\,\Big\{\mathbf{y}:\;\;\widehat{\boldsymbol{\tau}}(\mathbf{y}) = \boldsymbol{\tau}\,,\,\widehat{\mathbf{s}}(\mathbf{y}) = \mathbf{s}\Big\} = \bigcup_{\mathbf{s}}\,\Big\{\mathbf{y}:\;\;\mathbf{A}_{\mathbf{s}}\,\mathbf{y}\,\ge\,\mathbf{q}_{\mathbf{s}}\Big\}, \tag{4.14}$$

where the union is over all sign vectors, $\mathbf{s}\,\in\,\{-1,\,1\}^{\,|\mathrm{row}(\mathbf{A})|}$*. Observe that the number of elements for this union is* $2^{\,|\mathrm{row}(\mathbf{A})|}$ *which can grow fast and become intractable when* $|\mathrm{row}(\mathbf{A})|$ *is moderately large.*

4.4 Post-Detection Inference

Having detected change points using the PRUTF algorithm, an appealing follow-up goal would be performing statistical inference on the significance of the changes at these locations. In this section, using the post-selection inference framework, we provide inference tools to apply after implementing the PRUTF algorithm. Given the vector of observations $\mathbf{y}$, we assume that $\widehat{\boldsymbol{\tau}} = \left\{\widehat{\tau}_1, \ldots, \widehat{\tau}_J\right\}$ is the set of ordered change points, $1 < \widehat{\tau}_1 < \widehat{\tau}_2 < \ldots < \widehat{\tau}_J < n - r - 1$, detected using the PRUTF algorithm. Additionally, assume that $\widehat{\mathbf{s}} = \left\{\widehat{s}_1, \ldots, \widehat{s}_J\right\}$ is the set of signs associated with $\widehat{\boldsymbol{\tau}}$. For notational convenience, we denote $\widehat{\tau}_0 = 0$ and $\widehat{\tau}_{J+1} = n - r$. Our focus is on conducting valid statistical inference for the significance of changes at the locations $\widehat{\tau}_j$, $j = 1\,, \ldots\,, J$.

Hypothesis tests to determine the significance of the detected change points can be cast in a general linear form $H_0 : \boldsymbol{\eta}^T\mathbf{f} = 0$, for an arbitrarily selected nonzero contrast vector $\boldsymbol{\eta} \in \mathbb{R}^n$. Of particular interest in this thesis is the hypothesis $H_0 : \mathbf{D}_{\widehat{\tau}_j}\mathbf{f} = 0$, which tests the significance of a change right at its estimated location $\widehat{\tau}_j$, where $\mathbf{D}_{\widehat{\tau}_j}$ is the $\widehat{\tau}_j$-th row of the penalty matrix $\mathbf{D}$. However, our inferential framework is not specific to the choice $\boldsymbol{\eta}^T = \mathbf{D}_{\widehat{\tau}_j}$. Many other types of contrast vectors are possible, as long as $\boldsymbol{\eta}$ is fixed by conditioning on the detection procedure. For example, the segment contrast, proposed in [77], considers the difference between the averages of two neighboring segments $\left(\widehat{\tau}_{j-1}\,, \widehat{\tau}_j\right]$ and $\left(\widehat{\tau}_j\,, \widehat{\tau}_{j+1}\right]$. Also, [82] have used a window contrast of size h, which considers the difference between averages of h consecutive points just before the estimated change point, i.e., $\left(\widehat{\tau}_j - h - 1\,, \widehat{\tau}_j\right]$ and h consecutive points just after that, i.e., $\left(\widehat{\tau}_j\,, \widehat{\tau}_j + h\right]$. The window contrast is suitable for checking whether a true change point exists near $\widehat{\tau}_j$. It is worth mentioning that the choice of $\boldsymbol{\eta}$ ultimately depends on the researcher's objective.

The data-dependent nature of change point detection methodologies leads to a random change point set $\widehat{\boldsymbol{\tau}} = \widehat{\boldsymbol{\tau}}\,(\mathbf{y})$. Associated with this randomly chosen $\widehat{\boldsymbol{\tau}}$ is the vector $\boldsymbol{\eta} = \boldsymbol{\eta}(\widehat{\boldsymbol{\tau}})$ which, in turn, is a random object. This randomness invalidates classical theory for conducting statistical inference about $\boldsymbol{\eta}^T\mathbf{f}$; see [22] for a thorough discussion. In such cases, post-detection inference allows us to carry out our analysis. Specifically, post-detection inference revolves around the conditional distribution of $\boldsymbol{\eta}^T\mathbf{y}$ conditional on the selected change points. This conditioning makes $\boldsymbol{\eta} = \boldsymbol{\eta}(\widehat{\boldsymbol{\tau}})$ become a fixed vector. Now, the goal is to test a hypothesis that controls the conditional type-I error rate at level α as well as to build a conditional confidence interval $I_{\boldsymbol{\eta}}$, such that

$$\Pr\left(\boldsymbol{\eta}^T\mathbf{f} \in I_{\boldsymbol{\eta}} \,\middle|\, \widehat{\boldsymbol{\tau}} = \boldsymbol{\tau}\right) \geq 1 - \alpha,$$

for all fixed $\boldsymbol{\tau}$.

In a change point detection setting, [77, 78] have exploited post-selection inference to compute p-values for the significance of change points found by fused lasso, Binary Segmentation, Wild Binary Segmentation, Circular Binary Segmentation. To boost the power of tests, [82] have suggested a post-detection approach which attempts to reduce the size of conditioning events. This approach covers change points detected in a piecewise constant model with identical and independent Gaussian random noises. In [46], the authors have also considered post-selection inference for change points estimated by using a dynamic programming for a ℓ_0-penalization.

Conducting inference for $\boldsymbol{\eta}^T\mathbf{f}$ in the post-detection framework requires knowledge about the conditional distribution of $\boldsymbol{\eta}^T\mathbf{y}$ given $\{\widehat{\boldsymbol{\tau}} = \boldsymbol{\tau}\}$. As shown in Section 4.3, $\{\widehat{\boldsymbol{\tau}} = \boldsymbol{\tau}\,, \widehat{\mathbf{s}} = \mathbf{s}\}$ creates the polyhedron of the form $\big\{\,\mathbf{y}\,:\;\mathbf{A}\,\mathbf{y} \geq \mathbf{q}\big\}$, and $\{\widehat{\boldsymbol{\tau}} = \boldsymbol{\tau}\}$ is a union of such polyhedra over all possible sign vectors $\mathbf{s}$. Therefore, it is easier to obtain the conditional distribution of $\boldsymbol{\eta}^T\mathbf{y}$ given $\{\widehat{\boldsymbol{\tau}} = \boldsymbol{\tau}\,, \widehat{\mathbf{s}} = \mathbf{s}\}$, a single polyhedron. Observe that inferences that are valid conditional on this finer event will also be valid conditional on $\{\widehat{\boldsymbol{\tau}} = \boldsymbol{\tau}\}$, [93, 153]. In what follows, we illustrate techniques to compute valid post-detection p-values for the hypothesis $H_0 : \boldsymbol{\eta}^T\mathbf{f} = 0$ and to construct a post-detection confidence intervals for the parameter $\boldsymbol{\eta}^T\mathbf{f}$.

Ignoring the post-detection framework for a moment, recall that when the change points are assumed fixed, the inference about $\boldsymbol{\eta}^T\mathbf{f}$, depending on whether the error variance is known or unknown, is based on the normal or t distributions, respectively. More specifically, when σ^2 is known, the statistic $Z = \boldsymbol{\eta}^T\mathbf{y}/\sigma\|\boldsymbol{\eta}\|$ and when σ^2 is unknown, the statistic $T = \boldsymbol{\eta}^T\mathbf{y}/\widehat{\sigma}\|\boldsymbol{\eta}\|$, are employed to make inference for $\boldsymbol{\eta}^T\mathbf{f}$. We will essentially use the same statistics Z and T in post-detection inference and focus on determining their respective conditional distributions. We emphasize that these distributions are no longer the usual normal or t distributions as they must be conditioned on the detected change points.

To define our proposed test statistics, again assume that the PRUTF algorithm has detected J change points at locations $\widehat{\boldsymbol{\tau}} = \big\{\widehat{\tau}_1, \ldots, \widehat{\tau}_J\big\}$ with the corresponding signs $\widehat{\mathbf{s}} = \big\{\widehat{s}_1, \ldots, \widehat{s}_J\big\}$. These change points partition the entire signal $\mathbf{f}$ into $J+1$ segments, with each segment having its distinct polynomial signal of order r, namely $\mathbf{f}_j$, $j = 0\,, \ldots\,, J$. Also, let $\mathbf{y}_j$ denote the subvector of observations corresponding to the j-th segment; thus,

$$\mathbf{f}^T\mathbf{y} = \sum_{j=0}^{J} \mathbf{f}_j^T\mathbf{y}_j.$$

Implementing the least square approach to estimate $\mathbf{f}_j$ results in $\widehat{\mathbf{f}}_j = \mathbf{X}_j^T\,\left(\mathbf{X}_j^T\mathbf{X}_j\right)^{-1}\mathbf{X}_j^T\mathbf{y}_j$,

for $j = 0, \ldots, J$, where $\mathbf{X}_j$ is defined as

$$\mathbf{X}_j = \begin{pmatrix} 1 & \frac{\widehat{\tau}_j+1}{n} & \left(\frac{\widehat{\tau}_j+1}{n}\right)^2 & \cdots & \left(\frac{\widehat{\tau}_j+1}{n}\right)^r \\ 1 & \frac{\widehat{\tau}_j+2}{n} & \left(\frac{\widehat{\tau}_j+2}{n}\right)^2 & \cdots & \left(\frac{\widehat{\tau}_j+2}{n}\right)^r \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \frac{\widehat{\tau}_{j+1}}{n} & \left(\frac{\widehat{\tau}_{j+1}}{n}\right)^2 & \cdots & \left(\frac{\widehat{\tau}_{j+1}}{n}\right)^r \end{pmatrix}. \tag{4.15}$$

In words, $\mathbf{X}_j$ is the design matrix of the r-th polynomial regression of $\mathbf{y}_j$ on the indices of j-th segment, $\widehat{\tau}_j + 1, \ldots, \widehat{\tau}_{j+1}$. We also denote the projection matrix onto the column space of $\mathbf{X}_j$ as $\mathbf{P}_j = \mathbf{X}_j \left(\mathbf{X}_j^T \mathbf{X}_j\right)^{-1} \mathbf{X}_j^T$. Observe that $\mathbf{P}_j \, \mathbf{f}_j = \mathbf{f}_j$, for $j = 0, \ldots, J$, or equivalently $\mathbf{P}\,\mathbf{f} = \mathbf{f}$, where $\mathbf{P}$ is a block diagonal matrix whose diagonal entries are the submatrices $\mathbf{P}_j$. With these notations, an unbiased estimator of the error variance, when σ^2 is unknown, is given by

$$\widehat{\sigma}^2 = \frac{1}{d} \sum_{j=0}^{J} \left\| (\mathbf{I}_j - \mathbf{P}_j)\, \mathbf{y}_j \right\|^2 = \frac{1}{d} \left\| (\mathbf{I} - \mathbf{P})\, \mathbf{y} \right\|^2, \tag{4.16}$$

where $d = n - (J + 1)$.

According to the Gaussian model of (4.2), $\mathbf{y}$ follows the exponential family of the form

$$\mathbf{y} \sim \exp\left\{ \frac{1}{\sigma^2}\, \mathbf{f}^T \mathbf{y} - \frac{1}{2\sigma^2} \|\mathbf{y}\|^2 - \frac{\|\mathbf{f}\|^2}{2\sigma^2} \right\}.$$

Decomposing the data into the direction of $\boldsymbol{\eta}$ and orthogonal to $\boldsymbol{\eta}$, as well as using the fact that $\mathbf{P}\,\mathbf{f} = \mathbf{f}$, yield

$$\begin{aligned} \mathbf{y} &\sim \exp\left\{ \frac{1}{\sigma^2}\, \mathbf{f}^T \left(\mathbf{P}_\eta + \mathbf{P} - \mathbf{P}_\eta\right) \mathbf{y} - \frac{1}{2\,\sigma^2} \|\mathbf{y}\|^2 - \frac{\|\mathbf{f}\|^2}{2\sigma^2} \right\} \\ &= \exp\left\{ \left(\frac{1}{\sigma \|\boldsymbol{\eta}\|}\right) (\boldsymbol{\eta}^T \mathbf{f})^T \left(\frac{\boldsymbol{\eta}^T \mathbf{y}}{\sigma \|\boldsymbol{\eta}\|}\right) + \frac{1}{\sigma^2}\, \mathbf{f}^T \left(\mathbf{P} - \mathbf{P}_\eta\right) \mathbf{y} - \frac{1}{2\,\sigma^2} \|\mathbf{y}\|^2 - \frac{\|\mathbf{f}\|^2}{2\,\sigma^2} \right\}, \end{aligned} \tag{4.17}$$

where $\mathbf{P}_\eta$ is the orthogonal projection on the space spanned by $\boldsymbol{\eta} \in \mathbb{R}^n$, defined as

$$\mathbf{P}_\eta = \boldsymbol{\eta} \left(\boldsymbol{\eta}^T \boldsymbol{\eta}\right)^{-1} \boldsymbol{\eta}^T = \frac{\boldsymbol{\eta}\, \boldsymbol{\eta}^T}{\|\boldsymbol{\eta}\|^2}.$$

Consider a multi-parameter exponential family in which the vector of natural parameters can be split into two subvectors: the target and the nuisance parameters. The conditional distribution, given sufficient statistics associated with nuisance parameters, depends only on the target parameters. It is known that this conditional distribution belongs to an exponential family with the same target parameters, the same sufficient statistics, but a different reference measure and the normalizing constant. This fact allows us to derive the conditional distribution of the statistics Z and T, given the detected change points, their signs and the sufficient statistics of the nuisance parameters. Consequently, we can perform statistical inference for the target parameter $\boldsymbol{\eta}^T\mathbf{f}$ based on the derived distributions. We divide our presentation into two parts: known error variance and unknown error variance to perform such inferences.

4.4.1 Known Error Variance

When the error variance σ^2 is known, the statement (4.17) reveals that $\boldsymbol{\eta}^T\mathbf{y}$ and $\mathbf{V} = \big(\mathbf{P} - \mathbf{P}_{\eta}\big)\,\mathbf{y}$ are sufficient statistics for $\boldsymbol{\eta}^T\mathbf{f}$ and the nuisance parameters, respectively. As previously explained, the conditional distribution of $\boldsymbol{\eta}^T\mathbf{y}$ given $\mathbf{V}$ eliminates nuisance parameters from our analysis. Therefore, when σ^2 is known, we base our analysis on the statistic

$$Z = \frac{\boldsymbol{\eta}^T\mathbf{y}}{\sigma\,\|\boldsymbol{\eta}\|}\,. \tag{4.18}$$

We further seek to specify the conditional distribution of this statistic given $\{\,\mathbf{V},\ \mathbf{A}\mathbf{y} \geq \mathbf{q}\}$. Recall that the polyhedron $\{\mathbf{A}\mathbf{y} \geq \mathbf{q}\}$ is the substitute for $\big\{\widehat{\boldsymbol{\tau}}\,,\,\widehat{\mathbf{s}}\big\}$, estimated using PRUTF. The following theorem illustrates that this conditional distribution is indeed a truncated normal distribution with an explicitly specified truncation set.

Theorem 4.2 *Suppose* $\mathbf{y}$ *follows model* (4.2)*, where* σ^2 *is assumed known. For a nonzero contrast vector* $\boldsymbol{\eta}$*,*

a) the conditional distribution of Z *in* (4.18)*, given*

$$\Big\{\,\mathbf{V},\ \mathbf{A}\mathbf{y} \geq \mathbf{q}\Big\}, \tag{4.19}$$

is a normal distribution truncated to the interval $\big[\mathcal{V}_Z^-\,,\,\mathcal{V}_Z^+\big]$*, provided* $\mathcal{V}_Z^0 \geq 0$*. This distribution is denoted by* $\mathrm{TN}\big(\boldsymbol{\eta}^T\mathbf{f},\,1,\,[\mathcal{V}_Z^-,\mathcal{V}_Z^+]\big)$*. The truncation boundaries* $\mathcal{V}_Z^- =$

$\mathcal{V}_Z^-(\mathbf{V})$, $\mathcal{V}_Z^+ = \mathcal{V}_Z^+(\mathbf{V})$ *and* $\mathcal{V}_Z^0 = \mathcal{V}_Z^0(\mathbf{V})$ *are given by*

$$\mathcal{V}_Z^- = \max_{i:\,\rho_i>0} \frac{\big[\mathbf{q}-\mathbf{AV}\big]_i}{\sigma\rho_i}\,, \qquad \mathcal{V}_Z^+ = \min_{i:\,\rho_i<0} \frac{\big[\mathbf{q}-\mathbf{AV}\big]_i}{\sigma\rho_i}\,, \qquad \mathcal{V}_Z^0 = \min_{i:\,\rho_i=0} \big[\mathbf{AV}-\mathbf{q}\big]_i\,, \quad (4.20)$$

where $\rho_i = \big[\mathbf{A}\boldsymbol{\eta}/\|\boldsymbol{\eta}\|\big]_i$, *for* $i = 0\,,1\,,\ldots,\big|\mathrm{row}(\mathbf{A})\big|$. *Here* $\big|\mathrm{row}(\mathbf{A})\big|$ *denotes the number of rows in matrix* $\mathbf{A}$.

b) Moreover, let $\Phi^{[a\,,b]}(\cdot)$ *be the cumulative distribution function of* $\mathrm{TN}\big(0,\,1,\,[a\,,\,b]\big)$, *thus, under the null hypothesis* $H_0 : \boldsymbol{\eta}^T\mathbf{f} = 0$, *the conditional distribution of* $1 - \Phi^{[\mathcal{V}_Z^-,\mathcal{V}_Z^+]}(Z)$ *given the event* $\big\{\,\mathbf{Ay} \geq \mathbf{q}\big\}$ *is uniform on the unit interval* $[0,\,1]$, *that is,*

$$1 - \Phi^{[\mathcal{V}_Z^-,\mathcal{V}_Z^+]}(Z)\,\Big|\,\big\{\,\mathbf{Ay} \geq \mathbf{q}\big\} \sim \mathrm{U}(0\,,\,1)\,. \quad (4.21)$$

We refer to this statistic as TN *statistic.*

The proof is given in Appendix B.1.

Theorem 4.2 enables us to compute a post-detection p-value for the hypothesis $H_0 : \boldsymbol{\eta}^T\mathbf{f} = 0$ as well as to construct a post-detection confidence interval for $\boldsymbol{\eta}^T\mathbf{f}$. These tasks can be carried out using the TN statistic in (4.21). In particular, for the two-sided hypothesis testing problem

$$H_0 : \boldsymbol{\eta}^T\mathbf{f} = 0 \qquad \text{vs} \qquad H_1 : \boldsymbol{\eta}^T\mathbf{f} \neq 0, \quad (4.22)$$

the value of $2\min\Big\{1 - \Phi^{[\mathcal{V}_Z^-,\mathcal{V}_Z^+]}(Z)\,,\,\Phi^{[\mathcal{V}_Z^-,\mathcal{V}_Z^+]}(Z)\Big\}$ serves as a valid post-detection p-value, since under the null hypothesis

$$\Pr\left(2\min\Big\{1 - \Phi^{[\mathcal{V}_Z^-,\mathcal{V}_Z^+]}(Z)\,,\,\Phi^{[\mathcal{V}_Z^-,\mathcal{V}_Z^+]}(Z)\Big\} \leq \alpha\,\middle|\,\mathbf{Ay} \geq \mathbf{q}\right) = \alpha,$$

for all $0 \leq \alpha \leq 1$. To construct a two-sided post-detection confidence interval, define the confidence limits $L_Z(Z)$ and $U_Z(Z)$ such that

$$1 - \Phi^{[\mathcal{V}_Z^-,\mathcal{V}_Z^+]}\left(\frac{\boldsymbol{\eta}^T\mathbf{y} - L_Z(Z)}{\sigma\|\boldsymbol{\eta}\|}\right) = \frac{\alpha}{2}, \qquad \text{and} \qquad \Phi^{[\mathcal{V}_Z^-,\mathcal{V}_Z^+]}\left(\frac{\boldsymbol{\eta}^T\mathbf{y} - U_Z(Z)}{\sigma\|\boldsymbol{\eta}\|}\right) = \frac{\alpha}{2}\,. \quad (4.23)$$

These limits are well characterized since the survival function of $\mathrm{TN}\big(\mu,\,1,\,[a,\,b]\big)$ monotonically increases with respect to μ. Hence, the interval $\big[L_Z(Z)\,,\,U_Z(Z)\big]$ is a valid two-sided

post-detection confidence interval at a $1-\alpha$ level for $\boldsymbol{\eta}^T\mathbf{f}$. This confidence interval can be interpreted in the following manner. If $\mathbf{y}$ is repeatedly drawn from model (4.2) and the PRUTF algorithm is run, among those cases in which $\left\{\widehat{\boldsymbol{\tau}}\,,\,\widehat{\mathbf{s}}\right\}$ are detected, the interval $\left[L_{_Z}(Z)\,,\,U_{_Z}(Z)\right]$ contains the parameter $\boldsymbol{\eta}^T\mathbf{f}$ with a relative frequency approaching $1-\alpha$.

Remark 4.3 *The PRUTF algorithm estimates sings of change points in addition to their locations. We can incorporate this knowledge in forming the alternative hypothesis, that is,* $H_1 : \widehat{s}_{\widehat{\tau}_j}\,\mathbf{D}_{\widehat{\tau}_j}\,\mathbf{f} > 0$. *This alternative hypothesis means that* $\widehat{\tau}_{_j}$ *is a significant change point whose jump is in the direction of* $\widehat{s}_{\widehat{\tau}_j}$. *Note that this test is more powerful than its two-sided counterpart. [153] have provided a comparison between one-sided and two-sided tests for the significance of the selected variables using lasso. In general, suppose we are interested in testing the one-sided hypothesis* $H_0 : \boldsymbol{\eta}^T\mathbf{f} = 0$ *against* $H_1 : \boldsymbol{\eta}^T\mathbf{f} > 0$. *As with the two-sided hypotheses,* $1-\Phi^{[\mathcal{V}_Z^-,\mathcal{V}_Z^+]}(Z)$ *is a valid post-detection p-value for the one-sided test. Additionally,* $\left[L_{_Z}(Z)\,,\,\infty\right)$ *is a one-sided post-detection confidence interval for* $\boldsymbol{\eta}^T\mathbf{f}$ *where*

$$1-\Phi^{[\mathcal{V}_Z^-,\mathcal{V}_Z^+]}\left(\frac{\boldsymbol{\eta}^T\mathbf{y}-L_{_Z}(Z)}{\sigma\|\boldsymbol{\eta}\|}\right)=\alpha.$$

4.4.2 Unknown Error Variance

This section concerns post-detection inference after estimating change points, in the more realistic case, when the error variance σ^2 is unknown. The main methods proposed for post-selection inference such as in [93], [153], and [142] proceed with a known σ^2, with the exception of [52]. In the case of an unknown σ^2, we must further condition our inference on the sufficient statistic associated with the nuisance parameter σ^2.

According to (4.17), in the case of an unknown σ^2, the term $\boldsymbol{\eta}^T\mathbf{y}$ is the sufficient statistic for the target parameter $\boldsymbol{\eta}^T\mathbf{f}/\sigma^2$ and $\left(\mathbf{V}\,,\,\|\mathbf{y}\|^2\right)$ is a joint sufficient statistic for the nuisance parameters. Since testing $\boldsymbol{\eta}^T\mathbf{f}/\sigma^2=0$ is equivalent to testing $\boldsymbol{\eta}^T\mathbf{f}=0$, hence, we construct our analysis based on the statistic

$$T=\frac{\boldsymbol{\eta}^T\mathbf{y}}{\widehat{\sigma}\,\|\boldsymbol{\eta}\|}, \tag{4.24}$$

where $\widehat{\sigma}^2 = d^{-1}\left\|(\mathbf{I}-\mathbf{P})\,\mathbf{y}\right\|^2$, with $d=n-(J+1)$. Notice that $\widehat{\sigma}^2$ is simply a pooled estimate of the error variance σ^2 using $J+1$ segments created by the detected change points $\widehat{\boldsymbol{\tau}} = \left\{\widehat{\tau}_{_1},\,\ldots,\,\widehat{\tau}_{_J}\right\}$. The next step is to find the conditional distribution of T, given

the sufficient statistics associated with the nuisance parameters and the polyhedron event identified by the detection procedure. Clearly, this statistic is distributed as a t distribution constrained to the set $\big\{\,\mathbf{V},\, \|\mathbf{y}\|^2,\, \mathbf{A}\mathbf{y} \geq \mathbf{q}\big\}$. We establish the corresponding distribution in the following theorem, whose proof is given in Appendix B.2.

Theorem 4.4 *Suppose* $\mathbf{y}$ *follows model* (4.2) *and* σ^2 *is assumed unknown. For a nonzero contrast vector* $\boldsymbol{\eta}$*,*

a) the conditional distribution of T *given*

$$\Big\{\,\mathbf{V},\, \big\|\mathbf{y}\big\|^2,\, \mathbf{A}\mathbf{y} \geq \mathbf{q}\Big\}, \tag{4.25}$$

is a generalized (location-scale) t *distribution with mean* $\boldsymbol{\eta}^T\mathbf{f}$*, variance 1 and degrees of freedom* $d = n - (J+1)(r+1)$*, truncated to the interval* $[\mathcal{V}_T^-\,,\,\mathcal{V}_T^+]$*, denoted by* $\mathrm{Tt}\big(\boldsymbol{\eta}^T\mathbf{f},\, 1,\, d,\, [\mathcal{V}_T^-, \mathcal{V}_T^+]\big)$*. The truncation boundaries* $\mathcal{V}_T^- = \mathcal{V}_T^-\big(\mathbf{V},\, W\big)$ *and* $\mathcal{V}_T^+ = \mathcal{V}_T^+\big(\mathbf{V},\, W\big)$ *are given by*

$$\Big[\mathcal{V}_T^-\,,\,\mathcal{V}_T^+\Big] = \bigcap_{i=1}^{|\mathrm{row}(\mathbf{A})|} \Bigg\{ t \in \mathbb{R} : \Big[\mathbf{A}\mathbf{V} - \mathbf{q}\Big]_i\, t^2 + \left(2\,\frac{\boldsymbol{\eta}^T(\mathbf{I}-\mathbf{P})\,\mathbf{y}}{\widehat{\sigma}\|\boldsymbol{\eta}\|}\,\Big[\mathbf{A}\mathbf{V} - \mathbf{q}\Big]_i + \frac{W\rho_i}{\widehat{\sigma}} \right) t + \Big[\mathbf{A}\mathbf{V} - \mathbf{q}\Big]_i\, d \geq 0 \Bigg\}, \tag{4.26}$$

where $W = \big\|\big(\mathbf{I} - \mathbf{P} + \mathbf{P}_\eta\big)\,\mathbf{y}\big\|^2$*.*

b) In addition, let $G_d^{[a,\,b]}(\cdot)$ *denote the cumulative distribution function of* $\mathrm{Tt}\big(0\,,\,1\,,\,d\,,[a\,,\,b]\big)$*, then under the null hypothesis*

$$1 - G_d^{[\mathcal{V}_T^-,\,\mathcal{V}_T^+]}\,(T)\,\Big|\,\big\{\mathbf{A}\mathbf{y} \geq \mathbf{q}\big\} \sim U(0\,,\,1). \tag{4.27}$$

We refer this statistic as the Tt *statistic.*

Theorem 4.4 allows us to perform post-detection tests and construct post-detection confidence intervals for $\boldsymbol{\eta}^T\mathbf{f}$, when σ^2 is unknown. In the same fashion as in the truncated normal case, the quantity

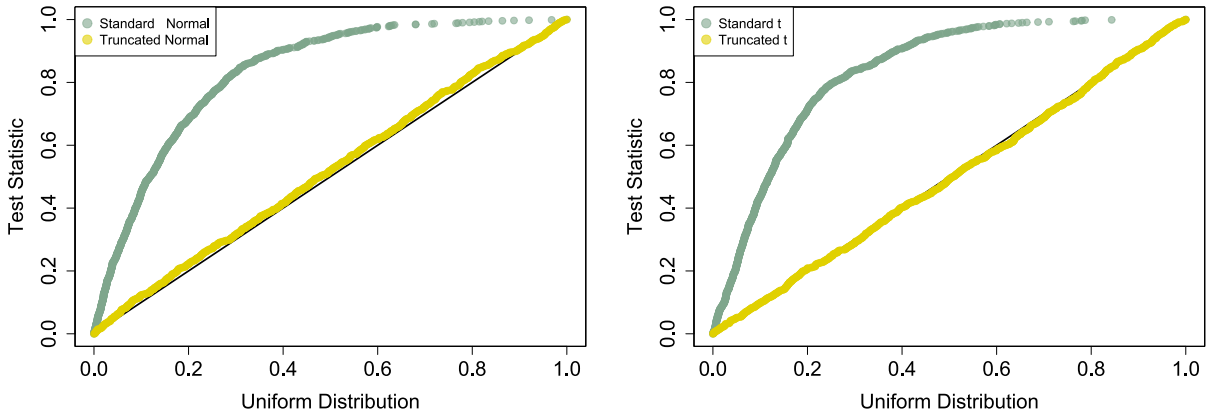
$$2\min\Big\{1 - G_d^{[\mathcal{V}_T^-,\,\mathcal{V}_T^+]}\,(T)\,,\, G_d^{[\mathcal{V}_T^-,\,\mathcal{V}_T^+]}\,(T)\Big\}$$

is a post-detection p-value for the two-sided hypothesis problem (4.22). A post-detection confidence interval for $\boldsymbol{\eta}^T\mathbf{f}$, when σ^2 is unknown, is also given by $\big[L_T(T)\,,\,U_T(T)\big]$, where

$$1-G_d^{[\mathcal{V}_Z^-,\mathcal{V}_Z^+]}\left(\frac{\boldsymbol{\eta}^T\mathbf{y}-L_T(T)}{\widehat{\sigma}\|\boldsymbol{\eta}\|}\right)=\alpha/2\,,\qquad\text{and}\qquad G_d^{[\mathcal{V}_Z^-,\mathcal{V}_Z^+]}\left(\frac{\boldsymbol{\eta}^T\mathbf{y}-U_T(T)}{\widehat{\sigma}\|\boldsymbol{\eta}\|}\right)=\alpha/2\,. \tag{4.28}$$

The same technique explained in Remark 4.3 can be used to construct a one-sided confidence interval for $\boldsymbol{\eta}^T\mathbf{f}$.

Figure 4.1 demonstrates the distribution of TN and Tt statistics, stated in (4.21) and (4.27), by displaying their quantiles versus those of $U(0\,,\,1)$. The figure also represents the distribution of the two statistics when the truncated normal and truncated t distributions in (4.21) and (4.27) are replaced with their untruncated counterparts. The figure certifies that the distribution of Z and T change from normal and t distributions to truncated normal and truncated t distributions, respectively, while accounting for the detection procedure.



(a) Q-Q plot of survival function of Z.

(b) Q-Q plot of survival function of T.

Figure 4.1: The Q-Q plots of survival functions of Z and T for standard and truncated normal and t distributions. A piecewise constant signal of size $n = 100$ and a true change point at $\boldsymbol{\tau} = \{50\}$ is considered. The left panel displays Q-Q plot, constructed over 1000 simulations, of the distribution of the survival function of Z in (4.21), conditional on the algorithm having made an incorrect selection in the second step, under standard normal (green dots) and truncated normal (yellow dots) distributions. The right panel shows Q-Q plots of the distribution of the survival function of T in (4.27) for the first detected change point, under t distribution (green dots) and truncated t distribution (yellow dots).

Remark 4.5 *We would like to emphasize that the confidence intervals constructed by using (4.23) and (4.28) employ the knowledge of the conditional distributions given the estimated change points and their signs,* $\{\widehat{\boldsymbol{\tau}} = \boldsymbol{\tau}\,, \widehat{\mathbf{s}} = \mathbf{s}\}$*, and therefore, have to be interpreted conditionally. These confidence intervals can also be interpreted simultaneously after applying the Bonferroni correction to properly adjust for multiplicity. That is, we would compute confidence intervals at the level* $1 - \dfrac{\alpha}{\widehat{J}}$*, where* $\widehat{J}$ *is the number of change points estimated by mPRUTF.*

One approach for constructing simultaneous confidence intervals in post-detection framework is to use the universal valid post-selection inference of [22], by considering all possible detection procedures. More precisely, for any change point detection procedure $\widehat{\boldsymbol{\tau}} : \mathbf{y} \longrightarrow \mathcal{T}$*, where* $\mathcal{T}$ *is the collection of all possible change point sets, the goal of the universal post-selection inference, referred to as simultaneous PoSI, is to construct confidence intervals for* $\boldsymbol{\eta}_{j,\,\widehat{\tau}}^T \mathbf{f}$*. In this notation,* $\boldsymbol{\eta}_{j,\,\widehat{\tau}}$ *denotes a nonzero contrast vector associated with* $\widehat{\tau}_j \in \widehat{\boldsymbol{\tau}}$*. According to the simultaneous PoSI method, a valid confidence interval for* $\boldsymbol{\eta}_{j,\,\widehat{\tau}}^T \mathbf{f}$ *takes the form*

$$\widehat{I}_{j,\,\widehat{\tau}}\,(K_\alpha) = \Big(\boldsymbol{\eta}_{j,\,\widehat{\tau}}^T \mathbf{y} \;\; \pm \;\; K_\alpha\, \sigma\, \|\boldsymbol{\eta}_{j,\,\widehat{\tau}}\| \Big),$$

where the constant K_α *is derived such that*

$$\Pr\Big(\boldsymbol{\eta}_{j,\,\widehat{\tau}}^T \mathbf{f} \;\in\; \widehat{I}_{j,\,\widehat{\tau}}(K_\alpha)\,, \quad \forall\, \widehat{\tau}_j \in \widehat{\boldsymbol{\tau}} \Big) \;\geq\; 1 - \alpha, \qquad\qquad \forall\, \widehat{\boldsymbol{\tau}} \in \mathcal{T}.$$

Note that the universal valid post-selection inference provides simultaneity both over the change points and over all detection procedures. This method has the advantage that it does not depend on the change point detection procedure. However, unless the number of possible change point sets is fairly small, this method is computationally challenging as the collection of all possible change point sets becomes intractable. Additionally, the confidence intervals derived from this method are unnecessarily wide as it disregards the knowledge of how the change points are estimated. To improve the universal valid post-detection inference, a method based on a simultaneous over selection (SoS) criterion was proposed by [175]. The method aims to construct confidence intervals for selected variables in a fixed stable model selection procedure in the regression context. We note that computing simultaneous confidence intervals is not among the goals of our inference and refer interested readers to [22], [21], [5] and [175].

Remark 4.6 *We have thus far explained how to carry out statistical inference for the detected change points after completing the PRUTF algorithm. In other words, we have*

computed post-selection p-values and confidence intervals conditional on the entire set of estimated change points $\widehat{\boldsymbol{\tau}}$. An alternative way is to perform statistical inferences for the change points in a sequential manner. In particular, suppose $\widehat{\boldsymbol{\tau}}_{j-1}$ and $\widehat{\mathbf{s}}_{j-1}$ are the vectors of detected change points and their corresponding signs at the iteration $j-1$ of the PRUTF algorithm. Additionally, suppose $\widehat{\tau}_j$ is the detected change point at step j. In the sequential scheme, the interest is in making inference about the significance of change at the location $\widehat{\tau}_j$, given $\left\{\widehat{\boldsymbol{\tau}}_j\,,\,\widehat{\mathbf{s}}_j\right\}$. Notice that the tools provided in Theorems 4.2 and 4.4 are still applicable in the sequential scheme, but the matrix $\mathbf{A}$ and vector $\mathbf{q}$ in the polyhedron representation must be adjusted accordingly. See [102] and [153] which have pursued statistical inference in a sequential approach.

4.5 A Critique of Post-Detection Inference Methods

In post-detection inference, we lose part of the information from the data because it has already been used to estimate change points. As a result, the post-detection distributions change from normal or t distributions to truncated normal or truncated t distributions (Theorem 4.2 and Theorem 4.4). The performance of post-detection inference heavily depends on the amount of information in the data used for change point detection. Over-conditioning, which uses too much information for the detection procedure and leaves little information for inference, leads to a significant loss in the power of tests and unacceptably wide confidence intervals; see [51], [93] and [88].

The over-conditioning issue in post-selection inference has motivated researchers to suggest various approaches that preserve higher amounts of left-over information for inference. One solution to the problem is data splitting (sometimes called data carving), in which the data is divided into two parts, one part for model selection and the other for inference [51]. Another approach, put forward by [146], is the idea of randomization, which selects the model based not on the actual dataset but on a noise-perturbed version of it. Also, [101] have suggested dividing the final selected model into a high-value and a low-value submodel. The post-selection inference will then be conducted by conditioning only on the high-value submodel. Although these approaches improve the performance of post-selection inference, they also share some drawbacks; see [51].

The specific problem of wide length confidence intervals in post-selection inference has been investigated by [88] for models chosen by lasso. They have established that a confidence interval produced by the truncated Gaussian distribution with a finite truncation boundary (either the upper or lower bound) always has an infinite expected length, Theorem 4.7, part (a). In the next theorem, part (b), we will show that the same property also

holds for a truncated t distribution. More precisely, let a truncation set $\mathcal{S}$ be the union of finitely many open intervals, where the intervals might be unbounded. In other words, $\mathcal{S}$ can be represented in the form of $\mathcal{S} = \bigcup_{i=1}^{k} (a_i\,,\, b_i)$, for a finite value k, and for some $a_1 < b_1 < a_2 < b_2 < \ldots < a_k < b_k$ in $\mathbb{R}$.

Theorem 4.7

a) Let $Z \sim \mathrm{TN}(\mu,\, \sigma^2,\, \mathcal{S}_Z)$, where the truncation set $\mathcal{S}_Z$ is of the form $\mathcal{S}_Z = \bigcup_{i=1}^{k} (a_i\,,\, b_i)$. Also, let $\big[L_Z(Z)\,,\, U_Z(Z)\big]$ be a conditional confidence interval for μ, given $\mathcal{S}_Z$. If the truncation set $\mathcal{S}_Z$ is bounded, either from below $(-\infty < a_1)$ or from above $(b_k < \infty)$, then

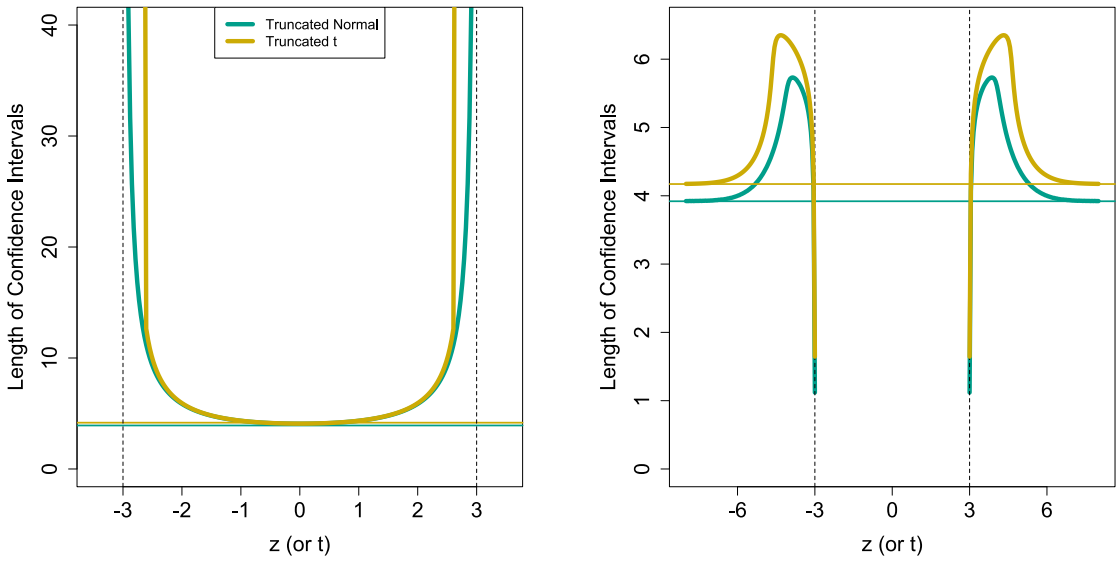
$$E\Big[U_Z(Z) - L_Z(Z)\Big] = \infty. \tag{4.29}$$

b) Let $T \sim \mathrm{Tt}(\mu,\, \sigma^2,\, d,\, \mathcal{S}_T)$, where the truncation set $\mathcal{S}_T$ is of the form $\mathcal{S}_T = \bigcup_{i=1}^{m} (c_i\,,\, d_i)$. Also, let $\big[L_T(T)\,,\, U_T(T)\big]$ be a conditional confidence interval for μ, given $\mathcal{S}_T$. Similar to part (a), if the truncation set $\mathcal{S}_T$ is bounded, either from below $(-\infty < c_1)$ or from above $(d_m < \infty)$, then

$$E\Big[U_T(T) - L_T(T)\Big] = \infty. \tag{4.30}$$

A proof is provided in Appendix B.3.

Theorem 4.7 states that the truncation set is crucial in constructing a desirable confidence interval in post-selection inference. To figure out why post-selection confidence intervals are sometimes exceedingly wide, assume $T \sim \mathrm{Tt}(\mu,\, \sigma^2,\, d,\, \mathcal{S}_T)$. When the truncation set $\mathcal{S}_T$ is bounded, values of T which are close to the endpoints of $\mathcal{S}_T$ leads to wide confidence intervals. This behaviour is because there are many values of μ that would be consistent with the data [93, 88]. On the other hand, when $\mathcal{S}_T$ is unbounded, the interval length is always bounded regardless of the values of T. Similar arguments hold for a truncated normal distribution. Figure 4.2 displays the lengths of confidence intervals derived using truncated normal and truncated t distributions for a bounded and an unbounded truncation set. The left panel indicates that the length of confidence intervals for truncated normal and truncated t distributions with bounded truncation sets diverge as the values of z or t approach the edges of the truncation set. This observation certifies the results derived in Theorem 4.7. On the other hand, when the truncation set is unbounded, the right panel indicates that the length of confidence intervals for both distributions are bounded. Moreover, the interval length converges to the length of confidence interval obtained from a



(a) Truncation set $\mathcal{S} = [-3\,,\,3]$.

(b) Truncation set $\mathcal{S} = (-\infty\,,\,-3) \cup (3\,,\,\infty)$.

Figure 4.2: Lengths of confidence intervals for a truncated normal distribution and a truncated t distribution with a bounded truncation set (left panel) and an unbounded truncation set (right panel). The solid horizontal lines in both panels show the length of confidence intervals (at the level 0.95) for the usual (untruncated) normal and t distributions.

usual (untruncated) normal or t distributions as the values of the statistics diverge from the edges of the unbounded truncation sets. Additionally, for both cases of bounded and unbounded truncation set, the lengths of confidence intervals for the truncated t distribution are greater or equal than those of the truncated normal distribution.

To measure the quality of the confidence intervals built in (4.23) and (4.28), we inspect their corresponding truncation sets to determine whether they are unbounded. The next theorem verifies that the truncation boundaries provided in Theorems 4.2 and 4.4 are in fact bounded.

Theorem 4.8 *Consider the settings of Theorems 4.2 and 4.4. Also, let* $\left[\mathcal{V}_Z^-(\mathbf{V})\,,\,\mathcal{V}_Z^+(\mathbf{V})\right]$ *and* $\left[\mathcal{V}_T^-(\mathbf{V},\,W)\,,\,\mathcal{V}_T^+(\mathbf{V},\,W)\right]$ *be the truncation sets for the truncated normal and truncated t distributions, derived in* (4.20) *and* (4.26)*, respectively. Then,*

(a) either one or both truncation boundaries in (4.20) are finite, that is,

$$-\infty < \mathcal{V}_Z^-(\mathbf{V}) \qquad \text{or} \qquad \mathcal{V}_Z^+(\mathbf{V}) < \infty.$$

(b) either one or both truncation boundaries in (4.26) are finite, that is,

$$-\infty < \mathcal{V}_T^-(\mathbf{V}, W) \qquad \text{or} \qquad \mathcal{V}_T^+(\mathbf{V}, W) < \infty.$$

A proof is given in appendix B.4.

Remark 4.9 *Theorem 4.8 states that the truncation sets derived in Theorems 4.2 and 4.4, for both truncated normal and truncated t distributions, are bounded. This result along with the result of Theorem 4.7 lead to*

$$E\Big[\, U_Z(Z) - L_Z(Z) \,\Big|\, \{\widehat{\boldsymbol{\tau}}\,, \widehat{\mathbf{s}}\}\Big] = \infty,$$

$$E\Big[\, U_T(T) - L_T(T) \,\Big|\, \{\widehat{\boldsymbol{\tau}}\,, \widehat{\mathbf{s}}\}\Big] = \infty. \tag{4.31}$$

In words, the expected lengths of the confidence intervals for $\boldsymbol{\eta}^T\mathbf{f}$ with known σ^2, given in (4.23), and with unknown σ^2, given in (4.28), are infinite.

4.6 Optimal Post-Detection Inference

Motivated by the fact that post-detection confidence intervals might become extremely wide, we seek approaches to attain confidence intervals with narrower length properties. Given a target change point, we can essentially think of $\widehat{\boldsymbol{\tau}}$ as containing two types of change points: those relevant to the discovery of the target change point and those irrelevant to it. The relevant change points are those change points that directly influence the detection of the target change point. In most cases, such points are located around the target change points. Therefore, one solution for improving the performance of post-detection inference is to allow only the relevant change points to participate in the analysis. In other words, the post-detection inference is carried out by only conditioning on a set containing relevant change points. We note that the set which includes these relevant change points is not necessarily a polyhedron. In the following, we will show that by narrowing down the conditional event to the set containing relevant change points, the resulting intervals

become much shorter. This procedure will also lead to improved powers for the associated hypothesis tests.

In the next two sections, we elaborate on two algorithms for conditioning on the relevant change point sets, leading to more powerful post-detection tests and narrower post-detection confidence intervals. We introduce two different setups. Section 4.6.1 involves post-detection inference by conditioning on the event that the given target change point is included in the estimated change point set. We refer to this setup as *global post-detection* because it verifies the significance of the target change point globally over the entire signal. In contrast, Section 4.6.2 involves statistical inference under the condition that the target change point and its adjacent change points are only included in the model. This setup is called *local post-detection* as it tests the significance of the target change points locally over its adjacent segments.

4.6.1 Global Post-Detection Inference

As explained before, global post-detection assumes that the target change point has been detected by the PRUTF algorithm and information about the rest of the change points is disregarded. In other words, for the given target change point $\widehat{\tau}_j$, we only assume that $\widehat{\tau}_j \in \widehat{\boldsymbol{\tau}}$. Now, the goal is to carry out post-detection inference by conditioning on the event $\{\, \widehat{\tau}_j \in \widehat{\boldsymbol{\tau}} \,\}$. Observe that this conditioning is different from conditioning on the entire change point set $\widehat{\boldsymbol{\tau}}$. We then attempt to compute a post-detection p-value for the null hypothesis $H_0 : \mathbf{D}_{\widehat{\tau}_j}\mathbf{f} = 0$, as well as building a post-detection confidence interval for $\mathbf{D}_{\widehat{\tau}_j}\mathbf{f}$. Depending on whether σ^2 is known or unknown, the post-detection tests are based on the statistics

$$Z_j^{\text{glo}} = \frac{\mathbf{D}_{\widehat{\tau}_j}\mathbf{y}}{\sigma\|\mathbf{D}_{\widehat{\tau}_j}^T\|} \qquad \text{and} \qquad T_j^{\text{glo}} = \frac{\mathbf{D}_{\widehat{\tau}_j}\mathbf{y}}{\widehat{\sigma}_j^{(\text{glo})}\|\mathbf{D}_{\widehat{\tau}_j}^T\|} \tag{4.32}$$

which, under the null hypothesis, are distributed as truncated normal and truncated t distributions, respectively. Given $\{\widehat{\tau}_j \in \widehat{\boldsymbol{\tau}}\}$, we split $\mathbf{y}$ into two subvectors $\mathbf{y}_{1,\,j}^{\text{glo}} = \big(y_1\,,\,\ldots\,,\,y_{\widehat{\tau}_j}\big)$ and $\mathbf{y}_{2,\,j}^{\text{glo}} = \big(y_{\widehat{\tau}_j+1}\,,\,\ldots\,,\,y_n\big)$. The projection matrices $\mathbf{P}_{k,\,j}^{\text{glo}}$ associated with the subvectors $\mathbf{y}_{k,\,j}^{\text{glo}}$, $k = 1\,,\,2$, and the block diagonal matrix $\mathbf{P}_j^{\text{glo}} = \text{diag}\big(\mathbf{P}_{1,\,j}^{\text{glo}}\,,\,\mathbf{P}_{2,\,j}^{\text{glo}}\big)$ can be defined accordingly (see Section 4.4). The next theorem explains how to specify the truncation sets for the underlying distributions. We provide the proof of the theorem in Appendix B.5.

Theorem 4.10 *Consider the statistics Z_j^{glo} and T_j^{glo} as defined in (4.32).*

(a) The conditional distribution of Z_j^{glo} *given* $\{\widehat{\tau}_j \in \widehat{\boldsymbol{\tau}}\}$*, under the null hypothesis, is the standard normal truncated to the set* $\big(-\infty\,,\,\mathcal{V}_{Z_j}^{-(glo)}\big] \bigcup \big[\mathcal{V}_{Z_j}^{+(glo)}\,,\,\infty\big)$*, where*

$$\mathcal{V}_{Z_j}^{-(glo)} = \frac{-\lambda_j\,\mathbf{D}_{\widehat{\tau}_j}\,\mathbf{D}_{\widehat{\tau}_j}^T + \mathbf{D}_{\widehat{\tau}_j}\,\mathbf{D}_{-\widehat{\tau}_j}^T\,\widehat{\mathbf{u}}_{\lambda_j,\,-\widehat{\tau}_j}}{\sigma\,\|\mathbf{D}_{\widehat{\tau}_j}^T\|},$$

$$\mathcal{V}_{Z_j}^{+(glo)} = \frac{\lambda_j\,\mathbf{D}_{\widehat{\tau}_j}\,\mathbf{D}_{\widehat{\tau}_j}^T + \mathbf{D}_{\widehat{\tau}_j}\,\mathbf{D}_{-\widehat{\tau}_j}^T\,\widehat{\mathbf{u}}_{\lambda_j,\,-\widehat{\tau}_j}}{\sigma\,\|\mathbf{D}_{\widehat{\tau}_j}^T\|}. \tag{4.33}$$

(b) The conditional distribution of T_j^{glo} *given* $\{\widehat{\tau}_j \in \widehat{\boldsymbol{\tau}}\}$*, under the null hypothesis, is* t *distribution with* $d^{(glo)} = n - 2(r+1)$ *degrees of freedom and truncated to the set* $\big(-\infty\,,\,\mathcal{V}_{T_j}^{-(glo)}\big] \bigcup \big[\mathcal{V}_{T_j}^{+(glo)}\,,\,\infty\big)$*, where*

$$\mathcal{V}_{T_j}^{-(glo)} = \frac{-\lambda_j\,\mathbf{D}_{\widehat{\tau}_j}\,\mathbf{D}_{\widehat{\tau}_j}^T + \mathbf{D}_{\widehat{\tau}_j}\,\mathbf{D}_{-\widehat{\tau}_j}^T\,\widehat{\mathbf{u}}_{\lambda_j,\,-\widehat{\tau}_j}}{\widehat{\sigma}^{\,(glo)}\,\|\mathbf{D}_{\widehat{\tau}_j}^T\|},$$

$$\mathcal{V}_{T_j}^{+(glo)} = \frac{\lambda_j\,\mathbf{D}_{\widehat{\tau}_j}\,\mathbf{D}_{\widehat{\tau}_j}^T + \mathbf{D}_{\widehat{\tau}_j}\,\mathbf{D}_{-\widehat{\tau}_j}^T\,\widehat{\mathbf{u}}_{\lambda_j,\,-\widehat{\tau}_j}}{\widehat{\sigma}^{\,(glo)}\,\|\mathbf{D}_{\widehat{\tau}_j}^T\|}, \tag{4.34}$$

and $\widehat{\sigma}_j^{2\,(glo)} = \big\|\big(\mathbf{I} - \mathbf{P}_j^{glo}\big)\,\mathbf{y}\big\|^2/d^{(glo)}$.

We note that the TN and Tt statistics, provided in (4.21) and (4.27), respectively, can be constructed using the corresponding distributions in both parts (a) and (b) of Theorem 4.10. Therefore, we can apply the same procedure as in Section 4.4 to compute corresponding post-detection p-values and confidence intervals.

It is worth mentioning that $\widehat{\sigma}_j^{2\,(\mathrm{glo})}$ may not be a suitable estimator of σ^2. This is because global post-detection only assumes that $\widehat{\tau}_j \in \widehat{\boldsymbol{\tau}}$. Therefore, in the case that there exit other change points in $\widehat{\boldsymbol{\tau}}$, $\widehat{\sigma}_j^{2\,(\mathrm{glo})}$ is unable to accurately estimate the variation in the observations. As an alternative, we can apply Median Absolute Deviation (MAD) proposed by [65], to robustly estimate σ^2. Also, see [16] for more details about MAD estimation.

Let $\big[L_{Z_j}^{\mathrm{glo}}\,,\,U_{Z_j}^{\mathrm{glo}}\big]$ and $\big[L_{T_j}^{\mathrm{glo}}\,,\,U_{T_j}^{\mathrm{glo}}\big]$ be confidence intervals derived in the same manner as (4.23) and (4.28) using the pivotal quantities Z_j^{glo} and T_j^{glo}. As the lower and upper

bounds associated with their distributions are unbounded, we suspect that the underlying confidence intervals have finite expected lengths. In such a case, [88] have provided an upper bound for the confidence intervals derived from a truncated normal distribution. In the following theorem, we will also give such an upper bound for confidence intervals derived from a truncated t distribution.

Theorem 4.11 *Let* $\left[L_{Z_j}^{glo}, U_{Z_j}^{glo}\right]$ *and* $\left[L_{T_j}^{glo}, U_{T_j}^{glo}\right]$ *be* $(1-\alpha)\%$ *confidence intervals derived from truncated normal and truncated* t *distributions given in Theorem 4.10.*

(a) The length of confidence intervals derived from the truncated normal distribution is always upper bounded by

$$U_{Z_j}^{glo} - L_{Z_j}^{glo} \overset{a.s.}{\leq} 2\,\sigma\,\Phi^{-1}\left(1-\frac{\alpha}{2}\right) + \mathcal{V}_{Z_j}^{+(glo)} - \mathcal{V}_{Z_j}^{-(glo)}. \tag{4.35}$$

(b) For $d^{(glo)} \geq 3$ *and under the condition*

$$\alpha\; G_{d^{(glo)}}\left(-\frac{\mathcal{V}_{T_j}^{+(glo)} - \mathcal{V}_{T_j}^{-(glo)}}{2\,\widehat{\sigma}_j^{(glo)}}\right) \geq G_{d^{(glo)}}\left(G_{d^{(glo)}}^{-1}\left(\frac{\alpha}{2}\right) - \frac{\mathcal{V}_{T_j}^{+(glo)} - \mathcal{V}_{T_j}^{-(glo)}}{2\,\widehat{\sigma}_j^{(glo)}}\right), \tag{4.36}$$

the length of confidence intervals derived from the truncated t *distribution is upper bounded by*

$$U_{T_j}^{glo} - L_{T_j}^{glo} \overset{a.s.}{\leq} 2\,\widehat{\sigma}_j^{(glo)}\, G_{d^{(glo)}}^{-1}\left(1-\frac{\alpha}{2}\right) + \mathcal{V}_{T_j}^{+(glo)} - \mathcal{V}_{T_j}^{-(glo)}. \tag{4.37}$$

A proof is provided in Appendix B.6.

4.6.2 Local Post-Detection Inference

Unlike in Section 4.6.1, here we are dealing with situations where the most relevant estimated change points to the target change point are involved in our inferential analysis. Obviously, the selection of $\widehat{\tau}_j$ as a change point relies on other estimated change points. Thus, conditioning only on $\{\widehat{\tau}_j \in \widehat{\boldsymbol{\tau}}\}$ is insufficient to decide whether $\widehat{\tau}_j$ is a meaningful change point. This fact indicates the need for knowledge about other change points. However, as previously discussed, conditioning our inference on the entire change point set yields an undesirable output such as wide confidence intervals. Note that the local nature of change point setting [112] dictates that inferences for a change point should depend on

its adjacent change points. For instance, only the immediate neighboring change points $\widehat{\tau}_{j-1}$ and $\widehat{\tau}_{j+1}$ play a role in the detection of the target change point $\widehat{\tau}_j$, and the rest remain irrelevant. This fact has motivated us to develop the local post-detection algorithm.

Following the local property of change points, we suggest a method for post-detection inference which leads to higher-powered tests and shorter confidence intervals. The idea is to condition the post-detection inference on the target change point as well as on its adjacent change points. In particular, for the given target change point $\widehat{\tau}_j$, the goal is to test the hypothesis $H_0 : \mathbf{D}_{\widehat{\tau}_j}\,\mathbf{f} = 0$, given $\big\{\{\widehat{\tau}_{j-1},\, \widehat{\tau}_j,\, \widehat{\tau}_{j+1}\} \in \widehat{\boldsymbol{\tau}}\big\}$. In fact, this conditioning creates a new change point problem over the shorter subsignal $\big(f_{\widehat{\tau}_{j-1}+1},\, \ldots,\, f_{\widehat{\tau}_{j+1}}\big)$. We first define some notations for ease of exposition. Let $\mathbf{y}_j^{\text{loc}} = \big(\mathbf{y}_{j-1}\,,\, \mathbf{y}_j\big) = \big(y_{\widehat{\tau}_{j-1}+1}\,,\, \ldots\,,\, y_{\widehat{\tau}_{j+1}}\big)$ and $\mathbf{f}_j^{\text{loc}} = \big(\mathbf{f}_{j-1}\,,\, \mathbf{f}_j\big) = \big(f_{\widehat{\tau}_{j-1}+1}\,,\, \ldots\,,\, f_{\widehat{\tau}_{j+1}}\big)$ denote the subvectors of $\mathbf{y}$ and $\mathbf{f}$ from $\widehat{\tau}_{j-1}+1$ to $\widehat{\tau}_{j+1}$. Also, let $\mathbf{D}_j$ be the $\big(\widehat{\tau}_{j+1}-\widehat{\tau}_{j-1}-r-1\big)\times\big(\widehat{\tau}_{j+1}-\widehat{\tau}_{j-1}\big)$ version of matrix $\mathbf{D}$. Applying these notations, the hypothesis of interest can be re-expressed as $H_0 : \mathbf{D}_{j,\,\Delta_j}\,\mathbf{f}_j^{\text{loc}} = 0$, where $\Delta_j = \widehat{\tau}_j - \widehat{\tau}_{j-1}$ and $\mathbf{D}_{j,\,\Delta_j}$ is the Δ_j-th row of $\mathbf{D}_j$ and $\mathbf{P}_j^{\text{loc}} = \text{diag}\big(\mathbf{P}_{1,\,j}^{\text{loc}}\,,\, \mathbf{P}_{2,\,j}^{\text{loc}}\big)$. For inference, similar approach taken in the global post-detection can be applied, but this time using $\mathbf{y}_j^{\text{loc}}$ and $\mathbf{D}_j$. Depending on whether σ^2 is known or unknown, the post-detection tests are based on statistics

$$Z_j^{\,\text{loc}} = \frac{\mathbf{D}_{j,\,\Delta_j}\,\mathbf{y}_j^{\text{loc}}}{\sigma\,\big\|\mathbf{D}_{j,\,\Delta_j}\big\|} \qquad , \qquad T_j^{\,\text{loc}} = \frac{\mathbf{D}_{j,\,\Delta_j}\,\mathbf{y}_j^{\text{loc}}}{\widehat{\sigma}_j^{\,(\text{loc})}\big\|\mathbf{D}_{j,\,\Delta_j}\big\|}. \tag{4.38}$$

We obtain the distributions of these statistics in the next theorem whose proof is similar to that of Theorem 4.10 by replacing $\mathbf{y}$, $\mathbf{D}$, $\widehat{\tau}_j$, $\mathbf{P}_j^{\text{glo}}$ and $\widehat{\sigma}_j^{\,2\,(\text{glo})}$ with $\mathbf{y}_j^{\text{loc}}$, $\mathbf{D}_j$, Δ_j, $\mathbf{P}_j^{\text{loc}}$ and $\widehat{\sigma}_j^{\,2\,(\text{loc})}$.

Theorem 4.12 *Consider the statistics $Z_j^{\,\text{loc}}$ and $T_j^{\,\text{loc}}$ defined in (4.38).*

(a) Given $\big\{\{\widehat{\tau}_{j-1}\,,\, \widehat{\tau}_j\,,\, \widehat{\tau}_{j+1}\} \in \widehat{\boldsymbol{\tau}}\big\}$, the conditional distribution of $Z_j^{\,\text{loc}}$, under the null hypothesis is the standard normal truncated to $\big(-\infty\,,\, \mathcal{V}_{Z_j}^{-(\text{loc})}\big] \bigcup \big[\mathcal{V}_{Z_j}^{+(\text{loc})}\,,\, \infty\big)$, where

$$\mathcal{V}_{Z_j}^{-(\text{loc})} = \frac{-\lambda_j\,\mathbf{D}_{j,\,\Delta_j}\,\mathbf{D}_{j,\,\Delta_j}^T \;+\; \mathbf{D}_{j,\,\Delta_j}\,\mathbf{D}_{j,\,-\Delta_j}^T\,\widehat{\mathbf{u}}_{\lambda_j,\,-\widehat{\tau}_j}}{\sigma\,\big\|\mathbf{D}_{j,\,\Delta_j}^T\big\|},$$

$$\mathcal{V}_{Z_j}^{+(\text{loc})} = \frac{\lambda_j\,\mathbf{D}_{j,\,\Delta_j}\,\mathbf{D}_{j,\,\Delta_j}^T \;+\; \mathbf{D}_{j,\,\Delta_j}\,\mathbf{D}_{j,\,-\Delta_j}^T\,\widehat{\mathbf{u}}_{\lambda_j,\,-\widehat{\tau}_j}}{\sigma\,\big\|\mathbf{D}_{j,\,\Delta_j}^T\big\|}. \tag{4.39}$$

(b) *Given* $\big\{\{\widehat{\tau}_{j-1}\,,\,\widehat{\tau}_j\,,\,\widehat{\tau}_{j+1}\}\,\in\,\widehat{\boldsymbol{\tau}}\big\}$, *the conditional distribution of* $T_j^{\rm loc}$, *under the null hypothesis, is* t *distribution with* $d_j^{(\rm loc)} = (\widehat{\tau}_{j+1} - \widehat{\tau}_{j-1}) - 2(r+1)$ *truncated to the set* $\big(-\infty\,,\,\mathcal{V}_{T_j}^{-(\rm loc)}\big]\,\bigcup\,\big[\mathcal{V}_{T_j}^{+(\rm loc)}\,,\,\infty\big)$, *where*

$$\mathcal{V}_{T_j}^{-(\rm loc)} = \frac{-\lambda_j\,\mathbf{D}_{j,\,\Delta_j}\,\mathbf{D}_{j,\,\Delta_j}^T + \mathbf{D}_{j,\,\Delta_j}\,\mathbf{D}_{j,\,-\Delta_j}^T\,\widehat{\mathbf{u}}_{\lambda_j,\,-\widehat{\tau}_j}}{\widehat{\sigma}_j^{\,(loc)}\,\|\mathbf{D}_{j,\,\Delta_j}^T\|},$$

$$\mathcal{V}_{T_j}^{+(\rm loc)} = \frac{\lambda_j\,\mathbf{D}_{j,\,\Delta_j}\,\mathbf{D}_{j,\,\Delta_j}^T + \mathbf{D}_{j,\,\Delta_j}\,\mathbf{D}_{j,\,-\Delta_j}^T\,\widehat{\mathbf{u}}_{\lambda_j,\,-\widehat{\tau}_j}}{\widehat{\sigma}_j^{\,(loc)}\,\|\mathbf{D}_{j,\,\Delta_j}^T\|}, \tag{4.40}$$

and $\widehat{\sigma}_j^{\,2\,(\rm loc)} = \big\|\big(\mathbf{I} - \mathbf{P}_j^{\rm loc}\big)\,\mathbf{y}_j^{\rm loc}\big\|^2/d_j^{(\rm loc)}$.

In the same manner as described in Section 4.4, the TN and Tt statistics can be employed to perform hypothesis testings and to construct confidence intervals.

Remark 4.13 *From Theorem 4.12, it turns out that the lower and upper limits of truncation sets using local post-detection inference are infinite. Therefore, similar to global post-detection inference, the expected lengths of confidence intervals associated with such distributions are upper bounded. In particular, let* $\big[L_{Z_j}^{\rm loc}\,,\,U_{Z_j}^{\rm loc}\big]$ *and* $\big[L_{T_j}^{\rm loc}\,,\,U_{T_j}^{\rm loc}\big]$ *be* $(1-\alpha)\%$ *confidence intervals derived from truncated normal and truncated* t *distributions, respectively, provided in Theorem 4.12. Hence, the lengths of such intervals are upper bounded by*

- *when* σ^2 *is known, the length of confidence intervals is always upper bounded by*

$$U_{Z_j}^{\rm loc} - L_{Z_j}^{\rm loc} \overset{a.s.}{\leq} 2\,\sigma\,\Phi^{-1}\left(1 - \frac{\alpha}{2}\right) + \mathcal{V}_{Z_j}^{+(\rm loc)} - \mathcal{V}_{Z_j}^{-(\rm loc)}\,. \tag{4.41}$$

- *when* σ^2 *is unknown, for* $d^{\,(loc)} \geq 3$ *and under the condition*

$$\alpha\;G_{\,d\,(loc)}\left(-\frac{\mathcal{V}_{T_j}^{+(loc)} - \mathcal{V}_{T_j}^{-(loc)}}{2\,\widehat{\sigma}_j^{\,(loc)}}\right) \geq\; G_{\,d\,(loc)}\left(G_{\,d\,(loc)}^{-1}\left(\frac{\alpha}{2}\right) - \frac{\mathcal{V}_{T_j}^{+(loc)} - \mathcal{V}_{T_j}^{-(loc)}}{2\,\widehat{\sigma}_j^{\,(loc)}}\right), \tag{4.42}$$

the length of confidence intervals is upper bounded by

$$U_{T_j}^{\rm loc} - L_{T_j}^{\rm loc} \overset{a.s.}{\leq} 2\,\widehat{\sigma}_j^{\,(\rm loc)}\,G_{\,d\,(\rm loc)}^{-1}\left(1 - \frac{\alpha}{2}\right) + \mathcal{V}_{T_j}^{+(\rm loc)} - \mathcal{V}_{T_j}^{-(\rm loc)}\,. \tag{4.43}$$

4.7 Numerical Studies

In this section, we investigate the performance of our proposed post-detection inference approaches for a piecewise constant signal, $r = 0$, and a piecewise linear signal, $r = 1$. We compare the performance of the approaches in terms of the empirical power for hypothesis testings and coverage probabilities for confidence intervals. We also apply the proposed methods to the three real datasets, which have been used to estimate change points using the PRUTF algorithm in Chapter 3.

4.7.1 Simulation Study

In order to investigate our proposed approaches for post-detection inference, we consider two signals: piecewise constant and piecewise linear. Suppose that $\mathbf{f}$ is a piecewise constant or linear signal of size $n = 500$, with four change points, $J_0 = 4$, at locations $\boldsymbol{\tau} = \{ 100, 200, 300, 400 \}$. For the piecewise constant signal, we consider a signal with the starting point 0 and the jump sizes δ. That is,

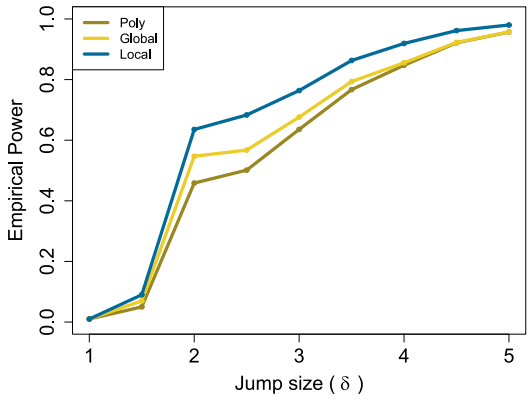
$$f_t^{\text{con}} = \begin{cases} 0 & \text{for} \quad 1 \le t \le 100 \quad \text{or} \quad 201 \le t \le 300 \quad \text{or} \quad 401 \le t \le 500, \\ \delta & \text{for} \quad 101 \le t \le 200 \quad \text{or} \quad 301 \le t \le 400, \end{cases} \tag{4.44}$$

where $\delta \in \{ 1, 1.5, 2, \ldots, 4.5, 5 \}$. Also, for the piecewise linear signal, we consider

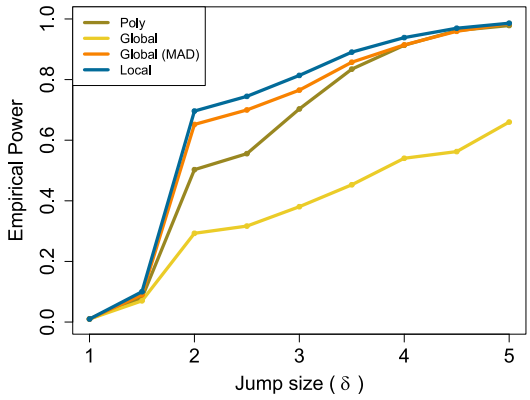
$$f_t^{\text{lin}} = \begin{cases} \delta\,(-0.5 + t) & \text{for} \quad 1 \le t \le 100 \quad \text{or} \quad 201 \le t \le 300 \quad \text{or} \quad 401 \le t \le 500, \\ \delta\,(0.5 - t) & \text{for} \quad 101 \le t \le 200 \quad \text{or} \quad 301 \le t \le 400. \end{cases} \tag{4.45}$$

We have generated a sample $\mathbf{y}$ from $N\big(\mathbf{f}, \sigma^2\,\mathbf{I}\big)$, with $\sigma^2 = 1$, for both piecewise constant and piecewise linear signals, and implemented PRUTF to estimate change points. Next, we have applied our proposed approaches: polyhedron post-detection (*poly*), global post-detection (*global*) and local post-detection (*local*) inference procedures. Over $N = 5000$ repetitions, we have computed the empirical power of a change point hypothesis using the three methods *poly, global* and *local*, for the cases of both known and unknown σ^2. Recall that the empirical power is defined as the ratio between the number of times the true change point is correctly detected and has p-value less than α, and the number of times the true change point is correctly detected, see [78]. Here, α is set to 0.05. The results of the estimation of the second change point, $\tau_2 = 200$, are displayed in Figure 4.3. For the piecewise constant signal, the empirical powers are computed over repetitions that contain

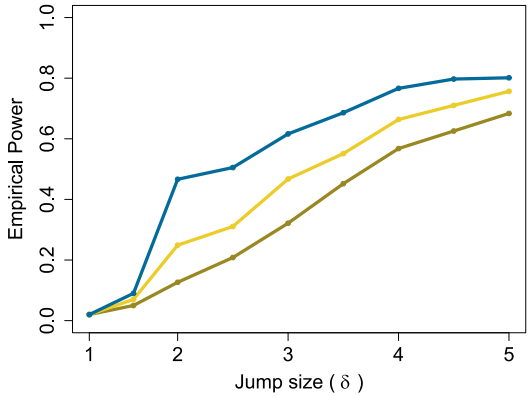
the underlying change point. This is the spike contrast put forward by [77]. Whereas the window contrast, used in [82], with $h = 15$ is employed for the computation in the piecewise



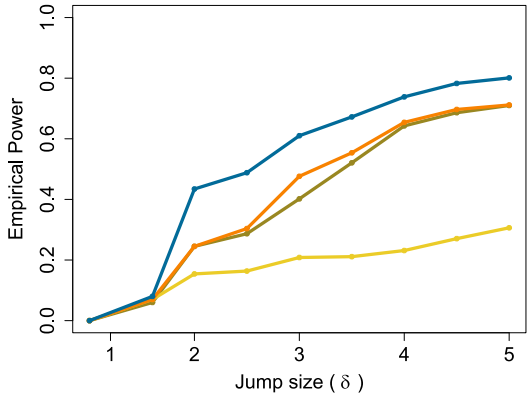
(a) Piecewise constant signal, known σ^2.



(b) Piecewise constant signal, unknown σ^2.



(c) Piecewise linear signal known σ^2.



(d) Piecewise linear signal unknown σ^2.

Figure 4.3: The empirical powers of the three proposed approaches, polyhedron, global and local post-detection inference. The results are provided for both cases when σ^2 is assumed known and unknown. The two top panels display the empirical powers for the piecewise constant signal. Also, the two bottom panels show the empirical powers for the piecewise linear signal. The solid orange lines in both panels (b) and (d) display the empirical powers of the global post-detection approach when $\widehat{\sigma}_j^{2\,(\mathrm{glo})}$, an estimation of σ^2, is replaced with MAD estimation.

linear signal. Window contrast is chosen because PRUTF mostly estimates change points near but not exactly at change points when the polynomial degree r increases.

Panels (a) and (b) of Figure 4.3 provide the empirical powers for the piecewise constant signal with both known and unknown σ^2. When σ^2 is known, the local and global post-detection approaches outperform the polyhedron post-detection approach, as expected. In the case of unknown σ^2, the local post-detection approach again performs better than the polyhedron one. However, the global post-detection approach performs poorly due to inaccurate σ^2 estimation, see Section 4.6.1. By replacing $\widehat{\sigma}_j^{2\,(\mathrm{glo})}$ in Theorem 4.10 with the MAD estimation, the empirical power of the global post-detection approach has improved. The results for the piecewise linear signal are also plotted in panels (c) and (d) of Figure 4.3. Similar patterns are observed for this signal, as well.

Moreover, we have computed the coverage probability for confidence intervals, obtained using the polyhedron, global and local post-detection methods. The outputs are reported in Table 4.1.

		Known σ^2			Unknown σ^2			
Piecewise Constant	δ	Poly	Global	Local	Poly	Global	Global (MAD)	Local
	2	0.9515	0.9527	0.9515	0.9515	0.9708	0.9527	0.9504
	3	0.9554	0.9564	0.9554	0.9554	0.9817	0.9577	0.9564
	4	0.9547	0.9586	0.9547	0.9605	0.9874	0.9601	0.9551
	5	0.9543	0.9531	0.9543	0.9618	0.9889	0.9553	0.9545
		Known σ^2			Unknown σ^2			
Piecewise Linear	δ	Poly	Global	Local	Poly	Global	Global (MAD)	Local
	2	0.9473	0.9543	0.9473	0.9473	0.9660	0.9450	0.9473
	3	0.9550	0.9555	0.9550	0.9570	0.9737	0.9581	0.9555
	4	0.9453	0.9491	0.9453	0.9491	0.9779	0.9514	0.9441
	5	0.9431	0.9419	0.9431	0.9506	0.9786	0.9426	0.9431

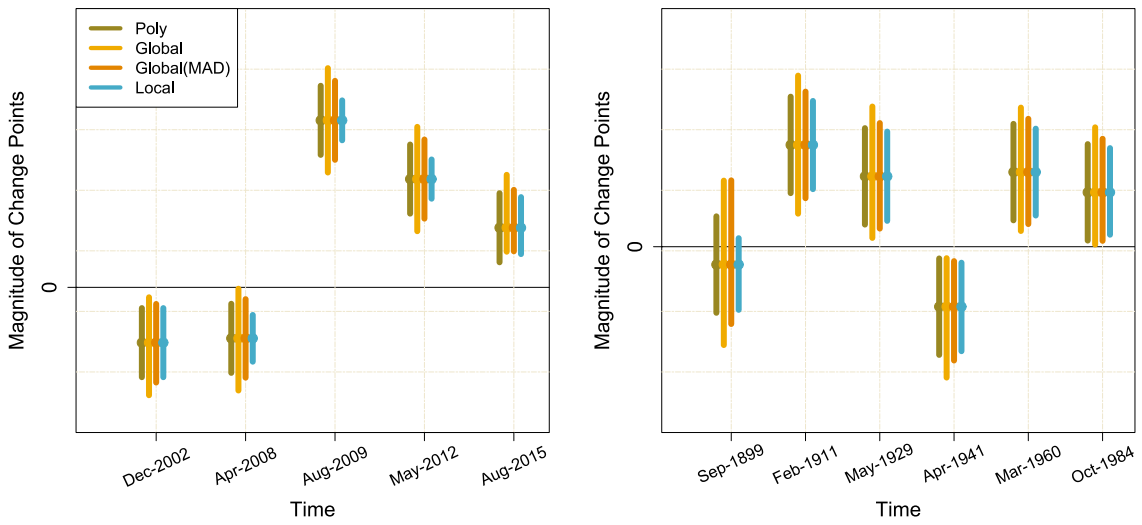
Table 4.1: Coverage probabilities of confidence intervals obtained using polyhedron, global and local post-detection approaches. The results are reported for four values of $\delta \in \{\,2\,,\,3\,,\,4\,,\,5\,\}$.

4.7.2 Real Data Analysis

In Chapter 3, we have analyzed three real datasets: UK House Price Index, the GISS Surface Temperature and COVID-19. In the following, we will apply our proposed post-detection inference approaches to evaluate the significance of change points estimated using PRUTF for the datasets.

Example 4.14 (UK HPI Data) *In Example 3.17, we have applied the mPRUTF algorithm to find change point locations in the UK HPI at Tower Hamlets from January 1996 to November 2018. Our algorithm has estimated five change points located at months: December 2002, April 2008 and August 2009, May 2012 and August 2015. We have provided the results of post-detection inference for these change points in panel (a) of Figure 4.4. The plot displays %95 post-detection confidence intervals for the change points. Based on the polyhedron and local post-detection results, we have concluded that all five estimated change points are significant as their corresponding intervals excluded zero.*

Example 4.15 (GISTEMP Data) *We have provided the change point results of the mPRUTF algorithm for the GISTEMP dataset in Example 3.18. The example has considered the monthly land-ocean temperature anomalies recorded from January 1880 to August 2019. For this dataset, the algorithm has estimated six change points located at months: September 1899, February 1911, May 1929, April 1941, March 1960, October 1984. We*



(a) CIs For HPI Data. **(b)** CIs For GISTEMP Data.

Figure 4.4: Valid post-detection confidence intervals for both the UK HPI dataset (left panel) and GISTEMP dataset (right panel). For each estimated change point, valid %95 post-detection confidence intervals, obtained from polyhedron, global using both pooled and MAD variances, and local post-detection inference approaches, are displayed.

have applied our proposed post-detection inference approaches to assess the significance of these estimated change points by computing the associated confidence intervals. The results are provided in Figure 4.4, panel (b). The outputs obtained from both polyhedron and local post-detection inference have confirmed the significance of all estimated change points except the one in September 1899.

Example 4.16 (COVID-19 Data) *The logarithm of the cumulative daily number of COVID-19 confirmed cases has been analyzed in Chapter 3 using a piecewise linear model. The choice of the piecewise linear model is natural because the slope of each segment, estimated using the detected change points, indicates the growth rate of the COVID-19 virus. Consequently, these slopes allow us to compare the virus growth rate among estimated segments and evaluate the effectiveness of undertaken strategies. Additionally, the linear trend of the last segment can be used to predict the status of the pandemic for future dates.*

In Example 3.19, we have applied the mPRUTF algorithm to detect change points that have occurred in the transformed COVID-19 datasets from March 10, 2020 until April 30, 2021, for Australia, Canada, the United Kingdom and the United States. We have applied our proposed post-detection inference approaches to these datasets. The %95 post-detection confidence intervals for the estimated change points are provided in Figure 4.5. For example, for Canada, based on the polyhedron and local post-detection inference approaches, the change points located on March 26, 2020; April 9, 2020; May 11, 2020; August 31, 2020 and January 12, 2021 are significant. For the United Kingdom, based on the polyhedron and local post-detection approaches, only the change point located on June 22, 2020 is insignificant. Moreover, the figure shows that the confidence intervals for March 18, 2021 in Canada and for February 23, 2021 in the United States derived from polyhedron post-detection method are very wide and skewed. These observations certify the results provided in Theorem 4.7.

4.8 More on Post-Detection Inference Versus Classical Inference

As discussed in Section 2.3, classical inference computes p-values and constructs confidence intervals for the selected change points without accounting for the fact that the data have been used in estimating these change points. Therefore, it seems clear that this naive inference is problematic. For example, the confidence intervals constructed by using

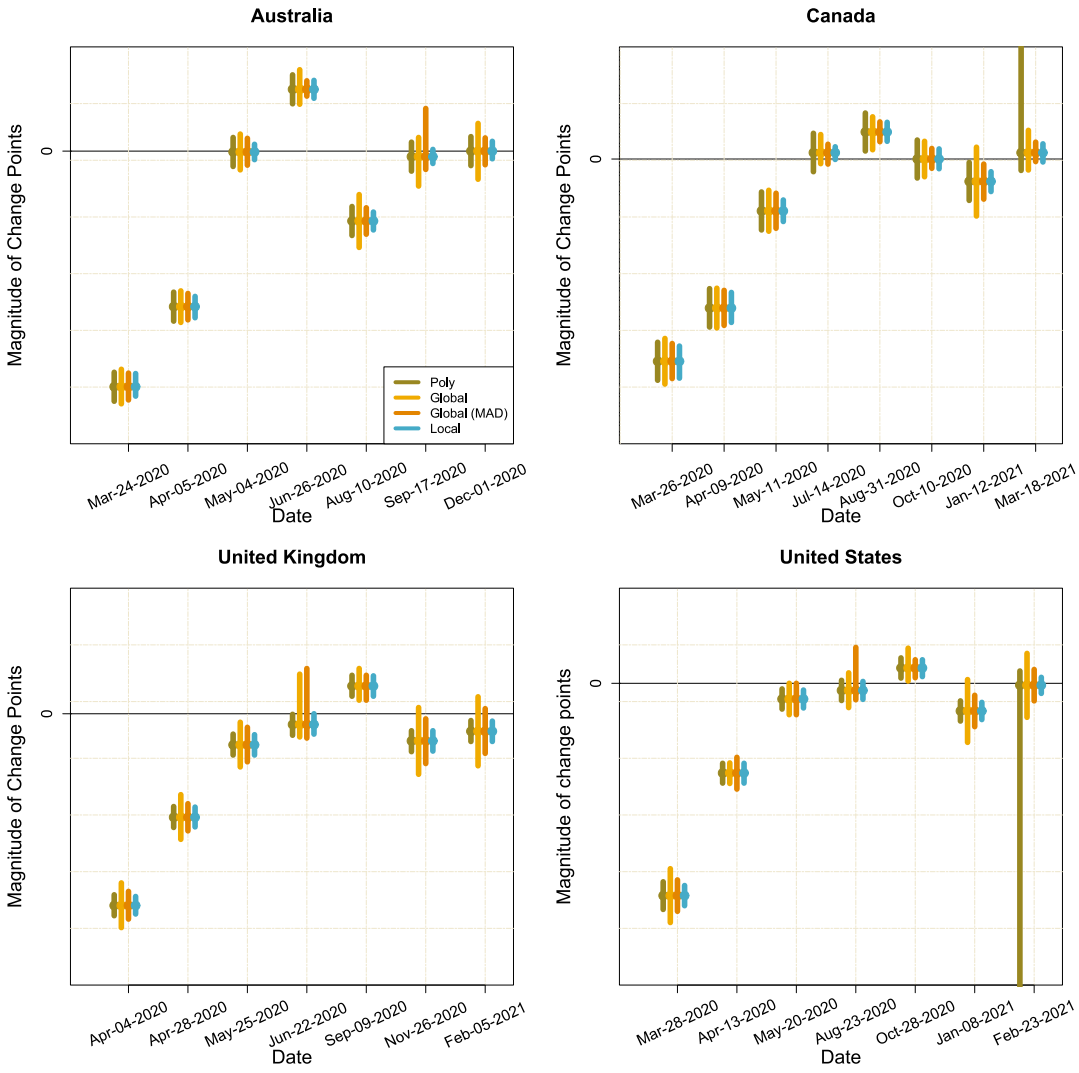
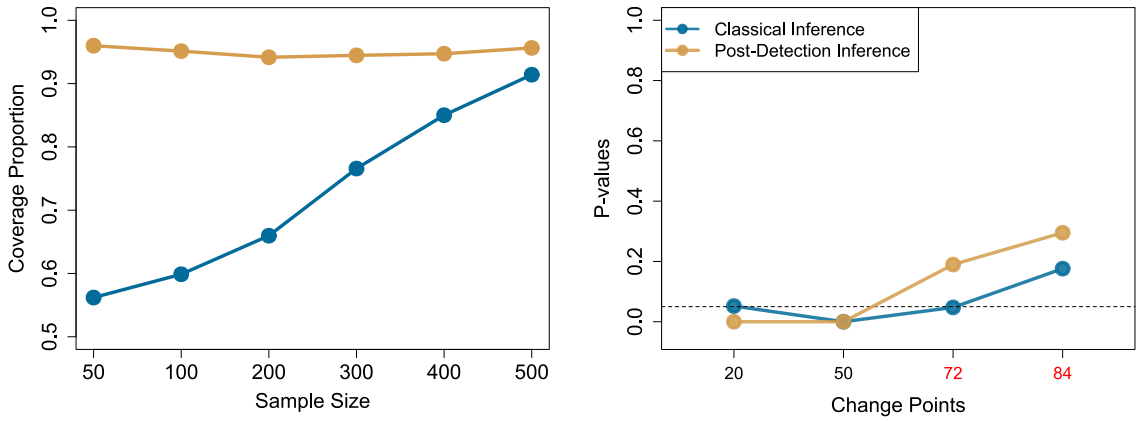


Figure 4.5: Valid post-detection confidence intervals for COVID-19 datasets.. For each estimated change point, % 95 post-detection confidence intervals obtained from polyhedron, global with both pooled and MAD variance estimations, and local post-detection inference approaches are displayed.

classical inference are narrower and have smaller coverage probabilities. In other words, classical confidence intervals with nominal $1-\alpha$ coverage, may no longer cover the target parameter at this level. Many researchers including [96, 22, 143, 93, 174] and [77] have

investigated this problem in the literature. For more details, see Section 2.3. In the following, we empirically show that classical confidence intervals may fail to have the correct coverage properties.

We first conduct a simulation study as follows. We consider piecewise constant signals with various sizes n, but with four change points, equally spaced along each signal, that is, $\boldsymbol{\tau} = \{\frac{n}{5}, \frac{2n}{5}, \frac{3n}{5}, \frac{4n}{5}\}$. We simulate 1000 samples from the model in (4.2) and run the mPRUTF algorithm for each sample to estimate change points. We then construct confidence intervals using both classical (naive Z-tests) and the local post-detection methods. Panel (a) of Figure 4.6 shows the coverage proportion of 95% classical confidence intervals and 95% of post-detection confidence intervals for various signal sizes. The results show that the coverage proportion of the post-detection confidence intervals are always near the nominal level 0.95, whereas the coverage proportions for the classical confidence intervals may be far below 0.95, specially when the signal sizes are small. However, as the signal size increases, the coverage proportions of the classical confidence intervals approaches this nominal confidence level.



(a) Coverage proportions for the simulation study.

(b) P-values for the simulated data.

Figure 4.6: Plots of the coverage proportions across a range of sample size (panel a) as well as p-values for the simulated example (panel b). As can be seen from panel (a), the coverage proportion of the classical confidence intervals can be far below the nominal level of 0.95. While the post-detection confidence intervals always have coverage proportion 0.95. In panel (b), for the false detected change points at 72 and 84 (shown in red), p-values derived from post-detection method are larger than those of the classical method.

Additionally, we compare the performance of both classical and post-detection inference for a simulated example. To this end, we generate a sample from a piecewise constant signal with $n = 100$ and two true change points at $\boldsymbol{\tau} = \{20, 50\}$. For this example, we do not apply the stopping rule in mPRUTF, and run the algorithm to detect four change points. Thus, in addition to the true change points, mPRUTF has detected two false change points located at 72 and 84. If we treat the detected change points as fixed and naively ignore their adaptive detection, then the usual Z-test would be a natural choice for determining the significance of each change point. Panel (b) of Figure 4.6 reports p-values computed from the naive Z-test as well as those from the post-detection inference. As seen from the figure, both classical and post-detection methods produce small p-values for the true change points 20 and 50. For the false detected change point located at 72 and 84, the post-detection method gives p-values away from 0.05 and correctly rejects the significance of these points. However, p-values obtained by classical inference are closer to 0.05, specifically the one for location 72, and may lead to incorrectly verifying it as a change point.

Chapter 5

Conclusions and Future Research

5.1 Conclusions

This thesis has introduced a novel methodology based on trend filtering to detect change points in a mean model. The PRUTF method has been developed to estimate the number and locations of change points in piecewise polynomial signals. Unlike the vast majority of methods currently available in the literature, PRUTF extends the problem of mean change point detection, going beyond just focusing on piecewise constant signals. Once able to estimate change points using PRUTF, we proposed polyhedron, global and local post-detection methods to test the significance of the estimates. To evaluate the performance of the proposed methods, we have executed simulation studies and compared the methods to some state-of-the-art approaches in the literature.

In Chapter 3, we proposed the PRUTF method for detecting change points in piecewise polynomial signals by using trend filtering. We demonstrated that the dual solution path produced by the PRUTF algorithm forms a Gaussian bridge process for any given value of the regularization parameter λ. This conclusion has allowed us to derive an efficient stopping rule for terminating the search algorithm, which is vital in the change point analysis. We then proved that when there is no staircase block in the signal, the method guarantees consistent pattern recovery. However, it fails to do so when there is a staircase in the underlying signal. To address this shortcoming, we have suggested a modification in the procedure of constructing the solution path, one that effectively prevents false change-point discovery. Evidence from both simulation studies and real-world data analyses confirms the accuracy and high detection power of the proposed method.

In Chapter 4, we attempted to quantify the uncertainty of change points estimated using PRUTF. We have provided a post-detection inference framework to compute valid p-values for the significance of change points estimated with PRUTF. We have also constructed confidence intervals for the magnitude of such estimated change points. These inferences have been executed by implementing conditional inferences through three distinct conditional detection events, giving the polyhedron, global and local post-detection inference procedures. We have also shown that these global and local post-detection inferences lead to higher-power tests by conditioning on much smaller detection events.

5.2 Future Research

For future work in this area, there are several possible ideas. We want to apply our proposed post-detection approaches to the extensively used change point detection algorithms such as binary segmentation, extensions, and ℓ_0-norm segmentation. Moreover, our focus in Chapters 3 and 4 has been on the independently and identically distributed random noises to precisely highlight the main ideas. However, our proposed approaches are not confined only to these types of random noises. One possible extension would be to incorporate more complex random noises. For example, we can consider heavy-tailed distributions or even auto-correlated random noises. The extension of the proposed approaches will allow us to provide inferential tools for a wide range of applications.

Our extension of the change point analysis to the generalized linear model framework is the subject of a forthcoming paper. As discussed, Gaussianity is a crucial assumption for the analysis in both Chapters 3 and 4. This assumption is required for developing the stopping rule in PRUTF and deriving the exact distributions of the test statistics for post-detection inferences. Regarding the Gaussianity assumption, a natural question arises: what happens if our model does not satisfy the Gaussian assumption? To address this question, we are working on the change point analysis of the canonical parameter of an exponential family with a piecewise polynomial structure.

The extension of PRUTF to dependent random noises is another interesting research topic. We should keep in mind that different types of random noises do not impact the dual solution path of trend filtering. However, the distribution of random noises plays a crucial role in the stopping rule of PRUTF. This is because the stopping rule is built based on the Gaussian bridge processes. As our real examples might be weakly dependent, we have executed simulation studies to explore the influence of weakly dependent random noises on PRUTF. Based on these simulation results, provided in Section 3.10, it appears

that PRUTF is robust against dependent noises. More detailed studies of PRUTF for dependent random noises will be the subject of our future research.

References

[1] Robert J Adler and Jonathan E Taylor. *Random fields and geometry*. Springer, 2009.

[2] Alnur Ali, Ryan J Tibshirani, et al. The generalized lasso problem and uniqueness. *Electronic Journal of Statistics*, 13(2):2307–2347, 2019.

[3] 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.

[4] Andreas Anastasiou and Piotr Fryzlewicz. Detecting multiple generalized change-points by isolating single ones. *arXiv preprint arXiv:1901.10852*, 2019.

[5] Isaiah Andrews, Toru Kitagawa, and Adam McCloskey. Inference on winners. Technical report, National Bureau of Economic Research, 2019.

[6] Taylor B Arnold and Ryan J Tibshirani. Efficient implementations of the generalized lasso dual path algorithm. *Journal of Computational and Graphical Statistics*, 25(1):1–27, 2016.

[7] Alexander Aue, Siegfried Hörmann, Lajos Horváth, Matthew Reimherr, et al. Break detection in the covariance structure of multivariate time series models. *The Annals of Statistics*, 37(6B):4046–4087, 2009.

[8] Alexander Aue and Lajos Horváth. Structural breaks in time series. *Journal of Time Series Analysis*, 34(1):1–16, 2013.

[9] Ivan E Auger and Charles E Lawrence. Algorithms for the optimal identification of segment neighborhoods. *Bulletin of mathematical biology*, 51(1):39–54, 1989.

[10] Valeriy Avanesov, Nazar Buzun, et al. Change-point detection in high-dimensional covariance structure. *Electronic Journal of Statistics*, 12(2):3254–3294, 2018.

[11] José A Azar, Jean-Francois Kagy, and Martin C Schmalz. Can changes in the cost of carry explain the dynamics of corporate "cash" holdings? *The Review of Financial Studies*, 29(8):2194–2240, 2016.

[12] François Bachoc, David Preinerstorfer, Lukas Steinberger, et al. Uniformly valid confidence intervals post-model-selection. *The Annals of Statistics*, 48(1):440–463, 2020.

[13] Jushan Bai. Estimating multiple breaks one at a time. *Econometric theory*, 13(3):315–352, 1997.

[14] Jushan Bai. Common breaks in means and variances for panel data. *Journal of Econometrics*, 157(1):78–92, 2010.

[15] Jushan Bai and Pierre Perron. Computation and analysis of multiple structural change models. *Journal of applied econometrics*, 18(1):1–22, 2003.

[16] Rafal Baranowski, Yining Chen, and Piotr Fryzlewicz. Narrowest-over-threshold detection of multiple change points and change-point-like features. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(3):649–672, 2019.

[17] Lawrence Bardwell, Paul Fearnhead, et al. Bayesian detection of abnormal segments in multiple time series. *Bayesian Analysis*, 12(1):193–218, 2017.

[18] Michele Basseville, Igor V Nikiforov, et al. *Detection of abrupt changes: theory and application*, volume 104. Prentice hall Englewood Cliffs, 1993.

[19] L Beghin and E Orsingher. On the maximum of the generalized brownian bridge. *Lithuanian Mathematical Journal*, 39(2):157–167, 1999.

[20] Patrick Bélisle, Lawrence Joseph, Brenda MacGibbon, David B Wolfson, and Roxane Du Berger. Change-point analysis of neuron spike train data. *Biometrics*, pages 113–123, 1998.

[21] Yoav Benjamini, Yotam Hechtlinger, and Philip B Stark. Confidence intervals for selected parameters. *arXiv preprint arXiv:1906.00505*, 2019.

[22] Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, Linda Zhao, et al. Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837, 2013.

[23] Marco Bianchi, Martin Boyle, and Deirdre Hollingsworth. A comparison of methods for trend estimation. *Applied Economics Letters*, 6(2):103–109, 1999.

[24] Patrick Billingsley. *Convergence of probability measures*. John Wiley and Sons, 2013.

[25] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

[26] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[27] Haim Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Springer Science & Business Media, 2010.

[28] E Brodsky and Boris S Darkhovsky. *Non-parametric statistical diagnosis: problems and methods*, volume 509. Springer Science & Business Media, 2013.

[29] Jie Chen and Arjun K Gupta. *Parametric statistical change point analysis: with applications to genetics, medicine, and finance*. Springer Science & Business Media, 2011.

[30] Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159, 2001.

[31] Haeran Cho and Piotr Fryzlewicz. Multiscale and multilevel technique for consistent segmentation of nonstationary time series. *Statistica Sinica*, pages 207–229, 2012.

[32] Haeran Cho and Piotr Fryzlewicz. Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(2):475–507, 2015.

[33] Haeran Cho and Piotr Fryzlewicz. Multiple change point detection under serial dependence: Wild energy maximisation and gappy schwarz criterion. *arXiv preprint arXiv:2011.13884*, 2020.

[34] Haeran Cho and Claudia Kirch. Data segmentation algorithms: Univariate mean change and beyond. *arXiv preprint arXiv:2012.12814*, 2020.

[35] Gabriela Ciuperca. A general criterion to determine the number of change-points. *Statistics & Probability Letters*, 81(8):1267–1275, 2011.

[36] Gabriela Ciuperca. Model selection by lasso methods in a change-point model. *Statistical Papers*, 55(2):349–374, 2014.

[37] David R Cox. A note on data-splitting for the evaluation of significance levels. *Biometrika*, 62(2):441–444, 1975.

[38] Miklos Csorgo and Lajos Horváth. *Limit theorems in change-point analysis*. John Wiley & Sons Chichester, 1997.

[39] IC Demetriou and EA Lipitakis. Certain positive definite submatrices that arise from binomial coefficient matrices. *Applied numerical mathematics*, 36(2-3):219–229, 2001.

[40] Holger Dette, Theresa Eckle, and Mathias Vetter. Multiscale change point detection for dependent data. *Scandinavian Journal of Statistics*, 47(4):1243–1274, 2020.

[41] 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.

[42] Holger Dette, Guangming Pan, and Qing Yang. Estimating a change point in a sequence of very high-dimensional covariance matrices. *Journal of the American Statistical Association*, pages 1–11, 2020.

[43] Francis X Diebold and Glenn D Rudebusch. Measuring business cycles: A modern perspective. Technical report, National Bureau of Economic Research, 1994.

[44] David L Donoho et al. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

[45] Monroe David Donsker. *An invariance principle for certain probability limit theorems*. 1951.

[46] Vo Nguyen Le Duy, Hiroki Toda, Ryota Sugiyama, and Ichiro Takeuchi. Computing valid p-value for optimal changepoint by selective inference using dynamic programming. *arXiv preprint arXiv:2002.09132*, 2020.

[47] Idris A Eckley, Paul Fearnhead, and Rebecca Killick. Analysis of changepoint models. *Bayesian Time Series Models*, pages 205–224, 2011.

[48] Farida Enikeeva, Zaid Harchaoui, et al. High-dimensional change-point detection under sparse alternatives. *Annals of statistics*, 47(4):2051–2079, 2019.

[49] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

[50] Paul Fearnhead, Robert Maidstone, and Adam Letchford. Detecting changes in slope with an l 0 penalty. *Journal of Computational and Graphical Statistics*, 28(2):265–275, 2019.

[51] William Fithian, Dennis Sun, and Jonathan Taylor. Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*, 2014.

[52] William Fithian, Jonathan Taylor, Robert Tibshirani, and Ryan Tibshirani. Selective sequential model selection. *arXiv preprint arXiv:1512.02565*, 2015.

[53] Klaus Frick, Axel Munk, and Hannes Sieling. Multiscale change point inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(3):495–580, 2014.

[54] Piotr Fryzlewicz. Detecting possibly frequent change-points: Wild binary segmentation 2 and steepest-drop model selection. *arXiv preprint arXiv:1812.06880*, 2018.

[55] Piotr Fryzlewicz. Narrowest significance pursuit: inference for multiple change-points in linear models. *arXiv preprint arXiv:2009.05431*, 2020.

[56] Piotr Fryzlewicz et al. Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*, 42(6):2243–2281, 2014.

[57] Piotr Fryzlewicz et al. Tail-greedy bottom-up data decompositions and fast multiple change-point detection. *The Annals of Statistics*, 46(6B):3390–3421, 2018.

[58] Wenjiang J Fu. Penalized regressions: the bridge versus the lasso. *Journal of computational and graphical statistics*, 7(3):397–416, 1998.

[59] Andreas Futschik, Thomas Hotz, Axel Munk, and Hannes Sieling. Multiscale dna partitioning: statistical evidence for segments. *Bioinformatics*, 30(16):2255–2262, 2014.

[60] Enric Galceran, Alexander G Cunningham, Ryan M Eustice, and Edwin Olson. Multipolicy decision-making for autonomous driving via changepoint-based behavior prediction: Theory and experiment. *Autonomous Robots*, 41(6):1367–1382, 2017.

[61] Dario Gasbarra, Tommi Sottinen, and Esko Valkeila. Gaussian bridges. In *Stochastic analysis and applications*, pages 361–382. Springer, 2007.

[62] Edit Gombay, Lajos Horváth, and Marie Husková. Estimators and tests for change in variances. *Statistics & Risk Modeling*, 14(2):145–160, 1996.

[63] John R Graham and Mark T Leary. The evolution of corporate cash. *The Review of Financial Studies*, 31(11):4288–4344, 2018.

[64] Peter J Green and Bernard W Silverman. *Nonparametric regression and generalized linear models: a roughness penalty approach*. Chapman and Hall/CRC, 1993.

[65] Frank R Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393, 1974.

[66] Bruce E Hansen. The new econometrics of structural change: dating breaks in us labour productivity. *Journal of Economic perspectives*, 15(4):117–128, 2001.

[67] James Hansen, Reto Ruedy, Mki Sato, and Ken Lo. Global surface temperature change. *Reviews of Geophysics*, 48(4), 2010.

[68] Zaıd Harchaoui and Céline Lévy-Leduc. Multiple change-point estimation with a total variation penalty. *Journal of the American Statistical Association*, 105(492):1480–1493, 2010.

[69] Douglas M Hawkins. Fitting multiple change-point models to data. *Computational Statistics & Data Analysis*, 37(3):323–341, 2001.

[70] Kaylea Haynes, Paul Fearnhead, and Idris A Eckley. A computationally efficient nonparametric approach for changepoint detection. *Statistics and Computing*, 27(5):1293–1305, 2017.

[71] David V Hinkley and Elizabeth A Hinkley. Inference about the change-point in a sequence of binomial variables. *Biometrika*, 57(3):477–488, 1970.

[72] Holger Hoefling. A path algorithm for the fused lasso signal approximator. *Journal of Computational and Graphical Statistics*, 19(4):984–1006, 2010.

[73] Lajos Horváth and Marie Hušková. Change-point detection in panel data. *Journal of Time Series Analysis*, 33(4):631–648, 2012.

[74] Lajos Horváth and Gregory Rice. Extensions of some classical methods in change point analysis. *Test*, 23(2):219–255, 2014.

[75] WD Hoskins and PJ Ponzo. Some properties of a class of band matrices. *Mathematics of Computation*, 26(118):393–400, 1972.

[76] Tao Huang, Baolin Wu, Paul Lizardi, and Hongyu Zhao. Detection of dna copy number alterations using penalized least squares regression. *Bioinformatics*, 21(20):3811–3817, 2005.

[77] Sangwon Hyun, Max G'Sell, Ryan J Tibshirani, et al. Exact post-selection inference for the generalized lasso path. *Electronic Journal of Statistics*, 12(1):1053–1097, 2018.

[78] Sangwon Hyun, Kevin Lin, Max G'Sell, and Ryan J Tibshirani. Post-selection inference for changepoint detection algorithms with application to copy number variation data. *arXiv preprint arXiv:1812.03644*, 2018.

[79] Brad Jackson, Jeffrey D Scargle, David Barnes, Sundararajan Arabhi, Alina Alt, Peter Gioumousis, Elyus Gwin, Paungkaew Sangtrakulcharoen, Linda Tan, and Tun Tao Tsai. An algorithm for optimal partitioning of data on an interval. *IEEE Signal Processing Letters*, 12(2):105–108, 2005.

[80] Nicholas A James and David S Matteson. Change points via probabilistically pruned objectives. *arXiv preprint arXiv:1505.04302*, 2015.

[81] X Jessie Jeng, T Tony Cai, and Hongzhe Li. Simultaneous discovery of rare and common segment variants. *Biometrika*, 100(1):157–172, 2013.

[82] Sean Jewell, Paul Fearnhead, and Daniela Witten. Testing for a change in mean after changepoint detection. *arXiv preprint arXiv:1910.04291*, 2019.

[83] Jinzhu Jia, Karl Rohe, et al. Preconditioning the lasso for sign consistency. *Electronic Journal of Statistics*, 9(1):1150–1172, 2015.

[84] Moritz Jirak et al. Uniform change point tests in high dimension. *Annals of Statistics*, 43(6):2451–2483, 2015.

[85] Rebecca Killick, Paul Fearnhead, and Idris A Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598, 2012.

[86] Seung-Jean Kim, Kwangmoo Koh, Stephen Boyd, and Dimitry Gorinevsky. \ell_1 trend filtering. *SIAM review*, 51(2):339–360, 2009.

[87] Claudia Kirch, Silke Weber, et al. Modified sequential change point procedures based on estimating functions. *Electronic Journal of Statistics*, 12(1):1579–1613, 2018.

[88] Danijel Kivaranovic and Hannes Leeb. On the length of post-model-selection confidence intervals conditional on polyhedral constraints. *Journal of the American Statistical Association*, (just-accepted):1–33, 2020.

[89] T Suneel Kumar, Vivek Kanhangad, and Ram Bilas Pachori. Classification of seizure and seizure-free eeg signals using multi-level local patterns. In *2014 19th International Conference on Digital Signal Processing*, pages 646–650. IEEE, 2014.

[90] Marc Lavielle. Using penalized contrasts for the change-point problem. *Signal processing*, 85(8):1501–1510, 2005.

[91] Chung-Bow Lee. Estimating the number of change points in a sequence of independent normal random variables. *Statistics and probability letters*, 25(3):241–248, 1995.

[92] Chung-Bow Lee. Estimating the number of change points in exponential families distributions. *Scandinavian Journal of Statistics*, 24(2):201–210, 1997.

[93] Jason D Lee, Dennis L Sun, Yuekai Sun, Jonathan E Taylor, et al. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016.

[94] Youngjo Lee and Hee-Seok Oh. A new sparse variable selection via random-effect model. *Journal of Multivariate Analysis*, 125:89–99, 2014.

[95] Hannes Leeb and Benedikt M Pötscher. The finite-sample distribution of post-model-selection estimators and uniform versus nonuniform approximations. *Econometric Theory*, 19(1):100–142, 2003.

[96] Hannes Leeb and Benedikt M Pötscher. Can one estimate the conditional distribution of post-model-selection estimators? *The Annals of Statistics*, 34(5):2554–2591, 2006.

[97] Hannes Leeb and Benedikt M Pötscher. Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory*, 24(2):338–376, 2008.

[98] Nathan JL Lenssen, Gavin A Schmidt, James E Hansen, Matthew J Menne, Avraham Persin, Reto Ruedy, and Daniel Zyss. Improvements in the uncertainty model in the goddard institute for space studies surface temperature (gistemp) analysis. *Journal of Geophysical Research: Atmospheres*, 2019.

[99] Housen Li, Axel Munk, Hannes Sieling, et al. Fdr-control in multiscale change-point segmentation. *Electronic Journal of Statistics*, 10(1):918–959, 2016.

[100] Dedi Liu, Xiaohong Chen, Yanqing Lian, and Zhanghua Lou. Impacts of climate change and human activities on surface runoff in the dongjiang river basin of china. *Hydrological Processes: An International Journal*, 24(11):1487–1495, 2010.

[101] Keli Liu, Jelena Markovic, and Robert Tibshirani. More powerful post-selection inference, with application to the lasso. *arXiv preprint arXiv:1801.09037*, 2018.

[102] Richard Lockhart, Jonathan Taylor, Ryan J Tibshirani, and Robert Tibshirani. A significance test for the lasso. *Annals of statistics*, 42(2):413, 2014.

[103] Joshua R Loftus and Jonathan E Taylor. A significance test for forward stepwise model selection. *arXiv preprint arXiv:1405.3920*, 2014.

[104] Alexandre Lung-Yut-Fong, Céline Lévy-Leduc, and Olivier Cappé. Distributed detection/localization of change-points in high-dimensional network traffic data. *Statistics and Computing*, 22(2):485–496, 2012.

[105] Robert Maidstone, Toby Hocking, Guillem Rigaill, and Paul Fearnhead. On optimal multiple changepoint algorithms for large data. *Statistics and Computing*, 27(2):519–533, 2017.

[106] Julien Mairal and Bin Yu. Complexity analysis of the lasso regularization path. *arXiv preprint arXiv:1205.0079*, 2012.

[107] Enno Mammen, Sara van de Geer, et al. Locally adaptive regression splines. *The Annals of Statistics*, 25(1):387–413, 1997.

[108] Reza Mehrizi and Shoja Chenouri. Detection of change points in piecewise polynomial signals using trend filtering. *Electronic Journal of Statistics*, 9(1):1150–1172, 2020.

[109] Alan Miller. *Subset selection in regression*. Chapman and Hall/CRC, 2002.

[110] Jessica Minnier, Lu Tian, and Tianxi Cai. A perturbation method for inference on regularized regression estimates. *Journal of the American Statistical Association*, 106(496):1371–1382, 2011.

[111] Chi Tim Ng, Woojoo Lee, Youngjo Lee, et al. Change-point estimators with true identification property. *Bernoulli*, 24(1):616–660, 2018.

[112] Yue S Niu and Heping Zhang. The screening and ranking algorithm to detect dna copy number variations. *The annals of applied statistics*, 6(3):1306, 2012.

[113] Yosihiko Ogata. Detection of anomalous seismicity as a stress change sensor. *Journal of Geophysical Research: Solid Earth*, 110(B5), 2005.

[114] Adam B Olshen, ES Venkatraman, Robert Lucito, and Michael Wigler. Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics*, 5(4):557–572, 2004.

[115] Michael R Osborne, Brett Presnell, and Berwin A Turlach. On the lasso and its dual. *Journal of Computational and Graphical statistics*, 9(2):319–337, 2000.

[116] Laurent Oudre, Alexandre Lung-Yut-Fong, and Pascal Bianchi. Segmentation of accelerometer signals recorded during continuous treadmill walking. In *2011 19th European Signal Processing Conference*, pages 1564–1568. IEEE, 2011.

[117] ES Page. A test for a change in a parameter occurring at an unknown point. *Biometrika*, 42(3/4):523–527, 1955.

[118] Ewan S Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954.

[119] Jianmin Pan and Jiahua Chen. Application of modified information criterion to multiple change point problems. *Journal of multivariate analysis*, 97(10):2221–2241, 2006.

[120] Florian Pein, Hannes Sieling, and Axel Munk. Heterogeneous change point inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1207–1227, 2017.

[121] Gianni B Pezzatti, Thomas Zumbrunnen, Matthias Bürgi, Paolo Ambrosetti, and Marco Conedera. Fire regime shifts as a consequence of fire policy and socio-economic development: an analysis based on the change point approach. *Forest Policy and Economics*, 29:7–18, 2013.

[122] Benedikt M Pötscher. Effects of model selection on inference. *Econometric Theory*, 7(2):163–185, 1991.

[123] Junyang Qian and Jinzhu Jia. On stepwise pattern recovery of the fused lasso. *Computational Statistics & Data Analysis*, 94:221–237, 2016.

[124] Aaditya Ramdas and Ryan J Tibshirani. Fast and flexible admm algorithms for trend filtering. *Journal of Computational and Graphical Statistics*, 25(3):839–858, 2016.

[125] Ananth Ranganathan. Pliss: labeling places using online changepoint detection. *Autonomous Robots*, 32(4):351–368, 2012.

[126] Stephen Reid, Jonathan Taylor, and Robert Tibshirani. Post-selection point and interval estimation of signal sizes in gaussian samples. *Canadian Journal of Statistics*, 45(2):128–148, 2017.

[127] Guillem Rigaill. A pruned dynamic programming algorithm to recover the best segmentations with 1 to k_max change-points. *Journal de la Société Française de Statistique*, 156(4):180–205, 2015.

[128] Alessandro Rinaldo. Corrections to properties and refinements of the fused lasso. 2014.

[129] Alessandro Rinaldo et al. Properties and refinements of the fused lasso. *The Annals of Statistics*, 37(5B):2922–2952, 2009.

[130] Alessandro Rinaldo, Daren Wang, Qin Wen, Rebecca Willett, and Yi Yu. Localizing changes in high-dimensional regression models. In *International Conference on Artificial Intelligence and Statistics*, pages 2089–2097. PMLR, 2021.

[131] Michael W Robbins, Robert B Lund, Colin M Gallagher, and QiQi Lu. Changepoints in the north atlantic tropical cyclone record. *Journal of the American Statistical Association*, 106(493):89–99, 2011.

[132] Cristian R Rojas and Bo Wahlberg. On change point detection using the fused lasso method. *arXiv preprint arXiv:1401.5408*, 2014.

[133] Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268, 1992.

[134] Eric Ruggieri. A bayesian approach to detecting change points in climatic records. *International Journal of Climatology*, 33(2):520–528, 2013.

[135] Galen R Shorack and Jon A Wellner. *Empirical processes with applications to statistics*. SIAM, 2009.

[136] Vasilios A Siris and Fotini Papagalou. Application of anomaly detection algorithms for detecting syn flooding attacks. In *IEEE Global Telecommunications Conference, 2004. GLOBECOM'04.*, volume 4, pages 2050–2054. IEEE, 2004.

[137] Yong Sheng Soh and Venkat Chandrasekaran. High-dimensional change-point estimation: Combining filtering with convex optimization. *Applied and Computational Harmonic Analysis*, 43(1):122–147, 2017.

[138] Won Son and Johan Lim. Modified path algorithm of fused lasso signal approximator for consistent recovery of change points. *Journal of Statistical Planning and Inference*, 200:223–238, 2019.

[139] DM Stasinopoulos and RA Rigby. Detecting break points in generalised linear models. *Computational Statistics & Data Analysis*, 13(4):461–471, 1992.

[140] Gabriele Steidl, Stephan Didas, and Julia Neumann. Splines in higher order tv regularization. *International journal of computer vision*, 70(3):241–255, 2006.

[141] Alexander Tartakovsky. *Sequential Change Detection and Hypothesis Testing: General Non-iid Stochastic Models and Asymptotically Optimal Rules*. CRC Press, 2019.

[142] Jonathan Taylor and Robert Tibshirani. Post-selection inference for-penalized likelihood models. *Canadian Journal of Statistics*, 46(1):41–61, 2018.

[143] Jonathan Taylor and Robert J Tibshirani. Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, 112(25):7629–7634, 2015.

[144] Jonathan E Taylor, Joshua R Loftus, Ryan J Tibshirani, et al. Inference in adaptive regression via the kac–rice formula. *The Annals of Statistics*, 44(2):743–770, 2016.

[145] Xiaoying Tian and Jonathan Taylor. Asymptotics of selective inference. *Scandinavian Journal of Statistics*, 44(2):480–499, 2017.

[146] Xiaoying Tian, Jonathan Taylor, et al. Selective inference with a randomized response. *The Annals of Statistics*, 46(2):679–710, 2018.

[147] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[148] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

[149] Ryan J Tibshirani. Divided differences, falling factorials, and discrete splines: Another look at trend filtering and related problems. *arXiv preprint arXiv:2003.03886*, 2020.

[150] Ryan J Tibshirani et al. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7:1456–1490, 2013.

[151] Ryan J Tibshirani et al. Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics*, 42(1):285–323, 2014.

[152] Ryan J Tibshirani, Alessandro Rinaldo, Rob Tibshirani, Larry Wasserman, et al. Uniform asymptotic inference and the bootstrap after model selection. *Annals of Statistics*, 46(3):1255–1287, 2018.

[153] Ryan J Tibshirani, Jonathan Taylor, Richard Lockhart, and Robert Tibshirani. Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111(514):600–620, 2016.

[154] Ryan Joseph Tibshirani. *The solution path of the generalized lasso*. Stanford University, 2011.

[155] Charles Truong, Laurent Oudre, and Nicolas Vayatis. A review of change point detection methods. *arXiv preprint arXiv:1801.00718*, 2018.

[156] AW Van Der Vaart and JA Wellner. Weak convergence and empirical processes: With applications to statistics springer series in statistics. *Springer*, 58:59, 1996.

[157] Ennapadam Seshan Venkatraman. Consistency results in multiple change-point problems. 1993.

[158] Lyudmila Yur'evna Vostrikova. Detecting "disorder" in multidimensional random processes. In *Doklady Akademii Nauk*, volume 259, pages 270–274. Russian Academy of Sciences, 1981.

[159] Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE transactions on information theory*, 55(5):2183–2202, 2009.

[160] Daren Wang, Yi Yu, Alessandro Rinaldo, et al. Univariate mean change point detection: Penalization, cusum and optimality. *Electronic Journal of Statistics*, 14(1):1917–1961, 2020.

[161] Daren Wang, Yi Yu, Alessandro Rinaldo, et al. Optimal covariance change point localization in high dimensions. *Bernoulli*, 27(1):554–575, 2021.

[162] Daren Wang, Yi Yu, Alessandro Rinaldo, and Rebecca Willett. Localizing changes in high-dimensional vector autoregressive processes. *arXiv preprint arXiv:1909.06359*, 2019.

[163] Tengyao Wang and Richard J Samworth. High dimensional change point estimation via sparse projection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):57–83, 2018.

[164] Yu-Xiang Wang, Alex Smola, and Ryan Tibshirani. The falling factorial basis and its statistical applications. In *International Conference on Machine Learning*, pages 730–738, 2014.

[165] Larry Wasserman and Kathryn Roeder. High dimensional variable selection. *Annals of statistics*, 37(5A):2178, 2009.

[166] Weichi Wu and Zhou Zhou. Multiscale jump testing and estimation under complex temporal dynamics. *arXiv preprint arXiv:1909.06307*, 2019.

[167] Yuehua Wu. Simultaneous change point analysis and variable selection in a regression problem. *Journal of Multivariate Analysis*, 99(9):2154–2171, 2008.

[168] Yi-Ching Yao. Estimating the number of change-points via schwarz'criterion. *Statistics & Probability Letters*, 6(3):181–189, 1988.

[169] Yi-Ching Yao and Siu-Tong Au. Least-squares estimation of a step function. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 370–381, 1989.

[170] Yi Yu and Sabyasachi Chatterjee. Localising change points in piecewise polynomials of general degrees. *arXiv preprint arXiv:2007.09910*, 2020.

[171] Nancy R Zhang and David O Siegmund. A modified bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, 63(1):22–32, 2007.

[172] Nancy R Zhang, David O Siegmund, Hanlee Ji, and Jun Z Li. Detecting simultaneous changepoints in multiple sequences. *Biometrika*, 97(3):631–645, 2010.

[173] Peng Zhao and Bin Yu. On model selection consistency of lasso. *Journal of Machine learning research*, 7(Nov):2541–2563, 2006.

[174] Sen Zhao, Daniela Witten, and Ali Shojaie. In defense of the indefensible: A very naive approach to high-dimensional inference. *arXiv preprint arXiv:1705.05543*, 2017.

[175] Tijana Zrnic and Michael I Jordan. Post-selection inference via algorithmic stability. *arXiv preprint arXiv:2011.09462*, 2020.

APPENDICES

Appendix A

Appendix of Chapter 3

A.1 Proof of Theorem 3.7

For $\varepsilon_1, \ldots, \varepsilon_n, \ldots \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$, and a sequence $q_1, \ldots, q_n, \ldots$ of real numbers, let

$$\nu_k = \mathrm{Var}\left(\sum_{i=1}^{k} q_i\, \varepsilon_i\right) = \sigma^2 \sum_{i=1}^{k} q_i^2 \qquad \text{for} \quad k \geq 1\,.$$

Define the partial weighted sum process $\left\{S_n(t) \,:\, 0 \leq t \leq 1\right\}$ by

$$S_n(t) = \frac{1}{\sqrt{\nu_n}} \sum_{i=1}^{\lfloor nt \rfloor} q_i\, \varepsilon_i\,, \qquad \text{for} \quad 0 \leq t \leq 1\,.$$

Obviously, for any $k \geq 1$, and any $0 < t_1 < t_2 < \cdots < t_k \leq 1$, the vector $\big(S_n(t_1), \ldots, S_n(t_k)\big)$ has a multivariate normal distribution, and therefore $\left\{S_n(t) \,:\, 0 \leq t \leq 1\right\}$ is a Gaussian process for any given n.

a) In our case, first note that

$$\begin{aligned}\Big[\big(\mathbf{D}_{-\mathcal{A}}\mathbf{D}_{-\mathcal{A}}^T\big)^{-1}\mathbf{D}_{-\mathcal{A}}\Big]_{\lfloor mt \rfloor}\,\mathbf{y} \;&=\; \Big[\big(\mathbf{D}_{-\mathcal{A}}\mathbf{D}_{-\mathcal{A}}^T\big)^{-1}\mathbf{D}_{-\mathcal{A}}\Big]_{\lfloor mt \rfloor}\,\big(\mathbf{y} - \mathbf{f}\,\big) \\ &=\; \Big[\big(\mathbf{D}_{-\mathcal{A}}\mathbf{D}_{-\mathcal{A}}^T\big)^{-1}\mathbf{D}_{-\mathcal{A}}\Big]_{\lfloor mt \rfloor}\,\boldsymbol{\varepsilon}, \qquad \text{(A.1)}\end{aligned}$$

which is a partial weighted sum process of independent and identical Gaussian random variables $\varepsilon_i,\ i = 1, \ldots, n$. The first equality in (A.1) is derived from the fact that the structure of the true signal $\mathbf{f}$ remains unchanged within the j-th block, meaning that $\left[\mathbf{D}_{-\mathcal{A}}\right]_{\lfloor mt \rfloor} \mathbf{f} = 0$, for $(\tau_j + r_a)/m \leq t \leq (\tau_{j+1} - r_b)/m$, which in turn implies

$$\left[\left(\mathbf{D}_{-\mathcal{A}}\mathbf{D}_{-\mathcal{A}}^T\right)^{-1}\mathbf{D}_{-\mathcal{A}}\right]_{\lfloor mt \rfloor} \mathbf{f} = 0.$$

Thus, from the aforementioned argument for $\{S_n(t)\}$, the process $\mathbf{W}_j = \{W_j(t) : (\tau_j + r_a)/m \leq t \leq (\tau_{j+1} - r_b)/m\}$ is a Gaussian process, where

$$W_j(t) = \left(\tau_{j+1} - \tau_j - r\right)^{-(2r+1)/2} \left[\left(\mathbf{D}_{-\mathcal{A}}\mathbf{D}_{-\mathcal{A}}^T\right)^{-1}\mathbf{D}_{-\mathcal{A}}\right]_{\lfloor mt \rfloor} \mathbf{y}.$$

Additionally, with the conditions given in (3.27), $\mathbf{W}_j$ is a Gaussian bridge process over the interval $(\tau_j + r_a)/m \leq t \leq (\tau_{j+1} - r_b)/m$. Furthermore, from (A.1), the mean vector and covariance matrix of $\mathbf{W}_j$ can be computed as $\mathbf{0}$ and $\sigma^2 \left(\mathbf{D}_{-\mathcal{A}}\mathbf{D}_{-\mathcal{A}}^T\right)^{-1}$.

b) Recall that the covariance matrix $\left(\mathbf{D}_{-\mathcal{A}}\mathbf{D}_{-\mathcal{A}}^T\right)^{-1}$ is a block diagonal matrix which states that the covariance matrix between two distinct blocks is zero. This completes the proof of the theorem.

A.2 Proof of Theorem 3.10

a) For $t = 1, \ldots, \tau_1 - r_b$, and both signs ± 1, according to the KKT conditions, the dual variables $\widehat{u}(t)$ must lie between $-\lambda$ and λ, that is,

$$-\lambda \leq \widehat{u}_0^{\text{st}}(t) - \lambda \left[\left(\mathbf{D}_{-\mathcal{A}}\mathbf{D}_{-\mathcal{A}}^T\right)^{-1}\mathbf{D}_{-\mathcal{A}}\right]_t \mathbf{D}_{A_1}^T \mathbf{1} \leq \lambda \tag{A.2}$$

which yields the constraint for the first block in (3.39).

b) Similar to the first block, for $t = \tau_{J_0} + r_a, \ldots, m$, the constraint becomes

$$-\lambda \leq \widehat{u}_{J_0}^{\text{st}}(t) + \lambda \left[\left(\mathbf{D}_{-\mathcal{A}}\mathbf{D}_{-\mathcal{A}}^T\right)^{-1}\mathbf{D}_{-\mathcal{A}}\right]_t \mathbf{D}_{A_{J_0}}^T \mathbf{1} \leq \lambda, \tag{A.3}$$

which leads to the result of (3.40).

c) For $t = \tau_j + r_a, \ldots, \tau_{j+1} - r_b$, and $j = 1, \ldots, J_0 - 1$ the constraint for the exact pattern recovery becomes

$$\lambda\, s_j \leq \widehat{u}_j^{\,\mathrm{st}}(t) - \lambda \left[\left(\mathbf{D}_{-\mathcal{A}} \mathbf{D}_{-\mathcal{A}}^T \right)^{-1} \mathbf{D}_{-\mathcal{A}} \right]_t \left(\mathbf{D}_{A_{j+1}}^T \mathbf{s}_{j+1} + \mathbf{D}_{A_j}^T \mathbf{s}_j \right) \leq \lambda\, s_{j+1}\,. \tag{A.4}$$

Since the stochastic process $\widehat{u}_j^{\,\mathrm{st}}(t)$ is symmetric around zero, when s_{j+1} and s_j have the opposite signs, this constraint reduces to (3.41). Otherwise, when $s_{j+1} = s_j$, which accounts for the staircase in block j, from (3.19) the constraint becomes $\widehat{u}_j^{\,\mathrm{st}}(t) \leq 0$ or $\widehat{u}_j^{\,\mathrm{st}}(t) \geq 0$.

A.3 Proof of Theorem 3.12

(a) The PRUTF algorithm is consistent in pattern recovery if the event

$$\Big\{ \widehat{\boldsymbol{\tau}} = \boldsymbol{\tau} \Big\} \bigcap \Big\{ \operatorname{sign}\big(\mathbf{D}_t \widehat{\mathbf{f}}_n \big) \;=\; \operatorname{sign}\big(\mathbf{D}_t \mathbf{f} \big),\; \forall\, t \in \boldsymbol{\tau} \Big\}, \tag{A.5}$$

occurs with probability approaching one. For ease of exposition, we first compute the probability of the statement in (A.5) for the piecewise constant case, $r = 0$. We then extend this probability computation to an arbitrary piecewise polynomial $r \in \mathbb{N}$.

Case $r = 0$**:** In this case, the event in (A.5) is equivalent to $\{ A_n \cap B_n \}$ where

$$A_n = \Big\{ \min_{t \in \boldsymbol{\tau}} \big| \mathbf{D}_t \widehat{\mathbf{f}}_n \big| \;>\; 0 \Big\}, \tag{A.6}$$

and

$$B_n = \Big\{ \max_{t \in \boldsymbol{\tau}^c} \big| \mathbf{D}_t\, \overline{\boldsymbol{\varepsilon}}_n \big| \;\leq\; 4\, \lambda_n \Big\}. \tag{A.7}$$

For $t \in \boldsymbol{\tau}^c = \{1, \ldots, m\} \backslash \boldsymbol{\tau}$, observe that $\mathbf{D}_t \big(\widehat{\mathbf{f}}_n - \mathbf{f} \big) = 0$; therefore,

$$\begin{aligned} \big| \mathbf{D}_t\, \overline{\boldsymbol{\varepsilon}}_n \big| \;&=\; \big| \mathbf{D}_t \left(\overline{\mathbf{y}}_n - \widehat{\mathbf{f}}_n \right) + \mathbf{D}_t \left(\widehat{\mathbf{f}}_n - \mathbf{f} \right) \big| = \big| \mathbf{D}_t \left(\overline{\mathbf{y}}_n - \widehat{\mathbf{f}}_n \right) \big| \\ &=\; \big| \mathbf{D}_t\, \mathbf{D}^T \widehat{\mathbf{u}} \big| \;\leq\; 4\, \lambda_n, \end{aligned}$$

which is captured in event B_n. The last inequality in the above equation occurs because, from Theorem 3.10, we have $|\widehat{\mathbf{u}}| \leq \lambda_n$ as well as the fact that $\sum_{i=1}^{m} \big| \big[\mathbf{D}_t\, \mathbf{D}^T \big]_i \big| = 2^{r+2}$, for an arbitrary r. In the following, we derive the conditions under which the probabilities of both events A_n and B_n converge to 1.

- To compute the probability of A_n, we first note that, for every $t \in \boldsymbol{\tau}$,

$$\begin{aligned} \left|\mathbf{D}_t \widehat{\mathbf{f}}_n\right| &= \left|\overline{\mathbf{y}}_{n,\,t} - \overline{\mathbf{y}}_{n,\,t-1} - \lambda_n\big(s_t - s_{t-1}\big)\right| \\ &= \left|\mathbf{D}_t\,\overline{\boldsymbol{\varepsilon}}_n + \mathbf{D}_t\,\mathbf{f} - \lambda_n\big(s_t - s_{t-1}\big)\right| \\ &\leq \left|\mathbf{D}_t\,\overline{\boldsymbol{\varepsilon}}_n - \lambda_n\big(s_t - s_{t-1}\big)\right| + \left|\mathbf{D}_t\,\mathbf{f}\right|, \end{aligned}$$

where $\overline{\mathbf{y}}_{n,\,t}$ is the average of observations in the segment created by block t. The last inequality in the above statement is derived from the triangular inequality. Therefore, in order to verify A_n, it is enough to show that, with the probability approaching one,

$$\max_{t\in\boldsymbol{\tau}} \left|\mathbf{D}_t\,\overline{\boldsymbol{\varepsilon}}_n - \lambda_n\big(s_t - s_{t-1}\big)\right| \leq \delta_n, \tag{A.8}$$

where $\delta_n = \min_{t\,\in\,\boldsymbol{\tau}} \left|\mathbf{D}_t\mathbf{f}\right|$ is the minimum jump between change points. Equivalently, it suffices to show that

$$\max_{t\in\boldsymbol{\tau}}\ \lambda_n \left|s_t - s_{t-1}\right| \leq \delta_n/2, \tag{A.9}$$

and

$$\max_{t\in\boldsymbol{\tau}}\ \left|\mathbf{D}_t\,\overline{\boldsymbol{\varepsilon}}_n\right| \leq \delta_n/2. \tag{A.10}$$

The inequality in (A.9) holds if $\lambda_n \leq \dfrac{\delta_n\,\underline{L}_n}{4}$, where $\underline{L}_n = \min_{j=0,\,\ldots,\,J_0} \left|\tau_{j+1} - \tau_j\right|$. Applying the union and Gaussian tail bounds, the probability of the complement of the event in (A.10) can be computed as

$$\begin{aligned} \Pr\Big(\max_{t\in\boldsymbol{\tau}} \left|\mathbf{D}_t\,\overline{\boldsymbol{\varepsilon}}_n\right| \,\geq\, \delta_n/2\Big) \;&\leq\; \sum_{t\in\boldsymbol{\tau}} \Pr\Big(\left|\mathbf{D}_t\,\overline{\boldsymbol{\varepsilon}}_n\right| \geq \delta_n/2\Big) \\ &\leq\; \sum_{t\in\boldsymbol{\tau}} \Pr\left(\left|Z_n\right| \geq \frac{\delta_n\,\sqrt{\underline{L}_n}}{2\,\sqrt{2}\,\sigma_n}\right) \;\leq\; 2\,J_0\,\exp\left(-\frac{\delta_n^2\,\underline{L}_n}{16\,\sigma_n^2}\right) \\ &=\; 2\,\exp\left(-\frac{\delta_n^2\,\underline{L}_n}{16\,\sigma_n^2} + \log(J_0)\right). \end{aligned} \tag{A.11}$$

The probability in (A.11) converges to zero if, for some $\xi > 0$,

$$\frac{\delta_n\,\sqrt{\underline{L}_n}}{\sigma_n} \longrightarrow \infty \qquad \text{and} \qquad \frac{\delta_n\,\sqrt{\underline{L}_n}}{\sigma_n\,\sqrt{\log(J_0)}} \;>\; \sqrt{16}\,(1+\xi). \tag{A.12}$$

- Next, we verify conditions under which $\Pr(B_n) \longrightarrow 1$. Equivalently, it is enough to determine the conditions under which the following probability converges to zero.

$$\begin{aligned}
\Pr\left(B_n^c\right) &= \Pr\left(\max_{t\in\boldsymbol{\tau}^c} \left|\mathbf{D}_t\overline{\boldsymbol{\varepsilon}}_n\right| \geq 4\,\lambda_n\right) \leq \sum_{t\in\boldsymbol{\tau}^c} \Pr\left(\left|\mathbf{D}_t\,\overline{\boldsymbol{\varepsilon}}_n\right| \geq 4\,\lambda_n\right) \\
&\leq \sum_{t\in\boldsymbol{\tau}^c} \Pr\left(\left|Z_n\right| \geq \frac{4\,\lambda_n\,\sqrt{\underline{L}_n}}{\sqrt{2}\,\sigma_n}\right) \\
&\leq 2\left(n - J_0\right)\,\exp\left(-\frac{4\,\lambda_n^2\,\underline{L}_n}{\sigma_n^2}\right) \\
&= 2\,\exp\left(-\frac{4\,\lambda_n^2\,\underline{L}_n}{\sigma_n^2} + \log\left(n - J_0\right)\right). \qquad \text{(A.13)}
\end{aligned}$$

  The above probability converges to zero if, for some $\xi > 0$, the following conditions hold,

$$\frac{\lambda_n\,\sqrt{\underline{L}_n}}{\sigma_n} \longrightarrow \infty \qquad \text{and} \qquad \frac{2\,\lambda_n\,\sqrt{\underline{L}_n}}{\sigma_n\,\sqrt{\log(n - J_0)}} \;>\; (1+\xi). \qquad \text{(A.14)}$$

Case arbitrary r : For the piecewise polynomial of order r ($r \in \mathbb{N}$), we note that, for any $t \in [\tau_j\,,\,\tau_{j+1})$,

$$\begin{aligned}
\left|\mathbf{D}_t\,\widehat{\mathbf{f}}_n\right| &= \left|\mathbf{D}_t\left(\mathbf{y}_n - \mathbf{D}_{-\mathcal{A}}^T\,\widehat{\mathbf{u}}\right)\right| \\
&= \left|\mathbf{D}_t\left[\left(\mathbf{I} - \mathbf{P}_D\right)\mathbf{y}_n \;-\; \lambda_n\,\mathbf{P}_D\left(\mathbf{D}_{A_{j+1}}^T\,\mathbf{s}_{j+1} + \mathbf{D}_{A_j}^T\,\mathbf{s}_j\right)\right]\right| \\
&= \left|\mathbf{D}_t\left[\left(\mathbf{I} - \mathbf{P}_D\right)\left(\mathbf{f} + \boldsymbol{\varepsilon}_n\right) \;-\; \lambda_n\,\mathbf{P}_D\left(\mathbf{D}_{A_{j+1}}^T\,\mathbf{s}_{j+1} + \mathbf{D}_{A_j}^T\,\mathbf{s}_j\right)\right]\right|,
\end{aligned}$$

where $\mathbf{P}_D = \mathbf{D}_{-\mathcal{A}}^T\left(\mathbf{D}_{-\mathcal{A}}\mathbf{D}_{-\mathcal{A}}^T\right)^{-1}\mathbf{D}_{-\mathcal{A}}$ is the projection map onto the row space of $\mathbf{D}_{-\mathcal{A}}$. In the preceding statement, the second equality is derived by plugging in the statement in (3.22) in place of $\widehat{\mathbf{u}}$. From (3.20), recall that $\mathbf{I} - \mathbf{P}_D$ is equivalent to the prediction matrix in the r-th polynomial regression of $\mathbf{y}$ onto indices $\tau_j + 1\,,\,\tau_j + 1\,,\,\ldots\,,\,\tau_{j+1}$. This fact allows us to derive an upper bound for the variance of

$\mathbf{D}_t \left(\mathbf{I} - \mathbf{P}_D\right) \boldsymbol{\varepsilon}_n$ [170],

$$\max_{t \in \boldsymbol{\tau}} \ \mathrm{Var}\Big(\mathbf{D}_t \left(\mathbf{I} - \mathbf{P}_D\right) \boldsymbol{\varepsilon}_n\Big) \ \leq \ 2^{r+1} \frac{n^{2r}\, \sigma_n^2}{\underline{L}_n^{2r+1}}.$$

Following a procedure similar to that used in the case $r = 0$,

$$\Pr\left(\max_{t \in \boldsymbol{\tau}} \left|\mathbf{D}_t \left(\mathbf{I} - \mathbf{P}_D\right) \boldsymbol{\varepsilon}_n\right| \ \geq \ \frac{\delta_n}{2}\right) \ \leq \ 2 \exp\left(-\frac{\delta_n^2 \ \underline{L}_n^{2r+1}}{2^{r+4}\, n^{2r}\, \sigma_n^2} + \log\left(J_0\right)\right). \quad \text{(A.15)}$$

For the case of an arbitrary r, there is a slight modification in the definition of event B_n:

$$B_n = \Big\{ \max_{t \in \boldsymbol{\tau}^c} \ \left|\mathbf{D}_t \left(\mathbf{I} - \mathbf{P}_D\right) \boldsymbol{\varepsilon}_n\right| \ \leq \ 2^{r+2}\, \lambda_n \Big\}. \quad \text{(A.16)}$$

Again, in the same manner

$$\Pr\left(B_n^c\right) \ \leq \ 2 \exp\left(-\frac{2^{r+2}\, \lambda_n^2 \ \underline{L}_n^{2r+1}}{n^{2r}\, \sigma_n^2} + \log\left(n - J_0\right)\right). \quad \text{(A.17)}$$

Therefore, for an arbitrary r, the PRUTF algorithm is consistent in pattern recovery if, in addition to

$$\lambda_n < \frac{\delta_n \ \underline{L}_n^{2r+1}}{n^{2r}\, 2^{r+2}},$$

the conditions in (3.43) and (3.44) hold.

(b) As shown in part (c) of Theorem 3.10, in staircase blocks, the violation of the KKT conditions boils down to crossing the zero line for a Gaussian bridge process. Suppose j-th block is a staircase block; therefore, PRUTF can attain the exact discovery if $\widehat{u}_j^{\,\mathrm{st}}(t) \leq 0$ or $\widehat{u}_j^{\,\mathrm{st}}(t) \geq 0$, for all $(\tau_j + r_a)/m \ \leq \ t \ \leq \ (\tau_{j+1} - r_b)/m$. Hence the probability of this event occurring is equal to $\Pr\Big(\max_{0 \leq t \leq L_j} \widehat{u}_j^{\,\mathrm{st}}(t) \leq 0\Big)$. According to [19],

$$\Pr\Big(\max_{0 \leq t \leq L_j} \widehat{u}_j^{\,\mathrm{st}}(t) \leq a\Big) = 1 - \exp\Big(-\frac{2\, a^2}{S_r^2(L_j)}\Big), \quad \text{(A.18)}$$

where $S_r^2(L_j)$ is the L_j-th diagonal element of the matrix $\sigma^2 \big(\mathbf{D}_{A_j} \mathbf{D}_{A_j}^T\big)^{-1}$. As a result, the probability converges to zero as a vanishes. This result implies that the PRUTF algorithm fails to consistently recover the true pattern in the presence of staircase patterns.

A.4 Residual Analysis For Real Datasets

In this section, we provide the residual analysis for UK HPI and GISTEMP and COVID-19 datasets. We compute the residuals for each dataset by subtracting their observations from the associated signal estimated using the mPRUTF algorithm. The results of auto-correlation function for these datasets are presented in Figure A.1. As can be seen from the figure, there is an auto-correlation among observations in GISTEMP dataset as well as the COVID-19 datasets for Canada, the United Kingdom and the United States.

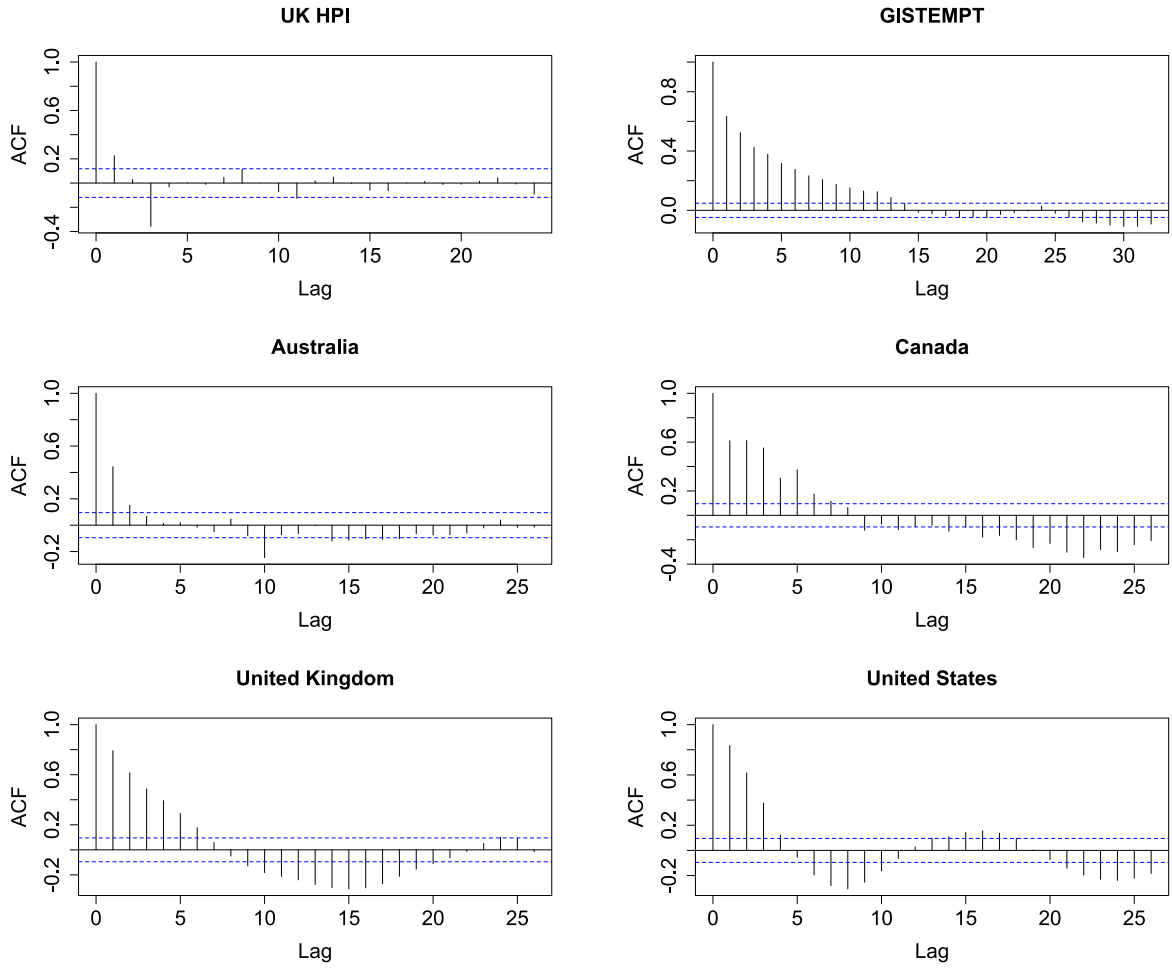


Figure A.1: Auto-correlation Function For the Real Datasets

Appendix B

Appendix of Chapter 4

B.1 Proof of Theorem 4.2

a) Note that

$$\mathbf{A}\mathbf{y} = \mathbf{A}\big(\mathbf{P}_{\eta} + \mathbf{P} - \mathbf{P}_{\eta}\big)\,\mathbf{y} = \left(\frac{\mathbf{A}\boldsymbol{\eta}}{\|\boldsymbol{\eta}\|}\right)\frac{\boldsymbol{\eta}^T\mathbf{y}}{\|\boldsymbol{\eta}\|} + \mathbf{A}\mathbf{V}, \tag{B.1}$$

which allows us to rewrite $\big\{\mathbf{A}\mathbf{y} \geq \mathbf{q}\big\}$ as

$$\Big\{\mathbf{A}\mathbf{y} \geq \mathbf{q}\Big\} = \Big\{\,\sigma\left[\mathbf{A}\boldsymbol{\eta}/\|\boldsymbol{\eta}\|\right]Z + \Big[\mathbf{A}\mathbf{V} - \mathbf{q}\Big] \;\geq\; 0\Big\}.$$

This representation leads to the truncation set

$$\big[\mathcal{V}_Z^-\,,\,\mathcal{V}_Z^+\big] = \bigcap_{i=1}^{|\mathrm{row}(\mathbf{A})|}\left\{z \in \mathbb{R}:\;\sigma\left[\mathbf{A}\boldsymbol{\eta}/\|\boldsymbol{\eta}\|\right]_i z + \Big[\mathbf{A}\mathbf{V} - \mathbf{q}\Big]_i \geq 0\right\}.$$

Observe that solving the inequality in the above set with respect to z depends on the sign of $\rho_i = \left[\mathbf{A}\boldsymbol{\eta}/\|\boldsymbol{\eta}\|\right]_i$. Therefore,

$$\Big\{z \in \mathbb{R}:\;\sigma\boldsymbol{\rho}_i\, z + \big[\mathbf{A}\mathbf{V} - \mathbf{q}\big]_i \geq 0\,,\; i = 1,\ldots,|\mathrm{row}(\mathbf{A})|\Big\} = \begin{cases} z \geq \frac{\left[\mathbf{q}-\mathbf{A}\mathbf{V}\right]_i}{\sigma\boldsymbol{\rho}_i} & i:\;\boldsymbol{\rho}_i > 0, \\ z \leq \frac{\left[\mathbf{q}-\mathbf{A}\mathbf{V}\right]_i}{\sigma\boldsymbol{\rho}_i} & i:\;\boldsymbol{\rho}_i < 0, \\ 0 \leq \big[\mathbf{A}\mathbf{V}-\mathbf{q}\big]_i & i:\;\boldsymbol{\rho}_i = 0. \end{cases}$$

The preceding statement leads to the interval $\left[\mathcal{V}_Z^-\,,\,\mathcal{V}_Z^+\right]$, provided $\mathcal{V}_Z^0 > 0$, where $\mathcal{V}_Z^- = \mathcal{V}_Z^-(\mathbf{V})$, $\mathcal{V}_Z^+ = \mathcal{V}_Z^+(\mathbf{V})$ and $\mathcal{V}_Z^0 = \mathcal{V}_Z^0(\mathbf{V})$ are provided in (4.20). Given $\mathbf{V}$, these truncation boundaries are fixed. Therefore, the distribution of Z given $\left\{\mathbf{V},\, \mathbf{A}\mathbf{y} \geq \mathbf{q}\right\}$ is equivalent to the distribution of a normal distribution constrained to the interval $\left[\mathcal{V}_Z^-\,,\,\mathcal{V}_Z^+\right]$. Hence

$$Z \;\Big|\; \left\{\mathbf{V},\, \mathbf{A}\mathbf{y} \geq \mathbf{q}\right\} \sim \mathrm{TN}\big(\boldsymbol{\eta}^T\mathbf{f}\,,\,1\,,\,\left[\,\mathcal{V}_Z^-\,,\,\mathcal{V}_Z^+\,\right]\big).$$

b) Applying the probability integral transform for (4.19) yields

$$1 - \Phi^{[\mathcal{V}_Z^-,\mathcal{V}_Z^+]}(Z) \;\Big|\; \left\{\mathbf{V},\, \mathbf{A}\mathbf{y} \geq \mathbf{q}\right\} \;\sim\; U\big(\,0\,,\,1\,\big).$$

For every $0 \leq u \leq 1$, by marginalizing the above statement over $\mathbf{V}$, we have

$$\begin{aligned}
&\Pr\left(1-\Phi^{[\mathcal{V}_Z^-,\mathcal{V}_Z^+]}(Z) \leq u \;\Big|\; \mathbf{A}\mathbf{y} \geq \mathbf{q}\right) \\
&\qquad = \int_{\mathbf{v}} \Pr\left(1 - \Phi^{[\mathcal{V}_Z^-,\mathcal{V}_Z^+]}(Z) \leq u \;\Big|\; \mathbf{V}\,,\,\mathbf{A}\mathbf{y} \geq \mathbf{q}\right) \Pr\left(\mathbf{V} \;\Big|\; \mathbf{A}\mathbf{y} \geq \mathbf{q}\right) \mathbf{dv} \\
&\qquad = u \int_{\mathbf{v}} \Pr(\mathbf{V} \;\Big|\; \mathbf{A}\mathbf{y} \geq \mathbf{q})\, \mathbf{dv} \;=\; u
\end{aligned}$$

which establishes Equation (4.21).

B.2 Proof of Theorem 4.4

a) Let $W = \left\|\left(\mathbf{I} - \mathbf{P} + \mathbf{P}_\eta\right)\mathbf{y}\right\|^2$, since both projection matrices $\mathbf{P}$ and $\mathbf{P}_\eta$ are symmetric and idempotent, then

$$\begin{aligned}
W &= \left\|(\mathbf{I} - \mathbf{P})\,\mathbf{y}\right\|^2 + \left(\frac{\boldsymbol{\eta}^T\mathbf{y}}{\|\boldsymbol{\eta}\|}\right)^2 + 2\left(\frac{\boldsymbol{\eta}^T(\mathbf{I} - \mathbf{P})\,\mathbf{y}}{\|\boldsymbol{\eta}\|}\right)\left(\frac{\boldsymbol{\eta}^T\mathbf{y}}{\|\boldsymbol{\eta}\|}\right) \\
&= \widehat{\sigma}^2\left[d + T^2 + 2\left(\frac{\boldsymbol{\eta}^T(\mathbf{I} - \mathbf{P})\,\mathbf{y}}{\widehat{\sigma}\|\boldsymbol{\eta}\|}\right)T\right].
\end{aligned}$$

From (B.1), the polyhedron set can be rewritten as

$$\Big\{\mathbf{A}\mathbf{y} \;\geq\; \mathbf{q}\Big\} = \bigg\{\Big(\frac{\mathbf{A}\boldsymbol{\eta}}{\|\boldsymbol{\eta}\|}\Big)\,\frac{\boldsymbol{\eta}^T\mathbf{y}}{\|\boldsymbol{\eta}\|} + \mathbf{A}\mathbf{V} \;\geq\; \mathbf{q}\bigg\}.$$

Multiplying both sides of the above inequality by $W/\,\widehat{\sigma}^2$ yields the truncation set

$$\begin{aligned}\big[\mathcal{V}_T^-\,,\,\mathcal{V}_T^+\big] &= \bigcap_{i=1}^{|\mathrm{row}(\mathbf{A})|}\bigg\{t\in\mathbb{R}:\;\Big(\frac{W\rho_i}{\widehat{\sigma}}\Big)\,t + \Big[\mathbf{A}\mathbf{V}-\mathbf{q}\Big]_i\Big(d + t^2 + 2\,\frac{\boldsymbol{\eta}^T(\mathbf{I}-\mathbf{P})\,\mathbf{y}}{\widehat{\sigma}\|\boldsymbol{\eta}\|}\,t\Big)\geq 0\bigg\}\\ &= \bigcap_{i=1}^{|\mathrm{row}(\mathbf{A})|}\bigg\{t\in\mathbb{R}:\;\Big[\mathbf{A}\mathbf{V}-\mathbf{q}\Big]_i\,t^2 + \bigg(2\,\frac{\boldsymbol{\eta}^T(\mathbf{I}-\mathbf{P})\,\mathbf{y}}{\widehat{\sigma}\|\boldsymbol{\eta}\|}\,\Big[\mathbf{A}\mathbf{V}-\mathbf{q}\Big]_i + \frac{W\rho_i}{\widehat{\sigma}}\bigg)\,t\\ &\qquad + \Big[\mathbf{A}\mathbf{V}-\mathbf{q}\Big]_i\,d\geq 0\bigg\}. \end{aligned}\tag{B.2}$$

In the above intersection, for each $i,\ i=1\,,\,\ldots\,,\,|\mathrm{row}(\mathbf{A})|$, the inequality can be solved explicitly. Note that $\big[\mathcal{V}_T^-\,,\,\mathcal{V}_T^+\big]$ is a bounded interval (see Theorem 4.8).

On the other hand, W can also be decomposed as

$$W = \big\|\big(\mathbf{I}-\mathbf{P}+\mathbf{P}_{\boldsymbol{\eta}}\big)\,\mathbf{y}\big\|^2 = \big\|\mathbf{y}\big\|^2 \;-\; \big\|\big(\mathbf{P}-\mathbf{P}_{\boldsymbol{\eta}}\big)\,\mathbf{y}\big\|^2.\tag{B.3}$$

Therefore, conditioning on $\big\{\mathbf{V}\,,\,\|\mathbf{y}\|^2\big\}$ is equivalent to conditioning on $\big\{\mathbf{V},\,W\big\}$. This means that, given $\big\{\mathbf{V}\,,\,\|\mathbf{y}\|^2\big\}$, the truncation boundaries $\mathcal{V}_T^- = \mathcal{V}_T^-\big(\mathbf{V},\,W\big)$ and $\mathcal{V}_T^+ = \mathcal{V}_T^+\big(\mathbf{V},\,W\big)$ are fixed. Therefore, the distribution of T given $\big\{\mathbf{V}\,,\,\|\mathbf{y}\|^2\big\}$ is equivalent to the distribution of a t distribution with d degrees of freedom, constrained to $\big[\mathcal{V}_T^-\,,\,\mathcal{V}_T^+\big]$. Hence,

$$T\;\Big|\;\Big\{\mathbf{V},\,\big\|\mathbf{y}\big\|^2\,,\,\mathbf{A}\mathbf{y}\geq\mathbf{q}\Big\}\;\sim\;\mathrm{Tt}\big(\boldsymbol{\eta}^T\mathbf{f},\,1,\,d,\,[\mathcal{V}_T^-,\mathcal{V}_T^+]\big)$$

b) Similar to part (b) of Theorem 4.2, by marginalization of (4.25) over $\big\{\mathbf{V}\,,\,W\big\}$, we obtain the statement in (4.27)

B.3 Proof of Theorem 4.7

(a) The result of this part has been proved in [88]. Here, we give the proof in our notation. Lemma A.4 of the latter reference shows that for $b_k < \infty$,

$$\lim_{z \longrightarrow b_k^-} \big(b_k - z\big)\, L_Z(z) \;=\; -\sigma^2 \log\big(1-\alpha/2\big),$$

$$\lim_{z \longrightarrow b_k^-} \big(b_k - z\big)\, U_Z(z) \;=\; -\sigma^2 \log\big(\alpha/2\big).$$

The above equations together leads to

$$\lim_{z \longrightarrow b_k^-} \big(b_k - z\big)\big[\, U_Z(z) - L_Z(z)\,\big] = \sigma^2 \log\,\big((2-\alpha)/\alpha\big)$$

This limit states that there exits an $\epsilon > 0$, such that

$$U_Z(z) - L_Z(z) \;\geq\; \frac{\sigma^2\,\log\,\big((2-\alpha)/\alpha\big)}{2\,\big(b_k - z\big)} \qquad\qquad \text{for any} \;\; z \in \big(b_k - \epsilon\,,\, b_k\big) \,\bigcap\, \mathcal{S}_Z.$$

Now, let $f^* = \inf\Big\{ f^{\mathcal{S}_Z}_{\mu,\,\sigma^2}(z) :\; z \in \big(b_k - \epsilon\,,\, b_k\big) \,\bigcap\, \mathcal{S}_Z \Big\}$. Clearly, $f^* > 0$ as $f^{\mathcal{S}_Z}_{\mu,\,\sigma^2}$ is a probability density function. Therefore,

$$\begin{aligned}
E\Big[U_Z(Z) - L_Z(Z) \,\Big|\, Z \in \mathcal{S}_Z\Big] \;&=\; \int_{z\,\in\,\mathcal{S}_Z} \big[U_Z(z) - L_Z(z)\big]\, f^{\mathcal{S}_Z}_{\mu,\,\sigma^2}(z)\, dz \\
&\geq\; \frac{\sigma^2 \log\big((2-\alpha)/\alpha\big)}{2} \int_{z\,\in\,(b_k-\epsilon\,,\,b_k)\,\cap\,\mathcal{S}_Z} \frac{f^{\mathcal{S}_Z}_{\mu,\,\sigma^2}(z)}{b_k - z}\, dz \\
&\geq\; \frac{\sigma^2 \log\big((2-\alpha)/\alpha\big)\, f^*}{2} \int_{z\,\in\,(b_k-\epsilon\,,\,b_k)\,\cap\,\mathcal{S}_Z} \frac{1}{b_k - z}\, dz = \infty.
\end{aligned} \tag{B.4}$$

This can be similarly shown for $a_1 > -\infty$.

(b) From the result in part (a), it suffices to show that

$$E\Big[U_T(T) - L_T(T)\Big] \;\geq\; E\Big[U_Z(Z) - L_Z(Z)\Big].$$

To this end, we show that a random variable $T \sim \mathrm{Tt}\left(\mu\,,\,\sigma^2\,,\,d\,,\,\mathcal{S}_T\right)$ has a heavier tail than a random variable $Z \sim \mathrm{TN}\left(\mu\,,\,\sigma^2\,,\,\mathcal{S}_T\right)$. It is known that a random variable Y has heavier tails than X if and only if

$$\lim_{t\longrightarrow\infty}\frac{1-F_Y(t)}{1-F_X(t)}=\infty.$$

This equation is equivalent to

$$\lim_{t\longrightarrow\infty}\frac{f_Y(t)}{f_X(t)}=\infty.$$

Now, let $f^{\mathcal{S}_T}_{\mu,\,\sigma^2}(\cdot)$ and $g^{\mathcal{S}_T}_{\mu,\,\sigma^2,\,d}(\cdot)$, denote the density function of Z and T, respectively. Therefore,

$$\begin{aligned}\lim_{x\longrightarrow\infty}\frac{g^{\mathcal{S}_T}_{\mu,\,\sigma^2,\,d}(x)}{f^{\mathcal{S}_T}_{\mu,\,\sigma^2}(x)}&=\lim_{y\longrightarrow\infty}\frac{g^{\mathcal{S}_T}_{d}(y)}{f^{\mathcal{S}_T}_{0,\,1}(y)}=\lim_{y\longrightarrow\infty}\exp\left\{\log g^{\mathcal{S}_T}_{d}(y)-\log f^{\mathcal{S}_T}_{0,\,1}(y)\right\}\\&=\exp\left\{\lim_{y\longrightarrow\infty}-\frac{d+1}{2}\,\log\left(1+\frac{y^2}{d}\right)+\frac{1}{2}\,y^2+C\right\}\\&=\exp\left\{\lim_{y\longrightarrow\infty}y^2\left[-\frac{d+1}{2\,y^2}\log\left(1+\frac{y^2}{d}\right)+\frac{1}{2}\right]+C\right\},\end{aligned}$$

where in the first equality we use the change variable $y=(x-\mu)/\sigma$. Since

$$\lim_{y\longrightarrow\infty}\frac{1}{y^2}\,\log\left(1+\frac{y^2}{d}\right)=0,$$

we have

$$\lim_{x\longrightarrow\infty}\frac{g^{\mathcal{S}_T}_{\mu,\,\sigma^2,\,d}(x)}{f^{\mathcal{S}_T}_{\mu,\,\sigma^2}(x)}=\infty.$$

The preceding result proves that the truncated t distribution has heavier tails than truncated normal.

Now, since $\left[L_Z(Z)\,,\,U_Z(Z)\right]$ and $\left[L_T(T)\,,\,U_T(T)\right]$ are confidence intervals at the same level for truncated normal and truncated t distributions, respectively, and $U_Z(Z)-L_Z(Z)$ and $U_T(T)-L_T(T)$ are two non-negative random variables such that

$$U_T\big(T\big)\;-\;L_T\big(T\big)\;\overset{a.s.}{\geq}\;U_Z\big(Z\big)\;-\;L_Z\big(Z\big)\,.$$

Hence,

$$E\Big[\,U_{\scriptscriptstyle T}(T)\;-\;L_{\scriptscriptstyle T}(T)\,\Big]\;\geq\;E\Big[\,U_{\scriptscriptstyle Z}(Z)\;-\;L_{\scriptscriptstyle Z}(Z)\,\Big]\,.$$

The above equation and Equation (B.4) together lead to the statement in (4.30).

B.4 Proof of Theorem 4.8

(a) From Equation (4.20), it can be viewed that $\mathcal{V}_{\scriptscriptstyle Z}^{-}$ and $\mathcal{V}_{\scriptscriptstyle Z}^{+}$ are unbounded if both sets $\big\{i\,:\,\rho_i\,<\,0\big\}$ and $\big\{i\,:\,\rho_i\,>\,0\big\}$ are empty, where $\rho_i\,=\,\big[\mathbf{A}\boldsymbol{\eta}/\|\boldsymbol{\eta}\|\big]_i$, for $i\,=\,1\,,\,\ldots\,,\,\big|\mathrm{row}(\mathbf{A})\big|$. This condition implies $\rho_i\,=\,0$ for all i which in turn requires that $\mathbf{A}\boldsymbol{\eta}\,=\,\mathbf{0}$ for all $\boldsymbol{\eta}\,\neq\,\mathbf{0}$. Hence every vector $\boldsymbol{\eta}\,\neq\,\mathbf{0}$ is orthogonal to the row space of the matrix $\mathbf{A}$, which contradicts with the fact that the conditioning set is a polyhedron.

(b) From (4.26), $\mathcal{V}_{\scriptscriptstyle T}^{-}$ and $\mathcal{V}_{\scriptscriptstyle T}^{+}$ are unbounded if the inequality

$$\Big[\mathbf{AV}-\mathbf{q}\Big]_i\,t^2+\left(2\,\frac{\boldsymbol{\eta}^T(\mathbf{I}-\mathbf{P})\,\mathbf{y}}{\widehat{\sigma}\|\boldsymbol{\eta}\|}\,\Big[\mathbf{AV}-\mathbf{q}\Big]_i+\frac{W\rho_i}{\widehat{\sigma}}\right)t+\,\Big[\mathbf{AV}-\mathbf{q}\Big]_i\,d\geq 0 \qquad \text{(B.5)}$$

always holds for any $i=1,\ldots,\big|\mathrm{row}(\mathbf{A})\big|$. This requires that the quadratic equation in (B.5) to have at most one real root, i.e.,

$$\left(2\,\frac{\boldsymbol{\eta}^T(\mathbf{I}-\mathbf{P})\,\mathbf{y}}{\widehat{\sigma}\|\boldsymbol{\eta}\|}\,\Big[\mathbf{AV}-\mathbf{q}\Big]_i\,+\,\frac{W\rho_i}{\widehat{\sigma}}\right)^2\,-\,4\,d\,\Big[\mathbf{AV}-\mathbf{q}\Big]_i^2\leq 0\,,$$

as well as the coefficient sign of t^2 must be positive, i.e., $\big[\mathbf{AV}-\mathbf{q}\big]_i>0$. According to (4.20), this latter condition occurs if and only if $\rho_i=0$, for all $i=1,\ldots,\big|\mathrm{row}(\mathbf{A})\big|$, which is impossible as shown in part (a).

B.5 Proof of Theorem 4.10

(a) Since $s_{\hat{\tau}_j}$ is the sign of $\big[\mathbf{D}\widehat{\mathbf{f}}\big]_{\hat{\tau}_j}\,=\,\mathbf{D}_{\hat{\tau}_j}\widehat{\mathbf{f}}$, hence, $s_{\hat{\tau}_j}\,\mathbf{D}_{\hat{\tau}_j}\widehat{\mathbf{f}}\,\geq\,0$. From the primal-dual relationship in (3.8), we get

$$s_{\hat{\tau}_j}\,\mathbf{D}_{\hat{\tau}_j}\big(\mathbf{y}-\mathbf{D}^T\widehat{\mathbf{u}}_{\lambda_j}\big)\geq 0, \qquad \text{(B.6)}$$

where λ_j is the value of the regularization parameter associated with $\widehat{\tau}_j$. Now, the goal is to find a range for Z_j^{glo} for which $\widehat{\tau}_j$ is a change point. Recall that from KKT conditions of the dual problem (3.6), if $\widehat{\tau}_j$ is a change point then $\widehat{u}_{\lambda_j,\widehat{\tau}_j} = s_{\widehat{\tau}_j}\,\lambda_j$. Applying the decomposition $\mathbf{y} = \mathbf{P}_{\boldsymbol{\eta}}\,\mathbf{y} + \big(\mathbf{I} - \mathbf{P}_{\boldsymbol{\eta}}\big)\,\mathbf{y}$ into (B.6) and setting $\widehat{u}_{\lambda_j,\widehat{\tau}_j} = s_{\widehat{\tau}_j}\,\lambda_j$, we have

$$s_{\widehat{\tau}_j}\,\mathbf{D}_{\widehat{\tau}_j}\left(\frac{\boldsymbol{\eta}}{\|\boldsymbol{\eta}\|^2}\,\boldsymbol{\eta}^T\,\mathbf{y} + \big(\mathbf{I} - \mathbf{P}_{\boldsymbol{\eta}}\big)\,\mathbf{y} - s_{\widehat{\tau}_j}\,\lambda_j\,\mathbf{D}_{\widehat{\tau}_j}^T - \mathbf{D}_{-\widehat{\tau}_j}^T\,\widehat{\mathbf{u}}_{\lambda_j,-\widehat{\tau}_j}\right).$$

Hence, setting $\boldsymbol{\eta} = \mathbf{D}_{\widehat{\tau}_j}^T$ in the preceding equation, for $s_{\widehat{\tau}_j} = 1$,

$$\mathbf{D}_{\widehat{\tau}_j}\mathbf{y} \;\geq\; \lambda_j\,\mathbf{D}_{\widehat{\tau}_j}\,\mathbf{D}_{\widehat{\tau}_j}^T + \mathbf{D}_{\widehat{\tau}_j}\,\mathbf{D}_{-\widehat{\tau}_j}^T\,\widehat{\mathbf{u}}_{\lambda_j,-\widehat{\tau}_j}, \tag{B.7}$$

and, for $s_{\widehat{\tau}_j} = -1$,

$$\mathbf{D}_{\widehat{\tau}_j}\mathbf{y} \;\leq\; -\lambda_j\,\mathbf{D}_{\widehat{\tau}_j}\,\mathbf{D}_{\widehat{\tau}_j}^T + \mathbf{D}_{\widehat{\tau}_j}\,\mathbf{D}_{-\widehat{\tau}_j}^T\,\widehat{\mathbf{u}}_{\lambda_j,-\widehat{\tau}_j}. \tag{B.8}$$

Therefore, given $\mathbf{D}_{\widehat{\tau}_j}$, λ_j, $\widehat{\mathbf{u}}_{\lambda_j}$, and σ^2, the quantity $Z_j^{\text{glo}} = \frac{\mathbf{D}_{\widehat{\tau}_j}\,\mathbf{y}}{\sigma\,\|\mathbf{D}_{\widehat{\tau}_j}^T\|}$ is restricted to regions $\big(-\infty\,,\,\mathcal{V}_{Z_j}^{-(glo)}\big]$ and $\big[\mathcal{V}_{Z_j}^{+(glo)}\,,\,\infty\big)$ where $\mathcal{V}_{Z_j}^{-(glo)}$ and $\mathcal{V}_{Z_j}^{+(glo)}$ are given in (4.33).

(b) In the same fashion as part (a), given $\mathbf{D}_{\widehat{\tau}_j}$, λ_j and $\widehat{\mathbf{u}}_{\lambda_j}$, the quantity $T_j^{\text{glo}} = \frac{\mathbf{D}_{\widehat{\tau}_j}\mathbf{y}}{\widehat{\sigma}_j^{(\text{glo})}\,\|\mathbf{D}_{\widehat{\tau}_j}^T\|}$ is restricted to the regions $\big(-\infty\,,\,\mathcal{V}_{T_j}^{-(glo)}\big]$ and $\big[\mathcal{V}_{T_j}^{+(glo)}\,,\,\infty\big)$ where $\mathcal{V}_{T_j}^{-(glo)}$ and $\mathcal{V}_{T_j}^{+(glo)}$ are given in (4.34).

B.6 Proof of Theorem 4.11

We will prove the inequality for the truncated t distribution. A similar proof can be applied to the truncated normal distribution which is also provided in [88]. Suppose $T \sim t\,(\mu\,,\,\sigma^2\,,\,d)$ is a location-scale t distribution with cumulative distribution function $G_{\mu,\,\sigma^2,\,d}\,(\cdot)$. For any $0 < \gamma < 1$, define $Q_\gamma(t)$ via

$$G_{Q_\gamma,\,\sigma^2,\,d}\big(t\big) = \Pr{}_{Q_\gamma,\,\sigma^2,\,d}\big(T \leq t\big) = \gamma,$$

or equivalently, $Q_\gamma(t) = t - \sigma\, G_d^{-1}(\gamma)$. Thus, the cumulative distribution function of T truncated to $\mathcal{S} = \bigcup_{i=1}^{m} (c_i\,,\, d_i)$, is upper bounded by

$$\begin{aligned} G^{\mathcal{S}}_{Q_\gamma,\,\sigma^2,\,d}(t) &= \frac{\Pr_{Q_\gamma,\,\sigma^2,\,d}\big(T \le t \;\bigcap\; T \in \mathcal{S}\big)}{\Pr_{Q_\gamma,\,\sigma^2,\,d}\big(T \in \mathcal{S}\big)} \\ &\le \frac{\Pr_{Q_\gamma,\,\sigma^2,\,d}\,(T \le t)}{\Pr_{Q_\gamma,\,\sigma^2,\,d}\,(T \in \mathcal{S})} \le \frac{\gamma}{\underline{g}}, \end{aligned} \tag{B.9}$$

where $\underline{g} = \inf_{\mu}\ \Pr_{\mu,\,\sigma^2,\,d}\big(T \in \mathcal{S}\big)$. On the other hand, from the inequality $\Pr(A \bigcap B) \ge \Pr(A) + \Pr(B) - 1$, we notice that

$$\begin{aligned} G^{\mathcal{S}}_{Q_\gamma,\,\sigma^2,\,d}(t) &= \frac{\Pr_{Q_\gamma,\,\sigma^2,\,d}\big(T \le t \;\bigcap\; T \in \mathcal{S}\big)}{\Pr_{Q_\gamma,\,\sigma^2,\,d}\big(T \in \mathcal{S}\big)} \\ &\ge \frac{\Pr_{Q_\gamma,\,\sigma^2,\,d}\big(T \le t\big) \;+\; \Pr_{Q_\gamma,\,\sigma^2,\,d}\big(T \in \mathcal{S}\big) \;-1}{\Pr_{Q_\gamma,\,\sigma^2,\,d}\big(T \in \mathcal{S}\big)} \\ &\ge \inf_{\mu}\ \frac{\Pr_{Q_\gamma,\,\sigma^2,\,d}\big(T \le t\big) \;+\; \Pr_{\mu,\,\sigma^2,\,d}\big(T \in \mathcal{S}\big) \;-1}{\Pr_{\mu,\,\sigma^2,\,d}\big(T \in \mathcal{S}\big)} \\ &= \frac{\gamma + \overline{g} - 1}{\overline{g}}, \end{aligned} \tag{B.10}$$

where $\overline{g} = \sup_{\mu}\ \Pr_{\mu,\,\sigma^2,\,d}\big(T \in \mathcal{S}\big)$. Equations (4.28) and (B.9) imply that

$$G^{\mathcal{S}}_{Q_{\frac{\alpha}{2}\,\underline{g}},\,\sigma^2,\,d}(t) \;\le\; \frac{\alpha}{2} \;=\; G^{\mathcal{S}}_{U(T),\,\sigma^2,\,d}(t).$$

Because $G^{\mathcal{S}}_{\mu,\,\sigma^2,\,d}(t)$ is a decreasing function with respect to μ, hence,

$$U(T) \;\overset{a.s.}{\le}\; Q_{\frac{\alpha}{2}\,\underline{g}}(T) = T - \sigma\, G_d^{-1}\Big(\frac{\alpha}{2}\,\underline{g}\Big).$$

Similarly, from (4.28) and (B.10), we obtain

$$L(T) \;\overset{a.s.}{\ge}\; Q_{1-\frac{\alpha}{2}\,\overline{g}}(T) = T - \sigma\, G_d^{-1}\Big(1 - \frac{\alpha}{2}\,\overline{g}\Big).$$

Therefore, from the last two equations, we have

$$\begin{aligned} U(T) - L(T) &\overset{a.s.}{\leq} \sigma\left[G_d^{-1}\Big(1 - \frac{\alpha}{2}\,\overline{g}\Big) - G_d^{-1}\Big(\frac{\alpha}{2}\,\underline{g}\Big)\right] \\ &= \sigma\left[G_d^{-1}\Big(1 - \frac{\alpha}{2}\,\overline{g}\Big) + G_d^{-1}\Big(1 - \frac{\alpha}{2}\,\underline{g}\Big)\right] \\ &\leq 2\,\sigma\, G_d^{-1}\Big(1 - \frac{\alpha}{2}\,\underline{g}\Big)\,. \end{aligned} \tag{B.11}$$

Note that the last equality in the preceding statement is obtained from $1 - \frac{\alpha}{2}\,\overline{g} \leq 1 - \frac{\alpha}{2}\,\underline{g}$. Now, consider the minimization problem

$$\inf_{\mu}\ \Pr{}_{\mu,\,\sigma^2,\,d}\Big(T < d_1 \ \text{ or } \ T > c_m\Big),$$

which minimizes at $\frac{d_1+c_m}{2}$, with the minimum value $g_0 = 2\,G_d\Big(-\frac{c_m-d_1}{2\,\sigma}\Big)$. Observe that $g_0 \leq \underline{g}$, and since $G^{-1}(\cdot)$ is an increasing function, then

$$G_d^{-1}\Big(1 - \frac{\alpha}{2}\,\underline{g}\Big) \ \leq \ G_d^{-1}\Big(1 - \frac{\alpha}{2}\,g_0\Big).$$

Lastly, to obtain the final upper bound in (4.37), it suffices to show

$$G_d^{-1}\Big(1 - \frac{\alpha}{2}\,g_0\Big) \ \leq \ G_d^{-1}\Big(1 - \frac{\alpha}{2}\Big) \ + \ \frac{c_m - d_1}{2\,\sigma}\,, \tag{B.12}$$

or equivalently,

$$G_d^{-1}\Big(\frac{\alpha}{2}\,g_0\Big) \ \geq \ G_d^{-1}\Big(\frac{\alpha}{2}\Big) \ - \ \frac{c_m - d_1}{2\,\sigma}\,, \tag{B.13}$$

Given the condition (4.36), we note that,

$$\alpha\ G_d\Big(-\frac{c_m - d_1}{2\,\sigma}\Big) \ \geq\ G_d\Big(G_d^{-1}(\alpha/2) - \frac{c_m - d_1}{2\,\sigma}\Big), \qquad \text{for} \quad d \geq 3. \tag{B.14}$$

Hence, by plugging in the value of g_0, we have

$$\frac{\alpha}{2}\,g_0 = \alpha\ G_d\left(-\frac{c_m - d_1}{2\,\sigma}\right) \ \geq\ G_d\left(G_d^{-1}\left(\frac{\alpha}{2}\right) - \frac{c_m - d_1}{2\,\sigma}\right).$$

Applying the increasing function $G_d^{-1}(\cdot)$ to both sides of the above inequality leads to (B.13), and in turn (B.12). Now, Equations (B.11) and (B.12) together yield

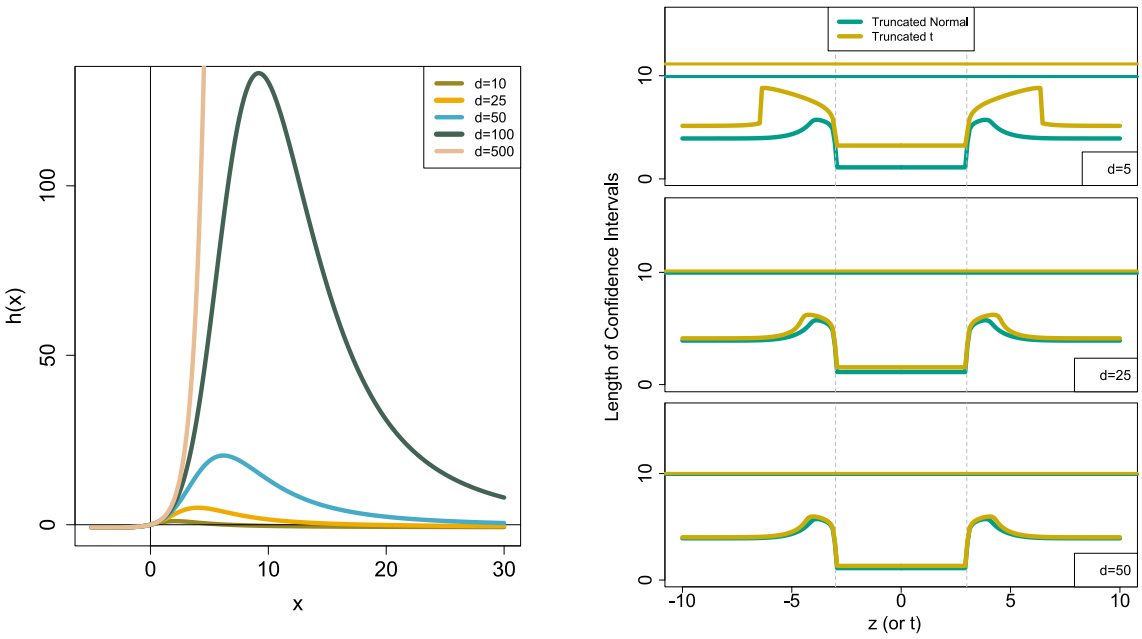
$$U(T) - L(T) \overset{a.s.}{\leq} 2\,\sigma\, G_d^{-1}\Big(1 - \frac{\alpha}{2}\Big) + \big(c_m - d_1\big).$$

Finally, the upper bound associated with the truncated t distribution in (4.37) will be achieved by replacing $L(T)$, $U(T)$, σ, d_1, c_m and d with $L_{T_j}^{\text{glo}}$, $U_{T_j}^{\text{glo}}$, $\widehat{\sigma}_j^{(\text{glo})}$, $d^{(\text{glo})}$, $\mathcal{V}_{T_j}^{-(\text{glo})}$, $\mathcal{V}_{T_j}^{+(\text{glo})}$ and $d^{(\text{glo})}$, respectively.

To explore the condition (B.14), define

$$h(x) = \frac{\alpha\, G_d(-x)}{G_d\left(G_d^{-1}(\alpha/2) - x\right)} - 1\,, \qquad \text{for} \quad d \geq 3\,.$$

The function $h(x)$ has two roots, one at zero and one at a positive value x_0. Moreover, $h(x) > 0$, for $x \in \big(0\,,\, x_0\big)$. Notice that the positive root x_0 diverges as degrees of freedom d increases. More specifically, for large d, the truncated t distribution behaves similar to normal one, and hence $h(x)$ becomes an increasing function for all $x > 0$. See Figure B.1, panel (a). Therefore, for large d, the statement in (B.14) always holds. For small to moderate values of d, (B.14) holds under the condition $\frac{c_m - d_1}{2\,\sigma} \in \big(0\,,\, x_0\big]$. We have visualized the behaviour of lengths of confidence intervals and their upper bounds in Figure B.1, panel (b), for various values of d.



(a) Function $h(x)$. **(b)** Truncation set $\mathcal{S} = (-\infty\,,\,-3) \cup (3\,,\,\infty)$.

Figure B.1: Panel (a) displays the function $h(x)$ which, for moderate values of d, is a unimodal function with two roots, zero and a positive value x_0. Also, $h(x) \geq 0$ for $x \in [0\,,\,x_0]$. Panel (b) shows the lengths of confidence intervals for a truncated normal distribution and a truncated t distribution, under various values of d. The solid horizontal lines in panel (b) shows the upper bounds given in Theorem 4.11, for both distributions.