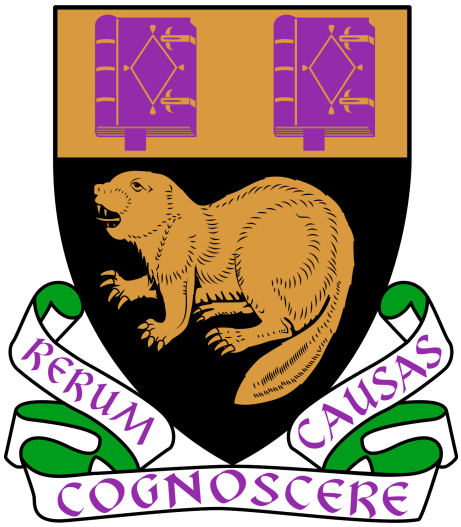
Tools for Model Selection for Mean-Nonstationary Time Series



Shuhan Yang

Department of Statistics

London School of Economics and Political Science

A thesis submitted for the degree of

Doctor of Philosophy

December 2023

Declaration

I certify that the thesis I have presented for examination for the MPhil/PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it). The copyright of this thesis rests with the author. Quotation from it is permitted, provided that full acknowledgement is made. This thesis may not be reproduced without my prior written consent. I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party. I declare that my thesis consists of around 28100 words.

Shuhan Yang

December 2023

Acknowledgements

First and foremost, I would like to express gratitude to my supervisor, Professor Piotr Fryzlewicz, for allowing me to be his student, and for his constant encouragement, remarkable patience, professional feedback and insightful guidance during the past four years. Without his help, this thesis would never have been completed. His vast knowledge, creativity and consistent support have not only helped me shape the direction of my research but also largely enriched my understanding of different topics. Meanwhile, his mentorship also provides me with much freedom that encourages me to develop independent and critical thinking. It is indeed a great honour for me to work under his supervision throughout my years of study. Additionally, I would like to extend my appreciation to my second supervisor, Dr. Yining Chen, for his feedback and suggestions on my thesis.

I would also like to convey my deep gratitude to the London School of Economics and Political Science and the Department of Statistics for the financial support and resources provided. This LSE scholarship has played a crucial role in enabling me to conduct this research, and I am sincerely grateful for this incredible opportunity. I thank the staff in the Department of Statistics and especially Penny Montague and Imelda Noble for the invaluable assistance throughout my PhD journey.

Last but not least, I extend my deepest appreciation to my parents, Li Juan and Jiandong Yang, for their ever-lasting love, trust, support and encouragement. In addition, I am indeed grateful to my friends and colleagues, Xiaolin Zhu, Zezhun Chen, Qin Fang, Yiliu Wang, Kaifang Zhou, Jinghan Tee, Xinyi Liu, Shakeel Gavioli-Akilagun, Mingwei Lin and everyone on the fifth floor in the Columbia House, for the help and enjoyable memories during my study. In particular, I am deeply grateful to Xiaolin Zhu for her precious guidance and support in both my research and personal life. Thank you all!

Abstract

This thesis studies the problem of detecting multiple change-points of the process with a piecewise-constant signal plus dependent noise, and analysing lead-lag relationships between nonstationary time series.

Motivated by the demand of long-run variance (LRV) in extending the application of existing change-point detection (CPD) approaches proposed for independent time series, the first part of the thesis introduces novel wavelet-based consistent LRV estimators to quantify the level of noise in mean nonstationary processes. In our proposed estimators, a particular blend of wavelets and well-suited thresholds make our methods lie somewhere in between the two broad classes of LRV estimators: residual- and difference-based estimators. Specifically, they bypass the difficulty in the pre-estimation of signals and can be robust to potential outliers that largely impact the performance many difference-based estimators. Several asymptotic properties of our estimators are proved, and their performance are illustrated through comparative simulation studies.

Secondly, we study the aspects of model selection for nonstationary time series with level change. In particular, we explore the possible extensions of the Narrowest-Over-Threshold (NOT) detection algorithm, hoping that it can show better performance for serially correlated data. Our attempts mainly consist of three parts and we provide more detailed discussion of the last two, including data-preprocessing and the modification of the strengthened Schwarz Information Criterion (sSIC) applied in NOT solution

path algorithm. Many simulations are conducted to demonstrate the practicability of our ideas.

Lastly, motivated by the dynamics of COVID-19 datasets, our interest shifts towards investigating lead-lag relationships between nonstationary time series. Relying the "scale-space" viewpoint employed in the SIgnificant ZERo crossings of derivatives (SiZer) map, we introduce an exploratory approach, Multi-scale Lead Lag Heatmap (MLLH), for providing an broad view of (possible) significant relations between two time series, which may serve as the first step for further lead-lag or causal analyses. Starting from simple examples, we develop and describe several heatmaps that display significant features of simulated bi-variate data over both locations and scales. Finally we assess the performance of MLLH on real-world COVID-19 data examples.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The problem of detecting and estimating abrupt change-points in the structural features of time series has been of interest to statisticians over many fields such as finance, economics and medicine for a long time. And this issue has been extensively investigated under the particular assumption of independent noise, see e.g. Yao (1988), Yao and Au (1989), Lee (1995), Vostrikova (1981) and Venkatraman (1992) for some early references for detection on mean shifts. Since the independence assumption is quite restrictive from a practical viewpoint, the focus, more recently, has been distinguishing change-points from the natural fluctuations in serial correlated processes, see e.g. Lavielle and Moulines (2000), Davis et al. (2006), Chakar et al. (2017) and Cho and Kirch (2022), or Aue and Horváth (2013) for an overview. In Chapter 2, we provide more reviews on the topics of change-point detection and estimation. The basic concepts of time series analysis are given in Section 2.1, and Section 2.2 introduces more methodologies especially developed for sequences with piecewise-constant mean.

For serially dependent data, one particular line of research is to make a suitable extension on the applicability of the existing approaches proposed for independent data.

And hence the estimation of long-run variance (LRV) can be particularly useful for overcoming this issue by quantifying the level of noise in applications with correlated time series, see e.g. Tecuapetla-Gómez and Munk (2017), Khismatullina and Vogt (2020) and Dette et al. (2019). This task becomes more challenging when the signal is discontinuous due to the difficulty arising from the pre-estimation of the signals, which is the core of residual-based LRV estimators. On the other hand, existing difference-based estimators can circumvent this notoriously difficult problem but can often be quite sensitive to the choice of the smoothing parameter, see formula (4.3) in Khismatullina and Vogt (2020) as an example. To tackle this issue, Chapter 3 proposes several asymptotically unbiased and consistent LRV estimators based on wavelets and the related idea of wavelet shrinkage. To eliminate the signals and better estimate the error process, the two stage procedure first conducts a wavelet transformation to remove the most of the piecewise-constant signals in series, and then thresholding is utilised to remove the remaining "outliers" containing more than noise. The theoretical results demonstrate the asymptotic properties of our robust estimators, and their good practical performances are illustrated in an extensive comparative simulation study as well.

Chapter 4 is motivated by the consideration whether LRV can be utilised to enhance the performance of Narrowest-Over-Threshold (NOT) change-point detection algorithm for dependent data. We provide an overall discussion on how NOT can be extended to time series produced by piecewise-constant signal and serial correlated error process. Starting from the threshold-based NOT algorithm, we assess the practicability of building the threshold proportional to the new LRV estimator. Due to the possible failure of finding the optimal threshold, Baranowski et al. (2019) produced the NOT solution path algorithm to allow for the automatic selection of threshold and the corresponding best candidate model via minimising the strengthened Schwarz Information Criterion. Therefore, we secondly investigate the potential development of this new algorithm, and begin with studying the two practical data preprocessing methods introduced to

reduce the dependence of data before employing NOT, see Section 4 in Baranowski et al. (2019). Furthermore, we explore the possible usefulness of extension via changing the measure of fit or penalty in sSIC. Simulations show the performance of our attempts over a variety of series with different number and locations of change-points in mean shifts plus dependent error processes.

In Chapter 5, we conduct an exploratory analysis on lead-lag relationships between bi-variate nonstationary time series. In the spirit of "SIgnificant ZERo crossings of derivatives" (SiZer) (Chaudhuri and Marron, 1999), our approach is developed on the scale-space viewpoint, with the aim of eliminating the requirement of choosing an optimal bandwidth by simultaneously studying a broad range of scales (bandwidths). Meanwhile, we construct several multi-scale lead-lag heatmaps to help graphically display the significance of (possible) relations. To make the model simpler, we avoid the problem of finding possible time lags by directly drawing inference from data at original time locations, and hence it can provide more accurate information for dataset knowing the direction of lead-lag relations. In the first part of this chapter, we present many simulated examples and provide detailed descriptions of the obtained heatmaps. Additional real-world examples of COVID-19 curves (Mathieu et al., 2020) are analysed in the remainder of the chapter.

Finally, Chapter 6 provides a brief conclusion of the contributions and discusses possible directions for future research.

Chapter 2

Literature Review

This chapter offers a comprehensive review of the literature in the two domains of statistics explored within this thesis: time series analysis and lead-lag relationships. And we shall attach particular importance to asymptotic estimation of long-run variance and change-point detection in time series.

2.1 Time series analysis

2.1.1 Introduction

The term *time series* generally refers to stochastic processes that consist of observations collected sequentially over time. A stochastic process is a family of random variables $\{X_t,\ t \in T\}$, where $T \neq \emptyset$ is an index set. Usually, a realisation of $\{X_t,\ t \in T\}$ is also considered as a time series in practice. For discrete time series analysis, the index set T is usually chosen to be the integers $\mathbb{Z}$, the non-negative integers $\mathbb{N}$, the positive integers $\mathbb{Z}^+$ or the commonly-used set $T = \{1, 2, \ldots, n\}$. Unlike regression analysis for

independent data, time series analysis aims to solve the statistical problems resulted from the time correlations of adjacent observations.

To capture the time correlations, we commonly analyse and utilise the moments of a time series, especially the *first and the second order moments*. The first moment, i.e. the mean function, is given by $\mu_t = \mathbb{E}(X_t)$, and together with the second moment it gives us the *autocovariance function (ACVF)*, defined as

$$\gamma_t(\tau) = \mathrm{Cov}(X_t, X_{t+\tau}) = \mathbb{E}[(X_t - \mu_t)(X_{t+\tau} - \mu_{t+\tau})]$$

and the *autocorrelation function (ACF)* is defined as

$$\rho_t(\tau) = \frac{\gamma_t(\tau)}{\sqrt{\mathrm{Var}(X_t)\mathrm{Var}(X_{t+\tau})}}.$$

These two functions can display the degree of linear dependence between two points on the same time series. The degree of persistence in one series can be reflected in the long-lasting large value of $\gamma_t(\tau)$, which holds even for large τ. Meanwhile, this series will generally fluctuate if the large $\gamma_t(\tau)$ is negative. On the other hand, the condition $\gamma_t(\tau) = 0$ tells us that the two time points X_t and $X_{t+\tau}$ can only be non-linearly correlated.

Imposing assumptions on the dependence or distribution of a time series is necessary for utilising some apparent features to describe the data. The idea of *stationarity* forms the basis of many statistical procedures employed in existing literature. A time series $\{X_t,\ t \in T\}$ can be defined as *strictly stationary* if the joint distribution of $(X_{t_1}, X_{t_2}, \ldots, X_{t_h})$ is the same as that of $(X_{t_1+\tau}, \ldots, X_{t_h+\tau})$ for all $t_1, \ldots, t_h \in T$ and τ such that $t_1 + \tau, \ldots, t_h + \tau \in T$. In particular, when taking $h = 1, 2$, we can know that mean and autocovariance functions are time-invariant and the latter one depends solely on the time lag τ.

Indeed, verifying properties for a given time series can be challenging, especially when the underlying distribution of the series is complex, i.e. it is often difficult to estimate numerous model parameters from the available data. Therefore, strict stationarity is often too restrictive in practice and so *weak stationarity*, or *second-order stationarity*, is introduced instead. A time series is defined as weakly stationary when the mean $\mathbb{E}(X_t) = \mu < \infty$, $\mathrm{Var}(X_t) < \infty$ and the autocovariance function $\gamma_t(\tau)$ depends on the time location only through the difference τ, i.e. $\gamma_t(\tau) = \gamma(\tau)$.

For time series analysis and forecasting, *white noise process*, a sequence of uncorrelated random variables with constant mean, is one of the most fundamental building components of complex time series data. Mathematically speaking, a white noise process $\{X_i\}$ satisfies $\mathbb{E}(X_i) = \mu < \infty$ and $\gamma(\tau) = 0$ for $\tau \neq 0$, i.e. $X_i \sim WN(\mu, \sigma^2)$. In the next section, we shall talk about the definition of Autoregressive Moving Average (ARMA) processes, which can be represented as a linear combination of white noise time series variables.

2.1.2 Autoregressive Moving Average (ARMA) processes

In statistical analysis, the ARMA process, a widely used model, offers an explicit representation of weakly stationary stochastic processes through two polynomials for the AR and the MA parts respectively. This process can be applied to forecast the observation at time $t+1$ based on the recorded historical data $(X_t, X_{t-1}, \dots)$. The notation ARMA(p, q) refers to an ARMA model with p autoregressive and q moving-average terms, and for the error process X_t, it is defined as

$$X_t - \sum_{j=1}^{p} \phi_j X_{t-j} = \epsilon_t + \sum_{l=1}^{q} \theta_l \epsilon_{t-l} \tag{2.1.1}$$

where ϕ_j and θ_l are all constants with $\phi_p \neq 0$ and $\theta_q \neq 0$, and the ϵ_t are independent, identically distributed (*iid*) zero-mean random variables, which are often assumed to be normally distributed. A stationary ARMA(p, q) model can have non-zero mean μ if

$$X_t = \alpha + \sum_{j=1}^{p} \phi_j X_{t-j} + \epsilon_t + \sum_{l=1}^{q} \theta_l \epsilon_{t-l} \tag{2.1.2}$$

with $\alpha = \mu/(1 - \sum_{j=1}^{p} \phi_j)$. Let $\Theta(B)$ and $\Phi(B)$ denote the *moving average operator* and *autoregressive operator* respectively, which are defined as

$$\begin{aligned} \Theta(B) &= 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q \\ \Phi(B) &= 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p \end{aligned} \tag{2.1.3}$$

where B is a *backward shift operator* giving $B^s X_t = X_{t-s}$. And hence the ARMA(p, q) model can also be represented as

$$\Phi(B) X_t = \Theta(B) \epsilon_t. \tag{2.1.4}$$

An ARMA(p, q) process is weakly stationary if the corresponding AR characteristic polynomial

$$\Phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p \tag{2.1.5}$$

has no complex roots inside the unit circle or on its boundary. Any process with an MA(∞) representation with absolute convergence of the sum of its coefficients can be considered as weak stationarity as well.

2.2 Nonstationary Time Series

Nonstationarity in time series can arise in different aspects, including shifts in the mean, fluctuations in the variance, changes in both aspects, shifts in the autocorrelation

function, or even transformations in the entire joint distribution. A time series can also possess more than one nonstationary behaviours.

When dealing with nonstationary data, there exist scenarios where it is possible to interpret the nonstationarity as smooth transitions. In such cases, the underlying trend or pattern in the data changes gradually over time rather than experiencing abrupt changes. For example, the house prices in the UK can be modelled as having a smoothly varying trend, see Figure 2.1. On the other hand, it is also natural to model time series in the way of containing abrupt change-points in the underlying patterns. This presents a substantial challenge and leads to extensive investigations even in recent years within various fields such as finance (Aminikhanghahi and Cook, 2017; Habibi, 2021; Kim et al., 2022), environment (Alyousifi et al., 2022; Getahun et al., 2021; He et al., 2022; Shi et al., 2022), medical sciences (Malladi et al., 2013; Liu et al., 2018; Chen et al., 2019; Ondrus et al., 2021), and engineering (Daly et al., 2020; Kamalabad et al., 2023).

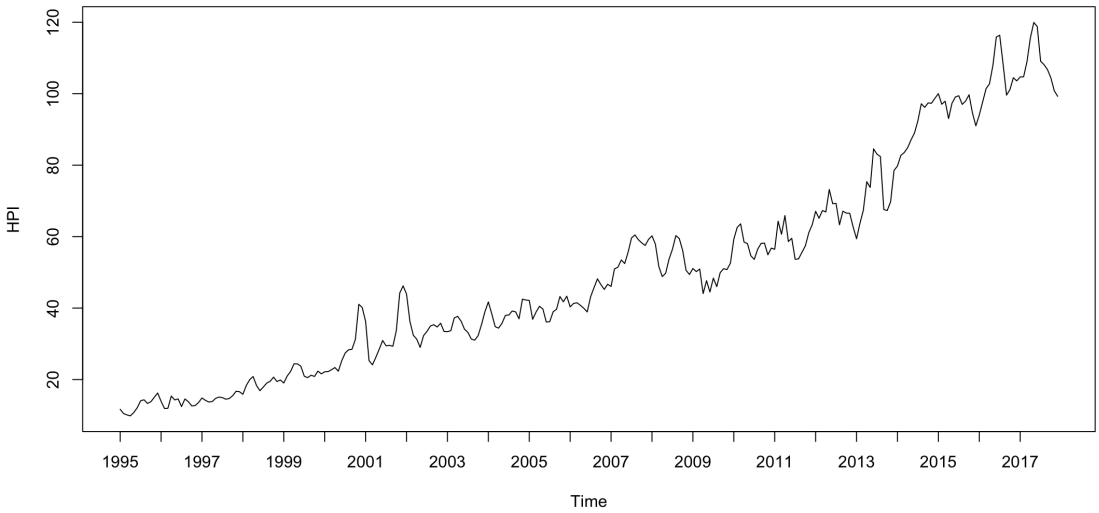


Figure 2.1: Monthly house price index (HPI) recorded from 1995-01-01 to 2017-12-01 in London, where the changes of the observed time series seem to occur gradually rather than abruptly.

Among all categories, *unit-root non-stationarity* occurs when a time series $\{Y_t\}_{t=1}^n$ contains a stochastic trend, i.e. the simple random walk process $Y_t = Y_{t-1} + \epsilon_t$, which does not converge to a constant mean, and the statistical properties such as mean and variance change over time. And hence the presence of a unit root indicates a lack of stationarity. Therefore, traditional methods designed for detecting abrupt change-points in the unknown signal of an otherwise stationary time series are not well-behaved for unit-root nonstationary time series. Instead of unit-root nonstationarity, this thesis intends to review the main existing approaches for modelling other kinds of non-stationarities in time series, especially mean shifts. T hen we shall focus on the particular class of nonstationary processes that will be of interest to us in the following chapters.

One common technique to address nonstationary data is to assume that the observed data can be "well approximated" by piecewise-stationarity over shorter intervals of time, which can be regarded as "local stationarity". For example, considering time dependence, Dahlhaus (1997) introduced the novel concept of *time-rescaling* and extended the Cramér representation to the class of *locally stationary Fourier (LSF) processes*, where the well-known *Cramér representation* (see e.g. Brockwell and Davis (2009), Chapter 4) allows for the decomposition of any zero-mean, uni-variate, discrete time series that possesses weak stationarity.

Inspired by Dahlhaus (1997), Nason et al. (2000) proposed the *locally stationary wavelet (LSW)* model after incorporating the concept of rescaled time. However, it deviates from the LSF approach by substituting the Fourier basis representation with a non-decimated wavelet basis representation (see its definition in Section 2.3.1). To complete the LSW model, Fryzlewicz (2003) modified the definition of model in Nason et al. (2000).

When dealing with situations where there exists rapid changes in second-order structure, the LSW processes are effective due to the compact support of the wavelets. Sev-

eral authors show their applications of the LSW framework: for example, Fryzlewicz (2003) proposed an algorithm for forecasting nonstationary time series and utilised the LSW model for analysing the financial log-return data; Knight et al. (2012) considered an LSW process with missing observations; Chapman et al. (2020) introduced a non-parametric technique to detect changes in variance for data with outliers and heavy tails. However, the LSW process has a limitation that it cannot handle time-varying first-order behavior, as it is applicable only to zero-mean time series. To overcome this, McGonigle et al. (2022) introduced a polynomial trend LSW process and Dette and Wu (2022) proposed a new estimator for the high-dimensional covariance matrix of a locally stationary process characterized by a smoothly varying trend.

In the following subsections, since we are particularly interested in multiple change-point detection for models with piecewise-constant signals, we assume that the data $\{Y_1, Y_2, \ldots, Y_n\}$ is identified in the following structure

$$Y_i = f_i + X_i, \ i = 1, 2, \ldots, n \tag{2.2.1}$$

where f_i is the unknown deterministic signal and $\{X_i\}_{i=1}^n$ represents the error process. Let $\eta_1, \ldots, \eta_q$ denote the change-points where there exist changes in features of interest and q represents the unknown number of change-points.

2.2.1 Change-Points in Piecewise Constant Signal

In this section, we consider the problem of detecting multiple change-points when f_t in (2.2.1) becomes piecewise-constant signal of nonstationary time series such that

$$f_t = \sum_{j=1}^{q+1} \theta_j \mathbb{1}_{(\eta_{j-1}, \eta_j]}(t), \quad t = 1, 2, \ldots, n$$

where the locations of q change-points satisfy $0 = \eta_0 < \eta_1 < \dots \eta_q < \eta_{q+1} = n$, q is either known or unknown, and $\theta_1, \dots, \theta_{q+1}$ are the function values of f_t.

One standard technique for change-point estimation is relying on minimising a model cost function of the form

$$L(Y_t, \eta_1, \dots, \eta_{\hat{q}}) + \text{penalty}(\hat{q}, \eta_1, \dots, \eta_{\hat{q}}) \tag{2.2.2}$$

where the likelihood-type function $L(\cdot)$ measures the quality of fit of estimated model (to find the change-point locations) and the "penalty" term penalty$(\cdot)$ discourages over-fitting (for setting the number of change points).

For continuously distributed *iid* noise $\{X_i\}_{i=1}^n$ such that $\mathbb{E}(X_i) = 0$ and $\mathbb{E}(X_i^6) < \infty$, Yao and Au (1989) applied the following least-squares method to estimate the $\eta_1, \dots, \eta_q$, $\theta_1, \dots, \theta_k$ and described their behaviour under the assumption of known q,

$$\underset{\eta_1,\dots,\eta_q}{\arg\min} \left\{ \sum_{k=1}^{q+1} \sum_{t=\eta_{k-1}+1}^{\eta_k} \left(Y_t - \bar{Y}_{(\eta_{k-1}+1):\eta_k}\right)^2 \right\} \tag{2.2.3}$$

where $\bar{Y}_{(\eta_{k-1}+1):\eta_k}$ represents the average of the observations $Y_{\eta_{k-1}+1}, \dots, Y_{\eta_k}$. In terms of the unknown number of change-points, Yao (1988) regarded q as the dimension of the model and proposed an estimator via minimising the *Schwarz criterion (SIC)*, i.e. *Bayesian information criterion (BIC)*, for independent Gaussian sequence with $X_i \overset{iid}{\sim} N(0, \sigma^2)$. Under the assumption of unknown σ^2, the criterion is defined as follows

$$\text{SIC}(q) = \underset{\eta_1,\dots,\eta_q}{\arg\min} \left[\frac{n}{2} \log \left\{ \frac{1}{n} \sum_{k=1}^{q+1} \sum_{t=\eta_{k-1}+1}^{\eta_k} \left(Y_t - \bar{Y}_{(\eta_{k-1}+1):\eta_k}\right)^2 \right\} + q \log(n) \right] \tag{2.2.4}$$

And the estimator of q is

$$\hat{q}_{SIC} = \underset{0 \leqslant q \leqslant q_U}{\arg\min}\ \text{SIC}(q) \tag{2.2.5}$$

where q_U is the pre-specified fixed upper bound of q. Moreover, for the selection of penalty, Lee (1995), Lavielle and Moulines (2000) and Boysen et al. (2009) decided to choose the ones built on the number of change-points whereas Pan and Chen (2006) and Zhang and Siegmund (2007) considered penalties relying on both the number and locations of change-points. Especially, Boysen et al. (2009) took account of the minimisers of the *Potts functional*, which is given below.

$$\arg\min_{\mathbf{f}} \left[\frac{1}{n} \sum_{t=1}^{n} (Y_t - f_t)^2 + \gamma |J(\mathbf{f})| \right] \tag{2.2.6}$$

where $J(\mathbf{f}) = \{t : 1 \leqslant t \leqslant n-1, f_t \neq f_{t+1}\}$ represents the set of change-points of the candidate signals $\mathbf{f} \in \mathbb{R}^n$ and $|\cdot|$ indicates the number of elements in the target set. However, the recommended parameter $\gamma = 2.5\hat{\sigma}^2 \log(n)/n$ makes this method indeed close to SIC. To clarify, if the noise variance σ^2 is assumed to be a known parameter, the basic setting of SIC can be represented as

$$\frac{1}{n} \sum_{k=1}^{q+1} \sum_{t=\eta_{k-1}+1}^{\eta_k} \left(Y_t - \bar{Y}_{(\eta_{k-1}+1):\eta_k} \right)^2 + 2\sigma^2 q \log(n)/n \tag{2.2.7}$$

We can see it deviates from (2.2.6) by the constant factor 2 instead of 2.5. When σ^2 is unknown, SIC would be more convenient since there is no need to separately provide a consistent estimator of σ^2. In order to overcome the issue due to the irregularities in the likelihood function, Zhang and Siegmund (2007) derived a *modified BIC (mBIC)* by providing a more detailed formula for the penalty function:

$$-\frac{1}{2} \left[3q \log(n) + \sum_{i=1}^{q+1} \log \left(\frac{\eta_i - \eta_{i-1}}{n} \right) \right]. \tag{2.2.8}$$

Compared to traditional BIC, the mBIC statistic also penalises the relative locations of change-points and hence may intuitively be more effective because of anomalies in the likelihood function. However, the mBIC actually inadvertently promotes change-

points locating closely to each other since it imposes a less severe penalty on this kind of change-points. Hence this estimator is more suitable for models without closely located change-points.

By balancing the asymptotic null distribution of the multi-scale test statistic (for controlling the probability of overestimating the true number of change-points) and the exponential bounds (for the probability of underestimation), Frick et al. (2014) provided an approach for the exponential family regression model instead of the basic Gaussian ones. In contrast, to allow for applications in broader classes of models, Du et al. (2016) attached importance to the flexibility of Bayesian approaches with prior distributions and proposed a related change-point estimation method in a marginal likelihood framework, whose penalty is also maximised, similar to mBIC, when change-point estimates are located as close as possible.

Although such least-square approaches using dynamic programming seem to be optimal in the case of a normal error process when minimising likelihood-type function, their efficacy were severely impacted by the slow computational speed. When turning to computing the theoretical minimum, the penalty-based optimisations often have an $O(n^2)$ cost for both space and time. To overcome this issue, Jackson et al. (2005) introduced the *Optimal Partitioning* algorithm that can find the exact global optimum with a linear computational cost in best scenarios. This dynamic programming algorithm, in terms of both the number q and locations $\eta_1, \ldots, \eta_q$ of change-points, recursively minimises the penalised cost, which often takes the form of

$$Q(N; \eta_1, \ldots, \eta_q) = \sum_{i=1}^{q+1} C(Y_{1:n}, \eta_{i-1}, \eta_i) + q\lambda \tag{2.2.9}$$

where the segment cost functions for the residual sum of squares are introduced as $C(Y_{1:n}, k, l) = \sum_{i=k+1}^{l} (Y_i - \bar{Y}_{(k+1):l})^2$, $l > k$. In order to further enhance the Optimal Partitioning algorithm, Killick et al. (2012) incorporated the pruning rule and proposed

an algorithm named the *Pruned Exact Linear Time (PELT)*, which can significantly enhance computational efficiency particularly for data containing short segments relative to the overall data length. Rigaill (2015) proposed the *Pruned Dynamic Programming Algorithm (pDPA)* as a solution to the computational problem. Guédon (2013) introduced the *Forward-backward Dynamic Programming* algorithm and *Smoothing-type Forward-backward Programming* algorithm for different types of change-points.

Binary Segmentation

Binary Segmentation (BS) proposed by Vostrikova (1981) can be considered the most straightforward *hierarchical*, *top-down* approach for detecting multiple change-points. Specifically, top-down change-point detection techniques start by examining the entire dataset to identify the most prominent change-point candidate and then progressively narrow the focus to two sub-intervals split by the detected candidate to obtain less prominent change-point candidates. This hierarchical approach helps efficiently identify multiple change-points within the dataset. At each stage of the BS procedure, change-point testing is carried out using a CUSUM (Cumulative Sum) statistic $\mathcal{C}_{s,b,e}(\mathbf{Y})$.

Given $1 \leqslant s \leqslant b < e \leqslant n$, the CUSUM statistic (Vostrikova, 1981) for any sequence $\mathbf{Y} = (Y_1, \ldots, Y_n)^\intercal$ on the interval $[s, e]$ is defined by

$$\begin{aligned} \mathcal{C}_{s,b,e}(\mathbf{Y}) &= \sqrt{\frac{e-b}{(e-s+1)(b-s+1)}} \sum_{t=s}^{b} Y_t - \sqrt{\frac{b-s+1}{(e-s+1)(e-b)}} \sum_{t=b+1}^{e} Y_t \\ &= \sqrt{\frac{(e-b)(b-s+1)}{e-s+1}} [\bar{Y}(s,b) - \bar{Y}(b+1,e)] \end{aligned} \tag{2.2.10}$$

The BS algorithm is defined recursively with the resulting change-point in $[s, e]$:

$$b^* = \arg\max_{s \leqslant b < e} |\mathcal{C}_{s,b,e}(\mathbf{Y})|, \quad \text{if } \max_{s \leqslant b < e} |\mathcal{C}_{s,b,e}(\mathbf{Y})| > \lambda, \tag{2.2.11}$$

where λ is a pre-specified threshold. To be specific, the BS algorithm calculates $\mathcal{C}_{1,b,n}(\mathbf{Y})$ in the beginning and considers $b_{1,1} = \arg\max_{1\leqslant b\leqslant n} |\mathcal{C}_{1,b,n}(\mathbf{Y})|$ as the first change-point candidate after judging its significance against the criterion. $\mathcal{C}_{1,b,b_{1,1}}(\mathbf{Y})$ and $\mathcal{C}_{b_{1,1}+1,b,n}(\mathbf{Y})$ are then computed in the next step. The whole algorithm terminates when no additional change-points are detected. This condition is met when the computed CUSUM values become lower than the specified threshold λ. Vostrikova (1981) proved the consistency of Binary Segmentation for a fixed number of breaks and Venkatraman (1992) provided the proof the consistency under weaker conditions of change-points (both number and location).

Compared to optimisation approaches, the BS algorithm is a "greedy" procedure. It operates sequentially, where each stage depends on the preceding ones and are never visited again. Also, each stage is straightforward and only involves one-dimensional optimization. These provides BS with the advantages of low computational complexity and conceptual simplicity.

While Binary Segmentation is widely used for multiple change-point detection, it is still limited in its ability to accurately handle segments $[s, e]$ with more than one true change-point since it only utilises a single change-point in the least-squares sense to fit a best piecewise-constant function. To eliminate the weakness of the BS algorithm, Olshen et al. (2004) proposed the *Circular Binary Segmentation*; Fryzlewicz (2014) proposed the *Wild Binary Segmentation (WBS)* where local CUSUM statistics, instead of a global one, are calculated over a group of pre-specified subsamples; *Narrowest-Over-Threshold (NOT)* detection method was developed in Baranowski et al. (2019); and Fryzlewicz (2020) investigated the reasons for the poor performance of Wild Binary Segmentation in some settings and proposed the *Wild Binary Segmentation 2 (WBS2)* and the *steepest-drop model selection* method.

To be specific, given $1 \leqslant s < e \leqslant n$, the WBS algorithm improved BS by calculating

the CUSUM statistic $\mathcal{C}_{s,b,e}(\mathbf{Y})$ of suitably many subsamples $(Y_s, Y_{s+1}, \ldots, Y_e)$ in the first stage, where s and e are integers selected uniformly, independently and with replacement instead of computing a global statistic $\mathcal{C}_{1,b,n}(\mathbf{Y})$ over the entire dataset $(Y_1, Y_2, \ldots, Y_n)$ (which is the idea behind the term 'Wild'). Then the first change-point candidate will be the largest maximiser among the entire collection of largest CUSUMs in all subsamples if it is larger than a pre-specified threshold λ. Similar to the Binary Segementation, the WBS algorithm conducts the same procedure recursively to the left and right of the chosen change-point candidate.

However, although the WBS algorithm outperform the BS method, it is still not optimal, especially when multiple change-points are present in close proximity. In detail, the interval selected by WBS at each stage is not necessary to contain a single change-point exclusively. Consequently, the CUSUM statistic obtained from that interval may not show us the most accurate estimator for the change-point location. Baranowski et al. (2019) proposed the NOT localisation approach to deal with this issue by favouring the shortest intervals with the significant contrast statistics (e.g. CUSUMs), see detailed literature review in Section 2.5. Mathematically speaking, the chosen change-point in the narrowest interval is defined by

$$b^* = \underset{s \leqslant \arg\max_{s \leqslant b < e} |\mathcal{C}_{s,b,e}(\mathbf{Y})| < e}{\arg\min} \{|e - s| : \max_b |\mathcal{C}_{s,b,e}(\mathbf{Y})| > \lambda\}, \tag{2.2.12}$$

where λ is a pre-specified threshold. Furthermore, the concentration on the narrowest intervals within the data enables NOT to go beyond change-point detection for piecewise-constant signals, which is the primary objective of the WBS method. The WBS and BS algorithms, lacking this narrowest-interval focus, may not be as suitable for broader feature detection scenarios as the NOT algorithm. We shall cover more underlying models relying on NOT in the following chapters.

On the other hand, Fryzlewicz (2020) proposed the WBS2 solution path algorithm to

directly tackle the data-nonadaptivity problem of the WBS algorithm when deciding interval choice, especially for signal with multiple change-points. Instead of dividing the data into many segments at the beginning, WBS2 applies a data-adaptive interval drawing scheme, where the subsamples are recursively drawn relying on the detected change-point candidates.

Bottom-up Techniques

Recognising that the top-down nature can be the primary cause for the commonly observed weak performance of Binary Segmentation, a "bottom-up" essence is employed as an alternative direction in several existing papers. This idea starts from the finest level of resolution of the data and successively merges the adjacent regions that are highly likely to represent the same locally constant underlying signal, i.e. rather than the "divisive" character of BS, it is an "agglomerative" algorithm. For hierarchically estimating the change-point locations, Matteson and James (2014) proposed an agglomeration algorithm that proceeds by optimising a goodness-of-fit statistic. Messer et al. (2014) built upon the concepts from the filtered derivative method and introduced a strategy involving the progressive merging of change-point candidates in a bottom-up manner, starting from those initially identified candidates using the smallest bandwidth. By introducing and entailing the idea "tail-greediness", Fryzlewicz (2018) constructed the discrete *Unbalanced Haar (UH)* basis (Fryzlewicz, 2007) in an inherently different way and then proposed the *Tail-Greedy Unbalanced Haar (TGUH)* transform that yields a multi-scale data-adaptive decomposition of the one-dimensional data. This approach is particularly attractive due to its ability to provide good practical performance and offer fast computation speed regardless of the number of change-points or the complex features of the signals. The TGUH decomposition algorithm determines both the number q and the locations $\eta_1, \ldots, \eta_q$ of change-points in the piecewise-constant signal f_t through four sequential stages (see Fryzlewicz (2018)).

The tail-greediness of the TGUH algorithm carries significant and extensive implications for the computational complexity by giving it an $O(n\log^2(n))$ upper bound. Also, since multiple merges over non-overlapping regions take place at each scale j of the transform, TGUH can guarantee the L_2 consistency of the estimated signal $\mathbf{f}$, and after some post-processing, can even derive another estimator consistent in detecting the number and locations of the change-points in $\mathbf{f}$.

Moving Sum

When addressing multiple mean shift problems, it can typically be more straightforward to sequentially consider subsamples that are expected to contain at most one change-point on a moving-window basis, and then correspondingly detect and estimate the single change-point. Eichinger and Kirch (2018) explored the characteristics of change-point estimators constructed based on *Moving Sum (MOSUM)* statistics. The paper considers the CUSUM-like statistic of the form

$$\begin{aligned} T_n(G) &= \max_{G\leqslant k\leqslant n-G} \frac{|T_{k,n}(G)|}{\sigma_*} \\ T_{k,n}(G) &= T_{k,n}(G;Y_1,\ldots,Y_n) = \frac{1}{\sqrt{2G}}\left(\sum_{i=k+1}^{k+G} Y_i - \sum_{i=k-G+1}^{k} Y_i\right) \end{aligned} \tag{2.2.13}$$

with the bandwidth G satisfying

$$\frac{G}{n} \to 0 \quad \text{and} \quad \frac{n^{2/(2+\nu)}\log(n)}{G} \to 0 \tag{2.2.14}$$

where ν and σ_* are as in the following assumption on the error distribution: For error process X_i, there exists a standard Wiener process $\{W(k):\ 1\leqslant k\leqslant n\}$ and $\nu > 0$ such that

$$\sum_{i=1}^{n} X_i - \sigma_* W(n) = O(n^{1/(2+\nu)}) \quad a.s.$$

with a strictly positive *long-run variance* (see more details in section 2.4)

$$\sigma_*^2 = \sigma^2 + 2\sum_{h=1}^{\infty}\gamma(h) > 0, \quad \gamma(h) = \mathrm{Cov}(X_0, X_h), \quad \sigma^2 = \mathrm{Var}(X_0)$$

where the standard *Wiener process* is a continuous-time stochastic process that starts at zero ($W(0) = 0$), and its independent increments follows $W(t) - W(s) \sim N(0, t-s)$ for any $t > s \geqslant 0$.

Since the testing interval $[k - G + 1, k + G]$ moves along the time series, MOSUM only performs one single test for each fixed interval, which compares the data over $[k - G + 1, k]$ with that over $[k + 1, k + G]$.

2.2.2 Change-Point Detection Methods for Dependent Data

Under the conditions where permitting serial correlated error processes, considerable developments have also been made in detecting multiple change-points in the mean of one-dimensional data. When dealing with such a detection problem, the main challenge arises from the fact that it is hard to distinguish change-points in signal from natural fluctuations in a dependent error process. In this section, we shall conduct a brief literature review on multiple change-point detection for dependent data with piecewise-constant signal.

In general, there are two lines of research: one is extending the test statistics proposed under the assumption of independence to a broader setting while the other one is to conduct a simultaneous estimation on the serial dependence in noise and change-point structures in signal by applying specific time series models such as the autoregressive (AR) model.

Regarding to the first line of research, the performance of existing approaches largely rely on finding a reliable estimate of quantified dependence structure in noise, which can be described with nuisance parameters such as the long-run variance (LRV). We have a more detailed review of LRV in Section 2.4. In particular, under general assumptions on the error process, Dette et al. (2020) extended the *simultaneous multi-scale change-point estimator (SMUCE)* introduced in Frick et al. (2014) by scaling the basic statistic with a difference-based LRV estimator (see Hall et al. (1990) and Tecuapetla-Gómez and Munk (2017)). Eichinger and Kirch (2018) extended the applicability of the MOSUM statistic (Hušková and Slabỳ, 2001) for data with possible error processes, which can have better performance for small samples.

In the presence of (possibly) multiple change points, it can be challenging to find the LRV estimators, which may be largely impacted by the selection of tuning parameters, such as the bandwidth parameter mentioned in Shao and Zhang (2010), that can be closely associated with the frequency of change-points. To avoid the direct estimation of LRV, Shao and Zhang (2010) conducted an extension on the *self-normalization (SN)* method to build a SN Kolmogorov–Smirnov test for detecting change-points in the mean of short-range dependent time series; For long-range dependent time series, Betken (2016) proposed a robust SN test relying on the Wilcoxon-statistic; Pešta and Wendler (2020) combined SN and *wild bootstrap* (Wu, 1986) without computing either nuisance or tuning parameters. To be specific, based on the well-known Kolmogorov-Smirnov test statistic

$$KS_n = \sup_{k=1,\dots,n} |T_n(k)/\hat{\sigma}_*| \tag{2.2.15}$$

where $T_n(k) = n^{-1/2}\sum_{i=1}^{k}(Y_i - \bar{Y})$, Shao and Zhang (2010) introduced a new self-normaliser $V_n(k)$, to replace the estimated LRV $\hat{\sigma}_*^2$. Given $S_{t_1,t_2} = \sum_{i=t_1}^{t_2} Y_i$ for $1 \leqslant t_1 \leqslant t_2 \leqslant n$, the normalisation process for $k = 1, \dots, n-1$ is defined as follows

$$V_n(k) = \frac{1}{n^2}\left[\sum_{i=1}^{k}\left(S_{1,i} - \frac{i}{k}S_{1,k}\right)^2 + \sum_{i=k+1}^{n}\left(S_{i,n} - \frac{n-i+1}{n-k}S_{k+1,n}\right)^2\right].$$

The final test statistic is given as

$$G_n = \sup_{k=1,\dots,n} T_n(k)^2 / V_n(k) \tag{2.2.16}$$

and the estimated change-point can then be represented as

$$\hat{k} = \arg\max_{k=1,\dots,n} T_n(k)^2 / V_n(k) \tag{2.2.17}$$

For further extension, Shao and Zhang (2010) also discussed the scenarios of multiple change-points or a more general framework with other quantities of interest besides the mean shift. Self-normalisation can successfully bypass the estimation of the LRV and hence avoid the selection of the bandwidth parameter, but their theoretical validity is often limited to testing the availability of the single change-point. To overcome this issue, Zhao et al. (2022) developed a novel framework for change-point estimation by combining the SN test with a nested local window-based algorithm. In particular, instead of the global SN test, this paper computes a maximal SN test over nested window sets $H_{1:n}(k)$ covering each $k = h, \dots, n-h$, where h denotes the window size, with the test statistic for each subsample $\{Y_i\}_{i=t_1}^{t_2}$ defined as

$$G_n(t_1, k, t_2) = T_n(t_1, k, t_2)^2 / V_n(t_1, k, t_2) \tag{2.2.18}$$

where

$$T_n(t_1, k, t_2) = \frac{(k - t_1 + 1)(t_2 - k)}{(t_2 - t_1 + 1)^{3/2}} (\hat{\theta}_{t_1,k} - \hat{\theta}_{k+1,t_2}) \tag{2.2.19}$$

$$\begin{aligned} V_n(t_1, k, t_2) = &\sum_{i=t_1}^{k} \frac{(i - t_1 + 1)^2 (k - i)^2}{(t_2 - t_1 + 1)^2 (k - t_1 + 1)^2} (\hat{\theta}_{t_1,i} - \hat{\theta}_{i+1,k})^2 \\ &+ \sum_{i=k+1}^{t_2} \frac{(t_2 - i + 1)^2 (i - 1 - k)^2}{(t_2 - t_1 + 1)^2 (t_2 - k)^2} (\hat{\theta}_{i,t_2} - \hat{\theta}_{k+1,i-1})^2 \end{aligned} \tag{2.2.20}$$

and where $\hat{\theta}_{a,b}$ represents the nonparametric estimator of the quantity of interest. Then,

Zhao et al. (2022) proposed an *SN-based multiple change-point estimation (SNCP)* algorithm replying on the maximal test statistic $G_{1,n}(k) = \max_{(t_1,t_2)\in H_{1:n}(k)} G_n(t_1, k, t_2)$ and a user-specified threshold λ, leading to the first estimated change-point

$$\hat{k} = \underset{k=1,\ldots,n}{\arg\max}\, G_{1,n}(k), \quad \text{if } \max_{k=1,\ldots,n} G_{1,n}(k) > \lambda \tag{2.2.21}$$

This algorithm can be recursively utilised on subsamples $\{Y_i\}_{i=1}^{\hat{k}}$ and $\{Y_i\}_{i=\hat{k}+1}^{n}$ until no test statistic exceeds the threshold λ. Compared to the change-point testing methods above, this change-point estimation framework further provides the estimation on the number and locations of change-points. Besides, based on Schwarz criterion (Schwarz, 1978), Cho and Kirch (2022) introduced a localised pruning algorithm permitting an exhaustive search on change-point candidates obtained from multi-scale methods and its estimation consistency was proved under general assumptions allowing for heavy tails and dependence.

The second direction has been extensively researched in the literature. For the simplest AR(1)-type dependence model, Fang and Siegmund (2020) utilised the maximum score statistic for the change-point estimation in the level, slope, or other features of data; Chakar et al. (2017) developed the dynamic programming algorithm, *AR(1) Segmentation (AR1Seg)*, for dependent data although it requires a post-processing step to avoid estimating more than one change around each change-point location; to bypass this step, Romano et al. (2022) proposed the *Detecting Changes in Autocorrelated and Fluctuating Signals (DeCAFS)* algorithm, another principled dynamic programming algorithm, to minimise the penalised cost function for estimating the number and location of change-points. In this paper, the authors detected abrupt changes in data in the presence of signal $f_i = f_{i-1} + \eta_i + \delta_i$ with $\eta_i \overset{iid}{\sim} N(0, \sigma_\eta^2)$ and $\delta_i = 0$ except at time i immediately after the occurrence of change-points. The stationary error process follows an AR(1) model $X_i = \phi_1 X_{i-1} + \epsilon_i$, $i = 2, \ldots, n$, with $\epsilon_i \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$ and $\epsilon_1 \sim N(0, \sigma_\epsilon^2/(1-\phi_1^2))$. Denote $f_{1:n} = \{f_1, \ldots, f_n\}$ and $\delta_{1:n} = \{\delta_1, \ldots, \delta_n\}$. Under such

structure, Romano et al. (2022) defined the function as the minimum penalised cost for data $\{Y_1, \ldots, Y_i\}$, conditional on $f_i = f$, with the following representation

$$Q_i(f) = \min_{f_{1:i}, \delta_{2:i}, f_i = f} \Bigg\{ (1-\phi_1^2)(Y_1 - f_1)^2/\sigma_\epsilon^2 + \sum_{t=1}^{i} \left[(f_t - f_{t-1} - \delta_i)^2/\sigma_\eta^2 + \left(Y_t - f_t - \phi_1(Y_{t-1} - f_{t-1})\right)^2/\sigma_\epsilon^2 + \beta \mathbb{1}_{\{\delta_i \neq 0\}} \right] \Bigg\}. \quad (2.2.22)$$

A recursion is defined for $Q_i(f)$ for $i = 2, \ldots, n$

$$Q_i(f) = \min_{u \in \mathbb{R}} \left\{ Q_{i-1}(u) + \min\{(f-u)^2/\sigma_\eta^2, \beta\} + \left(Y_i - f - \phi_1(Y_{i-1} - u)\right)^2/\sigma_\eta^2 \right\}$$

with the start $Q_1(f) = (1-\phi_1^2)(Y_{i-1} - f)^2/\sigma_\sigma^2$, which indicates that this algorithm first finds the minimum penalised cost for set $\{Y_1, \ldots, Y_i\}$ given $f_{i-1} = u$ and $f_i = f$, and then conducts a second minimisation over u. In practice, the estimate of f_i is obtained by minimising the penalised cost for data $\{Y_1, \ldots, Y_i\}$ conditional on $f_{i+1} = \hat{f}_{i+1}$ and the change-point k is detected by considering whether $(\hat{f}_{k+1} - \hat{f}_k)^2/\sigma_\eta^2 > \beta$. Moreover, compared to Fang and Siegmund (2020), Fryzlewicz (2023) bypassed the requirement of an accurate estimation of the nuisance AR coefficients and developed the *Narrowest Significance Pursuit (NSP)* for automatic detection of localised regions for a given data sequence that each must contain a change-point.

Considering the two directions, Cho and Fryzlewicz (2023) proposed a combined methodology, *WCM.gSa*, for estimating multiple change-points in the piecewise-constant mean of an otherwise stationary, linear and autocorrelated time series. On the one hand, the *wild contrast maximisation (WCM)* principle is applied to generate the solution path for dividing change-points in mean from fluctuations in a serially dependent process. It concentrates on WBS2 proposed in Fryzlewicz (2020), where subsamples are drawn recursively based on the already detected change-point candidates to ensure the completeness of the procedure. On the other, the *gappy Schwarz algorithm (gSa)* is

constructed based on Schwarz criterion in the presence of AR errors, which can help estimate the dependence structure and the number of change-points without any direct estimation of the level of the noise. In particular, this paper considers the model (2.2.1), $Y_t = f_t + X_t$, under the following assumption of errors

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + \epsilon_t \quad \text{such that} \quad Y_t = (1 - \phi(B))f_t + \sum_{i=1}^{p} \phi_i Y_{t-i} + \epsilon_t \tag{2.2.23}$$

where the independent zero-mean ϵ_t satisfies $\mathrm{Var}(\epsilon_t) = \sigma_\epsilon^2 < \infty$ and $\phi(B) = \sum_{i=1}^{p} \phi_i B^i$ is defined with the backshift operator B. Here the order p in the AR model is unknown and could be derived with a data-driven approach in the proposed model selection methodology. Write the estimated value of p as $r \geqslant 0$ and a set of candidate change-point estimates $\mathcal{T} = \{k_j, 1 \leqslant k \leqslant m : k_1 < \cdots < k_m\} \subset \{1, \ldots, n\}$ corresponding to a candidate model. The Schwarz criterion is represented as

$$\mathrm{SC}(\{Y_t\}_{t=1}^n, \mathcal{T}, r) = \frac{n}{2} \log(\hat{\sigma}_n^2(\{Y_t\}_{t=1}^n, \mathcal{T}, r)) + (|\mathcal{T}| + r)\xi_n \tag{2.2.24}$$

where the choice of penalty parameter ξ_n is related to the distribution of ϵ_t and the applied residual sum of squares $\hat{\sigma}_n^2(\{Y_t\}_{t=1}^n, \mathcal{T}, r)$ is defined as $1/n\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|$, where $\mathbf{Y} = (Y_1, \ldots, Y_n)^\intercal$ and

$$\mathbf{X} = \underset{n\times(r+m+1)}{\begin{bmatrix} \mathbf{L}(r) & \mathbf{R}(\mathcal{T}) \\ n\times r & n\times(m+1)) \end{bmatrix}} = \begin{bmatrix} Y_0 & \ldots & Y_{1-r} & 1 & 0 & 0 & \ldots & 0 \\ \vdots & & & & & & & \\ Y_{k_1-1} & \ldots & Y_{k_1-r} & 1 & 0 & 0 & \ldots & 0 \\ Y_{k_1} & \ldots & Y_{k_1-r+1} & 0 & 1 & 0 & \ldots & 0 \\ \vdots & & & & & & & \\ Y_{n-1} & \ldots & Y_{n-r} & 0 & 0 & 0 & \ldots & 1 \end{bmatrix} \tag{2.2.25}$$

with $Y_0, \ldots, Y_{1-r}$ satisfying $\mathbb{E}(Y_t) = \mathbb{E}(Y_1)$ for any $t \leqslant 0$. Here $\mathbf{L}(r)$ and $\mathbf{R}(\mathcal{T})$ stand for the AR part and the part of mean shift modelling respectively. From the lease

squares estimation, regression parameters can be estimated with $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\alpha}}(r)^\intercal, \hat{\boldsymbol{\mu}}(\mathcal{T})^\intercal)^\intercal$, where $\hat{\boldsymbol{\alpha}}(r)$ is the estimator of AR parameters and $\hat{\boldsymbol{\mu}}(\mathcal{T})$ represents the estimated level of signals. Given the pre-specified upper bound $p_{\max}$, the unknown order p can be estimated with

$$\hat{p} = \underset{r \in \{0,\ldots,p_{\max}\}}{\arg\min} \ \mathrm{SC}(\{Y_t\}_{t=1}^n, \mathcal{T}, r) \tag{2.2.26}$$

With the participation of the gappy model sequence, this algorithm stops directly minimising the Schwarz criterion. Instead, it begins with the largest model and compares backwards with simpler models by considering the increase in the measure of fit and the model complexity resulting from the introduction of new change-point estimators with Schwarz criterion.

2.3 Wavelets

This section provides an overview of wavelets, especially Haar wavelets, which are the basis of our robust estimators of long-run variance. In addition, we review the concept of wavelet shrinkage that is a motivation of our new estimators.

2.3.1 Wavelet Analysis

Let indices j and k denote scale (or dilation) and location (or translation) parameters respectively. Those wavelet functions whose dyadic dilations and translations

$$\psi_{j,k} = 2^{j/2}\psi(2^j x - k), \quad j, k \in \mathbb{Z} \tag{2.3.1}$$

form an orthonormal basis of $L^2(\mathbb{R})$, especially the *Haar wavelet*, shall be the focus of this thesis. In general, this thesis concentrates on discretely sampled time series and hence we shall attach more importance on discrete wavelets instead of continuous ones. *Discrete wavelet transform (DWT)* (Mallat, 1989) is one specific technique that utilizes wavelet functions to achieve the multiresolution representation, making it well-suited for various signal and image processing applications.

Discrete Wavelet Transform

Corresponding to the wavelet function $\psi(\cdot)$, Mallat (1999) introduced a scaling function $\phi(\cdot)$ and scaling coefficients $c_{j,k}$. Given observations $\mathbf{Y} = (Y_1, Y_2 \ldots, Y_n)^\intercal$ with $n = 2^J$, the scaling $c_{j,k}$ and detailed $d_{j,k}$ coefficients at scale j can be computed from the scaling coefficients $c_{j+1,k}$ using the relation

$$\begin{aligned} c_{j,k} &= \sum_l h_{l-2k} c_{j+1,l} \\ d_{j,k} &= \sum_l g_{l-2k} c_{j+1,l} \end{aligned} \tag{2.3.2}$$

for $j = 1, 2, \ldots, J$ and $k = 1, 2, \ldots, 2^j$, where h_k and g_k are typically referred to as *low-pass* and *high-pass* filters in the corresponding filter bank. The coefficients $d_{j,k}$ and $c_{j,k}$ are often named as *smoothing (scaling)* and *detailed (wavelet)* coefficients respectively.

In a general context, the low-pass coefficients $c_{j,k}$ display the trend while the high-pass ones $d_{j,k}$ monitor the fluctuations present in the time series. Following Vidakovic (2009) (Chapter 3.4), we can see a comprehensive discussion on a wide variety of wavelets that can be encountered in both mathematical and statistical applications. We shall describe one key example of wavelets, Haar wavelets, employed in the following chapters.

Haar Wavelets. Haar (1910) introduced the Haar wavelet function, almost the most

well-known shape of wavelet functions. The Haar father wavelet can be described as

$$\phi(x) = \begin{cases} 1 & x \in [0,1], \\ 0 & \text{otherwise.} \end{cases} \tag{2.3.3}$$

Simultaneously, given a low-pass filter ($h_0 = h_1 = 1/\sqrt{2},\ h_k = 0$ otherwise) and a high-pass filter ($g_0 = -g_1 = 1/\sqrt{2},\ g_k = 0$ otherwise), by simple algebra, the Haar mother wavelet is defined by

$$\psi(x) = \begin{cases} 1 & x \in [0, \frac{1}{2}), \\ -1 & x \in [\frac{1}{2}, 1), \\ 0 & \text{otherwise.} \end{cases} \tag{2.3.4}$$

Therefore, suppose that n is a power of two, let $J = \log_2(n)$ and $m_j = 2^{j-1}$, $j = 1, 2, \ldots, J$. In discrete Haar wavelet transformation, the detailed coefficients $d_{j,k}$ for scales $j = 1, 2, \ldots, \log_2(n)$ and, within the j^{th} scale, for indices $k = 1, 2, \ldots, n/(2m_j)$ can be represented as

$$d_{j,k} \coloneqq \frac{1}{\sqrt{2m_j}} \left(\sum_{i=(2k-1)m_j+1}^{2km_j} Y_i - \sum_{i=(2k-2)m_j+1}^{(2k-1)m_j} Y_i \right) \tag{2.3.5}$$

Despite the fact that their lack of continuity will make it hard to approximate smooth signals, Haar wavelet functions can be helpful for analysing signals with sudden changes, and they offer significant advantages for modeling and computational techniques owing to their inherent simplicity.

Maximal Overlap Discrete Wavelet Transform

Translation invariance, i.e. a shift in the position of the original signal does not result

in different wavelet coefficients, is a desirable property for statistical modeling like regression analysis. This represents the primary limitation of the standard DWT, as it restricts its capacity to extracting information from the input vector only at specific (dyadic) locations within any given scale. To tackle this problem, Percival and Walden (2000) described the *Maximal Overlap Discrete Wavelet Transform (MODWT)*, a modified version of the DWT, as a means of obtaining a more comprehensive representation of the analysed data. In wavelet literature, multiple very similar transforms were independently discovered and assigned different names such as the *Non-decimated wavelet transform (NDWT)* in Pesquet et al. (1996), the *shift-invariant DWT (SIDWT)* in Lang et al. (1995), the *overcomplete DWT (ODWT)* in Zaciu et al. (1996) and the *redundant DWT (RDWT)* in Fowler (2005). The MODWT computes all wavelet coefficients and is no longer sensitive to the origin point.

In this thesis, we also concentrate on Maximal Overlap Haar wavelets. Mathematically speaking, given $J_0 = \lfloor n/(2m_j) \rfloor$, as MODWT filters can be defined by renormalizing the DWT filters, the wavelet coefficients $\tilde{W}_{j,k}$ can similarly be defined here for scales $j = 1, 2, \ldots, J_0$ and, within the j^{th} scale, for indices $k = 1, 2, \ldots, n - 2m_j + 1$,

$$\tilde{W}_{j,k} \coloneqq \frac{1}{2m_j} \left(\sum_{i=k+m_j}^{k+2m_j-1} Y_i - \sum_{i=k}^{k+m_j-1} Y_i \right) \tag{2.3.6}$$

Compared to DWT, MODWT can be more informative because it can be defined for all sample sizes without the dyadic restriction and retain overlapping or redundant information between scales and locations, in addition to its translation invariant property.

2.3.2 Wavelet Smoothing

To estimate a function $f : [0, 1] \to \mathbb{R}$ given noisy observations $\{Y_1, Y_2, \ldots, Y_n\}$ observed on an equally spaced grid:

$$Y_i = f\left(\frac{i}{n}\right) + X_i, \ \ i = 1, 2, \ldots, n \tag{2.3.7}$$

where the noises X_i's are zero-mean random variables, non-linear smoothing methods can show better performance for less regular functions. In particular, Donoho and Johnstone (1995), Donoho (1995), and Donoho et al. (1995) introduced the principle of *wavelet shrinkage*, a non-linear smoothing method, in their seminal papers.

The main idea of the wavelet shrinkage method is taking the wavelet transform of Equation (2.3.7) to obtain $d_{j,k} = \mu_{j,k} + Z_{j,k}$, where $d_{j,k}$ and $(\mu_{j,k}, Z_{j,k})$ are the corresponding wavelet coefficients of Y_i and $(f(i/n), X_i)$. A proper threshold is chosen to separate significantly different coefficients, thus distinguishing signal from noise. In the final stage, the values of estimate $\hat{f}$ on $\{i/n\}_{i=1}^n$ are derived from the inverse DWT. There are two main reasons supporting this approach.

1. In the wavelet domain, the signal $f(i/n)$ can be efficiently represented with *sparsity*, i.e. the wavelet coefficients $\mu_{j,k}$ corresponding to locations with smooth signals will be close to zero because of the vanishing moments property of wavelets whereas those corresponding to locations with irregular signals will significantly differ from zero;

2. Since DWT is orthogonal, white noise in the time domain can be transformed into white noise in the wavelet domain.

Therefore, the wavelet shrinkage method is proposed with the expectation that the larger empirical wavelet coefficients can predominantly capture true signals whereas

the smaller coefficients only reflect noise. Two thresholding methods that have found widespread use and undergone thorough examination are defined in Donoho and Johnstone (1994) as

$$\begin{aligned} \hat{d}^H_{j,k} &= \eta_H(d_{j,k}, \lambda) = d_{j,k}\mathbb{1}_{(|d_{j,k}>\lambda|)} \\ \hat{d}^S_{j,k} &= \eta_S(d_{j,k}, \lambda) = \text{sgn}(d_{j,k})(|d_{j,k}| - \lambda)d_{j,k}\mathbb{1}_{(|d_{j,k}|>\lambda)} \end{aligned} \tag{2.3.8}$$

for a given λ, where $\hat{d}^H_{j,k}$ and $\hat{d}^S_{j,k}$ stand for *hard* and *soft thresholding* functions respectively, and $\mathbb{1}(\cdot)$ is the indicator function. Donoho and Johnstone (1994) also provides the definition of *universal threshold*, a particularly commonly used threshold, which is denoted by $\lambda = \sigma\sqrt{2\log(n)}$. Since the standard deviation σ of the noise is unknown in practical situations, an estimator $\hat{\sigma}$ should be derived first for further non-linear wavelet estimation. For $X_i \sim iid\ N(0, \sigma^2)$, the scaled *Median Absolute Deviation (MAD)* is employed by many authors on the series of wavelet coefficients at the finest level to compute $\hat{\sigma}$, and thus help in controlling the possible upward bias caused by the presence of signal at that particular level (see e.g. Donoho and Johnstone (1994) and Johnstone and Silverman (1997)).

2.4 Long-Run Variance Estimators

As mentioned in Section 2.2.2, one particular line of research for serially correlated data is to make an extension of the existing approaches proposed for independent data, especially those relying on CUSUM or MOSUM test statistics (Cho and Fryzlewicz, 2023). When detecting a single change-point, Paul and Piotr (2022) described the extended likelihood ratio test (or CUSUM test) for data whose error process is dependent or non-Gaussian. In particular, when testing a CUSUM statistic against a threshold in a correlated data context, this corresponding threshold should be multiplied by the long-run standard deviation σ_*. Moreover, given the unknown signal

$f_t = \sum_{j=1}^{q+1} \theta_j \mathbb{1}_{(\eta_{j-1},\eta_j]}(t),\ t = 1, 2, \ldots, n$ with the locations of q (unknown) change-points satisfying $0 = \eta_0 < \eta_1 < \ldots \eta_q < \eta_{q+1} = n$ and $\theta_1, \ldots, \theta_{q+1}$ being the function values of f_t. For simplicity, Dette et al. (2020) considered the estimators of the form $\hat{f}_t = \sum_{j=1}^{\hat{q}+1} \hat{\theta}_j \mathbb{1}_{(\hat{\eta}_{j-1},\hat{\eta}_j]}(t)$, where the estimates only take values at points $0, \frac{1}{n}, \ldots, \frac{n-1}{n}, 1$ and the set of these functions is represented by $\mathcal{S}_n$. To extend SMUCE to the models with piecewise-constant regression function and dependent error process, they embedded the LRV estimator $\hat{\sigma}_*$ into the multi-scale statistic

$$V_n(Y, f) = \max_{1 \leqslant j \leqslant q+1} \max_{\substack{n\eta_j \leqslant s \leqslant e < n\eta_{j+1} \\ e-s+1 \geqslant nc_n}} \left\{ \frac{1}{\hat{\sigma}_*} \sqrt{e-s+1} \left| \frac{1}{e-s+1} \sum_{l=s}^{e} Y_l - \theta_j \right| - \sqrt{2 \log \left(\frac{\exp(1) n}{e-s+1} \right)} \right\}$$

where $\{c_n > 0\}$ is a sequence that converges to 0. Next, for a fixed threshold λ, the number of change-points is estimated by

$$\hat{q} = \hat{q}(V_n, \lambda) = \inf_{f \in \mathcal{S}_n \ V_n(Y,f) \leqslant \lambda} |\{\eta_1, \ldots, \eta_q\}|.$$

The best candidate step function for data is then identified with the following formula

$$\hat{f} = \underset{f \in \mathcal{C}(V_n, \lambda)}{\arg\min} = \sum_{i=1}^{n} \left(Y_i - f(i/n) \right)^2$$

where $\mathcal{C}(V_n, \lambda) \coloneqq \{f \in \mathcal{S}_n : \ |\{\eta_1, \ldots, \eta_q\}| = \hat{q} \text{ and } V_n(Y, f) \leqslant \lambda\}$ represents the estimated set whose elements with the minimal number of change-points $\hat{q}$ satisfying the multi-scale criterion $V_n(Y, f) \leqslant \lambda$.

To distinguish abrupt change-points in mean shifts from the random fluctuations in noise, a well-defined *long-run variance estimator* can be particularly useful since it can quantify the level of noise without being largely impacted by the presence of change in features. Here we present the definition of LRV for the error process $\{X_i\}_{i=1}^n$:

$$\sigma_*^2 = \Sigma_{\tau \in \mathbb{Z}} \mathrm{Cov}(X_0, X_\tau) = \Sigma_{\tau \in \mathbb{Z}} \gamma(\tau) \tag{2.4.1}$$

which is also known as *time-average variance constant (TAVC)* or asymptotic variance of sample mean in many existing works.

When the mean changes are not present, there are, in general, three well-known classes of methods for deriving LRV estimators: the subsampling method, resampling method and kernel-based method (Chan and Yau, 2017). First, for a pre-specified $l \in \{1, 2, \ldots, n\}$, the subsampling approach starts with dividing the whole sample into overlapping batches, or namely subsamples, $\{X_1, \ldots, X_l\}, \{X_2, \ldots, X_{l+1}\}, \ldots, \{X_{n-l+1}, \ldots, X_n\}$. In addition, the *overlapping batch means (OLBM)* estimator, a possible extension of non-overlapping batch means (NOLBM) estimator (Carlstein et al., 1986), was proposed in Meketon and Schmeiser (1984) using the batch means, which is defined as

$$\hat{\sigma}^2_{*,OBM} := \frac{l}{n-l+1} \sum_{k=l}^{n} \left[\left(\frac{1}{l} \sum_{i=k-l+1}^{k} X_i \right) - \bar{X} \right]^2 \tag{2.4.2}$$

where $\bar{X}$ denotes the overall sample mean. Although utilising overlapping batches leads to large positive correlations, the OLBM estimator is still proved to be a good one due to the pre-specified batch size. Also, Welch (1987) introduced the partially overlapping batch means estimator. In addition, we can see more relevant studies in Song and Schmeiser (1995), Alexopoulos and Goldsman (2004) and Damerdji (1994). Alexopoulos et al. (2011) employed the idea of overlapping batches to standardized time series in Schruben (1983) and developed an . Second, to mimic the behaviour of the test statistic while simultaneously keeping the original data structure, Kunsch (1989) developed a bootstrapping-based procedure, i.e. resampling method, by considering each batch as a whole and doing *iid* resampling from different batches. Based on the same setting of batches in OLBM, a new sample (resample) of size $K := ml$ is generated by randomly drawing m batches with replacement. After repeating this action of $N^{\#}$ times, we can derive the sample mean of each resample and the overall sample mean

over the $N^{\#}$ resamples with the following definitions

$$\bar{X}^{\#}_{K,j} \coloneqq \frac{1}{K}\sum_{i=1}^{K} X^{\#}_{i,j} \text{ and } \tilde{X}^{\#}_{K,N^{\#}} \coloneqq \frac{1}{N^{\#}}\sum_{j=1}^{N^{\#}} \bar{X}^{\#}_{K,j} \tag{2.4.3}$$

for $j = 1, 2, \ldots, N^{\#}$. Furthermore, the jackknife and the bootstrap (JB) estimator (Kunsch, 1989) is given as follows

$$\hat{\sigma}^2_{*,JB} \coloneqq \frac{K}{N^{\#}}\sum_{j=1}^{N^{\#}} \left(\bar{X}^{\#}_{K,j} - \tilde{X}^{\#}_{K,N^{\#}}\right)^2. \tag{2.4.4}$$

Related works include the matched-block bootstrap estimator in Carlstein et al. (1998), circular block-resampling estimator in Romano (1992), dependent wild bootstrap estimator in Shao (2010a), tapered block bootstrap (TAB) estimator in Paparoditis and Politis (2001) and extended TAB in Shao (2010b). These subsampling estimators can result in small biases and mean square errors but require a pre-specified number of subsamples and often suffer from high computational complexity. The final kernel-based estimators can be constructed from a good combination of the sample autocovariance $\hat{\gamma}(\tau)$. Since the direct summation of $\hat{\gamma}(\tau)$ leads to inconsistent results, different kernels are applied to assign weights to $\hat{\gamma}(\tau)$. And the estimator may be defined as

$$\hat{\sigma}^2_{*} = \sum_{\tau=-(n-1)}^{n-1} W\left(\frac{\tau}{b_n}\right)\hat{\gamma}(\tau) \tag{2.4.5}$$

where $W : [-1, 1] \to \mathbb{R}$ is a user-specified kernel and b_n is the bandwidth parameter, which is hard to choose unless we are provided with the detailed dependence structure (Politis and Romano, 1995).

Later, considering the presence of mean changes, we face an increase in the difficulty of estimating the LRV, and it becomes even more complicated for series with (possibly) multiple change-points. Well-known LRV estimators for non-constant mean trend

can generally be divided into two broad classes: residual-based and difference-based estimators. Here we provide a literature review on both of them.

As indicated by the name, residual-based LRV estimators are constructed relying on the residuals $\hat{X}_i = Y_i - \hat{f}_h(i)$, where $\hat{f}_h(\cdot)$ is a non-parametric estimator of the unknown signal derived with the bandwidth or smoothing parameter h. Spline smoothing and kernel smoothing are two commonly used approaches for estimation while an AR process is commonly applied to model the noise. For example, to estimate the unknown smooth function $f(t_i)$ defined on $[0, 1]$ for $t_i = i/n$, Truong (1991) proposed a sequence of estimators based on local averages under the assumption of a Gaussian error process. Given $\{\delta_n\}_{n\geqslant 1}$, a sequence of positive numbers that gradually decreases towards zero, and $t \in [0, 1]$, define the blocks

$$I_n(t) = \{i : 0 \leqslant i \leqslant n |\ |t_i - t| \leqslant \delta_n\} \text{ and } N_n(t) = |I_n(t)|$$

The kernel estimator, or moving average estimator, is then represented by

$$\hat{f}(t) = \frac{1}{N_n(t)} \sum_{I_n(t)} Y_i$$

Due to computational expediency and the demand of single optimisation, Shao and Yang (2011) focused on the spline smoothing instead of kernel smoothing. This paper applies polynomial splines to estimate the trend function $f(t_i)$ by dividing the interval $[0, 1]$ into subintervals with width h. Moreover, compared to these two estimators without introducing an optimal smoothing parameter, Qiu et al. (2013) proposed a modified moving average estimator with an automatic selection of optimal parameter. Without the assumption of Gaussianity of data, the proposed trend estimator is defined

as follows

$$\hat{f}(t/n) = \begin{cases} \frac{1}{2q+1}\sum_{i=t-q}^{t+q} X_i, & q+1 \leqslant t \leqslant n-q \\ \frac{1}{N_{1t}}\sum_{i=1}^{t+q} X_i - \frac{1}{N_{2t}}\sum_{i=1}^{t+q}(i-t)X_i, & 1 \leqslant t \leqslant q \\ \frac{1}{N_{3t}}\sum_{i=t-q}^{n} X_i - \frac{1}{N_{4t}}\sum_{i=t-q}^{n}(i-t)X_i, & n-q+1 \leqslant t \leqslant n \end{cases} \tag{2.4.6}$$

where N_{1t}, N_{2t}, N_{3t} and N_{4t} are time-dependent values that follow

$$\begin{aligned} N_{1t}^{-1} &= \frac{4q^2 - 4qt + 6q + 4t^2 - 6t + 2}{(q+t)(q+t-1)(q+t+1)} \\ N_{2t}^{-1} &= \frac{6(q-t+1)}{(q+t)(q+t-1)(q+t+1)} \\ N_{3t}^{-1} &= \frac{4(n-t))^2 + 4q^2 - 4q(n-t) + 2(n+q-t)}{(n+q-t+2)(n+q-t+1)(n+q-t)} \\ N_{4t}^{-1} &= \frac{6(n+q-t)}{(n+q-t+2)(n+q-t+1)(n+q-t)} \end{aligned}$$

For all of the approaches above, an Yule–Walker estimator can then be derived for $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)^\intercal$ in AR(p) error process $X_i = \sum_{k=1}^{p} \phi_k X_{i-k} + \epsilon_i$ with $\mathbb{E}(\epsilon_i) = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma^2$ based on the residuals $\hat{\mathbf{X}} = \mathbf{Y} - \mathbf{f}$, which is given as

$$\hat{\boldsymbol{\phi}} = \hat{\Gamma}^{-1}\hat{\boldsymbol{\gamma}}\ , \quad \hat{\gamma}(k) = \frac{1}{n}\sum_{i=1}^{n-k} \hat{X}_i \hat{X}_{i+k}$$

where $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}(1), \ldots, \hat{\gamma}(p))$ and $\hat{\Gamma}$ is a $p \times p$ estimated covariance matrix with $\hat{\Gamma}_{ij} = \hat{\gamma}(i-j)$. Then the variance σ^2 and LRV σ_*^2 can be estimated by

$$\hat{\sigma}^2 = \hat{\gamma}(0) - \hat{\boldsymbol{\gamma}}^\intercal \hat{\boldsymbol{\phi}} \ \text{ and } \ \hat{\sigma}_*^2 = \frac{\hat{\sigma}^2}{(1 - \sum_{i=1}^{p} \hat{\phi}_i)^2} \tag{2.4.7}$$

On the other hand, the pre-estimation of signal function f can be indeed difficult when there are dominant random fluctuations from the error process (Tecuapetla-Gómez and Munk, 2017). To overcome this issue, difference-based techniques are developed

by building the whole estimator on the l-th differences $Y_i - Y_{i-l}$ of the observed time series, which as first introduced in Hall et al. (1990). In particular, a difference sequence $\{\Delta_j\}$ is a sequence of real numbers such that

$$\sum \Delta_j = 0, \text{ and } \sum \Delta_j^2 = 1 \tag{2.4.8}$$

Assume that $\Delta_j = 0$ for $j < -l_1$ and $j > l_2$ and $\Delta_{-l_1}\Delta_{l_2} \neq 0$ for $l_1, l_2 \geqslant 0$. The term $l = l_1 + l_2$ is then regarded as the order of the sequence, where $l_1 = 0$ and $l_2 = l$ are usually selected for simplicity. Based on the difference sequence, a difference-based estimator of $\gamma(0)$ is defined with the following format

$$\hat{\gamma}(0) = \frac{1}{n-l} \sum_{k=l_1+1}^{n-l_2} \left(\sum_{i=-l_1}^{l_2} \Delta Y_{i+k} \right)^2 .$$

Similar estimators have been explored extensively when there are no discontinuities or high fluctuation in signals and we can see more examples in Muller and Stadtmuller (1987), Dette et al. (1998) and Brown and Levine (2007), etc. In particular,Muller and Stadtmuller (1987) also introduced estimators for autocovariances in the presence of stationary m-dependent errors; Hall and Keilegom (2003) proposed autocovariance estimates of autoregressive noise. To avoid the restriction on the dependence structure in noise, Tecuapetla-Gómez and Munk (2017) considered the m-dependent errors and proposed biased-optimized estimates that only depend on m, where the estimator leading to the smallest bias is defined as

$$\hat{\gamma}^{(m)}(0) = \frac{1}{6n_m} \sum_{i=1}^{n_m} \left(Y_i - 2Y_{i+(m+1)} + Y_{i+2(m+1)} \right)^2, \quad n_m = n - 2(m+1)$$

The other values of the autocovariance function $\gamma(1), \ldots, \gamma(m)$ can also be represented with an m-dependent formula. However, to construct the LRV $\sigma_*^2 = \sum \gamma(\tau)$ from autocovariances, other methods like HAC-type estimation procedure are required to define the final version of estimated LRV (Khismatullina and Vogt, 2020); see equation

(2.4.5) as an example. Here *HAC* is the abbreviation of *"Heteroskedasticity and Autocorrelation Consistent"*. Because the performance of LRV estimators can be largely impacted by the choice of the smoothing parameter in HAC-type estimation, Khismatullina and Vogt (2020) mentioned that the assumption of a time series model on $\{X_i\}_{i=1}^n$ could be really helpful even though some biases possibly exists. Then they constructed a difference-based estimator by considering the error process as the class of AR(p) model.

Considering more general error processes, Wu and Zhao (2007) proposed three asymptotically consistent LRV estimators under the assumption of Lipschitz continuous function, i.e. given an interval I on $\mathbb{R}$, a function f is considered as Lipschitz continuous on I if $\sup_{x_1 \neq x_2} |f(x_1) - f(x_2)|/|x_1 - x_2| < \infty$ for any $x_1, x_2 \in I$. Before creating the estimators, the sample is first divided into $m_n = \lfloor n/k_n \rfloor$ blocks $\{Y_{1+ik_n}, \ldots, Y_{(i+1)k_n}\}$ for $i = 0, 1, \ldots, m_n - 1$ and the local averages are derived as follows

$$A_i = \frac{1}{k_n} \sum_{j=1}^{k_n} Y_{j+ik_n}$$

Then the (absolute) difference-based estimator is developed as

$$\hat{\sigma}_* = \frac{\sqrt{\pi k_n}}{2(m_n - 1)} \sum_{i=1}^{m_n - 1} |A_i - A_{i-1}| \tag{2.4.9}$$

which can be quite sensitive to abrupt changes in signals. Among these three estimators, the median-based one can be more robust to large jumps in signal, which is introduced as follows.

$$\hat{\sigma}_* = \frac{\sqrt{k_n}}{\sqrt{2 z_{1/4}}} \text{median}(|A_i - A_{i-1}|,\ 1 \leqslant i \leqslant m_n - 1) \tag{2.4.10}$$

where $z_{1/4}$ denotes the third quartile of $N(0, 1)$. When extending SMUCE for dependent data, Dette et al. (2020) applied the mean-based estimator in Wu and Zhao (2007)

and further proved its consistency in the presence of piecewise-constant signals. We can see its definition in the following

$$\hat{\sigma}_*^2 = \frac{k_n}{2(m_n - 1)} \sum_{i=1}^{m_n - 1} |A_i - A_{i-1}|^2. \tag{2.4.11}$$

Considering the possibly nonstationary error process, Dette et al. (2019) introduced the following LRV estimator for data with piecewise-constant signals. First, they define $S_{k,r} = \sum_{i=k}^{r} Y_i$ and $\Delta_{S,j} = (S_{j-m+1,j} - S_{j+1,j+m})/m$ for $m \geqslant 2$. Then, for $t \in [m/n, 1 - m/n]$,

$$\hat{\sigma}_*^2 = \sum_{j=1}^{n} \frac{m\Delta_{S,j}^2}{2} \omega(t, j) \tag{2.4.12}$$

where for some bandwidth $\tau_n \in (0, 1)$,

$$\omega(t, j) = K\left(\frac{j/n - t}{\tau_n}\right) \Big/ \sum_{j=1}^{n} K\left(\frac{j/n - t}{\tau_n}\right).$$

For constant variance, the estimator can be reduced to

$$\hat{\sigma}_*^2 = \sum_{j=1}^{n} \frac{m\Delta_{S,j}^2}{2n} \tag{2.4.13}$$

which is a kind of overlapped version of estimator (2.4.11). Based on the flat-top kernels in Politis and Romano (1995), Eichinger and Kirch (2018) proposed the following time dependent MOSUM version of it for estimating LRV in the presence of dependent data

$$\hat{\sigma}_{k,*}^2 = \hat{\gamma}_k(0) + 2 \sum_{h=1}^{\Lambda_n} \omega(h/\Lambda_n) \hat{\gamma}_k(h) \tag{2.4.14}$$

with pre-specified bandwidth Λ_n and suitable weights ω, where autocovariances are

estimated by

$$\hat{\gamma}_k(h) = \frac{1}{2G} \sum_{i=k-G+1}^{k-h} (X_i - \bar{X}_{k-G+1,k})(X_{i+h} - \bar{X}_{k-G+1,k}) + \frac{1}{2G} \sum_{i=k+1}^{k+G-h} (X_i - \bar{X}_{k+1,k+G})(X_{i+h} - \bar{X}_{k+1,k+G}).$$

Also, McGonigle and Cho (2023) introduced a robust scale-dependent TAVC for many existing multi-scale change-point detection approaches, see more details in Section 3.1.2. More recently, Chan (2022) proposed a general framework for LRV estimation for data with serially correlated noise and non-constant mean trends. Take the definition of difference sequence (2.4.8), then the mth order lag-h difference statistics are defined as

$$D_i = \sum_{j=0}^{m} \Delta_j Y_{i-jh}, \quad i = mh+1, \ldots, n$$

and the mth order difference-based estimator is then developed as

$$\hat{\sigma}_*^2 = \sum_{|k|<l} W\left(\frac{k}{l}\right) \hat{\gamma}^D(k), \quad \text{with } \hat{\gamma}^D(k) = \frac{1}{n} \sum_{i=mh+|k|+1}^{n} D_i D_{i-|k|} \tag{2.4.15}$$

where the kernel function $W(\cdot)$ is similar to that defined in (2.4.5).

In Chapter 3, we propose several robust wavelet-based LRV estimators and conduct comparative simulation studies with some existing LRV estimators, see Section 3.4 for detailed results.

2.5 Narrowest-Over-Threshold Technique

The change-point detection (CPD) method for dependent data discussed in Chapter 4 of this thesis is based upon the *Narrowest-Over-Threshold (NOT)* detection device

proposed in Baranowski et al. (2019). In this section, we review the underlying ideas of NOT algorithm and provide a detailed description of the *NOT solution path algorithm*.

2.5.1 The NOT algorithm

To search for the an unknown number of features in signal f_i, NOT skillfully conducts both "global" and "local" treatment of the observations Y_i's and hence help detect change-points in a multi-scale manner. From the "global" perspective, the NOT algorithm starts with a random selection of a group of special subsamples $(Y_{s+1}, \ldots, Y_e)^\intercal$, $0 \leqslant s < e \leqslant n$, which are assumed to contain at most one change-point. Various contrast functions are proposed to meet the requirement of searching for the most possible location of the change-point for different type of signals. On the other hand, the "local" stage concentrates on finding the "narrowest-over-threshold" interval that is highly likely to contain the single change-point, i.e. among all subsamples containing contrast larger than a preset threshold, NOT tends to pay more attention to the contrast derived from the interval with the smallest range. Such a technique makes NOT effective in detecting the unknown number and locations of change-points in f_i for data with different models of signal and noise.

Among all stages, the selection of a contrast function $\mathcal{C}^b_{(s,e]}(\cdot)$ is of great importance and it largely relies on the underlying model chosen for signal f_i and noise X_i within the data. We shall review the two particular tailor-made contrast functions corresponding to data with *iid* Gaussian noise plus precewise-constant or precewise-linear signals.

For example, when signal f_t is precewise-constant, for any integer triple (s, e, b) with $0 \leqslant s < b < e \leqslant n$, the contrast vector $\boldsymbol{\psi}^b_{(s,e]} = (\psi^b_{(s,e]}(1), \psi^b_{(s,e]}(2), \ldots, \text{`}\psi^b_{(s,e]}(n))^\intercal$ is

defined as

$$\psi^b_{(s,e]}(t) = \begin{cases} \sqrt{\frac{e-b}{(e-s)(b-s)}} & t = s+1, \ldots, b, \\ -\sqrt{\frac{b-s}{(e-s)(b-s)}} & t = b+1, \ldots, e, \\ 0 & \text{otherwise.} \end{cases} \tag{2.5.1}$$

In contrast, if $b \notin \{s+1, \ldots, e-1\}$, $\psi^b_{(s,e]}(t)$ is set to be zero for all t. For any vector $\mathbf{v} = (v_1, \ldots, v_n)^\intercal$, the contrast function is defined as

$$C^b_{(s,e]}(\mathbf{v}) = \left| \left\langle \mathbf{v}, \boldsymbol{\psi}^b_{(s,e]} \right\rangle \right|. \tag{2.5.2}$$

Also, when the continuous signal f_t is precewise-linear, for any integer triple (s, e, b) with $0 \leqslant s < e \leqslant n$ and $s+1 < b < e$, the contrast vector $\boldsymbol{\phi}^b_{(s,e]} = (\phi^b_{(s,e]}(1), \phi^b_{(s,e]}(2), \ldots, \phi^b_{(s,e]}(n))^\intercal$ is considered as

$$\phi^b_{(s,e]}(t) = \begin{cases} \alpha^b_{(s,e]}\beta^b_{(s,e]}[\{3(b-s)+(e-b)-1\}t - \{b(e-s-1)+2(s+1)(b-s)\}] & t = s+1, \ldots, b, \\ -\frac{\alpha^b_{(s,e]}}{\beta^b_{(s,e]}}[\{3(e-b)+(b-s)+1\}t - \{b(e-s-1)+2e(e-b+1)\}] & t = b+1, \ldots, e, \\ 0 & \text{otherwise.} \end{cases}$$

where $\alpha^b_{(s,e]} = \sqrt{\frac{6}{l(l^2-1)(1+(e-b+1)(b-s)+(e-b)(b-s-1))}}$, $\beta^b_{(s,e]} = \sqrt{\frac{(e-b+1)(e-b)}{(b-s-1)(b-s)}}$ and $l = e-s$. Similarly, if $b \notin \{s+2, \ldots, e-1\}$, $\psi^b_{(s,e]}(t)$ will be zero for all t. For any vector $\mathbf{v} = (v_1, \ldots, v_n)^\intercal$, the contrast function is defined as

$$C^b_{(s,e]}(\mathbf{v}) = \left| \left\langle \mathbf{v}, \boldsymbol{\phi}^b_{(s,e]} \right\rangle \right|.$$

In Figure 2.2, plots of $\boldsymbol{\psi}^b_{(s,e]}$ and $\boldsymbol{\phi}^b_{(s,e]}$ are presented as an illustration for contrast vectors over different (s, e, b).

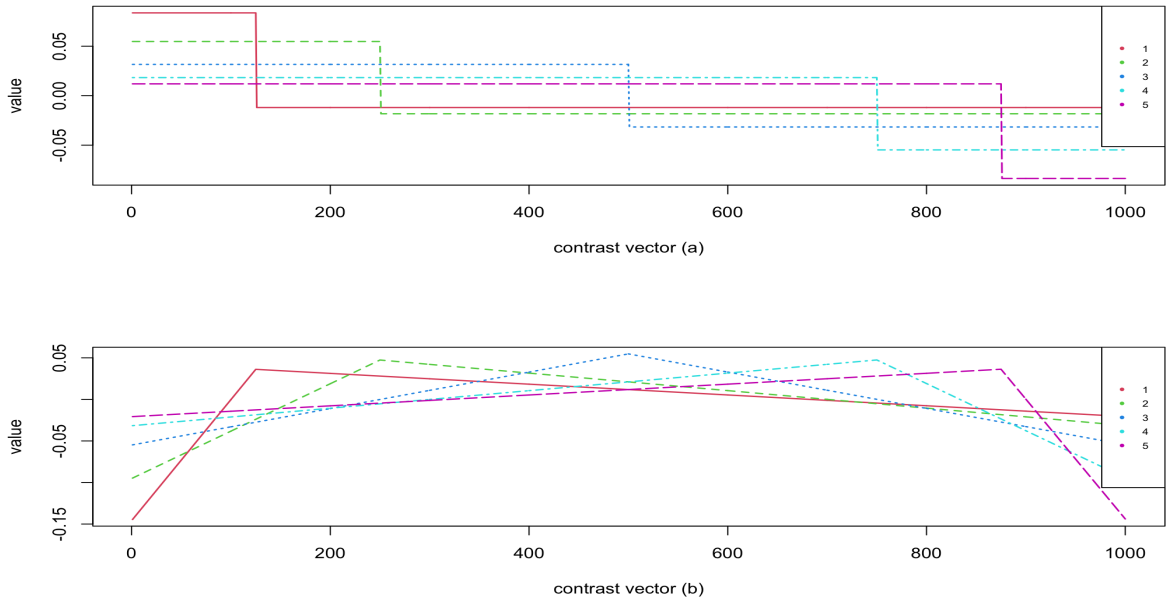


Figure 2.2: Plots (a) and (b) show respectively values of $\boldsymbol{\psi}^b_{(s,e]}$ and $\boldsymbol{\phi}^b_{(s,e]}$ for $s = 0, e = 1000$ and $b = 125, 250, 500, 750, 875$, where colour number 1 to 5 given on graphs indicate values of b from 125 to 875.

2.5.2 The NOT solution path algorithm

Besides the contrast function, two vital tuning parameters are required to be considered as well: a user-specified threshold satisfying $\zeta_n > 0$ and the number M of the intervals randomly drawn in the process. Notably, since the performance of the original NOT algorithm is largely impacted by the choice of ζ_n, it may not show a good performance without a well-selected threshold.

To overcome this issue, the NOT solution path algorithm combines the original NOT with a well-suited information-based criterion, allowing for the automatic determination of the threshold ζ_n. This innovation eliminates the requirement for a predefined threshold, rendering the algorithm entirely threshold-free. Simultaneously, when considering a fixed number of well-spaced change-points, Baranowski et al. (2019) mention

that the minimum required value for M increases with sample size n and 10000 is the number recommended for datasets of lengths of the order of thousands.

In the following, we present the detailed pseudo-code for the NOT solution path algorithm developed in the supplementary material of NOT (Baranowski et al., 2019); see Algorithm 1. Let $\mathcal{T}(\zeta_n) = \{\hat{\eta}_1(\zeta_n), \ldots, \hat{\eta}_{\hat{q}(\zeta_n)}(\zeta_n)\}$ denote the locations of change-points estimated by the NOT algorithm with threshold ζ_n. It is also assumed that there exists thresholds satisfying $0 = \zeta_n^{(1)} < \zeta_n^{(2)} < \cdots < \zeta_n^{(N)}$, where $\mathcal{T}(\zeta_n^{(i)}) \neq \mathcal{T}(\zeta_n^{(i+1)})$ for all $i = 1, 2, \ldots, N-1$, $\mathcal{T}(\zeta_n) = \mathcal{T}(\zeta_n^{(i)})$ for all $\zeta_n \in [\zeta_n^{(i)}, \zeta_n^{(i+1)})$ and $\mathcal{T}(\zeta_n) = \emptyset$ for all $\zeta_n \leqslant \zeta_n^{(N)}$. Algorithm 1 mainly works by relying on the iterative application of information from $\mathcal{T}(\zeta_n^{(i)})$ to obtain both $\zeta_n^{(i+1)}$ and the corresponding $\mathcal{T}(\zeta_n^{(i+1)})$ for any $i = 1, 2, \ldots, N-1$. To sum up, it produces the entire threshold-indexed solution path $\{\mathcal{T}(\zeta_n)\}$ for $\zeta_n \geqslant 0$ with low computational complexity, and it can be employed directly with the help of the **R** package **not** and **breakfast**.

In Algorithm 1, F_n^M represents a collection of M left-open and right-closed intervals that encompass all randomly selected sub-intervals for testing. In general, given a fixed threshold ζ_n, a binary tree structure can be constructed in accordance with the detected sequence of change-points, starting with $(s, e] = (0, n]$. And then the initial change-point identified within the $(0, n]$ interval is treated as the central node of the tree. Subsequently, its branches are expanded in a recursive manner.

To be specific, every tree node holds the following three features. First, the interval $(N.s, N.e]$ is the target to test for each node. Second, the narrowest-over-threshold sub-interval is chosen from the intervals resulting from the intersection of $(N.s, N.e]$ and elements in F_n^M. This algorithm denotes $N.c$ as the highest value of the contrast function among all potential data locations, and $N.b$ as the corresponding location within the same sub-interval. $N.Left$ and $N.Right$ represent the nodes of the subsequently detected change-points in sub-intervals $(N.s, N.b]$ and $(N.b, N.e]$, respectively.

Algorithm 1 NOT solution path

Input: Data vector $\mathbf{Y}$, all sub-intervals $(s_m, e_m] \in F_n^M$ together with
$b_m \coloneqq \arg\max_{s_m < b \leqslant e_m} \mathcal{C}^b_{(s_m,e_m]}(\mathbf{Y})$, $c_m \coloneqq \mathcal{C}^{b_m}_{(s_m,e_m]}(\mathbf{Y})$, $l_m \coloneqq e_m - s_m$
Output: Thresholds $0 = \zeta_n^{(1)} < \zeta_n^{(2)} < \cdots < \zeta_n^{(N)}$ and sets of estimated change-points $\mathcal{T}(\zeta_n^{(1)}), \mathcal{T}(\zeta_n^{(2)}) \ldots \mathcal{T}(\zeta_n^{(N)})$
To start the algorithm: Call SOLUTIONPATH()
procedure BUILDBINARYTREE($(s, e], \zeta_n$, N)
    $\mathcal{M}_{(s,e]} \coloneqq$ set of those $m \in \{1, 2, \ldots, M\}$ such that $(s_m, e_m] \subset (s, e]$
    $\mathcal{O}_{(s,e]} \coloneqq$ set of those $m \in \mathcal{M}_{(s,e]}$ such that $c_m > \zeta_n$
    if $\mathcal{O}_{(s,e]} = \emptyset$ **then**
        N=NULL
    else
        $k \coloneqq$ any element of $\arg\min_{m \in \mathcal{M}_{(s,e]}} l_m$
        N.b $\coloneqq b_k$, N.c $\coloneqq c_k$, N.Left $\coloneqq$ NULL, N.Right $\coloneqq$ NULL
        BUILDBINARYTREE($(s$, N.b$], \zeta_n$, N.Left)
        BUILDBINARYTREE((N.b, $e], \zeta_n$, N.Right)
    end if
end procedure
procedure UPDATEBINARYTREE($(s, e], \zeta_n$, N)
    if $N.c \leqslant \zeta_n$ **then**
        BUILDBINARYTREE($(s, e], \zeta_n$, N)
    else
        if N.Left $\neq$ NULL **then**
            UPDATEBINARYTREE($(s$, N.b$], \zeta_n$, N.Left)
        end if
        if N.Right $\neq$ NULL **then**
            UPDATEBINARYTREE((N.b, $e], \zeta_n$, N.Right)
        end if
    end if
end procedure
procedure SOLUTIONPATH()
    Set $\mathrm{N}_r \coloneqq$ NULL, $i \coloneqq 1$, $\zeta_n^{(1)} \coloneqq 0$
    BUILDBINARYTREE($(0, n], \zeta_n^{(1)}, \mathrm{N}_r$)
    while $\mathrm{N}_r \neq \emptyset$ **do**
        $\mathcal{D} \coloneqq \{\mathrm{N}_r$and all its children nodes$\}$
        $\mathcal{T}(\zeta_n^{(i)}) \coloneqq \{\mathrm{N.b} | \mathrm{N} \in \mathcal{D}\}$
        $\zeta_n^{(i+1)} \coloneqq \min_{\mathrm{N} \in \mathcal{D}}\{\mathrm{N.c}\}$
        UPDATEBINARYTREE($(0, n], \zeta_n^{(i+1)}, \mathrm{N}_r$)
        $i \coloneqq i + 1$
    end while
end procedure

Choice of ζ_n via the *strengthened Schwarz Information Criterion (sSIC)*

As mentioned in the beginning of Section 2.5.2, finding a proper information-based criterion supports the automatic selection of ζ_n. For $k = 1, \ldots, N$ let $\hat{q}_k = |\mathcal{T}(\zeta_n^k)|$, $\hat{\boldsymbol{\Theta}}_1, \hat{\boldsymbol{\Theta}}_2, \ldots, \hat{\boldsymbol{\Theta}}_{\hat{q}_k+1}$ denote the maximum likelihood estimators of the parameters obtained via the estimated change-points $\hat{\tau}_1, \hat{\tau}_2, \ldots, \hat{\tau}_{\hat{q}_k} \in \mathcal{T}(\zeta_n^k)$, and n_k denote the total number of estimated parameters, from $\hat{\tau}_1, \hat{\tau}_2, \ldots, \hat{\tau}_{\hat{q}_k}$ to free parameters in $\hat{\boldsymbol{\Theta}}_1, \hat{\boldsymbol{\Theta}}_2, \ldots, \hat{\boldsymbol{\Theta}}_{\hat{q}_k+1}$. In Algorithm 1, Baranowski et al. (2019) choose the threshold ζ_n^k and the associated estimated locations of change-points $\mathcal{T}(\zeta_n^k)$ by minimising the sSIC defined below.

$$\mathrm{sSIC}(k) = -2 \sum_{j=1}^{\hat{q}_k+1} \log(l(Y_{\hat{\tau}_{j-1}+1}, \ldots, Y_{\hat{\tau}_j}; \hat{\boldsymbol{\Theta}}_j)) + n_k \log^{\alpha}(n) \tag{2.5.3}$$

for a pre-specified $\alpha \geqslant 1$, $\hat{\tau}_0 = 0$ and $\hat{\tau}_{\hat{q}_k+1} = n$.

Besides the proved consistency and near-optimality of NOT for generalised change-points detection, Baranowski et al. (2019) also present the simulation results for different distributions of the error process X_t, from heavy-tailed noise to weakly dependent noise. In Chapter 4, we aim to practically explore several possible extensions of the NOT solution path algorithm for different serial correlated data.

2.6 Lead-lag Relationship

As mentioned in the introduction, Chapter 5 intends to propose an exploratory method for the analysis of the lead-lag or causal relationships between two nonstationary time series. To that end, we shall first provide a general review on basic concepts of these two patterns.

The *lead-lag effect* is a kind of phenomenon that sees changes in one time series influence

another time series, but usually with a time delay. Specifically, it characterises the order of the arrival of changes in two variables, i.e. the concept that the changes in one variable lead (lag) the changes in the other indicates that the the first (second) variable may impact the second (first). This phenomenon is widely examined in various fields, especially economics and finance. In particular, after figuring out how trends in prices of certain financial commodities follow that of some other commodities, conservative investors can potentially use the information of specific "time lag" to make more profits without taking more risks. The *lead-lag relationship* between stock index and stock index futures has been extensively studied to derive a (possible) price discovery function in a capital market without perfect efficiency, but the findings drawn from different markets and time periods via different approaches are indeed contradictory, referring to Gong et al. (2016) for a comprehensive review. For example, Kawaller et al. (1987), Abhyankar (1995) and Hasbrouck (2003) indicated that price movements of futures lead that of index in the short-run whereas some other studies such as Shyy et al. (1996) mentioned the opposite. More recently, Li et al. (2022) introduced the idea that the definition of the "lead–lag effect" slightly differs from that of "lead–lag relationship", where the latter one is a short-run event that can be random and hence may not offer us too much information. In this paper, Li et al. (2022) formally defined the lead-lag effect and developed a lead–lag investment strategy by exploring the availability of the well-known power-law function (Clauset et al., 2009) in stock trading and considering the corresponding conditions should be satisfied. Since we consider relatively long-run features in bi-variate data, we use the terms "lead–lag effect" and "lead–lag relationship" interchangeably in the following.

In order to identify (possible) lead-lag relationships between time series, many conventional approaches, such as the cross-correlation function (CCF), vector error correction (VEC) model, vector autoregressive (VAR) model, the generalized autoregressive conditional heteroskedasticity (GARCH) model and Granger-causality test, are commonly applied to detect the dependence of two series within a certain period. The early inves-

tigations into relationships between time series were conducted via simple correlation, where CCF is a typical measure of similarities. To be specific, the *cross-correlation function* can measure the correlation between two time series $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$ at different lags. At each time lag k, we have the sample CCF defined below

$$R_{XY}(k) = \begin{cases} \frac{1}{n}\sum_{i=1}^{n-k} \frac{(X_i-\bar{X})(Y_{i+k}-\bar{Y})}{s_X s_Y} & k = 1, 2, \ldots, n-1, \\ \frac{1}{n}\sum_{i=1}^{n+k} \frac{(Y_i-\bar{Y})(X_{i-k}-\bar{X})}{s_X s_Y} & k = -1, -2, \ldots, -(n-1) \end{cases} \tag{2.6.1}$$

where $\bar{X}$ and $\bar{Y}$ denote the sample mean, and s_X and s_Y represent the sample standard deviation. Its results are often interpreted through the sign and magnitude of the correlation $R_{XY(k)}$ at different lags k. In particular, when one or more $R_{XY}(k)$ are significant with positive k's, $\{X_i\}_{i=1}^n$ can be considered to lead $\{Y_i\}_{i=1}^n$; similarly, $\{X_i\}_{i=1}^n$ can be considered to lag $\{Y_i\}_{i=1}^n$ if one or more $R_{XY}(k)$ are significant with negative k's. Based on CCF, Chapter 11 in Box et al. (2015) introduces a straightforward lead–lag modeling method over assumed transfer function models. However, compared to regression that can summarise the association, it can be quickly recognised that correlation does not possess a natural direction (Hoover, 2008).

By employing a VAR or VEC model, the direction of lead-lag relationships is naturally described in the time domain and their existence among the involved time series can be indicated by the presence of non-zero or statistically significant lagged coefficients of the explanatory variables (Skoura, 2019). In econometrics, an *vector error correction model* is often estimated to describe the co-integration among nonstationary time series, which refers to the property of these time series sharing a common stochastic trend. The long-run association implied by co-integration is compatible with a lead-lag relationship in the short run (Kanas and Kouretas, 2005), and more examples are shown in, for example, Corhay et al. (1993) and Chung and Liu (1994). Moreover, the *Autoregressive Conditional Heteroskedasticity (ARCH)* model is often applied to

to capture the properties on volatility in time series data. The model is defined as

$$X_t = \sigma_t \epsilon_t, \quad \sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}^2 \tag{2.6.2}$$

where $\{\epsilon_t\}_{t=1}^n$ follows an *iid* distribution with zero mean and unit variance, and $\alpha_0 > 0$, $\alpha_i \geqslant 0$ for all $i > 0$. It was extended to be GARCH model by Bollerslev (1987) to further characterise the persistence nature of volatility, with the following definition

$$X_t = \sigma_t \epsilon_t, \quad \sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}^2 + \sum_{i=1}^{q} \beta_j \sigma_{t-j}^2 \tag{2.6.3}$$

with the same assumption of ϵ_t and α_0, together with $\alpha_i \geqslant 0$, $\beta_j \geqslant 0$ and $\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j < 1$. Kavussanos et al. (2008) introduced a VECM-GARCH-X model to allow for the time-varying variance and covariances of the series, see also the bi-variate error-correction EGARCH model in Zhong et al. (2004).

Granger (1988) demonstrated that for a pair of co-integrated series, there must be some Granger causation between them in at least one direction. Therefore, with the development of co-integration analysis, the studies of lead-lag relationships become closely related to the causal inference, especially Granger causality, with the application of a VAR model Skoura (2019). The Granger causality test is hence quite essential for the detection of lead-lag relationships, see examples in, for example, Jiang et al. (2019), Scherbina and Schlusche (2020), Otneim et al. (2022) and Zeng and Atta Mills (2023). In the next subsection, we shall have a brief review of the concept of causality in time series, especially the Granger causality. Although it is well-known that the Granger causality may not imply a substantive cause-effect relationship due to omission of relevant variables, nevertheless the tested associations between variables can still be useful for empirical causality investigation (Eichler, 2013).

In addition, to capture the dynamic nature of the lead-lag relationships, new approaches

including genetic programming (GP) (Lien et al., 2003), wavelet analysis (Reboredo and Rivera-Castro, 2013; Kim and In, 2005), the thermal optimal path (TOP) method (Gong et al., 2016; Meng et al., 2017), dynamic time warping (DTW) (Ma et al., 2022; Wang et al., 2012) and Aligned Correlation (AC) (Gupta and Chatterjee, 2020), etc. were developed for relation discovery.

2.6.1 Causality in time series

Generally speaking, *causality* represents the influence that changes of one variable may have by causing the changes of the other in a system. Many research questions are inherently causal and the increasing demand to understand such underlying patterns between variables largely promotes the development of the study of causality. In this section, we shall mainly provide an overview of causality in time series.

When defining causality, *temporal precedence* (causes happen before their effects) and *physical influence* (changes in causes lead to changes in effects) are the two most significant properties, where the second aspect often lies at the heart of existing literature on causal relationships (see more in Eichler (2012)). In time series analysis, since it is usually impractical to set controlled experiments, many existing approaches of causal inference tend to focus on temporal precedence, a readily available property in data. There are, in general, four possible existing definitions of causality utilised for time series, such as intervention causality in Eichler (2012), structural causality in White and Lu (2010), Granger causality and Sims causality in Sims (1972). In particular, *Granger causality* is a statistical concept of causality providing information about the capability of prediction and is probably the most widely used concept (see details in Granger (1969), Granger (1980) and Granger (1988)). In the original definition, Granger causality is a measure of the relationship between the variables and is supposed to be analysed with all relevant information. On the other hand, observed information can be incom-

plete and the there could be loss of crucial variables, and hence Hsiao (1982) proposed concepts for spurious causality among variables in a tri-variate model.

In the meantime, graphical models, and in particular directed acyclic graphs, have been widely used to describe and infer causal relationships in data (see Pearl (1995), Dawid (2000), Lavielle and Moulines (2000) and Pearl (2009)). In a time series setting, compared to the intervention-based causality in Pearl (1995), Dahlhaus and Eichler (2003) employed a more straightforward idea that relies on the principle of temporal precedence that is readily available in time series. Through analysing partial correlations at different time lags, noncausality relations among the variables can be studied as well. This approach also allows for showing directional edges in graphical models without assuming a pre-specified variable ordering. To be specific, they provided a detailed discussion on the application of Granger causality in a more general framework of modern graph-based causal inference.

Let $X = \{X_t\}_{t\in\mathbb{Z}}$ with $X_t = (X_{1,t}, \ldots, X_{d,t})^\intercal$ be a d-variate stationary process. Its mean-square convergent autoregressive representation can be defined as

$$X_t = \sum_{u=1}^{\infty} \Phi(u) X_{t-u} + \epsilon_t$$

where $\Phi(u)$ is a square summable sequence of $d \times d$ matrices and the errors ϵ_t are *iid* random variables with mean 0 and non-singular covariance matrix Σ. To visualize the dependence structure of the *vector autoregressive* process, the simplest method is to construct a graph from the conditional distribution of X_t given its past values, where the vertices a_t and b_{t-k} can be used to represent variable $X_{a,t}$ and lagged instances $X_{b,t-k}$ respectively. To be specific, for $\Sigma_{ab} \neq 0$, an undirected (dashed) edge can be used to link two distinct vertices a_t and b_t. Meanwhile, vertices b_{t-k} will be connected to nodes a_t for $\Phi_{ab}(k) \neq 0$ whereas they will be removed when $\Phi_{ab}(k) = 0$ for all a.

In Chapter 5, our introduced graphical approach is purely exploratory, which may serve as the first step in lead-lag or causal analyses, and it is in the spirit of SiZer, which is discussed in the next section.

2.7 SiZer – A Visual Tool for Time Series

In data analysis, smoothing methods for curve estimation offer a valuable technique for obtaining the underlying patterns or structures within data.

The *SiZer* in Chaudhuri and Marron (1999), a shortening for *"SIgnificant ZERo crossings of derivatives"*, is a useful approach that can be utilised to capture the statistical significance of features at different locations and scales (bandwidths), such as peaks and valleys, in uni-variate linear data. The developed SiZer map provides an obvious identification for significant features with colour red or blue whereas the remaining grey areas indicate the data with sparse information.

To make SiZer useful for different scenarios, several adaptations have been developed in the existing statistical literature. For example, Godtliebsen et al. (2002) extended SiZer to a two-dimensional setting of image analysis; Li and Marron (2005) proposed the local likelihood SiZer map, which can offer more powerful inferences and is more efficient for discrete data; Rondonotti et al. (2007) extended SiZer to deal with time series with dependent noise; to deal with the quartile composition instead of mean structure of the data, Park et al. (2010) introduced the *quartile SiZer* constructed on a local linear quartile smoother; under the assumption of Gaussian distributed residuals, Skrøvseth et al. (2012) proposed a change-point detection technique by imposing causality on the scale-space viewpoint applied in SiZer-type plots; there are also many adaptations to linear variables, see e.g. Rudge (2008) and Rydén (2010).

For circular data, Oliveira et al. (2014) introduced the *CircSiZer*, a new extension of SiZer that is constructed with the kernel density estimator and regression estimator to assess the statistical significance of observed features. In order to take advantage of the wrapped Gaussian kernel which yields circular "causality", i.e. the number of modes should be non-increasing with increasing bandwidth, Huckemann et al. (2016) proposed the *Wrapped SiZer (WiZer)* and provided a numerical foundation for choosing the number of wrappings for statistical tests. Inspired by Godtliebsen et al. (2002), Vuollo and Holmström (2018) developed *SphereSiZer* for investigating structure in spherical data.

Chapter 3

Multi-scale estimation of long-run variance

3.1 Introduction

Considering the aforementioned importance of long-run variance in change-point estimation, this chapter concentrates on developing well-suited LRV estimators for serial correlated time series modelled with piecewise-constant signal. To clarify, a simplified version of model (2.2.1) is formulated through

$$f_t = \sum_{j=1}^{q+1} \theta_j \mathbb{1}_{(\eta_{j-1},\eta_j]}(t) \tag{3.1.1}$$

which is the unknown signal being partitioned into $q+1$ segments $(\eta_{j-1}, \eta_j]$ for $j = 1, \ldots, q+1$, where $0 = \eta_0 < \eta_1 < \ldots \eta_q < \eta_{q+1} = n$ are distinct change-point locations, and $\theta_1, \ldots, \theta_q \in \mathbb{R}$ are the function values of f, satisfying $\theta_j \neq \theta_{j+1}$ for $j = 1, \ldots, q$.

Meanwhile, the stationary error process $\{X_i\}_{i=1}^n$ satisfies the properties $\mathbb{E}(X_i) = 0$ and

$\mathrm{Var}(X_i) = \sigma^2 < \infty$. To avoid much loss of generality, we consider the stationary error process of the general form

$$X_i = g(\ldots, \epsilon_{i-1}, \epsilon_i) \tag{3.1.2}$$

where the inputs ϵ_i, $i \in \mathbb{Z}$, are independent and identically distributed (*iid*) random variables and $g(\cdot)$ is an input-output filter or transformation leading to all the dependencies within the outputs X_i. Since dependence is an intrinsic feature of a stochastic process, how to measure such dependence has been discussed in many existing works, see especially the *strong-mixing coefficients* first proposed in the influential paper (Rosenblatt, 1956). There are various types of strong mixing conditions, including $\alpha-$, $\beta-$ and $\phi-$ conditions, etc (see a broad review in Doukhan (2012)). Although they are widely applied and continuously improved after being developed, their application can still be challenging due to the difficulty in computation and verification under some cases.

To provide dependence measures under mild and easily provable conditions, Wu (2005) introduced a physical system that shows a direct relationship with data-generating mechanisms. This representation is generally enough to cover a huge class of stochastic processes, which subsumes many well-known time series models such as the ARMA, ARCH (Engle, 1982) and GARCH (Bollerslev, 1986) models. Due to its easily workable measures, we establish the asymptotic consistency of our new wavelet-based LRV estimators under the assumption that error process $\{X_i\}_{i=1}^n$ follows the physical system proposed in Wu (2005).

In the following discussion, we write $\|X\|_p = (\mathbb{E}|X_i|^p))^{1/p}$, $p \geqslant 1$, for a random variable X (in the case of its existence). For two real sequences of positive numbers $\{a_n\}$ and $\{b_n\}$, write $a_n \asymp b_n$ if $0 < \liminf_{n\to\infty} a_n/b_n \leqslant \limsup_{n\to\infty} a_n/b_n < \infty$. Let $\epsilon_j^{'}$ be an *iid* copy of ϵ_j and $X_i^* = g(\ldots, \epsilon_{-1}, \epsilon_0^{'}, \epsilon_1, \ldots, \epsilon_{i-1}, \epsilon_i)$. To show how alterations in the inputs result in corresponding changes in the outputs, we consider the *functional*

dependence measure

$$\delta_{i,p} \coloneqq \|X_i - X_i^*\|_p$$

under the condition $\|X_i\|_p < \infty$. This measure intuitively calculates the dependence of X_i on the single innovation ϵ_0 by replacing ϵ_0, with all other inputs ϵ_i remaining the same.

In practice, the dependence in time series can describe diverse patterns and characteristics. A sequence is often assumed to have *short-range dependence (SRD)* and hence the impacts of such dependence can be constrained with standard extreme value limits, i.e. events X_i and X_j can be considered independent if the time points i and j are sufficiently distant from each other. With the functional dependence measure $\delta_{i,p}$, the process $\{X_i\}_{i=1}^n$ can be regarded as short-range dependence if

$$\Delta_{m,p} \coloneqq \sum_{i=m}^{\infty} \delta_{i,p} < \infty, \ m \geqslant 0$$

where $\Delta_{m,p}$ quantifies the cumulative impact of ϵ_0 on X_i for $i \geqslant m$ given fixed m. For the process $X_\cdot = \{X_i\}_{i=-\infty}^{\infty}$ with slow decay of $\Delta_{m,p}$, the *dependence adjusted norm (DAN)* (Wu and Wu, 2016) is further applied to account for the serial correlation in this stronger dependence case, which is defined as follows

$$\|X_\cdot\|_{p,\upsilon} \coloneqq \sup_{m\geqslant 0}(m+1)^\upsilon \Delta_{m,p} = \sup_{m\geqslant 0}(m+1)^\upsilon \sum_{i=m}^{\infty} \delta_{i,p}, \ \upsilon \geqslant 0$$

In general, this chapter introduces several wavelet-based LRV estimators that can be well-suited for dependent data with signal (3.1.1) and the error process (3.1.2). Their asymptotic unbiasedness and consistency are proved under basic assumptions related to the above mentioned measures (see detailed assumptions in section 3.3). In particular, ARMA model, a commonly used time series model, also follows the above mentioned assumptions and the related dependence measures are presented in Remark 3.1. To

compare the effectiveness of several LRV estimators, we conduct complete simulation tests on the following ARMA model, which has an explicit expression for the true LRV.

Remark 3.1 *The ARMA model is a vital special class of linear process* $X_t = \sum_{i=0}^{\infty} a_i \epsilon_{t-i}$ *and takes the form* $X_t - \sum_{j=1}^{p} \phi_j X_{t-j} = \epsilon_t + \sum_{l=1}^{q} \theta_l \epsilon_{t-l}$, *where* ϕ_j *and* θ_l *are autoregressive and moving average parameters. The linear process carries the convenient results for functional dependence measure that* $\delta_{i,p} = |a_i| \|\epsilon_0 - \epsilon_0^{'}\|_p$, *where* $\|\epsilon_0 - \epsilon_0^{'}\|_p < \infty$, *see Example 1 in Wu (2011). When representing the ARMA model as a linear process, the corresponding* a_i *should be the coefficient of the series* $(1 + \sum_{l=1}^{q} \theta_l z^l)/(1 - \sum_{j=1}^{p} \phi_j z^j)$. *Let* $\lambda_1, \ldots, \lambda_p$ *denote the roots of the equation* $\lambda_p - \sum_{j=1}^{p} \varphi_j \lambda_{p-j} = 0$. *After adding the assumption* $\lambda_* = \max_{m \leqslant p} |\lambda_m| < 1$, *Example 2 in Wu (2011) shows the coefficients* $|a_i| = O(r^i)$ *hold for all* $r \in (\lambda_*, 1)$.

And hence we can prove that both assumptions (A2) and (A3) introduced in Section 3.3 are satisfied, i.e. $\Delta_{0,4} = \sum_{i=0}^{\infty} |a_i| \|\epsilon_0 - \epsilon_0^{'}\|_4 = O(1) \|\epsilon_0 - \epsilon_0^{'}\|_4 < \infty$, $\sum_{i=1}^{\infty} i\delta_{0,2} = \sum_{i=1}^{\infty} i|a_i| \|\epsilon_0 - \epsilon_0^{'}\|_2 \leqslant \check{C} r/(1-r)^2 \|\epsilon_0 - \epsilon_0^{'}\|_2 < \infty$ *and the DAN* $\|X_{\cdot}\|_{p,\upsilon} = \sup_{m \geqslant 0}\{(m+1)^{\upsilon} \sum_{i=m}^{\infty} |a_i|\} \|\epsilon_0 - \epsilon_0^{'}\|_p = \sup_{m \geqslant 0}\{(m+1)^{\upsilon} r^m/(1-r)\} \|\epsilon_0 - \epsilon_0^{'}\|_p < \infty$ *for all* $\upsilon \geqslant 0$, *where* $\check{C}$ *is a finite constant.*

Overall, this chapter is organised as follows. In the remainder of this section, we provide the underlying ideas for our methodology and review more related works. Section 3.2 describes our proposed estimators, while Section 3.3 presents their theoretical properties. Simulation results are shown in Section 3.4, 3.5 and Appendix 3.6, and the proofs of theorems are offered in Section 3.7.

3.1.1 Basic Ideas for Wavelet-Based Estimators

In general, in the presence of non-constant mean trend, there are two broad classes of LRV estimators: difference- and residual-based estimators. Difference-based techniques are built on the l-th differences $Y_i - Y_{i-l}$ of the observed time series whereas the residual-based ones attach great importance to residuals $\hat{X}_i = Y_i - \hat{f}_h(i)$, where $\hat{f}_h(\cdot)$ is a non-parametric estimator of the unknown signal derived with the bandwidth or smoothing parameter h. To be specific, before proposing the estimator for the whole LRV $\sigma_*^2 = \sum_{\tau\in\mathbb{Z}} \gamma(\tau)$, it is quite common to first consider the estimation of autocovariance $\gamma(\tau)$. The two classes of estimators usually take the following forms respectively

$$\hat{\gamma}^D(0) = \frac{1}{2(n-l)} \sum_{i=l+1}^{n} (Y_i - Y_{i-l})^2,\ \hat{\gamma}^D(h) = \hat{\gamma}^D(0) - \frac{1}{n-|m|} \sum_{i=|m|+1}^{n} (Y_i - Y_{i-|m|})^2$$

$$\hat{\gamma}^R(0) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}_h(i))^2,\ \hat{\gamma}^R(h) = \frac{1}{n}\sum_{i=1}^{n-m}(Y_i - \hat{f}_h(i))(Y_{i+m} - \hat{f}_h(i+m))$$

where $m = 1, \ldots, n-1$. The widely used moving average estimator is defined as $\hat{f}_h(i) = N_{i,q}^{-1} \sum_{|j-i|\leqslant q} Y_j$, where q is the chosen lag, $h = q/n$, and $N_{i,q}$ denotes the number of j satisfying $|j-i| \leqslant q$ (see Qiu et al. (2013) and Khismatullina and Vogt (2020)).

By applying a threshold to the wavelet coefficients in chosen scales, we develop, in what follows, several asymptotically unbiased and consistent wavelet-based estimators lying somewhere in between the two broad classes. From the perspective of difference-based measures, our idea is quite clear since Haar wavelet coefficients can be regarded as functions built on m_j-th differences $Y_i - Y_{i-m_j}$, i.e. $d_{j,k} = \frac{1}{\sqrt{2m_j}}(\sum_{i=(2k-1)m_j+1}^{2km_j} Y_i - \sum_{i=(2k-2)m_j+1}^{(2k-1)m_j} Y_i) = \frac{1}{\sqrt{2m_j}} \sum_{i=(2k-1)m_j+1}^{2km_j}(Y_i - Y_{i-m_j})$.

On the other hand, wavelets can be viewed as analogous to difference in local averages, which can provide difference-based estimation of the signal f_i, i.e. Applying DWT

$d_{j,k} = \mu_{j,k} + Z_{j,k}$, we have large $d_{j,k}$ stands for the estimated signal. Meanwhile, similar to the idea of separating the signal and noise in $\hat{\gamma}^R(h)$, utilising wavelet shrinkage (thresholding) can successfully eliminate the signals from the observations. Here we provide a brief explanation of this idea based on the Haar discrete wavelet transform. Figure 3.1 illustrates this point by presenting the reconstructed signal and error process obtained from *discrete inverse wavelet transform (IDWT)* after removing large coefficients, where the implementation of IDWT is provided in the publicly available **R** package **wavethresh** (Nason et al., 2022). This example considers data generated by adding an independent Gaussian process (a) $X_i^{(1)} = \epsilon_i$ and an AR(1) process (b) $X_i^{(2)} = 0.3X_{i-1}^{(2)} + \epsilon_i$, where $\epsilon_i \overset{iid}{\sim} N(0,1)$, to a piecewise-constant signal with $q = 7$ regular change-points at locations $i = 64, 128, 192, 256, 320, 384, 448$ ($n = 512$).

Intuitively, similar to the ideas applied in wavelet shrinkage, the wavelet-based estimators are constructed via transforming Equation (2.3.7) into wavelet coefficients such as $d_{j,k} = \mu_{j,k} + Z_{j,k}$, with the expectation that $\mu_{j,k}$ corresponding to smooth signals will be close to zero while those corresponding to irregular signals will significantly differ from zero. And we can choose a threshold to separate all information into two parts, i.e. signal (coefficients over threshold) and noise (coefficients below threshold). Figure 3.1 demonstrates that for the simple example, adopting a suitable threshold can successfully divide the original simulated data as we can see the reconstructed signals in black are close to the true signal in red.

Also, in order to capture the central tendency of time series, we consider building our estimators with two basic measures: mean and median, for simplicity (see examples in Dette et al. (2019) and Wu and Zhao (2007)).

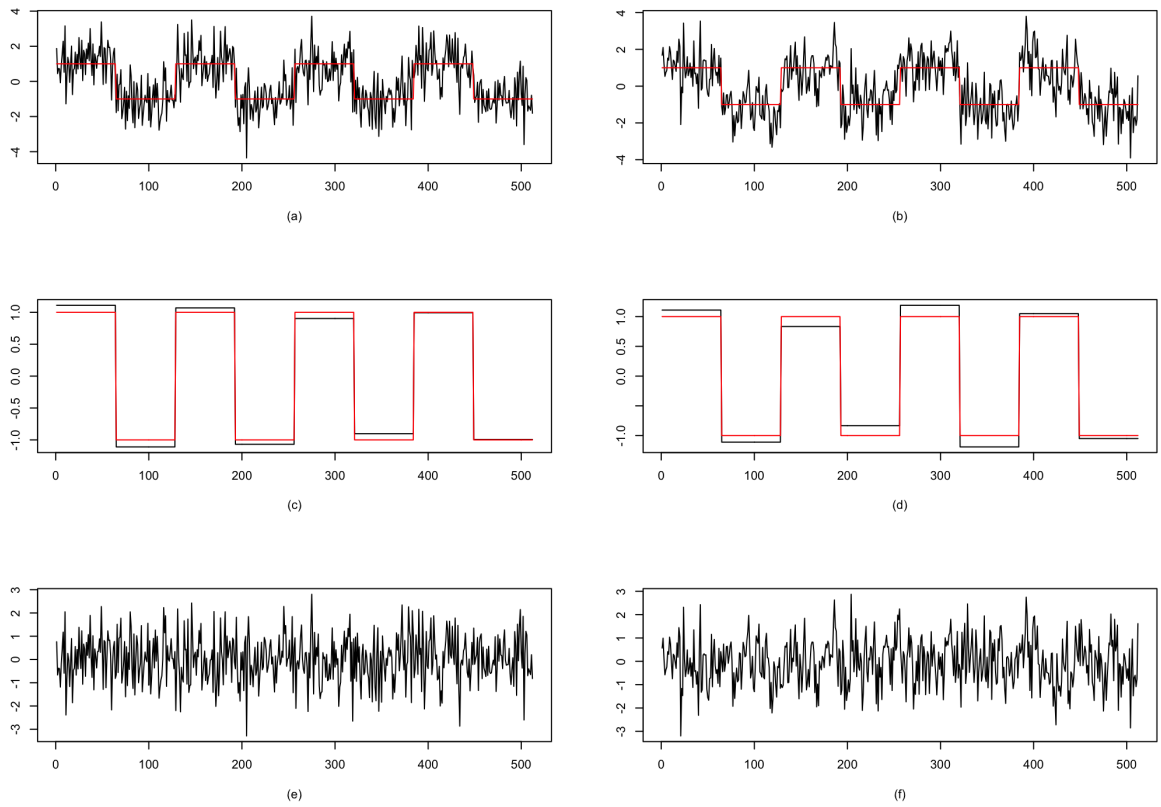


Figure 3.1: A simple example for the underlying idea of our wavelet-based LRV estimators. Y_t. Plots (a) and (b) display the simulated data generated by adding $\{X_i^{(1)}\}_{i=1}^n$ and $\{X_i^{(2)}\}_{i=1}^n$ to piecewise-constant signal with 7 regular change-points ($n = 512$). After applying discrete inverse wavelet transform, we have (c) and (d) show the obtained signal (black line) and true signal (red line) while (e) and (f) present obtained noise.

3.1.2 Related Work

In the following, to bypass the challenge of pre-estimating the signal with dominant fluctuation, we choose to employ difference-based estimators as the foundation for our subsequent construction. We now describe two asymptotically consistent estimators that serve as the motivation for our wavelet-based approaches. When estimating σ_*, Wu and Zhao (2007) first divide the time series into $m_n = \lfloor n/k_n \rfloor$ blocks with $\{Y_{1+ik_n}, \ldots, Y_{(i+1)k_n}\}$ for $i = 0, 1, \ldots, m_n - 1$ and use the difference of local averages

$A_i - A_{i-1}$, where

$$A_i = \frac{1}{k_n}\sum_{j=1}^{k_n} Y_{j+ik_n}$$

to eliminate the impact of signal and concurrently characterise the dependence structure of the sample. For dependent data with piecewise-constant mean, two estimates are formulated as follows

$$\begin{aligned}\hat{\sigma}_*^2 &= \frac{k_n}{2(m_n-1)}\sum_{i=1}^{m_n-1}|A_i - A_{i-1}|^2 \\ \hat{\sigma}_* &= \frac{\sqrt{k_n}}{\sqrt{2z_{1/4}}}\text{median}(|A_i - A_{i-1}|,\ 1 \leqslant i \leqslant m_n - 1)\end{aligned} \tag{3.1.3}$$

where the first estimate is developed closely related to the subseries variance estimate in Carlstein et al. (1986) and proved (the second one by conjecture) to have the same optimal *Mean Square Error (MSE)*, i.e. $\mathbb{E}[(\hat{\sigma}_*^2 - \sigma_*^2)^2] = O(n^{-2/3})$, when $k_n \asymp n^{1/3}$.

On the other hand, in the presence of serial dependence, McGonigle and Cho (2023) argued that it seems inappropriate to estimate the noise level with one single LRV estimator. Instead, to gauge the noise level within the data section of certain length, they introduced a robust estimator for the *scale-dependent time-average variance constant (TAVC)* that is defined as follows,

$$\hat{\sigma}_L^2 = \text{Var}\left(\frac{1}{\sqrt{L}}\sum_{t=1}^{L} X_t\right)$$

under the given scale L. To be specific, suppose that block size satisfies $G = L/2$ where L is an even number. For any $b \in \{0, 1, \ldots, G-1\}$ and corresponding number of blocks $N_1(b) = \lfloor (n - b - G)/G \rfloor$, local averages and their difference are similarly defined as

$$\bar{Y}_{j,b} = \frac{1}{G}\sum_{t=jG+b+1}^{(j+1)G+b} Y_t, \quad \text{and} \quad \xi_{j,b} = \frac{G(\bar{Y}_{j,b} - \bar{Y}_{j-1,b})^2}{2}$$

for $j = 1, 2, \ldots, N_1(b)$. McGonigle and Cho (2023) then adopt *M-estimators* to truncate observations with the solution

$$\hat{\theta} = \arg\min_{\theta} \left(\frac{1}{n} \sum_{i=1}^{n} \rho(Y_i, \theta) \right)$$

where the real-valued objective function of (Y_i, θ) is a sample average. *Least squares estimation (LSE)* and *maximum likelihood estimation (MLE)* are two well-studied special cases. In particular, the introduced robust estimator $\hat{\sigma}^2_{L,b}$ is provided as the solution to the following M-estimation formula proposed in Catoni (2012)

$$h_{L,b}(u) = \frac{1}{N_1(b)} \sum_{j=1}^{N_1(b)} \phi_v(\xi_{j,b} - u) = 0 \tag{3.1.4}$$

where $\phi_v(x) = v^{-1}\phi(vx)$ for some specific $v > 0$; and the non-decreasing influence function $\phi(x)$ is defined as

$$\phi(x) = \begin{cases} -\log(2) & x \leqslant -1, \\ \log(1 + x + x^2/2) & -1 < x \leqslant 0, \\ -\log(1 - x + x^2/2) & 0 < x \leqslant 1, \\ \log(2) & x > 1. \end{cases}$$

This scale-dependent TAVC estimator is also built relying on the physical system in Wu (2005) with assumptions similar to our models. And McGonigle and Cho (2023) shows that the estimator can successfully approximate the LRV when scale L is large enough. A maximum time-scale M is also pre-specified to balance the errors resulted from changes in time-scale L. Meanwhile, McGonigle and Cho (2023) also pointed out the possible failure of a global LRV resulted from its lack of adaptivity to the scale of local data sections when computing test statistics in multi-scale algorithms. In particular, taking MOSUM as an example, the application of a global LRV σ^2_* may

lead to a failure in true change-point detection when σ_*^2 is very small while spurious information may be detected when σ_*^2 is indeed a large value.

Unlike the scale-dependent TAVC, our estimation takes into account multiple scales of wavelets, making it more robust to one particular scale. Concurrently, maximal overlap wavelets are employed to better draw the serial correlated features from data. These two points, together with thresholding, make our estimator well-constructed and also robust to the presence of multiple mean shifts.

3.2 Estimation of the LRV

In this section, we shall describe our new estimators of the LRV σ_*^2 in (2.4.1) that are asymptotically unbiased and consistent, and are robust to the presence of multiple change-points in piecewise-constant signal. The detailed procedure of the development of these estimators will also be discussed in the first part of this section, from removing extreme scales to removing large coefficients.

This section generally contains four parts, in terms of different combinations of two measures of central tendency, i.e. mean and median, and two wavelet transform methods, i.e. DWT and MODWT.

1. Mean (Discrete Haar Wavelets)

Here we focus on and develop a two-step procedure for selecting coefficients as these factors play a central role in the estimation process. Specifically, after thresholding, the (scaled) squared coefficients $d_{j,k}^2$ or $2m_j\tilde{W}_{j,k}^2$ in suitable scales j can serve as asymptotically unbiased estimators of LRV σ_*^2, see Lemma 3.1 and Section 3.7 for further details.

In the first step, we decide to remove some of the finest and coarsest scales to realise the "concentration of power". To clarify, optimising the performance of wavelet thresholding, in reality, largely relies on selecting scales where the transformed data show a sparsity pattern in the wavelet domain. This means it would be better to focus on scales with only a few coefficients having notable magnitudes while most coefficients take on smaller values, which mainly reflect noise. In practice, it is often observed that the dominant Haar coefficients are mainly situated at scales that are coarser, though typically not at the coarsest scales.

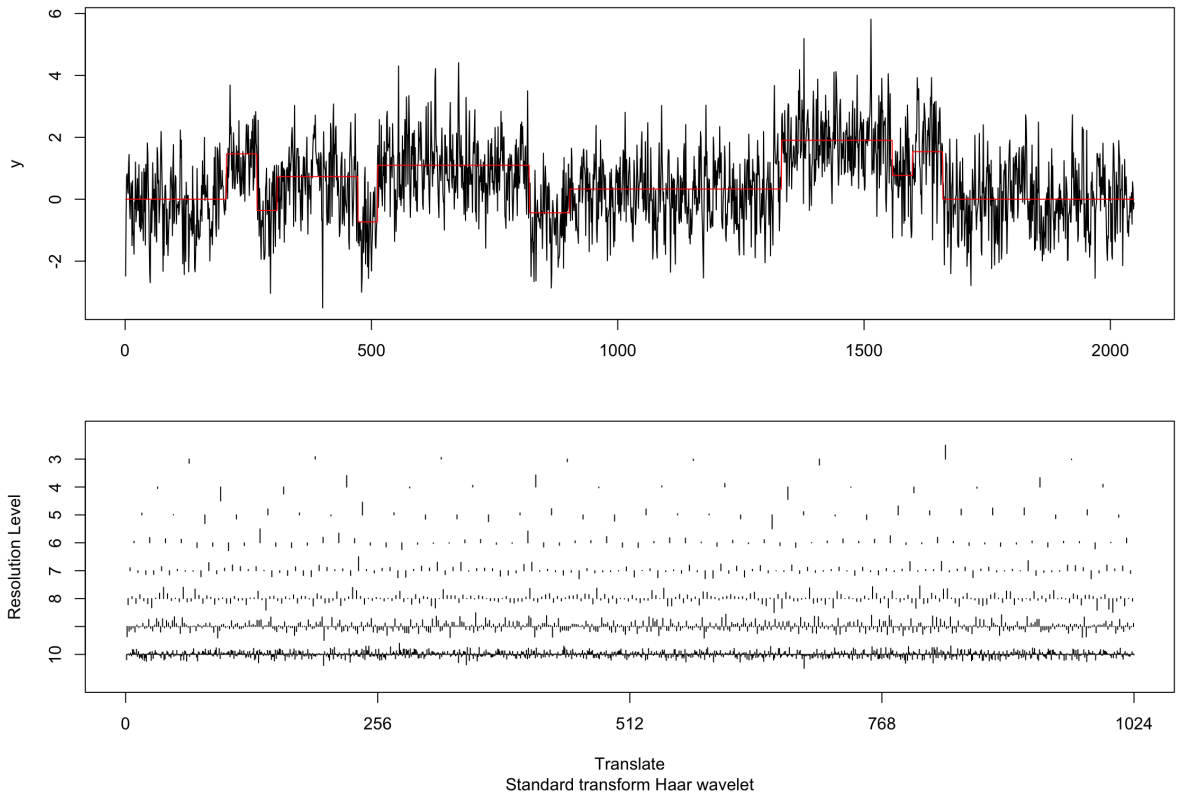


Figure 3.2: Scaled Haar DWT coefficients computed at scale 1-8 (multiresolution level 10-3) for the simulated data Y_t. The corresponding mean shifts are plotted with red line.

Using the implementation available from **R** package **wavethresh** (Nason et al., 2022), we study the scaled Haar coefficients computed at different scales, where the largest resolution level indicates the finest scale $j = 1$, and second largest resolution level is the

second finest scale $j = 2$, etc. As a simple example, Figure 3.2 plots the corresponding coefficients for simulated data with noise following an AR(1) process with resolution levels 3-10, i.e. scales 8 to 1. It indicates that the coarser scales, especially $j = 4, 5$, offer relatively more significant features in signals without completely excluding all information about noises. For the finest scales, such as $j = 1, 2$, we can see the magnitude of coefficients is largely impacted by the level of noise and hence it is hard to find any patterns related to the true signal. On the other hand, the coarsest scales, such as $j = 7, 8, 9, 10$, only contain a few elements, which carry less information and cannot even show us half of the features in the signal.

Therefore, when deciding the choice of scales, we exclude the coarsest scales because they contain too few coefficients to fully capture all features in mean shift signals, especially when there are closely located change-points. For example, in Figure 3.2, it is evident that at the coarsest scales, the change-points around $t = 500$ are not effectively reflected in the corresponding coefficient. Simultaneously, the finest scales are also eliminated due to the fact that although these scales mainly reflect noise, some signal still remains but we cannot find a clear separation between noise and signal. This goes against the target of only retaining the structure of the error process.

After removing the extreme scales, we introduce a wavelet-based estimator with coefficients $d_{j,k}$ (2.3.5), which is represented as follows

$$\hat{\sigma}_1^2 = \frac{1}{b-a+1} \sum_{j=a}^{b} \frac{1}{K_j} \sum_{k=1}^{K_j} d_{j,k}^2 \tag{3.2.1}$$

This basic estimator is somewhat similar to the first difference-based estimator introduced in (3.1.3), where the wavelet coefficients $d_{j,k}$'s are applied as a substitution for the difference in local averages $A_i - A_{i-1}$.

In the second step, with the aim of enhancing the sparsity of the representation, we

will eliminate scales at the extremes and further selectively reduce certain large coefficients within the remaining scales, in the hope that the chosen threshold λ can remove all coefficients corresponding to locations with discontinuities or other irregularities. Mathematically, let N_j^1 denote the set of all those indices $k = 1, 2, \ldots, K_j$ for which $|d_{j,k}| \leqslant \lambda$ in scale j, and denote $K_j^* = |N_j^1|$. We have the general formula

$$\hat{\sigma}_1^2(\lambda) = \frac{1}{b-a+1} \sum_{j=a}^{b} \frac{1}{K_j^*} \sum_{k=1}^{K_j} d_{j,k}^2 \mathbb{1}_{(|d_{j,k}| \leqslant \lambda)} \tag{3.2.2}$$

where $\mathbb{1}_{(\cdot)}$ represents the indicator function and λ is the pre-specified threshold for all scales. From the formula, it becomes evident that as the parameter λ grows increasingly larger, the estimator approaches closer and closer to $\hat{\sigma}_1^2$, i.e. $\hat{\sigma}_1^2(\infty) = \hat{\sigma}_1^2$.

2. Mean (Maximal Overlap Haar Wavelets)

Compared to DWT, Figure 3.3 demonstrates that MODWT coefficients can extract more information from the same dataset, especially at coarser scales. In order to remove the dyadic restriction and better retain the dependence structure of the error process, we adopt the MODWT instead of the DWT (see section 2.3.1). Similarly, with $a, b \in \{1, 2, \ldots, J_0\}$, where $a \leqslant b$ and $J_0 = \lfloor \log(n) \rfloor$, we have the estimators relying on the definition of $\tilde{W}_{j,k}^2$ in (2.3.6) that

$$\begin{aligned} \hat{\sigma}_2^2 &= \frac{1}{b-a+1} \sum_{j=a}^{b} \frac{2m_j}{n-2m_j+1} \sum_{k=1}^{n-2m_j+1} \tilde{W}_{j,k}^2 \\ \hat{\sigma}_2^2(\lambda) &= \frac{1}{b-a+1} \sum_{j=a}^{b} \frac{2m_j}{T_j^*} \sum_{k=1}^{n-2m_j+1} \tilde{W}_{j,k}^2 \mathbb{1}_{(|\sqrt{2m_j}\tilde{W}_{j,k}| \leqslant \lambda)} \end{aligned} \tag{3.2.3}$$

where $T_j^* = |N_j^2|$ and N_j^2 denotes the set of all those indices $k = 1, 2, \ldots, n - 2m_j + 1$ for which $|\sqrt{2m_j}\tilde{W}_{j,k}| \leqslant \lambda$.

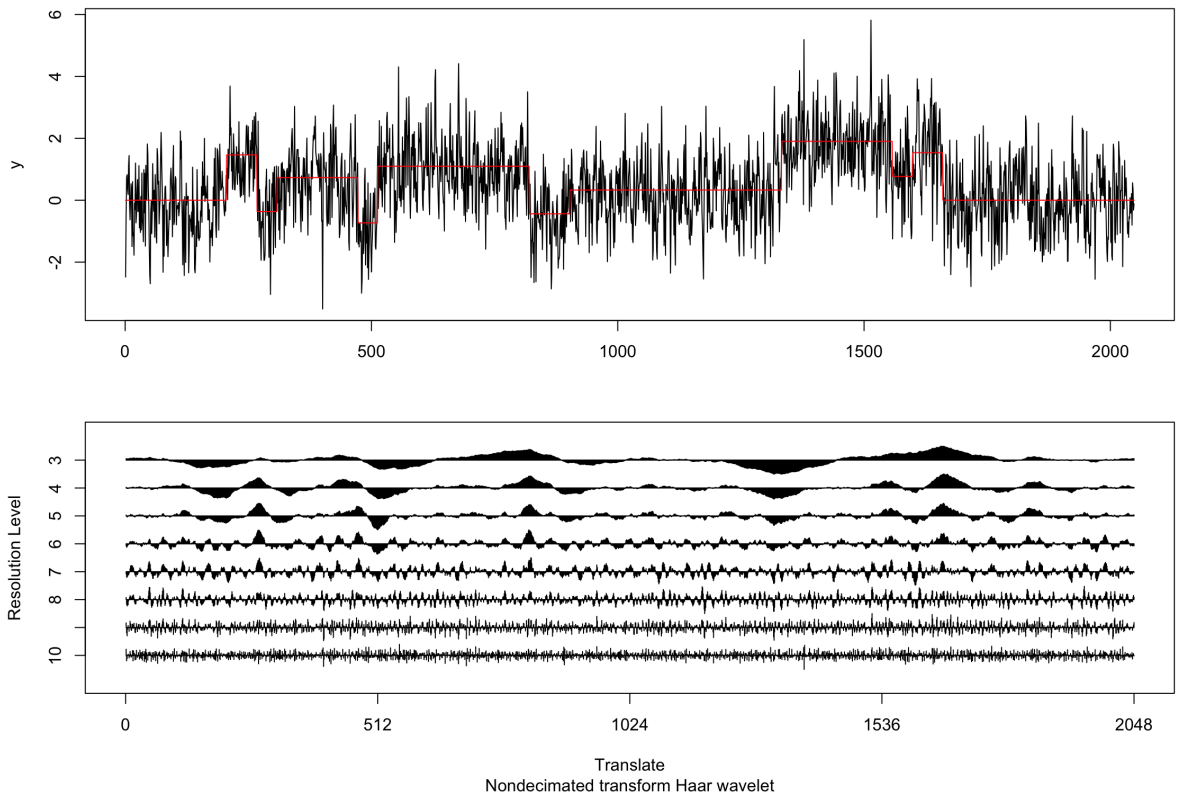


Figure 3.3: Scaled Haar MODWT coefficients computed at scale 1-8 (multiresolution level 10-3) for the simulated data Y_t. The corresponding mean shifts are plotted with red line.

3. Median (Discrete Haar Wavelets)

When analysing the coefficients at each scale, taking the mean usually gives us an overall picture of all values whereas taking the median can reduce the impact of extreme values and hence give us a more robust representation of these values. Similar to σ_1^2 and $\sigma_1^2(\lambda)$, the median-based estimator is also proposed following a two-step procedure. First, after removing extreme fine and coarse scales, one simple estimator of long-run standard deviation is given as

$$\hat{\sigma}_3 = \frac{1}{b-a}\sum_{j=a}^{b}\frac{1}{z_{1/4}}\text{median}(|d_{j,k}|,\ 1 \leqslant k \leqslant K_j) \tag{3.2.4}$$

which can be close to the second estimator in (3.1.3) if we replace the difference of local

averages $A_i - A_{i-1}$ with wavelet coefficients $d_{j,k}$. Moreover, although $\hat{\sigma}_3$ is comparatively robust to the absolutely large coefficients, we still consider removing information related to signals to test whether this estimator can be further improved. After removing both extreme scales and large coefficients, the formula (3.2.4) can be extended to

$$\hat{\sigma}_3(\lambda) = \frac{1}{b-a}\sum_{j=a}^{b}\frac{1}{z_{1/4}}\text{median}(|d_{j,k}|\mathbb{1}_{(|d_{j,k}|\leqslant\lambda)},\ 1 \leqslant k \leqslant K_j) \tag{3.2.5}$$

where $z_{1/4}$ is the third quartile of the standard normal distribution. Compared to the mean-based estimators $\hat{\sigma}_1$ and $\hat{\sigma}_1(\lambda)$, the performance of median-based estimators may not be heavily impacted when the pre-specified threshold is not good enough.

4. Median (Maximal Overlap Haar Wavelets)

Similarly, after extending (3.2.4) and (3.2.5) to the maximal overlap case, we have

$$\begin{aligned}
\hat{\sigma}_4 &= \frac{1}{b-a}\sum_{j=a}^{b}\frac{\sqrt{2m_j}}{z_{1/4}}\text{median}(|\tilde{W}_{j,k}|,\ 1 \leqslant k \leqslant n-2m_j+1) \\
\hat{\sigma}_4(\lambda) &= \frac{1}{b-a}\sum_{j=a}^{b}\frac{\sqrt{2m_j}}{z_{1/4}}\text{median}(|\tilde{W}_{j,k}|\mathbb{1}_{(|\sqrt{2m_j}\tilde{W}_{j,k}|\leqslant\lambda)},\ 1 \leqslant k \leqslant n-2m_j+1)
\end{aligned} \tag{3.2.6}$$

based on the definition (2.3.6).

Here scales a and b grow as the sample size increases to ensure the appropriate asymptotic behaviour. For details see Theorem 3.1 and Theorem 3.2 below, where we show the asymptotic consistency of our wavelet-based estimators. Meanwhile, guidance on the choice of threshold λ is given in Section 3.4.1.

3.3 Asymptotic Properties of Estimators

For the statement of the asymptotic properties in this section we shall make the following basic assumptions on the error process $\{X_i\}_{i=1}^n$.

(A1) $\|X_i\|_4 < \infty$

(A2) $\Delta_{0,4} < \infty$ and $\sum_{i=1}^{\infty} i\delta_{i,2} < \infty$

(A3) $\|X_\cdot\|_{p,\upsilon} < \infty$, where $p > 2$ and $\upsilon > 0$

Under Assumption (A1), the condition $\Delta_{0,4} < \infty$ implies that the autocovariance functions are absolutely summable, i.e. $\sum_{\tau\in\mathbb{Z}} |\gamma(\tau)| < \infty$, ensuring the existence of $\sum_{\tau\in\mathbb{Z}} \gamma(\tau)$. Hence, Assumption (A2) indicates the short-range dependence of the error process $\{X_i\}_{i=1}^n$ and is made to establish the asymptotic properties of differences in local sums, see Lemma 3.1. Assumption (A3) permits a sharp and easy-to-use bound of the wavelet coefficients, see Lemma 3.2 and Lemma 3.3. See Remark 3.1 for a particular example satisfying these assumptions.

In contrast to the common short-range dependent time series, the notion of *long-range dependence (LRD)* is basically associated with data whose ACVFs $\gamma(\tau)$ decays like τ^{2d-1} as $\tau \to \infty$, where $0 < d < 1/2$. Such fractional differencing parameter d guarantees that the corresponding ACVF is not absolutely summable. The study of LRD began with the work of Hurst (1951), Mandelbrot and Van Ness (1968) and Mandelbrot and Wallis (1968), and long memory processes have attracted substantial interest in the field of hydrology, geophysics, finance and economics (Gerstenberger, 2021), see Doukhan et al. (2002) for more literature on applications. Compared to LRV of SRD $\lim_{n\to\infty} \mathrm{Var}(n^{1/2}\bar{X}_{1:n}) = \lim_{n\to\infty} \mathrm{Var}(n^{-1/2}\sum_{i=1}^n X_i)$, Pipiras and Taqqu (2017) suggests that the sample mean $\bar{X}_{1:n}$ should be normalised by $n^{d-1/2}$ instead to achieve the LRV of long-range dependence, i.e. $\lim_{n\to\infty} \mathrm{Var}(n^{1/2-d}\bar{X}_{1:n}) \to \frac{c_2}{d(2d+1)}$, where c_2 is a constant for conditions $\gamma(\tau) = L_2(\tau)\tau^{2d-1}$ and $L_2(u) \sim c_2$. Abadir et al. (2009) discussed two possible LRV estimators to deal with long memory time series. Although

our LRV estimators are developed for SRD time series, they can be applied in testing procedures (such as calculating CUSUM statistics) for distinguishing between a SRD time series with abrupt mean shifts and a stationary LRD time series. For example, Berkes et al. (2006) proposed the CUSUM-type testing procedure by supposing SRD under the null hypothesis and LRD under the alternative; Gerstenberger (2021) introduced a similar Wilcoxon-type testing procedure that can outperforms the CUSUM-type one in the presence of outliers. Also, considering time series non-stationarity and long-range dependence, Bai and Wu (2024) developed a difference-based long-run covariance matrix estimator for time series observations following functional linear models with time-varying regression coefficients.

The following Lemmas present asymptotic results for the proof of consistency of our LRV estimators, see Theorem 3.1 and Theorem 3.2. To clarify, lemmas 3.2 and 3.3 provide theoretical justifications for $\hat{\sigma}_1^2(\lambda)$ and $\hat{\sigma}_2^2(\lambda)$ respectively.

Lemma 3.1 *Denote* $S_n = \sum_{i=1}^n X_i$ *and* $\sigma_*^2 = \sum_{i\in\mathbb{Z}} \mathrm{Cov}(X_0, X_i)$. *Under the assumption (A2), we have* $\|S_{2n} - 2S_n\|^2 = 2n\sigma_*^2 + O(1)$ *(see proof of Lemma 4 in Wu and Zhao (2007)). Also, let* $k = m + 2l$, *for* $l = 1, 2, \ldots, l'$, *we have*

$$\sum_{l=1}^{l'} \mathrm{Cov}(S_{(k+1)n} - 2S_{kn} + S_{(k-1)n},\ S_{(m+1)n} - 2S_{mn} + S_{(m-1)n}) < \infty \tag{3.3.1}$$

and the same result can be derived when we let $k = m - 2l$, *for* $l = 1, 2, \ldots, l'$.

Proof: Denote $\gamma(n) = \mathrm{Cov}(Y_0, Y_n) = \mathrm{Cov}(X_0, X_n)$ and $S_n = \sum_{i=1}^n X_i$. We have that when $|k - m| \geqslant 2$

$$\begin{aligned}
&\mathrm{Cov}(S_{(k+1)n} - 2S_{kn} + S_{(k-1)n},\ S_{(m+1)n} - 2S_{mn} + S_{(m-1)n}) \\
&= \mathrm{Cov}\left(\sum_{i=kn+1}^{(k+1)n} X_i - \sum_{i=(k-1)n+1}^{kn} X_i,\ \sum_{i=mn+1}^{(m+1)n} X_i - \sum_{i=(m-1)n+1}^{mn} X_i \right)
\end{aligned}$$

$$\begin{aligned}
&\leqslant \left|\mathrm{Cov}\left(\sum_{i=kn+1}^{(k+1)n} X_i, \sum_{i=mn+1}^{(m+1)n} X_i\right)\right| + \left|\mathrm{Cov}\left(\sum_{i=kn+1}^{(k+1)n} X_i, \sum_{i=(m-1)n+1}^{mn} X_i\right)\right| \\
&\quad + \left|\mathrm{Cov}\left(\sum_{i=(k-1)n+1}^{kn} X_i, \sum_{i=mn+1}^{(m+1)n} X_i\right)\right| + \left|\mathrm{Cov}\left(\sum_{i=(k-1)n+1}^{kn} X_i, \sum_{i=(m-1)n+1}^{mn} X_i\right)\right| \\
&\leqslant \sum_{h=1}^{2n} h|\gamma(n|k-m|-2n+h)| + \sum_{h=1}^{2n}(2n-h)|\gamma(n|k-m|+h)|
\end{aligned} \tag{3.3.2}$$

Let $k = m + 2l$, for $l = 1, 2, \ldots, l^{'}$. A straightforward calculation shows that

$$\begin{aligned}
&\sum_{l=1}^{l^{'}} \mathrm{Cov}(S_{(k+1)n} - 2S_{kn} + S_{(k-1)n},\ S_{(m+1)n} - 2S_{mn} + S_{(m-1)n}) \\
&\leqslant \sum_{h=1}^{2n} h|\gamma(h)| + 2n \sum_{h=1}^{(2l^{'}-2)n} |\gamma(2n+h)| + \sum_{h=1}^{2n}(2n-h)|\gamma(2l^{'}n+h)| \leqslant \sum_{h=1}^{\infty} h|\gamma(h)| < \infty
\end{aligned} \tag{3.3.3}$$

And the same result can be derived when let $k = m - 2l$, for $l = 1, 2, \ldots, l^{'}$.

Lemma 3.2 *Let $m_j = 2^{j-1}$ and $S_j^1 = \{1 \leqslant k \leqslant n/(2m_j) :\ d_{j,k}$ is such that $(2k-2)m_j + 1 < \eta_i < 2km_j$ for some $i = 1, 2, \ldots, q\}$ and $S_j^0 = \{1, 2, \ldots, n/(2m_j)\} \setminus S_j^1$. If assumption (A3) is satisfied, for any $j = 1, 2, \ldots, J$, $k \in S_j^0$, as $n \to \infty$*

$$P\left\{\frac{1}{\sqrt{2m_j}}\left|\sum_{i=(2k-1)m_j+1}^{2km_j} Y_i - \sum_{i=(2k-2)m_j+1}^{(2k-1)m_j} Y_i\right| > C\sqrt{\log(n)}\right\} \to 0 \tag{3.3.4}$$

i.e. $P\left\{|d_{j,k}| > C\sqrt{\log(n)}\right\} \to 0$ for some C, and we have $K_j/K_j^* - 1 = o_p(n^{-1+\nu_0})$ when $\nu_0 > \nu$ (detailed conditions of these parameters can be found in proof), for any $j = \nu \log_2(n)$.

Proof: We apply Theorem 2 in Wu and Wu (2016) and it follows that under the assumption (A3), (i) if $\upsilon > 1/2 - 1/p$, there exist some constants C_1, C_2 and C_3 only

depending on p and υ such that, for all $\sqrt{a}\lambda > 0$,

$$P(|S_a| > \sqrt{a}\lambda) \leqslant C_1 \frac{a\|X_\cdot\|_{p,\upsilon}^p}{(\sqrt{a}\lambda)^p} + C_2 \exp\left(-\frac{C_3 a\lambda^2}{a\|X_\cdot\|_{2,\upsilon}^2}\right) \tag{3.3.5}$$

(ii) If $\upsilon < 1/2 - 1/p$, for all $\sqrt{a}\lambda > 0$, there exists the inequality

$$P(|S_a| > \sqrt{a}\lambda) \leqslant C_1 \frac{a^{p/2-\upsilon p}\|X_\cdot\|_{p,\upsilon}^p}{(\sqrt{a}\lambda)^p} + C_2 \exp\left(-\frac{C_3 a\lambda^2}{a\|X_\cdot\|_{2,\upsilon}^2}\right). \tag{3.3.6}$$

After setting $a = n^\nu$, $0 < \nu < 1$ and $\lambda = C\sqrt{\log(n)}$ for certain C, we have

$$\begin{aligned} C_1 \frac{a\|X_\cdot\|_{p,\upsilon}^p}{(\sqrt{a}\lambda)^p} + C_2 \exp\left(-\frac{C_3 a\lambda^2}{a\|X_\cdot\|_{2,\upsilon}^2}\right) &= O(n^{\nu(1-p/2)}\log(n)^{-p/2}) + O(\exp(-C_4\log(n))) \\ &= O(n^{\nu(1-p/2)}\log(n)^{-p/2} + n^{-C_4}) \end{aligned} \tag{3.3.7}$$

where C_4 is a positive constant that depends on p, υ, C and the dependence condition $\|X_\cdot\|_{2,\upsilon}$. In addition,

$$C_1 \frac{a^{p/2-\upsilon p}\|X_\cdot\|_{p,\upsilon}^p}{(\sqrt{a}\lambda)^p} + C_2 \exp\left(-\frac{C_3 a\lambda^2}{a\|X_\cdot\|_{2,\upsilon}^2}\right) = O(n^{-\upsilon\nu p}\log(n)^{-p/2} + n^{-C_4}). \tag{3.3.8}$$

For $2m_j = n^\nu$, we can know that $P(|d_{j,k}| > \lambda)$ should have the same order as (3.3.7) or (3.3.8) for all $k \in S_j^0$. Given the fact that the signal f_t is piecewise constant with finite change-points, the set S_j^1 contains a finite number of elements, independently of $n \in \mathbb{N}$ and $|S_j^0| = O(n^{1-\nu})$. Then, applying a Bonferroni correction, we have that (i) when $\nu > \min(2/p, 1 - C_4)$,

$$\begin{aligned} P\left(K_j^* < |S_j^0|\right) = P\left\{\exists k \in S_j^0,\ |d_{j,k}| > \lambda\right\} &\leqslant \sum_{k\in S_j^0} P\left(|d_{j,k}| > \lambda\right) \\ &\leqslant |S_j^0| P\left(|d_{j,k}| > \lambda\right) \\ &= O(n^{1-\nu p/2}\log(n)^{-p/2} + n^{1-\nu-C4}) \\ &\to 0 \quad \text{as } n \to \infty \end{aligned} \tag{3.3.9}$$

(ii) Or when $\nu > \min(1/(1+\upsilon p), 1-C_4)$,

$$P\left(K_j^* < |S_j^0|\right) = O(n^{1-\nu-\upsilon\nu p}\log(n)^{-p/2} + n^{1-\nu-C4}) \to 0 \ \text{ as } n \to \infty \tag{3.3.10}$$

As $K_j = n^{1-\nu}$, for any $\delta_n > 0$, we can always find a value j' leading to $\delta_n K_j \geqslant |S_j^1|$ when $j > j'$ and $\delta_n = O(n^{-1+\nu_0})$, where $\nu_0 > \nu$. It hence gives

$$\begin{aligned}\lim_{n\to\infty} P\left(1 - \frac{K_j^*}{K_j} > \delta_n\right) &= \lim_{n\to\infty} P(K_j - K_j^* > \delta_n K_j) \\ &= \lim_{n\to\infty} P(K_j^* < (1-\delta_n)K_j) \\ &\leqslant \lim_{n\to\infty} P(K_j^* < |S_j^0|) = 0\end{aligned} \tag{3.3.11}$$

And we can also obtain that for any $\delta_n^{'} \geqslant \delta_n/(1-\delta_n)$,

$$\lim_{n\to\infty} P\left(\frac{K_j}{K_j^*} - 1 > \delta_n^{'}\right) = \lim_{n\to\infty} P\left(\frac{1 - K_j^*/K_j}{K_j^*/K_j} > \delta_n^{'}\right) = 0 \tag{3.3.12}$$

And hence we have $K_j/K_j^* - 1 = o_p(n^{-1+\nu_0})$ when $\nu_0 > \nu > \min(2/p, 1-C_4)$ if $\upsilon > 1/2 - 1/p$ (or $\nu_0 > \nu > \min(1/(1+\upsilon p), 1-C_4)$ if $\upsilon < 1/2 - 1/p$).

Lemma 3.3 *Let* $\tilde{S}_j^1 = \{1 \leqslant k \leqslant n - 2m_j + 1 :\ \tilde{W}_{j,k}$ *is such that* $k < \eta_i < k + 2m_j - 1$ *for some* $i = 1, 2, \ldots, N\}$ *and* $\tilde{S}_j^0 = \{1, 2, \ldots, n - 2m_j + 1\} \setminus \tilde{S}_j^1$. *Under assumption (A3), for any* $j = 1, 2, \ldots, J$, $k \in \tilde{S}_j^0$, *as* $n \to \infty$,

$$P\left\{\frac{1}{\sqrt{2m_j}}\left|\sum_{i=k+m_j}^{k+2m_j-1} Y_i - \sum_{i=k}^{k+m_j-1} Y_i\right| > \breve{C}\sqrt{\log(n)}\right\} \to 0 \tag{3.3.13}$$

i.e. $P\left\{\sqrt{2m_j}|\tilde{W}_{j,k}| > \breve{C}\sqrt{\log(n)}\right\} \to 0$ *for some* C, *and for any* $j = \nu\log_2(n)$, *we have* $(n - 2m_j + 1)/T_j^* - 1 = o_p(n^{-1+\nu_0})$ *when* $\nu_0 > \nu$ *(detailed conditions of these parameters can be found in the proof).*

Proof: Let $\lambda = C\sqrt{(1+\kappa)\log(n)}$ given some κ. From inequalities (3.3.5) and (3.3.6) in Lemma 2, we can know that $P(|\sqrt{2m_j}\tilde{W}_{j,k}| > \lambda) = O(n^{\nu(1-p/2)}\log(n)^{-p/2} + n^{-C_4(1+\kappa)})$ (or $O(n^{-\upsilon\nu p}\log(n)^{-p/2} + n^{-C_4(1+\kappa)})$), where C_4 is a positive constant depending on p, υ, C and the dependence condition $\|X_{\cdot}\|_{2,\upsilon}$, for all $k \in \tilde{S}_j^0$. Similarly, as $|\tilde{S}_j^0| = O(n)$, applying a Bonferroni correction, we have (i) when $\kappa > 1/C_4 - 1$ and $\nu \geqslant 2/(p-2)$,

$$\begin{aligned} P\left(T_j^* < |\tilde{S}_j^0|\right) = P\left\{\exists k \in \tilde{S}_j^0,\ |\sqrt{2m_j}\tilde{W}_{j,k}| > \lambda\right\} &\leqslant \sum_{k\in\tilde{S}_j^0} P\left(|\sqrt{2m_j}\tilde{W}_{j,k}| > \lambda\right) \\ &\leqslant |\tilde{S}_j^0| P\left(|\sqrt{2m_j}\tilde{W}_{j,k}| > \lambda\right) \\ &= O(n^{1+\nu-\nu p/2}\log(n)^{-p/2} + n^{1-C4(1+\kappa)}) \\ &\to 0 \quad \text{as } n \to \infty \end{aligned} \tag{3.3.14}$$

And (ii) when $\kappa > 1/C_4 - 1$ and $\nu \geqslant 1/(\upsilon p)$,

$$\begin{aligned} P\left(T_j^* < |\tilde{S}_j^0|\right) &= O(n^{1-\upsilon\nu p}\log(n)^{-p/2} + n^{1-C4(1+\kappa)}) \\ &\to 0 \quad \text{as } n \to \infty \end{aligned} \tag{3.3.15}$$

Let $T_j = n - 2m_j + 1$. Also, since $|\tilde{S}_j^1| = O(n^\nu)$, for any $\delta_n > 0$, we can always find a value j' leading to $\delta_n T_j > |\tilde{S}_j^1|$ when $j > j'$ and $\delta_n = O(n^{-1+\nu_0})$, where $\nu_0 > \nu$. It hence gives

$$\begin{aligned} \lim_{n\to\infty} P\left(1 - \frac{T_j^*}{T_j} > \delta_n\right) &= \lim_{n\to\infty} P(T_j - T_j^* > \delta_n T_j) \\ &= \lim_{n\to\infty} P(T_j^* < (1-\delta_n)T_j) \\ &\leqslant \lim_{n\to\infty} P(T_j^* < |\tilde{S}_j^0|) = 0 \end{aligned} \tag{3.3.16}$$

And we can also obtain that for any $\delta_n^{'} \geqslant \delta_n/(1-\delta_n)$,

$$\lim_{n\to\infty} P\left(\frac{T_j}{T_j^*} - 1 > \delta_n^{'}\right) = \lim_{n\to\infty} P\left(\frac{1 - T_j^*/T_j}{T_j^*/T_j} > \delta_n^{'}\right) = 0 \tag{3.3.17}$$

And hence we have $(n - 2m_j + 1)/T_j^* - 1 = o_p(n^{-1+\nu_0})$ when $\kappa > 1/C_4 - 1$ and

$\nu_0 > \nu \geqslant 2/(p-2)$ if $\upsilon > 1/2 - 1/p$ (or $\nu_0 > \nu \geqslant 1/(\upsilon p)$ if $\upsilon < 1/2 - 1/p$).

Throughout the paper, whenever we refer to the estimators $\hat{\sigma}^2(\lambda)$ as "consistent", we mean it in the sense that $(\hat{\sigma}^2(\lambda) - \sigma_*^2)^2 \to 0$ or simply $|\hat{\sigma}^2(\lambda) - \sigma_*^2| \to 0$, on a set with probability approaching 1 with sample size n. Here, we shall provide the asymptotic properties and theoretically compare the estimators. More information for constant C is given in lemmas 3.2 and 3.3, and that for constants $\tilde{a}$ and ν_0 is shown in the proof of Theorem 3.1 and Theorem 3.2, see Section 3.7.

For our mean-based asymptotically consistent estimators $\hat{\sigma}_1^2$ and $\hat{\sigma}_1^2(\lambda)$, we introduce a set $\mathcal{A}_n$ (3.3.18). Given the inequality in Lemma 3.2 such that $P\{|d_{j,k}| > C\sqrt{\log(n)}\} \to 0$ for some constant C when $k \in S_j^0$ as $n \to \infty$, we have $P(\mathcal{A}_n) \to 0$ for $\lambda = C\sqrt{2\log(n)}$ as well, which arises from the fact that $d_{j,k} = Z_{j,k}$ for any $k \in S_j^0$, $j = 1, \ldots, J$. Heuristically speaking, on the set $\mathcal{A}_n$, the estimator $\hat{\sigma}_1^2(\lambda)$ is well-behaved in the sense that the wavelet coefficients with or without the signals can be successfully separated with threshold λ.

Theorem 3.1 *Let Y_i follows the model (2.2.1) with piecewise-constant signal in (3.1.1) plus the stationary error process X_i in (3.1.2) satisfying Assumptions (A1)-(A3). Let a and b denote the minimum and maximum value of scales chosen for estimators respectively. Write $a = \alpha \log_2(n)$ and $b = \beta \log_2(n)$. Let the threshold parameter satisfy $\lambda = C\sqrt{2\log(n)}$ for certain constant C. On the set $\mathcal{A}_n$, defined by*

$$\mathcal{A}_n = \left\{ \forall\ 1 \leqslant k \leqslant n(2m_j)^{-1} \quad \frac{1}{\sqrt{2m_j}} \left| \sum_{i=(2k-1)m_j+1}^{2km_j} X_i - \sum_{i=(2k-2)m_j+1}^{(2k-1)m_j} X_i \right| \leqslant \lambda \right\} \tag{3.3.18}$$

which contains all $|Z_{j,k}| \leqslant \lambda$ at $k = 1, \ldots, n(2m_j)^{-1}$. When it satisfies $P(\mathcal{A}_n) \to 1$, we

have for some $0 < \alpha \leqslant \beta < 1$,

$$\begin{aligned} |\hat{\sigma}_1^2(\lambda) - \sigma_*^2| &= O_p\{(n^{-1+\beta+\tilde{a}} + n^{-\alpha}) \log_2(n)^{-1}\} \\ |\hat{\sigma}_1^2 - \sigma_*^2| &= O_p\{(n^{-1+\beta+\tilde{a}} + n^{-\alpha} + n^{-1+2\beta}) \log_2(n)^{-1}\} \end{aligned} \tag{3.3.19}$$

where $\alpha \geqslant 1/(2+2\upsilon)$ *and* $\tilde{a} > (1-\beta)/2$ *(*$\tilde{a} > (1-\beta)(1/2-1/\upsilon)$*) when* $\upsilon > 1/2 - 1/p$ *(*$\upsilon < 1/2 - 1/p$*).*

In general, for a small positive constant Δ_a, write $\tilde{a} = (1-\beta)/2 + \Delta_a$ ($\tilde{a} = (1-\beta)(1/2-1/\upsilon) + \Delta_a$) when $\upsilon > 1/2 - 1/p$ ($\upsilon < 1/2 - 1/p$). Given $1 \leqslant \alpha \leqslant \beta < 1$, we can see the error $|\hat{\sigma}_1^2(\lambda) - \sigma_*^2|$ goes to zero when $1/(2+2\upsilon) \leqslant \alpha \leqslant \beta < 1 - 2\Delta_a$ or $1/(2+2\upsilon) \leqslant \alpha \leqslant \beta < 1 - \Delta_a/(1/2+1/\upsilon)$ respectively. Meanwhile, the selection of α and β that helps the error $|\hat{\sigma}_1^2 - \sigma_*^2|$ converging to zero is similar but bounded more conservatively by $\min(1/2, 1-2\Delta_a)$ or $\min(1/2, 1-\Delta_a/(1/2+1/\upsilon))$. If $1/2 \leqslant \beta < 1 - 2\Delta_a$ ($1/2 \leqslant \beta < 1 - \Delta_a/(1/2+1/\upsilon)$) when $\upsilon > 1/2 - 1/p$ ($\upsilon < 1/2 - 1/p$), the estimator with thresholding $\hat{\sigma}_1^2(\lambda)$ will outperform $\hat{\sigma}_1^2$. Also, it is interesting to note that when $\upsilon > 1/2 - 1/p$, $\hat{\sigma}_1^2(\lambda)$ could achieve the optimal error $|\hat{\sigma}_1^2(\lambda) - \sigma_*^2| = O_p\{(n^{(-1+2\Delta_a)/3}) \log_2(n)^{-1}\}$ if $\alpha = \beta = (1-2\Delta_a)/3$, which is close to that of MSE optimal variance estimate introduced in Wu and Zhao (2007) and Dette et al. (2020) when Δ_a is small enough. Also, when $\upsilon < 1/2 - 1/p$, we have the optimal error $|\hat{\sigma}_1^2(\lambda) - \sigma_*^2| = O_p\{(n^{(-1/2-1/\upsilon+\Delta_a)/(3/2+1/\upsilon))}) \log_2(n)^{-1}\}$ with $\alpha = \beta = (1/2 + 1/\upsilon - \Delta_a)/(3/2+1/\upsilon)$, which seems that our estimators can outperform when there exists a tiny υ, i.e. under a strong dependence case. However, the following restriction is required to be discussed as well.

From a more detailed perspective, we shall also consider the inequality $\alpha > \min(2/p, 1-C_4)$ when $\upsilon > 1/2-1/p$ (or $\alpha > \min(1/(1+\upsilon p), 1-C_4)$ when $\upsilon < 1/2-1/p$) introduced in Lemma 3.2, where p, υ and C are pre-specified parameters, and C_4 is a positive constant that depends on p, υ, C and the dependence condition $\|X_\cdot\|_{2,\upsilon}$ with C being

a user-tailored constant in threshold λ. In particular, the value of C_4 should not be ignored (can be relatively small) when q and C are large and υ, together with $\|X_{\cdot}\|_{2,\upsilon}$, is small. However, because $\|X_{\cdot}\|_{2,\upsilon}$ is closely related to the underlying model of error process and no explicit formulas are provided for constants in Nagaev inequality (3.3.5), we can hardly provide the precise value of constant C_4 and hence cannot further narrow the range of α and β derived above.

In terms of the choice of scales following $a = \alpha \log_2(n)$ and $b = \beta \log_2(n)$, the existence of indicator function in both numerator and denominator of our estimators, such as $1/K_j^*$ and $\mathbb{1}_{(|d_{j,k}|\leqslant\lambda)}$ in $\hat{\sigma}_1^2(\lambda)$, complicates the computation of $\mathbb{E}[\hat{\sigma}_1^2(\lambda)]$. To be specific, even under the strict assumption that the noise X_i is a Gaussian process, we still have difficulty dealing with $\mathbb{E}[1/K_j^* \sum_{k=1}^{K_j} d_{j,k}^2 \mathbb{1}_{(|d_{j,k}|\leqslant\lambda)}]$ when the explicit formula of $\mathbb{E}[d_{j,k}^2 \mathbb{1}_{(|d_{j,k}|\leqslant\lambda)}]$ can be easily derived as follows

$$\begin{aligned}
\mathbb{E}[d_{j,k}^2 \mathbb{1}_{(|d_{j,k}|\leqslant\lambda)}] &= \mathrm{Var}(d_{j,k} \mathbb{1}_{(|d_{j,k}|\leqslant\lambda)}) \\
&= \sigma_j^2 \left[1 + \frac{-(\lambda/\sigma_j)\phi(-\lambda/\sigma_j) - (\lambda/\sigma_j)\phi(\lambda/\sigma_j)}{\Phi(\lambda/\sigma_j) - \Phi(-\lambda/\sigma_j)}\right] \\
&= \sigma_j^2 \left[1 - \frac{(\lambda/\sigma_j)\phi(\lambda/\sigma_j)}{\Phi(\lambda/\sigma_j) - 1/2}\right] = \sigma_j^2(1 - C_j)
\end{aligned}$$

where $\sigma_j = \sqrt{\mathrm{Var}(d_{j,k})}$, $\phi(\cdot)$ and $\Phi(\cdot)$ are probability density function and cumulative distribution function of standard normal distribution respectively. The function $x\phi(x)/(\Phi(x) - 1/2)$ is a monotonically decreasing function, which maps $(0, \infty)$ onto $(0, 1)$. Consequently, it becomes challenging to decide the optimal value of scales a and b with respect to bias-variance trade-off. Therefore, instead, we offer practical guidance on the selection of scales in Section 3.4.1.

Analogously, we provide a set $\mathcal{B}_n$ (3.3.18) for MODWT-based estimators $\hat{\sigma}_2^2$ and $\hat{\sigma}_2^2(\lambda)$. Similarly, the set satisfies that $P(\mathcal{B}_n) \to 0$ as $n \to \infty$ when $\lambda = C\sqrt{2\log(n)}$ due to the inequality $P\{|\tilde{W}_{j,k}| > C\sqrt{\log(n)}\} \to 0$, $k \in \tilde{S}_j^0$, $j = 1, \ldots, J_0$, for some constant C

proved in Lemma 3.3. And we have the following conclusion.

Theorem 3.2 *Let Y_i follows the model (2.2.1) with piecewise-constant signal in (3.1.1) plus the stationary error process X_i in (3.1.2) satisfying Assumptions (A1)-(A3). Let a and b denote the minimum and maximum value of scales chosen for estimators respectively. Write $a = \alpha \log_2(n)$ and $b = \beta \log_2(n)$. Let the threshold parameter satisfy $\lambda = C\sqrt{2\log(n)}$ for certain constant C. Considering the set $\mathcal{B}_n$,*

$$\mathcal{B}_n = \left\{ \forall\ 1 \leqslant k \leqslant n - 2m_j + 1 \quad \frac{1}{\sqrt{2m_j}} \left| \sum_{i=k+m_j}^{k+2m_j-1} X_i - \sum_{i=k}^{k+m_j-1} X_i \right| \leqslant \lambda \right\} \tag{3.3.20}$$

which contains all $|\tilde{Z}_{j,k}| \leqslant \lambda$ at $k = 1, \ldots, n(2m_j)^{-1}$. When it satisfies $P(\mathcal{B}_n) \to 1$ for some constant C (different from the one in Theorem 3.1), we have for some $0 < \alpha \leqslant \beta < 1$,

$$\begin{aligned} |\hat{\sigma}_2^2(\lambda) - \sigma_*^2| &= O_p\{(n^{-1+\tilde{a}} + n^{-\alpha} + n^{-1+\nu_0}) \log_2(n)^{-1}\} \\ |\hat{\sigma}_2^2 - \sigma_*^2| &= O_p\{(n^{-1+\tilde{a}} + n^{-\alpha} + n^{-1+2\beta}) \log_2(n)^{-1}\} \end{aligned} \tag{3.3.21}$$

when $\alpha \geqslant 1/(2+2\upsilon)$, $\nu_0 > \beta$ and $\tilde{a} > 1/2$.

Similarly, on the one hand, we discuss the choice of constants α, β, $\tilde{a}$ and ν_0 from a general perspective. Given $0 < \alpha \leqslant \beta < 1$, the error $|\hat{\sigma}_2^2(\lambda) - \sigma_*^2|$ (or $|\hat{\sigma}_2^2 - \sigma_*^2|$) is proved to converge to zero when $1/(2+2\upsilon) \leqslant \alpha < \nu_0 < 1$ (or $1/(2+2\upsilon) \leqslant \alpha \leqslant \beta < 1/2$) and $1/2 < \tilde{a} < 1$. If $\beta \geqslant 1/2$, the estimator $\hat{\sigma}_2^2(\lambda)$ will outperform $\hat{\sigma}_2^2$. Also, write $\nu_0 = \alpha + \Delta_\nu$, the estimator with thresholding $\hat{\sigma}_2^2(\lambda)$ could achieve the optimal error $|\hat{\sigma}_1^2(\lambda) - \sigma_*^2| = O_p\{(n^{(-1+\Delta_\nu)/2}) \log_2(n)^{-1}\}$ if $\alpha = 1 - \tilde{a} = (1 - \Delta_\nu)/2$ while $\hat{\sigma}_2^2$ could get $|\hat{\sigma}_1^2 - \sigma_*^2| = O_p\{(n^{-1/3} \log_2(n)^{-1}\}$ when $\alpha = \beta = 1/3$. On the other hand, the underlying condition from Lemma 3.3 for Theorem 3.2 is $\nu_0 > \nu \geqslant 2/(p-2)$ if $\upsilon > 1/2 - 1/p$ (or $\nu_0 > \nu \geqslant 1/(\upsilon p)$ if $\upsilon < 1/2 - 1/p$), which would be impacted by the pre-specified parameters p and υ. Moreover, compared to estimators built on DWT, Lemma 3.3 imposes stricter restrictions on constants depending on p and υ. These constraints can

significantly influence the optimal choice of constants α, β, $\tilde{a}$ and ν_0, thereby affecting the optimal error of estimators constructed with MODWT.

Furthermore, for the median-based estimators $\hat{\sigma}_3(\lambda)$ and $\hat{\sigma}_4(\lambda)$, while we do not present their analogous consistency proofs in this thesis, we do include them in our comparative simulation studies of Section 3.4.

3.4 Simulation Studies I

In this section, we shall run a series of tests on dependent data with piecewise-constant signals. To intuitively display the efficacy of our new estimators $\hat{\sigma}_2^2(\lambda)$ and $\hat{\sigma}_4^2(\lambda)$, we will set several estimators mentioned in Chapter 2 as the benchmark. To be specific, approaches (A1) and (A2) represent the estimators proposed in Wu and Zhao (2007), with definitions (2.4.9) and (2.4.10) respectively; also, the estimator (2.4.11) introduced in Dette et al. (2020) is recorded as approach (A3); (A4) indicates the general framework developed by Chan (2022), with definition (2.4.15). Meanwhile, we also consider the two estimators built on the assumption of an AR(p) error process: (A5) is the difference-based estimator proposed in Khismatullina and Vogt (2020), implemented in **R** package **dlrv** available on the author's website, and (A6) is the residual-based estimator introduced in Qiu et al. (2013), with definition (2.4.6), where all calculations are implemented by **R** package **aTSA** (Qiu, 2015) and **itsmr** (Weigt, 2022). (A7) stands for the estimator introduced in Dette et al. (2019), which is defined as (2.4.13). Only the MODWT-based estimators $\hat{\sigma}_2^2(\lambda)$ and $\hat{\sigma}_4^2(\lambda)$ are tested here since they can be applied without the restriction of dyadic sample size. All parameters are chosen as recommended in the corresponding papers.

As mentioned in Remark 3.1, the commonly used ARMA process is a special case in the general system applied for our dependent noise and it satisfies all of the assumptions

listed in the previous section. When the error process $\{X_i\}_{i=1}^n$ follows an ARMA(p, q) model represented with $\Phi(B)X_t = \Theta(B)\epsilon_t$ and $\text{Var}(\epsilon_t) = \sigma^2$, its LRV can also be written as $\sigma_*^2 = \sigma^2(\Theta(1)/\Phi(1))^2$ (Lee and Phillips, 1994). This explicit formula for true LRV makes it the concentration of our simulation.

To evaluate the performance of our estimators, we report the normalized absolute error (NAE) between true and estimated LRVs, which is given by $\tilde{\sigma}_* = |\sigma_* - \hat{\sigma}_*|/\sigma_*$. Also, though detailed rules have not been fully investigated, we shall first provide some guidance on the choice of important parameters, such as λ, a and b, for the new LRV estimators.

3.4.1 Practical considerations

The choice of λ. For the threshold $\lambda = C\sqrt{\log(n)}$, our decision is initiated from the MAD-based robust estimator in Johnstone and Silverman (1997) developed for computing the wavelet transformed noise level at particular scale j, which is defined as $\hat{\sigma}_j = \text{MAD}(d_{j,k}, k = 1, \ldots, n/(2m_j))/z_{1/4}$, where $z_{1/4}$ is the third quartile of the standard normal distribution. And for the MODWT case, we have $\hat{\sigma}_j = \text{MAD}(\tilde{W}_{j,k}, k = 1, \ldots, n - 2m_j + 1)/z_{1/4}$. For simplicity, we choose an overall threshold for all scales and recommend $\lambda = \dot{C}(\sqrt{2m_{J_*}}\hat{\sigma}_{J_*})\sqrt{2\log(n)}$, where $J_* = \lfloor \log_2(n)/2 \rfloor$. The value $\dot{C}$ is chosen to be 0.5 for $\hat{\sigma}_2^2(\lambda)$ and 0.75 for $\hat{\sigma}_4^2(\lambda)$. In practice, this constant can be somehow related to signal-to-noise ratio in data, which may be discussed in future work.

The choice of a and b. In our simulation studies, we report the results obtained with $a = 4$ and $b = 4$ for the smallest sample size $n = 100$ whereas $a = \lceil 2\log_2(n)/5 \rceil$ and $b = \lfloor 2\log_2(n)/3 \rfloor$ for larger n.

3.4.2 Settings

Here we consider a variety of ARMA(p, q) error structures $\{X_i\}_{i=1}^n$, see (N1)-(N18). To provide more results, we assess the performance of different estimators both in the case of no change-points ($q = 0$) and one or more change-points ($q \geqslant 1$) over different sample sizes. In the following, we shall present the outcomes of six typical settings and postpone the descriptions of the simulation results from the remaining scenarios to Appendix 3.6. We also consider the scenario where the signal $\theta_i = 0$ to evaluate the possible under- or over-estimation for independent data.

(M1) f_t undergoes $q = 0$ with $n = 200$;

(M2) f_t undergoes $q = 0$ with $n = 500$;

(M3) f_t undergoes $q = 1$ change-points at $\eta_1 = 30$ with $n = 100$ and $(\theta_1, \theta_2) = (2.5, -2.5)$;

(M4) f_t undergoes $q = 1$ change-points at $\eta_1 = 100$ with $n = 200$ and $(\theta_1, \theta_2) = (1.5, -1.5)$;

(M5) f_t undergoes $q = 2$ change-points at $(\eta_1, \eta_2) = (30, 80)$ with $n = 150$ and $(\theta_1, \theta_2, \theta_3) = (-1, 3, -3)$;

(M6) f_t undergoes $q = 2$ change-points at $(\eta_1, \eta_2) = (80, 200)$ with $n = 300$ and $(\theta_1, \theta_2, \theta_3) = (0, 2, -2)$;

and the ARMA(p, q) noise follows

(N1) AR$(1)^{SP}$: $\phi = 0.20$, $\sigma = 1.00$;

(N2) AR$(1)^{LP}$: $\phi = 0.80$, $\sigma = 0.50$;

(N3) AR$(1)^{SN}$: $\phi = -0.20$, $\sigma = 1.00$;

(N4) AR$(1)^{LN}$: $\phi = -0.80$, $\sigma = 0.50$;

(N5) AR$(2)^{P}$: $\phi = c(0.75, -0.15)$, $\sigma = 0.50$;

(N6) AR$(2)^{N}$: $\phi = c(-0.75, 0.15)$, $\sigma = 0.50$;

(N7) MA$(1)^{SP}$: $\theta = 0.20$, $\sigma = 1.00$;

(N8) MA(1)^{LP}: $\theta = 0.80$, $\sigma = 1.00$;

(N9) MA(1)^{SN}: $\theta = -0.20$, $\sigma = 1.00$;

(N10) MA(1)^{LN}: $\theta = -0.80$, $\sigma = 1.00$;

(N11) MA(6)^{1}: $\theta = c(0.80, 0.70, 0.60, 0.50, 0.40, 0.30)$, $\sigma = 0.75$;

(N12) MA(6)^{2}: $\theta = c(0.80, -0.70, 0.60, 0.50, -0.40, 0.30)$, $\sigma = 1.00$;

(N13) ARMA$(1,1)^{1}$: $\phi = 0.50$, $\theta = 0.75$, $\sigma = 0.75$;

(N14) ARMA$(1,1)^{2}$: $\phi = 0.50$, $\theta = -0.75$, $\sigma = 1.00$;

(N15) ARMA$(1,1)^{3}$: $\phi = -0.50$, $\theta = 0.75$, $\sigma = 1.00$;

(N16) ARMA$(1,1)^{4}$: $\phi = -0.50$, $\theta = -0.75$, $\sigma = 0.50$;

(N17) ARMA$(2,6)^{1}$: $\phi = c(0.75, -0.15)$, $\theta = c(0.80, 0.70, 0.60, 0.50, 0.40, 0.30)$, $\sigma = 0.25$;

(N18) ARMA$(2,6)^{2}$: $\phi = c(-0.75, 0.15)$, $\theta = c(0.80, 0.70, 0.60, 0.50, 0.40, 0.30)$, $\sigma = 0.50$.

where ϕ, θ and σ represent the parameters we simulate in **R**.

Models (M1) and (M3) to (M7) consider relatively shorter time series with $n \in [100, 300]$. Models (M7), (M9), (M11) and (M13) contain relatively more frequent change-points which can be close to each other. On the other hand, noise models (N1)-(N18) represent ARMA(p, q) stationary error scenarios. Among all models, AR models (N4) and (N6), MA model (N10) and ARMA models (N14) and (N16) have LRV closer to zero, making its accurate estimation difficult. Models (A2), (A5) and (A17) show stronger autocorrelations in the error process $\{X_i\}_{i=1}^{n}$.

In general, we evaluate the performance of all methods when the signal series satisfies models (M1) and (M4)-(M9). When sample size increases, algorithms (A5) and (A6), proposed under the assumption of an AR(p) error process tend to possess a larger computation burden with the **R** packages mentioned in their papers. Hence, for simulated data produced by the other signal models, we only test the effectiveness of estimators (A1)-(A4), (A7) and the two introduced wavelet-based estimators.

3.4.3 Results

Table 3.1 to 3.6 and Table 3.10 to 3.17 summarise the results of the comparative simulation studies from 100 realisations under each combination of signal and noise, i.e. (M1)-(M14) and (N1)-(N18). Besides NAEs for all models, we also provide simple summations on NAEs and record them under "Total" and "Subtotal", where "Total" is for the sum of all elements above while "Subtotal" shows that of elements under different model choices. These two measures can help us illustrate the effectiveness of estimators on different noises in general and also demonstrates their performance on some of the error processes that most estimators can work well with.

The reason for computing "Subtotal" is that we can see most tested estimators show poor performance on estimating data with the negatively correlated noise such as $\text{AR}(1)^{LN}$ and $\text{MA}(1)^{LN}$. Such failure in estimation is largely resulted from the nature of the true LRV of the tested noise. For example, $\text{AR}(1)^{LN}$ $X_i = -0.8X_{i-1} + \epsilon_i$ with $\sigma = 0.5$ has $\sigma_*^2 \approx 0.1389$ and $\text{MA}(1)^{LN}$ $X_i = \epsilon_i - 0.8\epsilon_{i-1}$ with $\sigma = 1$ leads to $\sigma_*^2 = 0.2$. These values are too small in practice, making the bias more obvious. Therefore, to be specific, for AR models, "Subtotal" only contains $\text{AR}(1)^{SP}$, $\text{AR}(1)^{LP}$, $\text{AR}(1)^{SN}$, and $\text{AR}(2)^{P}$; for MA models, "Subtotal" represents all models except $\text{MA}(1)^{LN}$; for ARMA models, it denotes $\text{ARMA}(1,1)^{1}$, $\text{ARMA}(1,1)^{3}$, $\text{ARMA}(2,6)^{1}$, and $\text{ARMA}(2,6)^{2}$; i.e. not consider the case with large negative coefficients.

Table 3.1 to 3.6 and Table 3.10 to 3.17 summarise the simulation results of the LRV estimators (A1)-(A7) for all combinations of signal and noise, where the methods that achieve the relatively best performance are highlighted in bold for the corresponding scenario. Overall, these tables indicate that our wavelet-based estimators can perform at least as well as the existing ones when the sample size is small and have relatively better performance for data with larger sample size, especially for dataset with MA parameters. In addition, although the median-based estimator cannot outperform the

mean-based one when applying the chosen λ, it can generally be more robust to the dependence structure of the time series.

Table 3.1: Normalized absolute error (NAE) between true and estimated Long-Run Variance (LRV) of the sequences with signal (M1), where $n = 200$ and $q = 0$, and noises following an ARMA model (N1)-(N18) based on approaches (A1)-(A7), $\hat{\sigma}_2(\lambda)$ and $\hat{\sigma}_4(\lambda)$, which stand respectively for "Mean" and "Median" MODWT-based estimators. Here, both of the chosen scales are $4:5$.

	(A1)	(A2)	(A3)	(A4)	(A5)	(A6)	(A7)	$\hat{\sigma}_2(\lambda)$	$\hat{\sigma}_4(\lambda)$
$\text{AR}(1)^{SP}$	**0.083**	0.126	0.089	0.298	**0.081**	0.096	**0.082**	0.116	0.130
$\text{AR}(1)^{LP}$	0.554	0.557	0.554	**0.166**	0.197	0.609	0.617	0.264	0.258
$\text{AR}(1)^{SN}$	0.109	0.151	0.115	0.195	**0.053**	0.101	0.079	0.108	0.158
$\text{AR}(1)^{LN}$	0.718	0.716	0.693	0.740	**0.037**	0.515	0.346	0.186	0.208
$\text{AR}(2)^{P}$	0.258	0.256	0.261	0.142	**0.116**	0.574	0.291	0.141	0.146
$\text{AR}(2)^{N}$	1.076	1.099	1.039	0.664	**0.090**	0.513	0.184	0.182	0.222
Subtotal	1.004	1.090	1.019	0.801	**0.447**	1.380	1.069	0.629	0.692
Total	2.798	2.905	2.751	2.205	**0.574**	2.408	1.599	0.997	1.122
$\text{MA}(1)^{SP}$	0.130	0.181	0.122	0.236	**0.089**	0.108	**0.089**	0.132	0.141
$\text{MA}(1)^{LP}$	0.133	0.158	0.143	0.582	0.143	0.260	**0.107**	0.133	0.161
$\text{MA}(1)^{SN}$	0.133	0.233	**0.104**	0.262	**0.102**	0.145	0.120	0.140	0.165
$\text{MA}(1)^{LN}$	2.859	2.851	2.785	**0.638**	3.206	3.351	3.100	1.335	1.283
$\text{MA}(6)^{L}$	0.386	0.372	0.395	1.118	0.368	0.232	0.496	**0.175**	**0.165**
$\text{MA}(6)^{M}$	0.293	0.288	0.297	0.383	0.247	0.264	0.332	**0.129**	0.153
Subtotal	1.075	1.232	1.061	2.581	0.949	1.009	1.144	**0.709**	0.785
Total	3.934	4.083	3.846	3.219	4.155	4.360	4.244	**2.044**	**2.068**
$\text{ARMA}(1,1)^{1}$	0.246	0.246	0.252	1.128	0.194	0.151	0.303	**0.115**	**0.108**
$\text{ARMA}(1,1)^{2}$	0.595	0.618	0.597	0.331	0.800	0.747	0.794	**0.300**	0.338
$\text{ARMA}(1,1)^{3}$	**0.068**	0.118	0.067	0.237	0.117	0.116	0.082	0.102	0.129
$\text{ARMA}(1,1)^{4}$	4.218	4.452	4.086	**0.923**	3.647	1.395	4.059	2.064	1.997
$\text{ARMA}(2,6)^{1}$	0.490	0.496	0.495	0.481	0.258	0.812	0.583	**0.171**	**0.182**
$\text{ARMA}(2,6)^{2}$	0.348	0.338	0.348	0.278	0.481	0.345	0.475	**0.136**	0.191
Subtotal	1.152	1.198	1.162	2.124	1.050	1.424	1.443	**0.524**	0.610
Total	5.965	6.268	5.845	3.378	5.497	3.566	6.296	**2.888**	2.945

Table 3.2: Normalized absolute error (NAE) between true and estimated Long-Run Variance (LRV) of the sequences with signal (M2), where $n = 500$ and $q = 0$, and noises following an ARMA model (N1)-(N18) based on approaches (A1)-(A5), $\hat{\sigma}_2(\lambda)$ and $\hat{\sigma}_4(\lambda)$, which stand respectively for "Mean" and "Median" MODWT-based estimators. Here, both of the chosen scales are $4:5$.

	(A1)	(A2)	(A3)	(A4)	(A7)	$\hat{\sigma}_2(\lambda)$	$\hat{\sigma}_4(\lambda)$
$\text{AR}(1)^{SP}$	0.099	0.128	0.095	0.194	**0.077**	**0.080**	0.101
$\text{AR}(1)^{LP}$	0.471	0.470	0.472	**0.189**	0.564	0.267	0.278
$\text{AR}(1)^{SN}$	0.107	0.150	0.094	0.192	**0.057**	**0.058**	0.074
$\text{AR}(1)^{LN}$	0.410	0.376	0.428	0.700	0.686	**0.208**	**0.212**
$\text{AR}(2)^{P}$	0.156	0.164	0.166	0.179	0.230	**0.088**	**0.090**
$\text{AR}(2)^{N}$	0.846	0.877	0.823	0.671	1.076	0.159	**0.149**
Subtotal	0.833	0.912	0.827	0.754	0.928	**0.493**	0.543
Total	2.089	2.165	2.078	2.125	2.690	**0.860**	0.904
$\text{MA}(1)^{SP}$	0.067	0.095	0.074	0.124	**0.062**	0.073	0.093
$\text{MA}(1)^{LP}$	0.092	0.107	0.095	0.698	**0.082**	0.087	0.092
$\text{MA}(1)^{SN}$	0.099	0.125	0.099	0.185	0.100	**0.059**	0.066
$\text{MA}(1)^{LN}$	2.101	2.094	2.107	**0.684**	2.567	1.339	1.270
$\text{MA}(6)^{L}$	0.262	0.267	0.261	1.296	0.400	**0.143**	**0.144**
$\text{MA}(6)^{M}$	0.212	0.220	0.213	0.506	0.259	**0.132**	**0.133**
Subtotal	0.732	0.814	0.742	2.809	0.903	**0.494**	0.528
Total	2.833	2.908	2.849	3.493	3.470	1.833	**1.798**
$\text{ARMA}(1,1)^{1}$	0.162	0.197	0.160	1.137	0.243	**0.113**	**0.120**
$\text{ARMA}(1,1)^{2}$	0.565	0.589	0.554	0.331	0.743	0.282	**0.249**
$\text{ARMA}(1,1)^{3}$	0.087	0.098	0.089	0.173	**0.055**	0.075	0.086
$\text{ARMA}(1,1)^{4}$	2.978	3.132	2.948	**0.896**	3.791	1.983	1.898
$\text{ARMA}(2,6)^{1}$	0.321	0.314	0.327	0.703	0.467	0.174	**0.154**
$\text{ARMA}(2,6)^{2}$	0.278	0.291	0.272	0.167	0.332	**0.146**	**0.156**
Subtotal	0.848	0.900	0.848	2.180	1.097	**0.508**	**0.516**
Total	4.391	4.621	4.350	3.407	5.631	2.773	**2.663**

Table 3.3: Normalized absolute error (NAE) between true and estimated Long-Run Variance (LRV) of the sequences with signal (M3), where $n = 100$ and $q = 1$, and noises following an ARMA model (N1)-(N18) based on approaches (A1)-(A7), $\hat{\sigma}_2(\lambda)$ and $\hat{\sigma}_4(\lambda)$, which stand respectively for "Mean" and "Median" MODWT-based estimators. Here, both of the chosen scales are $4:4$.

	(A1)	(A2)	(A3)	(A4)	(A5)	(A6)	(A7)	$\hat{\sigma}_2(\lambda)$	$\hat{\sigma}_4(\lambda)$
$\text{AR}(1)^{SP}$	**0.123**	0.138	0.192	1.201	0.782	0.825	0.148	0.168	0.246
$\text{AR}(1)^{LP}$	0.496	0.547	0.432	0.524	0.483	**0.317**	0.518	0.563	0.552
$\text{AR}(1)^{SN}$	0.420	0.332	0.634	1.455	1.139	1.442	0.552	**0.159**	**0.161**
$\text{AR}(1)^{LN}$	1.564	0.545	2.846	0.851	3.626	3.637	2.864	0.219	**0.118**
$\text{AR}(2)^{P}$	0.099	0.181	**0.073**	0.283	1.350	0.183	**0.082**	0.232	0.270
$\text{AR}(2)^{N}$	1.247	0.340	2.466	1.790	8.990	6.433	3.142	0.186	**0.157**
Subtotal	**1.138**	1.198	1.331	3.463	3.754	2.767	1.300	**1.122**	1.229
Total	3.949	2.083	6.643	6.104	16.370	12.837	7.306	1.527	**1.504**
$\text{MA}(1)^{SP}$	0.289	0.295	0.361	3.269	1.102	1.135	0.252	**0.160**	0.241
$\text{MA}(1)^{LP}$	0.124	0.188	0.095	2.187	0.672	0.923	**0.076**	0.167	0.217
$\text{MA}(1)^{SN}$	0.517	0.258	0.741	1.966	1.330	1.673	0.673	**0.140**	0.197
$\text{MA}(1)^{LN}$	4.286	3.309	5.392	2.759	6.529	10.218	5.701	2.065	**1.639**
$\text{MA}(6)^{L}$	0.397	0.431	0.373	3.679	0.811	0.570	0.495	0.413	**0.368**
$\text{MA}(6)^{M}$	0.160	0.222	0.130	3.986	**0.096**	0.907	0.253	0.178	0.257
Subtotal	1.487	1.394	1.700	15.087	4.011	5.208	1.749	**1.058**	1.280
Total	5.773	4.703	7.092	17.846	10.540	15.426	7.450	3.123	**2.919**
$\text{ARMA}(1,1)^{1}$	0.257	0.294	0.240	3.689	0.484	0.394	0.326	0.289	**0.185**
$\text{ARMA}(1,1)^{2}$	1.065	0.700	1.436	1.948	2.356	2.887	1.546	0.523	**0.423**
$\text{ARMA}(1,1)^{3}$	0.274	0.298	0.326	3.256	1.101	1.088	0.264	**0.137**	0.162
$\text{ARMA}(1,1)^{4}$	8.101	4.675	11.969	9.175	13.965	15.726	11.772	2.765	**1.981**
$\text{ARMA}(2,6)^{1}$	0.451	0.487	0.414	2.867	1.924	**0.317**	0.537	0.432	0.368
$\text{ARMA}(2,6)^{2}$	0.214	0.358	**0.094**	0.561	1.958	0.880	0.128	0.306	0.281
Subtotal	1.196	1.437	1.074	10.373	5.467	2.679	1.255	1.164	**0.996**
Total	10.362	6.812	14.479	21.496	21.788	21.292	14.573	4.452	**3.400**

Table 3.4: Normalized absolute error (NAE) between true and estimated Long-Run Variance (LRV) of the sequences with signal (M4), where $n = 200$ and $q = 1$, and noises following an ARMA model (N1)-(N18) based on approaches (A1)-(A5), $\hat{\sigma}_2(\lambda)$ and $\hat{\sigma}_4(\lambda)$, which stand respectively for "Mean" and "Median" MODWT-based estimators. Here, both of the chosen scales are $4:5$.

	(A1)	(A2)	(A3)	(A4)	(A5)	(A6)	(A7)	$\hat{\sigma}_2(\lambda)$	$\hat{\sigma}_4(\lambda)$
$\text{AR}(1)^{SP}$	0.141	0.161	0.151	0.620	**0.121**	0.233	0.089	0.160	0.147
$\text{AR}(1)^{LP}$	0.461	0.470	0.440	0.390	**0.196**	0.455	0.507	0.417	0.413
$\text{AR}(1)^{SN}$	0.178	0.239	0.208	0.195	0.114	0.387	0.220	**0.076**	**0.079**
$\text{AR}(1)^{LN}$	0.693	0.352	1.355	0.627	0.339	0.854	1.621	0.286	**0.236**
$\text{AR}(2)^{P}$	0.130	0.141	**0.111**	0.450	0.207	0.230	**0.118**	0.155	0.192
$\text{AR}(2)^{N}$	0.612	0.307	1.118	2.081	1.535	1.917	1.831	0.186	**0.163**
Subtotal	0.910	1.011	0.910	1.655	0.638	1.305	0.934	**0.808**	0.831
Total	2.215	1.670	3.383	4.363	2.512	4.076	4.386	1.280	**1.230**
$\text{MA}(1)^{SP}$	0.100	0.136	0.107	0.880	0.152	0.330	**0.066**	0.110	0.131
$\text{MA}(1)^{LP}$	0.071	0.125	0.067	1.109	0.234	0.515	0.077	**0.058**	0.073
$\text{MA}(1)^{SN}$	0.261	0.217	0.310	0.382	0.154	0.509	0.226	0.112	**0.084**
$\text{MA}(1)^{LN}$	2.690	2.385	3.152	**0.733**	3.585	5.489	3.432	1.868	1.758
$\text{MA}(6)^{L}$	0.297	0.274	0.308	1.501	**0.214**	0.231	0.366	0.248	**0.226**
$\text{MA}(6)^{M}$	0.227	0.233	0.202	1.156	0.229	0.539	0.237	**0.188**	0.197
Subtotal	0.956	0.985	0.994	5.028	0.983	2.124	0.972	**0.716**	**0.711**
Total	3.646	3.370	4.146	5.761	4.568	7.613	4.404	2.584	**2.469**
$\text{ARMA}(1,1)^{1}$	0.202	0.245	0.180	1.567	0.209	0.236	0.215	**0.171**	**0.163**
$\text{ARMA}(1,1)^{2}$	0.768	0.712	0.933	**0.490**	1.000	1.468	0.951	0.531	**0.496**
$\text{ARMA}(1,1)^{3}$	0.154	0.195	0.158	0.855	0.282	0.477	**0.130**	**0.120**	**0.130**
$\text{ARMA}(1,1)^{4}$	4.845	3.666	7.036	3.222	4.547	5.919	6.867	2.840	**2.603**
$\text{ARMA}(2,6)^{1}$	0.321	0.314	0.331	1.296	**0.159**	0.651	0.409	0.233	0.194
$\text{ARMA}(2,6)^{2}$	0.263	0.309	0.227	**0.217**	0.507	**0.207**	0.223	0.222	0.231
Subtotal	0.940	1.063	0.896	3.935	1.157	1.571	0.977	0.746	**0.718**
Total	6.553	5.441	8.865	7.647	6.704	8.958	8.795	4.117	**3.817**

Table 3.5: Normalized absolute error (NAE) between true and estimated Long-Run Variance (LRV) of the sequences with signal (M5), where $n = 150$ and $q = 2$, and noises following an ARMA model (N1)-(N18) based on approaches (A1)-(A7), $\hat{\sigma}_2(\lambda)$ and $\hat{\sigma}_4(\lambda)$, which stand respectively for "Mean" and "Median" MODWT-based estimators. Here, both of the chosen scales are $4:4$.

	(A1)	(A2)	(A3)	(A4)	(A5)	(A6)	(A7)	$\hat{\sigma}_2(\lambda)$	$\hat{\sigma}_4(\lambda)$
$\text{AR}(1)^{SP}$	0.410	**0.054**	0.953	1.178	1.226	1.163	0.489	0.223	0.215
$\text{AR}(1)^{LP}$	0.295	0.467	0.074	1.331	0.688	**0.196**	0.331	0.358	0.475
$\text{AR}(1)^{SN}$	0.826	0.311	1.794	0.875	1.539	1.666	1.005	0.216	**0.100**
$\text{AR}(1)^{LN}$	3.038	0.840	6.877	2.205	4.604	4.489	4.298	0.415	**0.170**
$\text{AR}(2)^{P}$	0.271	0.220	0.866	0.684	2.214	0.287	0.339	**0.125**	0.245
$\text{AR}(2)^{N}$	3.176	1.435	6.106	1.383	11.894	7.352	3.706	0.290	**0.153**
Subtotal	1.802	1.052	3.687	4.068	5.667	3.312	2.164	**0.922**	1.035
Total	8.016	3.327	16.670	7.656	22.165	15.153	10.168	1.627	**1.358**
$\text{MA}(1)^{SP}$	0.483	**0.192**	1.055	1.052	1.470	1.197	0.534	**0.184**	0.197
$\text{MA}(1)^{LP}$	0.254	0.136	0.493	4.659	0.930	1.256	0.218	0.262	**0.142**
$\text{MA}(1)^{SN}$	0.961	0.299	1.940	1.283	1.700	1.773	1.113	0.323	**0.157**
$\text{MA}(1)^{LN}$	5.824	2.982	10.217	3.419	7.625	10.513	7.237	2.914	**1.356**
$\text{MA}(6)^{L}$	0.197	0.301	**0.128**	3.592	0.633	0.275	0.319	0.265	0.342
$\text{MA}(6)^{M}$	0.107	0.234	0.264	3.059	**0.102**	0.820	**0.093**	0.138	0.216
Subtotal	2.002	1.162	3.880	13.645	4.835	5.321	2.277	1.172	**1.054**
Total	7.826	4.144	14.097	17.064	12.460	15.834	9.514	4.086	**2.410**
$\text{ARMA}(1,1)^{1}$	0.129	0.188	0.122	4.470	0.595	0.420	0.131	**0.051**	0.155
$\text{ARMA}(1,1)^{2}$	1.964	0.933	3.491	3.054	3.271	3.583	2.367	0.989	**0.304**
$\text{ARMA}(1,1)^{3}$	0.545	0.260	1.069	4.009	1.341	1.274	0.588	0.333	**0.135**
$\text{ARMA}(1,1)^{4}$	12.384	5.469	24.867	3.180	17.115	18.209	16.753	3.867	**1.988**
$\text{ARMA}(2,6)^{1}$	0.225	0.382	**0.073**	1.595	1.840	0.243	0.298	0.306	0.390
$\text{ARMA}(2,6)^{2}$	**0.130**	0.308	0.712	0.876	2.267	1.105	0.167	0.315	0.463
Subtotal	1.029	1.138	1.976	10.950	6.043	3.042	1.184	**1.005**	1.143
Total	15.377	7.540	30.334	17.184	26.429	24.834	20.304	5.861	**3.435**

Table 3.6: Normalized absolute error (NAE) between true and estimated Long-Run Variance (LRV) of the sequences with signal (M6), where $n = 300$ and $q = 2$, and noises following an ARMA model (N1)-(N18) based on approaches (A1)-(A7), $\hat{\sigma}_2(\lambda)$ and $\hat{\sigma}_4(\lambda)$, which stand respectively for "Mean" and "Median" MODWT-based estimators. Here, both of the chosen scales are $4:5$.

	(A1)	(A2)	(A3)	(A4)	(A5)	(A6)	(A7)	$\hat{\sigma}_2(\lambda)$	$\hat{\sigma}_4(\lambda)$
$\text{AR}(1)^{SP}$	**0.119**	**0.124**	0.228	0.583	0.342	0.819	0.190	0.181	0.159
$\text{AR}(1)^{LP}$	0.418	0.457	0.360	0.538	**0.206**	0.272	0.420	0.254	0.315
$\text{AR}(1)^{SN}$	0.328	0.145	0.576	0.211	0.437	1.196	0.528	0.203	**0.088**
$\text{AR}(1)^{LN}$	1.436	0.579	2.836	1.121	1.272	2.976	2.680	0.370	**0.117**
$\text{AR}(2)^{P}$	**0.054**	0.103	0.165	0.381	0.749	0.229	0.104	0.129	0.093
$\text{AR}(2)^{N}$	1.082	0.273	2.410	1.614	4.859	5.595	2.533	0.311	**0.112**
Subtotal	0.919	0.829	1.329	1.713	1.734	2.516	1.242	0.767	**0.655**
Total	3.437	1.681	6.575	4.448	7.865	11.087	6.455	1.448	**0.884**
$\text{MA}(1)^{SP}$	0.180	0.202	0.249	0.656	0.395	0.853	0.229	0.186	**0.093**
$\text{MA}(1)^{LP}$	0.089	0.174	**0.075**	1.604	0.353	0.826	**0.064**	0.163	0.126
$\text{MA}(1)^{SN}$	0.346	0.293	0.564	0.345	0.446	1.228	0.572	0.211	**0.078**
$\text{MA}(1)^{LN}$	3.570	2.647	4.901	2.230	4.362	8.803	4.947	1.984	**1.245**
$\text{MA}(6)^{L}$	0.285	0.315	0.274	1.791	0.250	0.157	0.357	**0.142**	0.198
$\text{MA}(6)^{M}$	0.153	0.180	0.103	1.118	**0.136**	0.772	**0.124**	**0.135**	0.187
Subtotal	1.053	1.164	1.265	5.514	1.580	3.836	1.346	0.837	**0.682**
Total	4.623	3.811	6.166	7.744	5.942	12.639	6.293	2.821	**1.927**
$\text{ARMA}(1,1)^{1}$	**0.114**	0.171	**0.108**	2.502	0.345	0.490	0.134	0.123	0.203
$\text{ARMA}(1,1)^{2}$	1.181	0.857	1.624	1.095	1.607	3.043	1.633	0.593	**0.254**
$\text{ARMA}(1,1)^{3}$	0.254	0.282	0.364	0.676	0.490	0.971	0.311	0.141	**0.068**
$\text{ARMA}(1,1)^{4}$	6.621	3.732	11.634	1.529	7.117	13.286	11.006	2.638	**1.776**
$\text{ARMA}(2,6)^{1}$	0.309	0.347	0.282	1.138	0.658	0.465	0.356	**0.095**	0.128
$\text{ARMA}(2,6)^{2}$	0.121	0.218	**0.059**	0.291	1.026	0.878	0.057	0.115	0.214
Subtotal	0.798	1.018	0.813	4.607	2.519	2.804	0.858	**0.474**	0.613
Total	8.600	5.607	14.071	7.231	11.243	19.133	13.497	3.705	**2.643**

3.5 Simulation Studies II

To assess the efficacy of our wavelet-based estimators in change-point estimation, we incorporate them into DepSMUCE (Dette et al., 2020) to substitute the introduced LRV estimator and then compare the finite sample performance of this adapted change-point estimator with the original DepSMUCE. For simplicity, we shall use the notations DepSMUCE^1 and DepSMUCE^2 respectively for the procedure adjusted by MODWT-based LRV estimators $\hat{\sigma}_2^2(\lambda)$ and $\hat{\sigma}_4^2(\lambda)$, whose parameters are set according to the guidelines outlined in Section 3.4.1. We will now consider the three scenarios taken from Section 4 in Dette et al. (2020) and maintain the same block length 10 for DepSMUCE. The sample size is $n = 1000$ and 1000 realisations are generated under each setting. Likewise, we will demonstrate the performance of different DepSMUCE when working with three specific significance levels, i.e. $\alpha = 0.1, 0.5, 0.9$. In practice, these estimators can be efficiently computed using a dynamic programming approach and are readily available for implementation through the function "**stepFit**" in the **R** package "**stepR**".

Table 3.7: Distribution of $\hat{q} - q$ obtained by the DepSMUCE algorithm for data generated according to (2.2.1) with the signal $(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6) = (0, 1, 0, 2, 0, -1)$ and noise from an MA(1) process with $\theta = 0.3$ and $\sigma = 1.00$, the average Mean Square Error (MSE) of the resulting estimate of the signal and the average Hausdorff distance d_H over 1000 simulations.

	$\hat{q} - q$								
Method	$\leqslant -3$	-2	-1	0	1	2	$\geqslant 3$	MSE	d_H
DepSMUCE(0.1)	0.004	0.031	0.395	0.570	0.000	0.000	0.000	0.079	0.822
DepSMUCE(0.1)^1	0.000	0.015	0.281	0.704	0.000	0.000	0.000	0.063	0.595
DepSMUCE(0.1)^2	0.001	0.037	0.334	0.628	0.000	0.000	0.000	0.074	0.721
DepSMUCE(0.5)	0.000	0.002	0.058	0.934	0.006	0.000	0.000	0.039	0.217
DepSMUCE(0.5)^1	0.000	0.000	0.033	0.956	0.011	0.000	0.000	0.037	0.186
DepSMUCE(0.5)^2	0.000	0.002	0.046	0.937	0.014	0.001	0.000	0.039	0.214
DepSMUCE(0.9)	0.000	0.000	0.003	0.914	0.080	0.003	0.000	0.035	0.184
DepSMUCE(0.9)^1	0.000	0.000	0.005	0.953	0.039	0.003	0.000	0.035	0.158
DepSMUCE(0.9)^2	0.000	0.000	0.005	0.915	0.073	0.006	0.001	0.035	0.193

For each scenario, a frequency table is provided for the distribution of $\hat{q} - q$, where $\hat{q}$ and q denote the number of the estimated and true change-points respectively. We also compute the estimated Mean Squared Error of the estimated signal $\hat{f}_t$, which is defined as

$$\text{MSE} = \mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n}(f_t - \hat{f}_t)^2\right]$$

Moreover, an estimate of the *(scaled) Hausdorff distance* is defined below for assessing the accuracy of estimated locations of change-points. For $0 \leqslant d_H \leqslant 1$,

$$d_H = \frac{1}{n}\mathbb{E}\left[\max\left\{\max_{j=0,\ldots,q+1}\min_{k=0,\ldots,\hat{q}+1}|\eta_j - \hat{\eta}_k|, \max_{k=0,\ldots,\hat{q}+1}\min_{j=0,\ldots,q+1}|\eta_j - \hat{\eta}_k|\right\}\right] \tag{3.5.1}$$

where the true and estimated locations of change-points respectively satisfy the constraints $0 = \eta_0 < \eta_1 < \cdots < \eta_q < \eta_{q+1} = n$ and $0 = \hat{\eta}_0 < \hat{\eta}_1 < \cdots < \hat{\eta}_{\hat{q}} < \hat{\eta}_{\hat{q}+1} = n$. In general, smaller d_H indicates better performance of the algorithm. Since d_H can be largely influenced by distinct estimated change-point locations or an under-estimated change-point number, we can analyse the performance of different estimators by finding a balance among the above three measures.

First, since SMUCE often fails to produce satisfactory results when the error process displays stronger dependencies, we consider a sequence generated according to (2.2.1) with signal (3.1.1) and noise (3.1.2) given by $(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6) = (0, 1, 0, 2, 0, -1)$ and MA(1) process with $\theta = 0.3$ and $\sigma = 1.00$ respectively. This underlying model exhibits $q = 5$ change-points at locations $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (101, 301, 501, 551, 751)$. Table 3.7 indicates that after replacing the LRV estimator, the adjusted DepSMUCE methods perform at least as well as the original DepSMUCE and DepSMUCE[1] tends to outperform the others. Specifically, we can observe higher rate of correctly estimated change-point numbers, improved signal estimation (lower MSE) and more satisfactory estimation of change-point locations (smaller d_H).

Second, Dette et al. (2020) considers another example building on $(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6) = (0, 3, 0, 4, 0, -3)$ and an MA(4) process with $\theta = c(0.90, 0.80, 0.70, 0.60)$ and $\sigma = 1.00$, where the number and location of change-points remain unchanged. Moreover, the last example follows the model that contains a piecewise-constant signal produced by $(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6) = (0, 5, 1, 8, 1, -2)$ and an ARMA(2,6) error process following $\phi = c(0.75, -0.5)$, $\theta = c(0.80, 0.70, 0.60, 0.50, 0.40, 0.30)$ and $\sigma = 1.00$. Both cases involve strong dependencies, further confirming the substantial superiority of DepSMUCE over SMUCE for serial correlated data.

Table 3.8: Distribution of $\hat{q} - q$ obtained by the DepSMUCE algorithm for data generated according to (2.2.1) with the signal $(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6) = (0, 3, 0, 4, 0, -3)$ and an MA(4) error process with $\theta = c(0.90, 0.80, 0.70, 0.60)$ and $\sigma = 1.00$, the average Mean Square Error (MSE) of the resulting estimate of the signal and the average Hausdorff distance d_H over 1000 simulations.

	$\hat{q} - q$								
Method	$\leqslant -3$	-2	-1	0	1	2	$\geqslant 3$	MSE	d_H
DepSMUCE(0.1)	0.008	0.136	0.525	0.331	0.000	0.000	0.000	0.851	0.617
DepSMUCE(0.1)^1	0.007	0.069	0.365	0.557	0.002	0.000	0.000	0.651	0.422
DepSMUCE(0.1)^2	0.017	0.082	0.305	0.586	0.009	0.001	0.000	0.632	0.460
DepSMUCE(0.5)	0.000	0.003	0.158	0.826	0.013	0.000	0.000	0.411	0.211
DepSMUCE(0.5)^1	0.000	0.001	0.083	0.864	0.048	0.004	0.000	0.357	0.210
DepSMUCE(0.5)^2	0.000	0.006	0.081	0.774	0.110	0.024	0.005	0.368	0.286
DepSMUCE(0.9)	0.000	0.000	0.021	0.832	0.135	0.012	0.000	0.329	0.275
DepSMUCE(0.9)^1	0.000	0.000	0.037	0.803	0.143	0.015	0.002	0.338	0.292
DepSMUCE(0.9)^2	0.000	0.001	0.072	0.788	0.111	0.024	0.004	0.361	0.288

Table 3.8 and 3.9 show the efficacy of DepSMUCE algorithms in estimating the number and locations of change-points over highly dependent data. Considering the first three rows in these two tables, we can see that the adjusted change-point estimators perform better when $\alpha = 1$. However such advantage cannot be guaranteed for $\alpha = 0.5$ and 0.9. For example, we observe that only DepSMUCE^1 outperforms while DepSMUCE^2 consistently leads to poorer results if $\alpha = 0.5$; on the other hand, when $\alpha = 0.9$, conclusions for DepSMUCE^2 can vary a lot between Table 3.8 and 3.9.

Table 3.9: Distribution of $\hat{q} - q$ obtained by the DepSMUCE algorithm for data generated according to (2.2.1) with the signals (M1) and (M2) and the noises (N1)-(N9), the average Mean Square Error (MSE) of the resulting estimate of the signal and the average Hausdorff distance d_H over 1000 simulations.

	$\hat{q} - q$								
Method	$\leqslant -3$	-2	-1	0	1	2	$\geqslant 3$	MSE	d_H
DepSMUCE(0.1)	0.004	0.056	0.361	0.579	0.000	0.000	0.000	1.411	0.685
DepSMUCE(0.1)[1]	0.002	0.035	0.299	0.664	0.000	0.000	0.000	1.234	0.563
DepSMUCE(0.1)[2]	0.005	0.087	0.314	0.593	0.001	0.000	0.000	1.455	0.667
DepSMUCE(0.5)	0.000	0.000	0.056	0.933	0.011	0.000	0.000	0.671	0.255
DepSMUCE(0.5)[1]	0.000	0.000	0.049	0.934	0.016	0.001	0.000	0.662	0.238
DepSMUCE(0.5)[2]	0.000	0.002	0.085	0.889	0.023	0.001	0.000	0.743	0.273
DepSMUCE(0.9)	0.000	0.000	0.002	0.890	0.101	0.007	0.000	0.610	0.269
DepSMUCE(0.9)[1]	0.000	0.000	0.005	0.909	0.083	0.003	0.000	0.611	0.253
DepSMUCE(0.9)[2]	0.000	0.000	0.011	0.917	0.070	0.002	0.000	0.622	0.246

In general, after adjusting the current change-point detection procedure DepSMUCE, we can see our MODWT-based LRV estimators, especially the mean-based one $\hat{\sigma}_2^2(\lambda)$, can enhance the performance of change-point estimation under certain pre-specified significance levels α, such as 0.1 or 0.5.

3.6 Appendix – Complete Simulated Results

In the section, we display the remaining simulation results summarised in Section 3.4.3 of the main text. Here are the simulated models.

(M7) f_t undergoes $q = 3$ change-points at $(\eta_1, \eta_2, \eta_3) = (50, 200, 250)$ with $n = 300$ and $(\theta_1, \theta_2, \theta_3, \theta_4) = (-2, 4, -3, 3.5)$;

(M8) f_t undergoes $q = 3$ change-points at $(\eta_1, \eta_2, \eta_3) = (100, 250, 400)$ with $n = 500$ and $(\theta_1, \theta_2, \theta_3, \theta_4) = (0, 2.5, -1.5, 2)$;

(M9) f_t undergoes $q = 5$ change-points at $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (50, 200, 230, 300, 410)$ with $n = 500$ and $(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6) = (-2, 3, -4, 2.5, -3.5, 3)$;

(M10) f_t undergoes $q = 5$ change-points at $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (50, 200, 300, 500, 650)$ with $n = 750$ and $(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6) = (0, 2.5, -2, 2, -2, 1.5)$;

(M11) f_t undergoes $q = 7$ change-points at $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5, \eta_6, \eta_7) = (50, 180, 260, 300, 500, 570, 850)$ with $n = 900$ and $(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_7, \theta_8) = (0, 4, -3, 3.5, -2.5, 3.5, -3, 3)$;

(M12) f_t undergoes $q = 7$ change-points at $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5, \eta_6, \eta_7) = (50, 180, 300, 500, 750, 910, 1050)$ with $n = 1200$ and $(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_7, \theta_8) = (0, 2, -3, 1.5, -2, 2, -2, 2.5)$.

(M13) f_t undergoes $q = 12$ change-points at $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5, \eta_6, \eta_7, \eta_8, \eta_9, \eta_{10}, \eta_{11}, \eta_{12}) = (50, 180, 220, 300, 500, 570, 850, 960, 1000, 1200, 1380, 1420)$ with $n = 1500$ and $(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_7, \theta_8, \theta_9, \theta_{10}, \theta_{11}, \theta_{12}, \theta_{13}) = (0, 4, -4.5, 4, -2.5, 3.5, -3.5, 3, -3, 4, -3, 4, -4)$;

(M14) f_t undergoes $q = 12$ change-points at $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5, \eta_6, \eta_7, \eta_8, \eta_9, \eta_{10}, \eta_{11}, \eta_{12}) = (50, 300, 500, 850, 960, 1000, 1200, 1380, 1480, 1600, 1750, 1900)$ with $n = 2000$ and $(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_7, \theta_8, \theta_9, \theta_{10}, \theta_{11}, \theta_{12}, \theta_{13}) = (0, 2, -2.5, 3, -2.5, 2.5, -2.5, 2, -2, 2.5, -3, 2, -2.5)$;

Table 3.10: Normalized absolute error (NAE) between true and estimated Long-Run Variance (LRV) of the sequences with signal (M7), where $n = 300$ and $q = 3$, and noises following an ARMA model (N1)-(N18) based on approaches (A1)-(A7), $\hat{\sigma}_2(\lambda)$ and $\hat{\sigma}_4(\lambda)$, which stand respectively for "Mean" and "Median" MODWT-based estimators. Here, both of the chosen scales are $4:5$.

	(A1)	(A2)	(A3)	(A4)	(A5)	(A6)	(A7)	$\hat{\sigma}_2(\lambda)$	$\hat{\sigma}_4(\lambda)$
$\text{AR}(1)^{SP}$	0.566	0.226	0.898	1.018	1.478	2.438	0.808	0.229	**0.190**
$\text{AR}(1)^{LP}$	0.234	0.460	**0.080**	0.225	0.922	0.602	0.111	0.197	0.411
$\text{AR}(1)^{SN}$	0.970	0.350	1.715	0.489	1.740	3.599	1.549	0.287	**0.229**
$\text{AR}(1)^{LN}$	3.161	0.504	6.550	1.345	6.034	10.115	6.106	0.307	**0.242**
$\text{AR}(2)^{P}$	0.436	**0.077**	0.867	0.145	2.601	1.756	0.742	**0.087**	0.288
$\text{AR}(2)^{N}$	2.726	0.357	5.764	3.558	13.436	17.301	5.522	0.298	**0.206**
Subtotal	2.206	1.113	3.560	1.877	6.741	8.395	3.210	**0.800**	1.118
Total	8.093	1.974	15.874	6.780	26.211	35.811	14.838	**1.405**	1.566
$\text{MA}(1)^{SP}$	0.584	0.219	0.953	0.431	1.446	2.315	0.844	**0.129**	0.278
$\text{MA}(1)^{LP}$	0.366	**0.170**	0.538	1.517	1.080	2.119	0.425	0.283	**0.167**
$\text{MA}(1)^{SN}$	1.077	0.339	1.868	0.724	1.909	3.925	1.666	**0.223**	**0.221**
$\text{MA}(1)^{LN}$	6.282	2.962	9.991	3.093	8.421	18.964	9.419	2.008	**0.914**
$\text{MA}(6)^{L}$	0.171	0.247	0.121	3.662	0.674	0.595	0.174	**0.097**	0.174
$\text{MA}(6)^{M}$	**0.075**	0.207	0.235	1.543	0.161	1.743	0.171	0.118	0.196
Subtotal	2.273	1.182	3.715	7.877	5.270	10.697	3.280	**0.850**	1.036
Total	8.555	4.144	13.706	10.970	13.691	29.661	12.699	2.858	**1.950**
$\text{ARMA}(1,1)^{1}$	**0.103**	0.145	0.121	3.801	0.754	1.149	0.063	0.182	0.179
$\text{ARMA}(1,1)^{2}$	2.330	1.179	3.555	1.931	3.740	7.178	3.279	0.675	**0.119**
$\text{ARMA}(1,1)^{3}$	0.641	0.231	1.056	0.541	1.621	2.805	0.949	0.292	**0.133**
$\text{ARMA}(1,1)^{4}$	13.144	4.583	24.100	3.406	20.183	37.627	22.579	2.793	**1.352**
$\text{ARMA}(2,6)^{1}$	0.157	0.324	**0.044**	0.719	2.035	0.398	0.100	0.104	0.304
$\text{ARMA}(2,6)^{2}$	0.257	0.170	0.652	0.544	2.664	3.073	0.555	**0.127**	0.367
Subtotal	1.158	0.870	1.873	5.605	7.074	7.425	1.667	**0.705**	0.983
Total	16.632	6.632	29.528	10.942	30.997	52.230	27.525	4.173	**2.454**

Table 3.11: Normalized absolute error (NAE) between true and estimated Long-Run Variance (LRV) of the sequences with signal (M8), where $n = 500$ and $q = 3$, and noises following an ARMA model (N1)-(N18) based on approaches (A1)-(A7), $\hat{\sigma}_2(\lambda)$ and $\hat{\sigma}_4(\lambda)$, which stand respectively for "Mean" and "Median" MODWT-based estimators. Here, both of the chosen scales are $4:5$.

	(A1)	(A2)	(A3)	(A4)	(A5)	(A6)	(A7)	$\hat{\sigma}_2(\lambda)$	$\hat{\sigma}_4(\lambda)$
$\text{AR}(1)^{SP}$	0.171	0.171	0.267	1.033	0.405	0.521	0.098	**0.075**	**0.077**
$\text{AR}(1)^{LP}$	0.353	0.407	0.290	0.709	0.985	**0.265**	0.443	0.387	0.359
$\text{AR}(1)^{SN}$	0.314	0.159	0.606	0.771	0.391	0.865	0.371	**0.065**	0.086
$\text{AR}(1)^{LN}$	1.588	0.680	3.090	0.304	1.209	2.254	2.267	0.172	**0.094**
$\text{AR}(2)^{P}$	0.081	0.134	0.202	0.637	1.216	0.109	**0.041**	0.125	0.134
$\text{AR}(2)^{N}$	1.698	1.065	2.810	0.371	13.026	4.299	2.267	0.119	**0.062**
Subtotal	0.919	0.871	1.365	3.150	2.997	1.760	0.953	**0.652**	**0.656**
Total	4.205	2.616	7.265	3.825	17.232	8.313	5.487	0.943	**0.812**
$\text{MA}(1)^{SP}$	0.584	0.219	0.953	0.431	1.446	2.315	**0.844**	0.129	0.278
$\text{MA}(1)^{LP}$	0.366	**0.170**	0.538	1.517	1.080	2.119	0.425	0.283	**0.167**
$\text{MA}(1)^{SN}$	1.077	0.339	1.868	0.724	1.909	3.925	1.666	**0.223**	**0.221**
$\text{MA}(1)^{LN}$	6.282	2.962	9.991	3.093	8.421	18.964	9.419	2.008	**0.914**
$\text{MA}(6)^{L}$	0.171	0.247	0.121	3.662	0.674	0.595	0.174	**0.097**	0.174
$\text{MA}(6)^{M}$	**0.075**	0.207	0.235	1.543	0.161	1.743	0.171	0.118	0.196
Subtotal	2.273	1.182	3.715	7.877	5.270	10.697	3.280	**0.850**	1.036
Total	8.555	4.144	13.706	10.970	13.691	29.661	12.699	2.858	**1.950**
$\text{ARMA}(1,1)^{1}$	0.077	0.170	**0.067**	2.460	0.411	0.437	0.172	0.127	**0.068**
$\text{ARMA}(1,1)^{2}$	1.102	0.728	1.640	0.735	1.360	2.278	1.243	0.369	**0.297**
$\text{ARMA}(1,1)^{3}$	0.166	0.122	0.321	0.960	0.487	0.803	0.171	**0.089**	0.105
$\text{ARMA}(1,1)^{4}$	6.787	3.557	12.605	3.222	6.672	10.549	9.482	2.094	**1.717**
$\text{ARMA}(2,6)^{1}$	0.257	0.255	0.252	2.240	1.636	0.473	0.410	0.263	**0.195**
$\text{ARMA}(2,6)^{2}$	**0.093**	0.232	**0.086**	0.368	2.898	0.633	0.104	0.230	0.219
Subtotal	**0.593**	0.779	0.726	6.028	5.432	2.346	0.857	0.709	**0.587**
Total	8.482	5.064	14.971	9.985	13.464	15.173	11.582	3.172	**2.601**

Table 3.12: Normalized absolute error (NAE) between true and estimated Long-Run Variance (LRV) of the sequences with signal (M9), where $n = 500$ and $q = 5$, and noises following an ARMA model (N1)-(N18) based on approaches (A1)-(A7), $\hat{\sigma}_2(\lambda)$ and $\hat{\sigma}_4(\lambda)$, which stand respectively for "Mean" and "Median" MODWT-based estimators. Here, both of the chosen scales are $4:5$.

	(A1)	(A2)	(A3)	(A4)	(A5)	(A6)	(A7)	$\hat{\sigma}_2(\lambda)$	$\hat{\sigma}_4(\lambda)$
$\text{AR}(1)^{SP}$	0.655	0.139	1.207	0.436	3.315	3.960	0.734	**0.089**	0.204
$\text{AR}(1)^{LP}$	0.107	0.345	**0.134**	0.215	3.549	1.335	0.151	0.317	0.384
$\text{AR}(1)^{SN}$	1.152	0.230	2.186	0.682	3.834	6.044	1.442	**0.059**	0.236
$\text{AR}(1)^{LN}$	3.918	0.739	8.076	0.765	17.391	17.087	5.792	0.204	**0.162**
$\text{AR}(2)^{P}$	0.553	**0.093**	1.147	0.175	6.409	3.229	0.650	**0.093**	0.268
$\text{AR}(2)^{N}$	3.843	1.198	7.144	0.344	49.041	29.256	5.239	**0.129**	0.194
Subtotal	2.467	0.807	4.674	1.508	17.107	14.568	2.977	**0.558**	1.092
Total	10.228	2.744	19.894	2.617	83.539	60.911	14.008	**0.891**	1.448
$\text{MA}(1)^{SP}$	0.665	0.165	1.240	0.656	3.269	4.124	0.788	**0.062**	0.227
$\text{MA}(1)^{LP}$	0.392	0.110	0.694	1.306	1.933	3.157	0.394	**0.063**	0.182
$\text{MA}(1)^{SN}$	1.189	0.251	2.300	0.399	3.775	6.241	1.523	**0.061**	0.208
$\text{MA}(1)^{LN}$	6.820	2.836	11.790	2.185	11.125	28.762	8.820	1.599	**0.879**
$\text{MA}(6)^{L}$	**0.065**	0.198	0.088	2.883	2.448	1.435	0.167	0.184	0.198
$\text{MA}(6)^{M}$	**0.145**	0.170	0.415	1.759	0.375	2.782	0.122	0.203	0.308
Subtotal	2.456	0.894	4.737	7.003	11.800	17.739	2.994	**0.573**	1.123
Total	9.276	3.730	16.527	9.188	22.925	46.501	11.814	2.172	**2.002**
$\text{ARMA}(1,1)^{1}$	**0.094**	0.116	0.205	3.277	1.488	1.974	0.050	0.138	0.196
$\text{ARMA}(1,1)^{2}$	2.465	0.910	4.286	0.893	6.564	10.945	3.078	0.448	**0.128**
$\text{ARMA}(1,1)^{3}$	0.729	0.121	1.371	0.360	3.163	4.473	0.854	**0.076**	0.285
$\text{ARMA}(1,1)^{4}$	14.841	3.888	29.392	5.041	42.656	60.478	21.494	2.192	**1.247**
$\text{ARMA}(2,6)^{1}$	**0.060**	0.207	0.123	1.498	4.489	0.868	0.140	0.200	0.254
$\text{ARMA}(2,6)^{2}$	0.435	**0.141**	1.017	0.343	10.864	5.992	0.541	**0.143**	0.310
Subtotal	1.318	0.585	2.716	5.478	20.004	13.307	1.585	**0.557**	1.045
Total	18.624	5.383	36.394	11.412	69.224	84.730	26.157	3.197	**2.420**

Table 3.13: Normalized absolute error (NAE) between true and estimated Long-Run Variance (LRV) of the sequences with signal (M10), where $n = 750$ and $q = 5$, and noises following an ARMA model (N1)-(N18) based on approaches (A1)-(A5), $\hat{\sigma}_2(\lambda)$ and $\hat{\sigma}_4(\lambda)$, which stand respectively for "Mean" and "Median" MODWT-based estimators. Here, both of the chosen scales are $4:5$.

	(A1)	(A2)	(A3)	(A4)	(A7)	$\hat{\sigma}_2(\lambda)$	$\hat{\sigma}_4(\lambda)$
$\text{AR}(1)^{SP}$	0.325	0.174	0.543	0.961	0.259	**0.104**	**0.096**
$\text{AR}(1)^{LP}$	0.196	0.260	**0.138**	1.570	0.340	0.220	0.239
$\text{AR}(1)^{SN}$	0.609	0.220	1.038	0.984	0.601	**0.125**	**0.123**
$\text{AR}(1)^{LN}$	2.276	0.632	4.413	2.453	3.017	0.262	**0.109**
$\text{AR}(2)^{P}$	0.232	0.133	0.438	0.752	0.170	**0.073**	0.172
$\text{AR}(2)^{N}$	2.181	0.777	3.879	2.418	2.554	0.202	**0.153**
Subtotal	1.362	0.787	2.157	4.267	1.370	**0.522**	0.630
Total	5.819	2.196	10.449	9.138	6.941	0.986	**0.892**
$\text{MA}(1)^{SP}$	0.330	0.116	0.601	0.887	0.339	0.109	**0.061**
$\text{MA}(1)^{LP}$	0.169	0.145	0.255	2.315	**0.082**	0.121	**0.082**
$\text{MA}(1)^{SN}$	0.632	0.209	1.086	0.858	0.672	**0.100**	0.135
$\text{MA}(1)^{LN}$	4.243	2.076	6.814	3.843	5.193	1.605	**0.854**
$\text{MA}(6)^{L}$	**0.058**	0.078	**0.049**	4.297	0.224	0.071	0.100
$\text{MA}(6)^{M}$	**0.063**	0.145	0.112	2.828	0.114	**0.075**	0.119
Subtotal	1.252	0.693	2.103	11.185	1.431	**0.476**	0.497
Total	5.495	2.769	8.917	15.028	6.624	2.081	**1.351**
$\text{ARMA}(1,1)^{1}$	**0.080**	0.147	**0.063**	4.109	0.113	0.091	0.121
$\text{ARMA}(1,1)^{2}$	1.396	0.674	2.241	1.088	1.636	0.468	**0.101**
$\text{ARMA}(1,1)^{3}$	0.362	0.143	0.568	1.381	0.337	0.158	**0.107**
$\text{ARMA}(1,1)^{4}$	9.120	3.235	16.978	17.343	12.272	2.119	**1.211**
$\text{ARMA}(2,6)^{1}$	0.096	0.149	**0.062**	3.437	0.270	**0.067**	**0.073**
$\text{ARMA}(2,6)^{2}$	0.186	0.083	0.367	1.119	0.051	**0.039**	0.182
Subtotal	0.724	0.522	1.060	10.046	0.771	**0.355**	0.483
Total	11.240	4.431	20.279	28.477	14.679	2.942	**1.795**

Table 3.14: Normalized absolute error (NAE) between true and estimated Long-Run Variance (LRV) of the sequences with signal (M11), where $n = 900$ and $q = 7$, and noises following an ARMA model (N1)-(N18) based on approaches (A1)-(A5), $\hat{\sigma}_2(\lambda)$ and $\hat{\sigma}_4(\lambda)$, which stand respectively for "Mean" and "Median" MODWT-based estimators. Here, both of the chosen scales are $4:6$.

	(A1)	(A2)	(A3)	(A4)	(A7)	$\hat{\sigma}_2(\lambda)$	$\hat{\sigma}_4(\lambda)$
$\text{AR}(1)^{SP}$	0.433	0.105	1.035	1.034	0.406	**0.076**	0.171
$\text{AR}(1)^{LP}$	**0.118**	0.336	**0.127**	0.801	0.245	0.214	0.268
$\text{AR}(1)^{SN}$	0.794	0.128	1.923	0.361	0.891	**0.077**	0.177
$\text{AR}(1)^{LN}$	2.742	0.423	7.265	3.267	3.943	**0.150**	**0.162**
$\text{AR}(2)^{P}$	0.386	0.099	1.005	0.421	0.336	**0.074**	0.196
$\text{AR}(2)^{N}$	2.364	0.277	6.389	8.628	3.575	**0.114**	0.192
Subtotal	1.731	0.668	4.090	2.617	1.878	**0.441**	0.812
Total	6.837	1.368	17.744	14.512	9.396	**0.705**	1.166
$\text{MA}(1)^{SP}$	0.501	0.107	1.151	0.912	0.459	**0.060**	0.142
$\text{MA}(1)^{LP}$	0.304	0.108	0.610	2.238	0.213	**0.068**	0.087
$\text{MA}(1)^{SN}$	0.890	0.178	2.071	0.488	0.977	**0.083**	0.151
$\text{MA}(1)^{LN}$	4.861	1.731	10.717	5.677	6.256	1.024	**0.586**
$\text{MA}(6)^{L}$	**0.053**	0.129	0.091	4.348	0.158	0.085	0.091
$\text{MA}(6)^{M}$	0.152	0.104	0.396	2.713	**0.042**	0.072	0.107
Subtotal	1.900	0.626	4.319	10.699	1.849	**0.368**	0.578
Total	6.761	2.357	15.036	16.376	8.105	1.392	**1.164**
$\text{ARMA}(1,1)^{1}$	0.079	0.108	0.209	3.631	**0.052**	0.075	0.088
$\text{ARMA}(1,1)^{2}$	1.726	0.517	3.784	1.047	2.051	0.239	**0.062**
$\text{ARMA}(1,1)^{3}$	0.522	0.138	1.188	0.804	0.510	**0.057**	0.144
$\text{ARMA}(1,1)^{4}$	10.546	2.547	26.529	11.028	15.285	1.532	**0.935**
$\text{ARMA}(2,6)^{1}$	**0.053**	0.154	0.138	3.329	0.172	0.089	0.114
$\text{ARMA}(2,6)^{2}$	0.278	0.117	0.864	1.048	0.238	**0.101**	0.232
Subtotal	0.932	0.517	2.399	8.812	0.972	**0.322**	0.578
Total	13.204	3.581	32.712	20.887	18.308	2.093	**1.575**

Table 3.15: Normalized absolute error (NAE) between true and estimated Long-Run Variance (LRV) of the sequences with signal (M12), where $n = 1200$ and $q = 7$, and noises following an ARMA model (N1)-(N18) based on approaches (A1)-(A5), $\hat{\sigma}_2(\lambda)$ and $\hat{\sigma}_4(\lambda)$, which stand respectively for "Mean" and "Median" MODWT-based estimators. Here, both of the chosen scales are $4:6$.

	(A1)	(A2)	(A3)	(A4)	(A7)	$\hat{\sigma}_2(\lambda)$	$\hat{\sigma}_4(\lambda)$
$\text{AR}(1)^{SP}$	0.429	0.096	0.998	0.721	0.390	**0.060**	0.139
$\text{AR}(1)^{LP}$	0.121	0.276	**0.102**	1.192	0.252	0.176	0.208
$\text{AR}(1)^{SN}$	0.747	0.125	1.826	0.570	0.843	**0.048**	0.154
$\text{AR}(1)^{LN}$	2.523	0.388	6.920	2.245	3.767	**0.126**	**0.131**
$\text{AR}(2)^{P}$	0.345	**0.085**	0.918	0.383	0.288	**0.086**	0.199
$\text{AR}(2)^{N}$	2.193	0.292	6.118	6.052	3.424	**0.121**	0.139
Subtotal	1.642	0.582	3.844	2.866	1.773	**0.370**	0.700
Total	6.358	1.262	16.882	11.163	8.964	**0.617**	0.970
$\text{MA}(1)^{SP}$	0.461	0.098	1.063	0.994	0.426	**0.056**	0.153
$\text{MA}(1)^{LP}$	0.282	0.077	0.555	2.496	0.183	**0.063**	**0.074**
$\text{MA}(1)^{SN}$	0.792	0.143	1.940	0.794	0.909	**0.063**	0.147
$\text{MA}(1)^{LN}$	4.869	1.978	10.270	4.571	6.013	1.038	**0.674**
$\text{MA}(6)^{L}$	**0.062**	0.130	**0.075**	3.804	0.176	0.112	0.108
$\text{MA}(6)^{M}$	0.140	0.109	0.366	2.662	**0.032**	0.085	0.120
Subtotal	1.737	0.557	3.999	10.750	1.726	**0.379**	0.602
Total	6.606	2.535	14.269	15.321	7.739	1.417	**1.276**
$\text{ARMA}(1,1)^{1}$	0.076	0.100	0.205	3.641	**0.049**	0.086	0.083
$\text{ARMA}(1,1)^{2}$	1.660	0.560	3.601	1.203	1.972	**0.260**	0.065
$\text{ARMA}(1,1)^{3}$	0.483	**0.102**	1.140	0.766	0.474	**0.091**	0.152
$\text{ARMA}(1,1)^{4}$	10.008	2.798	25.389	9.917	14.651	1.475	**0.987**
$\text{ARMA}(2,6)^{1}$	**0.040**	0.154	0.133	2.768	0.177	0.135	0.134
$\text{ARMA}(2,6)^{2}$	0.253	0.118	0.805	0.890	0.202	**0.137**	0.244
Subtotal	0.852	0.474	2.283	8.065	0.902	**0.449**	0.613
Total	12.520	3.832	31.273	19.185	17.525	2.184	**1.665**

Table 3.16: Normalized absolute error (NAE) between true and estimated Long-Run Variance (LRV) of the sequences with signal (M13), where $n = 1500$ and $q = 12$, and noises following an ARMA model (N1)-(N18) based on approaches (A1)-(A5), $\hat{\sigma}_2(\lambda)$ and $\hat{\sigma}_4(\lambda)$, which stand respectively for "Mean" and "Median" MODWT-based estimators. Here, both of the chosen scales are $4:6$.

	(A1)	(A2)	(A3)	(A4)	(A7)	$\hat{\sigma}_2(\lambda)$	$\hat{\sigma}_4(\lambda)$
$\text{AR}(1)^{SP}$	1.224	0.192	2.383	1.646	1.504	**0.067**	0.285
$\text{AR}(1)^{LP}$	0.283	0.189	0.766	0.951	0.300	**0.164**	0.377
$\text{AR}(1)^{SN}$	1.997	0.263	3.998	2.480	2.628	**0.053**	0.331
$\text{AR}(1)^{LN}$	6.527	0.653	13.707	7.646	9.494	**0.156**	0.310
$\text{AR}(2)^{P}$	1.170	0.114	2.381	0.854	1.463	**0.081**	0.337
$\text{AR}(2)^{N}$	5.900	0.763	12.068	10.186	8.314	**0.132**	0.331
Subtotal	4.674	0.758	9.528	5.931	5.895	**0.365**	1.330
Total	17.101	2.174	35.303	23.763	23.703	**0.653**	1.971
$\text{MA}(1)^{SP}$	1.321	0.231	2.536	1.600	1.614	**0.083**	0.271
$\text{MA}(1)^{LP}$	0.777	0.171	1.459	1.899	0.870	**0.082**	0.195
$\text{MA}(1)^{SN}$	2.083	0.267	4.163	2.968	2.755	**0.075**	0.323
$\text{MA}(1)^{LN}$	10.018	2.164	19.439	15.737	13.733	1.129	**0.415**
$\text{MA}(6)^{L}$	0.254	0.074	0.506	4.050	0.176	**0.058**	0.142
$\text{MA}(6)^{M}$	0.575	0.084	1.118	3.499	0.605	**0.062**	0.224
Subtotal	5.010	0.827	9.782	14.016	6.020	**0.360**	1.155
Total	15.028	2.991	29.221	29.753	19.753	**1.489**	1.570
$\text{ARMA}(1,1)^{1}$	0.442	0.095	0.789	2.326	0.405	**0.071**	0.161
$\text{ARMA}(1,1)^{2}$	3.731	0.738	7.275	3.044	4.997	0.295	**0.150**
$\text{ARMA}(1,1)^{3}$	1.345	0.246	2.593	1.647	1.665	**0.066**	0.272
$\text{ARMA}(1,1)^{4}$	23.228	3.212	47.812	26.779	33.911	1.667	**0.737**
$\text{ARMA}(2,6)^{1}$	0.340	**0.070**	0.717	1.628	0.296	0.084	0.254
$\text{ARMA}(2,6)^{2}$	1.030	0.099	2.132	1.913	1.279	**0.084**	0.379
Subtotal	3.157	0.510	6.231	7.514	3.645	**0.305**	1.066
Total	30.116	4.460	61.318	37.337	42.553	2.267	**1.953**

Table 3.17: Normalized absolute error (NAE) between true and estimated Long-Run Variance (LRV) of the sequences with signal (M14), where $n = 2000$ and $q = 12$, and noises following an ARMA model (N1)-(N18) based on approaches (A1)-(A5), $\hat{\sigma}_2(\lambda)$ and $\hat{\sigma}_4(\lambda)$, which stand respectively for "Mean" and "Median" MODWT-based estimators. Here, both of the chosen scales are $4:6$.

	(A1)	(A2)	(A3)	(A4)	(A7)	$\hat{\sigma}_2(\lambda)$	$\hat{\sigma}_4(\lambda)$
$\text{AR}(1)^{SP}$	0.730	0.133	1.450	1.110	0.730	**0.062**	0.162
$\text{AR}(1)^{LP}$	**0.067**	0.170	0.318	0.895	0.085	0.207	0.292
$\text{AR}(1)^{SN}$	1.207	0.190	2.487	1.251	1.365	**0.041**	0.180
$\text{AR}(1)^{LN}$	3.969	0.383	9.110	4.682	5.521	**0.123**	0.153
$\text{AR}(2)^{P}$	0.686	**0.086**	1.423	0.543	0.674	**0.073**	0.230
$\text{AR}(2)^{N}$	3.505	0.347	8.026	8.881	4.808	**0.108**	0.175
Subtotal	2.690	0.579	5.678	3.799	2.854	**0.383**	0.864
Total	10.164	1.309	22.814	17.362	13.183	**0.614**	1.192
$\text{MA}(1)^{SP}$	0.779	0.167	1.515	1.082	0.756	**0.047**	0.176
$\text{MA}(1)^{LP}$	0.461	0.127	0.829	2.037	0.367	**0.051**	0.116
$\text{MA}(1)^{SN}$	1.219	0.144	2.622	1.249	1.452	**0.044**	0.219
$\text{MA}(1)^{LN}$	6.573	1.782	13.204	9.932	8.344	1.023	**0.628**
$\text{MA}(6)^{L}$	0.102	**0.063**	0.228	4.545	**0.050**	0.072	0.077
$\text{MA}(6)^{M}$	0.301	**0.084**	0.596	2.928	0.178	**0.073**	0.156
Subtotal	2.862	0.585	5.790	11.841	2.803	**0.287**	0.744
Total	9.435	2.367	18.994	21.773	11.147	**1.310**	1.372
$\text{ARMA}(1,1)^{1}$	0.217	**0.056**	0.416	3.463	0.094	**0.054**	0.071
$\text{ARMA}(1,1)^{2}$	2.366	0.610	4.730	3.153	2.855	0.236	**0.033**
$\text{ARMA}(1,1)^{3}$	0.816	0.156	1.601	1.825	0.822	**0.046**	0.184
$\text{ARMA}(1,1)^{4}$	14.885	2.714	32.633	15.418	20.617	1.539	**1.000**
$\text{ARMA}(2,6)^{1}$	0.125	0.088	0.302	2.492	**0.038**	0.097	0.143
$\text{ARMA}(2,6)^{2}$	0.578	**0.070**	1.254	1.554	0.533	0.094	0.264
Subtotal	1.736	0.370	3.573	9.334	1.487	**0.291**	0.662
Total	18.987	3.694	40.936	27.905	24.959	2.066	**1.695**

3.7 Proofs

3.7.1 Proof of Theorem 3.1

Let $d_{j,k} = \mu_{j,k} + Z_{j,k}$, where $\mu_{j,k}$ and $Z_{j,k}$ represent the scaled signal and noise parts respectively. We have $\mu_{j,k} = 0$ for detailed coefficients in S_j^0. By applying Lemma 1, the definition of $d_{j,k}^2$ gives that $\mathbb{E}(d_{j,k}^2)$ follows

$$\begin{aligned}\mathbb{E}(d_{j,k}^2) &= \mathbb{E}[(\mu_{j,k} + Z_{j,k})^2] = \mu_{j,k}^2 + \mathbb{E}(Z_{j,k}^2) \\ &= \mu_{j,k}^2 + \frac{1}{2m_j}\,\mathbb{E}\left[\left(\sum_{i=(2k-1)m_j+1}^{2km_j} X_i \;-\; \sum_{i=(2k-2)m_j+1}^{(2k-1)m_j} X_i\right)^2\right] \\ &= \mu_{j,k}^2 + \sigma_*^2 + O\left(\frac{1}{2m_j}\right)\end{aligned} \tag{3.7.1}$$

Hence we have $\sigma_*^2 = \mathbb{E}(d_{j,k}^2) + O(\frac{1}{2m_j})$ for any $k \in S_j^0$. By Lemma 2, elementary calculations show that on the set $\mathcal{A}_n$,

$$\begin{aligned}|\hat{\sigma}_1^2(\lambda) - \sigma_*^2| &= \left|\frac{1}{b-a+1}\sum_{j=a}^{b}\frac{1}{K_j^*}\left(\sum_{k=1}^{K_j}\left[d_{j,k}^2\mathbb{1}_{(|d_{j,k}|\leqslant\lambda)} - \sigma_*^2\right] + (K_j - K_j^*)\sigma_*^2\right)\right| \\ &= \left|\frac{1}{b-a+1}\sum_{j=a}^{b}\frac{1}{K_j^*}\left[\left(\sum_{k\in S_j^0} + \sum_{k\in S_j^1}\right)\left[d_{j,k}^2\mathbb{1}_{(|d_{j,k}|\leqslant\lambda)} - \sigma_*^2\right] + (K_j - K_j^*)\sigma_*^2\right]\right| \\ &\leqslant \frac{1}{b-a+1}\sum_{j=a}^{b}\frac{1}{K_j^*}\left(\left|\sum_{k\in S_j^0}(d_{j,k}^2 - \sigma_*^2)\right| + \left|\sum_{k\in S_j^1}(d_{j,k}^2\mathbb{1}_{(|d_{j,k}|\leqslant\lambda)} - \sigma_*^2)\right|\right) \\ &\quad + \frac{1}{b-a+1}\sum_{j=a}^{b}\frac{1}{K_j^*}(K_j - K_j^*)\sigma_*^2 \\ &\leqslant \frac{1}{b-a+1}\sum_{j=a}^{b}\frac{1}{K_j^*}\left(\left|\sum_{k\in S_j^0}(Z_{j,k}^2 - \mathbb{E}(Z_{j,k}^2))\right| + O\left(\frac{|S_j^0|}{2m_j}\right)\right)\end{aligned}$$

$$+\frac{1}{b-a+1}\sum_{j=a}^{b}\frac{1}{K_j^*}|S_j^1|(\lambda^2+\sigma_*^2)+\frac{1}{b-a+1}\sum_{j=a}^{b}\frac{1}{K_j^*}(K_j-K_j^*)\sigma_*^2 \tag{3.7.2}$$

Here, since the error process $X_t = g(\ldots,\epsilon_{t-1},\epsilon_t)$ is assumed to be a physical system proposed in Wu (2005), $Z_{j,k}^2 - \mathbb{E}(Z_{j,k}^2)$ also follows the same system and we can write $D_{j,k} = Z_{j,k}^2 - \mathbb{E}(Z_{j,k}^2) = \hat{g}(\ldots,\epsilon_0,\epsilon_1,\ldots,\epsilon_{2km_j-1},\epsilon_{2km_j})$. Then, let $Z_{j,k}^* \coloneqq \frac{1}{\sqrt{2m_j}}\left(\sum_{i=(2k-1)m_j+1}^{2km_j} X_i^* - \sum_{i=(2k-2)m_j+1}^{(2k-1)m_j} X_i^*\right)$, $D_{j,k}^* = (Z_{j,k}^*)^2 - \mathbb{E}((Z_{j,k}^*)^2)$, $\Delta_{m,j,p}^D \coloneqq \sum_{k=m}^{\infty}\delta_{j,k,p}^D = \sum_{k=m}^{\infty}\|D_{j,k}-D_{j,k}^*\|_p$ and the DAN for the process $D_{j\cdot} = \{D_{j,k}\}_{k=-\infty}^{\infty}$,

$$\|D_{j\cdot}\|_{p,\upsilon} \coloneqq \sup_{m\geqslant 0}(m+1)^{\upsilon}\Delta_{m,p}^D = \sup_{m\geqslant 0}(m+1)^{\upsilon}\sum_{k=m}^{\infty}\delta_{k,p}^D,\ \upsilon\geqslant 0$$

By Hölder inequality, we know that

$$\begin{aligned}
&\|D_{j,k}-D_{j,k}^*\|_p \\
&= \|Z_{j,k}^2 - \mathbb{E}(Z_{j,k}^2) - (Z_{j,k}^*)^2 + \mathbb{E}((Z_{j,k}^*)^2)\|_p \\
&= \|Z_{j,k}^2 - (Z_{j,k}^*)^2\|_p \\
&= \frac{1}{2m_j}\left\|\left(\sum_{i=(2k-1)m_j+1}^{2km_j} X_i - \sum_{i=(2k-2)m_j+1}^{(2k-1)m_j} X_i\right)^2 - \left(\sum_{i=(2k-1)m_j+1}^{2km_j} X_i^* - \sum_{i=(2k-2)m_j+1}^{(2k-1)m_j} X_i^*\right)^2\right\|_p \\
&= \frac{1}{2m_j}\left\|\sum_{i=(2k-1)m_j+1}^{2km_j}(X_i-X_i^*) - \sum_{i=(2k-2)m_j+1}^{(2k-1)m_j}(X_i-X_i^*)\right\|_{2p} \\
&\quad\times\left\|\sum_{i=(2k-1)m_j+1}^{2km_j}(X_i+X_i^*) - \sum_{i=(2k-2)m_j+1}^{(2k-1)m_j}(X_i+X_i^*)\right\|_{2p} \\
&\leqslant \frac{1}{2m_j}\left(\sum_{i=(2k-2)m_j+1}^{2km_j}\|X_i-X_i^*\|_{2p}\right)\left\|\sum_{i=(2k-1)m_j+1}^{2km_j}(X_i+X_i^*) - \sum_{i=(2k-2)m_j+1}^{(2k-1)m_j}(X_i+X_i^*)\right\|_{2p}
\end{aligned} \tag{3.7.3}$$

By Theorem 1 (iii) in Wu et al. (2007), we then have

$$\|D_{j,k} - D^*_{j,k}\|_p \leqslant \frac{1}{2m_j}\left(\sum_{i=(2k-2)m_j+1}^{2km_j} \|X_i - X_i^*\|_{2p}\right)\frac{2pB_{2p}}{2p-1}n^{1/2}\Theta_{0,2p} \tag{3.7.4}$$

Therefore, under the assumption $\|X_\cdot\|_{p,\upsilon} = \sup_{m\geqslant 0}(m+1)^\upsilon \sum_{i=m}^\infty \|X_i - X_i^*\|_p < \infty$, we have $\|D_{j\cdot}\|_{p,\upsilon} = \sup_{m\geqslant 0}(m+1)^\upsilon \lim_{n\to\infty}\sum_{k=m}^{K_j}\|D_{j,k} - D^*_{j,k}\|_p \leqslant \sup_{m\geqslant 0}(m+1)^\upsilon \lim_{n\to\infty}\sum_{k=m}^{K_j}\frac{1}{2m_j}\sum_{i=(2k-2)m_j}^{2km_j-1}\|X_i - X_i^*\|_{2p}\frac{2pB_{2p}}{2p-1}n^{1/2}\Theta_{0,2p} = \sup_{m\geqslant 0}(m+1)^\upsilon \lim_{n\to\infty}\sum_{i=(2m-2)m_j}^{n-1}\|X_i - X_i^*\|_{2p}\frac{2pB_{2p}}{2p-1}\, n^{1/2-\nu}\Theta_{0,2p} = \sup_{m\geqslant 0}\lim_{n\to\infty}((2m-2)m_j+1)^\upsilon\sum_{i=(2m-2)m_j}^{n-1}\|X_i - X_i^*\|_{2p}\frac{2pB_{2p}}{2p-1}n^{1/2-(\upsilon+1)\nu}\Theta_{0,2p} < \infty$ when $\nu \geqslant \frac{1}{2(\upsilon+1)}$.

Let $j = \nu\log_2(n)$. Due to the finite number of change-points, we have $|S_j^0| = O(n^{1-\nu})$. When $\|D_{j\cdot}\|_{p,\upsilon} < \infty$, where $p > 2$ and $\upsilon > 1/2 - 1/p$, it follows from Wu and Wu (2016) that for any $k \in S_j^0$, there exist some $\tilde{a}$, for all $\tilde{\delta} > 0$,

$$\begin{aligned} P\left(\left|\sum_{k\in S_j^0}(Z_{j,k}^2 - \mathbb{E}(Z_{j,k}^2))\right| > n^{\tilde{a}}\tilde{\delta}\right) &= P\left(\left|\sum_{k\in S_j^0} D_{j,k}\right| > n^{\tilde{a}}\tilde{\delta}\right) \\ &\leqslant \tilde{C}_1\frac{|S_j^0|\|D_{j\cdot}\|_{p,\upsilon}^p}{(n^{\tilde{a}}\tilde{\delta})^p} + \tilde{C}_2\exp\left(-\frac{\tilde{C}_3|(n^{\tilde{a}}\tilde{\delta})^2}{|S_j^0|\|D_{j\cdot}\|_{2,\nu}^2}\right) \\ &= O_p(n^{1-\nu-p\tilde{a}} + \exp(-\tilde{C}_4 n^{2\tilde{a}-1+\nu})) \end{aligned} \tag{3.7.5}$$

where $\tilde{C}_4$ is a positive constant that depend of p, υ and the dependence condition $\|D_{j\cdot}\|_{2,\nu}$. It tells us when $\tilde{a} > (1-\nu)/2$, the term above converges to 0 as $n \to \infty$. We can then get that $\sum_{k\in S_j^0}(Z_{j,k}^2 - \mathbb{E}(Z_{j,k}^2)) = o_p(n^{\tilde{a}})$. Also, Lemma 2 tells us $K_j/K_j^* - 1 = o_p(n^{-1+\nu_0})$ when $\nu_0 > \nu$. It gives $1/|K_j^*| = O_p(n^{-1+\nu})$. When $\lambda = O(\sqrt{\log(n)})$, it can be obtained that

$$|\hat{\sigma}_1^2(\lambda) - \sigma_*^2| = \frac{1}{b-a+1}\sum_{j=a}^{b} O_p(n^{-1+\nu})\left(o_p(n^{\tilde{a}}) + O(n^{1-2\nu})\right)$$

$$
\begin{aligned}
&+ \frac{1}{b-a+1}\sum_{j=a}^{b} o_p(n^{-1+\nu}\log(n)) + \frac{1}{b-a+1}\sum_{j=a}^{b} o_p(n^{-1+\nu_0}) \\
&= O_p\{\log_2(n)^{-1}(n^{-1+\beta}n^{\tilde{a}} + n^{-\alpha} + n^{-1+\nu_0})\} \\
&= O_p\{(n^{-1+\beta+\tilde{a}} + n^{-\alpha})\log_2(n)^{-1}\} \qquad (3.7.6)
\end{aligned}
$$

where $\tilde{a} > (1-\beta)/2$. Similarly, when $p > 2$ and $\upsilon < 1/2 - 1/p$, it follows that for any $k \in S_j^0$, there exist some $\tilde{a}$, for all $\tilde{\delta} > 0$,

$$
P\left(\left|\sum_{k\in S_j^0}(Z_{j,k}^2 - \mathbb{E}(Z_{j,k}^2))\right| > n^{\tilde{a}}\tilde{\delta}\right) = O_p(n^{(1-\nu)(p/2-p\upsilon)-p\tilde{a}} + \exp(-\tilde{C}_4 n^{2\tilde{a}-1+\nu})) \qquad (3.7.7)
$$

Here we can get $\sum_{k\in S_j^0}(Z_{j,k}^2 - \mathbb{E}(Z_{j,k}^2)) = o_p(n^{\tilde{a}})$ when $\tilde{a} > (1-\nu)(1/2-1/\upsilon)$ and hence we have (3.7.6) when $\tilde{a} > (1-\beta)(1/2-1/\upsilon)$. On the other hand, we know that

$$
\mu_{j,k}^2 = \frac{1}{2m_j}\left[\sum_{i=(2k-1)m_j+1}^{2km_j} f\left(\frac{i}{n}\right) - \sum_{i=(2k-2)m_j+1}^{(2k-1)m_j} f\left(\frac{i}{n}\right)\right]^2 \qquad (3.7.8)
$$

which indicates $\mu_{j,k}^2 = \frac{1}{2m_j}O(m_j^2) = O(\frac{m_j}{2})$ uniformly over $k = 1, 2, \ldots, K_j$. Given the fact that f_t is piecewise constant with finite change-points, the set $\{k \in \{1, 2, \ldots, K_j\}|$ $\mu_{j,k} \neq 0\}$ contains a finite number of elements, independently of $n \in \mathbb{N}$. It then follows that $\frac{1}{b-a+1}\sum_{j=a}^{b}\frac{1}{K_j}\sum_{k=1}^{K_j}\mu_{j,k}^2 = \frac{1}{b-a+1}\sum_{j=a}^{b}\frac{1}{K_j}O(\frac{m_j}{2}) = O(n^{-1+2\beta}\{\log_2(n)\}^{-1})$. By $\sum_{k\in S_j^0}(Z_{j,k}^2 - \mathbb{E}(Z_{j,k}^2)) = o_p(n^{\tilde{a}})$, we can then similarly derive that

$$
\begin{aligned}
|\hat{\sigma}_1^2 - \sigma_*^2| &= \left|\frac{1}{b-a+1}\sum_{j=a}^{b}\frac{1}{K_j}\sum_{k=1}^{K_j}\left(d_{j,k}^2 - \sigma_*^2\right)\right| \\
&= \left|\frac{1}{b-a+1}\sum_{j=a}^{b}\frac{1}{K_j}\sum_{k=1}^{K_j}\left(Z_{j,k}^2 + 2Z_{j,k}\mu_{j,k} + \mu_{j,k}^2 - \sigma_*^2\right)\right| \\
&\leqslant \frac{1}{b-a+1}\sum_{j=a}^{b}\frac{1}{K_j}\left(\left|\sum_{k=1}^{K_j}(Z_{j,k}^2 - \sigma_*^2)\right| + \left|\sum_{k\in S_j^1} 2Z_{j,k}\mu_{j,k}\right| + \left|\sum_{k\in S_j^1}\mu_{j,k}^2\right|\right)
\end{aligned}
$$

$$
\begin{aligned}
&\leqslant \frac{1}{b-a+1}\sum_{j=a}^{b}\frac{1}{K_j}\left(\left|\sum_{k=1}^{K_j}(Z_{j,k}^2-\mathbb{E}(Z_{j,k}^2))\right|+O\left(\frac{K_j}{2m_j}\right)\right)\\
&\quad+\frac{1}{b-a+1}\sum_{j=a}^{b}\frac{|S_j^1|}{K_j}O_p(\log(n)^{1/2})*O(n^{\nu/2})+\frac{1}{b-a+1}\sum_{j=a}^{b}\frac{1}{K_j}\sum_{k\in S_j^1}\mu_{j,k}^2\\
&=O_p\{(n^{-1+\beta+\tilde{\alpha}}+n^{-\alpha}+n^{-1+2\beta})\log_2(n)^{-1}\}
\end{aligned}
\tag{3.7.9}
$$

3.7.2 Proof of Theorem 3.2

Let $\sqrt{2m_j}\tilde{W}_{j,k}=\tilde{\mu}_{j,k}+\tilde{Z}_{j,k}$, where $\tilde{\mu}_{j,k}$ and $\tilde{Z}_{j,k}$ represent the scaled signal and noise parts respectively. We have $\tilde{\mu}_{j,k}=0$ for detailed coefficients in $\tilde{S}_j^0$. By applying Lemma 1, the definition of $\tilde{W}_{j,k}^2$ gives that $2m_j\mathbb{E}(\tilde{W}_{j,k}^2)$ follows

$$
\begin{aligned}
2m_j\mathbb{E}(\tilde{W}_{j,k}^2)&=\mathbb{E}[(\tilde{\mu}_{j,k}+\tilde{\mu}_{j,k})^2]=\tilde{\mu}_{j,k}^2+\mathbb{E}(\tilde{Z}_{j,k}^2)\\
&=\tilde{\mu}_{j,k}^2+\sigma_*^2+O\left(\frac{1}{2m_j}\right)
\end{aligned}
\tag{3.7.10}
$$

Hence we also have $\sigma_*^2=2m_j\mathbb{E}(\tilde{W}_{j,k}^2)+O(\frac{1}{2m_j})$ for any $k\in\tilde{S}_j^0$. Elementary calculations show that

$$
\begin{aligned}
|\hat{\sigma}_2^2(\lambda)-\sigma_*^2|&=\left|\frac{1}{b-a+1}\sum_{j=a}^{b}\frac{1}{T_j^*}\left(\sum_{k=1}^{T_j}\left[2m_j\tilde{W}_{j,k}^2\mathbb{1}_{(|\sqrt{2m_j}\tilde{W}_{j,k}|\leqslant\lambda)}-\sigma_*^2\right]+(T_j-T_j^*)\sigma_*^2\right)\right|\\
&\leqslant\frac{1}{b-a+1}\sum_{j=a}^{b}\frac{1}{T_j^*}\left(\left|\sum_{k\in\tilde{S}_j^0}(\tilde{Z}_{j,k}^2-\mathbb{E}(\tilde{Z}_{j,k}^2))\right|+O\left(\frac{|\tilde{S}_j^0|}{2m_j}\right)\right)\\
&\quad+\frac{1}{b-a+1}\sum_{j=a}^{b}\frac{1}{T_j^*}|\tilde{S}_j^1|(\lambda^2+\sigma_*^2)+\frac{1}{b-a+1}\sum_{j=a}^{b}\frac{1}{T_j^*}(T_j-T_j^*)\sigma_*^2
\end{aligned}
\tag{3.7.11}
$$

When $\|\tilde{D}_j.\|_{p,\upsilon}<\infty$, where $p>2$ and $\upsilon>1/2-1/p$, it follows from Wu and Wu

(2016) that for any $k \in \tilde{S}_j^0$, there exist some $\tilde{a}$, for all $\tilde{\delta} > 0$,

$$
\begin{aligned}
P\left(\left|\sum_{k\in\tilde{S}_j^0}(\tilde{Z}_{j,k}^2 - \mathbb{E}(\tilde{Z}_{j,k}^2))\right| > n^{\tilde{a}}\tilde{\delta}\right) &= P\left(\left|\sum_{k\in\tilde{S}_j^0}\tilde{D}_{j,k}\right| > n^{\tilde{a}}\tilde{\delta}\right) \\
&\leqslant \tilde{C}_1\frac{|\tilde{S}_j^0|\|\tilde{D}_{j\cdot}\|_{p,\upsilon}^p}{(n^{\tilde{a}}\tilde{\delta})^p} + \tilde{C}_2\exp\left(-\frac{\tilde{C}_3|(n^{\tilde{a}}\tilde{\delta})^2}{|\tilde{S}_j^0|\|\tilde{D}_{j\cdot}\|_{2,\nu}^2}\right) \\
&= O_p(n^{1-p\tilde{a}} + \exp(-\tilde{C}_4 n^{2\tilde{a}-1})) \qquad (3.7.12)
\end{aligned}
$$

This tells us that when $\tilde{a} > 1/2$, the term above converges to 0 as $n \to \infty$. We can then get that $\sum_{k\in\tilde{S}_j^0}(\tilde{Z}_{j,k}^2 - \mathbb{E}(\tilde{Z}_{j,k}^2)) = o_p(n^{\tilde{a}})$. Similarly, when $p > 2$ and $\upsilon < 1/2 - 1/p$, it follows that for any $k \in \tilde{S}_j^0$, there exist some $\tilde{a}$, for all $\tilde{\delta} > 0$,

$$
P\left(\left|\sum_{k\in\tilde{S}_j^0}(\tilde{Z}_{j,k}^2 - \mathbb{E}(\tilde{Z}_{j,k}^2))\right| > n^{\tilde{a}}\tilde{\delta}\right) = O_p(n^{p/2-p\upsilon-p\tilde{a}} + \exp(-\tilde{C}_4 n^{2\tilde{a}-1+\nu})) \qquad (3.7.13)
$$

Here we can also get $\sum_{k\in\tilde{S}_j^0}(\tilde{Z}_{j,k}^2 - \mathbb{E}(\tilde{Z}_{j,k}^2)) = o_p(n^{\tilde{a}})$ when $\tilde{a} > 1/2$. Also, Lemma 2 tells us $T_j/T_j^* - 1 = o_p(n^{-1+\nu_0})$ when $\nu_0 > \nu$. In addition, we know $1/T_j^* = O_p(n^{-1})$. When $\lambda = O(\sqrt{\log(n)})$, it can be obtained that

$$
\begin{aligned}
|\hat{\sigma}_2^2(\lambda) - \sigma_*^2| &= \frac{1}{b-a+1}\sum_{j=a}^{b} O_p(n^{-1})\left(o_p(n^{\tilde{a}}) + O(n^{1-\nu})\right) \\
&\quad + \frac{1}{b-a+1}\sum_{j=a}^{b} o_p(n^{-1+\nu}\log(n)) + \frac{1}{b-a+1}\sum_{j=a}^{b} o_p(n^{-1+\nu_0}) \qquad (3.7.14) \\
&= O_p\{\log_2(n)^{-1}(n^{-1+\tilde{a}} + n^{-\alpha} + n^{-1+\nu_0})\}
\end{aligned}
$$

where $\nu_0 > \beta$ and $\tilde{a} > 1/2$. In addition, write $\frac{1}{b-a+1}\sum_{j=a}^{b}\frac{1}{T_j}\sum_{k=1}^{T_j}\tilde{\mu}_{j,k}^2 = \frac{1}{b-a+1}\sum_{j=a}^{b}$

$\frac{2m_j}{T_j} O(\frac{m_j}{2}) = O(n^{-1+2\beta}\{\log_2(n)\}^{-1})$. Using lemma 3, we can similarly derive that

$$
\begin{aligned}
|\hat{\sigma}_2^2 - \sigma_*^2| &= \left| \frac{1}{b-a+1} \sum_{j=a}^{b} \frac{1}{T_j} \sum_{k=1}^{T_j} \left[2m_j \tilde{W}_{j,k}^2 - \sigma_*^2 \right] \right| \\
&= \left| \frac{1}{b-a+1} \sum_{j=a}^{b} \frac{1}{T_j} \sum_{k=1}^{T_j} \left(\tilde{Z}_{j,k}^2 + 2\tilde{Z}_{j,k}\tilde{\mu}_{j,k} + \tilde{\mu}_{j,k}^2 - \sigma_*^2 \right) \right| \\
&\leqslant \frac{1}{b-a+1} \sum_{j=a}^{b} \frac{1}{T_j} \left(\left| \sum_{k=1}^{T_j} (\tilde{Z}_{j,k}^2 - \sigma_*^2) \right| + \left| \sum_{k \in \tilde{S}_j^1} 2\tilde{Z}_{j,k}\tilde{\mu}_{j,k} \right| + \left| \sum_{k \in \tilde{S}_j^1} \tilde{\mu}_{j,k}^2 \right| \right) \\
&\leqslant \frac{1}{b-a+1} \sum_{j=a}^{b} \frac{1}{T_j} \left(\left| \sum_{k=1}^{T_j} (\tilde{Z}_{j,k}^2 - \mathbb{E}(\tilde{Z}_{j,k}^2)) \right| + O\left(\frac{T_j}{2m_j} \right) \right) \\
&\quad + \frac{1}{b-a+1} \sum_{j=a}^{b} \frac{|\tilde{S}_j^1|}{T_j} O_p(\log(n)^{1/2}) * O(n^{\nu/2}) + \frac{1}{b-a+1} \sum_{j=a}^{b} \frac{1}{T_j} \sum_{k \in \tilde{S}_j^1} \tilde{\mu}_{j,k}^2 \\
&= O_p\{(n^{-1+\tilde{a}} + n^{-\alpha} + n^{-1+2\beta}) \log_2(n)^{-1}\}
\end{aligned}
\tag{3.7.15}
$$

Chapter 4

Aspects of model selection for nonstationary time series with level change

4.1 Introduction

In this chapter, our focus is the application of *Narrowest-Over-Threshold (NOT) algorithm* for dependent data generated with piecewise-constant signal. The NOT algorithm, together with NOT solution path algorithm, was introduced in Baranowski et al. (2019) to identify the number and locations of multiple changes in the uni-variate statistical model

$$Y_t = f_t + X_t, \ i = 1, 2, \ldots, n \tag{4.1.1}$$

where the unknown deterministic signal f_t can show some regularity across time t, especially "features" like jumps or kinks, and $\{X_i\}_{i=1}^n$ represents the independent Gaussian error process exactly or approximately centred at zero. Generally speaking, the most

valuable contribution of the NOT approach arises from the idea of considering the observations $\{Y_i\}_{i=1}^n$ by concentrating on narrowest candidate intervals, which make it effective for detection problems with more general features, besides the change-points in the piecewise-constant signal. Also, the limited number of randomly drawn intervals M and the proposed threshold-indexed solution path lead to the low computational complexity of this algorithm.

Due to the aforementioned benefits, it would be indeed valuable to find possible extensions of the NOT algorithm for serial correlated data. Therefore, the present chapter is devoted to the analysis of NOT in dependent time series starting with that produced by piecewise-constant signal under more general assumptions on the error process, see detailed definitions of signal (3.1.1) and noise (3.1.2) in Chapter 3.

As discussed in Section 4.1 in Baranowski et al. (2019), although there might be information loss, the NOT algorithm can still be utilised as a quasi-likelihood-type procedure for dependent data. The consistency of the basic NOT algorithm can still be proved if the assumption $X_t \overset{iid}{\sim} N(0, \sigma_t^2)$ is replaced by a short-memory stationary error process, where σ_t represents the standard deviation of noise at time t; see Corollary 1 and 2 in Baranowski et al. (2019). For the dependent error setting, the simplest extension the can be made on the basic NOT algorithm is to instead choose the threshold ζ_n relying on LRV estimators. Its results will be presented first in the next section.

Also, some numerical results for NOT solution path algorithm are displayed in Section E of the online supplementary material (Baranowski et al., 2019). However, it seems that the algorithm tends not to work so well for dependent data, the main reason being that the defined strengthened Schwarz Information Criterion is no longer well-suited in the presence of serial correlated noise. Therefore, in the following sections, we shall experiment on several practical ways to overcome this issue.

- Preprocessing data to eliminate the dependence and increase its similarity to a Gaussian series (see Baranowski et al. (2019)), which is discussed in Section 4.3.

  (1). Add an additional zero-mean independent Gaussian error process to the original data, hoping that the new series will be closer to Gaussian and the serial correlated noise can be reduced.

  (2). Divide the time series into $\lfloor T/h \rfloor$ blocks of length h and regard the calculated local averages as the new dataset, with a similar expectation that the pre-averaged noise can approximately follow a Gaussian distribution by the law of large numbers and become less dependent.

- Modifying the sSIC to make it more suitable for dependent data, see Section 4.4.

  (1). Directly replace the maximum likelihood estimator of the residual variance with segmented LRV estimators or the sum of the estimated autocovariances; the hope is that the substituted estimators can describe the information of both dependence and the candidate change-points.

  (2). Choose an appropriate constant α in the original sSIC based on the strength of autocorrelation and signal-to-noise ratio obtained from data; following this, we can see how the independent NOT algorithm works for dependent data and when it breaks down.

In order to check the potential improvement in performance, we first analyse the behaviour of different NOT algorithms using the same simulation models introduced in the online supplementary materials of Baranowski et al. (2019):

(M1) **teeth:** piecewise-constant f_t, $n = 512$, $q = 7$ change-points at $t = 64, 128, ..., 448$, with the corresponding jump sizes $-2, 2, -2, ..., -2$, starting intercept $f_1 = 1$ for $t = 1, \ldots, n$.

(M2) **blocks**: piecewise-constant f_t, $n = 2024$, $q = 11$ change-points at $t = 205, 267, 308, 472, 512, 820, 902, 1332, 1557, 1598, 1659$, with the corresponding jump sizes $1.464, -1.830, 1.098, -1.464, 1.830, -1.537, 0.768, 1.574, -1.135, 0.769, -1.537$, starting intercept $f_1 = 0$ for $t = 1, ..., n$.

Besides (M1) and (M2), we also simulate more examples under various scenarios and the description of the full simulation models (from twelve scenarios with different sample size, number and location of change-points and serial dependence features) is deferred to Appendix 4.5

In general, this chapter is aimed at discussing the possible improvements we can make on the original NOT or NOT solution path algorithm. The main contribution is as follows. First, Section 4.2 discusses the performance of the basic NOT algorithm by choosing the preset threshold ζ_n on the estimated LRV estimators. Since the basic NOT highly depends on the choice of threshold and its performance is roughly moderate even for independent noise, we attach more importance to the NOT solution path algorithm. In Section 4.3, we illustrate the effectiveness of two data preprocessing approaches applied before the NOT solution path algorithm, while freezing all prespecified parameters in the algorithm. We also consider extending the NOT solution path algorithm itself in Section 4.4 via changing the information criterion or parameters in the penalty function. The Supplementary Appendix 4.5 contains complete simulation models.

4.2 Threshold-based NOT Algorithm

4.2.1 Choice of Threshold

For a stationary error process $\{X_i\}_{i=1}^n$, the autocorrelation function $\rho_t(\tau)$ depends only on the difference τ and hence we write $\rho_t(\tau) = \rho(\tau)$ for any lag τ. As shown in Corollary 1 in Baranowski et al. (2019), the basic NOT algorithm still works when the dependent Gaussian process $\{X_i\}_{i=1}^n$ satisfies $\sum_{\tau=-\infty}^{\infty} |\rho(\tau)| < \infty$. In particular, the ARMA models can satisfy this condition because we know that they can be represented as a linear process with coefficients $|a_i| = O(r^i)$ for all $r \in (\lambda_*, 1)$, see Remark 3.1 in Chapter 3. This means that $\sum_{\tau=-\infty}^{\infty} |\rho(\tau)| = 1 + 2\sum_{\tau=1}^{\infty} |\rho(\tau)| = 1 + 2\sum_{i=0}^{\infty}\sum_{j=i+1}^{\infty} |a_i a_j| / \sum_{i=0}^{\infty} a_i^2 < \infty$. Here, we shall focus on the ARMA model when constructing the serial correlated noise.

For independent noise $X_i \sim N(0, \sigma^2)$, the basic NOT algorithm employs a pre-specified threshold ζ_n proportional to the the standard deviation σ. Its MAD-based estimator (Hampel, 1974) applied in NOT is defined as $\hat{\sigma} = \text{Median}(|Y_2 - Y_1|, \ldots, |Y_n - Y_{n-1}|)/(\sqrt{2} z_{1/4})$, where $z_{1/4}$ is the third quartile of the standard normal distribution. Under the new dependence assumption, the threshold ζ_n is permitted to be proportional to $\sigma\sqrt{\sum_{\tau=-\infty}^{\infty} |\rho(\tau)|}$. For positively correlated noise, this term can naturally be regarded as long-run standard deviation and hence it could be estimated relying on our new LRV estimators.

4.2.2 Simulation Results

In practice, the success of NOT is largely impacted by the selection of the pre-specified constants in ζ_n, i.e. the constant C_{not} in $\zeta_n = C_{not}\hat{\sigma}\sqrt{2\log(n)}$ for NOT and the constant

C_{lrv} in $\zeta_n = C_{lrv}\hat{\sigma}_2(\lambda)\sqrt{2\log(n)}$ for the possible extended NOT.

Therefore, we first consider the simplest examples, data with independent noise, to see whether our LRV estimators can reduce the impacts of the unobserved quantities. In other words, we choose proper constants C_{not} and C_{lrv} based on simulated data with signal (M1) and will not change them when we detect change-points in data with signal (M2). To be specific, we compare the performance of NOT and extended NOT ('NOT LR') using the test signals (M1) and (M2) introduced in Section 4.1, accompanying the error process following (a) $X_t \overset{iid}{\sim} N(0,1)$ and (b) $X_t \overset{iid}{\sim} N(0,2)$; here results for all cases are summarised in Table 4.1.

For each case, we show a frequency table for the distribution of $\hat{q} - q$, where $\hat{q}$ and q denote the number of the estimated and true change-points respectively. We also provide the estimated Mean Squared Error of the estimated signal $\hat{f}_t$, which is defined as

$$\text{MSE} = \mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n}(f_t - \hat{f}_t)^2\right]$$

In the NOT algorithm, the values of the estimated piecewise-constant mean of the vector are computed by taking sample means of the sequence between each pair of consecutive detected candidate change-points. Meanwhile, to assess the accuracy of estimating the locations of change-points, we calculate the estimates of the *(scaled) Hausdorff distance* defined below, which follows $0 \leqslant d_H \leqslant 1$

$$d_H = \frac{1}{n}\mathbb{E}\left[\max\left\{\max_{j=0,\ldots,q+1}\min_{k=0,\ldots,\hat{q}+1}|\eta_j - \hat{\eta}_k|, \max_{k=0,\ldots,\hat{q}+1}\min_{j=0,\ldots,q+1}|\eta_j - \hat{\eta}_k|\right\}\right] \tag{4.2.1}$$

where the true and estimated locations of change-points respectively satisfy the constraints $0 = \eta_0 < \eta_1 < \cdots < \eta_q < \eta_{q+1} = n$ and $0 = \hat{\eta}_0 < \hat{\eta}_1 < \cdots < \hat{\eta}_{\hat{q}} < \hat{\eta}_{\hat{q}+1} = n$.

In Table 4.1, we observe that based on an optimal choice of threshold, utilising the

Table 4.1: Distribution of $\hat{q}-q$ obtained by NOT and NOT LR for data generated according to (2.2.1) with the signals (M1) and (M2), together with the noise $X_t \overset{iid}{\sim} N(0,1)$ and $N(0,2)$, the average Mean Square Error of the resulting estimate of the signal and the average Hausdorff distance d_H over 100 simulations.

			$\hat{q}-q$								
Method	Signal	Noise	$\leqslant -3$	-2	-1	0	1	2	$\geqslant 3$	MSE	d_H
NOT	(M1)	$N(0,1)$	0	0	9	72	15	4	0	0.984	0.125
NOT LR			0	0	16	82	2	0	0	0.999	0.135
NOT		$N(0,2)$	0	0	8	80	10	2	0	1.968	0.139
NOT LR			0	0	12	82	6	0	0	1.973	0.140
NOT	(M2)	$N(0,1)$	2	23	57	17	1	0	0	0.995	0.441
NOT LR			0	8	59	25	6	2	0	0.991	0.498
NOT		$N(0,2)$	51	30	19	0	0	0	0	2.004	0.845
NOT LR			20	30	30	15	4	1	0	1.988	0.821

new LRV estimator can slightly enhance the performance of estimating the number of change-points while it leads to a slight increase in MSE and Hausdorff distance. However, the last four rows show us that without any adaptation, the two algorithms using the same constants C_{not} and C_{lrv} are not able to give good estimation results in the case of irregular change-points even for independent noise. Therefore, finding a mechanism for choosing the optimal threshold is indeed vital for the performance of NOT. In practice, while Theorem 1 and Corollary 1 in Baranowski et al. (2019) assert the existence of threshold ζ_n that ensures consistent estimation of the change-points in dependent data, determining this parameter in experiments often relies on many unobserved quantities. Generally speaking, the NOT algorithm tends to be sensitive to the choice of threshold even under the independent case.

To overcome this issue, Baranowski et al. (2019) proposed the NOT solution path algorithm that can automatically select the optimal threshold and the corresponding candidate model from a threshold-indexed solution path by minimising the sSIC, see Section 2.5. To take full advantage of this existing flexible CPD technique, in the following sections, we shall pay more attention to the NOT solution path algorithm.

4.3 Data Preprocessing

In the development of NOT solution path algorithm, the most straightforward way is to conduct a data preprocessing procedure first to make the tested sequence closer to an independent Gaussian distributed series. Following the aforementioned data preprocessing approaches, we now discuss how the NOT solution path algorithm can be extended in practice to handle serial correlated error processes. In this section, we tend to focus on analysing the practical behaviour of the algorithm on the new series, with the readily available function in **R** package "**breakfast**" for Gaussian mean shift model. To avoid the possible estimation error resulting from the noise type, we conduct the analysis under the assumption of a Gaussian distributed error process.

4.3.1 Adding zero-mean *iid* Gaussian distributed error process

Here we consider adding an additional series $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ to the original model $Y_i = f_i + X_i$, where f_i is piecewise-constant and X_i follows an AR model $X_i = \sum_{j=1}^{p} \phi_j X_{i-j} + \epsilon_i$ with $\epsilon_i \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$. This method is applied in the hope that the suitably chosen *iid* Gaussian noise ε_i can lead to weaker serial correlation within the new observations $Y_i + \varepsilon_i$. We motivate this idea with the following example:

Example 1. (a) Signal (M1), noise $X_i = 0.3X_{i-1} + \epsilon_i$ with $\epsilon_i \overset{iid}{\sim} N(0, 1)$; (b) Signal (M2), noise $X_i = 0.3X_{i-1} + \epsilon_i$ with $\epsilon_i \overset{iid}{\sim} N(0, 1)$; (c) Add $\varepsilon_i \overset{iid}{\sim} N(0, 1)$ on (a); (d) Add $\varepsilon_i \overset{iid}{\sim} N(0, 1)$ on (d).

Figure 4.1 shows that the new error process $X_i + \varepsilon_i$ seems to be closer to an independent series and somehow leads to the dependence reduction in the original series, which is more obvious in (a) and (c). Heuristically speaking, the definition of long-run variance $\sigma_*^2 = \sum_{i \in \mathbb{Z}} \gamma(i)$ indicates that the main difference between LRV σ_*^2 and variance $\gamma(0)$

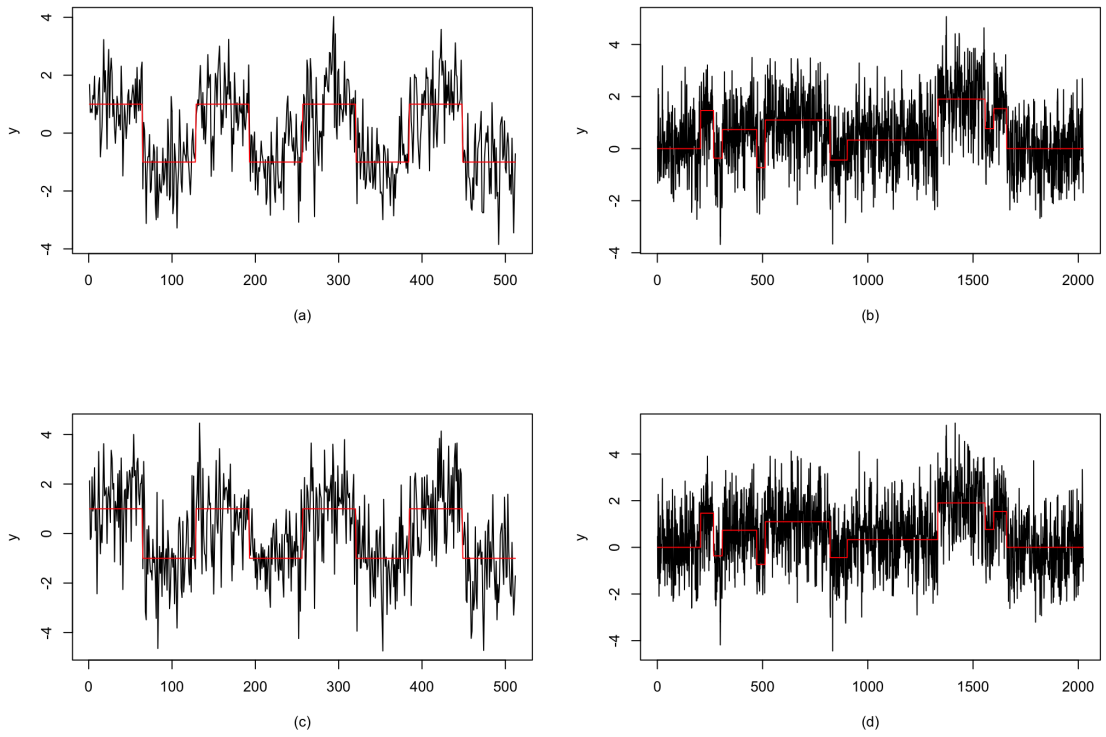


Figure 4.1: Plots for original simulated data [(a) and (b)] and plots for the corresponding preprocessed simulated data [(c) and (d)] after adding proper independent Gaussian series. The red lines represent the true signal f_i, see (M1) and (M2).

originates from the autocorrelations of the series, where independent sequences will always have $\gamma(0)/\sigma_*^2 = 1$. Hence the underlying idea of "correlation reduction" is to make the ratio of traditional variance $\gamma(0)$ to long-run variance σ_*^2 closer to 1. It can be easily derived that for any stationary error process X_i and new *iid* series $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, we have

$$\frac{\mathrm{Var}(X_i + \varepsilon_i)}{\mathrm{LRV}(X_i + \varepsilon_i)} = \frac{\mathrm{Var}(X_i) + \mathrm{Var}(\varepsilon_i)}{\mathrm{LRV}(X_i) + \mathrm{LRV}(\varepsilon_i)} = \frac{\gamma(0) + \sigma^2}{\sigma_*^2 + \sigma^2} \tag{4.3.1}$$

which always satisfies

$$\left| \frac{\gamma(0) + \sigma^2}{\sigma_*^2 + \sigma^2} - 1 \right| < \left| \frac{\gamma(0)}{\sigma_*^2} - 1 \right| \tag{4.3.2}$$

for any $\sigma^2 > 0$. The real challenging issue is to find an optimal σ^2 to make the NOT solution path algorithm well-suited for dependent data.

Before considering the the optimal value of σ^2, we first assess the effectiveness of this preprocessing approach in the case of signal (M1) under the following AR error structures (N1)-(N5) and (M2) under (N1), (N6) and (N7).

(N1) $X_i = 0.3X_{i-1} + \epsilon_i$, where $\epsilon_i \overset{iid}{\sim} N(0,1)$;

(N2) $X_i = 0.8X_{i-1} + \epsilon_i$, where $\epsilon_i \overset{iid}{\sim} N(0,1)$;

(N3) $X_i = 0.8X_{i-1} + \epsilon_i$, where $\epsilon_i \overset{iid}{\sim} N(0,0.35^2)$;

(N4) $X_i = -0.3X_{i-1} + \epsilon_i$, where $\epsilon_i \overset{iid}{\sim} N(0,1)$;

(N5) $X_i = -0.8X_{i-1} + \epsilon_i$, where $\epsilon_i \overset{iid}{\sim} N(0,1)$;

(N6) $X_i = 0.3X_{i-1} + \epsilon_i$, where $\epsilon_i \overset{iid}{\sim} N(0,0.5^2)$;

(N7) $X_i = 0.8X_{i-1} + \epsilon_i$, where $\epsilon_i \overset{iid}{\sim} N(0,0.1^2)$;

Models (N2), (N3), (N5) and (N7) represent strongly autocorrelated error process. And models (N6) and (N7) somehow present the error processes with the largest possible σ_ϵ^2 under which NOT can work well after data preprocessing. Table 4.2 and Table 4.3 generally demonstrate that adding additional *iid* Gaussian noise can successfully enhance the performance of the NOT solution path algorithm on dependent data. However, the scenarios (N2) in Table 4.2 and (N1) in Table 4.3 also indicate that further investigations are still required to discover a suitable formula of the possible variance σ^2 of the *iid* noise with respect to the sample sizes, change-point configurations and strength of dependence in the original data. That is, we can see that after the preprocessing procedure, NOT works well for data produced by signal (M1) and noise (N1), but fails to provide good results for data with the same noise but different signals. Therefore, the key point remaining here is to find out what kind of dependent data can potentially be estimated with NOT and what the optimal choices of σ^2 are for different kinds of dependent data.

Theoretically speaking, we know from the supplementary material of NOT (Baranowski et al., 2019) that the consistency of the NOT solution path algorithm is proved when the set E_n holds. Write $\mathbf{1}_{(s,e]} = (\mathbf{1}_{(s,e]}(1), \ldots, \mathbf{1}_{(s,e]}(n))^\intercal$ with

$$\mathbf{1}_{(s,e]}(t) = \begin{cases} (e-s)^{-1/2} & t = s+1, \ldots, e, \\ 0 & \text{otherwise} \end{cases} \tag{4.3.3}$$

and $\mathbf{X} = (X_1, \ldots, X_n)^\intercal$. The set E_n is defined as

$$E_n = \left\{ \max_{0 \leqslant s < e \leqslant n} \langle \mathbf{1}_{(s,e]}, \mathbf{X} \rangle \leqslant \sqrt{6 \log(n)} \right\}. \tag{4.3.4}$$

This provides a restriction to both the dependence and σ_ϵ^2 in the original noise and the σ^2 for newly added noise ε_i. Johnstone and Silverman (1997) presented that given Gaussian random variables $X_1, \ldots, X_n$ distributed with mean 0 and variances σ_i^2, we have

$$p\left\{ \max_{1 \leqslant i \leqslant n} |X_i/\sigma_i| > \sqrt{2 \log(n)} \right\} \to 0 \tag{4.3.5}$$

regardless of the level of dependence in $\{X_i\}_{i=1}^n$. Although the inequalities (4.3.4) and (4.3.5) do not necessarily share the same right bound, this still gives us a rough idea for setting the maximum value of σ_ϵ^2. Take the AR(1) error process as a special example, when adding a new *iid* Gaussian noise $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ to the original model with error process $X_i = \phi_1 X_{i-1} + \epsilon_i$, $\epsilon_i \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$. We can see that the newly created noise $X_i^N = X_i + \varepsilon_i$ has the properties $\mathrm{Var}(X_i^N) = \sigma_\epsilon^2/(1-\phi_1^2) + \sigma^2$. Therefore, we may derive from (4.3.4) and (4.3.5) that the largest possible value of $\sigma_\epsilon^2/(1-\phi_1^2) + \sigma^2$ may not locate distinctly away from 3. However, the specific expression for choosing the optimal σ^2 should be explored further and we attempt to start from a more practical perspective. Intuitively speaking, this issue can be closely related to the other problem with figuring out the combination of σ_ϵ^2 and ϕ_1 under which NOT works or breaks down, which is briefly discussed in Section 4.4.3.

Table 4.2: Distribution of $\hat{q} - q$ obtained by NOT solution path algorithm after pre-adding an *iid* error process following $N(0, \sigma^2)$ for data generated according to (2.2.1) with the signals (M1) and the noises (N1) to (N5), the average Mean Square Error of the resulting estimate of the signal and the average Hausdorff distance d_H over 100 simulations.

		$\hat{q} - q$								
Noise	σ	$\leqslant -3$	-2	-1	0	1	2	$\geqslant 3$	MSE	d_H
(N1)	0.00	0	0	0	67	10	15	8	1.481	0.124
	0.50	0	0	0	78	7	8	7	1.264	0.105
	0.75	0	0	0	82	13	3	2	1.506	0.100
	1.00	0	0	0	89	10	1	0	2.892	0.100
	1.25	0	1	0	83	10	4	2	2.557	0.138
	1.50	0	4	1	86	8	1	0	2.994	0.205
(N2)	0.00	0	0	0	0	0	0	100	4.676	0.463
	0.50	0	0	0	1	2	0	97	1.487	0.448
	0.75	0	0	0	0	0	2	98	4.687	0.459
	1.00	0	0	0	1	3	7	89	5.901	0.450
	1.25	0	1	1	2	10	13	73	2.840	0.460
	1.50	3	2	3	11	6	10	65	5.863	0.468
	1.75	4	8	5	12	20	11	40	8.356	0.553
	2.00	8	8	12	19	6	17	30	5.482	0.690
(N3)	0.00	0	0	0	0	0	0	100	0.774	0.460
	0.50	0	0	0	4	2	8	86	0.419	0.385
	0.75	0	0	0	28	18	16	38	0.767	0.266
	1.00	0	0	0	44	28	11	17	2.097	0.195
	1.25	0	0	0	69	14	8	9	1.739	0.165
	1.50	0	2	1	79	12	3	3	2.170	0.182
	1.75	1	9	3	71	15	1	0	4.719	0.272
	2.00	8	13	9	56	10	2	2	4.135	0.536
(N4)	0.00	0	0	0	100	0	0	0	0.033	0.022
(N5)	0.00	0	0	0	100	0	0	0	0.028	0.013

4.3.2 Pre-averaging the sequence over non-overlapping moving windows

In this method, we divide the time series into $\lfloor T/h \rfloor$ blocks of length h and consider the calculated local averages as the new dataset, hoping that the pre-averaged noise can approximately follow a Gaussian distribution by the law of large numbers. In the

Table 4.3: Distribution of $\hat{q} - q$ obtained by NOT solution path algorithm after pre-adding an *iid* error process following $N(0, \sigma^2)$ for data generated according to (2.2.1) with the signals (M2) and the noises (N1), (N6) and (N7), the average Mean Square Error of the resulting estimate of the signal and the average Hausdorff distance d_H over 100 simulations.

		$\hat{q} - q$								
Noise	σ	$\leqslant -3$	-2	-1	0	1	2	$\geqslant 3$	MSE	d_H
(N1)	0.00	1	15	36	20	18	5	5	1.406	0.829
	0.50	8	13	41	26	7	3	2	1.325	0.783
	0.75	26	18	27	21	3	3	2	1.432	0.949
	1.00	31	37	19	10	3	0	0	2.751	0.869
	1.25	48	33	11	4	3	1	0	2.642	1.251
	1.50	84	10	6	0	0	0	0	2.803	1.696
(N6)	0.00	0	0	4	71	17	2	6	0.293	0.307
	0.50	0	0	2	81	11	3	3	0.279	0.267
	0.75	0	0	2	72	17	8	1	0.294	0.313
	1.00	0	0	6	74	12	6	2	0.447	0.228
	1.25	0	0	11	75	11	2	1	0.429	0.310
	1.50	0	0	21	59	14	4	2	0.453	0.437
(N7)	0.00	0	0	0	0	0	0	100	0.2816	2.397
	0.25	0	0	0	19	21	20	40	1.035	1.27
	0.50	0	0	3	74	17	5	1	0.272	0.251
	0.75	0	0	27	64	8	1	0	0.294	0.307
	1.00	2	12	45	38	3	0	0	1.019	0.436

new series, we often have two consecutive change-points located in the block containing the original change-point, see sub-figures (a) and (b) in Figure 4.3.

To study whether this approach can be helpful, here we display Figures 4.2 to 4.5 that summarise the results for different AR(1) error processes defined in Section 4.3.1. In particular, the raw data shown in (a) and (c) of Figure 4.2 and 4.3 is built on signal (M1) and noise (N1) while (b) and (d) presenting results for signal (M1) and noise (N2). Also, Figure 4.4 shows the results for signal (M2) and the same noises (N1) and (N2) whereas Figure 4.5 is plotted for signal (M2) and noise (N6) for sub-figures (a) and (c), or noise (N7) for (b) and (d).

These graphs show that the pre-averaging stage can be useful for NOT to estimate the

change-points in correlated series but its effectivess also depends on the sample sizes, change-point configurations and strength of dependence mentioned in the last section. The problem goes back to identifying the characteristics of the raw data. Moreover, compared to adding *iid* noise, the application of the pre-averaging approach can face more than one issues. First, even if the change-points are successfully detected in the pre-averaged data, it is still hard to find the true change-point location within the detected block. Although utilising moving windows may help reduce this problem, it will also increase the computational complexity. The second issue arises from the optimal choice of bandwidth h, and the difficulty in detecting the two consecutive change-points resulted from pre-averaging; see red dots in Figure 4.3 and 4.5.

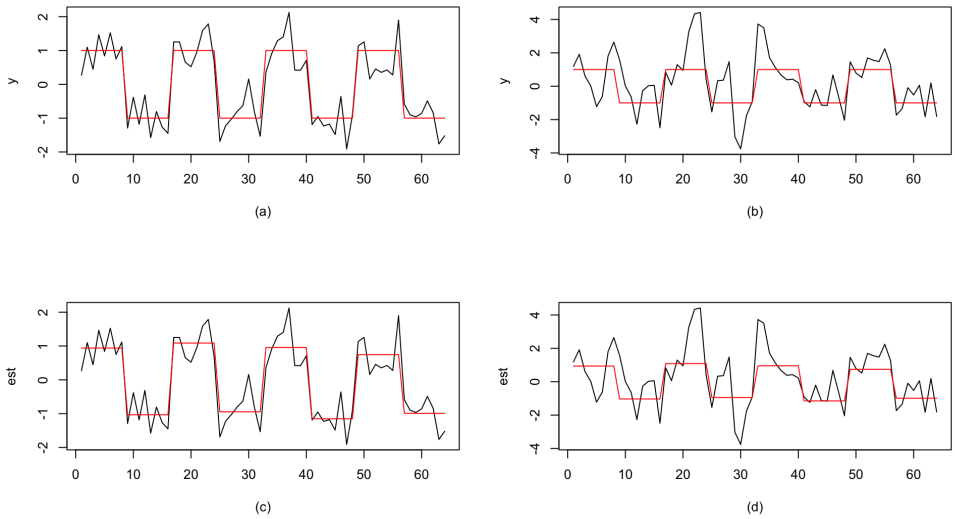


Figure 4.2: Under original signal (M1), plots for pre-averaged data with low correlated noise (N1) [(a) and (c)] and plots for the pre-averaged data with high correlated noise (N2) [(c) and (d)], where the bandwidth is set to be 8. The red lines in (a)-(b) and (c)-(d) represent the true and estimated signals in pre-averaged data respectively.

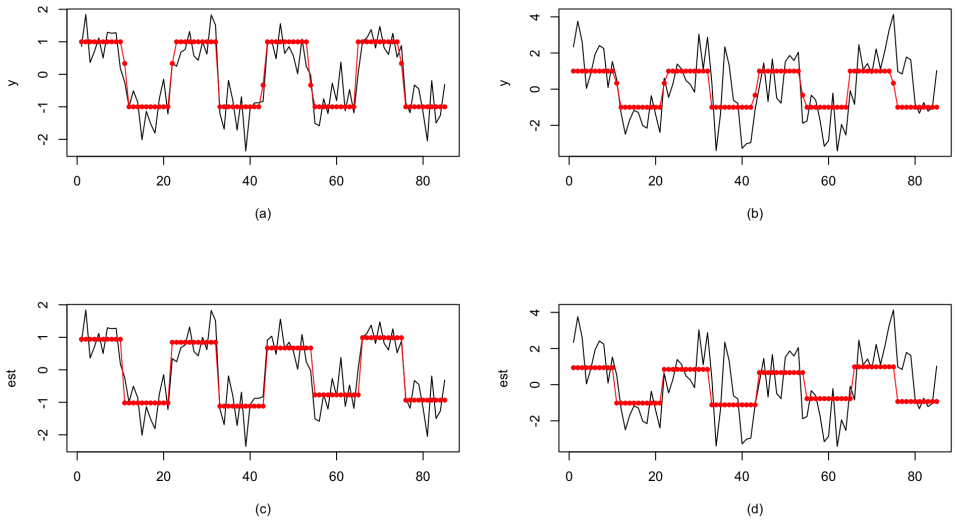


Figure 4.3: Under original signal (M1), plots for pre-averaged data with low correlated noise (N1) [(a) and (c)] and plots for the pre-averaged data with high correlated noise (N2) [(c) and (d)], where the bandwidth is set to be 6. The red lines in (a)-(b) and (c)-(d) represent the true and estimated signals in pre-averaged data respectively

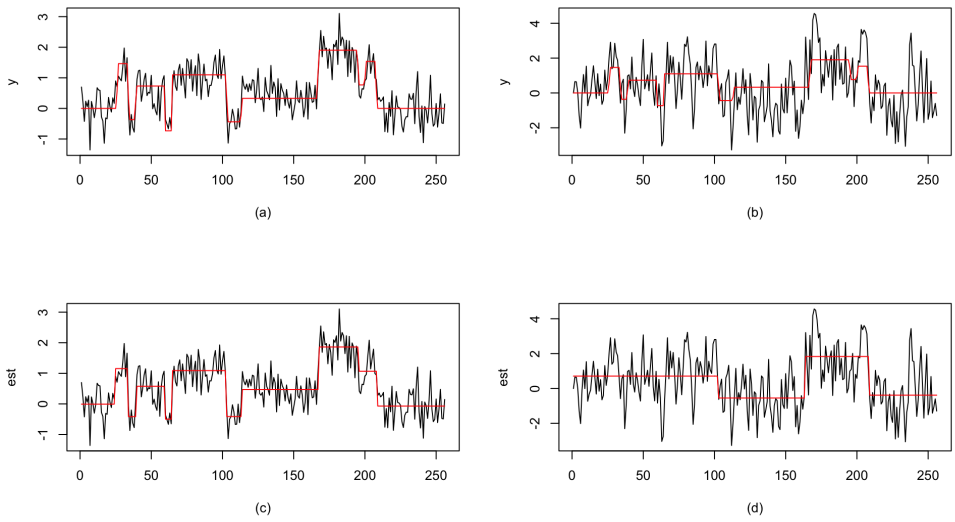


Figure 4.4: Under original signal (M2), plots for pre-averaged data with low correlated noise (N1) [(a) and (c)] and plots for the pre-averaged data with high correlated noise (N2) [(c) and (d)]. The red lines in (a)-(b) and (c)-(d) represent the true and estimated signals in pre-averaged data respectively

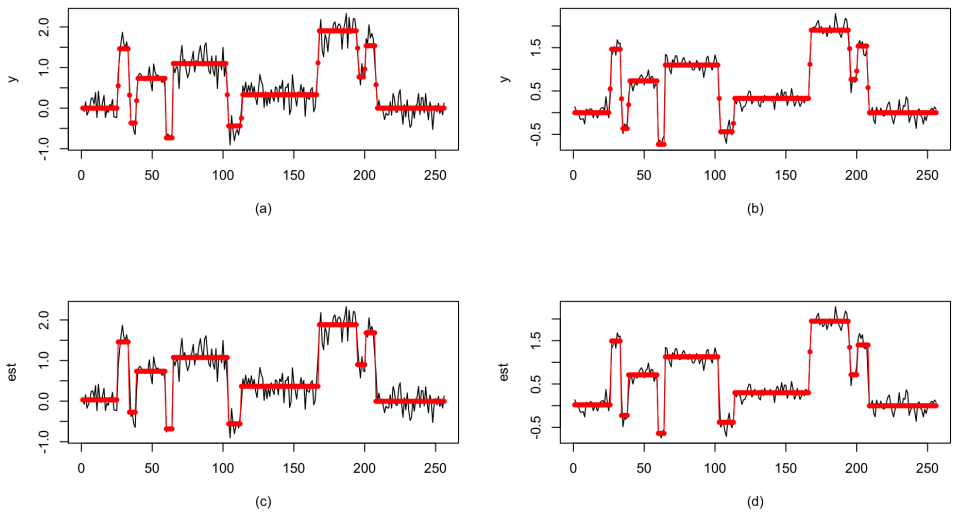


Figure 4.5: Under original signal (M2), plots for pre-averaged data with low correlated noise (N6) [(a) and (c)] and plots for the pre-averaged data with high correlated noise (N7) [(c) and (d)]. The red lines in (a)-(b) and (c)-(d) represent the true and estimated signals in pre-averaged data respectively

4.4 Extended NOT Solution Path Algorithm

In this section, we shall concentrate on discovering the possible extensions on the NOT solution path algorithm itself instead of data preprocessing. There are generally two straightforward ways that can be utilised to build the information-based criterion for candidate model selection under the assumption of dependent noise: adjusting the measure of fit to the data or the penalty of the strengthen Schwarz Information Criterion in NOT. These two parts are discussed separately in the following sections.

4.4.1 New Information Criterion

In order to extend the NOT solution path algorithm to dependent data, we attempt to develop new information-based criterion for better performance. For any candidate

model $\mathcal{T}(\zeta_n^{(k)})$, let $\hat{f}_t^k$ denote the estimated signal f with the definition $\hat{f}_t^k = (\hat{\eta}_{i+1} - \hat{\eta}_i)^{-1} \sum_{j=\hat{\eta}_i+1}^{\hat{\eta}_{i+1}} Y_j$ for $\hat{\eta}_i + 1 \leqslant t < \hat{\eta}_{i+1}$, $i = 1, \ldots, \hat{q}(\zeta_n^{(k)})$. Let $\hat{\sigma}_k^2 = n^{-1} \sum_{t=1}^n (Y_t - \hat{f}_t^k)^2$ represent the maximum likelihood estimator of the residual variance. Under the assumption of *iid* Gaussian noise, the sSIC function (2.5.3) is then reduced to

$$\mathrm{sSIC}(k) = \frac{n}{2} \log \hat{\sigma}_k^2 + n_k \log^\alpha(n) \tag{4.4.1}$$

Considering the serial correlation in the Gaussian error process, we can somehow represent the measure of fit as a function of quantified dependence. We first attempt to start with the log-likelihood function of a multivariate normal distribution, but we find it hard to theoretically simplify the expression and even difficult to estimate the determinant of the unknown covariance matrix. Therefore, we turn to thinking about whether there are other measures acting as a possible substitution for the log-likelihood function.

For simplicity, we decide to include the serial correlation in noise by straightforwardly replacing the estimated residual variance $\hat{\sigma}_k^2$ in (2.5.3). Note that for given observations $\{Y_i\}_{i=1}^n$ and set of random intervals F_n^M, each threshold ζ_n and the corresponding candidate model are selected by minimising sSIC. This indicates that we should retain the features in the signal when they are not tested in the candidate model. Here, we shall provide two possible solutions: segmented LRV estimators $\hat{\sigma}_2^2$ and sum of estimated autocovariances. Mathematically speaking, for any candidate model $\mathcal{T}(\zeta_n^{(k)})$, we define

$$\begin{aligned}
\mathrm{IC}^{LRV}(k) &= \frac{n}{2} \log \frac{1}{n} \sum_{i=1}^{\hat{q}(\zeta_n^{(k)})} (\hat{\eta}_{i+1} - \hat{\eta}_i + 1) \hat{\sigma}_*^2(Y_{\hat{\eta}_i+1}, \ldots, Y_{\hat{\eta}_{i+1}}) + n_k \log^\alpha(n) \\
\mathrm{IC}^{ACV}(k) &= \frac{n}{2} \log \frac{1}{n} \sum_{i=1}^{\hat{q}(\zeta_n^{(k)})} (\hat{\eta}_{i+1} - \hat{\eta}_i + 1) \left(\sum_{\tau=1}^{K} \left[2 |\hat{\gamma}_{[\hat{\eta}_i, \hat{\eta}_{i+1})}(\tau)| + 1 \right] \right)^2 + n_k \log^\alpha(n)
\end{aligned} \tag{4.4.2}$$

where $\hat{\sigma}_*^2(Y_{\hat{\eta}_i+1}, \ldots, Y_{\hat{\eta}_{i+1}})$ denotes the local LRV estimated on interval $[\hat{\eta}_i, \hat{\eta}_{i+1})$ and $\hat{\gamma}_{[\hat{\eta}_i, \hat{\eta}_{i+1})}(\tau)$ represents the estimated autocovariance at lag τ over interval $[\hat{\eta}_i, \hat{\eta}_{i+1})$.

First, it is natural to think of the proposed wavelet-based estimators of LRV, which provide a good description of dependence in data and have good performance for different cases. To keep the undetected changes in the signal, we choose the non-thresholded estimators and tend not to use the most robust median-based ones. In $\mathrm{IC}^{LRV}(k)$, the locally estimated LRV $\hat{\sigma}_*^2(Y_{\hat{\eta}_i+1}, \ldots, Y_{\hat{\eta}_{i+1}})$ acts as a replacement of the estimated residual variance. Then we construct a simple weighted average to provide the whole picture of the candidate model. Analogously, the sum of the absolute estimated autocovariances can somehow represent the LRV whereas the absolute sign increases the strength of dependence extracted from the candidate model. Before studying the theoretical development of the new information-criterion-based methods, we shall, in the next subsection, first analyse the practical behaviour of the NOT solution path algorithm.

4.4.2 Simulation Results I

We test the performance of the three NOT approaches in the case of regular change-points, i.e. signal (M1), under the assumptions of error structures (N1) to (N5) defined in Section 4.3.1. We report the results of three methods based on $\mathrm{IC}^{LRV}(k)$, $\mathrm{IC}^{ACV}(k)$ and sSIC referred to as approaches (A1), (A2) and (A3), respectively. The number of randomly drawn intervals and the maximum number of change-points for sSIC and two new information-based criterion are pre-specified to be $M = 10000$ and $q_{max} = 20$ respectively. Table 4.4 summarises the results of the simulation study from 100 samples of a time series with length $n = 512$ produced by signal (M1) plus noise (N1)-(N5) when having a good choice of α.

Overall, it shows that the original NOT solution path algorithm can generally out-

Table 4.4: Distribution of $\hat{q}-q$ obtained by approach (A1)-(A3) for data generated according to (2.2.1) with the signals (M1) and the noises (N1) to (N5) in Section 4.3, the average Mean Square Error of the resulting estimate of the signal and the average Hausdorff distance d_H over 100 simulations.

		$\hat{q}-q$								
Noise	Method	$\leqslant -3$	-2	-1	0	1	2	$\geqslant 3$	MSE	d_H
(N1)	(A1)	0	2	5	82	9	2	0	1.058	0.051
	(A2)	0	0	0	100	0	0	0	0.002	5.360
	(A3)	0	0	0	100	0	0	0	0.003	0.015
(N2)	(A1)	22	9	13	13	21	6	16	2.129	0.169
	(A2)	15	10	13	29	19	8	6	0.004	57.600
	(A3)	27	6	8	18	16	8	17	1.975	0.187
(N3)	(A1)	1	5	10	65	16	2	1	0.313	0.037
	(A2)	0	0	0	99	1	0	0	0.001	3.480
	(A3)	0	0	0	99	1	0	0	0.295	0.014
(N4)	(A1)	0	0	0	100	0	0	0	1.064	0.004
	(A2)	0	0	0	100	0	0	0	0.002	2.190
	(A3)	0	0	0	100	0	0	0	1.063	0.003
(N5)	(A1)	0	0	0	100	0	0	0	2.692	0.002
	(A2)	1	5	3	72	13	5	1	0.005	14.360
	(A3)	0	0	0	100	0	0	0	2.691	0.002

perform the two extended NOT methods if an optimal α is selected. Additionally, it is evident that the two proposed measures cannot effectively replace the estimated determinant of the unknown covariance matrix. Such failure arises from their inability to achieve a satisfactory balance between change-point features and dependence conditions in serial correlated data. In particular, the NOT based on $\mathrm{IC}^{ACV}(k)$ tends to overestimate the change-points as we can see the small MSE and large d_H when the number of change-points is correct. One potential explanation is that the summation of absolute autocovariances incorporates the dependence in noise to a greater extent than necessary. One the other hand, the distribution of $\hat{q}-q$ for the $\mathrm{IC}^{LRV}(k)$-based NOT suggests that the LRV estimators are still too robust to multiple mean shifts even without removing large coefficients in moderate scales, which somehow aligns with the proved asymptotic consistency of $\hat{\sigma}_2^2$. Specifically, although the maximum of absolute difference in local averages can serve as a global measure of the discrepancy, see Wu and

Zhao (2007), the difference-based LRV estimators are usually developed to eliminate the signal. We leave the vital problem of finding an optimal information criterion as one of the possible extensions of NOT for future investigation.

4.4.3 Extension on Existing Information Criterion

In this section, we analyse the practical performance of the NOT solution path algorithm built on the original sSIC (2.5.3) by simply adapting the value of α. In general, the choice of α intuitively depends on four main perspectives: (a) sample size n, (b) strength of autocorrelation, (c) the number of estimated change-points and (d) signal-to-noise ratio. However, the number of estimated change-points should be unknown before we conduct the change-point detection. Therefore, we would like to have a discussion on how to choose the parameter α based on the remaining three elements.

In particular, to summarise the strength of autocorrelation of one process, we apply a straightforward *autocorrelation index (ACI)* ϖ_Y defined as

$$\varpi_Y^2 = \sum_{\tau=-\infty}^{\infty} |\rho(\tau)| \tag{4.4.3}$$

Considering real-world applications, the positively correlated error process is widely employed and capable of describing the persistent features of data in many fields, such as economic data (stock prices, GDP growth rate, inflation rates) and environmental data (daily temperature, precipitation). Hence we shall start with simulations on data following an AR(p) error process with positive serial dependence, where $X_i = \sum_{j=1}^{p} \phi_j X_{i-j} + \epsilon_i$ with $\epsilon_i \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$. Under such an assumption, ACI shares the same expression as the scaled long-run standard deviation, i.e. $\varpi_Y^2 = \sum_{\tau=-\infty}^{\infty} \rho(\tau) = \sigma_*^2/\gamma(0)$, which makes our wavelet-based estimators in Section 3 good measures for calculating ACI. Here we choose $\hat{\varpi}_Y = \hat{\sigma}_4(\lambda)/\hat{\gamma}(0)$; see its definition in (3.2.6). Given

a pre-specified window size h and the corresponding block number $k_n = \lfloor n/h \rfloor$, the estimated $\gamma(0)$ is defined as follows

$$\hat{\gamma}(0) = \text{median}\left(\frac{1}{h-1}\sum_{i=1}^{h}(Y_i - \bar{Y}_{(1+jh):(j+1)h})^2, \quad j = 1, \ldots, k_n\right) \tag{4.4.4}$$

where $\bar{Y}_{(1+jh):(j+1)h}$ stands for the average of observations $Y_{1+jh}, \ldots, Y_{(j+1)h}$.

Meanwhile, estimating the *signal-to-noise ratio (SNR)* is often helpful for comparing the level of the signal and noise in data, and its challenge mainly comes from quantifying the strength of the signal. In practice, they are frequently described with the corresponding standard deviation, power or the actual value directly. As $\hat{\sigma}_4(\lambda)$ is suitable for estimating the level of noise, it is then a good choice to quantify the signal in the form of "standard deviation" as well. Therefore, we consider the empirical signal-to-noise ratio computed by

$$\widehat{\text{SNR}}_Y = \frac{s_Y}{\hat{\sigma}_4(\lambda)} \tag{4.4.5}$$

where s_Y represents the sample standard deviation of $\{Y_i\}_{i=1}^n$.

Table 4.5: Possible choices of α obtained by running the NOT solution path algorithm for data generated according to (2.2.1) with the signal (M1) and the noises following AR(1) model, whose sample signal-to-noise ratio and autocorrelation index are presented under SNR and ACI. Meanwhile, ϕ_1 denotes the coefficient of AR(1) noise. And $\hat{\sigma}_\epsilon$ represents the maximum value of σ_ϵ in $\epsilon_i \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$ where NOT still works.

Noise	ϕ_1	$\hat{\sigma}_\epsilon$	α	SNR	LRSD	ACI	$\hat{\gamma}(0)$
AR(1)	0.1	1.80	1.00	0.857	2.412	1.435	1.687
AR(1)	0.2	1.60	1.05-1.15	0.794	2.386	1.596	1.496
AR(1)	0.3	1.40	1.15-1.30	0.760	2.328	1.783	1.304
AR(1)	0.4	1.20	1.30-1.45	0.722	2.290	2.012	1.128
AR(1)	0.5	1.00	1.40-1.60	0.681	2.228	2.353	0.946
AR(1)	0.6	0.80	1.65	0.666	2.103	2.767	0.766
AR(1)	0.7	0.65	1.70	0.639	2.063	3.260	0.633
AR(1)	0.8	0.45	1.90	0.667	1.857	4.131	0.448
AR(1)	0.9	0.20	2.20	1.198	0.923	4.643	0.201

Figure 4.6 and 4.7 display the estimated signal-to-noise ratio and autocorrelation index

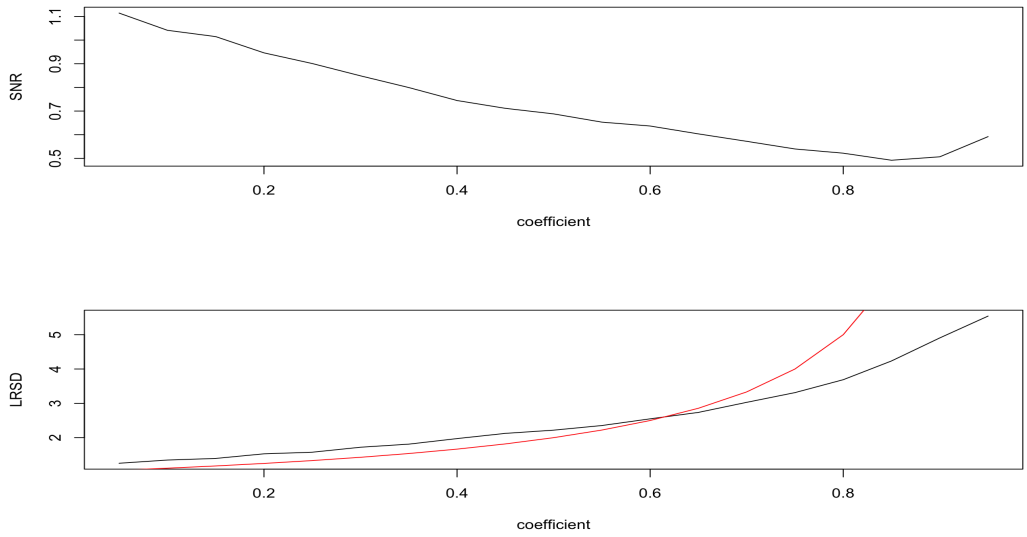


Figure 4.6: Under original signal (M1), plots for estimated signal-to-noise ratio (SNR) and autocorrelation index (ACI) with AR(1) Gaussian noise $X_i = \phi_i X_{i-1} + \epsilon_i$, where $\phi_i > 0$, $\epsilon_i \overset{iid}{\sim} N(0,1)$ $(n = 512)$. The red line represents the true long-run standard deviation for the corresponding error process.

Table 4.6: Possible choices of α obtained by running the NOT solution path algorithm for data generated according to (2.2.1) with the signal (M2) and the noises following AR(1) model, whose sample signal-to-noise ratio and autocorrelation index are presented under SNR and ACI. Meanwhile, ϕ_1 denotes the coefficient of AR(1) noise. And $\hat{\sigma}_\epsilon$ represents the maximum value of σ_ϵ in $\epsilon_i \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$ where NOT still works.

Noise	ϕ_1	$\hat{\sigma}_\epsilon$	α	SNR	LRSD	ACI	$\hat{\gamma}(0)$
AR(1)	0.1	0.65	1.00-1.10	1.332	0.720	1.133	0.637
AR(1)	0.2	0.55	1.10-1.30	1.363	0.670	1.246	0.538
AR(1)	0.3	0.45	1.15-1.55	1.382	0.609	1.366	0.445
AR(1)	0.4	0.40	1.35-1.50	1.342	0.614	1.542	1.401
AR(1)	0.5	0.35	1.40-1.50	1.263	0.642	1.804	0.358
AR(1)	0.6	0.25	1.60	1.411	0.543	2.049	0.264
AR(1)	0.7	0.20	1.70	1.389	0.542	2.455	0.221
AR(1)	0.8	0.15	1.90	1.364	0.548	3.159	0.173
AR(1)	0.9	0.10	2.15	1.340	0.543	4.498	0.121

for data produced by positively correlated AR(1) noise with $\epsilon_i \overset{iid}{\sim} N(0,1)$ and coefficient $\phi_i \in [0.05, 0.95]$ but different signals (M1) and (M2). We can see that with the

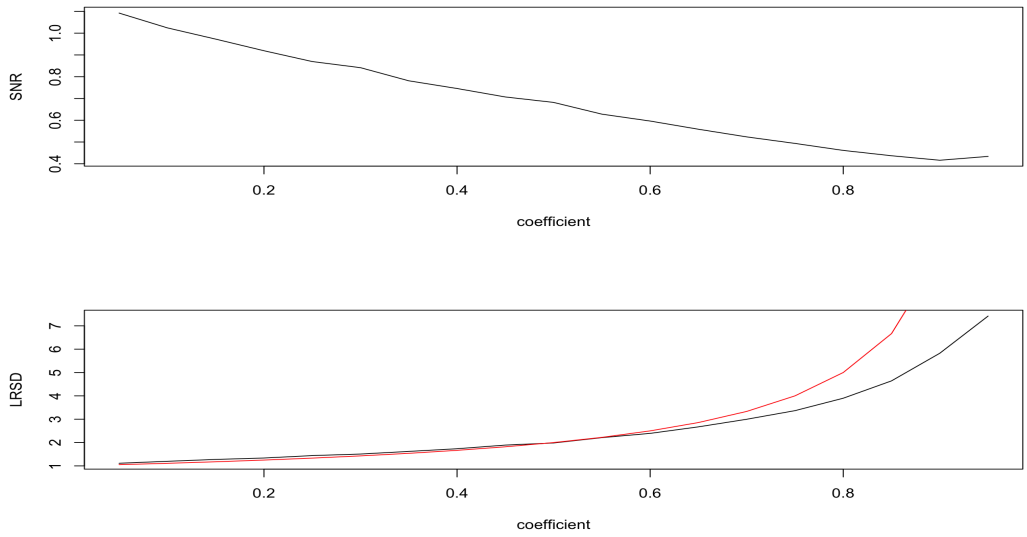


Figure 4.7: Under original signal (M2), plots for estimated signal-to-noise ratio (SNR) and autocorrelation index (ACI) with AR(1) Gaussian noise $X_i = \phi_i X_{i-1} + \epsilon_i$, where $\phi_i > 0$ $\epsilon_i \overset{iid}{\sim} N(0,1)$ $(n = 2024)$. The red line represents the true long-run standard deviation.

same $\sigma_\epsilon^2 = 1$, the estimation results of ACI are poorer for error process with higher correlation, which may, together with estimated SNR, act as an indicator of the data setting when NOT still works well or breaks down for dependent data.

To clarify, from the two tables, it is reasonable to presume that the optimal α for NOT solution path algorithm can be viewed as a value proportional to the AR coefficient ϕ_1. Also, except for the results of the cases with large ϕ_1, we can see that the ACI can somehow linearly reflect the value of ϕ_1. Meanwhile, since we have different relationships between α and ACI from the two tables, we consider SNR as a condition for classifying different scenarios.

Therefore, we conduct experiments on more simulated data produced by simpler signal models (M3) to (M6), with AR(1) error processes generated from the following two models for $\phi_1 = 0.1, 0.2 \ldots, 0.9$.

- $X_i = \phi_1 X_{i-1} + \epsilon_i$, where $\epsilon_i \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$ with $\sigma_\epsilon = 2(1-\phi_1)$
- $X_i = \phi_1 X_{i-1} + \epsilon_i$, where $\epsilon_i \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$ with $\sigma_\epsilon = \sqrt{1-\phi_1^2}$

During experiments, we first briefly summarise the conditions when the NOT solution path algorithm, after changing the α, could still have good performance and when it might break down. That is, for time series with larger mean shift size, the extended algorithm works well for any $\sigma_\epsilon \in [0.05, \max(2(1-\phi_1), \sqrt{1-\phi_1^2})]$, whereas it is better to have $\sigma_\epsilon \in [0.05, \min(2(1-\phi_1), \sqrt{1-\phi_1^2})]$ when the mean shift size is small.

After conducting several tests, we find two series of α that can roughly work well for all models above. And in particular, for data such as (A3) and (A5) with larger mean shift size, $\alpha^{(1)} = 1.2, 1.35, 1.5, 1.6, 1.7, 1.9, 2.0, 2.2, 2.5$ works better for $\phi_1 = 0.1, 0.2 \ldots, 0.9$; On the other hand, we tend to use $\alpha^{(2)} = 1.085, 1.225, 1.35, 1.45, 1.525, 1.7, 1.8, 1.975, 2.25$, see Figure 4.8. To test whether such α is also well-suited for other coefficients, we apply the simple Lagrangian function to derive the α for other ϕ_1's, i.e. we test the performance of data generated by AR(1) noise with different ϕ_1's based on α on the lines in Figure 4.8. The good results encourage us to find an explicit formula for these α's. To fulfill this aim, we first use signal-to-noise ratio and the autocorrelation index to categorise the data into different classes. We mainly consider two important criteria for the segmentation process: one is to divide the models with different mean shift size ($\widehat{\text{SNR}}_Y$) and the other is to separate data with weak or strong autocorrelations ($\hat{\varpi}$). Meanwhile, following the aforementioned combination of mean shift size and σ_ϵ allowing for good performance, we divide the samples into three categories and then derive the corresponding formulas.

Here we propose an empirical formula on the selection of parameter α required for improving the performance of NOT for dependent data produced by piecewise-constant

signal and an AR(p) error process.

$$\hat{\alpha} = \begin{cases} 0.65|\hat{\varpi}_Y - 0.8| + 1.2 & \widehat{\mathrm{SNR}}_Y \geqslant 1.4, \hat{\varpi} \leqslant 2, \\ 0.65|\hat{\varpi}_Y - 1.0| + 1.2 & \widehat{\mathrm{SNR}}_Y \geqslant 1.4, \hat{\varpi} > 2, \\ \min(2.5, 0.6|\hat{\varpi}_Y - 1.2| + 1.0) & 0.5 < \widehat{\mathrm{SNR}}_Y < 1.4 \end{cases} \tag{4.4.6}$$

When $|\hat{\varpi}_Y|$ is small, if $\widehat{\mathrm{SNR}}_Y/\hat{\varpi}_Y < 0.5$, the NOT solution path is highly likely to break down regardless of the choice of α.

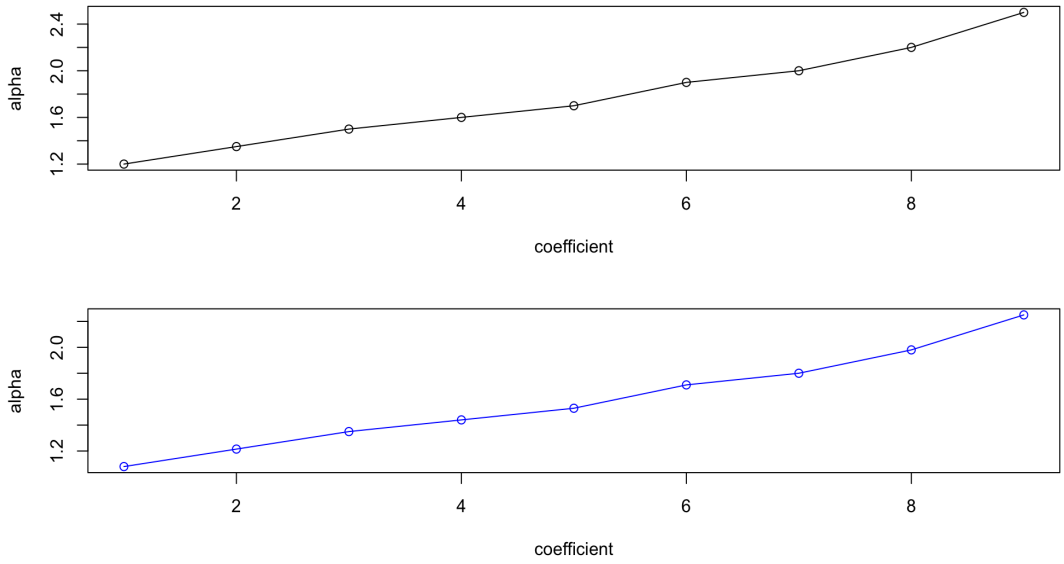


Figure 4.8: Plots for possible choices of α that works well for different values of coefficient ϕ_1 in AR(1) error process $X_i = \phi_1 X_{i-1} + \epsilon_i$, where the top (bottom) one is provided for data with larger (small) mean shift size.

4.4.4 Simulation Results II

In this section, we evaluate the performance of the NOT solution path algorithm applied with the guidance on the choice of the pre-specified α; see (4.4.6). We provide the simulation results in the case of examples following (M1)-(M14) introduced in Section

4.1, under a variety of error processes defined below. We assume that $\epsilon_t \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$ and different σ_ϵ's are randomly generated for different signals, which will be specified later. For each scenario, we consider the case with zero change-points ($q = 0$) in order to evaluate the proposed methodology on its possibility of false detection.

(N1) $X_t = \epsilon_t$ with $\sigma_\epsilon = 1$;

(N2) AR(1) model $X_t = \phi_1 X_{t-1} + \epsilon_t$, with $\phi_1 = 0.1$;

(N3) AR(1) model $X_t = \phi_1 X_{t-1} + \epsilon_t$, with $\phi_1 = 0.5$;

(N4) AR(1) model $X_t = \phi_1 X_{t-1} + \epsilon_t$, with $\phi_1 = 0.9$;

(N5) AR(2) model $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \epsilon_t$, with $\phi_1 = 0.5$ and $\phi_2 = 0.3$;

(N6) AR(2) model $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \epsilon_t$, with $\phi_1 = 0.7$ and $\phi_2 = 0.2$;

(N7) ARMA(1,1) model $X_t = \phi_1 X_{t-1} + \epsilon_t + \vartheta_1 \epsilon_{t-1}$, with $\phi_1 = 0.5$ and $\vartheta_1 = 0.3$ with $\sigma_\epsilon = 1/2.14285$;

(N8) ARMA(1,1) model $X_t = \phi_1 X_{t-1} + \epsilon_t + \vartheta_1 \epsilon_{t-1}$ with $\phi_1 = 0.7$ and $\vartheta_1 = 0.2$ with $\sigma_\epsilon = \sqrt{(1-\phi_1)^2/(1+\phi_1\vartheta_1+\vartheta_1^2)}$;

(N9) MA(1) model $X_t = \epsilon_t + \vartheta_1 \epsilon_{t-1}$, with $\vartheta_1 = -0.9$ and $\sigma_\epsilon = 1$.

For signal models with larger size of θ_i, i.e. the ones marked with even numbers within (M3)-(M14), we generate the corresponding noises σ_ϵ from $\sigma_\epsilon \sim \mathcal{U}[0.05, \max(0.2(1-\phi_1), \sqrt{1-\phi_1^2})]$. On the other hand, for the remaining cases with smaller signals or zero signal, the noises are selected relying on $\sigma_\epsilon \sim \mathcal{U}[0.05, \min(0.2(1-\phi_1), \sqrt{1-\phi_1^2})]$. The generated sample for simulation is recorded in Table 4.7 and the choice of parameters in models (N7) and (N8) are motivated by samples in Cho and Fryzlewicz (2023). The small LRV in (N9) increases the difficulty in its accurate estimation. Due to the comparatively small signals in (M1) and (M2), their corresponding noises σ_ϵ

for (N2)-(N9) are selected in advance as $0.65, 0.45, 0.10, 0.25, 0.15, 0.30, 0.20, 0.80$ and $0.55, 0.35, 0.05, 0.15, 0.05, 0.30, 0.20, 0.80$ respectively.

Table 4.7: The σ_ϵ's randomly generated for different signals (M3)-(M14).

σ_ϵ	(N2)	(N3)	(N4)	(N6)	σ_ϵ	(N2)	(N3)	(N4)	(N5)	(N6)
(M3)	1.42	0.75	0.42	0.37	(M4)	0.78	0.39	0.16	0.41	0.15
(M5)	1.53	0.47	0.34	0.26	(M6)	0.56	0.81	0.13	0.37	0.19
(M7)	1.65	0.90	0.25	0.34	(M8)	0.80	0.46	0.14	0.40	0.15
(M9)	1.51	0.80	0.29	0.21	(M10)	0.62	0.65	0.10	0.39	0.16
(M11)	1.28	0.59	0.23	0.17	(M12)	0.49	0.29	0.07	0.28	0.14
(M13)	1.47	0.35	0.27	0.28	(M14)	0.85	0.76	0.13	0.35	0.10

Similarly, we generate 100 replications and summarise the results of the simulation study in a frequency table with the distribution of $\hat{q} - q$, the estimated MSE of the estimated signal $\hat{f}_t$, the estimates of the (scaled) Hausdorff distance d_H and the proportion of spurious detection (evaluation of the size control performance).

Overall, based on the empirical formula of α (4.4.6), the NOT solution path algorithm displays good size control for dependent data. In particular, when the sample size is sufficiently large ($n \geqslant 200$), the proportion of the cases where the α-adapted NOT falsely detects any change-point in a dataset with $q = 0$ is strictly controlled below 0.20 (often around 0.01). In addition, when $q \geqslant 1$, Table 4.9 to Table 4.11 demonstrate that the α-adapted NOT algorithm consistently performs well for time series with relatively low or moderate serial correlations (N1)-(N3) regardless of the value of σ_ϵ (if it satisfies the aforementioned criterion). Also, this algorithm shows good performance of detecting both the number and locations of change-points in scenario (N9) where the LRV is close to 0, i.e. good model selection accuracy as shown by the distribution of $\hat{q} - q$ and good localisation accuracy from low d_H. However, the α-adapted NOT seems to have two potential pitfalls. First, when there are strong autocorrelations, the usefulness of this adapted algorithm can be impacted by the value of σ_ϵ or the number and localisation of change-points in the original data. Second, this empirical

result is not robust to the small size of θ_i and we can see this limitation in Table 4.8, where the algorithm finds it hard to show satisfactory results even for data with lower dependence.

For further inspection of the results, Figure 4.9 to Figure 4.22 present the histograms of the estimated change-point locations for different combinations of signal and noise models across 100 realisations, where the data correspond to those tested examples whose results are summarised in tables. Here all zeros shown in the histograms indicate the failure to detect any change-points when $q \geqslant 1$, which suggests an overestimate of α. We see that the α-adapted NOT solution path algorithm can often accurately detect the change-point locations where relatively large mean shift size exists. For example, Figure 4.14 shows that for the dataset with signal (M6), the localisation of the first change-point is not as precise as the second one, where the mean shift sizes are 2 and 5 respectively.

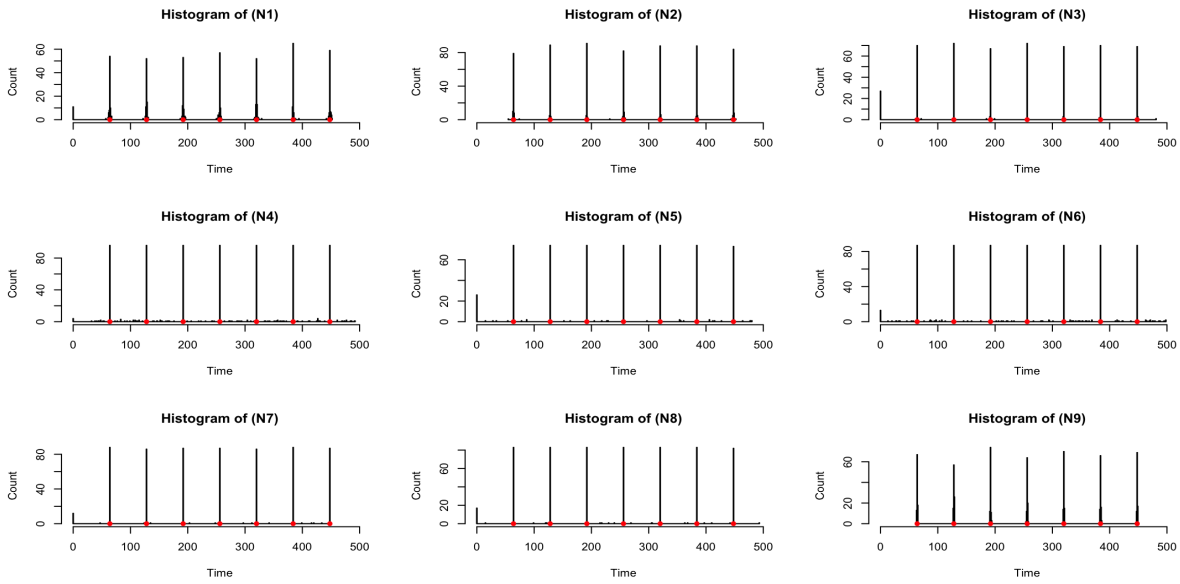


Figure 4.9: Histograms plotting the estimated change-points for data produced by signal (M1) and noises (N1)-(N9) by the α-adapted NOT solution path algorithm. The true change-points are coloured red.

Table 4.8: Distribution of $\hat{q}-q$ obtained by the α-adapted NOT algorithm for data generated according to (2.2.1) with the signals (M1) and (M2) and the noises (N1)-(N9), the average Mean Square Error (MSE) of the resulting estimate of the signal and the average Hausdorff distance d_H over 100 simulations. We also report the size, the proportion of replications with change-points being falsely detected when there are no change-points.

			$\hat{q}-q$								
Model	Noise	Size	$\leqslant 3$	-2	-1	0	1	2	$\geqslant 3$	MSE	d_H
(M1)	(N1)	0.000	12	0	0	88	0	0	0	0.164	10.836
	(N2)	0.010	0	0	0	100	0	0	0	0.017	0.172
	(N3)	0.030	28	0	0	69	1	1	1	0.295	24.709
	(N4)	0.000	4	0	0	51	21	5	19	0.057	5.670
	(N5)	0.000	27	0	0	51	17	4	1	0.280	23.969
	(N6)	0.010	13	0	0	46	17	18	6	0.162	13.701
	(N7)	0.020	12	0	0	81	4	2	1	0.130	10.805
	(N8)	0.000	17	0	0	70	9	4	0	0.180	15.453
	(N9)	0.000	0	0	0	100	0	0	0	0.018	0.221
(M2)	(N1)	0.000	0	1	36	63	0	0	0	0.015	0.951
	(N2)	0.010	0	0	13	87	0	0	0	0.008	0.425
	(N3)	0.050	0	0	5	84	10	1	0	0.006	1.082
	(N4)	0.000	61	0	0	39	0	0	0	0.274	47.371
	(N5)	0.000	22	3	43	32	0	0	0	0.053	12.371
	(N6)	0.010	58	0	0	42	0	0	0	0.275	46.489
	(N7)	0.030	0	0	18	75	5	2	0	0.009	0.970
	(N8)	0.000	2	0	30	66	1	1	0	0.010	1.550
	(N9)	0.000	0	0	100	0	0	0	0	0.009	2.024

Table 4.9: Distribution of $\hat{q}-q$ obtained by the α-adapted NOT algorithm for data generated according to (2.2.1) with the signals (M3)-(M6) and the noises (N1)-(N9), the average Mean Square Error (MSE) of the resulting estimate of the signal and the average Hausdorff distance d_H over 100 simulations. We also report the size, the proportion of replications with change-points being falsely detected when there are no change-points.

			$\hat{q}-q$								
Model	Noise	Size	$\leqslant 3$	-2	-1	0	1	2	$\geqslant 3$	MSE	d_H
(M3)	(N1)	0.000	0	0	0	100	0	0	0	0.023	0.020
	(N2)	0.080	0	0	0	100	0	0	0	0.064	0.080
	(N3)	0.200	0	0	0	98	2	0	0	0.047	0.910
	(N4)	0.370	0	0	11	66	17	4	2	0.890	12.770
	(N5)	0.420	0	0	0	75	15	7	3	0.282	10.210
	(N6)	0.410	0	0	3	64	22	4	7	0.434	13.060
	(N7)	0.150	0	0	0	97	3	0	0	0.035	0.630
	(N8)	0.160	0	0	2	92	4	1	1	0.274	2.880
	(N9)	0.000	0	0	0	100	0	0	0	0.011	0.040
(M4)	(N1)	0.010	0	0	0	100	0	0	0	0.012	0.067
	(N2)	0.030	0	0	0	100	0	0	0	0.006	0.007
	(N3)	0.100	0	0	0	100	0	0	0	0.003	0.000
	(N4)	0.080	0	0	23	76	1	0	0	0.472	7.963
	(N5)	0.110	0	0	0	94	6	0	0	0.033	1.340
	(N6)	0.120	0	0	13	82	4	1	0	0.273	5.657
	(N7)	0.070	0	0	0	100	0	0	0	0.010	0.000
	(N8)	0.040	0	0	1	94	4	1	0	0.104	2.143
	(N9)	0.000	0	0	0	100	0	0	0	0.006	0.067
(M5)	(N1)	0.030	0	0	0	99	1	0	0	0.025	0.307
	(N2)	0.080	0	0	0	100	0	0	0	0.083	0.173
	(N3)	0.150	0	0	0	93	5	1	1	0.024	1.600
	(N4)	0.350	0	1	1	58	22	14	4	0.364	8.853
	(N5)	0.470	0	0	0	78	18	1	3	0.282	4.600
	(N6)	0.380	0	1	0	50	22	13	14	0.213	11.253
	(N7)	0.190	0	0	0	94	5	0	1	0.031	1.253
	(N8)	0.260	0	0	1	88	6	4	1	0.216	2.567
	(N9)	0.000	0	0	0	100	0	0	0	0.014	0.060
(M6)	(N1)	0.010	0	0	0	100	0	0	0	0.018	0.240
	(N2)	0.030	0	0	0	99	1	0	0	0.006	0.123
	(N3)	0.130	0	0	0	97	3	0	0	0.045	0.727
	(N4)	0.100	0	19	0	76	3	2	0	0.882	13.410
	(N5)	0.140	0	2	4	91	3	0	0	0.155	3.457
	(N6)	0.050	0	21	5	68	4	0	2	1.025	17.027
	(N7)	0.100	0	0	0	99	1	0	0	0.019	0.167
	(N8)	0.040	0	0	18	76	6	0	0	0.231	9.257
	(N9)	0.000	0	0	0	100	0	0	0	0.016	0.150

Table 4.10: Distribution of $\hat{q}-q$ obtained by the α-adapted NOT algorithm for data generated according to (2.2.1) with the signals (M7)-(M10) and the noises (N1)-(N9), the average Mean Square Error (MSE) of the resulting estimate of the signal and the average Hausdorff distance d_H over 100 simulations. We also report the size, the proportion of replications with change-points being falsely detected when there are no change-points.

			$\hat{q}-q$								
Model	Noise	Size	$\leqslant 3$	-2	-1	0	1	2	$\geqslant 3$	MSE	d_H
(M7)	(N1)	0.000	0	0	0	99	1	0	0	0.016	0.027
	(N2)	0.030	0	0	0	99	1	0	0	0.092	0.247
	(N3)	0.100	0	0	0	97	3	0	0	0.05	0.217
	(N4)	0.100	20	0	0	68	8	2	2	1.897	18.623
	(N5)	0.130	2	0	0	96	2	0	0	0.318	2.003
	(N6)	0.100	23	0	0	62	11	2	2	2.249	21.617
	(N7)	0.150	0	0	0	93	4	2	1	0.023	1.037
	(N8)	0.090	0	0	0	97	1	1	1	0.109	0.537
	(N9)	0.000	0	0	0	100	0	0	0	0.007	0.017
(M8)	(N1)	0.000	0	0	0	100	0	0	0	0.017	0.102
	(N2)	0.020	0	0	0	100	0	0	0	0.009	0.044
	(N3)	0.200	0	0	0	98	2	0	0	0.008	0.218
	(N4)	0.030	67	0	0	33	0	0	0	1.925	53.600
	(N5)	0.020	2	0	0	96	2	0	0	0.092	1.830
	(N6)	0.010	47	0	0	51	1	0	1	1.356	37.794
	(N7)	0.040	0	0	0	100	0	0	0	0.016	0.062
	(N8)	0.000	7	0	1	88	4	0	0	0.311	6.954
	(N9)	0.000	0	0	0	100	0	0	0	0.013	0.096
(M9)	(N1)	0.030	0	0	0	100	0	0	0	0.013	0.002
	(N2)	0.000	0	0	0	100	0	0	0	0.041	0.032
	(N3)	0.150	0	0	0	99	1	0	0	0.033	0.160
	(N4)	0.010	31	0	0	65	4	0	0	2.796	25.628
	(N5)	0.120	20	0	0	76	2	2	0	1.905	16.734
	(N6)	0.030	16	0	0	74	5	5	0	1.462	13.960
	(N7)	0.070	0	0	0	99	1	0	0	0.016	0.094
	(N8)	0.000	4	0	0	94	2	0	0	0.448	3.538
	(N9)	0.000	0	0	0	100	0	0	0	0.008	0.024
(M10)	(N1)	0.010	0	0	0	100	0	0	0	0.015	0.066
	(N2)	0.030	0	0	0	99	1	0	0	0.004	0.018
	(N3)	0.040	0	0	0	100	0	0	0	0.013	0.057
	(N4)	0.010	6	0	0	87	5	1	1	0.546	5.612
	(N5)	0.000	2	0	0	93	5	0	0	0.222	2.118
	(N6)	0.000	9	0	0	83	4	4	0	0.829	8.192
	(N7)	0.030	0	0	0	99	1	0	0	0.012	0.040
	(N8)	0.000	8	0	0	92	0	0	0	0.791	6.572
	(N9)	0.000	0	0	0	100	0	0	0	0.012	0.053

Table 4.11: Distribution of $\hat{q}-q$ obtained by the α-adapted NOT algorithm for data generated according to (2.2.1) with the signals (M11)-(M14) and the noises (N1)-(N9), the average Mean Square Error (MSE) of the resulting estimate of the signal and the average Hausdorff distance d_H over 100 simulations. We also report the size, the proportion of replications with change-points being falsely detected when there are no change-points.

			$\hat{q}-q$								
Model	Noise	Size	$\leqslant 3$	-2	-1	0	1	2	$\geqslant 3$	MSE	d_H
(M11)	(N1)	0.000	0	0	0	100	0	0	0	0.010	0.009
	(N2)	0.040	0	0	0	100	0	0	0	0.025	0.028
	(N3)	0.050	0	0	0	98	2	0	0	0.011	0.126
	(N4)	0.000	24	0	0	75	1	0	0	2.182	22.741
	(N5)	0.000	18	0	1	80	2	0	0	1.484	15.761
	(N6)	0.010	3	0	0	81	8	4	4	0.296	4.431
	(N7)	0.010	0	0	0	98	1	0	1	0.013	0.147
	(N8)	0.000	0	0	0	100	0	0	0	0.066	0.033
	(N9)	0.000	0	0	0	100	0	0	0	0.007	0.027
(M12)	(N1)	0.000	0	0	0	100	0	0	0	0.013	0.075
	(N2)	0.020	0	0	0	100	0	0	0	0.002	0.081
	(N3)	0.050	0	0	0	97	2	1	0	0.003	0.202
	(N4)	0.000	10	0	0	89	1	0	0	0.779	15.769
	(N5)	0.010	2	0	0	98	0	0	0	0.100	1.757
	(N6)	0.000	42	0	0	58	0	0	0	2.206	44.625
	(N7)	0.010	0	0	0	100	0	0	0	0.013	0.034
	(N8)	0.000	0	0	51	49	0	0	0	0.123	5.622
	(N9)	0.000	0	0	0	100	0	0	0	0.013	0.056
(M13)	(N1)	0.000	0	0	0	100	0	0	0	0.011	0.007
	(N2)	0.030	0	0	0	100	0	0	0	0.033	0.037
	(N3)	0.050	0	0	0	79	16	4	1	0.006	0.223
	(N4)	0.000	64	0	0	36	0	0	0	7.509	67.213
	(N5)	0.010	40	0	0	60	0	0	0	4.739	37.275
	(N6)	0.010	55	0	0	42	2	0	1	6.422	51.237
	(N7)	0.010	0	0	0	96	3	0	1	0.016	0.093
	(N8)	0.000	2	0	1	97	0	0	0	0.301	2.004
	(N9)	0.000	0	0	0	100	0	0	0	0.004	0.014
(M14)	(N1)	0.000	0	0	0	100	0	0	0	0.009	0.014
	(N2)	0.010	0	0	0	100	0	0	0	0.007	0.008
	(N3)	0.050	0	0	0	100	0	0	0	0.017	0.029
	(N4)	0.000	27	0	0	73	0	0	0	1.744	25.650
	(N5)	0.000	0	0	0	98	1	1	0	0.010	0.047
	(N6)	0.000	6	0	0	93	1	0	0	0.391	5.730
	(N7)	0.030	0	0	0	100	0	0	0	0.009	0.002
	(N8)	0.020	0	0	1	99	0	0	0	0.054	0.217
	(N9)	0.000	0	0	0	100	0	0	0	0.008	0.031

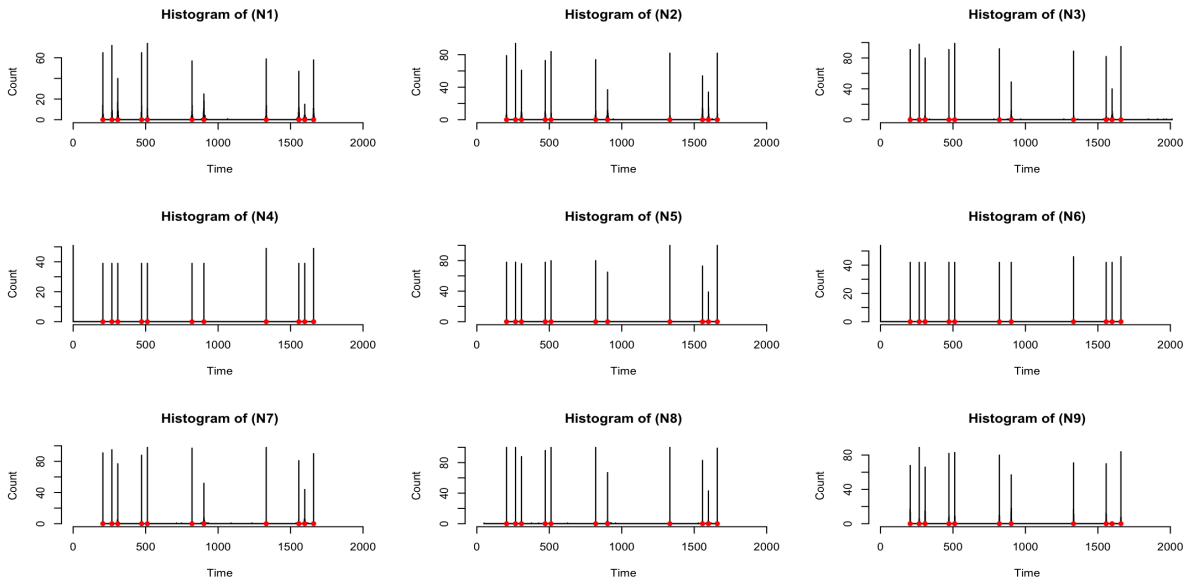


Figure 4.10: Histograms plotting the estimated change-points for data produced by signal (M2) and noises (N1)-(N9) by the α-adapted NOT solution path algorithm. The true change-points are coloured red.

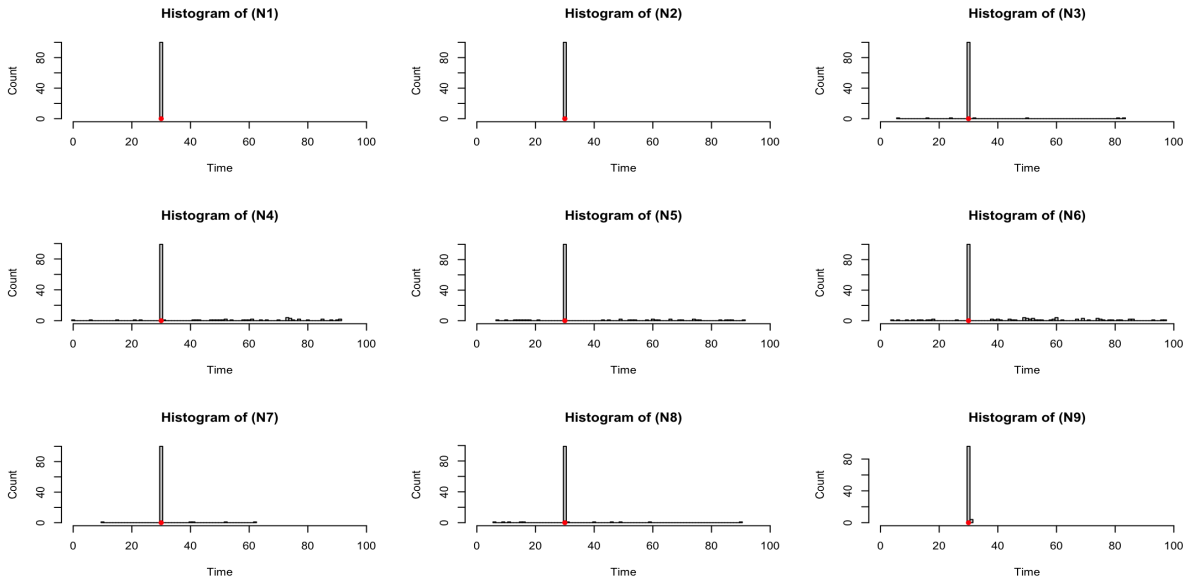


Figure 4.11: Histograms plotting the estimated change-points for data produced by signal (M3) and noises (N1)-(N9) by the α-adapted NOT solution path algorithm. The true change-points are coloured red.

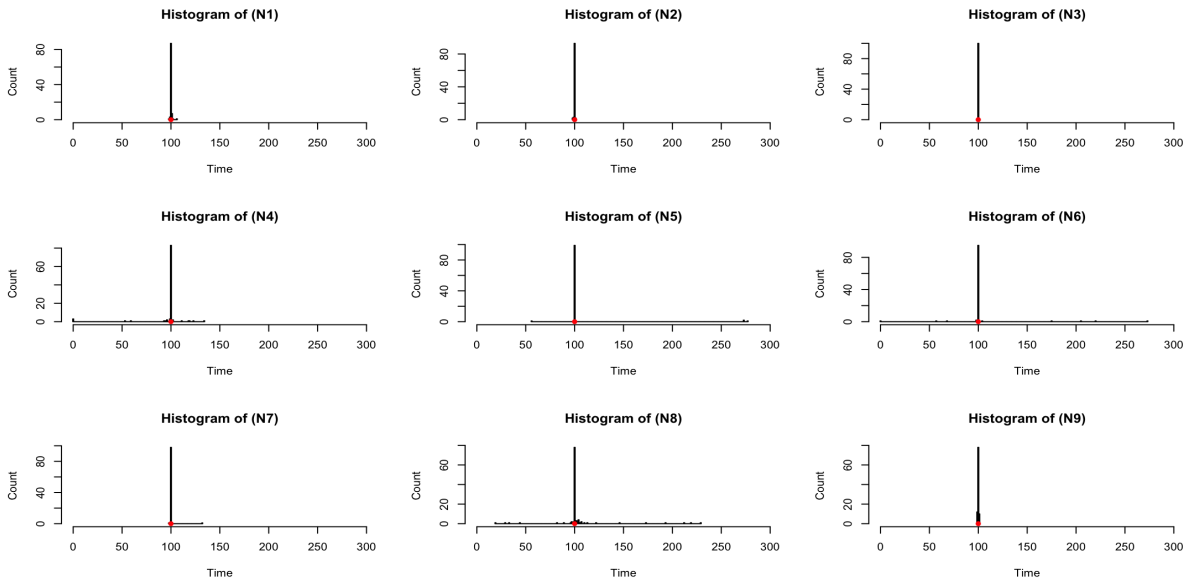


Figure 4.12: Histograms plotting the estimated change-points for data produced by signal (M4) and noises (N1)-(N9) by the α-adapted NOT solution path algorithm. The true change-points are coloured red.

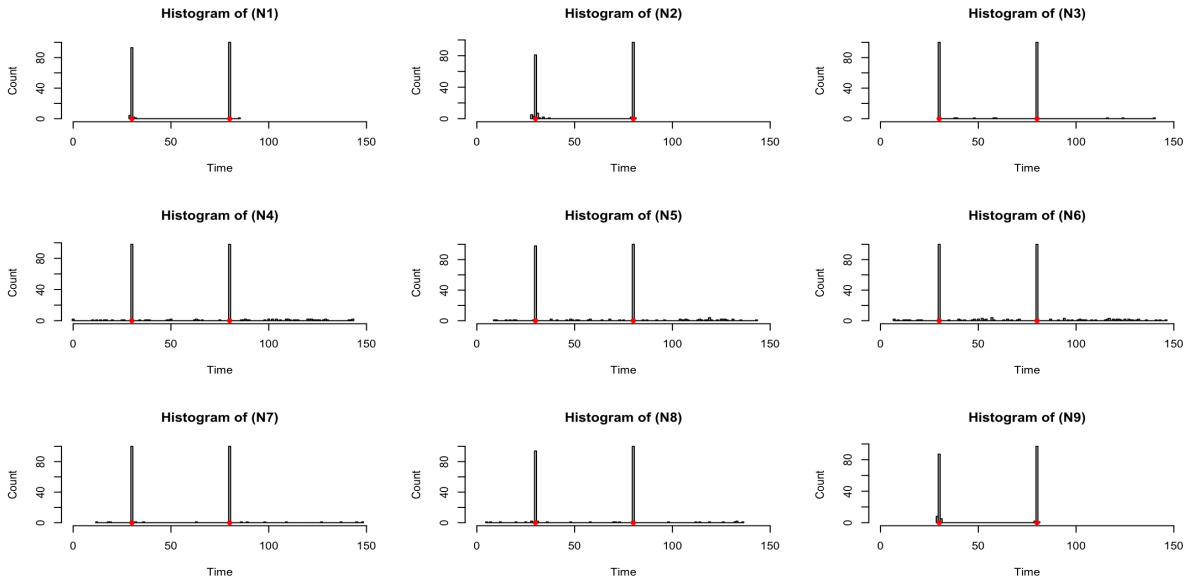


Figure 4.13: Histograms plotting the estimated change-points for data produced by signal (M5) and noises (N1)-(N9) by the α-adapted NOT solution path algorithm. The true change-points are coloured red.

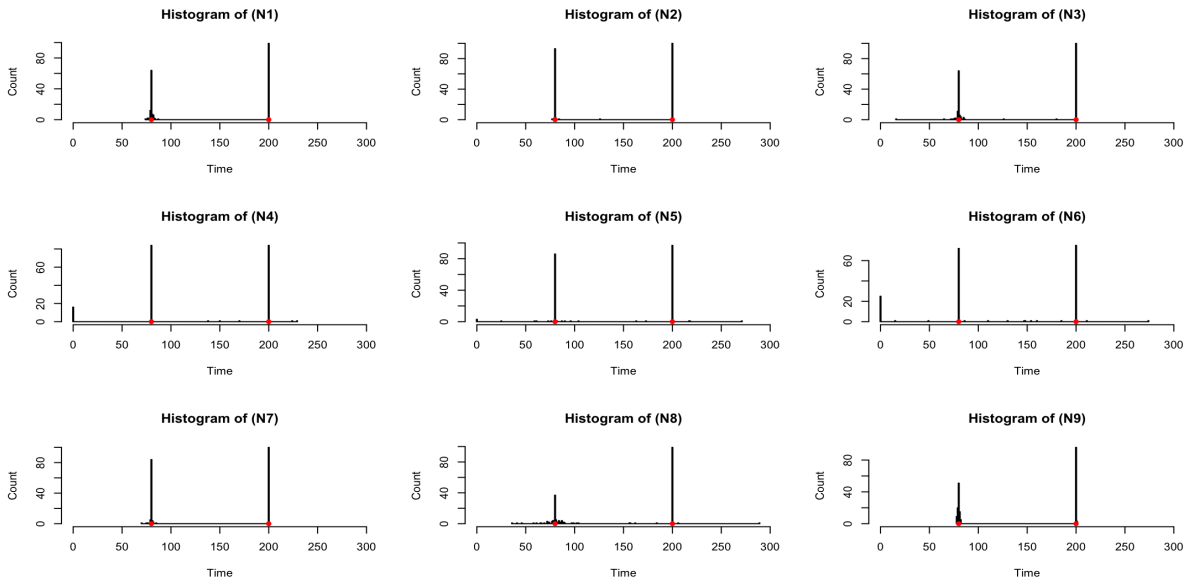


Figure 4.14: Histograms plotting the estimated change-points for data produced by signal (M6) and noises (N1)-(N9) by the α-adapted NOT solution path algorithm. The true change-points are coloured red.

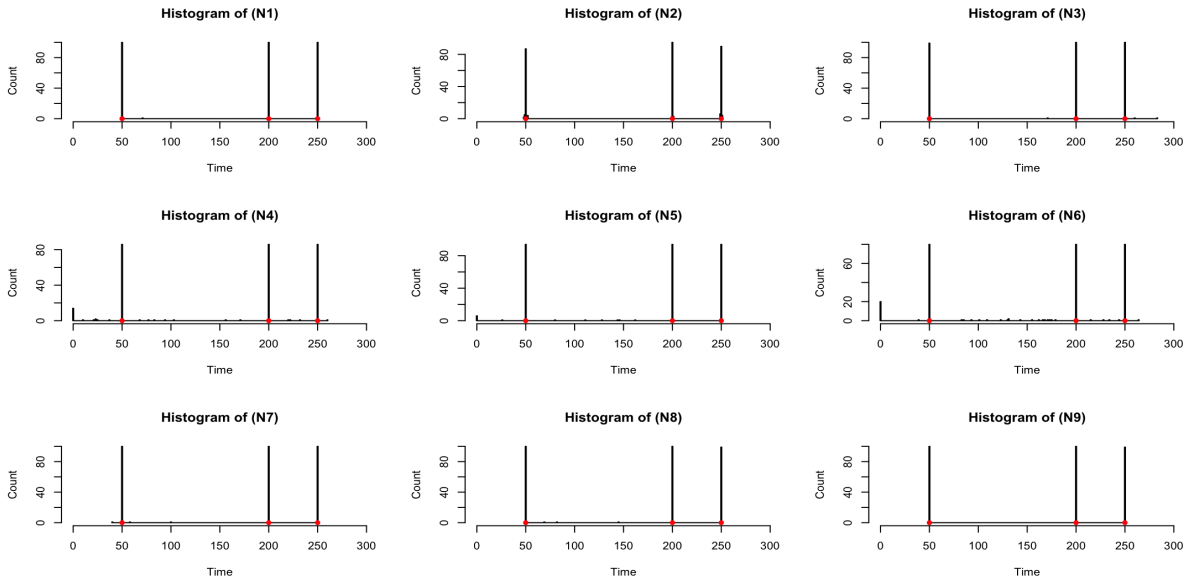


Figure 4.15: Histograms plotting the estimated change-points for data produced by signal (M7) and noises (N1)-(N9) by the α-adapted NOT solution path algorithm. The true change-points are coloured red.

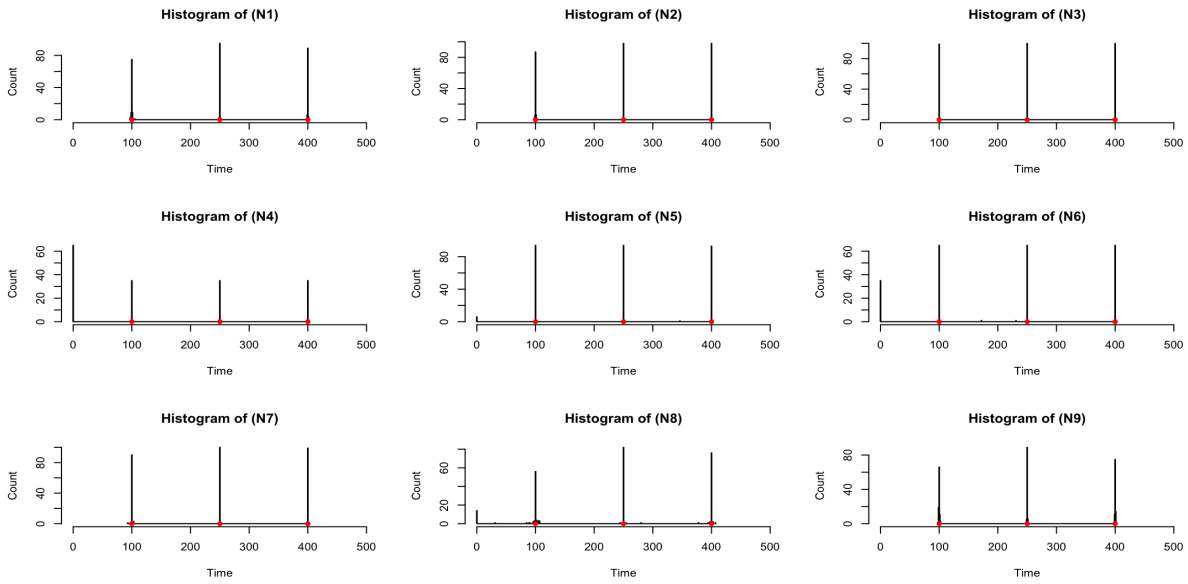


Figure 4.16: Histograms plotting the estimated change-points for data produced by signal (M8) and noises (N1)-(N9) by the α-adapted NOT solution path algorithm. The true change-points are coloured red.

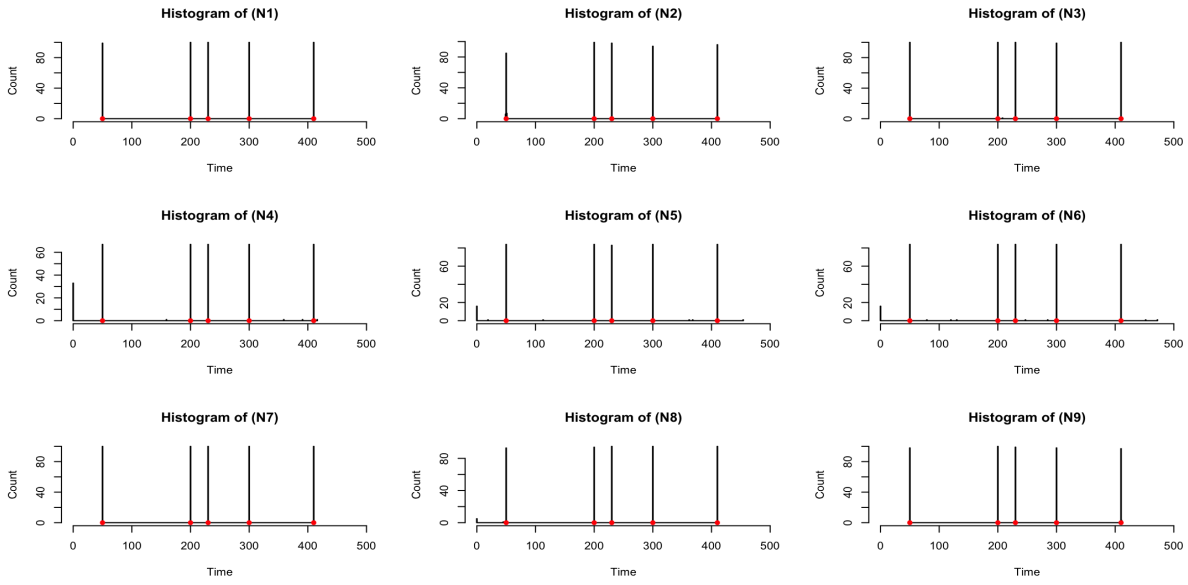


Figure 4.17: Histograms plotting the estimated change-points for data produced by signal (M9) and noises (N1)-(N9) by the α-adapted NOT solution path algorithm. The true change-points are coloured red.

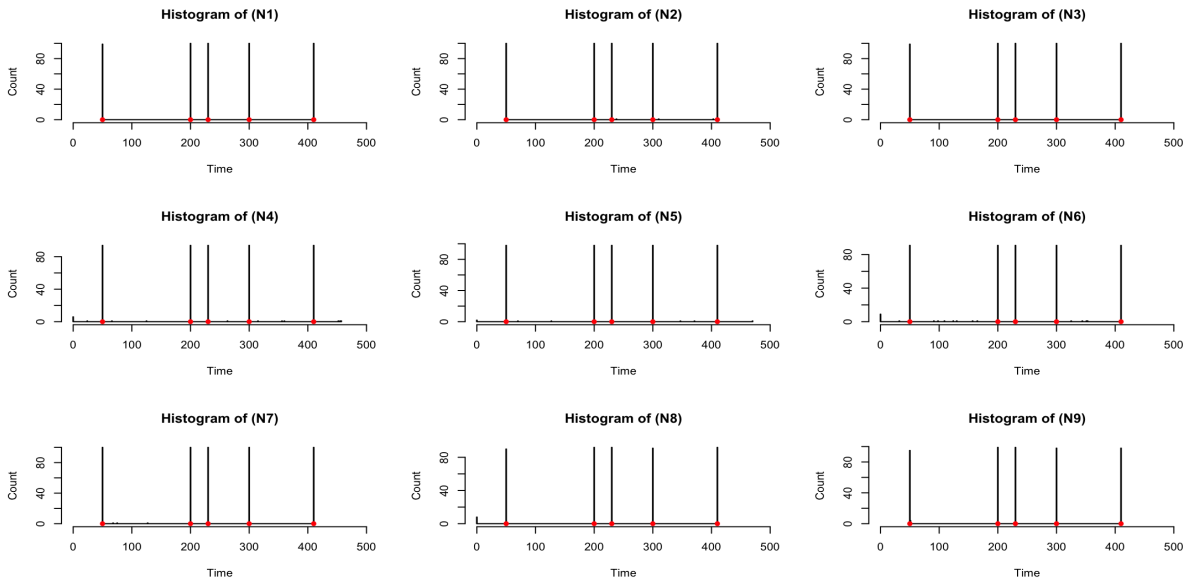


Figure 4.18: Histograms plotting the estimated change-points for data produced by signal (M10) and noises (N1)-(N9) by the α-adapted NOT solution path algorithm. The true change-points are coloured red.

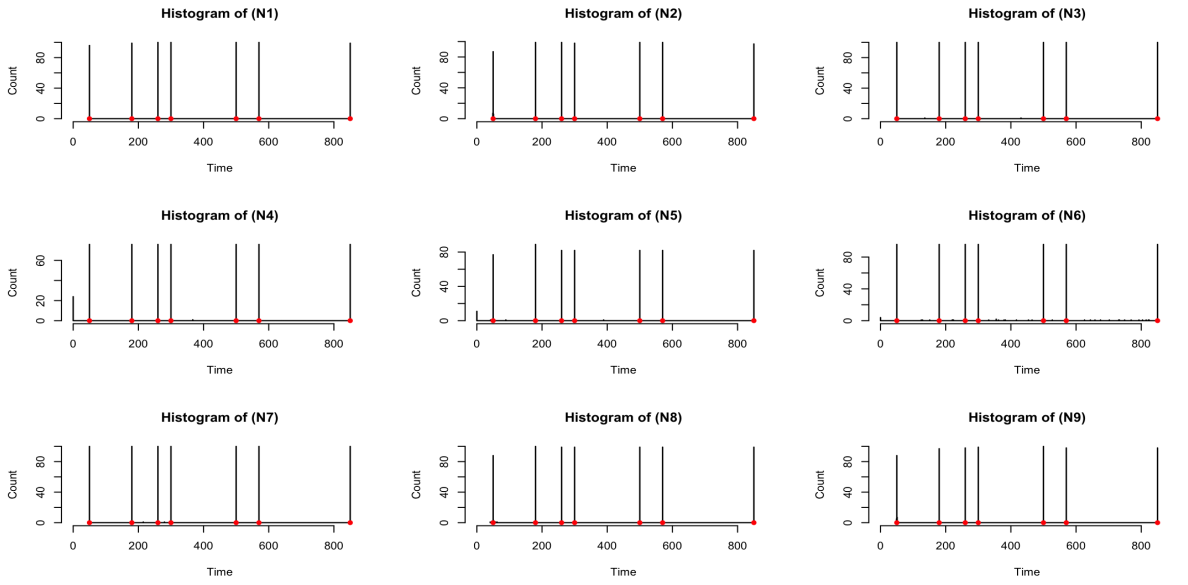


Figure 4.19: Histograms plotting the estimated change-points for data produced by signal (M11) and noises (N1)-(N9) by the α-adapted NOT solution path algorithm. The true change-points are coloured red.

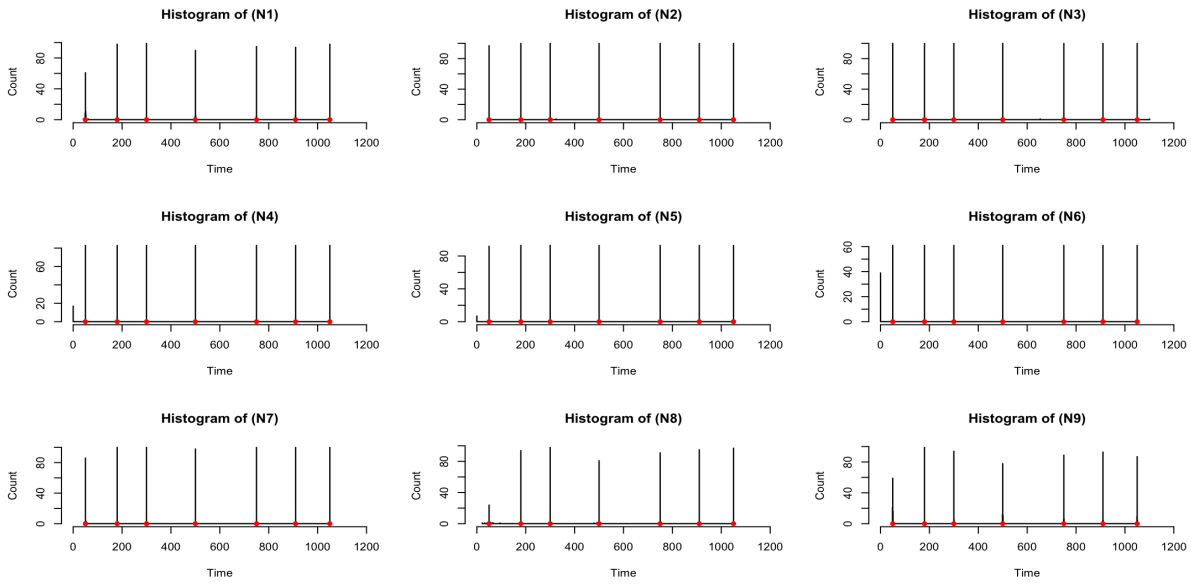


Figure 4.20: Histograms plotting the estimated change-points for data produced by signal (M12) and noises (N1)-(N9) by the α-adapted NOT solution path algorithm. The true change-points are coloured red.

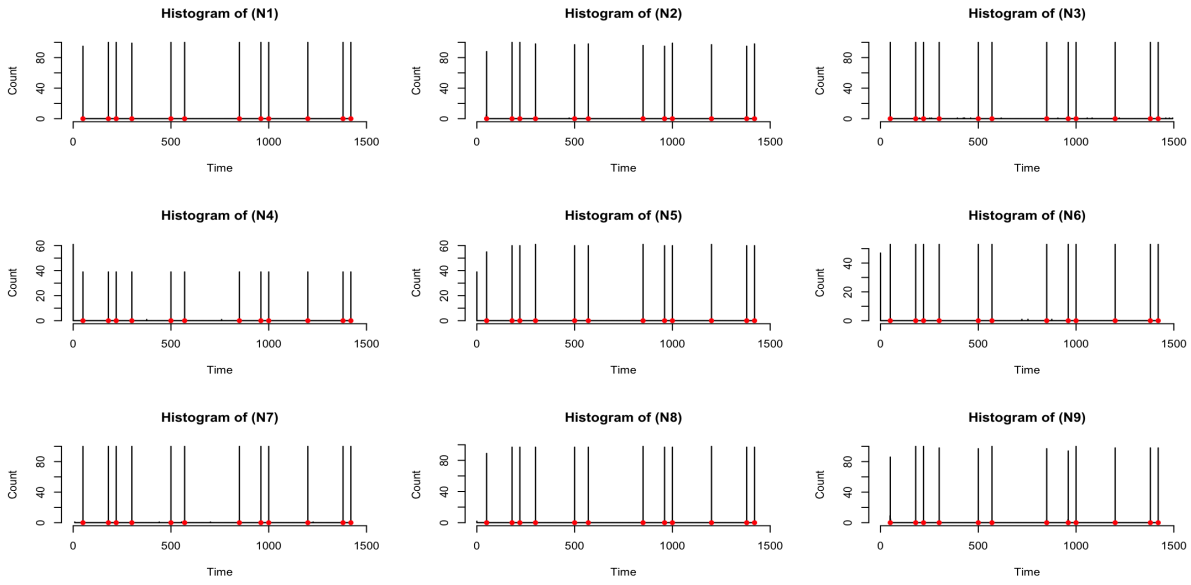


Figure 4.21: Histograms plotting the estimated change-points for data produced by signal (M13) and noises (N1)-(N9) by the α-adapted NOT solution path algorithm. The true change-points are coloured red.

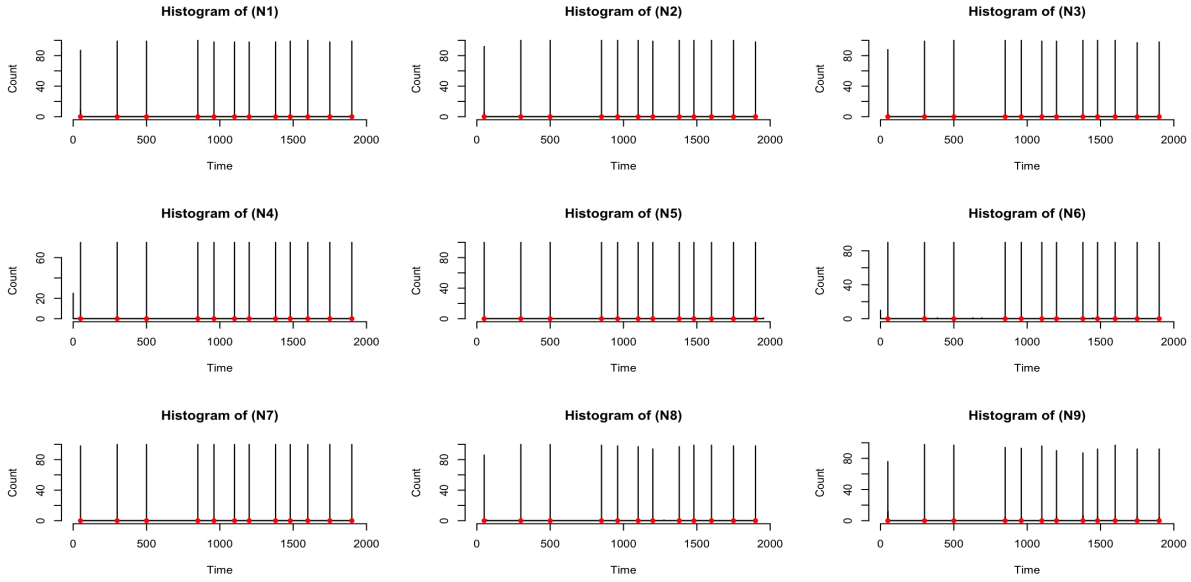


Figure 4.22: Histograms plotting the estimated change-points for data produced by signal (M14) and noises (N1)-(N9) by the α-adapted NOT solution path algorithm. The true change-points are coloured red.

4.5 Appendix – Complete Simulation Models

In addition to (M1) and (M2), we simulate more examples under the following scenarios. We assess the performance of our methods in the case of both no change-points ($q = 0$) and one or more change-points ($q \geqslant 1$) under different sample sizes, where $\theta_i = 0$ is considered in each scenario to show the false detection rate under different settings of the error processes.

(M3) f_t undergoes $q = 1$ change-points at $\eta_1 = 30$ with $n = 100$ and $(\theta_1, \theta_2) = (2.5, -2.5)$;

(M4) f_t undergoes $q = 1$ change-points at $\eta_1 = 100$ with $n = 200$ and $(\theta_1, \theta_2) = (1.5, -1.5)$;

(M5) f_t undergoes $q = 2$ change-points at $(\eta_1, \eta_2) = (30, 80)$ with $n = 150$ and $(\theta_1, \theta_2, \theta_3) = (-1, 3, -3)$;

(M6) f_t undergoes $q = 2$ change-points at $(\eta_1, \eta_2) = (80, 200)$ with $n = 300$ and $(\theta_1, \theta_2, \theta_3) = (0, 2, -2)$;

(M7) f_t undergoes $q = 3$ change-points at $(\eta_1, \eta_2, \eta_3) = (50, 200, 250)$ with $n = 300$ and $(\theta_1, \theta_2, \theta_3, \theta_4) = (-2, 4, -3, 3.5)$;

(M8) f_t undergoes $q = 3$ change-points at $(\eta_1, \eta_2, \eta_3) = (100, 250, 400)$ with $n = 500$ and $(\theta_1, \theta_2, \theta_3, \theta_4) = (0, 2.5, -1.5, 2)$;

(M9) f_t undergoes $q = 5$ change-points at $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (50, 200, 230, 300, 410)$ with $n = 500$ and $(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6) = (-2, 3, -4, 2.5, -3.5, 3)$;

(M10) f_t undergoes $q = 5$ change-points at $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (50, 200, 300, 500, 650)$ with $n = 750$ and $(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6) = (0, 2.5, -2, 2, -2, 1.5)$;

(M11) f_t undergoes $q = 7$ change-points at $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5, \eta_6, \eta_7) = (50, 180, 260, 300, 500, 570, 850)$ with $n = 900$ and $(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_7, \theta_8) = (0, 4, -3, 3.5, -2.5, 3.5, -3, 3)$;

(M12) f_t undergoes $q = 7$ change-points at $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5, \eta_6, \eta_7) = (50, 180, 300, 500, 750, 910, 1050)$ with $n = 1200$ and $(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_7, \theta_8) = (0, 2, -3, 1.5, -2, 2, -2, 2.5)$.

(M13) f_t undergoes $q = 12$ change-points at $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5, \eta_6, \eta_7, \eta_8, \eta_9, \eta_{10}, \eta_{11}, \eta_{12}) = (50, 180, 220, 300, 500, 570, 850, 960, 1000, 1200, 1380, 1420)$ with $n = 1500$ and $(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_7, \theta_8, \theta_9, \theta_{10}, \theta_{11}, \theta_{12}, \theta_{13}) = (0, 4, -4.5, 4, -2.5, 3.5, -3.5, 3, -3, 4, -3, 4, -4)$;

(M14) f_t undergoes $q = 12$ change-points at $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5, \eta_6, \eta_7, \eta_8, \eta_9, \eta_{10}, \eta_{11}, \eta_{12}) = (50, 300, 500, 850, 960, 1000, 1200, 1380, 1480, 1600, 1750, 1900)$ with $n = 2000$ and $(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_7, \theta_8, \theta_9, \theta_{10}, \theta_{11}, \theta_{12}, \theta_{13}) = (0, 2, -2.5, 3, -2.5, 2.5, -2.5, 2, -2, 2.5, -3, 2, -2.5)$;

Models (M4)-(M8) consider relatively shorter time series with sample size $n \in [100, 300]$. We generate 100 replications under each combination of the signal model above and the noise model defined in later sections (unless stated otherwise).

Chapter 5

Multi-scale viewpoint in estimating the strength of lead-lag relationships between nonstationary time series

5.1 Motivation

In this chapter, we intend to introduce an exploratory technique that could be able to serve as the first step in analyses for lead-lag relationships or even causality in time series. Since December 2019, the severe coronavirus disease 2019 (COVID-19) pandemic became a global outbreak and resulted in numerous cases, where the severest cases can finally lead to death. To manage or prevent such disease, various government policies were officially approved in different countries. The large amount of information available seems to be a challenge, but it provides us with informative datasets for statistical analyses. For example, Figure 5.1 shows the recorded new cases and deaths in Mathieu et al. (2020), where sub-figures (b) and (d) display the dataset recorded

at the subsequent 200 days from 2020-02-01. As seen, the common pattern observed in both curves (b) and (d), i.e. 200-day new cases and deaths, involves an generally upward movement, a peak, and then a downward trend. Considering the COVID data, it is reasonable to believe that the changes in the number of new deaths will react to that of new cases. Hence the tight association can heuristically indicate the lead-lag relationship between these two time series. However, the number of new cases exhibits larger but less frequent fluctuations than that of new deaths, which may somehow enhance the importance of choosing parameters such as the number of regression variables and order of regression in conventional approaches. Meanwhile, without data preprocessing, the distinct units (metrics) of the dataset for new cases and deaths also make lead-lag relationship less apparent. These issues demonstrate the vital role of the selection of window size before making any statistical inference on an interrelationship. Therefore, with the aim of obtaining a basic understanding of data, a multi-scale graphical device across time can be convenient for identifying the availability and location of the features of interest.

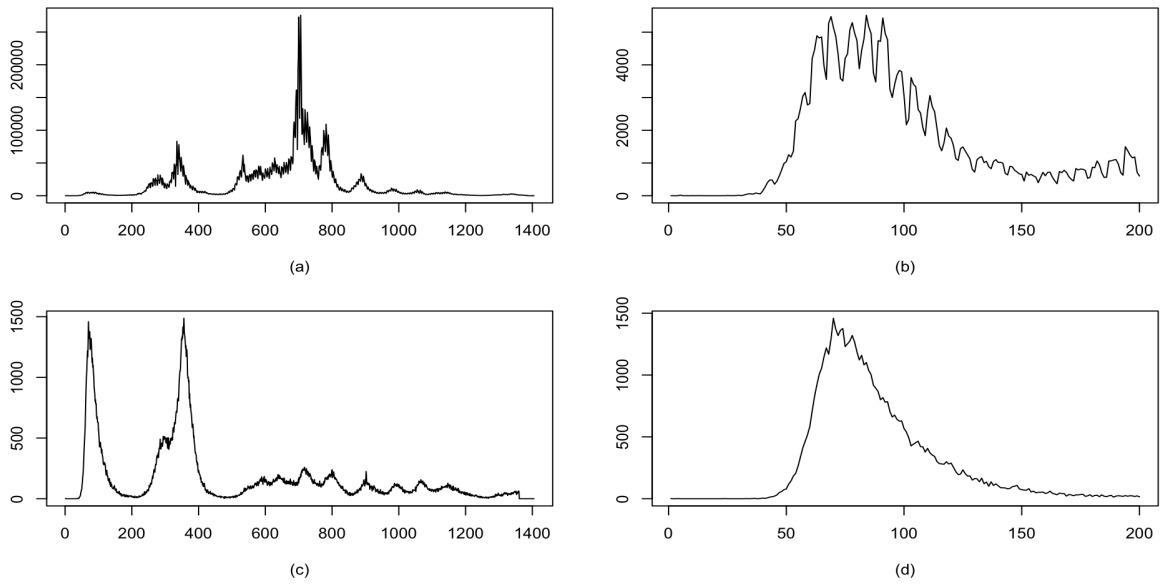


Figure 5.1: New cases (a) and new deaths (c) attributed to COVID-19 in United Kingdom recorded from 2020-02-01 to 2023-12-06, and (b) and (d) display the corresponding data recorded at the subsequent 200 days from 2020-02-01.

Motivated by such COVID curves collected by Mathieu et al. (2020), we consider the issue of discovering the lead-lag relationships among nonstationary time series, and attempt to introduce a graphical method to display the significant features captured from data without much preprocessing works. Meanwhile, instead of focusing on the direction of such a relationship, we pay more attention to figuring out the significance of the (possible) lead-lag phenomenon in bi-variate data with natural direction, hoping that it can serve as a reasonable first step in lead-lag or causal analyses.

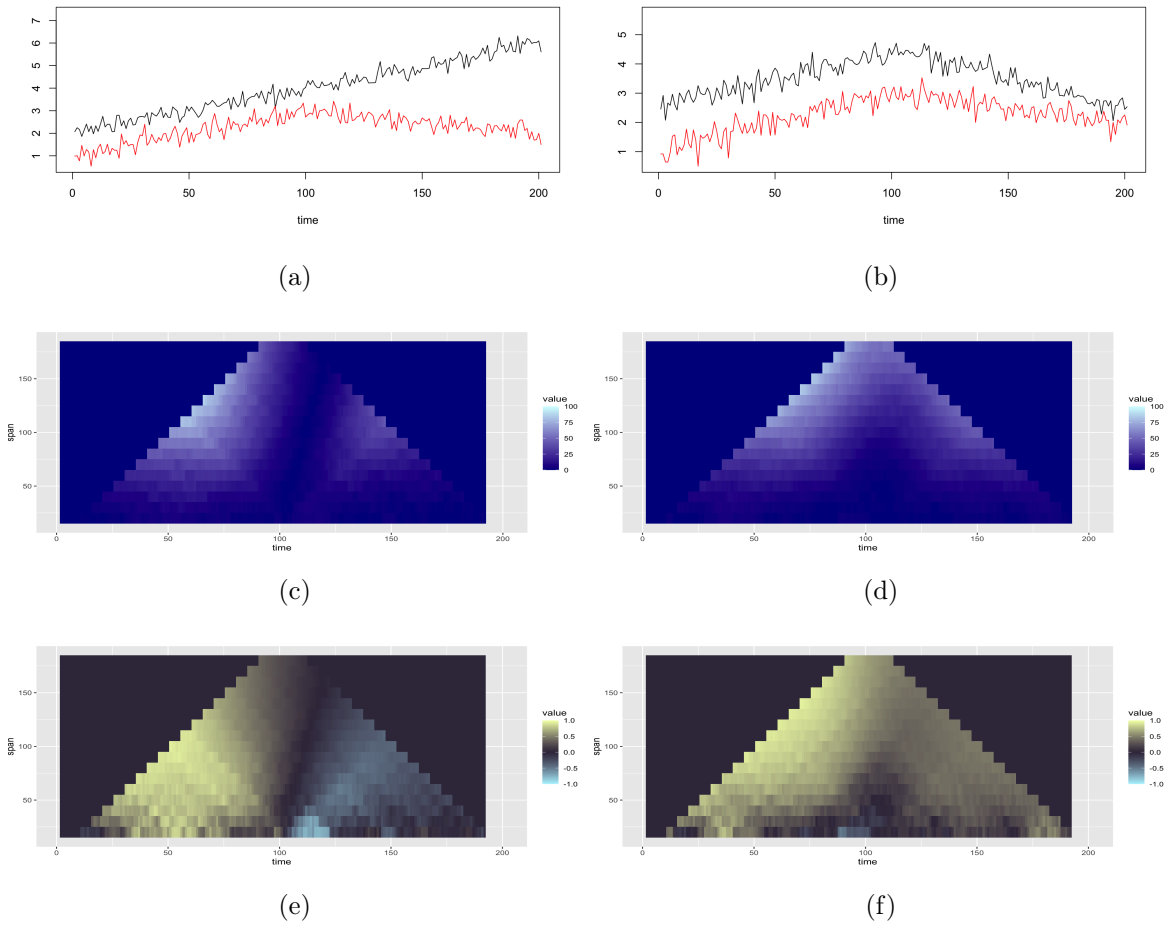
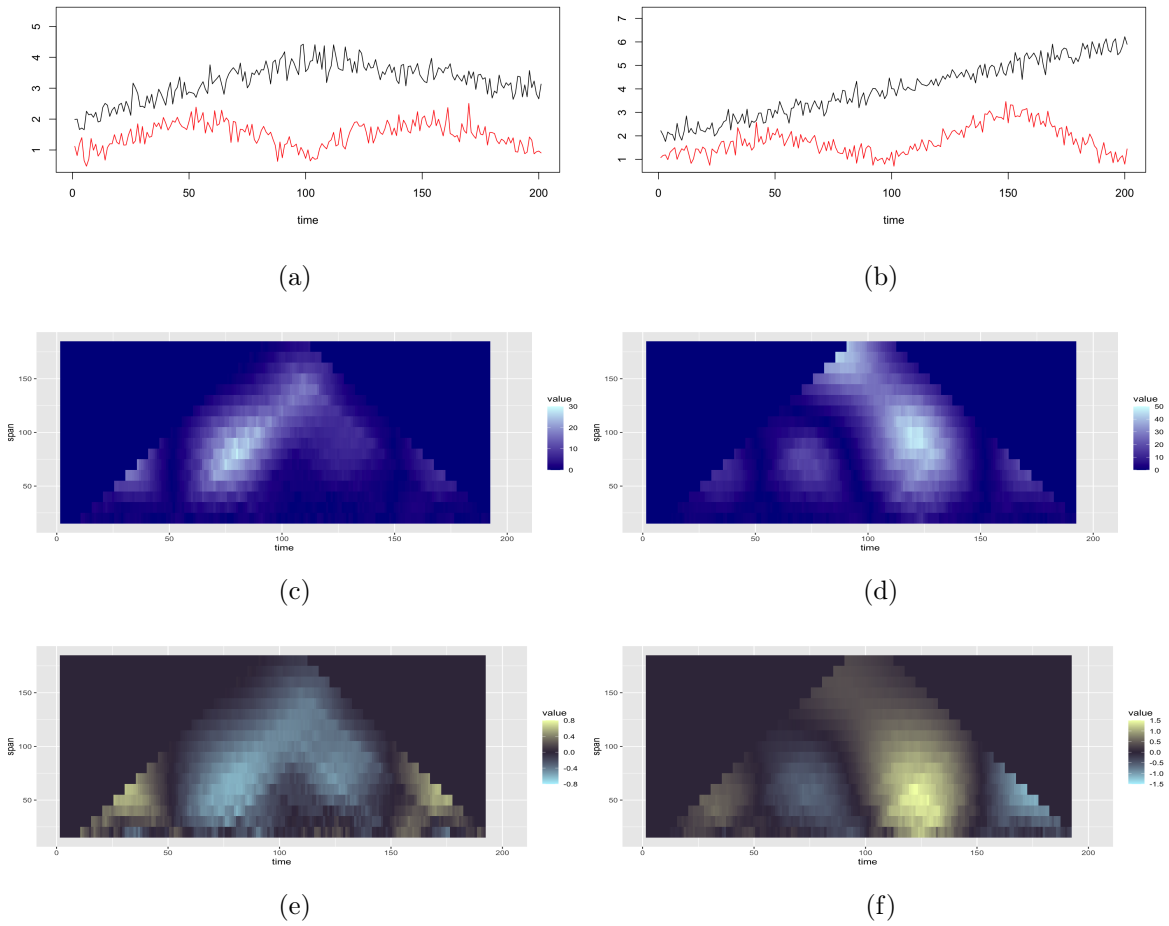


Figure 5.2: Combination of simulated samples with piecewise-linear signal and *iid* Gaussian noise [(a) and (b)], heatmaps of $-\log(p)$ [(c) and (d)] and heatmaps of coefficients [(e) and (f)]. The black (red) lines in (a) and (b) are the simulated regressors X_i (dependent variables Y_i).

In particular, we provide several simple examples of (X_i, Y_i) without any time lags in Figure 5.2 and 5.3 to show the basic patterns to be identified with our method proposed in the next section. From Figure 5.2, we see that the process X_i in sample (a) displays a consistent upward trend while Y_i experiences an initial increase and a subsequent decrease. In contrast, X_i and Y_i in sub-figure (b) demonstrate a similar pattern of change. It is evident that the co-movements of (X_i, Y_i) in either the same or opposite direction bring bright colours in the heatmaps of $-\log(p)$, indicating the presence of significant correlations and (possible) lead-lag relationships between sequences. Additionally, the occurrence of a change in the direction of these co-movements is reflected by a shift between yellow and blue in the heatmaps of coefficients. Furthermore, considering the largest scales, if no obvious colour shift happens in the coefficient heatmap, we can discover the existence and the potential overall direction of significant correlations between sequences X_i and Y_i over the tested range, see sub-figure (f) as an example; otherwise, the colour shown at those scales could indicate the dominance of the significant correlation between (X_i, Y_i) within its corresponding range. Specifically, the colour yellow at the largest scales in sub-figure (e) implies that the dependence tested within $[1, 100]$ is significant enough to dominate the entire sample (X_i, Y_i).

Following the identification process above, we shall investigate the mentioned patterns in Figure 5.3 with more detailed explanations. In sub-figure (a), the process X_i first experiences an upward trend and then follows a subsequent downward movement with a flatter slope whereas Y_i reaches the peak twice with the same slope of ascent and descent. Sub-figure (b) shows another example that X_i keeps increasing while Y_i goes through two peaks with consistent increasing and decreasing rates but the slope for the second peak is sharper. Besides the apparent shifts in colours in sub-figures (e) and (f), diagonal stripes at 45° angles are noticeable in all four heatmaps, especially in sub-figures (c) and (d), which help indicate the consistent co-movement of two sequences within different windows and the coincidence of the trends around the locations of the spikes.



(a) (b)

(c) (d)

(e) (f)

Figure 5.3: Combination of simulated time series with piecewise-linear signal and *iid* Gaussian noise [(a) and (b)], heatmaps of $-\log(p)$ [(c) and (d)] and heatmaps of coefficients [(e) and (f)]. The black (red) lines in (a) and (b) are the simulated regressors (dependent variables).

Meanwhile, the width and length (total area) of the brightest stripe can serve as an indicator of the strength of the most significant dependence between sequences. For example, the positive-sloping (negative-sloping) diagonal stripes starting around $i = 20$ and $i = 60$ (concluding around $i = 140$ and $i = 180$) displayed in sub-figure (c) align with the period where X_i and Y_i experiences prominent co-movements within approximate ranges of $[10, 50]$ and $[50, 100]$ ($[100, 150]$ and $[150, 190]$). Among all these

stripes, we can see the brightest one locates within the range $[60, 90]$ but does not vanish too much till around $i = 110$ at a large window size 140, which demonstrates that the correlation of processes over range $[60, 90]$ is indeed significant and can dominate the surrounding observations.

This overall picture provides us with two potential ways of separating the whole sample size into subsamples that contain similar information for further lead-lag analysis over each one. On the one hand, we could partition the time series into equal-sized overlapping or non-overlapping blocks by first identifying a "best" scale (window size) where we can observe comprehensive data features, such as scale 50 in Figure 5.3. Specifically, starting from this scale, the dependence patterns become more stable, indicated by much fewer changes in colour within coefficient heatmaps and increased significance in corresponding p-values; Meanwhile, all changes in co-movements have not been covered by more dominant patterns and hence remain discernible. On the other hand, we could simply focus on the ranges identified by colours in coefficient heatmaps, such as $[10, 50]$, $[50, 150]$ and $[150, 190]$ in sub-figure (e) and $[10, 50]$, $[50, 100]$, $[100, 150]$ and $[150, 190]$ in sub-figure (f). Different modelling methods are then encouraged to capture potential lead-lag relationships within each subsample. In the following sections, we shall employ our proposed algorithm to first investigate these basic features in different sorts of data. Furthermore, studies may also be conducted to see the possible choice of time lags between bi-variate time series.

The remainder of this chapter is organised as follows. Section 5.2 provides a full description of the multi-scale lead-lag heatmaps and the simple algorithm behind it. In Section 5.3, we present the a comprehensive simulation study over different pairs of time series and the performance of our algorithm is examined in Section 5.4 via data examples of historical data on the COVID-19 pandemic collected up to 2023-12-06. In Section 5.5, more visualisation examples are displayed and we provide a brief discussion in the final section.

5.2 Methodology

To extract features in data, many smoothing methods have been developed. Compared to classic approaches, SiZer (Chaudhuri and Marron, 1999) is produced relying on the scale-space viewpoint, i.e. analysing zero-crossings over a wide range of bandwidths at the same time, and attaches more importance to the observed data instead of estimating the true underlying model. Inspired by such a multi-scale idea, we decide to develop the heatmap over different moving windows to study the potential lead-lag relationship between two series over both location and scale in time, see Figure 5.5 as a simple example. This method is designed to highlight significant correlations between sequences by showing the areas where the minus log p-value, $-\log(p)$, corresponding to its non-zero regression coefficient is large enough. As stated in Algorithm 2, for the sake of simplicity, our methodology is built on simple linear regression under the Gaussian assumption. Mathematically speaking, given time series observations (X_i, Y_i), at $i = t-h, t-h+1, \ldots, t+h-1$, we assume a linear model with a rolling time window

$$Y_i = \beta^0_{h,t} + \beta^1_{h,t} X_i + \varepsilon_i \tag{5.2.1}$$

where $\beta^0_{h,t}$ and $\beta^1_{h,t}$ are unknown parameters and errors ε_i $(i = t-h, \ldots, t+h-1)$ are such that

- $\mathbb{E}(\varepsilon_i) = 0$ for any i;
- $\mathrm{Var}(\varepsilon_i) = \sigma^2$ for any i (homoskedasticity);
- $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for any $i \neq j$;
- $\mathbb{E}(\varepsilon_i | X_i) = 0$ (exogeneity);
- ε_i's are *iid* Gaussian errors.

To estimate $\beta^1_{h,t}$, we apply the unbiased *ordinary least squares (OLS)* estimator, and hence the difference $\hat{\beta}^1_{h,t} - \beta^1_{h,t}$ is Gaussian, leading to the test statistic following a

t-distribution. To clarify, under the null hypothesis $H_0 : \beta^1_{h,t} = 0$, the test statistic is $\hat{\beta}^1_{h,t}/\mathrm{SE}(\hat{\beta}^1_{h,t})$ and the corresponding $-\log(p)$ over each moving window is derived from

$$p = P\left(|T| > \left|\frac{\hat{\beta}^1_{h,t}}{\mathrm{SE}(\hat{\beta}^1_{h,t})}\right| \middle| H_0\right) \tag{5.2.2}$$

where $\mathrm{SE}(\hat{\beta}^1_{h,t})$ is the estimated standard error.

Algorithm 2 presents the general procedure of deriving the Multi-scale Lead-Lag Heatmaps (MLLH). The resulted heatmaps present the features simultaneously over location and scale (window size). To be specific, the blue colour scheme in Figure 5.5(b) is lighter where the p-value is closer to 0. In Figure 5.5(c), we can see the colour yellow (blue) when the coefficient is positive (negative) and the lighter the colour is, the larger the magnitude is. The heatmap as a whole follows the shape of a trapezium since we can provide features at fewer locations when considering a larger moving window size.

Algorithm 2 Multi-scale Lead-Lag Heatmap

1: **Input:** Data vectors $\mathbf{X}$ and $\mathbf{Y}$, a wide range of moving window sizes h, together with
2: $\mathbf{X}_{t-h,t+h-1} = (X_{t-h}, \ldots, X_{t+h-1})^\intercal$ and $\mathbf{Y}_{t-h,t+h-1} = (Y_{t-h}, \ldots, Y_{t+h-1})^\intercal$
3: For any $h, t > 0$ satisfying $0 < t - h < t + h - 1 \leqslant n$, build the linear regression model
4: (allowing for other models as well)
5: $Y_i = \beta^0_{h,t} + \beta^1_{h,t} X_i + \varepsilon_i$, $i = t - h, \ldots, t + h - 1$
6: **Output:** Coefficients $\hat{\beta}^1_{h,t}$ and their corresponding minus log p-value $-\log(p)$ produced under the Gaussian assumption

Compared to the traditionally applied cross-covariance function (CCF), this algorithm shares similar interpretations but simultaneously provides well-structured hypothesis testing for the significance of the relationship between sequences. Also, the linear regression model can be utilised for handling more than one regressors at the same time, allowing for more complex relationships. Meanwhile, the scale-space viewpoint can help us find out the existence of a lead-lag relationship from both "local" and "global" perspectives. To clarify, considering the colour of the coefficient heatmap

generally can help us get a basic understanding of the potential positive or negative relations between bi-variate time series; on the other hand, under various scales, we can see the different significance levels, and hence figure out the changing lead-lag relationships of the tested dataset and the (possibly) "best" window size (scale) for detecting the complete features in data.

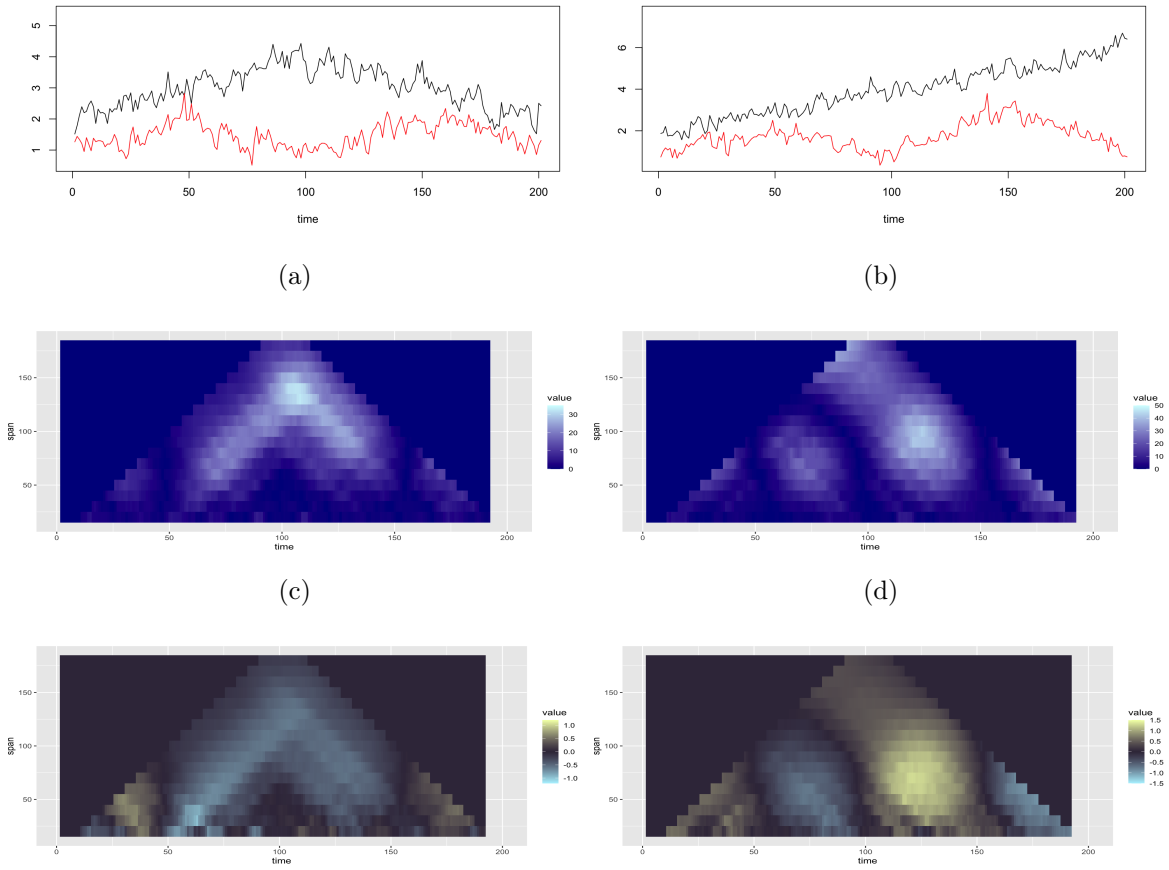


Figure 5.4: Combination of simulated time series built on piecewise-linear signal and serial correlated error process [(a) and (b)], heatmaps of $-\log(p)$ [(c) and (d)] and heatmaps of coefficients [(e) and (f)]. The black (red) lines in (a) and (b) are the simulated regressors (dependent variables).

In practical uses, our exploratory method aims to provide a concise overview of the

presence and potential locations of lead-lag relationship between bi-variate time series, and hence it remains applicable even when the data is dependent. Take Figure 5.4 as an example, based on the same signal, two examples generated by serial dependent processes are provided as control groups for the pair observations (X_i, Y_i) in Figure 5.3. After comparison, it is evident that except for the detailed strength of significance in $-\log(p)$ heatmaps, test results shown in Figure 5.4 share common patterns with those in Figure 5.3.

5.3 Simulation Study

In this section, we analyse the possible lead-lag relationship between trends and their changes for two time series shown by Algorithm 2. When looking at the lead-lag relationship between bi-variate time series, we tend to assume that the direction of the relation is known, i.e. $\{X_i\}_{i=t-h}^{t+h-1}$ leads $\{Y_i\}_{i=t-h}^{t+h-1}$, especially when applied for many real-world examples. For example, in COVID-19 dataset, it is natural to consider that changes in the number of new cases can impact that of new deaths, and hence we set cases as regressor X_i and deaths as dependent variable Y_i. Such a natural direction in turn encourages us to concentrate on examining the significance and the corresponding location of the (possible) lead-lag relationships.

Although our algorithm simply provides a family of regression coefficients and corresponding p-values indexed by the moving window size, it comes with one point that can be straightforward and convenient before considering the statistical literature. That is, when getting an overall picture of the bi-variate data, we avoid the choice of parameters such as the number of regression variables (or time lags), for simplicity, and simultaneously study a wide range of window sizes $2h$ due to the potential useful information available at different resolution levels of data.

To capture all information from the local to whole dataset, the sizes of the moving windows are chosen mainly relying on the sample itself. Specifically, we set the smallest window size $2h_{\min}$ based on the practical rule of thumb of linear regression that 10 to 20 observations are required for each regressor to guarantee reasonable power of estimation. The range of data can be applied as the largest window size $2h_{\max}$. However, it is heuristically "dangerous" to use the whole sample size as the largest window since the relationship can be evolving over time and hence applying this sort of a global tool may provide distorted information about the true dependence. To clarify, conducting a linear regression on the whole sample length is a bit close to the idea of taking the sample mean of a time series with a trend, which does not seem to be meaningful.

5.3.1 Analysis on sequences with similar patterns

In the following, we present different combinations of two sequences with the same patterns, i.e. for the two series $\{Y_i\}_{i=1}^n$ and $\{X_i\}_{i=1}^n$ following simple periodic models

$$\begin{aligned} Y_t &= A\sin(2\pi B(t - C^Y)) + D \\ X_t &= A\sin(2\pi B(t - C^X)) + D \end{aligned} \tag{5.3.1}$$

with the same frequency of oscillation $B > 0$, magnitude $A > 0$ and vertical shift $D \geqslant 0$, we consider the results over various differences in phase shifts C^Y and C^X, which satisfy $B(C^Y - C^X) \leqslant 1$ for some $C^Y, C^X \geqslant 0$, to illustrate the performance of heatmaps on discovering lead-lag relationships. In this context, the differences of phase shifts used in our test essentially corresponds to the time lags in traditional correlation functions. To be specific, we consider the following simulated sequences (M1)-(M7) representing the regressor X_t and dependent variable Y_t with differences $C^Y - C^X = 25, 50, 75, 100, 125, 150, 175$ respectively. Results together with the corresponding signals of another group of sequences with higher frequencies are given in

Section 5.5. Since we want to clearly display the patterns, in this section, we shall focus on the scenarios without noises. Also, given that the provided examples (M1)-(M7) are constructed without frequent fluctuations, we choose to set $h_{\min} = 25$ and $h_{\max} = 200$. For sequences (M12)-(M18), the scale set is pre-defined with $h_{\min} = 25$ and $h_{\max} = 300$ due to higher frequency but the same sample length.

(M1) $X_t = \sin(\pi t/200 - \pi/2) + 1$ for $t = 1, 2, \ldots, 825$ and $Y_t = \sin(\pi t/200 - 5\pi/8) + 1$ for $t = 26, 27, \ldots, 825$.

(M2) $X_t = \sin(\pi t/200 - \pi/2) + 1$ for $t = 1, 2, \ldots, 850$ and $Y_t = \sin(\pi t/200 - 3\pi/4) + 1$ for $t = 51, 52, \ldots, 850$.

(M3) $X_t = \sin(\pi t/200 - \pi/2) + 1$ for $t = 1, 2, \ldots, 875$ and $Y_t = \sin(\pi t/200 - 7\pi/8) + 1$ for $t = 76, 77, \ldots, 875$.

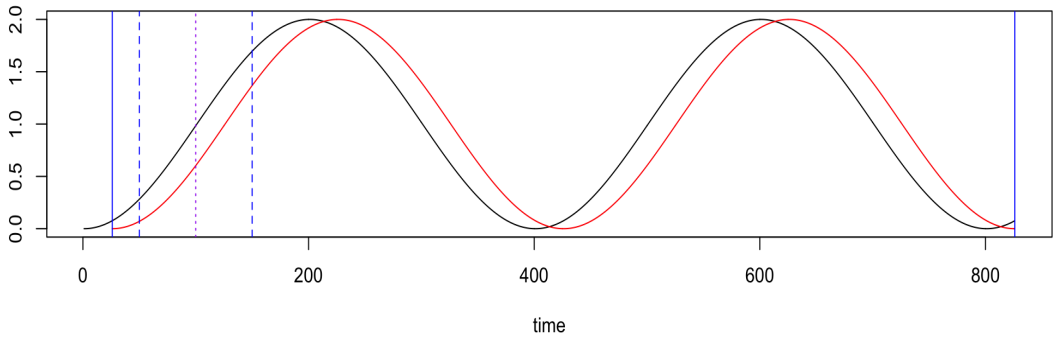
(M4) $X_t = \sin(\pi t/200 - \pi/2) + 1$ for $t = 1, 2, \ldots, 900$ and $Y_t = \sin(\pi t/200 - \pi) + 1$ for $t = 101, 52, \ldots, 900$.

(M5) $X_t = \sin(\pi t/200 - \pi/2) + 1$ for $t = 1, 2, \ldots, 925$ and $Y_t = \sin(\pi t/200 + 7\pi/8) + 1$ for $t = 126, 127, \ldots, 925$.

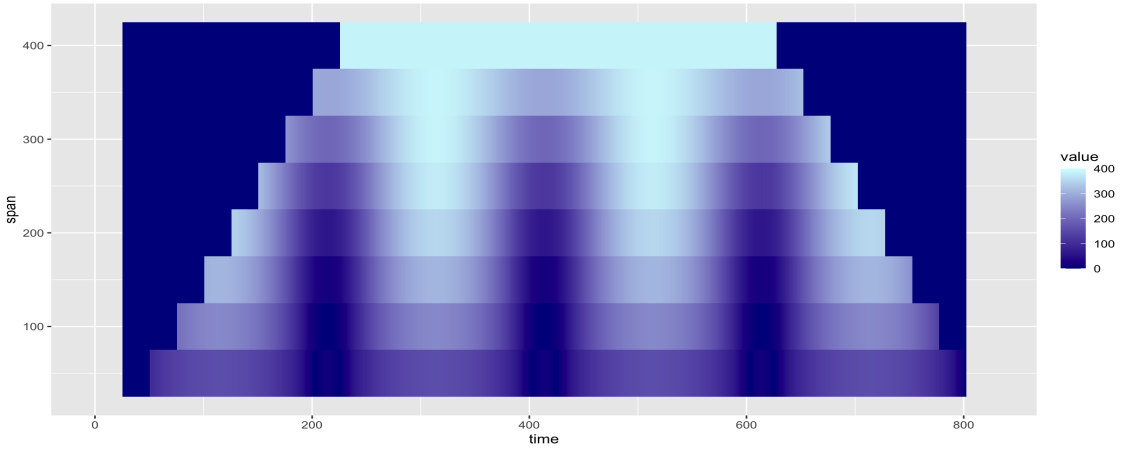
(M6) $X_t = \sin(\pi t/200 - \pi/2) + 1$ for $t = 1, 2, \ldots, 950$ and $Y_t = \sin(\pi t/200 + 3\pi/4) + 1$ at $t = 151, 152, \ldots, 950$

(M7) $X_t = \sin(\pi t/200 - \pi/2) + 1$ for $t = 1, 2, \ldots, 975$ and $Y_t = \sin(\pi t/200 + 5\pi/8) + 1$ for $t = 176, 177, \ldots, 975$.

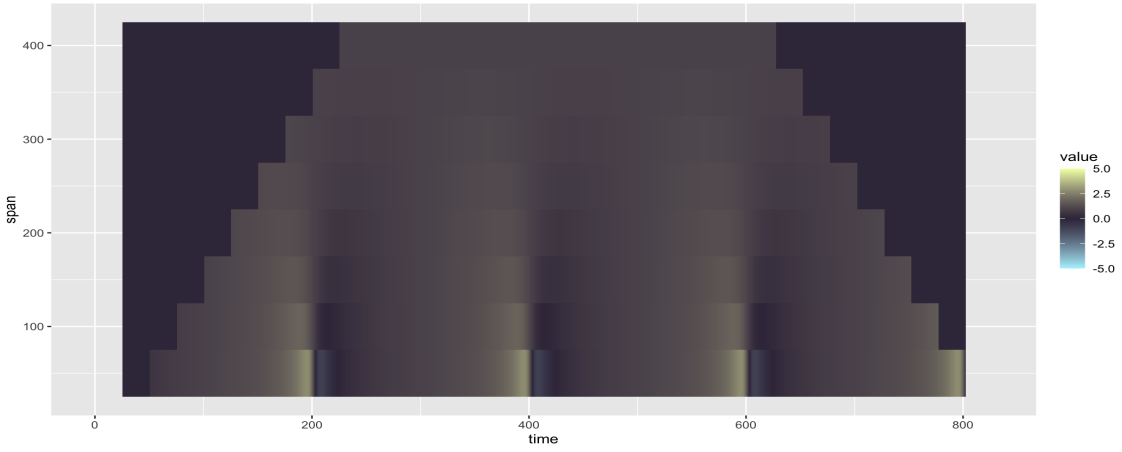
Before further discussion, we first provide a more detailed description of Figure 5.5. In sub-figure (a), the examined area is marked with blue vertical lines. The black (red) curve in (a) is the simulated regressor (dependent variable). We estimate the coefficients over a symmetric moving window around each location, which, for example, are shown with blue dashed lines and a purple dotted line respectively. Sub-figures (b) and (c) display the estimation results at the corresponding locations and sizes of moving window.



(a) Original data



(b) $-\log(p)$



(c) coefficients

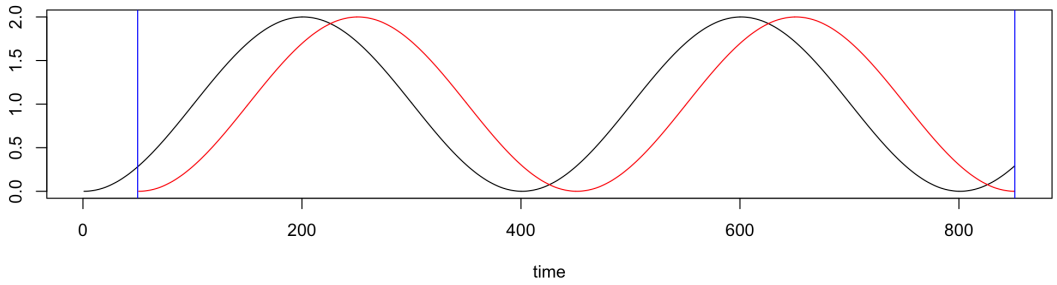
Figure 5.5: Combination of simulated signal (a) and heatmaps [(b) and (c)] for dataset (M1). The blue vertical lines bound the examined area. As an example, the blue dashed line and purple dotted lines stand for the location and the corresponding boundaries of the symmetric moving window.

For illustrative purposes, we display the figures 5.5 to 5.11 for all examples produced by models (M1) to (M7) respectively, and the same colour scale is applied to allow for the comparison between heatmaps. Overall, the difference of phase shifts $C^Y - C^X$, which can also be regarded as possible time lags between b-variate time series, is reported in the time length of the shifted colour in sub-figures (b), such as the colour blue in Figure 5.5 to 5.11. Meanwhile, these heatmaps provide reasonable evidence supporting the idea that for periodic bi-variate dataset with period 400, $2h = 200$ can be a suitable window size to discover the significance of dependence between time series. Then, we will try to identify the fundamental patterns discernible within each graph.

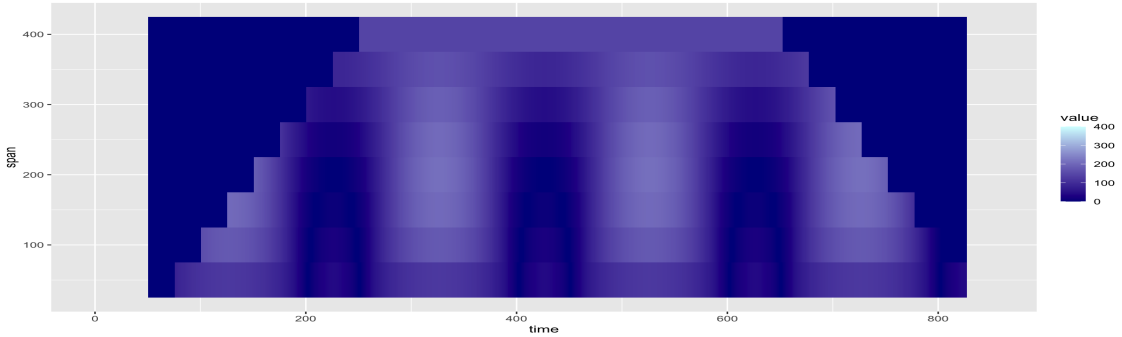
Starting from Figure 5.5, we can see the two time series X_t (black) and Y_t (red) regarded with lag 25 share very close movements and the sub-figure (c) indicates that such associations are overall positive except for the conditions where small moving windows covering the time intervals with comparatively different patterns, i.e. at $t \in [200, 225] \cup [440, 425] \cup [600, 625]$, within the smallest window size 50, X_t and Y_t will experience either a peak or a valley but with some time delays. Since the patterns continue changing within the small window, the detected relationship is not significant, as reflected by the small $-\log(p)$ in sub-figure (b). However, as the window size (scale) grows, the corresponding colour of $-\log(p)$ at the aforementioned time points becomes lighter and lighter owing to the simple reason that a larger window size can capture a more obvious co-movement in Model (M1). Overall, due to the small difference $C^Y - C^X$, the largest scale 400 use a sample length with longer co-movements and hence report the strongest association between two time series.

Secondly, with a small increase in the difference of phase shifts $C^Y - C^X$, Figure 5.6(c) shows that compared to Model (M1), the strength of dependence between two time series produced by Model (M2) is partly reduced (the colour blue looks darker) because of the decreased sample length of relatively strict co-movements, by comparing Figure 5.6(b) with Figure 5.5(b). On the other hand, the negative associations at small scales

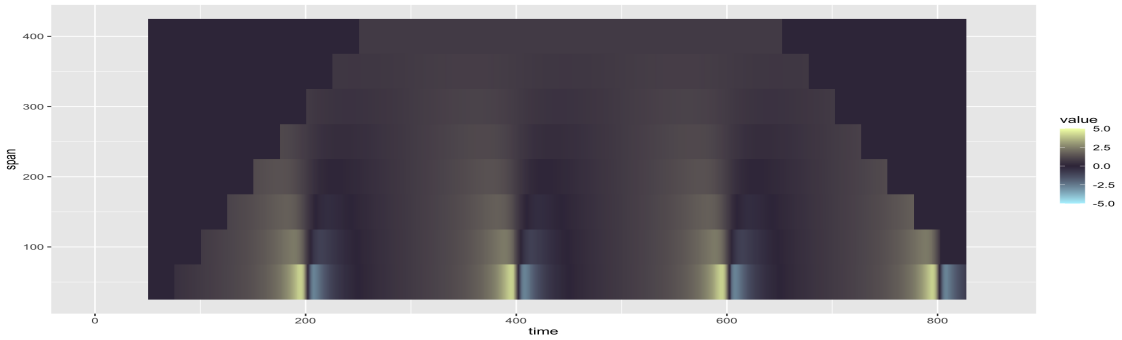
are detected over a larger range while the coefficients located at the middle of this range become a bit more significant (but not so obvious). Similarly, in Figure 5.7(b) and 5.8(b), we can see a further reduction in the strength of detected dependence at large scales and a slight increase at small scales.



(a) Original data

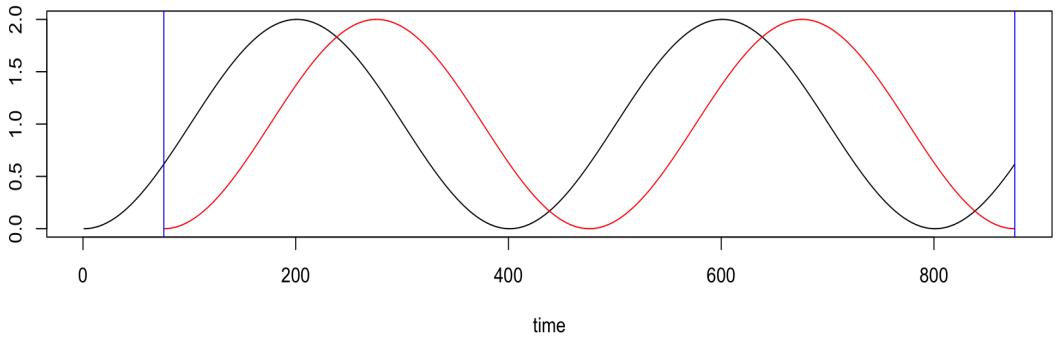


(b) $-\log(p)$

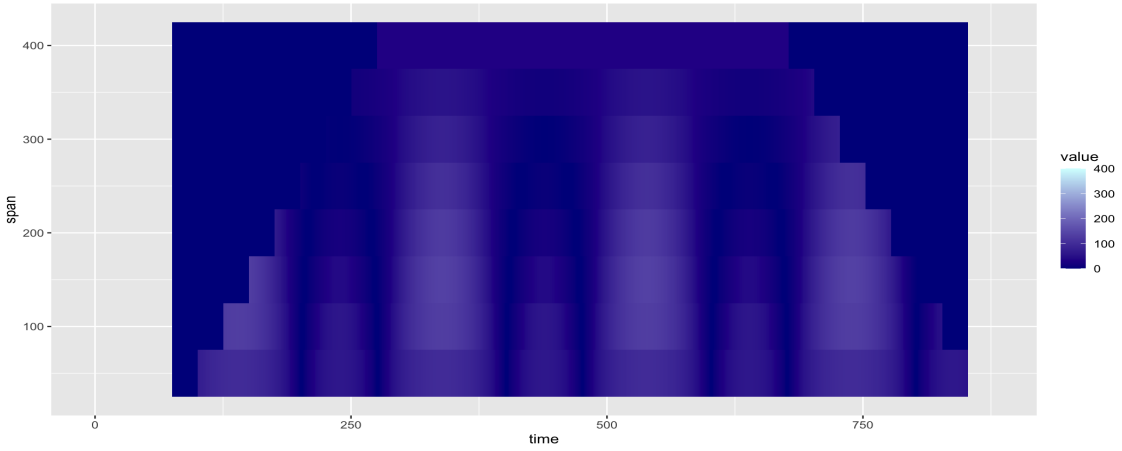


(c) coefficients

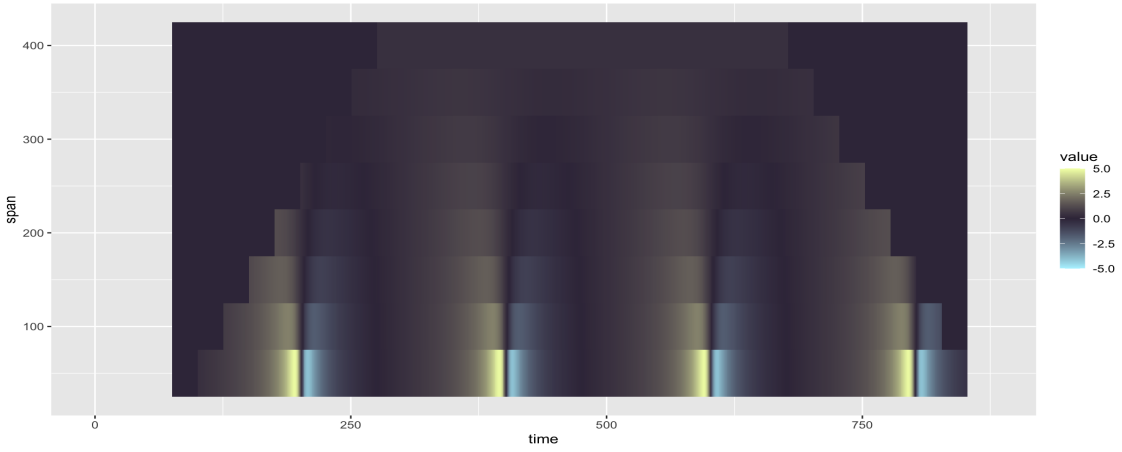
Figure 5.6: Combination of simulated signal (a) and heatmaps [(b) and (c)] for dataset (M2).



(a) Original data

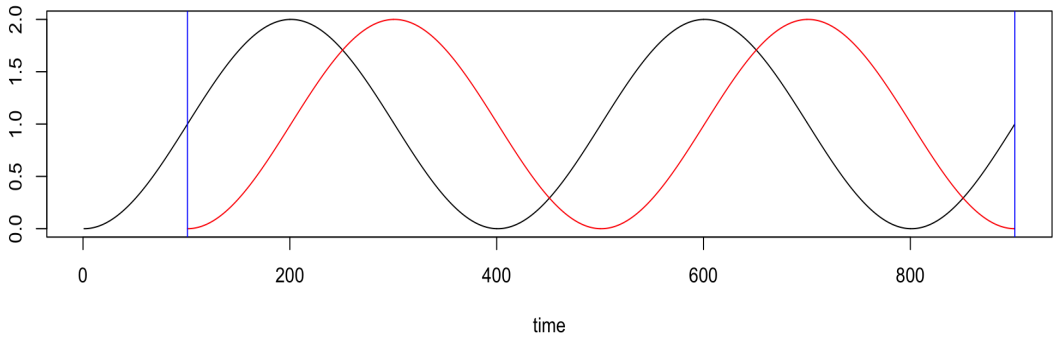


(b) $-\log(p)$

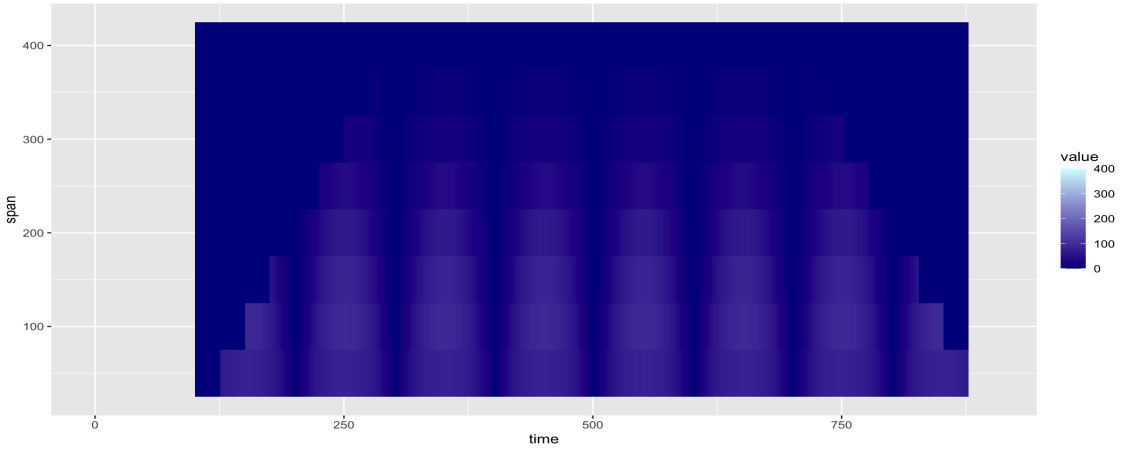


(c) coefficients

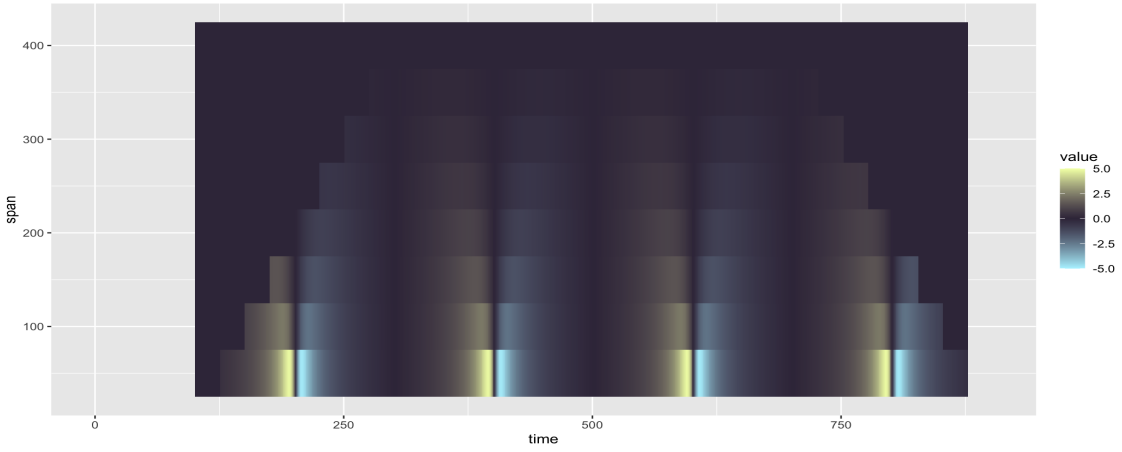
Figure 5.7: Combination of simulated data (a) and heatmaps [(b) and (c)] for dataset (M3).



(a) Original data



(b) $-\log(p)$

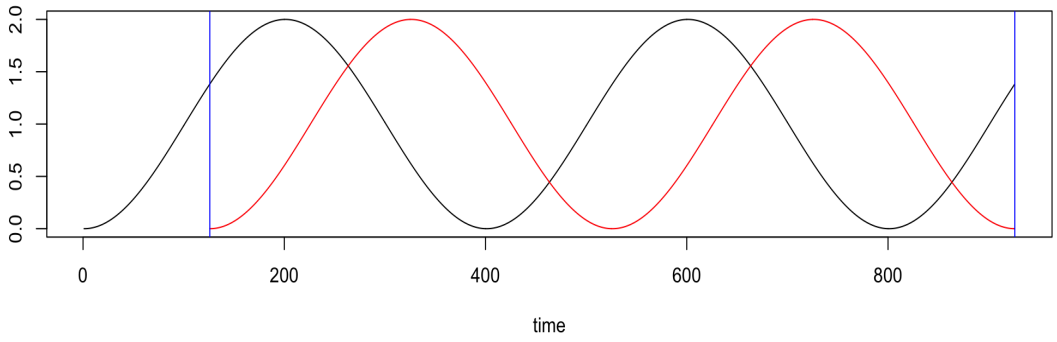


(c) coefficients

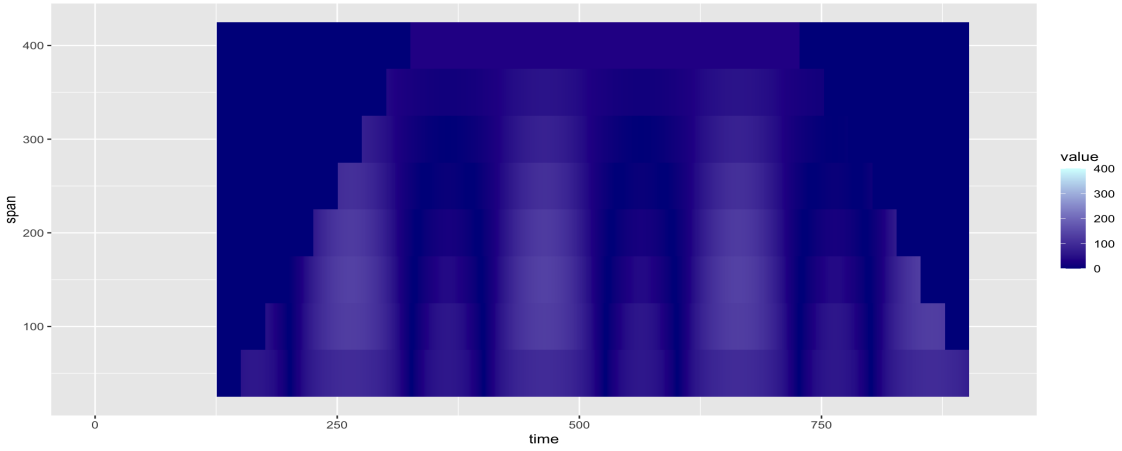
Figure 5.8: Combination of simulated data (a) and heatmaps [(b) and (c)] for dataset (M4).

Next, in graphs starting from Figure 5.9, the simulated bi-variate dataset experience a growing co-movement in the opposite direction and we can see the colour blue gradually dominates the heatmap of coefficients. Also, the more and more significant dependence detected at large scales are highlighted by lighter and lighter blue in sub-figures (b). In particular, following Model (M7), where the preset difference $C^Y - C^X$ is 175, Figure 5.11 shows that the co-movement in the opposite direction (and the corresponding strength of dependence) becomes even as significant as the co-movement displayed in the bi-variate time series introduced in Model (M1). We can also see similar features in sub-figures (b) for (M2) and (M6), and those for (M3) and (M5).

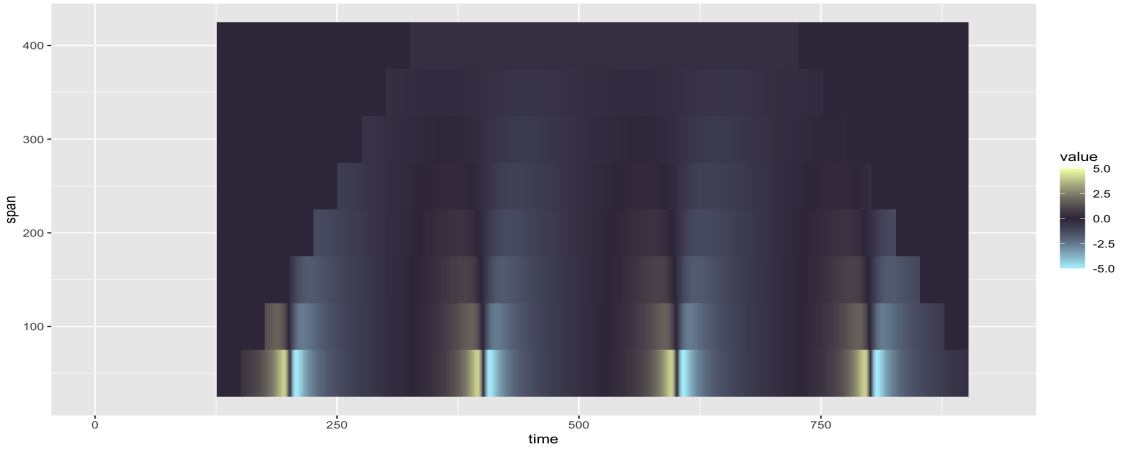
Moreover, we conduct another similar simulation study on another group of sequences with higher frequencies, see models and corresponding figures 5.35 to 5.41 in Section 5.5. Here, we choose the colour scale $[0, 400]$ for this series of $-\log(p)$ heatmaps to make comparison easier. Among all these figures, one special case is the "wrong" graph, Figure 5.38, which shows the test results of Model (M15). As seen in its sub-figure (a), the pair of observations experiences the strictly opposite patterns, and hence the coefficient is always equal to 1 and the standard error of our test statistic that quantifies the expected variability of estimated coefficient should be equivalent to 0, leading to the grey area in sub-figure (b) (all NA's). However, since real-word data usually contain noise, such a "wrong" feature is highly unlikely to occur in practice.



(a) Original data

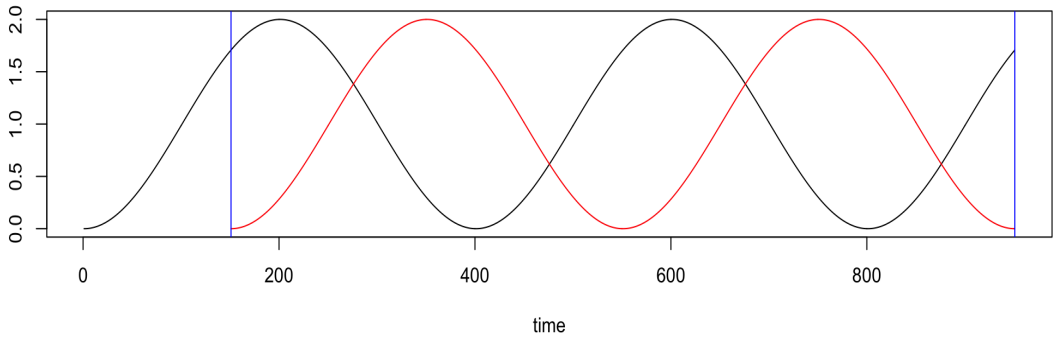


(b) $-\log(p)$

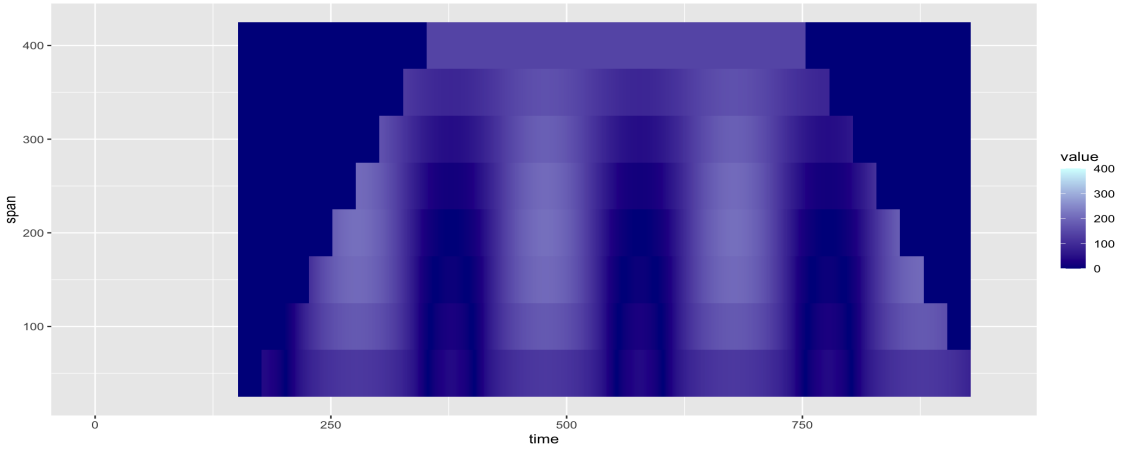


(c) coefficients

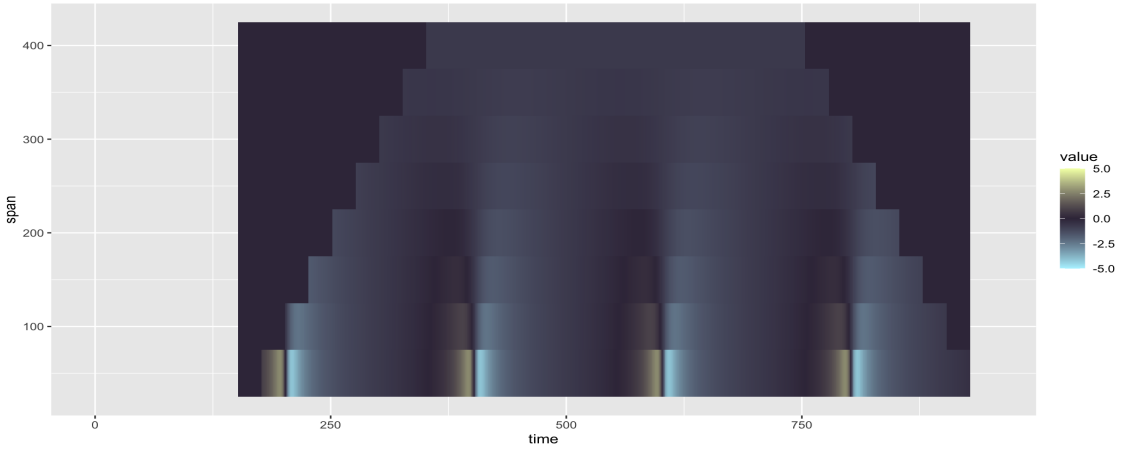
Figure 5.9: Combination of simulated data (a) and heatmaps [(b) and (c)] for dataset (M5).



(a) Original data

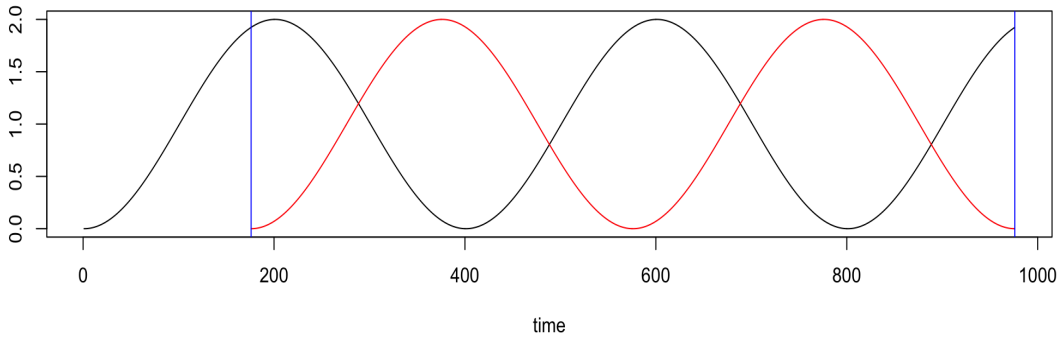


(b) $-\log(p)$

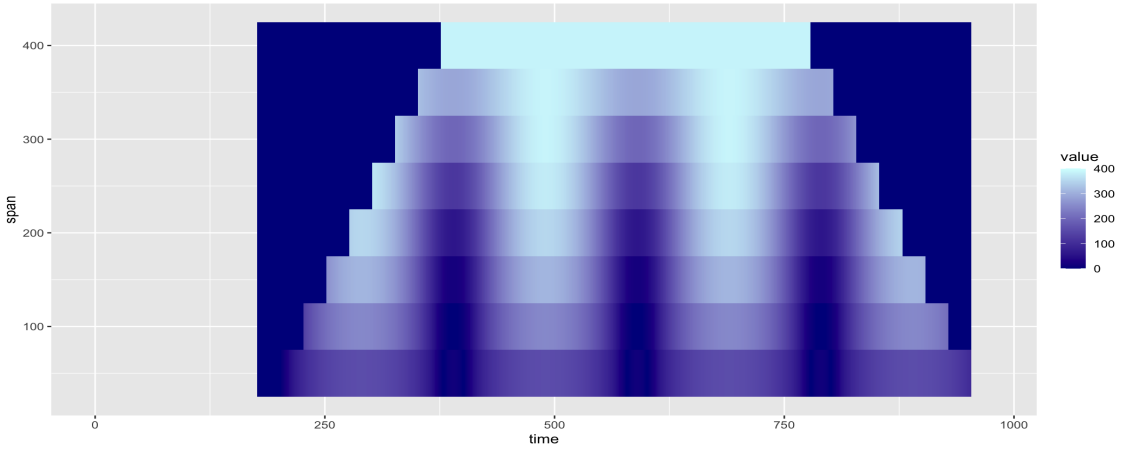


(c) coefficients

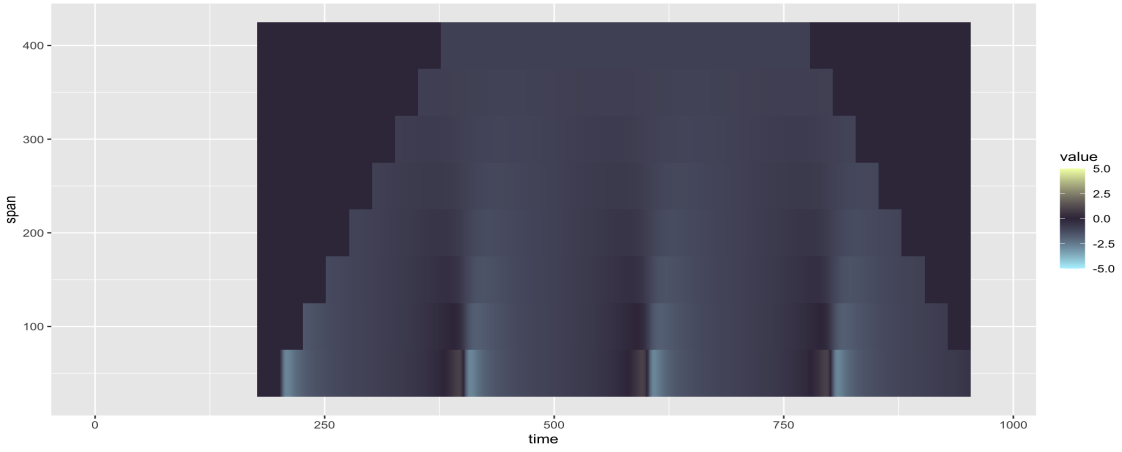
Figure 5.10: Combination of simulated data (a) and heatmaps [(b) and (c)] for dataset (M6).



(a) Original data



(b) $-\log(p)$



(c) coefficients

Figure 5.11: Combination of simulated data (a) and heatmaps [(b) and (c)] for dataset (M7).

In moving on to the remaining graphs, we shall first discuss Figure 5.35. Although models (M2) and (M12) are built on different frequencies, their corresponding difference in phase shifts $C^Y - C^X$ make the two models share similar patterns, and consequently we can see information reported by heatmaps over scales $2h = 50, 100, 150, 200$ for data from (M12) within $[26, 425]$ is even the same as that provided over scales $2h = 100, 200, 300, 400$ for series from (M2) within $[51, 850]$. Besides the similar parts, three more points are highlighted in Figure 5.35. First, the evidence against the null hypothesis is generally stronger in scales larger than 200. Second, different scales are likely to report different patterns of significance at the same location. For example, the scale 250 and 300 show a lighter colour in sub-figure (b) at different locations. Intuitively speaking, such a difference arises from the synchronised trends, i.e. co-movement in either the same or opposite directions, that can be covered within the moving window at different time points. Lastly, when window size grows to 600, the statistical significance obtained across time locations looks roughly the same, see at scales 200 and 400 as well. In this case, the period of our sine formula can be reflected in the value of the scales. We can also find this result from scale 400 in figures 5.5, 5.6 and 5.7, etc. Almost the same heatmaps are shown in figures 5.35, 5.37, 5.39 and 5.41, where the small difference is the shifted colour and its corresponding time length shown in sub-figures (c). This indicates that they exhibit a comparable degree of co-movements (in the same or opposite directions) over all window sizes.

In Figure 5.36 and Figure 5.40, due to the limited degree of co-movements, the evidence supporting the dependence between bi-variate time series is not so strong compared to the other scenarios but we can still find some characteristics at relatively smaller scales. On the whole, figures 5.35 to 5.41 reveal analogous features observed in figures 5.5 to 5.11, but $2h = 100$, rather than $2h = 200$, is a more suitable window size for the periodic bi-variate dataset with period 200.

In summary, from Figure 5.5 to Figure 5.11 for the first pair of observations and Figure

5.35 to Figure 5.41 for the second pair of observations, we make the following statements about the lead-lag relationship between Y_t and X_t:

1. Overall, the colour yellow or blue that represents the sign of coefficients at each location over all scales can provide an overview of the positive or negative correlation between features in two time series. The possible time lags between bi-variate time series is suggested by the time length of the shifted colour in sub-figure (b). And the width of the "bars" in sub-figure (b) can somehow indicate whether the relationship is long-run or short-run.

2. At location t and half scale h with significant p-value and non-zero coefficients, we can see the strong dependence between $\{X_i\}_{i=t-h}^{t+h-1}$ and $\{Y_i\}_{i=t-h}^{t+h-1}$, and in particular the scale indicates that more complete patterns can be found under this window size while the location points out where we can discover comparatively more obvious lead-lag relations.
   For example, Model (M1) in Figure 5.5(a) is built with $\{X_i\}_{i=t-h}^{t+h-1}$ and $\{Y_i\}_{i=t-h}^{t+h-1}$ sharing the same pattern with only a small lag of 25. At a time point around 200, due to the change in trends, the heatmap does not show us a significant relationship (co-movement in the same direction) between two sequences for smaller scales, but when the window size continues to increase, more similar patterns (concave downwards) are shown in the examined interval and the coefficient becomes more significant. Also, nearly opposite patterns can report similar results to us, see Figure 5.11. On the other hand, Figure 5.7 to Figure 5.9 display stronger relationships under smaller window sizes because the aforementioned patterns are more obvious in narrower intervals.

In the next section, we assess the performance of our method under relatively more generalised scenarios with the relationship changing over time, presenting both experiment results and the corresponding heatmaps.

5.3.2 Analysis on sequences with different patterns

While sinusoidal functions can capture a range of behaviours, the modelling of real-world applications may also require considering situations where the amplitude A or the frequency of oscillation B in (5.3.1) does not stay constant. For example, in physics and calculus textbooks, it is common to explore models following a fixed percentage reduction in amplitude per second. Therefore, we present additional examples illustrating the conclusions drawn in the last section and possibly more meaningful information on more generalised conditions. Model (M8) and (M9) describe the scenarios with changing amplitude and changing frequency of oscillation respectively. Moreover, we consider examples (M10) and (M11) with piecewise-linear signal and independent error process to show the performance of MLLH on data with abrupt changes in slopes.

(M8) $X_t = \sin(\pi t/100 - \pi/2) + 1$ for $t = 1, 2, \ldots, 600$ and $Y_t = \sin(\pi t/100 + 9\pi/10) + 1$ for $t = 61, 62, \ldots, 260$, $Y_t = \sin(\pi t/90 - 7\pi/18) + 1$ for $t = 261, 262, \ldots, 440$, with $Y_t = \sin(\pi t/80 - \pi) + 1$ for $t = 441, 442, \ldots, 600$.

(M9) $X_t = \sin(\pi t/100 - \pi/2) + 1$ for $t = 1, 2, \ldots, 600$ and $Y_t = \sin(\pi t/100 - 9\pi/10) + 1$ for $t = 41, 42, \ldots, 240$, $Y_t = (\sin(\pi t/100 - 9\pi/10) + 1)/2$ for $t = 241, 242, \ldots, 440$, with $Y_t = (\sin(\pi t/80 - 9\pi/10) + 1)/4$ for $t = 441, 442, \ldots, 600$.

(M10) $X_t = f_t^X + \epsilon_t^X$ where f_t^X undergoes 10 change-points at $100, 200, 300, 400, 500, 650, 800, 900, 1000, 1100$ for $t = 1, \ldots, 1200$ and the corresponding slopes $-1/50, 1/50, -1/50, 1/50, 1/150, -1/50, 1/50, -1/50, 1/50, -1/50$, starting intercept $f_1^X = 2$ and slope $1/50$, and $\epsilon_t^X \overset{iid}{\sim} N(0, 0.5^2)$; $Y_t = f_t^Y + \epsilon_t^Y$ where f_t^Y undergoes 7 change-points at $126, 226, 326, 426, 626, 776, 1026$ for $t = 26, \ldots, 1200$ and the corresponding slopes $-1/50, 1/100, -1/100, 1/100, -1/150, -1/250, 1/500$, starting intercept $f_1^Y = 1$ and slope $1/50$, and $\epsilon_t^Y \overset{iid}{\sim} N(0, 0.5^2)$.

(M11) $X_t = f_t^X + \epsilon_t^X$ where f_t^X undergoes 8 change-points at $t = 100, 200, 300, 400, 600,$

$800, 900, 1000$ for $t = 1, \ldots, 1200$ and the corresponding slopes $-1/50, 1/25, -1/25, 1/100, -1/100, 1/200, -1/200, 1/200$, starting intercept $f_1^X = 3$ and slope $1/50$, and $\epsilon_t^X \overset{iid}{\sim} N(0, 0.5^2)$; $Y_t = f_t^Y + \epsilon_t^Y$ where f_t^Y undergoes 11 change-points at $111, 211, 311, 411, 511, 611, 711, 811, 911, 1011, 1111$ for $t = 11, \ldots, 1200$ and the corresponding slopes $-1/50, 1/50, -1/50, 1/50, -1/50, 1/50, -1/50, 1/50, -1/50, 1/50, -1/50$, starting intercept $f_1^Y = 1$ and slope $1/50$, and $\epsilon_t^Y \overset{iid}{\sim} N(0, 0.5^2)$.

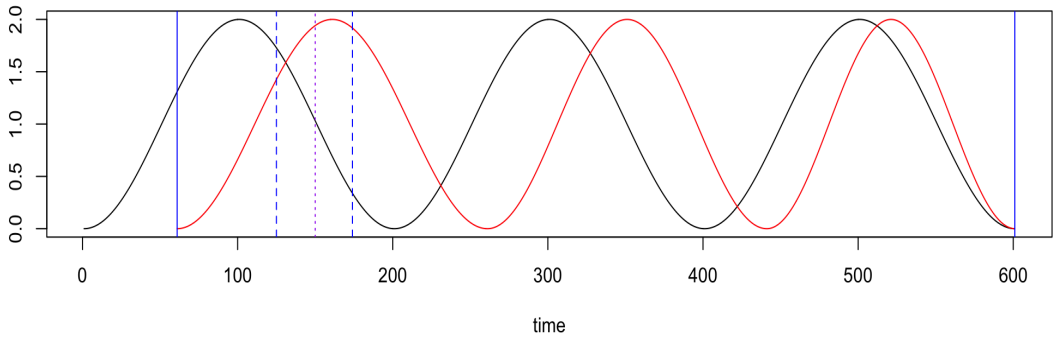
Model (M8) represents a sinusoidal function with increasing frequency and Model (M9) allows decreasing amplitude. To be specific, the examined area of Model (M8) can generally be divided into three parts, $[60, 260] \cup (260, 440] \cup (440, 600]$, and the bi-variate time series experiences a decreasing time lag within the last two intervals. On the other hand, the time delay included in Model (M9) remains unchanged. Models (M10) and (M11) present time series with changing trends and preset time lags of 25 and 10 respectively.

Figure 5.12 displays the bi-variate dataset produced by Model (M8) and analyses it with MLLH. Overall, the varying frequency (different patterns of co-movement) makes the detected coefficients less significant at large scales. The sub-figure (c) shows that there is a gradual shift in the colour at larger scales, transitioning from blue to yellow, implying an overall negative to positive dependence; Also, the "bars" in sub-figure (b) exhibit a greater width at the two sides and narrower width in the middle. Both of the two features align with the fact that the pair of observations starts with the co-movement in the opposite direction, which gradually diminishes over time, and later exhibits co-movement in the same direction. Second, we can see in (c) that the temporal range of the shifted colour becomes narrower across locations, which corresponds to the decreasing time lag of the original dataset. Moreover, Figure 5.12(b) shows lighter colour after $t = 440$ and hence provides stronger evidence in favour of a significant dependence between the bi-variate data. This also matches the enhanced co-movement shown in the latter part of sub-figure (a). For this example, we have an even higher

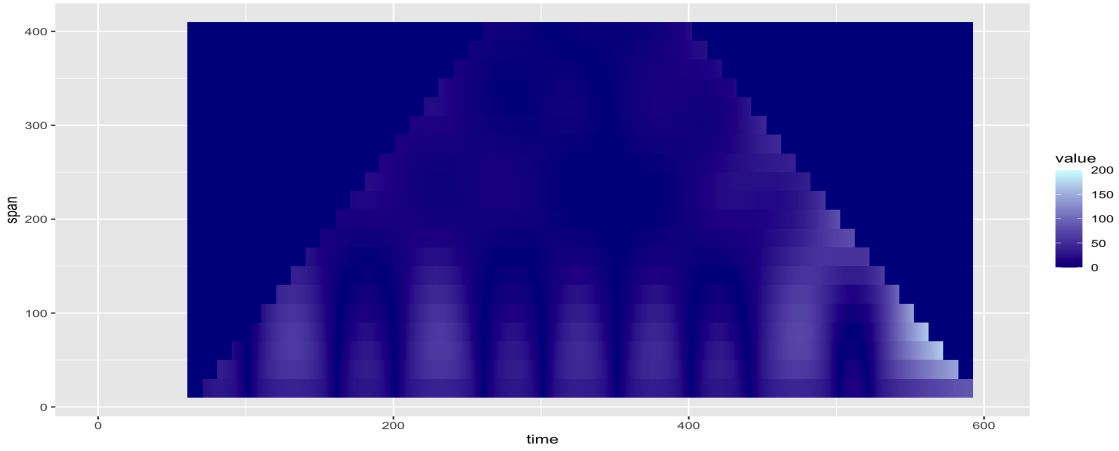
frequency compared to models (M12)-(M18), and hence the relatively best scale should be smaller, i.e. 80 or 60.

Next we study the performance of MLLH in the settings of Model (M9). In Figure 5.13, the colour of the heatmap (c) is less apparent due to the shrinking magnitude of the sinusoidal function applied in the dataset. Secondly, the "bars" in sub-figure (b) at scales under 200 generally show regular patterns, which aligns with the unchanged frequency in the original data. Also, the presence of both wider and narrower "bars" suggests that the time lag between the bi-variate time series cannot be half of the temporal range covering a complete pattern, which is the period of the sinusoidal function in this example. Thirdly, in Figure 5.13(c), the unchanged temporal range of the shifted colour corresponds to the time-invariant time lag of the original dataset while the diminishing magnitude of coefficients (darker colour) reflects the decreasing magnitude of data. In this example, we have the same frequency compared to models (M12)-(M18), and it shows that 100 is still a suitable scale while the smaller scales such as 80 or 60 could also provide strong evidence of dependence.

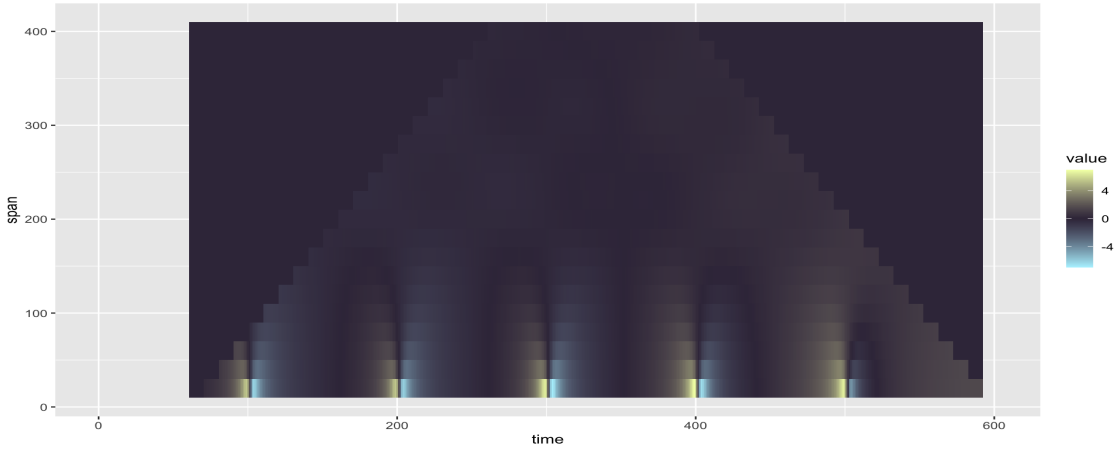
Overall, Figure 5.12 and Figure 5.13 demonstrate the effectiveness of MLLH on showing the significance of dependence (strength of lead-lag relationship) of the bi-variate datatset produced by the scenarios with a changing relationship over time. Besides, for data with underlying periodic functions, MLLH can provide information about possible periodic feature in data through regularity in heatmaps, together with potential changes in magnitude or frequency via different features across scale or space. To further illustrate this idea, we shall consider periodic data (M10) and (M11) with abrupt changes instead of the smooth changes in models (M8) and (M9).



(a) Original data

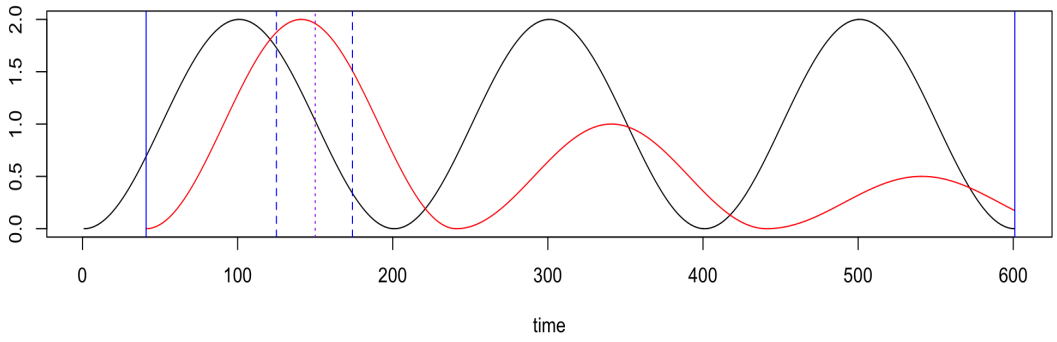


(b) $-\log(p)$

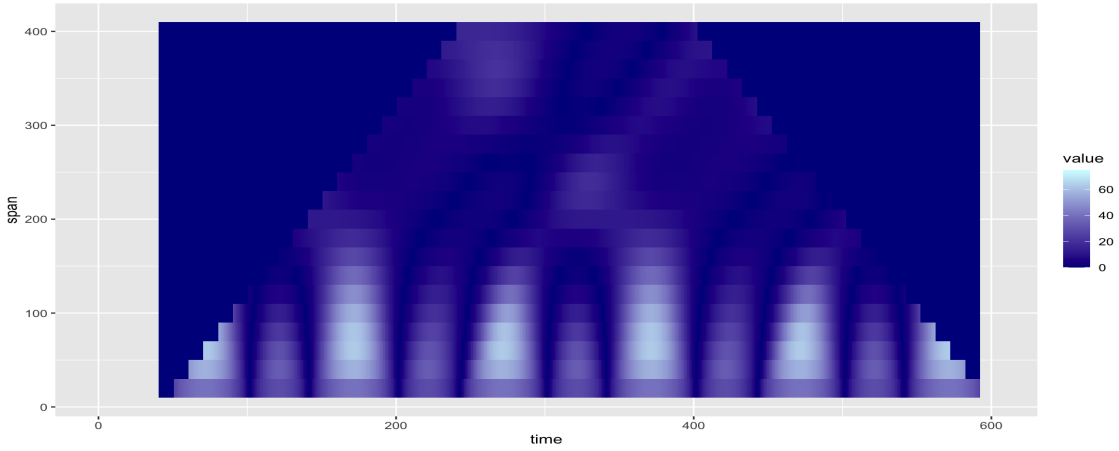


(c) coefficients

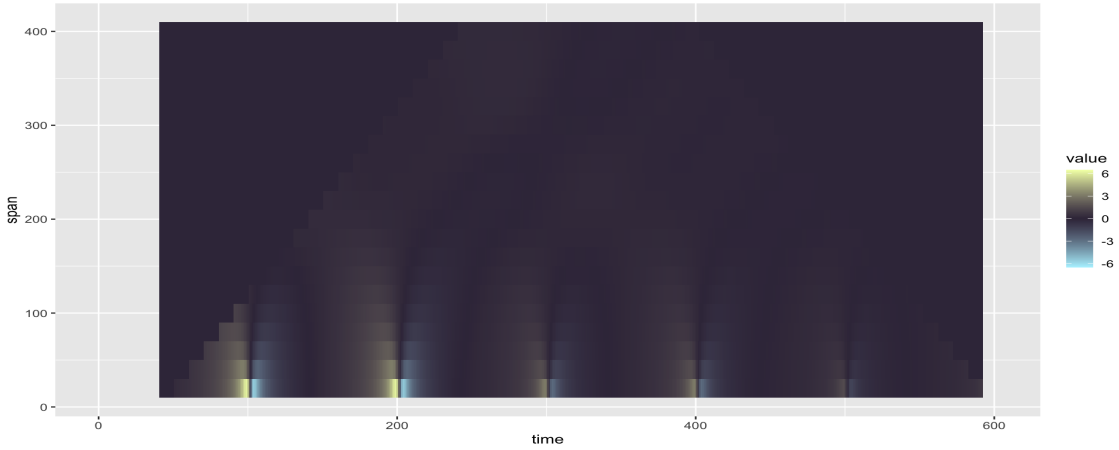
Figure 5.12: Combination of simulated signal (a) and heatmaps [(b) and (c)] for dataset (M8). The blue vertical lines bound the examined area. As an example, the blue dashed line and purple dotted lines stand for the location and the corresponding boundaries of the symmetric moving window.



(a) Original data



(b) $-\log(p)$



(c) coefficients

Figure 5.13: Combination of simulated signal (a) and heatmaps [(b) and (c)] for dataset (M9). The blue vertical lines bound the examined area. As an example, the blue dashed line and purple dotted lines stand for the location and the corresponding boundaries of the symmetric moving window.

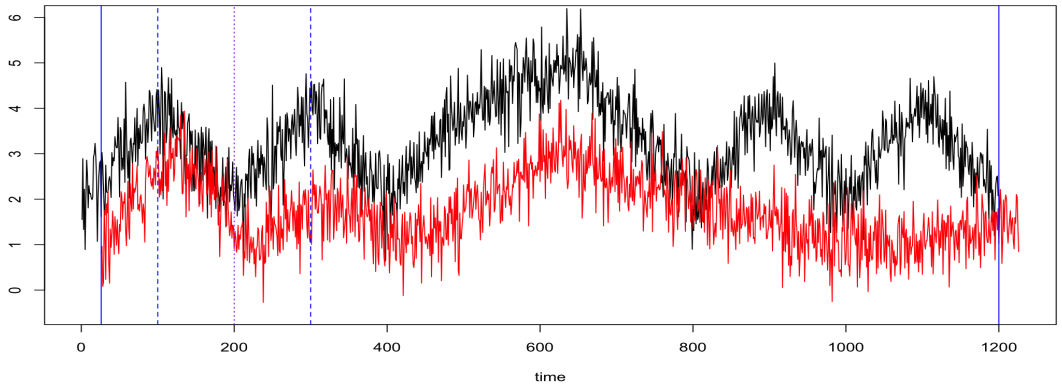
Figure 5.14 and Figure 5.15 are plotted to show the results of MLLH on bi-variate data with piecewise-linear signal and independent noise. Compared to sinusoidal functions, we can find another typical feature: the (45°) stripes in heatmaps, especially sub-figures (b), which help indicate the coincidence of the trends around the locations of the spikes. In particular, the brightest 45° thin region in Figure 5.14(b) is largely due to the significant co-movement of data around the spike at approximately 200.

Specifically, Figure 5.14 shows generally positive correlation between the bi-variate time series within areas $[35, 800]$ and $[900, 1000]$ and negative dependence within the remaining areas. Also, due to the presence of noises, the dependence tends to be less significant at the finest scales where the most significant feature locates within the areas around $t = 525$. This indicates the difficulty in finding a consistent pattern at small scales. In addition, the results in the sub-figure (b) can be divided into areas of nested "triangles" (or 45° stripes) via the pre-specified colour scheme, where different areas can somehow show dominant patterns and the (possible) shifts of the patterns in different regions of data. For example, the brightest parallelogram area whose sides are bounded by triangles with their base approximately at $[25, 625]$ and $[425, 825]$ present the strongest significance of dependence of time series. It indicates that we can find the dominant pattern within the region around $[425, 625]$, i.e. we are highly likely to discover the significant relationship when testing over regions containing this interval, and there could be apparent pattern shifts around $t = 425, 625$. This corresponds to the features we can observe in the original data. Similarly, the second dominant pattern of data lies somewhere within $[675, 825]$ and the pattern changes at $t = 825$. In this example, it shows that 200 could be a suitable scale that is able to provide most evident information of dependence.

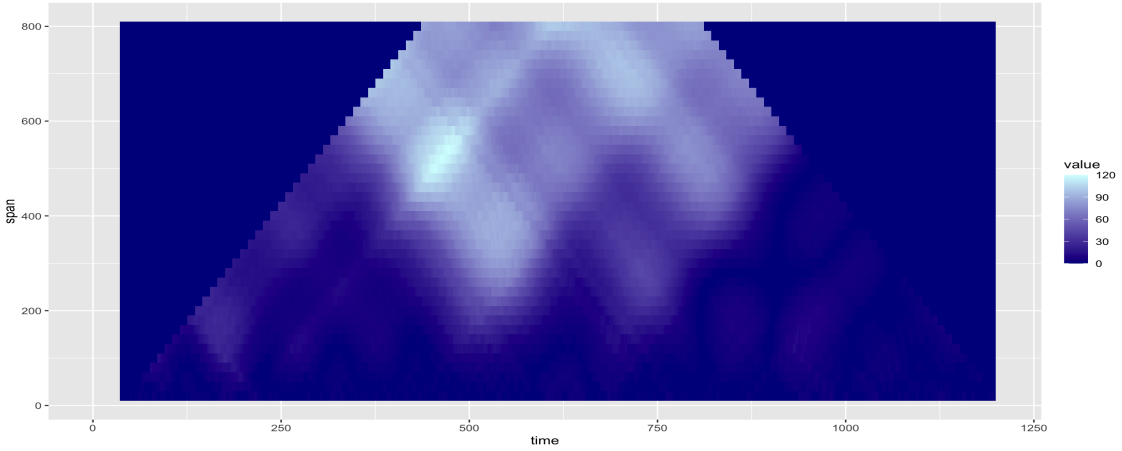
Analogously, Figure 5.15(c) presents the overall positive dependence of data and an particular change within the area $[550, 700]$. Also, Figure 5.15(b) shows that the most dominant significance locates somewhere around $[220, 320]$ and we can see consistent

correlation of the bi-variate time series over periods containing (part of) $[20, 320]$ in the long run. In addition, besides the obvious short-run shift in patterns after $t = 550$, there exists a particular long-run change in patterns after $t = 320$. On the other hand, although there are still some less significant short-run patterns after $t = 320$, we can hardly find apparent long-run patterns for the bi-variate time series. It partly contradicts with the features of the underlying signals, which should possess strongly significant long-run coincidence within region $[700, 1100]$. This indicates the large impact of noises on signal with relatively lower trends, i.e. in Model (M11), X_t has slopes $1/50$ and $-1/50$ in the beginning while the slopes within $[700, 1100]$ are merely $1/200$ and $-1/200$. Compared to Model (M10), the sub-figures (b) and (c) in Figure 5.15 indicate that 200 is also a suitable scale for Model (M11) while the smaller scales such as 180 or 160 could also provide the major patterns of dependence in the bi-variate time series.

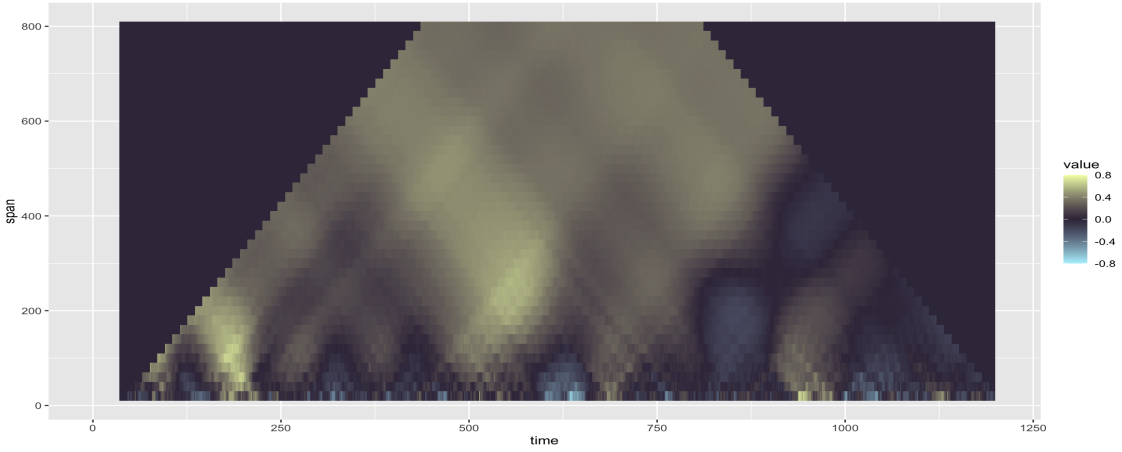
Overall, Figure 5.12 to Figure 5.15 demonstrate the effectiveness of MLLH as a tool for detecting the significance of changing lead-lag relationships between bi-variate time series with known direction. Meanwhile, MLLH provides information about dominant patterns of coincidence within the data and indicates the locations where (possible) pattern shifts exist, which can be useful for further data analysis. In addition, similar to the statements made on heatmaps for constant lead-lag relationships, the possible time lags between the tested bi-variate time series are suggested by the time length of the shifted colour in sub-figure (b). Considering the (possibly) "best" scale containing the majority of significant features in the data, the simulated results for time series with changing relations also imply the idea that larger scales should be more suitable for dataset with lower frequencies. Specifically, the highest frequency in (M8) and (M9) is approximately twice of that in (M10) and (M11) while the acceptable scale for (M8) and (M9) can be somehow half of that for (M10) and (M11).



(a) Original data

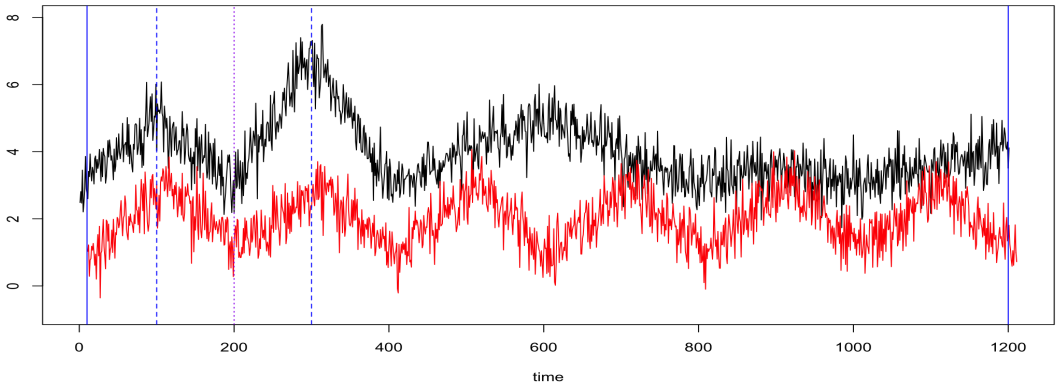


(b) $-\log(p)$

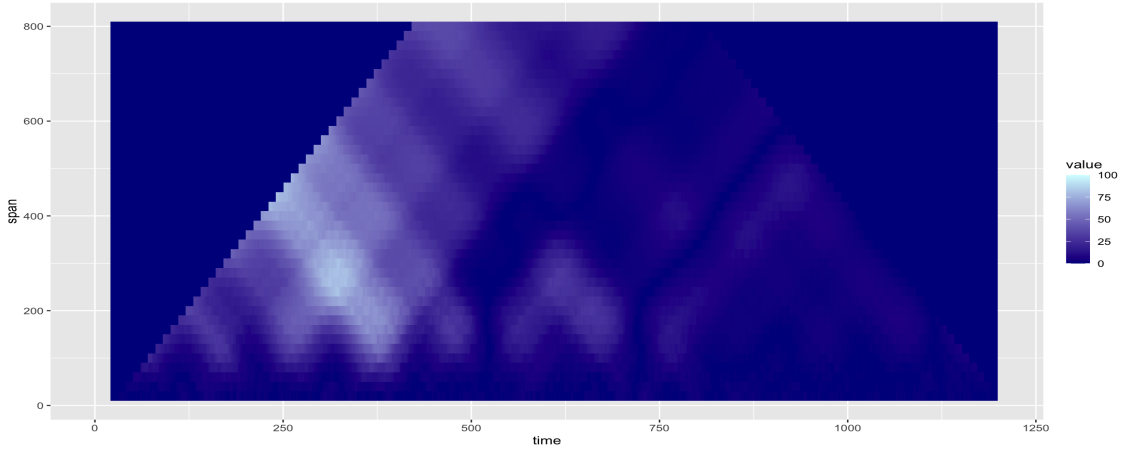


(c) coefficients

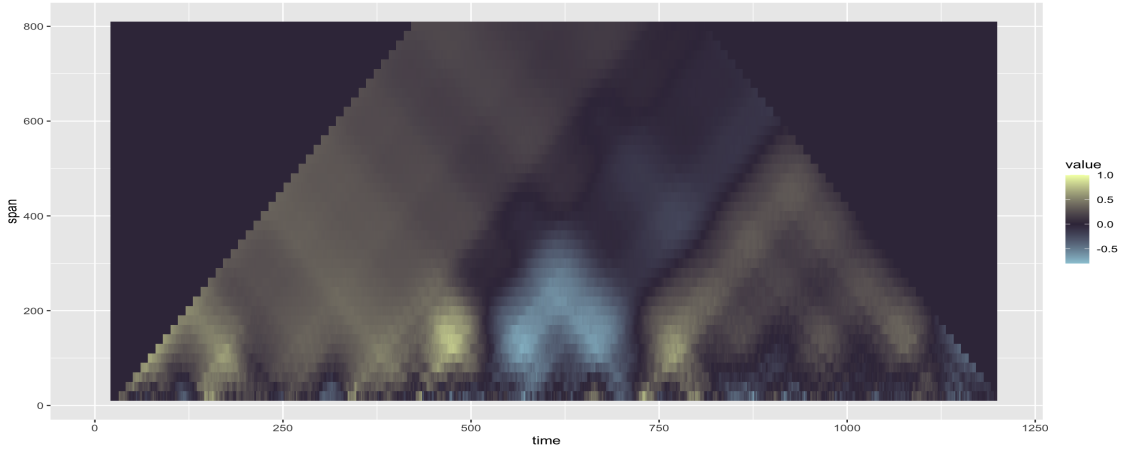
Figure 5.14: Combination of simulated data (a) and heatmaps [(b) and (c)] for dataset (M10). The blue vertical lines bound the examined area. As an example, the blue dashed line and purple dotted lines stand for the location and the corresponding boundaries of the symmetric moving window.



(a) Original data



(b) $-\log(p)$



(c) coefficients

Figure 5.15: Combination of simulated data (a) and heatmaps [(b) and (c)] for dataset (M11). The blue vertical lines bound the examined area. As an example, the blue dashed line and purple dotted lines stand for the location and the corresponding boundaries of the symmetric moving window.

5.4 Real-World Applications

In this section, we shall demonstrate the practical efficacy of MLLH through real-world examples, describing both its usefulness and potential pitfalls. We apply MLLH to COVID-19 datasets uploaded by (Mathieu et al., 2020) in order to visually present the lead-lag relationship between the number of new cases and deaths in a range of countries. Given the idea that the changes in the number of new cases will impact that of new deaths, i.e. the relation between time series have a natural direction, the strength of the potential lead-lag phenomenon can reasonably be reflected by the significance of dependence within this bi-variate dataset. As count data are commonly modelled as a Poisson-distributed series, we have decided to apply the Anscombe transform (Anscombe, 1948) to approximate the case-death bi-variate time series to a standard Gaussian distribution, i.e. we employ the transformation $x \to 2\sqrt{x + 3/8}$ to all observations.

Here we start with around four years of daily recorded time series between 03 January 2020 and 06 December 2023 in the UK. Of the provided series, we drop the data for the first month and the last two weeks since they contain all zeros for both new cases and deaths. The examined dataset is plotted in Figure 5.16(a), bounded by blue lines, where the new cases (deaths) is coloured black (red). To offer more information, we construct heatmaps at different scales, see Figure 5.16 and Figure 5.17 for results under window sizes $\{20, 40, \ldots, 800\}$ and $\{7, 14, \ldots, 280\}$ respectively.

Overall, the heatmaps in Figure 5.16 show that the bi-variate time series is generally positively correlated and the dependence between new cases and deaths is indeed significant. To specify, we can see many (nested) "triangular" areas in the heatmaps, especially for p-values. We can roughly find three regions across time locations, including $[0, 120]$, $[200, 420]$ and $[480, 1385]$ in the sub-figure (b). The robust evidence

at moderate scales in the first two regions aligns with the obvious co-movement in the original data, although the second region presents smaller coefficients. The darker area between the two regions indicates a locally low dependence and implies a change in the pattern of the bi-variate time series. In addition, Figure 5.18 displays the first 400 observations in the original dataset together with the corresponding MLLH, which again amplify the introduced idea. Moreover, within $[480, 1385]$, there are several apparent nested "triangular" areas delineated in different colours. For example, the dark triangles with their base across approximately $[480, 780]$ and $[780, 1200]$ indicate localised low significance of lead-lag relationships and suggests a dominant pattern for dependence within regions around $t = 780$. In addition, the left side of the largest triangle, together with the dark region between $[420, 480]$, highlights the changed pattern on two sides. Also, the obvious boundary dividing the two nested largest shapes with light colour presents a small change in characteristics of data around $t = 480$, see also time locations approximately at $t = 680, 780, 880$, etc. Such $(45°)$ stripes somehow suggest the presence of changing coincidence of the trends typically around the spikes of the bi-variate time series. For further details, we refer to Figure 5.17, which reports results obtained at finer scales. Notably, we can see frequent (slight) colour changes (stripes) within the interval $[680, 1100]$ across several the finest scales, and this corresponds to the weak coincidence of the trends at the locations of the spikes. To explore a potential reason for insignificant coefficients, we plot an additional sequence of the number of new vaccinations in Figure 5.19. It shows that the relatively insignificant relationship around $t = 700$ is intuitively related to the increase in vaccination rate.

Then, the second dataset we considered is the COVID curves recorded in China, which is indeed special due to the strict government policy employed over time, see (g) and (h) in Figure 5.20. Besides the number of new vaccinations, this figure also includes the government response stringency index, which is measured based on 9 response indicators such as school closures, workplace closures, and travel bans, etc (Mathieu et al., 2020). This index locates within the interval $[0, 100]$ with 100 representing the

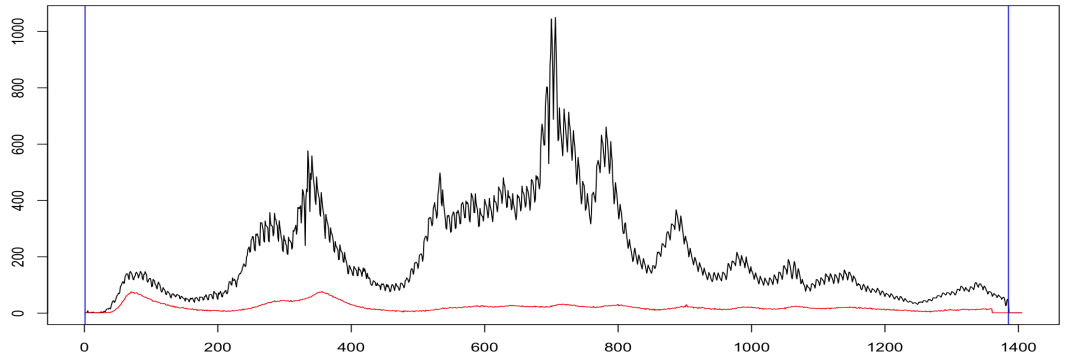
strictest response. Similarly, after removing time points where both new cases and deaths are all zeros, we present respectively the results of the entire remaining dataset in Figure 5.21 and those of the second half of data $[750, 1278]$ in Figure 5.22. Figure 5.21(b) generally highlights two main points: first, although the increase in new cases and deaths is largely controlled by the strict government policy, we can still observe their positive dependence. Also, considering the nested triangular areas (45° stripes) in the heatmaps, the coincidence of the trends at the locations of large spikes become more apparent in heatmap (b), see t around 30, 500, 800 and 1100 for example, and the test results will be more significant when such spikes are covered within the corresponding moving window, i.e. bounded by sides of various triangles. Second, from heatmaps (b) and (c), some obvious changes in the pattern of bi-variate time series are observed around $t = 60, 200, 750, 1100$, etc. In addition, the most significant dependence we can see in Figure 5.22 comes from the dramatic increase within the time length $[1070, 1130]$, which is caused by the sudden relaxation of government policy presented in 5.20(h).

Next we study the performance of MLLH in additional examples of case-death curves in other countries. Overall, Figure 5.23 to Figure 5.34 present the significance of lead-lag relationships between the recorded number of new cases and deaths in various countries, together with some different patterns in the original data highlighted via heatmaps (b). In the following, we continue to concentrate on the daily recorded dataset and hence conduct the tests over periods with different length and starting dates. For instance, since Mathieu et al. (2020) only provides weekly data in Brazil after 2023-03-06, our analysis is conducted over the period from 2020-02-27 to 2023-03-06, as bounded with blue vertical lines in Figure 5.23(a). Figure 5.23 indicates the consistently significant lead-lag relationship of the dataset recorded in Brazil, where an apparent shift in pattern is located around $t = 750$. For the dataset collected in Canada from 2020-01-26 to 2022-06-11, we can see in Figure 5.24 that there exists localised low significance of lead-lag relationships within regions $[200, 240]$, $[440, 550]$ and $[650, 740]$ while a dominant pattern of significant dependence can be observed within regions

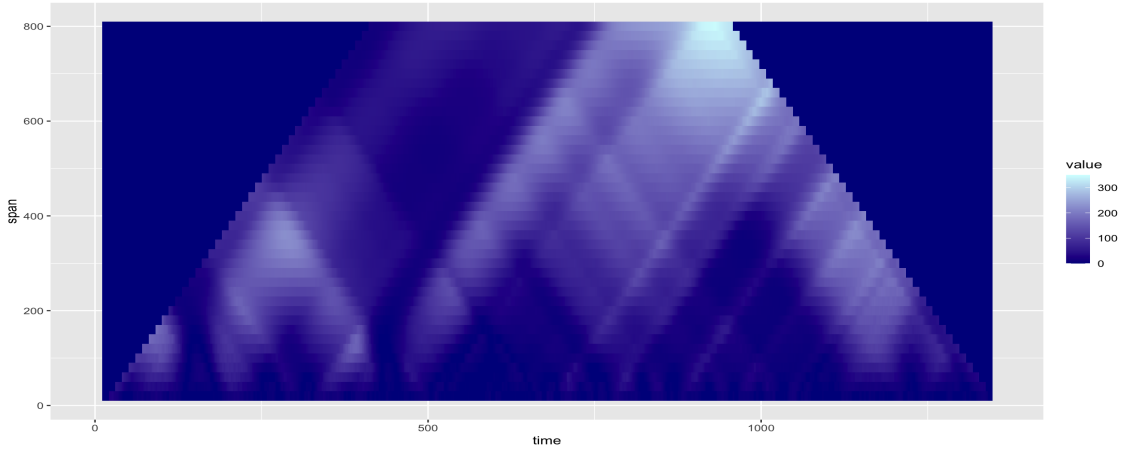
around $[250, 300]$. Also, an obvious abnormal pattern, i.e. the dark area between two nested triangles in light colour, is present around $t = 750$, which indicates the substantially changed coincidence of the trends at the spike around $t = 750$. In Figure 5.25, we plot heatmaps for the case-death curves in India from 2020-03-03 to 2023-12-03. The 45° stripes in the heatmaps suggest that the data experience considerable pattern shifts within regions around $t = 500, 780, 1200$, etc.

Additionally, Figure 5.26 to Figure 5.34 present tested results over case-death data recorded respectively from 2020-02-22 to 2023-11-16 in Italy, from 2020-02-11 to 2023-05-08 in Japan, from 2020-03-03 to 2023-03-10 in Malaysia, from 2020-03-07 to 2023-11-21 in Poland, from 2020-03-19 to 2023-05-15 in Russian, from 2020-02-05 to 2023-02-13 in Singapore, from 2020-03-06 to 2022-07-22 in South Africa, from 2020-02-16 to 2023-06-01 in South Korea, and from 2020-02-02 to 2022-10-19 in the USA. Besides the significant dependence shown within $[60, 150]$, Figure 5.26(b) indicates dominant patterns at locations around $t = 300, 530, 750, 1100$ and a notable shift in pattern at time around $t = 630$. Similar to the most obvious pattern change in Figure 5.16, we can not figure out the relatively more significant pattern on two sides. For data collected for Japan, Figure 5.27(b) presents evident 45° stripes indicating changes in coincidence of the trends within regions around spikes at approximately $t = 375, 600, 750, 900$, etc. In the remaining figures, we can also find the generally positive correlations, the areas with more significant dependence between bi-variate time series and the correspondence between stripes and the coincidence of the trends around the locations of the spikes, which helps decide the bounds of intervals with different patterns for further analysis.

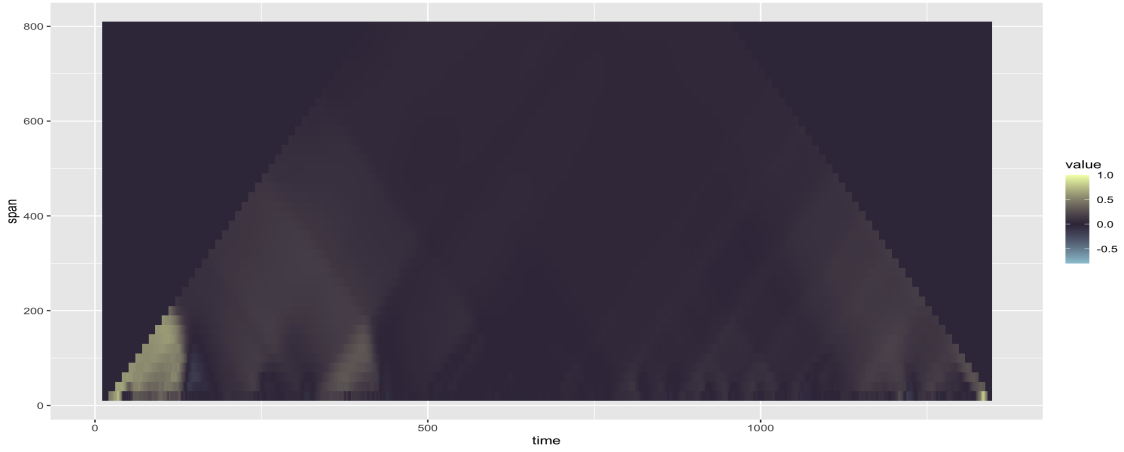
Additionally, relying on the empirical frequencies roughly observed from the real-world data, we can see that a suitable window size can be 200 for bi-variate time series recorded in India and Malaysia while we can choose scale 100 for analysis in the other tested countries.



(a) Original data

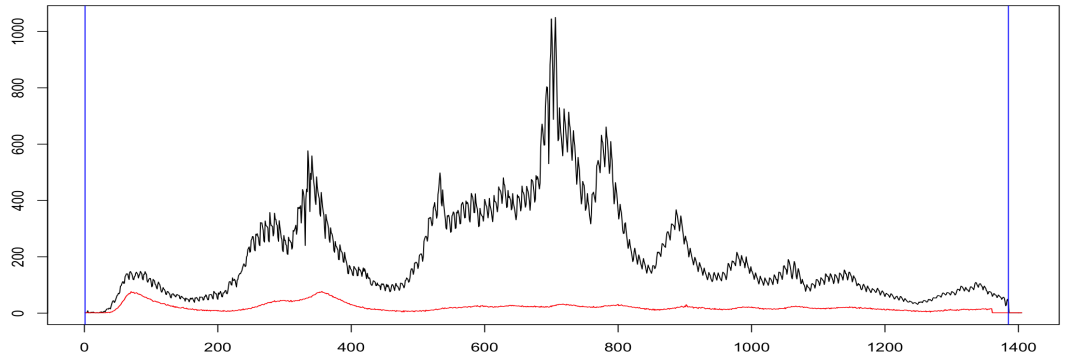


(b) $-\log(p)$

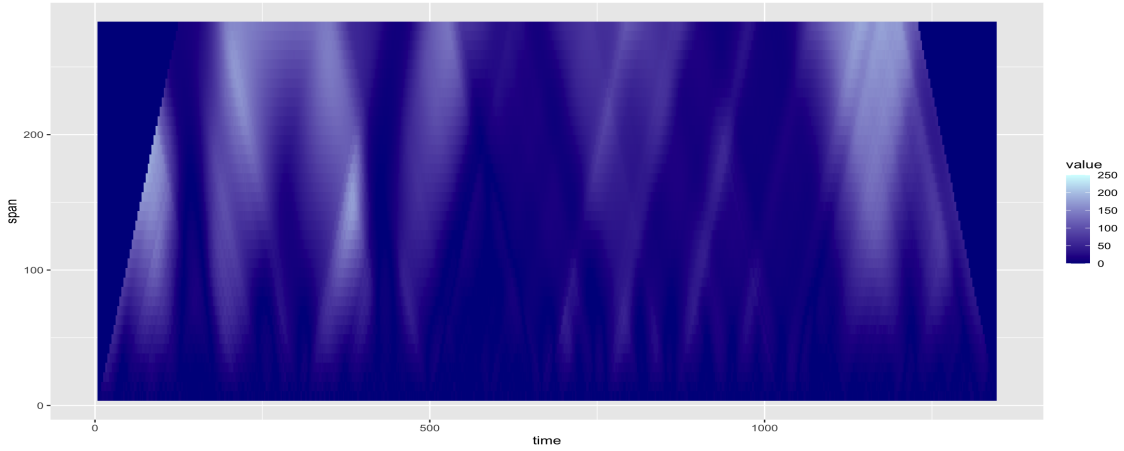


(c) coefficients

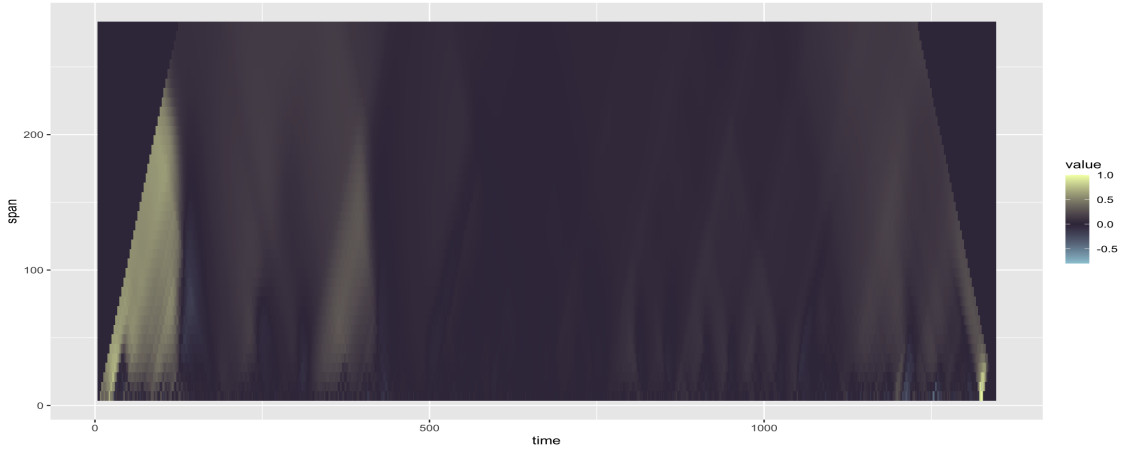
Figure 5.16: Combination of original data (a) and heatmaps [(b) and (c)] for new cases (black) and new deaths in UK recorded from 2020-02-01 to 2023-12-06. The chosen window sizes are $20, 40, \ldots, 800$. The blue vertical lines in (a) bound the examined area.



(a) Original data

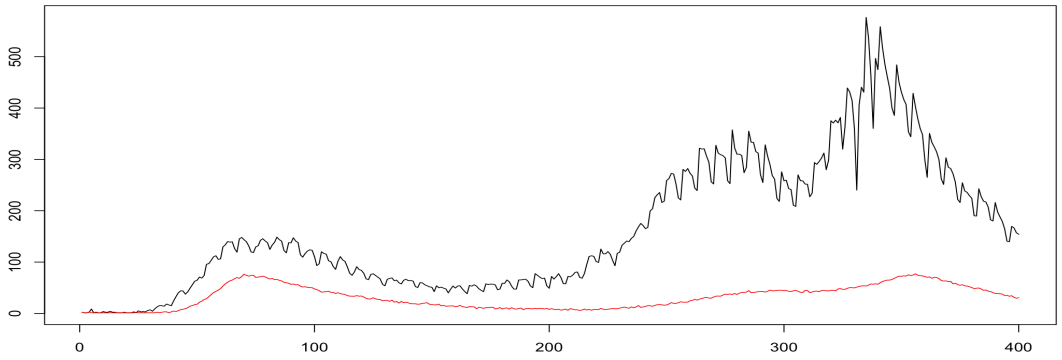


(b) $-\log(p)$

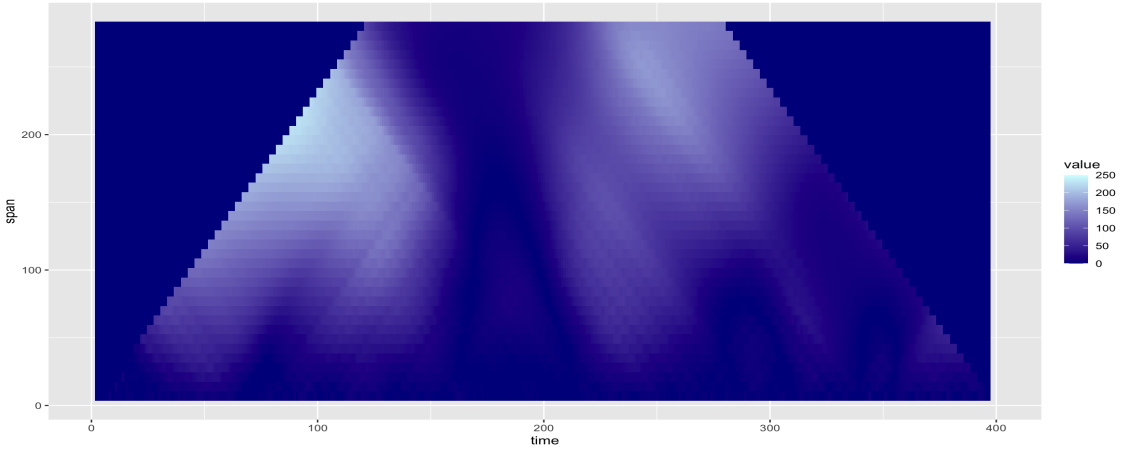


(c) coefficients

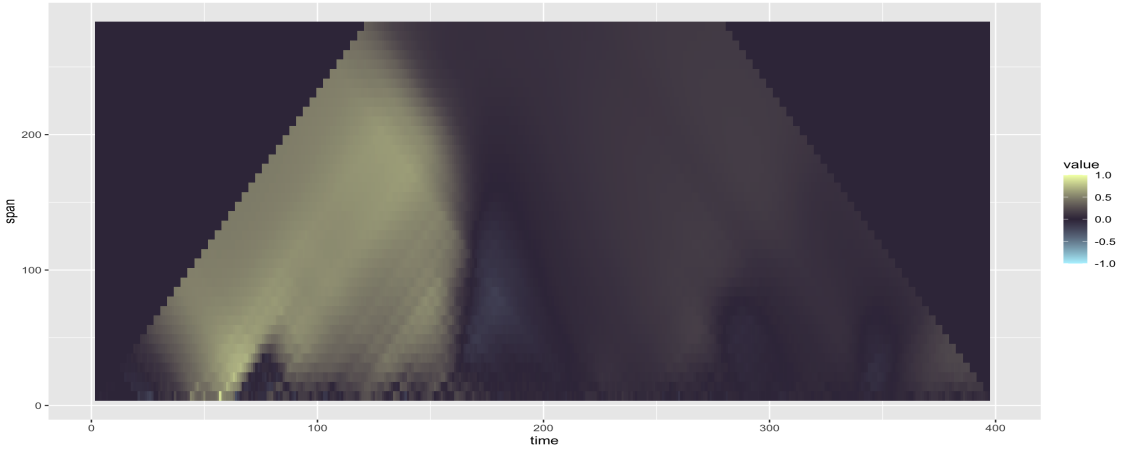
Figure 5.17: Combination of original data (a) and heatmaps [(b) and (c)] for new cases (black) and new deaths in UK recorded from 2020-02-01 to 2023-12-06. The chosen window sizes are $7, 14, \ldots, 270$. The blue vertical lines in (a) bound the examined area.



(a) Original data



(b) $-\log(p)$



(c) coefficients

Figure 5.18: Combination of original data (a) and heatmaps [(b) and (c)] for new cases (black) and new deaths in UK recorded from 2020-02-01 to 2021-03-06 (400 days). The chosen window sizes are $7, 14, \ldots, 270$. The blue vertical lines in (a) bound the examined area.

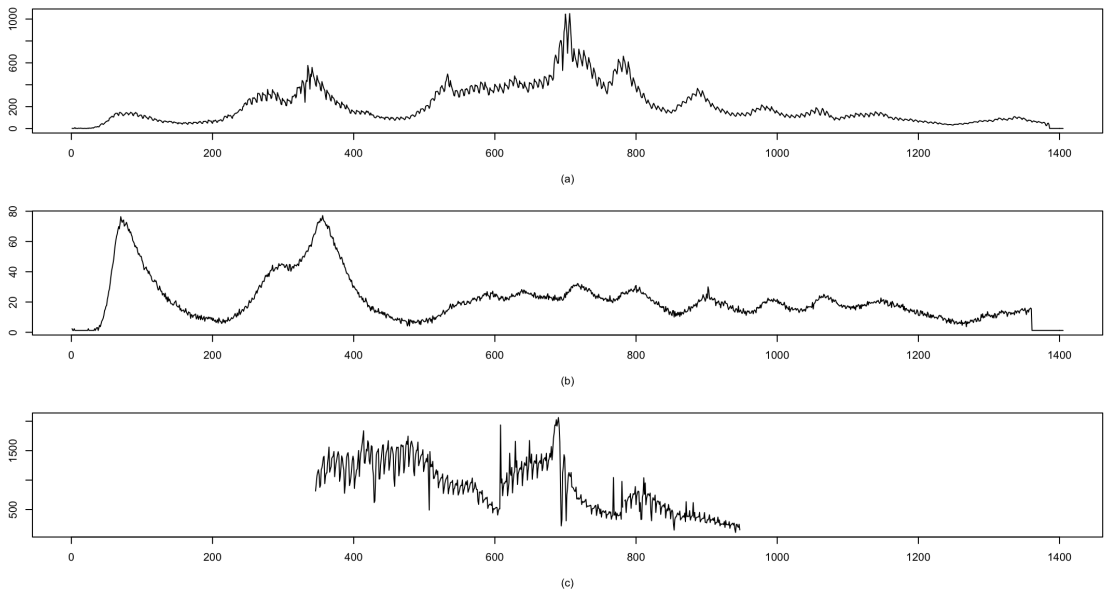


Figure 5.19: Combination of original data: (a) new cases, (b) new deaths and (c) new vaccinations in UK recorded from 2020-02-01 to 2023-12-06.

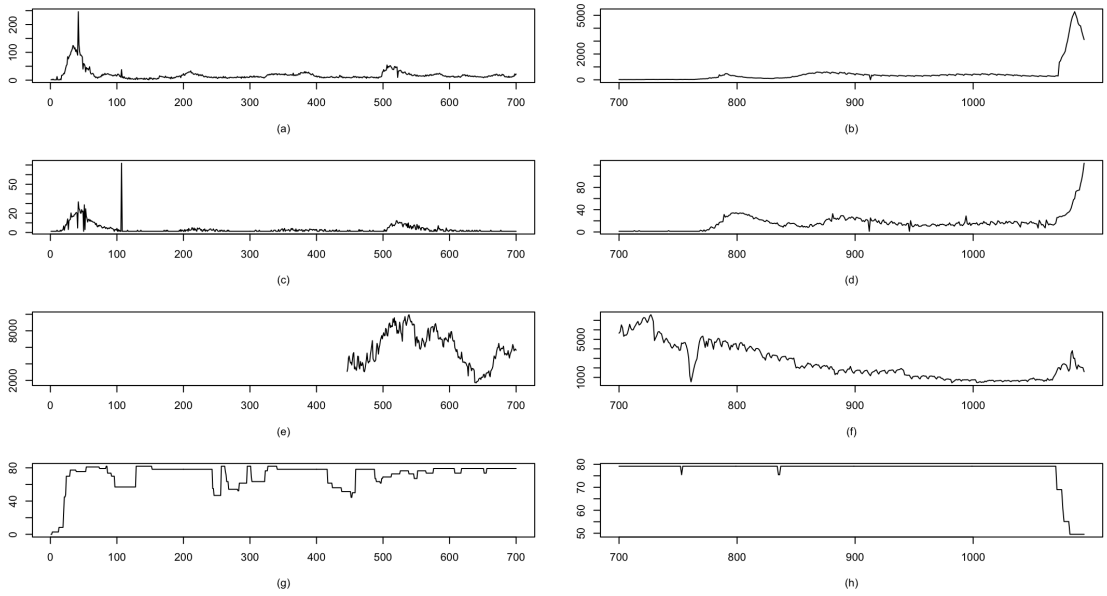
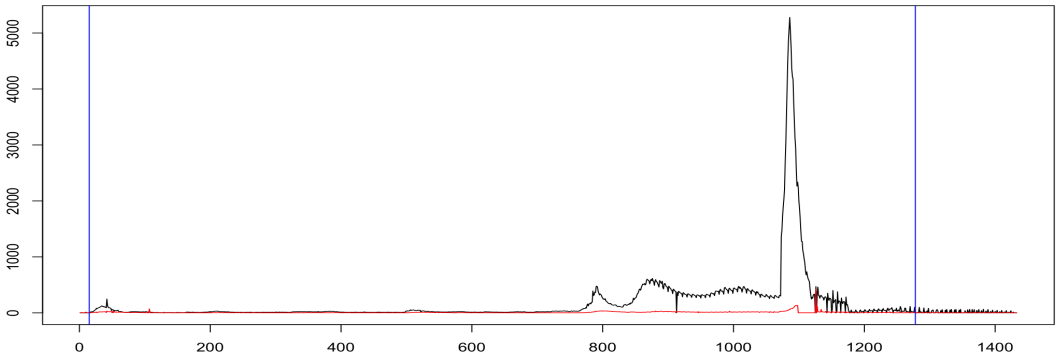
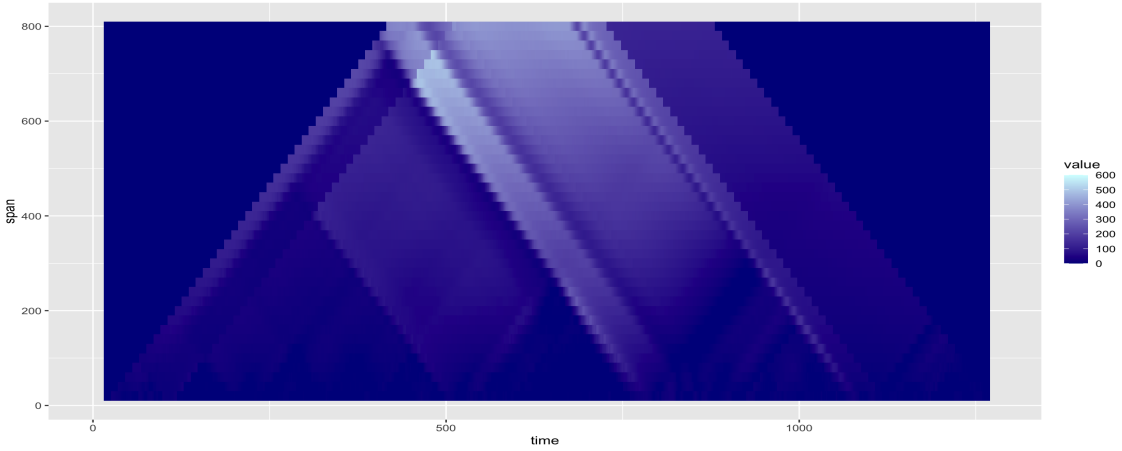


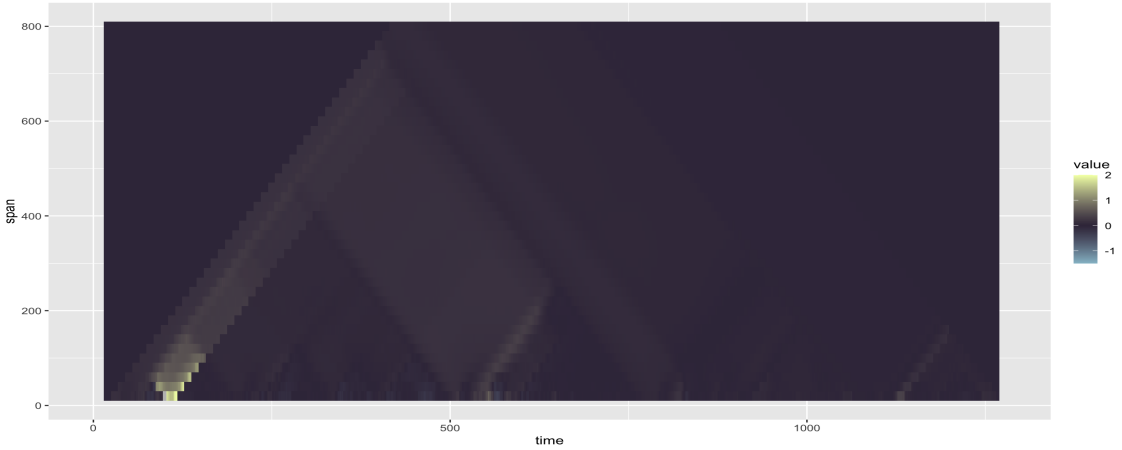
Figure 5.20: Combination of original data: (a)-(b) new cases, (c)-(d) new deaths, (e)-(f) new vaccinations and (g)-(h) government response stringency index in China recorded from 2020-01-03 to 2022-12-31.



(a) Original data

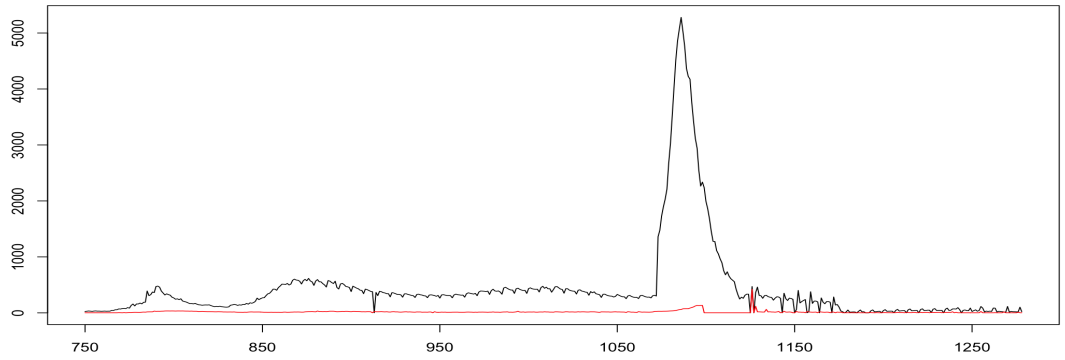


(b) $-\log(p)$

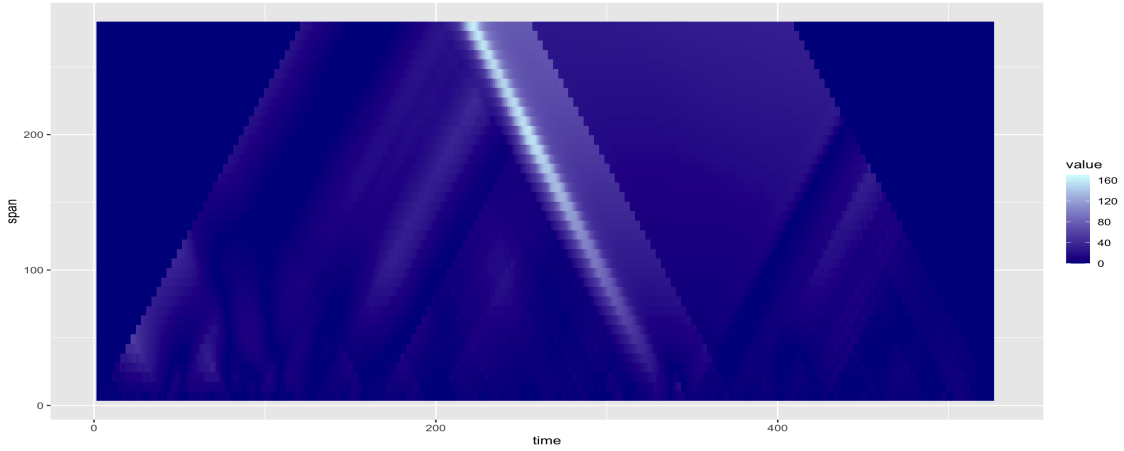


(c) coefficients

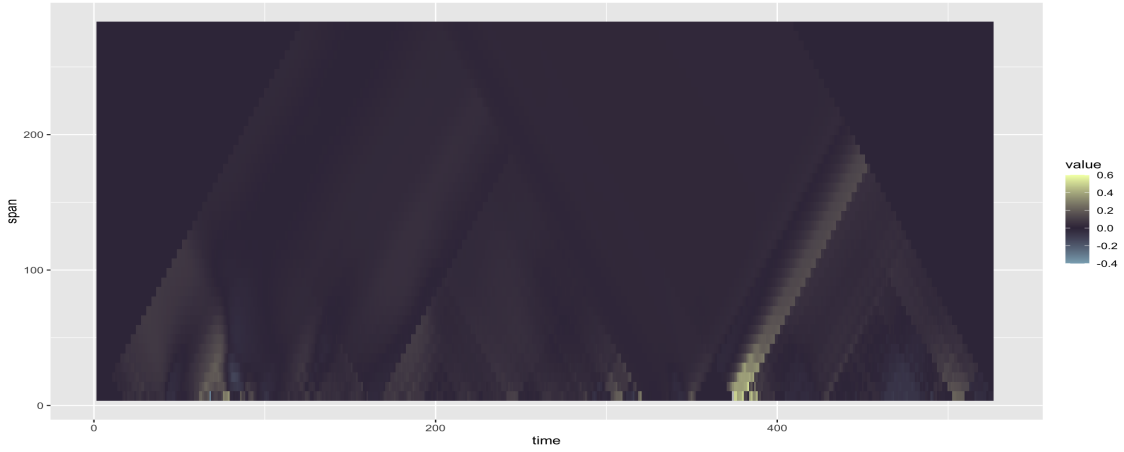
Figure 5.21: Combination of original data (a) and heatmaps [(b) and (c)] for new cases (black) and new deaths in China recorded from 2020-01-03 to 2023-12-06. The chosen window sizes are $20, 40, \ldots, 800$. The blue vertical lines in (a) bound the examined area.



(a) Original data

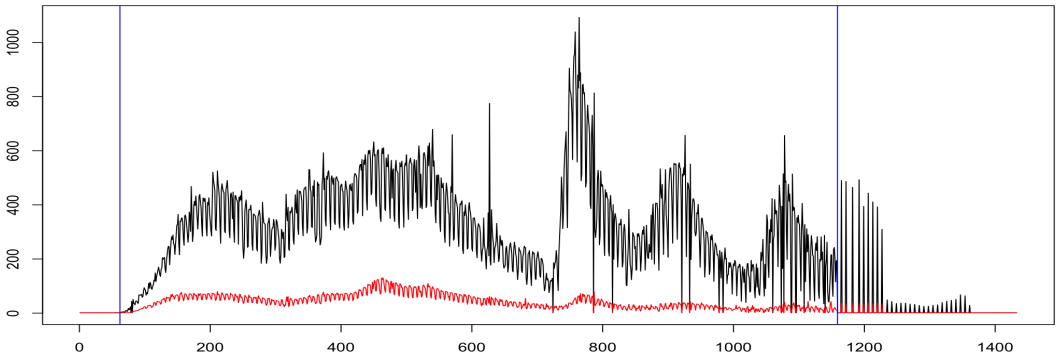


(b) $-\log(p)$

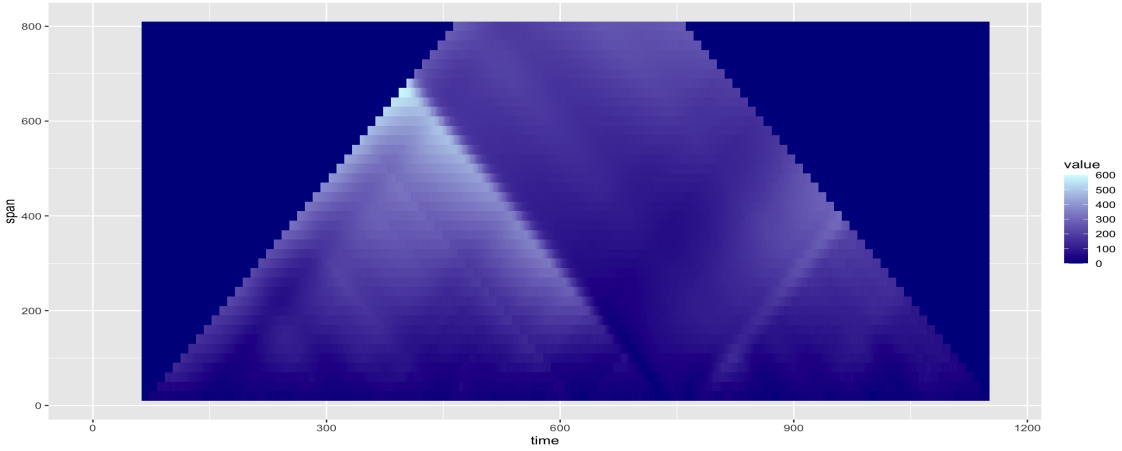


(c) coefficients

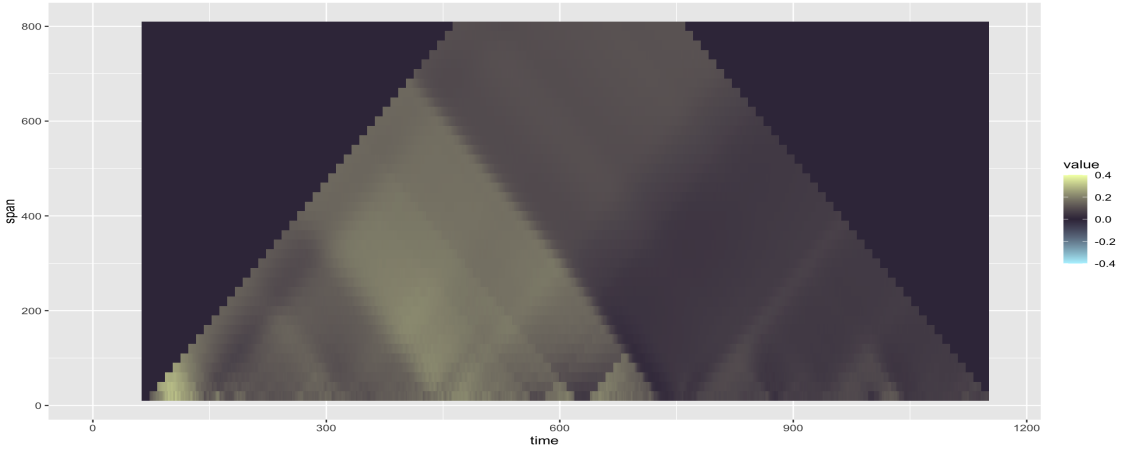
Figure 5.22: Combination of original data (a) and heatmaps [(b) and (c)] for new cases (black) and new deaths in China recorded from 2022-01-20 to 2023-07-03. The chosen window sizes are $7, 14, \ldots, 270$.



(a) Original data

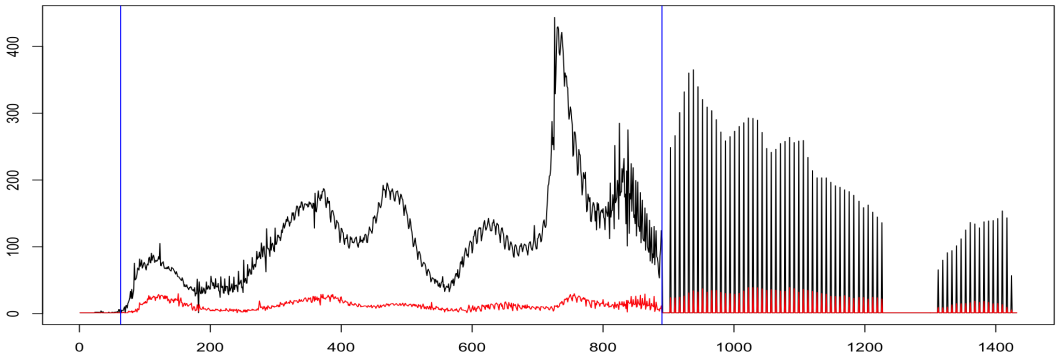


(b) $-\log(p)$

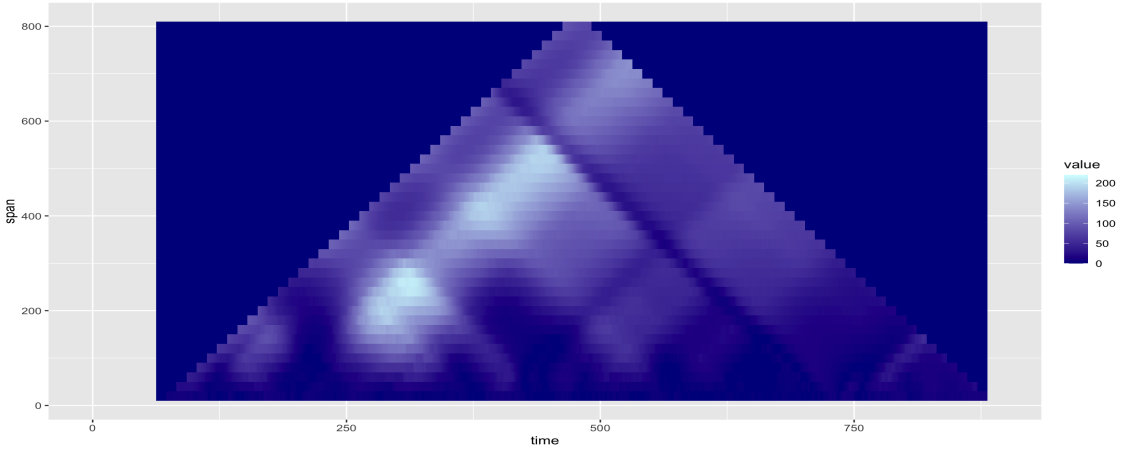


(c) coefficients

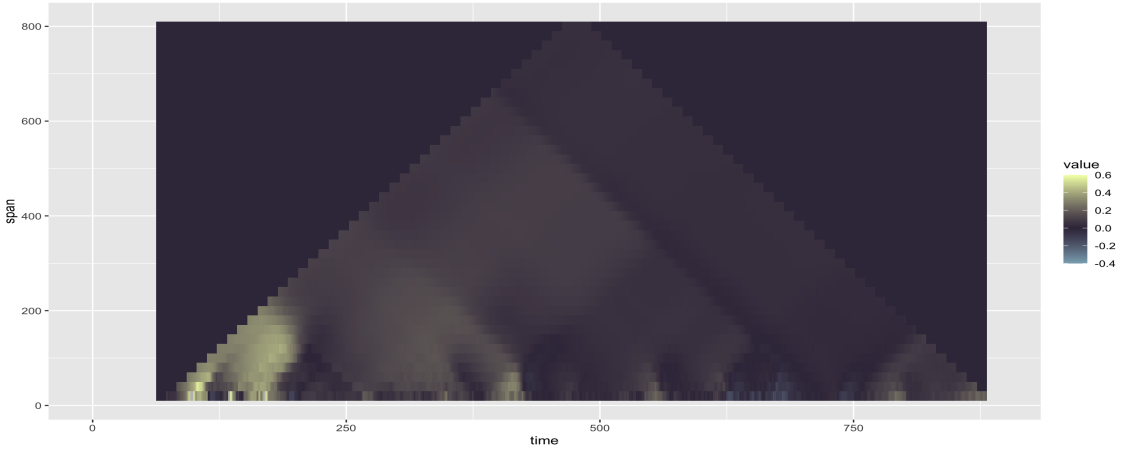
Figure 5.23: Combination of original data (a) and heatmaps [(b) and (c)] for new cases (black) and new deaths in Brazil recorded from 2020-01-03 to 2023-12-06. The chosen window sizes are $20, 40, \ldots, 800$. The blue vertical lines in (a) bound the examined area.



(a) Original data

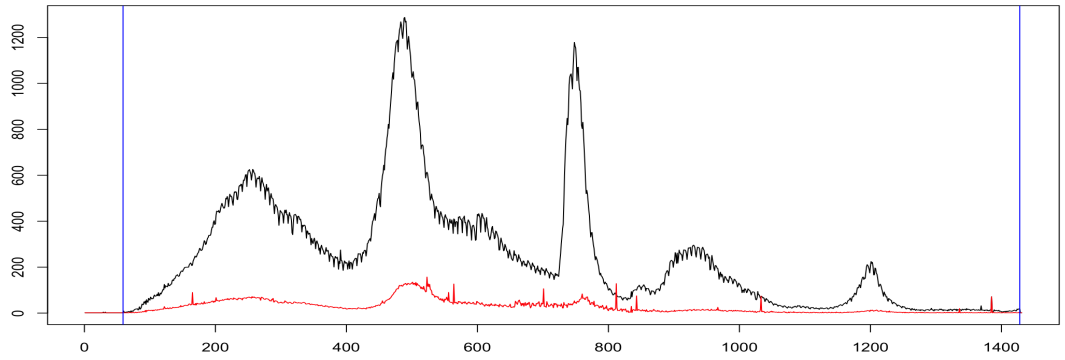


(b) $-\log(p)$

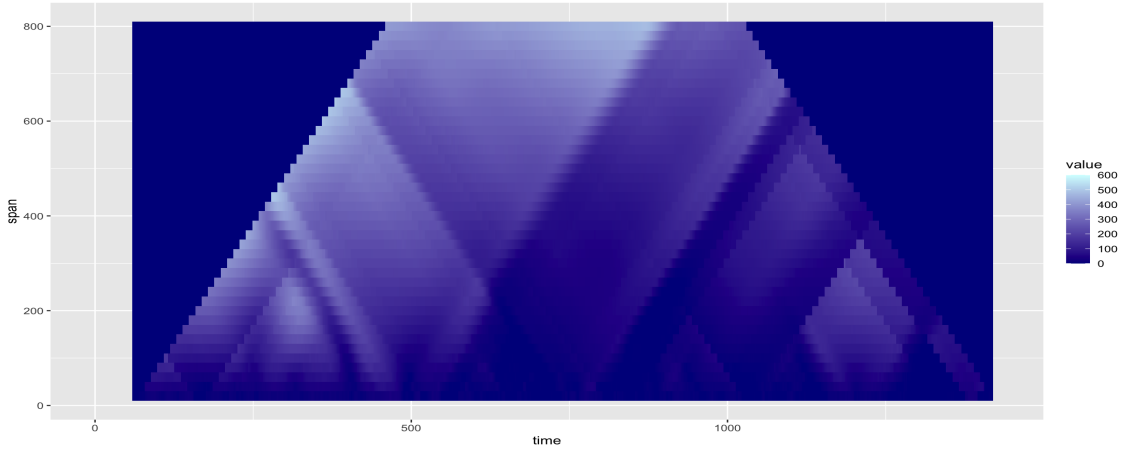


(c) coefficients

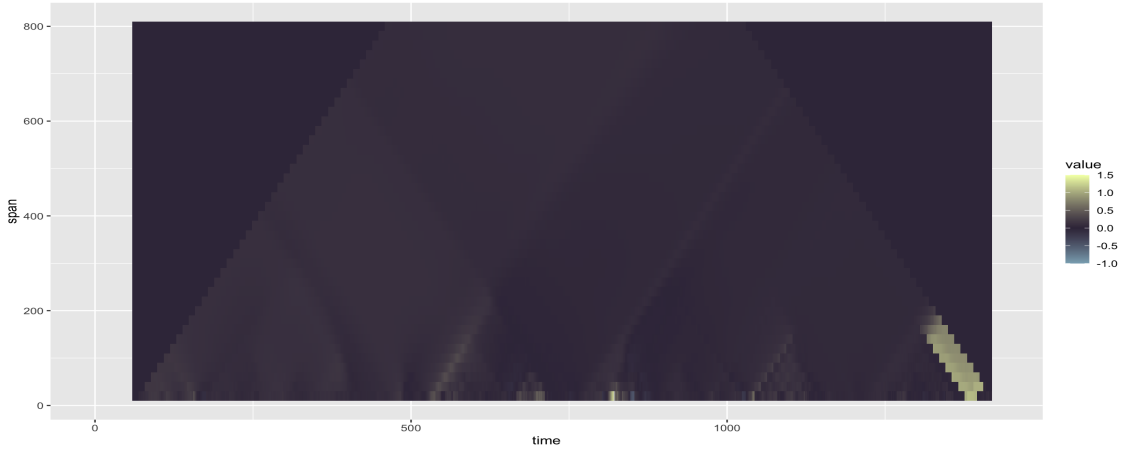
Figure 5.24: Combination of original data (a) and heatmaps [(b) and (c)] for new cases (black) and new deaths in Canada recorded from 2020-01-03 to 2023-12-06. The chosen window sizes are $20, 40, \ldots, 800$. The blue vertical lines in (a) bound the examined area.



(a) Original data

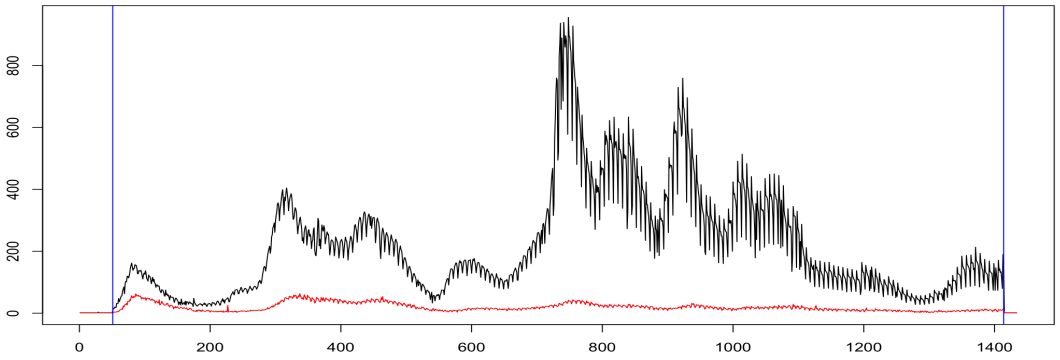


(b) $-\log(p)$

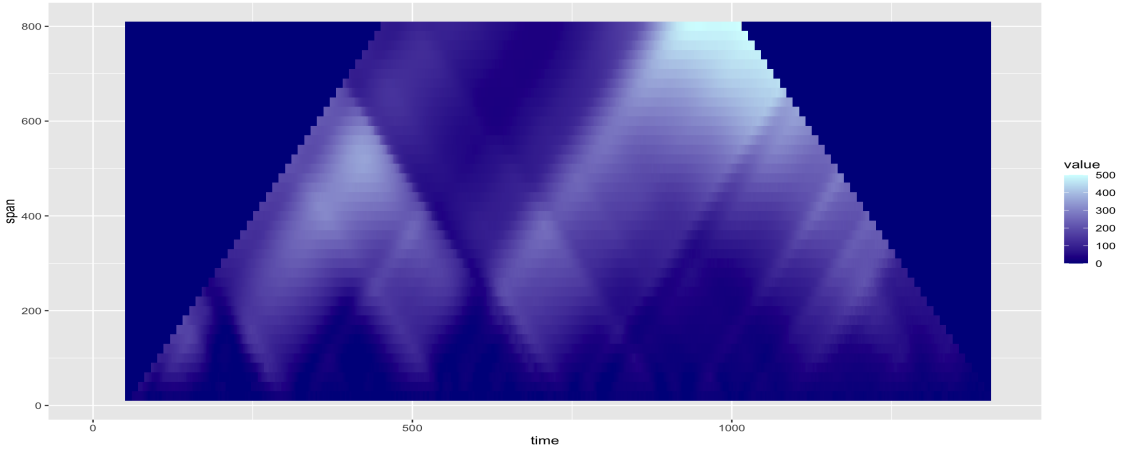


(c) coefficients

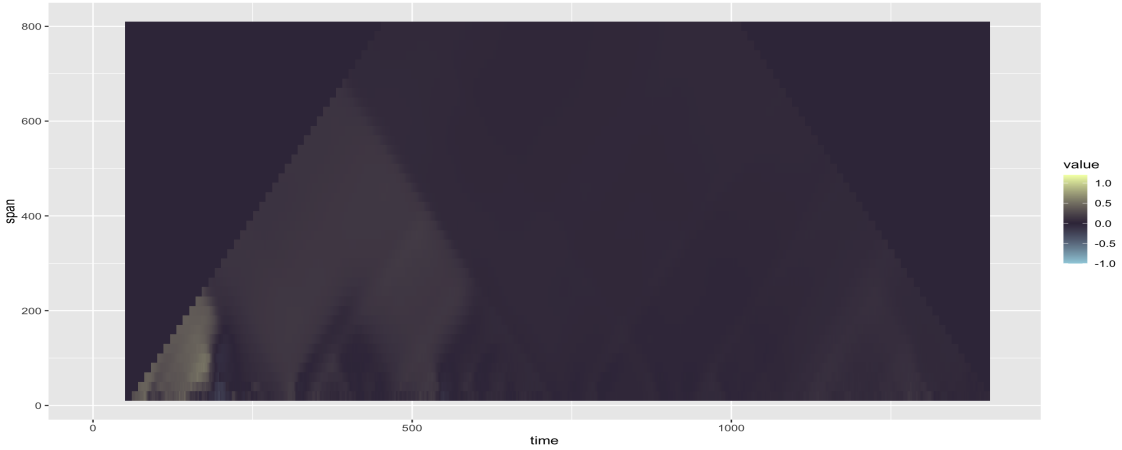
Figure 5.25: Combination of original data (a) and heatmaps [(b) and (c)] for new cases (black) and new deaths in India recorded from 2020-02-01 to 2023-12-06. The chosen window sizes are $20, 40, \ldots, 800$. The blue vertical lines in (a) bound the examined area.



(a) Original data

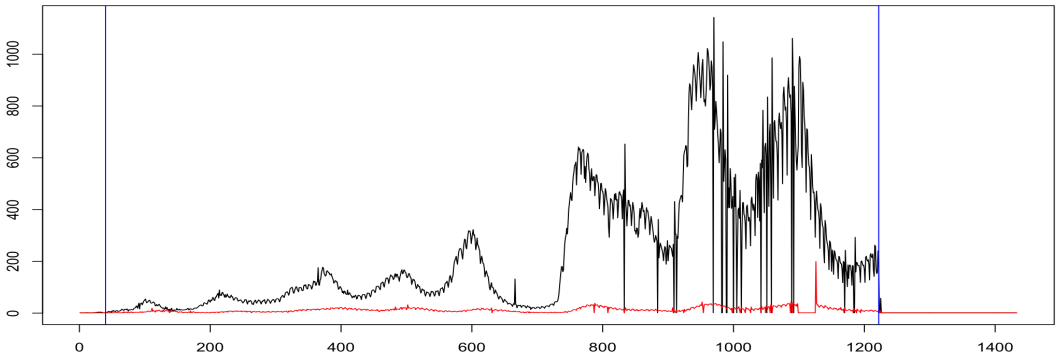


(b) $-\log(p)$

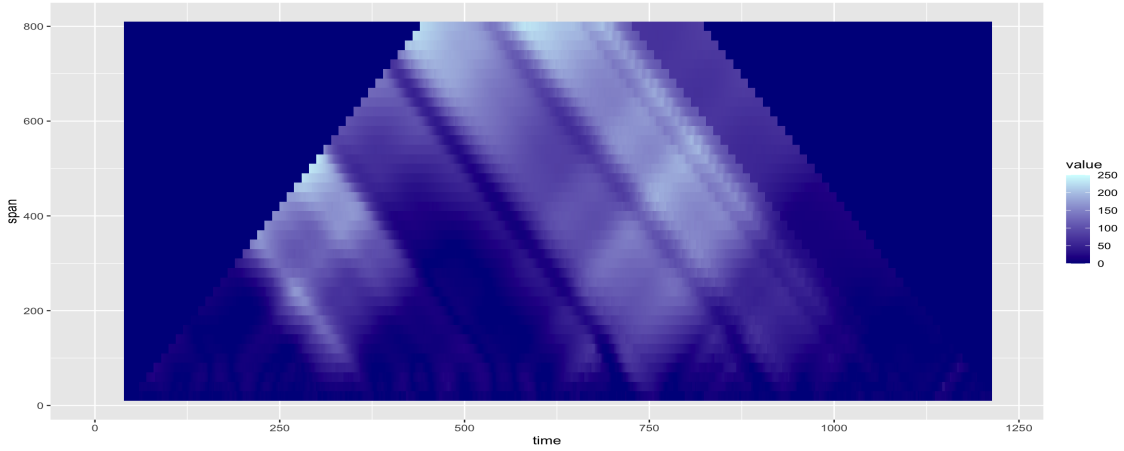


(c) coefficients

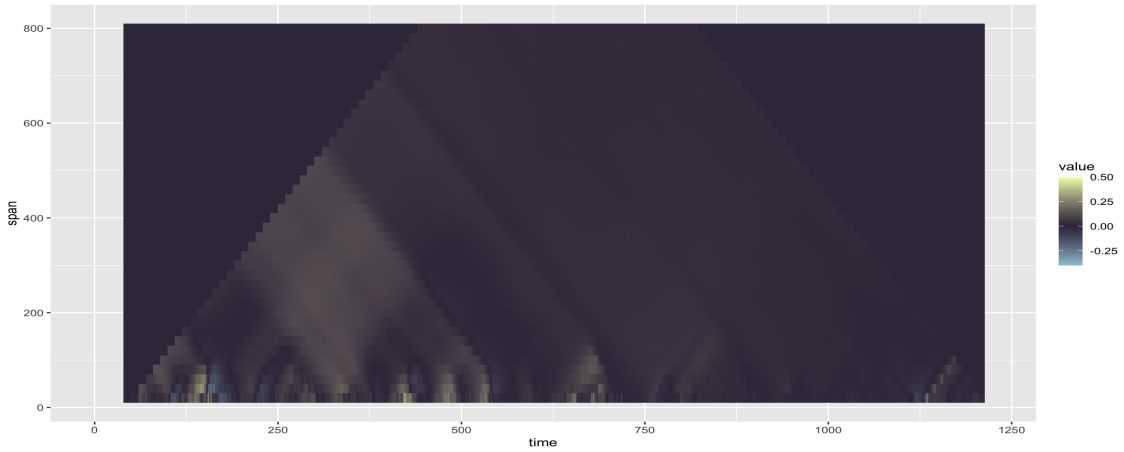
Figure 5.26: Combination of original data (a) and heatmaps [(b) and (c)] for new cases (black) and new deaths in Italy recorded from 2020-01-03 to 2023-12-06. The chosen window sizes are $20, 40, \ldots, 800$. The blue vertical lines in (a) bound the examined area.



(a) Original data

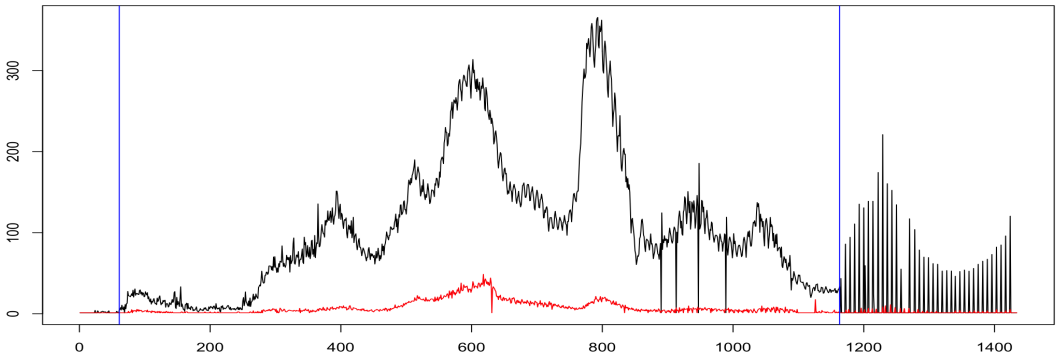


(b) $-\log(p)$

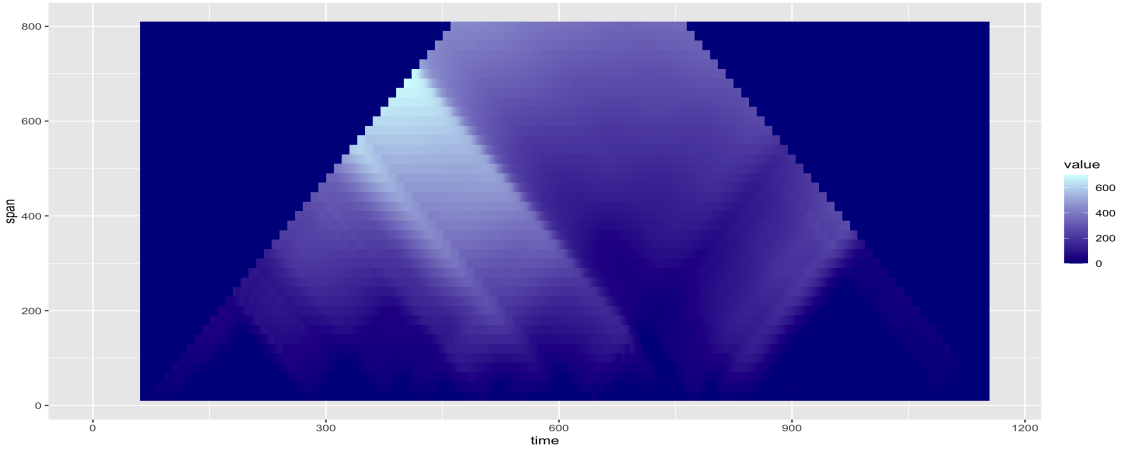


(c) coefficients

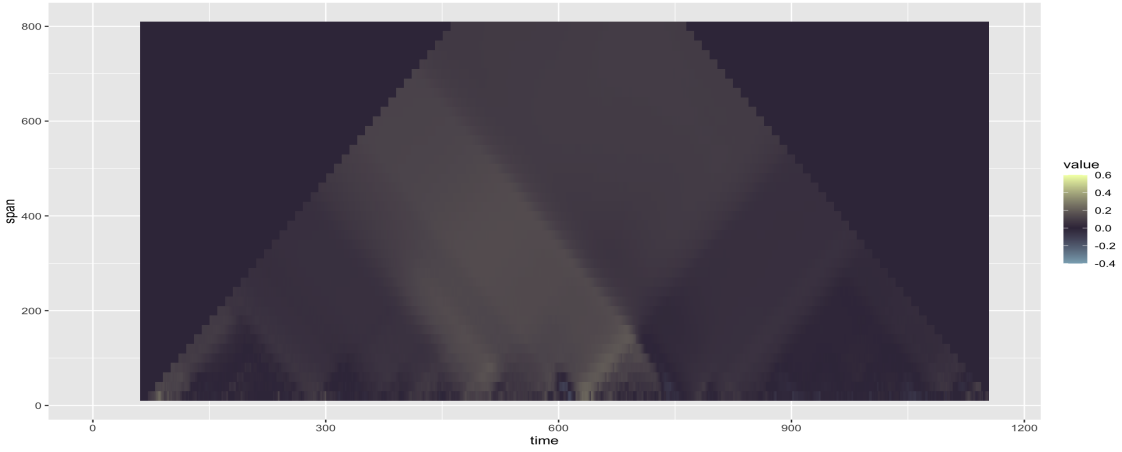
Figure 5.27: Combination of original data (a) and heatmaps [(b) and (c)] for new cases (black) and new deaths in Japan recorded from 2020-01-03 to 2023-12-06. The chosen window sizes are $20, 40, \ldots, 800$. The blue vertical lines in (a) bound the examined area.



(a) Original data

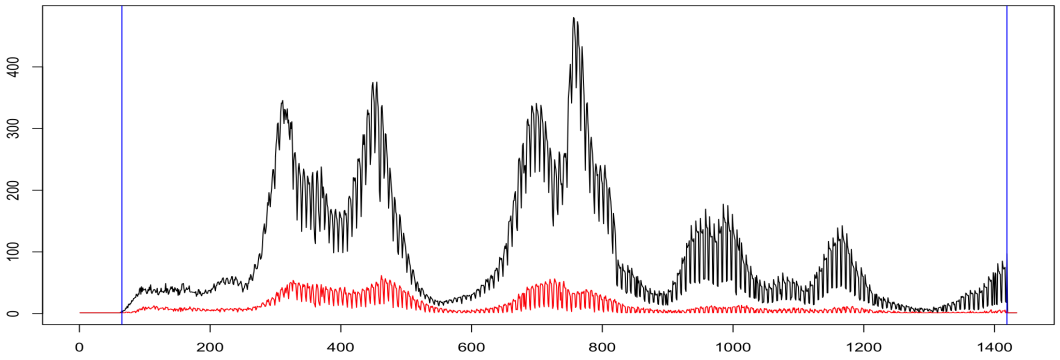


(b) $-\log(p)$

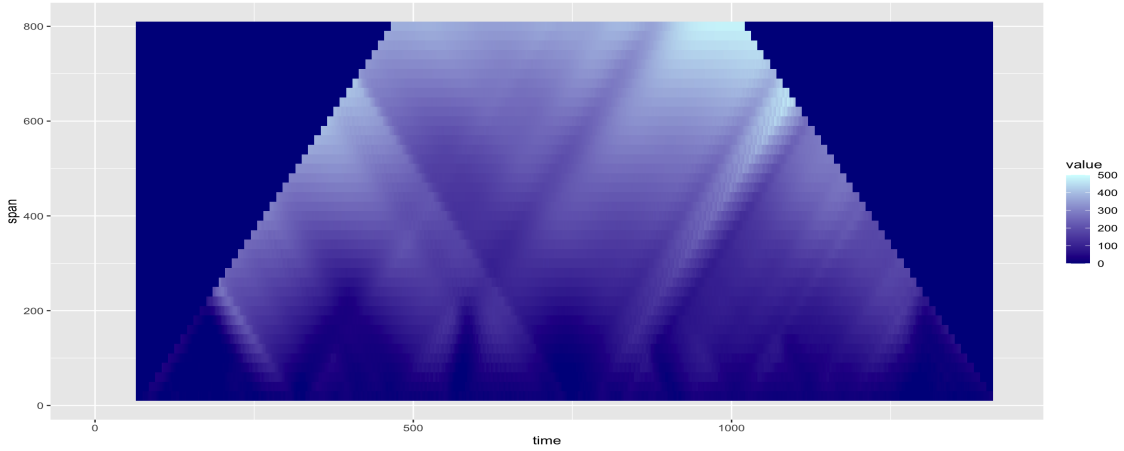


(c) coefficients

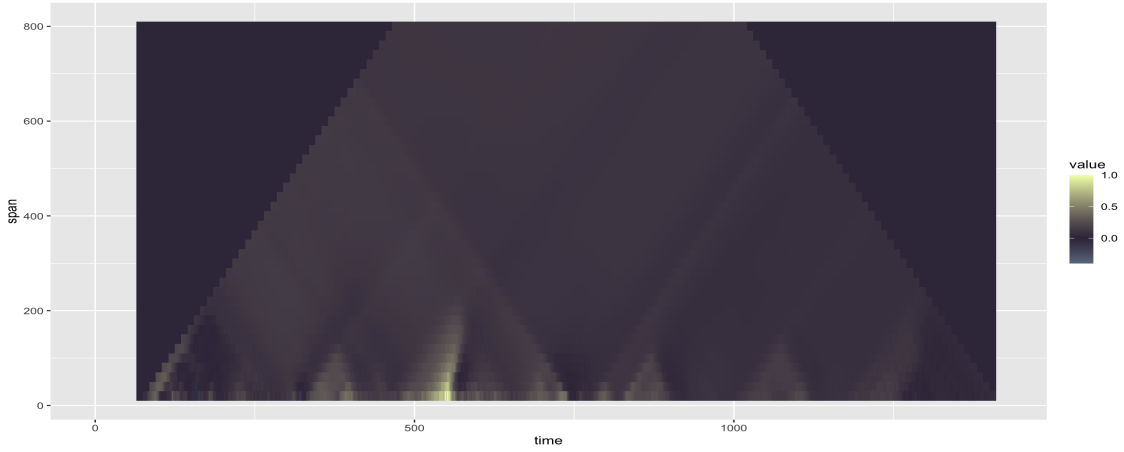
Figure 5.28: Combination of original data (a) and heatmaps [(b) and (c)] for new cases (black) and new deaths in Malaysia recorded from 2020-01-03 to 2023-12-06. The chosen window sizes are $20, 40, \ldots, 800$. The blue vertical lines in (a) bound the examined area.



(a) Original data

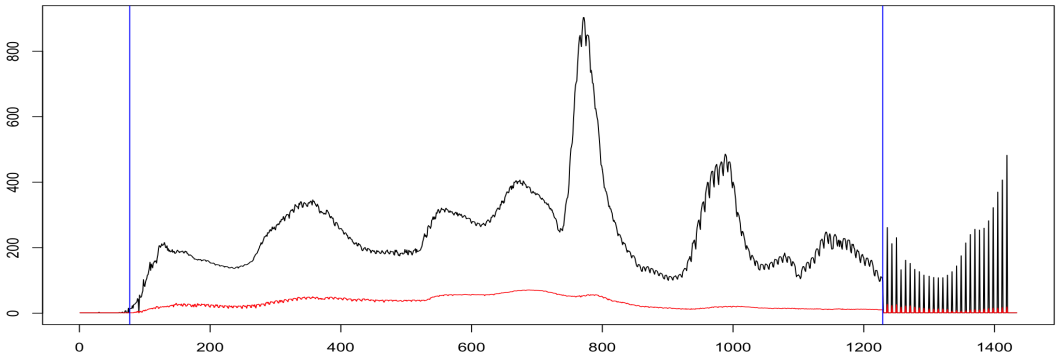


(b) $-\log(p)$

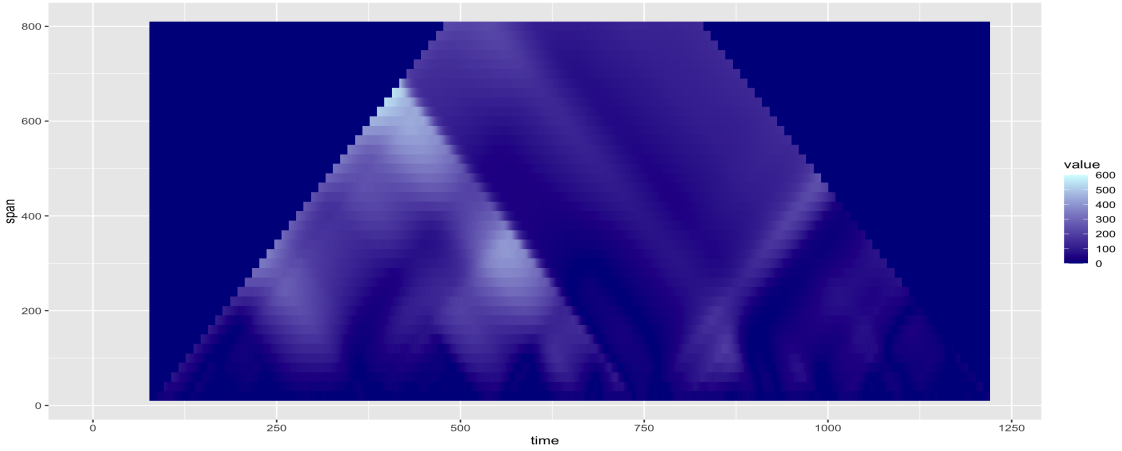


(c) coefficients

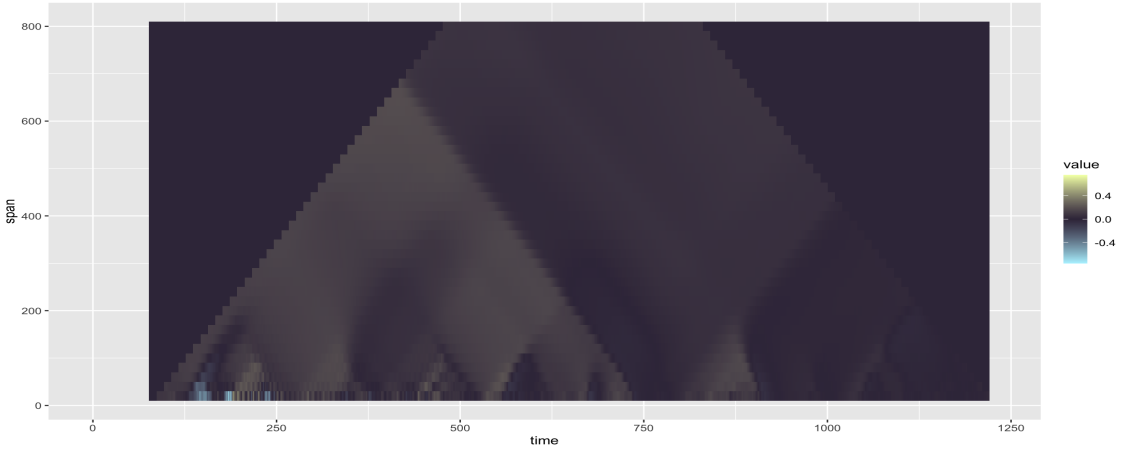
Figure 5.29: Combination of original data (a) and heatmaps [(b) and (c)] for new cases (black) and new deaths in Poland recorded from 2020-01-03 to 2023-12-06. The chosen window sizes are $20, 40, \ldots, 800$. The blue vertical lines in (a) bound the examined area.



(a) Original data

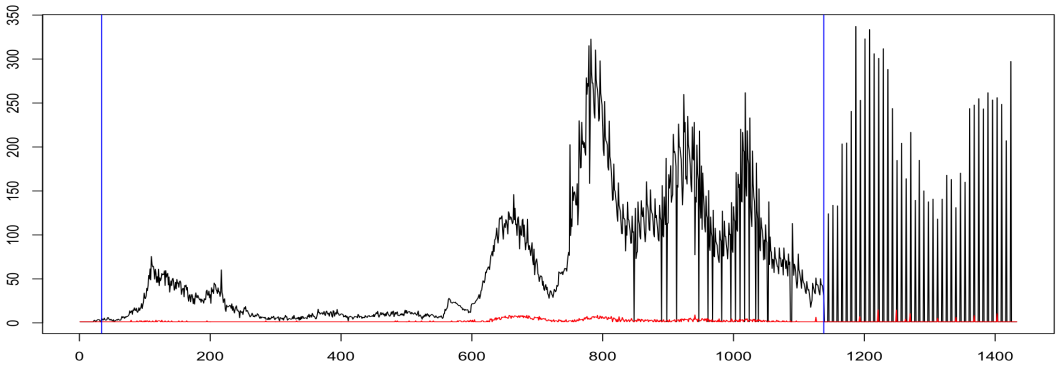


(b) $-\log(p)$

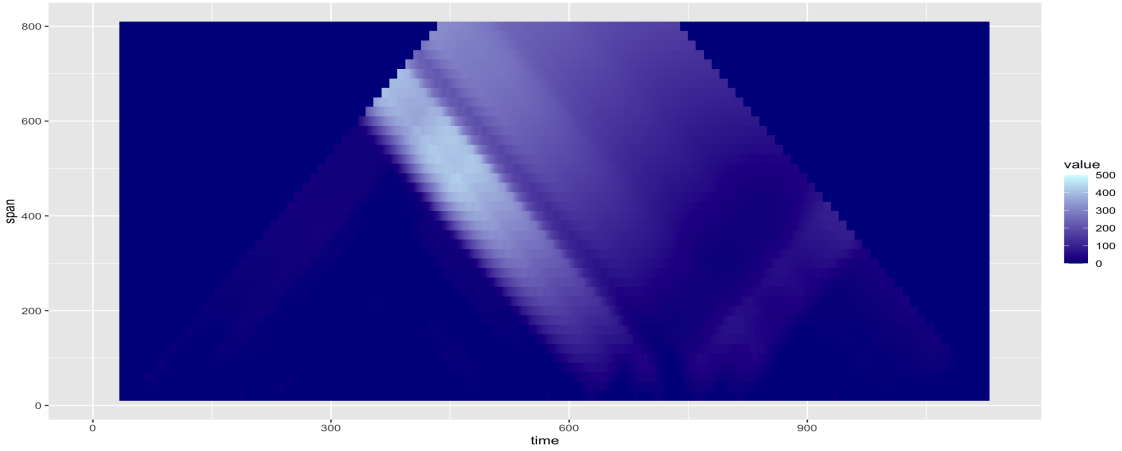


(c) coefficients

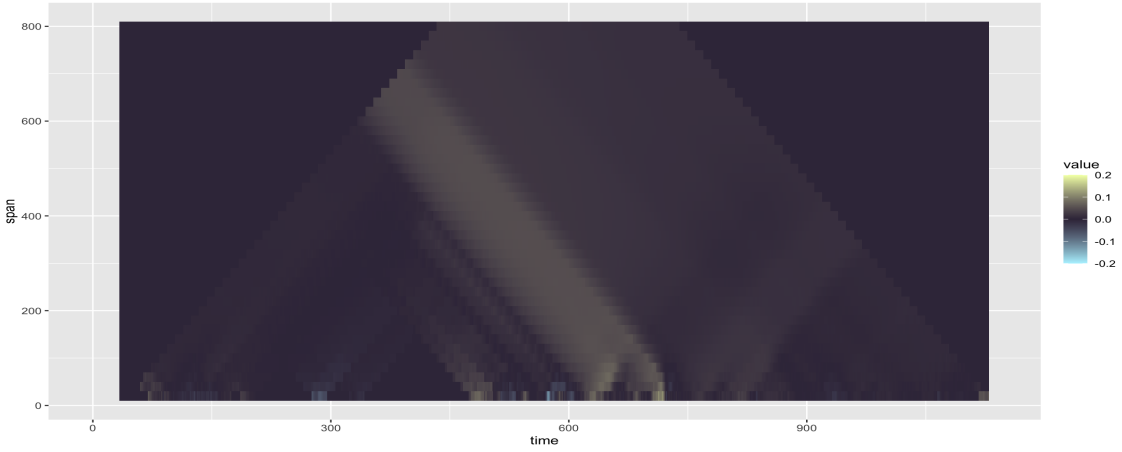
Figure 5.30: Combination of original data (a) and heatmaps [(b) and (c)] for new cases (black) and new deaths in Russia recorded from 2020-01-03 to 2023-12-06. The chosen window sizes are $20, 40, \ldots, 800$. The blue vertical lines in (a) bound the examined area.



(a) Original data

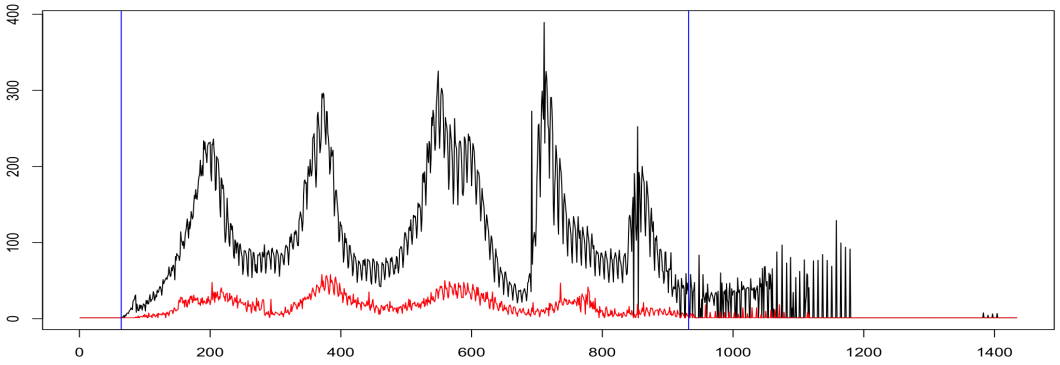


(b) $-\log(p)$

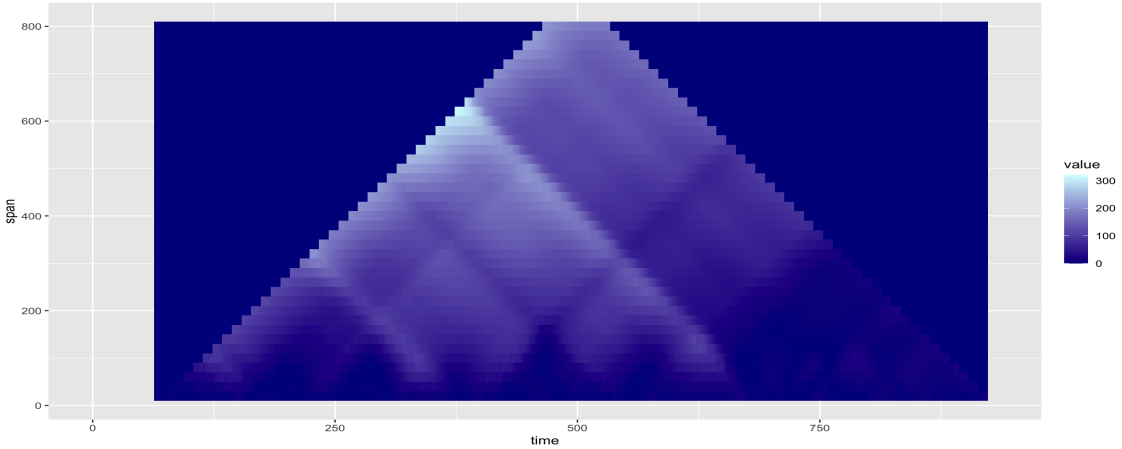


(c) coefficients

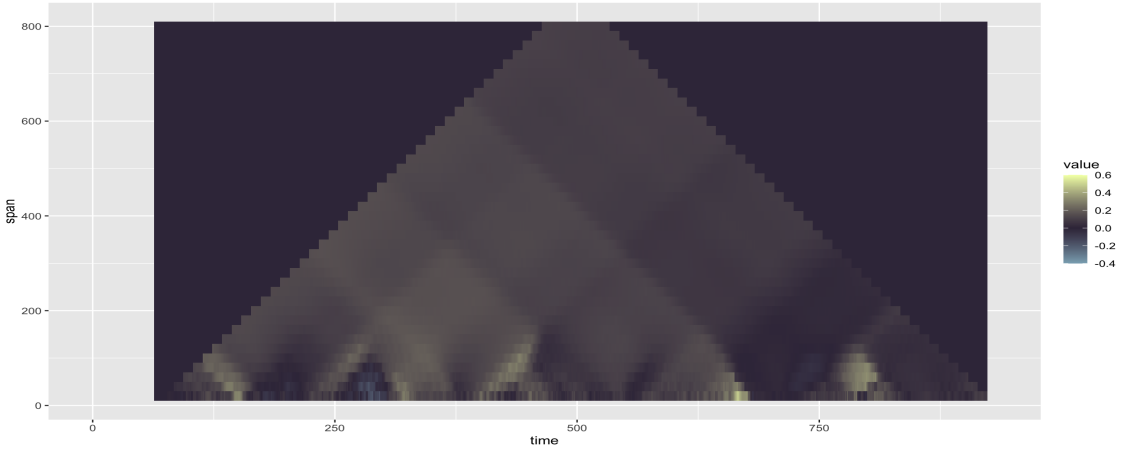
Figure 5.31: Combination of original data (a) and heatmaps [(b) and (c)] for new cases (black) and new deaths in Singapore recorded from 2020-01-03 to 2023-12-06. The chosen window sizes are $20, 40, \ldots, 800$. The blue vertical lines in (a) bound the examined area.



(a) Original data

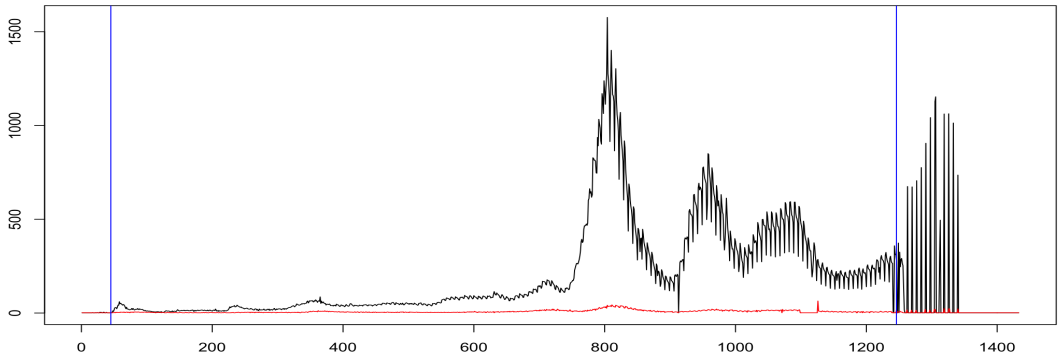


(b) $-\log(p)$

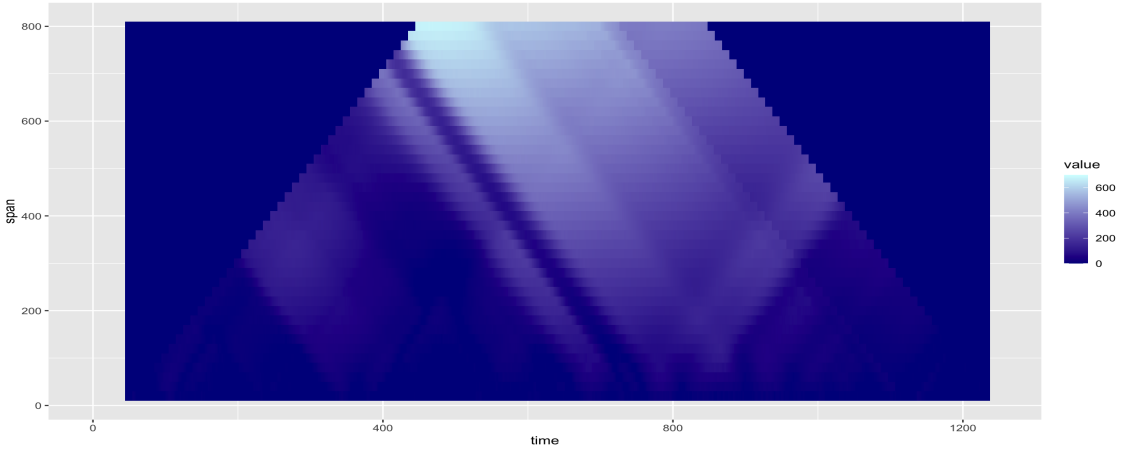


(c) coefficients

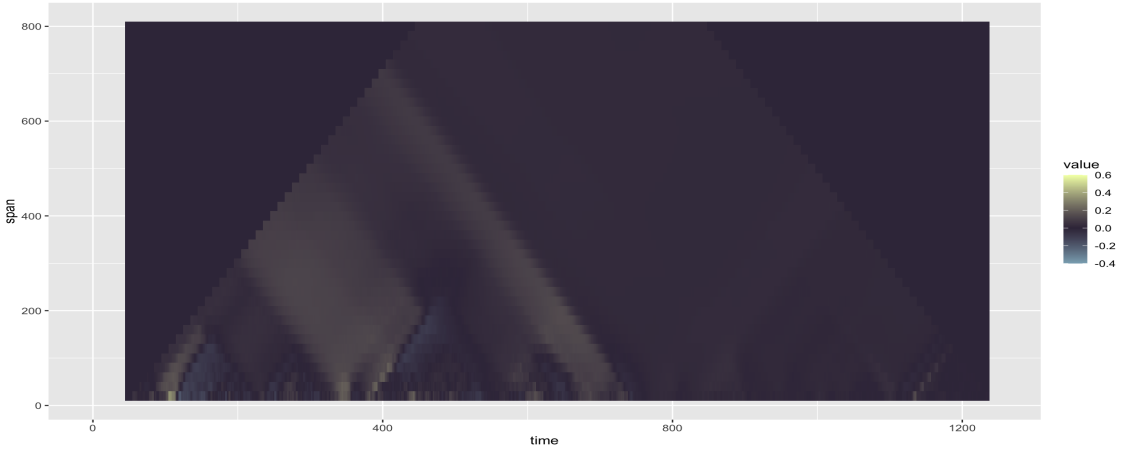
Figure 5.32: Combination of original data (a) and heatmaps [(b) and (c)] for new cases (black) and new deaths in South Africa recorded from 2020-01-03 to 2023-12-06. The chosen window sizes are $20, 40, \ldots, 800$. The blue vertical lines in (a) bound the examined area.



(a) Original data

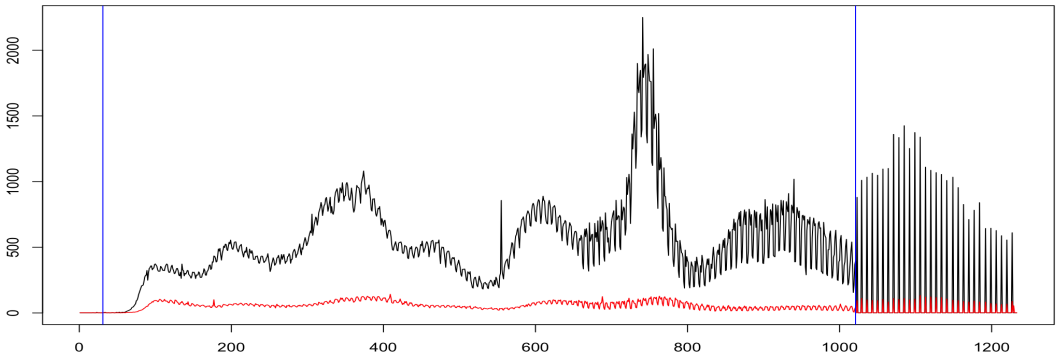


(b) $-\log(p)$

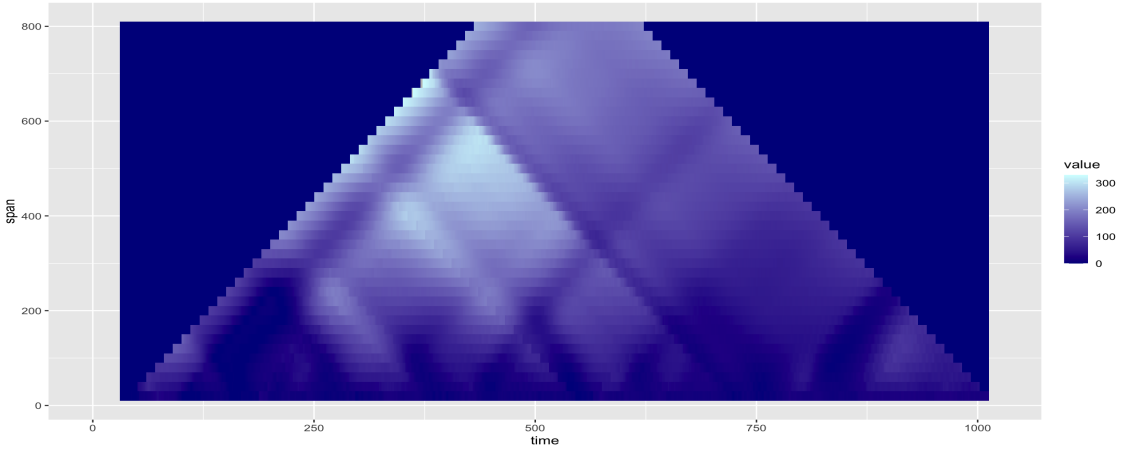


(c) coefficients

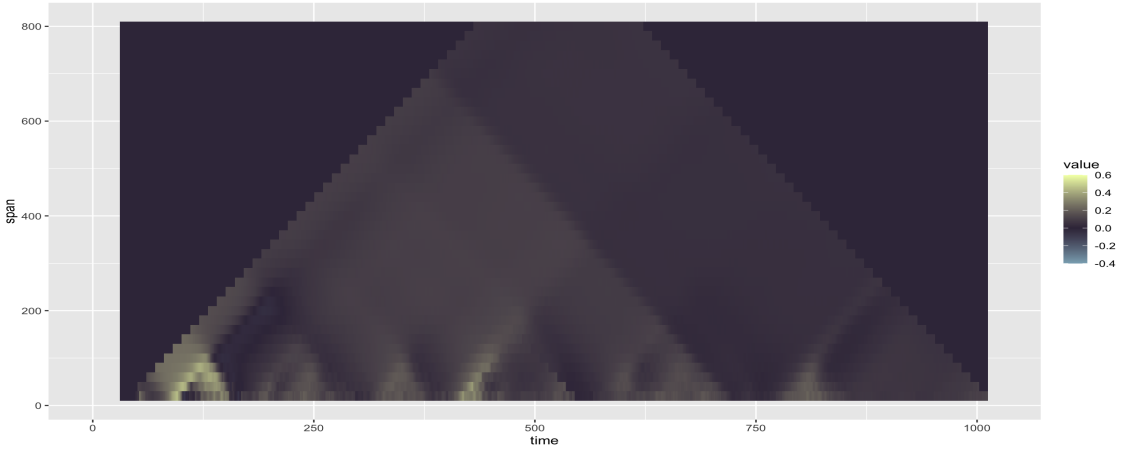
Figure 5.33: Combination of original data (a) and heatmaps [(b) and (c)] for new cases (black) and new deaths in South Korea recorded from 2020-01-03 to 2023-12-06. The chosen window sizes are $20, 40, \ldots, 800$. The blue vertical lines in (a) bound the examined area.



(a) Original data



(b) $-\log(p)$



(c) coefficients

Figure 5.34: Combination of original data (a) and heatmaps [(b) and (c)] for new cases (black) and new deaths in USA recorded from 2020-01-03 to 2023-05-20. The chosen window sizes are $20, 40, \ldots, 800$. The blue vertical lines in (a) bound the examined area.

5.5 Visualisation Examples

In this section, we introduce more simulated examples for testing the effectiveness of MLLH on the analysis of lead-lag relationships between bi-variate data. In particular, as supplements to the existing examples, we consider periodic functions with higher frequency.

(M12) $X_t = \sin(\pi t/100 - \pi/2) + 1$ for $t = 1, 2, \ldots, 825$ and $Y_t = \sin(\pi t/100 - \pi) + 1$ for $t = 26, 27, \ldots, 825$.

(M13) $X_t = \sin(\pi t/100 - \pi/2) + 1$ for $t = 1, 2, \ldots, 850$ and $Y_t = \sin(\pi t/100 - \pi) + 1$ for $t = 51, 52, \ldots, 850$.

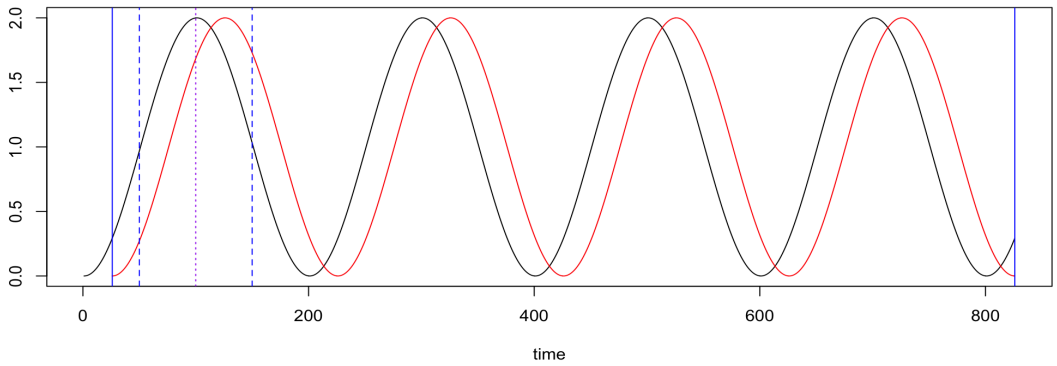
(M14) $X_t = \sin(\pi t/100 - \pi/2) + 1$ for $t = 1, 2, \ldots, 875$ and $Y_t = \sin(\pi t/100 - \pi) + 1$ for $t = 76, 77, \ldots, 875$.

(M15) $X_t = \sin(\pi t/100 - \pi/2) + 1$ for $t = 1, 2, \ldots, 900$ and $Y_t = \sin(\pi t/100 - 3\pi/2) + 1$ for $t = 101, 52, \ldots, 900$.

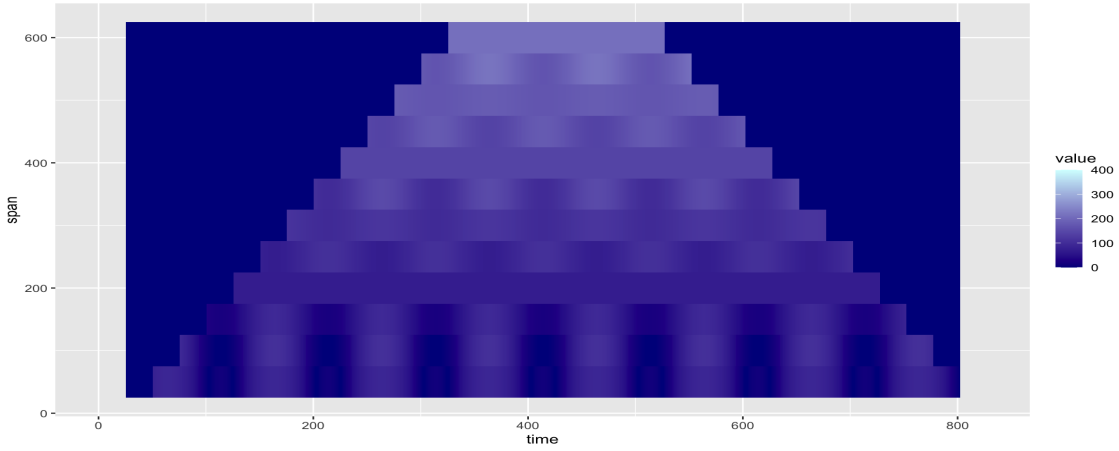
(M16) $X_t = \sin(\pi t/100 - \pi/2) + 1$ for $t = 1, 2, \ldots, 925$ and $Y_t = \sin(\pi t/100 - \pi) + 1$ for $t = 126, 127, \ldots, 925$.

(M17) $X_t = \sin(\pi t/100 - \pi/2) + 1$ for $t = 1, 2, \ldots, 950$ and $Y_t = \sin(\pi t/100 - 3\pi/2) + 1$ at $t = 151, 152, \ldots, 950$

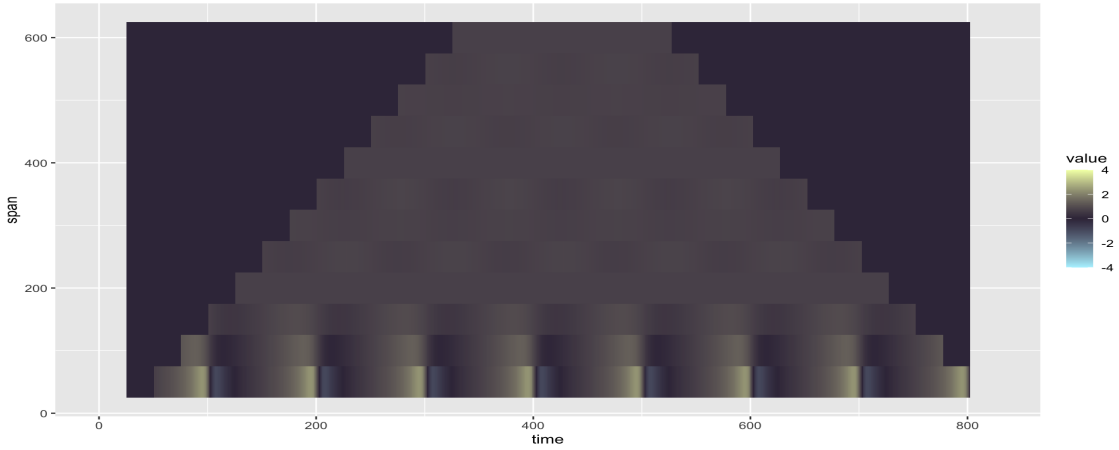
(M18) $X_t = \sin(\pi t/100 - \pi/2) + 1$ for $t = 1, 2, \ldots, 975$ and $Y_t = \sin(\pi t/100 - 3\pi/2) + 1$ for $t = 176, 177, \ldots, 975$.



(a) Original data

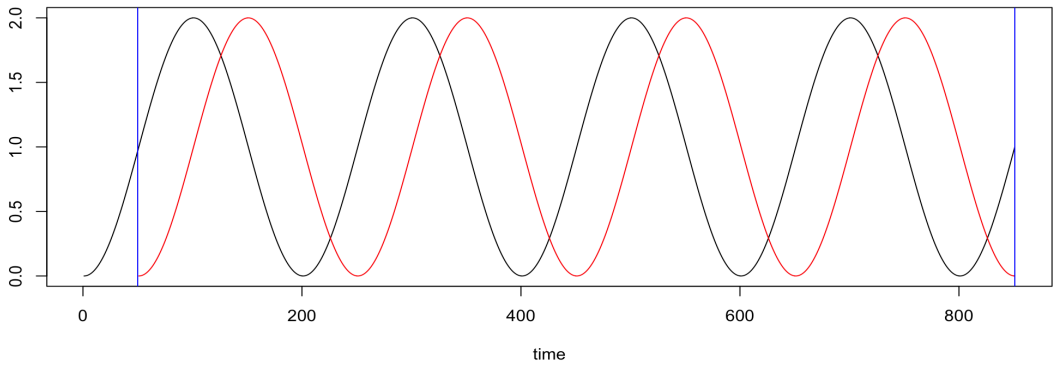


(b) $-\log(p)$

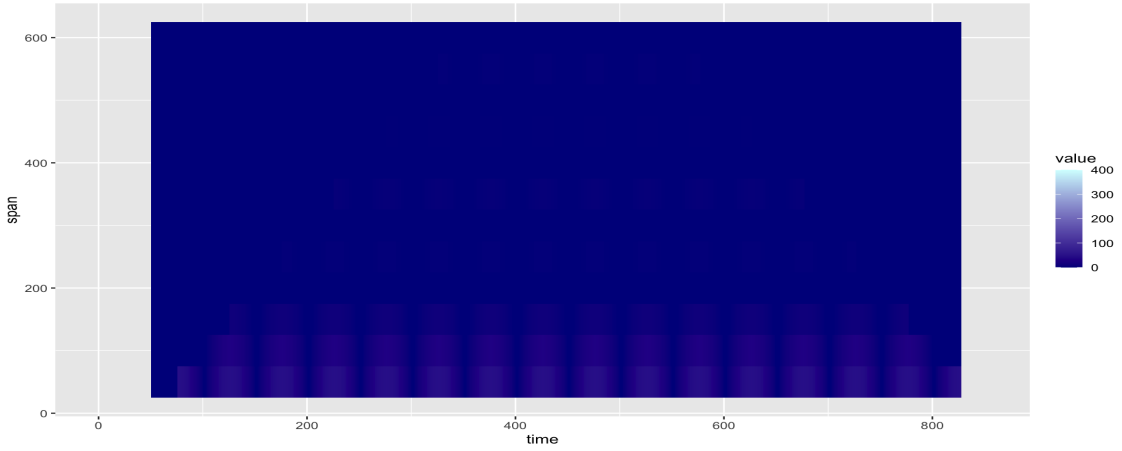


(c) coefficients

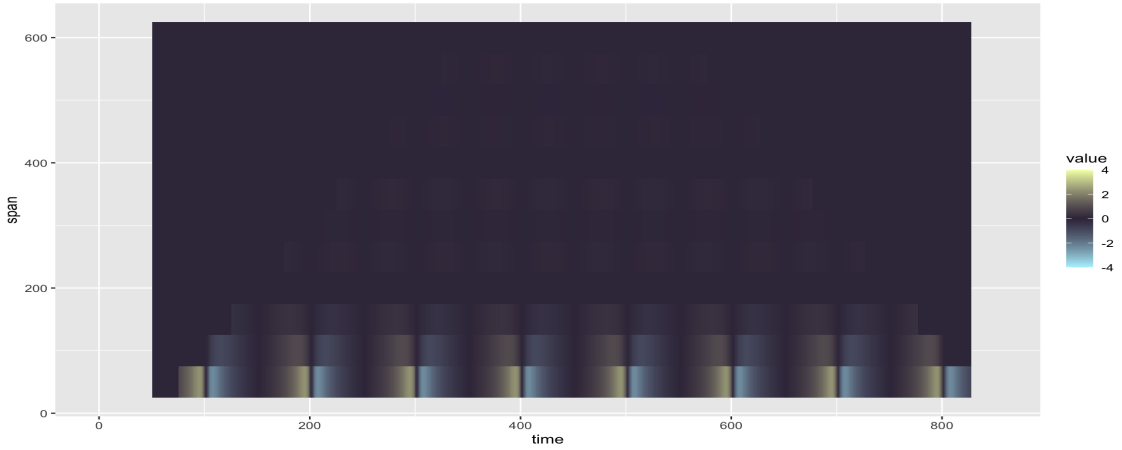
Figure 5.35: Combination of simulated signal (a) and heatmaps [(b) and (c)] for dataset (M12). The blue vertical lines bound the examined area. As an example, the blue dashed line and purple dotted lines stand for the location and the corresponding boundaries of the symmetric moving window.



(a) Original data

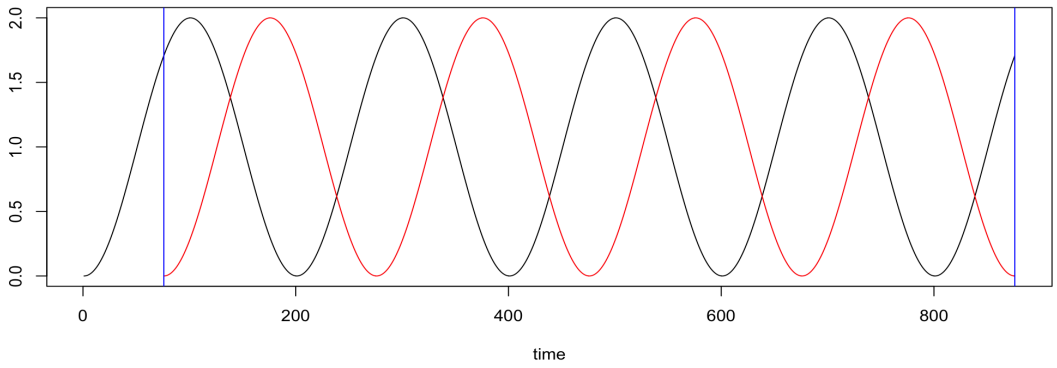


(b) $-\log(p)$

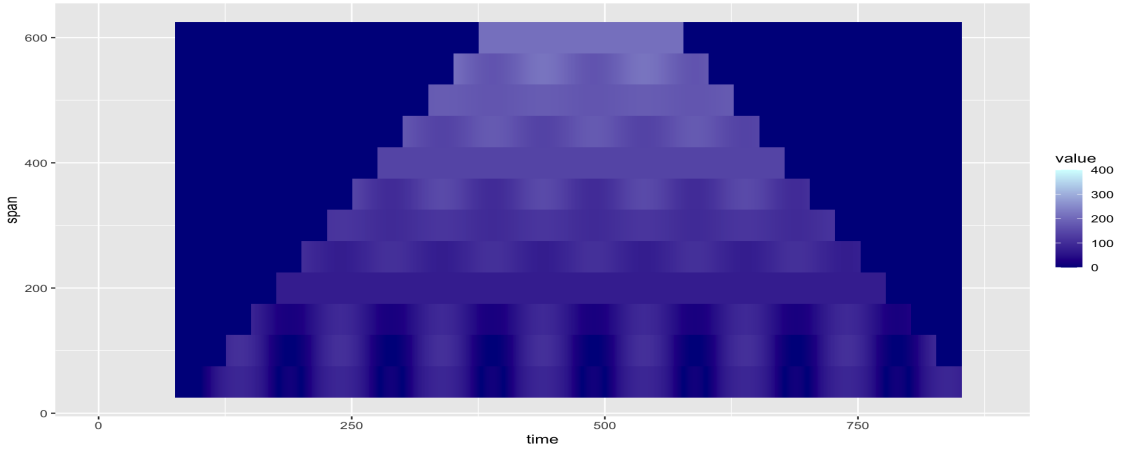


(c) coefficients

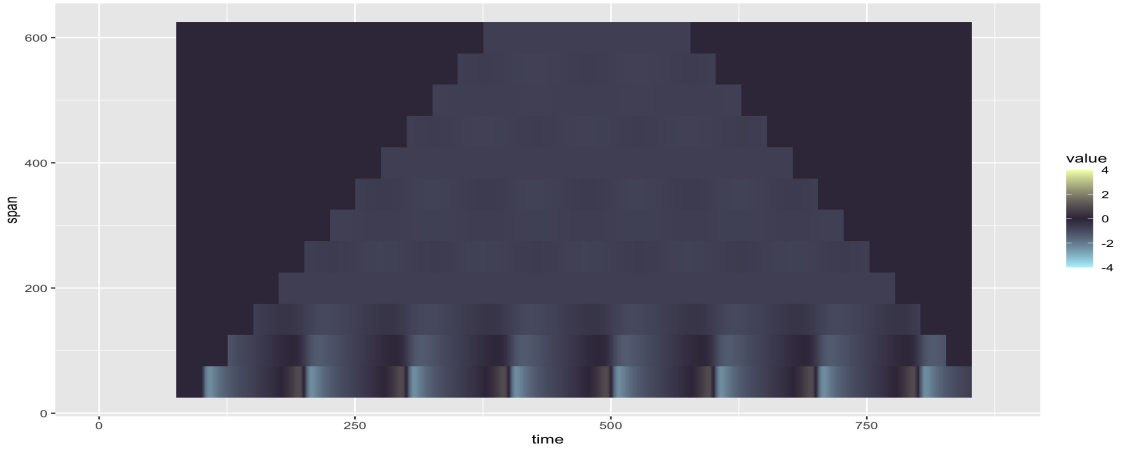
Figure 5.36: Combination of simulated signal (a) and heatmaps [(b) and (c)] for dataset (M13).



(a) Original data

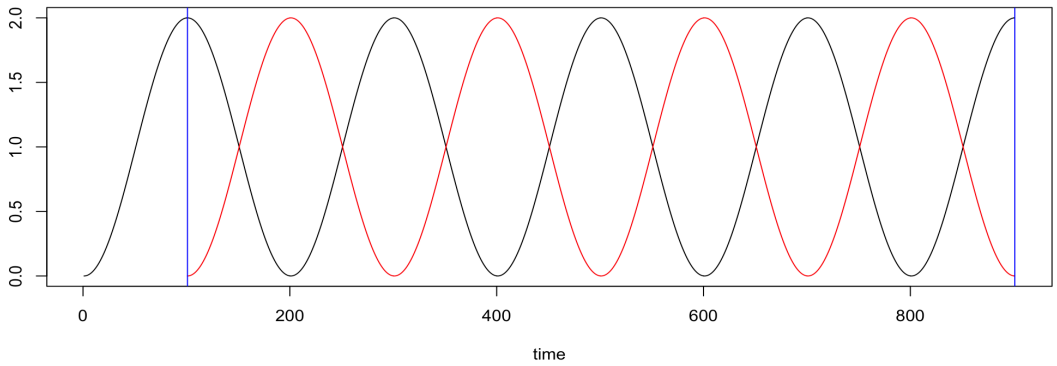


(b) $-\log(p)$

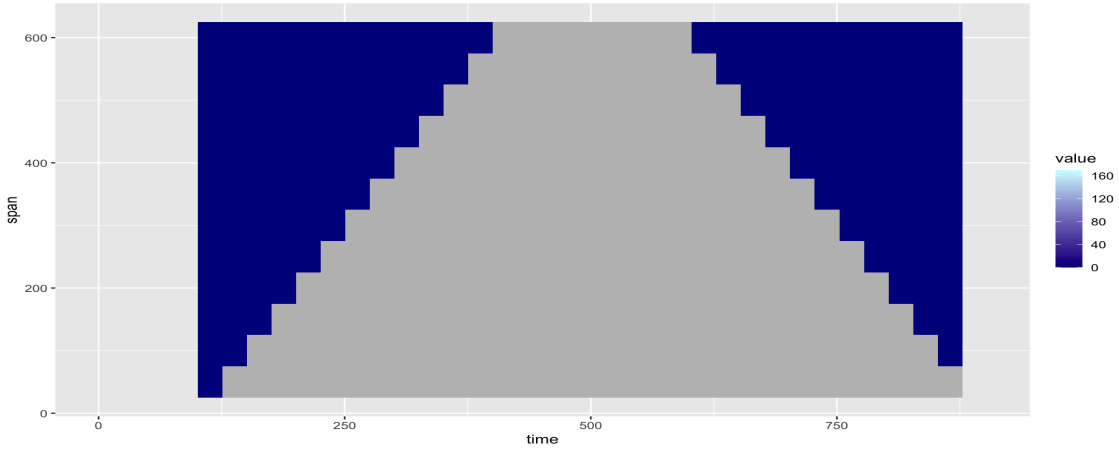


(c) coefficients

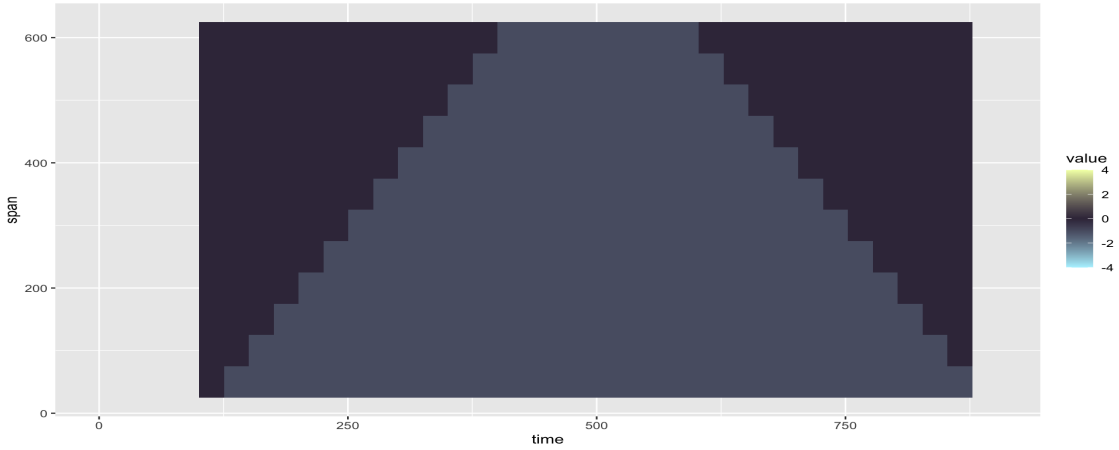
Figure 5.37: Combination of simulated data (a) and heatmaps [(b) and (c)] for dataset (M14).



(a) Original data

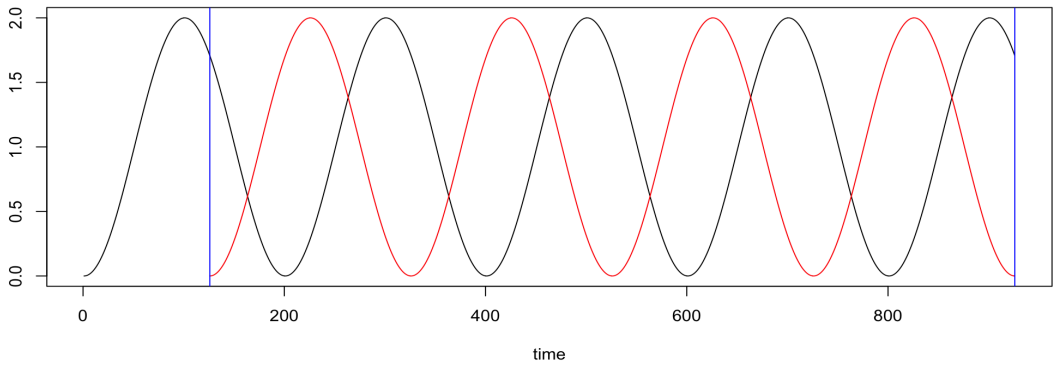


(b) $-\log(p)$

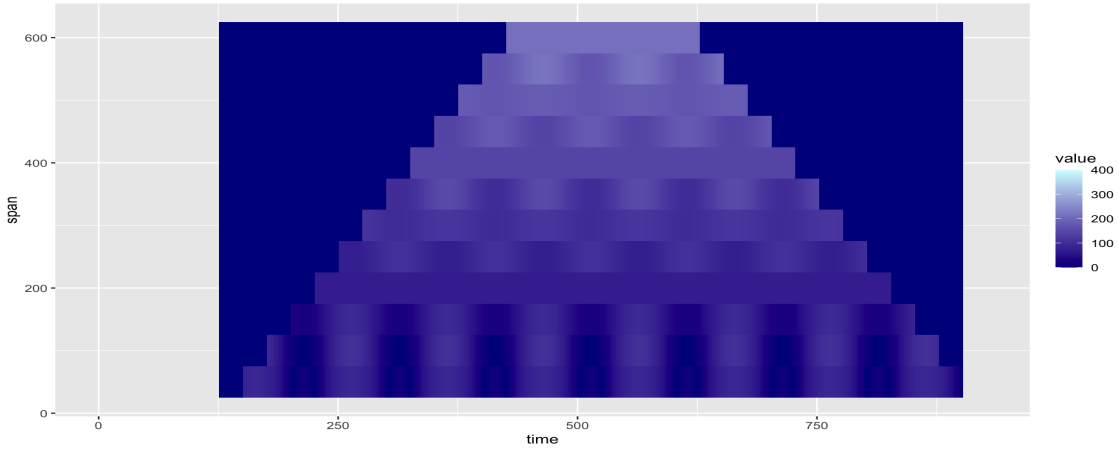


(c) coefficients

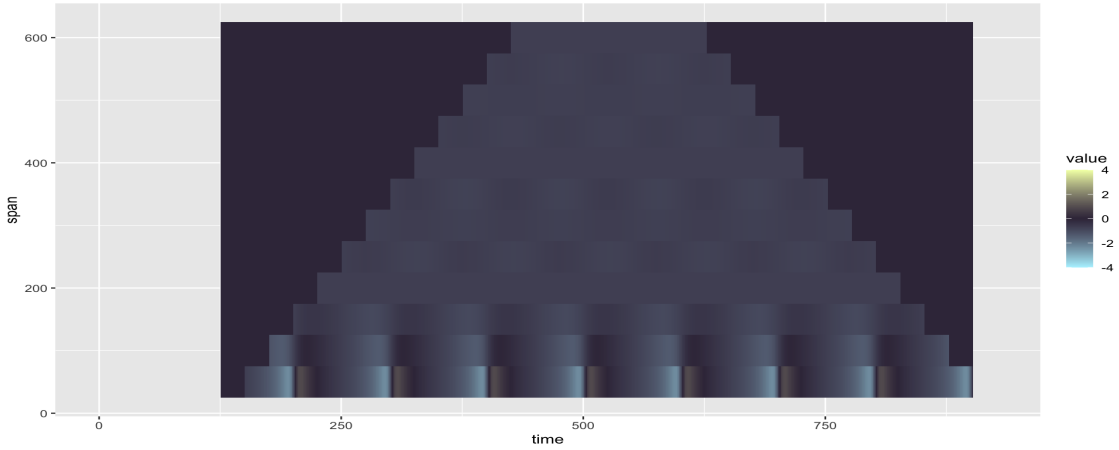
Figure 5.38: Combination of simulated data (a) and heatmaps [(b) and (c)] for dataset (M15).



(a) Original data

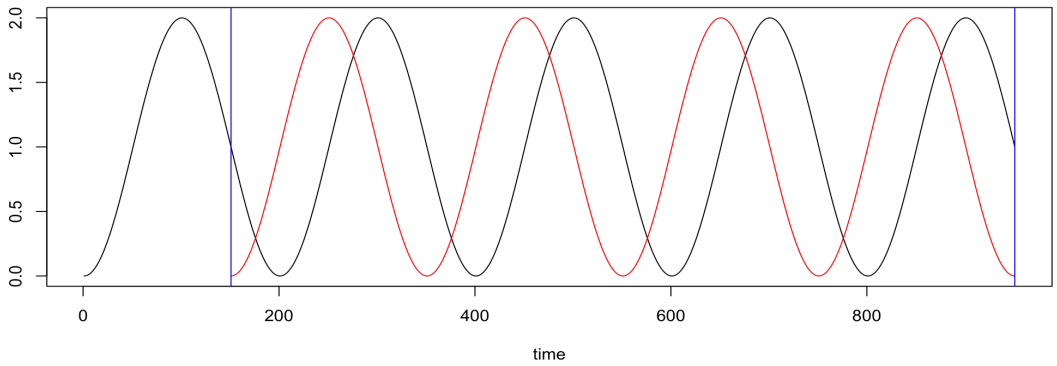


(b) $-\log(p)$

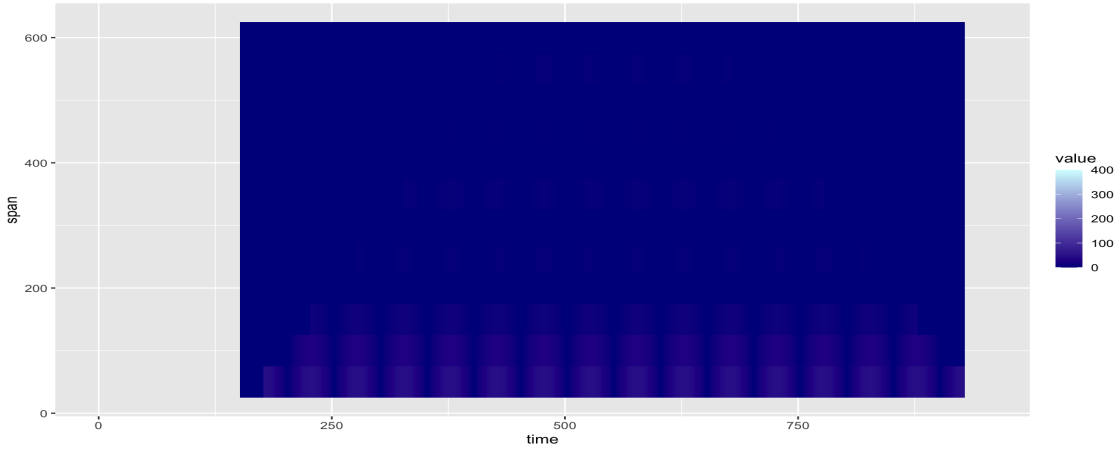


(c) coefficients

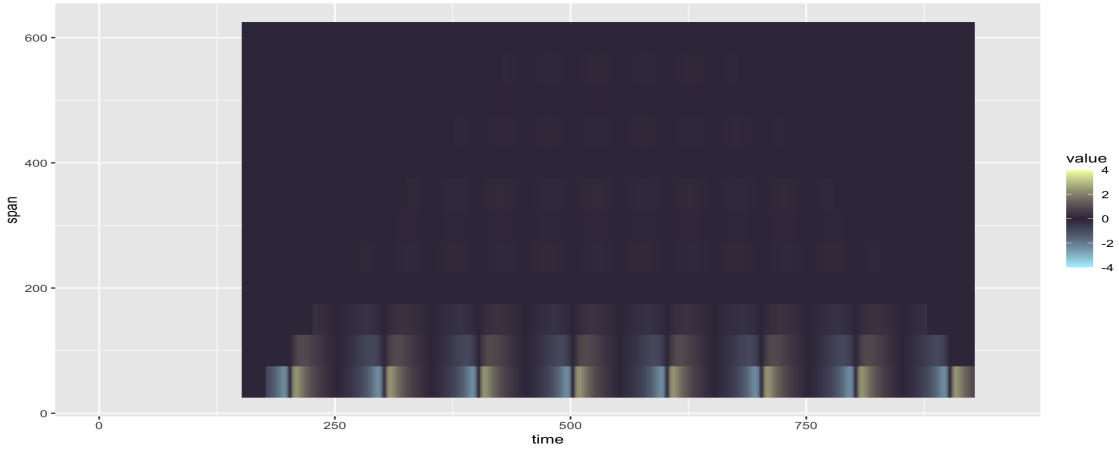
Figure 5.39: Combination of simulated data (a) and heatmaps [(b) and (c)] for dataset (M16).



(a) Original data

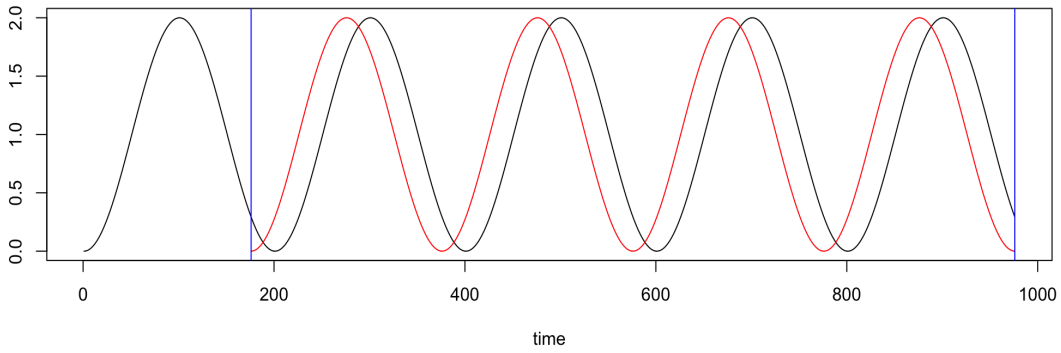


(b) $-\log(p)$

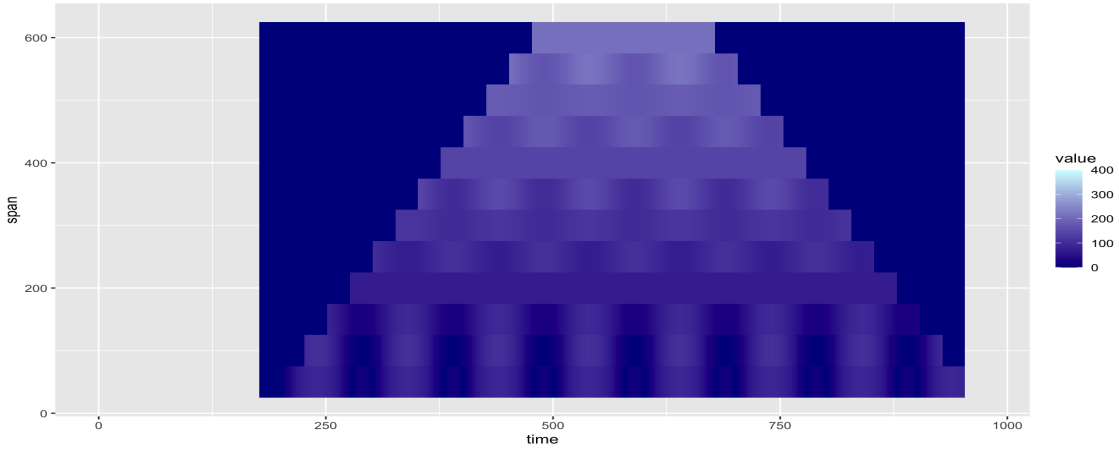


(c) coefficients

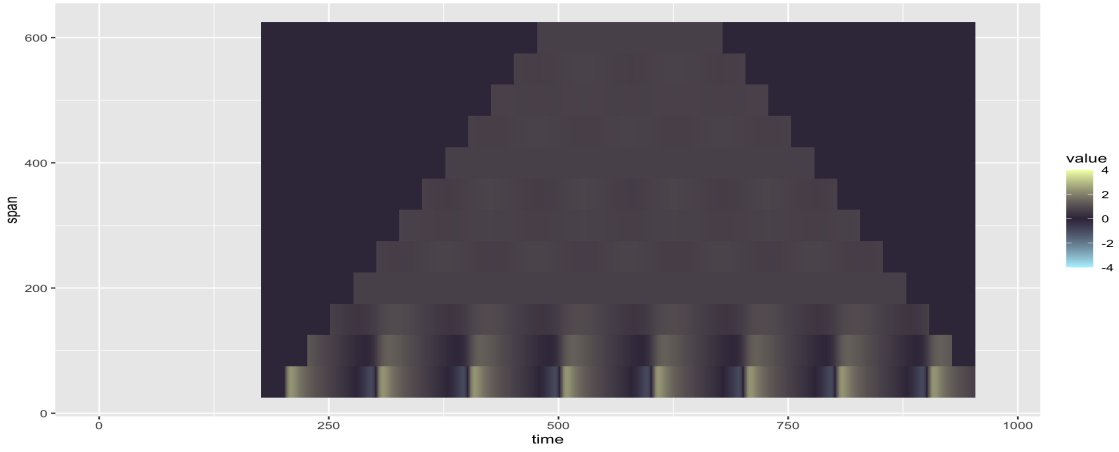
Figure 5.40: Combination of simulated data (a) and heatmaps [(b) and (c)] for dataset (M17).



(a) Original data



(b) $-\log(p)$



(c) coefficients

Figure 5.41: Combination of simulated data (a) and heatmaps [(b) and (c)] for dataset (M18).

5.6 Discussion

In this section, we discuss some closely relevant ideas of multi-scale lead-lag heatmaps for future research. First, considering the choice of time lags, the real-world data can be largely contaminated by the (possibly dependent) noise and hence makes it hard to figure out the time lags from the finest scales. Although we may still receive some information, such as the time length of the shifted colour in Figure 5.17(b), from test results obtained within small moving windows, a data preprocessing procedure might be helpful for getting more apparent features in heatmaps. Two possible methods for reducing the dependence in data are introduced in Section 4.3.

After adding the selected time lags $k \geqslant 1$, MLLH can then be extended for detecting the direction of the lead-lag relationships between bi-variate time series by switching X_t and Y_t, and even for analysing more complex relationships, i.e. multiple testing. It is of interest to start from the idea of the basic Granger causality test, i.e. conduct a multiple linear regression between dependent variable Y_t and regressors X_{t-k}, $\{Y_{t-1}, Y_{t-2}, \dots\}$. Seasonality indicators can also be included in regressors to assess the significance of such patterns in the real-world data. Then, with the new MLLH algorithm, we may be able to provide theoretical verification of the "best" window size for data analysis.

Chapter 6

Conclusions

In this thesis, we consider the problem of detecting multiple change-points of the dependent time series with abrupt mean shifts, and we investigate the problem of analysing lead-lag relationships between bi-variate nonstationary time series. This chapter provides a brief summary of the contributions presented in Chapter 3, 4 and 5 and offers a discussion on some possible directions for future studies.

Chapter 3 introduces new wavelet-based consistent LRV estimators that can help quantify the level of noise in processes with piecewise-constant signals and errors following the physical system proposed in Wu (2005). After applying a wavelet shrinkage idea to DWT- or MODWT-based Haar wavelets, our proposed estimators lie somewhere in between the two broad classes of LRV estimators: residual- and difference-based estimators. These robust estimators bypass the challenging task of pre-estimating the signals and will not be largely impacted by potential outliers that can possibly result in the poor performance of difference-based estimators. We employ the Theorem 2 in Wu and Wu (2016) to show the asymptotic unbiasedness and consistency of our estimators. The simulation results obtained from data with ARMA error processes

indicate that our MODWT-based approaches can generally outperform many existing LRV estimators. In recent years, more general change-point detection problems for data with higher-order polynomial trends, especially piecewise-linear segmentation, have attracted much interest in the literature. Therefore, one major possibility for future research is to extend the current LRV estimators to dependent data with linear trends.

In Chapter 4, we discuss the possible extensions of the threshold-based NOT algorithm and the NOT solution path algorithm for serially dependent data. First, we attempt to extend the former basic algorithm by constructing the threshold proportional to our new LRV estimators, which fail to show much efficacy. Secondly, we turn to focus on the NOT solution path algorithm and provide detailed analyses on the two data-preprocessing methods introduced in Baranowski et al. (2019). The test results indicate that adding an independent error process following $\varepsilon_t \overset{iid}{\sim} N(0, \sigma^2)$ can successfully enhance the performance of NOT but the choice of σ can indeed be a challenging issue. Additionally, although we can see the effectiveness of pre-averaging the sequence over non-overlapping moving windows, this approach faces three important issues that may arise in practice, i.e. the choice of window size, the detection of the two consecutive change-points resulting from pre-averaging, and the estimation of the true change-point within the detected moving window. Lastly, we study the modification of the strengthened Schwarz Information Criterion applied in the NOT solution path algorithm. The simulated results show that the new NOT algorithms with adjusted measures of fit cannot outperform the original NOT solution path algorithm if we choose an optimal α for the penalty function. Hence, we propose an empirical formula for α to make the original algorithm useful for dependent data, and the simulation studies demonstrate the practicability of this formula for data with various mean shifts and ARMA error processes. For future research, the first possible direction is to provide a theoretical formula for σ or α after figuring out the explicit conditions of dependent data where the NOT solution path algorithm breaks down regardless of the choice of σ or α. In

addition, we shall continue to consider the possible measures of fit, together with the corresponding penalty function, after determining the reasons leading to the failure of the two introduced measures $\mathrm{IC}^{LRV}(k)$ and $\mathrm{IC}^{ACV}(k)$.

In Chapter 5, we propose an exploratory approach, the Multi-scale Lead-Lag Heatmap, to investigate the lead-lag relationships between bi-variate nonstationary time series with natural directions. This approach is constructed based on the "scale-space" viewpoint applied in the SiZer map and hence can provide an broad view of relations between two time series, which is likely to serve as the first step for further lead-lag or causal analyses. After examining examples with similar or changing relationships, we present the information highlighted by the heatmaps of coefficients and the corresponding $-\log(p)$ from both the "local" and "global" perspectives, i.e. from the basic understanding of the overall dependence to the changing lead-lag relations and the (possibly) "best" window size for bi-variate time series. Then, we assess the practicability of MLLH on real-world COVID-19 data examples for many countries. As mentioned in Section 5.6, to make the heatmaps more informative, we can first decide the choice of time lags after conducting data preprocessing. The following possible extension comes from considering the direction of the lead-lag relationships and multiple testing for dependent data.

Bibliography

Abadir, K. M., W. Distaso, and L. Giraitis (2009). Two estimators of the long-run variance: beyond short memory. *Journal of Econometrics 150*(1), 56–70.

Abhyankar, A. N. (1995). Return and volatility dynamics in the FT–SE 100 stock index and stock index futures markets. *The Journal of Futures Markets (1986-1998) 15*(4), 457.

Alexopoulos, C. and D. Goldsman (2004). To batch or not to batch? *ACM Transactions on Modeling and Computer Simulation (TOMACS) 14*(1), 76–114.

Alexopoulos, C., D. Goldsman, and J. R. Wilson (2011). Overlapping batch means: something more for nothing? In *Proceedings of the 2011 Winter Simulation Conference (WSC)*, pp. 401–411. IEEE.

Alyousifi, Y., K. Ibrahim, W. Zin, and U. Rathnayake (2022). Trend analysis and change point detection of air pollution index in malaysia. *International Journal of Environmental Science and Technology 19*, 7679–7700.

Aminikhanghahi, S. and D. J. Cook (2017). A survey of methods for time series change point detection. *Knowledge and information systems 51*(2), 339–367.

Anscombe, F. J. (1948). The transformation of poisson, binomial and negative-binomial data. *Biometrika 35*(3/4), 246–254.

Aue, A. and L. Horváth (2013). Structural breaks in time series. *Journal of Time Series Analysis 34*(1), 1–16.

Bai, L. and W. Wu (2024). Difference-based covariance matrix estimate in time series nonparametric regression with applications to specification tests. *arXiv preprint arXiv:2303.16599*.

Baranowski, R., Y. Chen, and P. Fryzlewicz (2019). 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.

Berkes, I., L. Horváth, P. Kokoszka, and Q.-M. Shao (2006). On discriminating between long-range dependence and changes in mean. *The Annals of Statistics 34*(3), 1140–1165.

Betken, A. (2016). Testing for change-points in long-range dependent time series by means of a self-normalized wilcoxon test. *Journal of Time Series Analysis 37*(6), 785–809.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics 31*(3), 307–327.

Bollerslev, T. (1987). A conditionally heteroskedastic time series model for speculative prices and rates of return. *The review of economics and statistics 69*(3), 542–547.

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

Boysen, L., A. Kempe, V. Liebscher, A. Munk, and O. Wittich (2009). Consistencies and rates of convergence of jump-penalized least squares estimators. *The Annals of Statistics 37*, 157–183.

Brockwell, P. J. and R. A. Davis (2009). *Time series: theory and methods*. Springer science & business media.

Brown, L. D. and M. Levine (2007). Variance estimation in nonparametric regression via the difference sequence method. *The Annals of Statistics 35*(5), 2219–2232.

Carlstein, E. et al. (1986). The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *The annals of statistics 14*(3), 1171–1179.

Carlstein, E., K.-A. Do, P. Hall, T. Hesterberg, and H. R. Künsch (1998). Matched-block bootstrap for dependent data. *Bernoulli 4*(3), 305–328.

Catoni, O. (2012). Challenging the empirical mean and empirical variance: a deviation study. *48*(4), 1148–1185.

Chakar, S., E. Lebarbier, C. Lévy-Leduc, and S. Robin (2017). A robust approach for estimating change-points in the mean of an AR(1) process. *Bernoulli 23*(2), 1408–1447.

Chan, K. W. (2022). Optimal difference-based variance estimators in time series: A general framework. *The Annals of Statistics 50*(3), 1376–1400.

Chan, K. W. and C. Y. Yau (2017). High-order corrected estimator of asymptotic variance with optimal bandwidth. *Scandinavian Journal of Statistics 44*(4), 866–898.

Chapman, J.-L., I. Eckley, and R. Killick (2020). A nonparametric approach to detecting changes in variance in locally stationary time series. *Environmetrics 31*(1), e2576.

Chaudhuri, P. and J. S. Marron (1999). Sizer for exploration of structures in curves. *Journal of the American Statistical Association 94*(447), 807–823.

Chen, G., G. Lu, W. Shang, and Z. Xie (2019). Automated change-point detection of eeg signals based on structural time-series analysis. *IEEE Access 7*, 180168–180180.

Cho, H. and P. Fryzlewicz (2023). Multiple change point detection under serial dependence: Wild contrast maximisation and gappy schwarz algorithm. *arXiv preprint arXiv:2011.13884*.

Cho, H. and C. Kirch (2022). Two-stage data segmentation permitting multiscale change points, heavy tails and dependence. *Annals of the Institute of Statistical Mathematics 74*, 653–684.

Chung, P. J. and D. J. Liu (1994). Common stochastic trends in pacific rim stock markets. *The Quarterly Review of Economics and Finance 34*(3), 241–259.

Clauset, A., C. R. Shalizi, and M. E. Newman (2009). Power-law distributions in empirical data. *SIAM review 51*(4), 661–703.

Corhay, A., A. T. Rad, and J.-P. Urbain (1993). Common stochastic trends in european stock markets. *Economics Letters 42*(4), 385–390.

Dahlhaus, R. (1997). Fitting time series models to nonstationary processes. *The Annals of Statistics 25*(1), 1–37.

Dahlhaus, R. and M. Eichler (2003). Causality and graphical models in time series analysis. *Oxford Statistical Science Series*, 115–137.

Daly, D., W. Brown, H. Ingo, J. O'Leary, and D. Bradford (2020). The use of change point detection to identify software performance regressions in a continuous integration system. In *Proceedings of the ACM/SPEC International Conference on Performance Engineering*, pp. 67–75.

Damerdji, H. (1994). Strong consistency of the variance estimator in steady-state simulation output analysis. *Mathematics of Operations Research 19*(2), 494–512.

Davis, R. A., T. C. M. Lee, and G. A. Rodriguez-Yam (2006). Structural break estimation for nonstationary time series models. *Journal of the American Statistical Association 101*(473), 223–239.

Dawid, A. P. (2000). Causal inference without counterfactuals. *Journal of the American statistical Association 95*(450), 407–424.

Dette, H., T. Eckle, and M. Vetter (2020). Multiscale change point detection for dependent data. *Scandinavian Journal of Statistics 47*(4), 1243–1274.

Dette, H., A. Munk, and T. Wagner (1998). Estimating the variance in nonparametric regression—what is a reasonable choice? *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 60*(4), 751–764.

Dette, H. and W. Wu (2022). Prediction in locally stationary time series. *Journal of Business & Economic Statistics 40*(1), 370–381.

Dette, H., W. Wu, et al. (2019). Detecting relevant changes in the mean of nonstationary processes—a mass excess approach. *The Annals of Statistics 47*(6), 3578–3608.

Donoho, D. L. (1995). De-noising by soft-thresholding. *IEEE transactions on information theory 41*(3), 613–627.

Donoho, D. L. and I. M. Johnstone (1994). Ideal spatial adaptation by wavelet shrinkage. *biometrika 81*(3), 425–455.

Donoho, D. L. and I. M. Johnstone (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the american statistical association 90*(432), 1200–1224.

Donoho, D. L., I. M. Johnstone, G. Kerkyacharian, and D. Picard (1995). Wavelet shrinkage: asymptopia? *Journal of the Royal Statistical Society: Series B (Methodological) 57*(2), 301–337.

Doukhan, P. (2012). *Mixing: properties and examples*, Volume 85. Springer Science & Business Media.

Doukhan, P., G. Oppenheim, and M. Taqqu (2002). *Theory and applications of long-range dependence*. Springer Science & Business Media.

Du, C., C.-L. M. Kao, and S. Kou (2016). Stepwise signal extraction via marginal likelihood. *Journal of the American Statistical Association 111*(513), 314–330.

Eichinger, B. and C. Kirch (2018). A mosum procedure for the estimation of multiple random change points. *Bernoulli 24*(1), 526–564.

Eichler, M. (2012). Causal inference in time series analysis. *Causality: Statistical perspectives and applications*, 327–354.

Eichler, M. (2013). Causal inference with multiple time series: principles and problems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 371*(1997), 20110613.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica 50*(4), 987–1007.

Fang, X. and D. Siegmund (2020). Detection and estimation of local signals. *arXiv preprint arXiv:2004.08159*.

Fowler, J. E. (2005). The redundant discrete wavelet transform and additive noise. *IEEE Signal Processing Letters 12*(9), 629–632.

Frick, K., A. Munk, and H. Sieling (2014). Multiscale change point inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76*(3), 495–580.

Fryzlewicz, P. (2007). Unbalanced haar technique for nonparametric function estimation. *Journal of the American Statistical Association 102*(480), 1318–1327.

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

Fryzlewicz, P. (2018). Tail-greedy bottom-up data decompositions and fast multiple change-point detection. *The Annals of Statistics 46*(6B), 3390–3421.

Fryzlewicz, P. (2020). Detecting possibly frequent change-points: Wild binary segmentation 2 and steepest-drop model selection. *Journal of the Korean Statistical Society 49*, 1027–1070.

Fryzlewicz, P. (2023). Narrowest significance pursuit: inference for multiple change-points in linear models. *Journal of the American Statistical Association*, 1–14.

Fryzlewicz, P. Z. (2003). *Wavelet techniques for time series and Poisson data*. Ph. D. thesis, University of Bristol.

Gerstenberger, C. (2021). Robust discrimination between long-range dependence and a change in mean. *Journal of Time Series Analysis 42*(1), 34–62.

Getahun, Y. S., M.-H. Li, and I.-F. Pun (2021). Trend and change-point detection analyses of rainfall and temperature over the awash river basin of ethiopia. *Heliyon 7*(9), e08024.

Godtliebsen, F., J. Marron, and P. Chaudhuri (2002). Significance in scale space for bivariate density estimation. *Journal of Computational and Graphical Statistics 11*(1), 1–21.

Gong, C. C., S. D. Ji, L. L. Su, S. P. Li, and F. Ren (2016). The lead–lag relationship between stock index and stock index futures: A thermal optimal path method. *Physica A: Statistical Mechanics and its Applications 444*, 63–72.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society 37*(3), 424–438.

Granger, C. W. (1980). Testing for causality: A personal viewpoint. *Journal of Economic Dynamics and control 2*, 329–352.

Granger, C. W. (1988). Some recent development in a concept of causality. *Journal of econometrics 39*(1-2), 199–211.

Guédon, Y. (2013). Exploring the latent segmentation space for the assessment of multiple change-point models. *Computational Statistics 28*(6), 2641–2678.

Gupta, K. and N. Chatterjee (2020). Examining lead-lag relationships in-depth, with focus on fx market as covid-19 crises unfolds. *arXiv preprint arXiv:2004.10560*.

Haar, A. (1910). Zur theorie der orthogonalen funktionensysteme. *Mathematische Annalen 69*, 331–371.

Habibi, R. (2021). Bayesian online change point detection in finance. *Financial Internet Quarterly 17*(4), 27–33.

Hall, P., J. Kay, and D. Titterinton (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika 77*(3), 521–528.

Hall, P. and I. V. Keilegom (2003). Using difference-based methods for inference in nonparametric regression with time series errors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 65*(2), 443–456.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the american statistical association 69*(346), 383–393.

Hasbrouck, J. (2003). Intraday price formation in us equity index markets. *The Journal of Finance 58*(6), 2375–2400.

He, Y., K. A. Burghardt, and K. Lerman (2022). Leveraging change point detection to discover natural experiments in data. *EPJ Data Science 11*(1), 49.

Hoover, K. D. (2008). Causality in economics and econometrics. *The New Palgrave Dictionary of Economics*, 1–13.

Hsiao, C. (1982). Autoregressive modeling and causal ordering of economic variables. *Journal of economic Dynamics and Control 4*, 243–259.

Huckemann, S., K.-r. Kim, A. Munk, F. Rehfeldt, M. Sommerfeld, J. Weickert, and C. Wollnik (2016). The circular sizer, inferred persistence of shape parameters and application to early stem cell differentiation. *Bernoulli 22*(4), 2113–2142.

Hurst, H. E. (1951). Long-term storage capacity of reservoirs. *Transactions of the American society of civil engineers 116*(1), 770–799.

Hušková, M. and A. Slabỳ (2001). Permutation tests for multiple changes. *Kybernetika 37*(5), 605–622.

Jackson, B., J. D. Scargle, D. Barnes, S. Arabhi, A. Alt, P. Gioumousis, E. Gwin, P. Sangtrakulcharoen, L. Tan, and T. T. Tsai (2005). An algorithm for optimal partitioning of data on an interval. *IEEE Signal Processing Letters 12*(2), 105–108.

Jiang, T., S. Bao, and L. Li (2019). The linear and nonlinear lead–lag relationship among three sse 50 index markets: The index futures, 50etf spot and options markets. *Physica A: Statistical Mechanics and Its Applications 525*, 878–893.

Johnstone, I. M. and B. W. Silverman (1997). Wavelet threshold estimators for data with correlated noise. *Journal of the royal statistical society: series B (statistical methodology) 59*(2), 319–351.

Kamalabad, M. S., R. Leenders, and J. Mulder (2023). What is the point of change? change point detection in relational event models. *Social Networks 74*, 166–181.

Kanas, A. and G. P. Kouretas (2005). A cointegration approach to the lead–lag effect among size-sorted equity portfolios. *International Review of Economics & Finance 14*(2), 181–201.

Kavussanos, M. G., I. D. Visvikis, and P. D. Alexakis (2008). The lead-lag relationship between cash and stock index futures in a new market. *European Financial Management 14*(5), 1007–1025.

Kawaller, I. G., P. D. Koch, and T. W. Koch (1987). The temporal price relationship between s&p 500 futures and the s&p 500 index. *The Journal of Finance 42*(5), 1309–1329.

Khismatullina, M. and M. Vogt (2020). Multiscale inference and long-run variance estimation in non-parametric regression with time series errors. *Journal of the Royal Statistical Society Series B: Statistical Methodology 82*(1), 5–37.

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

Kim, K., J. H. Park, M. Lee, and J. W. Song (2022). Unsupervised change point detection and trend prediction for financial time-series using a new cusum-based approach. *IEEE Access 10*, 34690–34705.

Kim, S. and F. In (2005). The relationship between stock returns and inflation: new evidence from wavelet analysis. *Journal of empirical finance 12*(3), 435–444.

Knight, M. I., M. A. Nunes, and G. P. Nason (2012). Spectral estimation for locally stationary time series with missing observations. *Statistics and Computing 22*, 877–895.

Kunsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics 17*(3), 1217–1241.

Lang, M., H. Guo, J. E. Odegard, C. S. Burrus, and R. O. Wells Jr (1995). Nonlinear processing of a shift-invariant discrete wavelet transform (dwt) for noise reduction. In *Wavelet Applications II*, Volume 2491, pp. 640–651. SPIE.

Lavielle, M. and E. Moulines (2000). Least-squares estimation of an unknown number of shifts in a time series. *Journal of time series analysis 21*(1), 33–59.

Lee, C.-B. (1995). Estimating the number of change points in a sequence of independent normal random variables. *Statistics & probability letters 25*(3), 241–248.

Lee, C. C. and P. C. Phillips (1994). An ARMA prewhitened long-run variance estimator. *Manuscript, Yale University*.

Li, R. and J. Marron (2005). Local likelihood sizer map. *Sankhyā: The Indian Journal of Statistics 67*(3), 476–498.

Li, Y., T. Wang, B. Sun, and C. Liu (2022). Detecting the lead–lag effect in stock markets: definition, patterns, and investment strategies. *Financial Innovation 8*(1), 51.

Lien, D., Y. K. Tse, and X. Zhang (2003). Structural change and lead-lag relationship between the nikkei spot index and futures price: A genetic programming approach. *Quantitative Finance 3*(2), 136.

Liu, S., A. Wright, and M. Hauskrecht (2018). Change-point detection method for clinical decision support system rule monitoring. *Artificial intelligence in medicine 91*, 49–56.

Ma, C., R. Xiao, and X. Mi (2022). Measuring the dynamic lead–lag relationship between the cash market and stock index futures market. *Finance Research Letters 47*, 102940.

Malladi, R., G. P. Kalamangalam, and B. Aazhang (2013). Online bayesian change point detection algorithms for segmentation of epileptic activity. In *2013 Asilomar conference on signals, systems and computers*, pp. 1833–1837. IEEE.

Mallat, S. (1999). *A wavelet tour of signal processing*. Elsevier.

Mallat, S. G. (1989). Multiresolution approximations and wavelet orthonormal bases of $l_2(\mathbb{R})$. *Transactions of the American mathematical society 315*(1), 69–87.

Mandelbrot, B. B. and J. W. Van Ness (1968). Fractional brownian motions, fractional noises and applications. *SIAM review 10*(4), 422–437.

Mandelbrot, B. B. and J. R. Wallis (1968). Noah, joseph, and operational hydrology. *Water resources research 4*(5), 909–918.

Mathieu, E., H. Ritchie, L. Rodés-Guirao, C. Appel, C. Giattino, J. Hasell, B. Macdonald, S. Dattani, D. Beltekian, E. Ortiz-Ospina, and M. Roser (2020). Coronavirus pandemic (covid-19). *Our World in Data*. https://ourworldindata.org/coronavirus.

Matteson, D. S. and N. A. James (2014). A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association 109*(505), 334–345.

McGonigle, E. T. and H. Cho (2023). Robust multiscale estimation of time-average variance for time series segmentation. *Computational Statistics & Data Analysis 179*, 107648.

McGonigle, E. T., R. Killick, and M. A. Nunes (2022). Trend locally stationary wavelet processes. *Journal of Time Series Analysis 43*(6), 895–917.

Meketon, M. S. and B. Schmeiser (1984). Overlapping batch means: something for nothing? In *Proceedings of the 16th conference on Winter simulation*, pp. 226–230.

Meng, H., H.-C. Xu, W.-X. Zhou, and D. Sornette (2017). Symmetric thermal optimal path and time-dependent lead-lag relationship: novel statistical tests and application to uk and us real-estate and monetary policies. *Quantitative Finance 17*(6), 959–977.

Messer, M., M. Kirchner, J. Schiemann, J. Roeper, R. Neininger, and G. Schneider (2014). A multiple filter test for the detection of rate changes in renewal processes with varying variance. *The Annals of Applied Statistics 8*(4), 2027–2067.

Muller, H.-G. and U. Stadtmuller (1987). Estimation of heteroscedasticity in regression analysis. *The Annals of Statistics 15*(2), 610–625.

Nason, G., S. Barber, T. Downie, P. Frylewicz, A. Kovac, T. Ogden, B. Silverman, and M. G. Nason (2022). *wavethresh: Wavelets Statistics and Transforms*. R Foundation for Statistical Computing.

Nason, G. P., R. Von Sachs, and G. Kroisandt (2000). Wavelet processes and adaptive estimation of the evolutionary wavelet spectrum. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 62*(2), 271–292.

Oliveira, M., R. M. Crujeiras, and A. Rodríguez-Casal (2014). Circsizer: an exploratory tool for circular data. *Environmental and ecological statistics 21*, 143–159.

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

Ondrus, M., E. Olds, and I. Cribben (2021). Factorized binary search: change point detection in the network structure of multivariate high-dimensional time series. *arXiv preprint arXiv:2103.06347*.

Otneim, H., G. D. Berentsen, and D. Tjøstheim (2022). Local lead–lag relationships and nonlinear granger causality: An empirical analysis. *Entropy 24*(3), 378.

Pan, J. and J. Chen (2006). Application of modified information criterion to multiple change point problems. *Journal of multivariate analysis 97*(10), 2221–2241.

Paparoditis, E. and D. N. Politis (2001). Tapered block bootstrap. *Biometrika 88*(4), 1105–1119.

Park, C., T. C. Lee, and J. Hannig (2010). Multiscale exploratory analysis of regression quantiles using quantile sizer. *Journal of Computational and Graphical Statistics 19*(3), 497–513.

Paul, F. and F. Piotr (2022). Change-point detection and data segmentation chapter: detecting a single change-point. *arXiv preprint arXiv:2210.07066*.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika 82*(4), 669–688.

Pearl, J. (2009). *Causality*. Cambridge university press.

Percival, D. B. and A. T. Walden (2000). *Wavelet methods for time series analysis*, Volume 4. Cambridge university press.

Pesquet, J.-C., H. Krim, and H. Carfantan (1996). Time-invariant orthonormal wavelet representations. *IEEE transactions on signal processing 44*(8), 1964–1970.

Pešta, M. and M. Wendler (2020). Nuisance-parameter-free changepoint detection in non-stationary series. *Test 29*(2), 379–408.

Pipiras, V. and M. S. Taqqu (2017). *Long-range dependence and self-similarity*, Volume 45. Cambridge university press.

Politis, D. N. and J. P. Romano (1995). Bias-corrected nonparametric spectral estimation. *Journal of time series analysis 16*(1), 67–103.

Qiu, D. (2015). *aTSA: Alternative Time Series Analysis*. R Foundation for Statistical Computing.

Qiu, D., Q. Shao, and L. Yang (2013). Efficient inference for autoregressive coefficients in the presence of trends. *Journal of Multivariate Analysis 114*, 40–53.

Reboredo, J. C. and M. A. Rivera-Castro (2013). A wavelet decomposition approach to crude oil price and exchange rate dependence. *Economic Modelling 32*, 42–57.

Rigaill, G. (2015). 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.

Romano, G., G. Rigaill, V. Runge, and P. Fearnhead (2022). Detecting abrupt changes in the presence of local fluctuations and autocorrelated noise. *Journal of the American Statistical Association 117*(540), 2147–2162.

Romano, J. P. (1992). A circular block-resampling procedure for stationary data. *Exploring the Limits of Bootstrap 270*, 263.

Rondonotti, V., J. Marron, and C. Park (2007). Sizer for time series: a new approach to the analysis of trends. *Electron J Stat 1*, 268–289.

Rosenblatt, M. (1956). A central limit theorem and a strong mixing condition. *Proceedings of the national Academy of Sciences 42*(1), 43–47.

Rudge, J. F. (2008). Finding peaks in geochemical distributions: A re-examination of the helium-continental crust correlation. *Earth and Planetary Science Letters 274*(1-2), 179–188.

Rydén, J. (2010). Exploring possibly increasing trend of hurricane activity by a sizer approach. *Environmental and ecological statistics 17*, 125–132.

Scherbina, A. and B. Schlusche (2020). Follow the leader: using the stock market to uncover information flows between firms. *Review of Finance 24*(1), 189–225.

Schruben, L. (1983). Confidence interval estimation using standardized time series. *Operations Research 31*(6), 1090–1108.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics 6*(2), 461–464.

Shao, Q. and L. Yang (2011). Autoregressive coefficient estimation in nonparametric analysis. *Journal of time series analysis 32*(6), 587–597.

Shao, X. (2010a). The dependent wild bootstrap. *Journal of the American Statistical Association 105*(489), 218–235.

Shao, X. (2010b). Extended tapered block bootstrap. *Statistica Sinica 20*(2), 807–821.

Shao, X. and X. Zhang (2010). Testing for change points in time series. *Journal of the American Statistical Association 105*(491), 1228–1240.

Shi, X., C. Beaulieu, R. Killick, and R. Lund (2022). Changepoint detection: an analysis of the central england temperature series. *Journal of Climate 35*(19), 6329–6342.

Shyy, G., V. Vijayraghavan, and B. Scott-Quinn (1996). A further investigation of the lead-lag relationship between the cash market and stock index futures market with the use of bid/ask quotes: The case of france. *The Journal of Futures Markets (1986-1998) 16*(4), 405.

Sims, C. A. (1972). Money, income, and causality. *The American economic review 62*(4), 540–552.

Skoura, A. (2019). Detection of lead-lag relationships using both time domain and time-frequency domain; an application to wealth-to-income ratio. *Economies 7*(2), 28.

Skrøvseth, S. O., J. G. Bellika, and F. Godtliebsen (2012). Causality in scale space as an approach to change detection. *PloS one 7*(12), e52253.

Song, W. T. and B. W. Schmeiser (1995). Optimal mean-squared-error batch sizes. *Management Science 41*(1), 110–123.

Tecuapetla-Gómez, I. and A. Munk (2017). Autocovariance estimation in regression with a discontinuous signal and m-dependent errors: A difference-based approach. *Scandinavian Journal of Statistics 44*(2), 346–368.

Truong, Y. K. (1991). Nonparametric curve estimation with time series errors. *Journal of Statistical Planning and Inference 28*(2), 167–183.

Venkatraman, E. S. (1992). *Consistency results in multiple change-point problems*. Ph. D. thesis, Stanford University.

Vidakovic, B. (2009). *Statistical modeling by wavelets*. John Wiley & Sons.

Vostrikova, L. Y. (1981). Detecting "disorder" in multidimensional random processes. *259*(2), 270–274.

Vuollo, V. and L. Holmström (2018). A scale space approach for exploring structure in spherical data. *Computational Statistics & Data Analysis 125*, 57–69.

Wang, G.-J., C. Xie, F. Han, and B. Sun (2012). Similarity measure and topology evolution of foreign exchange markets using dynamic time warping method: Evidence from minimal spanning tree. *Physica A: Statistical Mechanics and its Applications 391*(16), 4136–4146.

Weigt, G. (2022). *itsmr: Time Series Analysis Using the Innovations Algorithm*. R Foundation for Statistical Computing.

Welch, P. D. (1987). On the relationship between batch means, overlapping means and spectral estimation. In *Proceedings of the 19th conference on Winter simulation*, pp. 320–323.

White, H. and X. Lu (2010). Granger causality and dynamic structural systems. *Journal of Financial Econometrics 8*(2), 193–243.

Wu, C.-F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *the Annals of Statistics 14*(4), 1261–1295.

Wu, W. B. (2005). Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences 102*(40), 14150–14154.

Wu, W. B. (2011). Asymptotic theory for stationary processes. *Statistics and its Interface 4*(2), 207–226.

Wu, W. B. et al. (2007). Strong invariance principles for dependent random variables. *The Annals of Probability 35*(6), 2294–2320.

Wu, W.-B. and Y. N. Wu (2016). Performance bounds for parameter estimates of high-dimensional linear models with correlated errors. *Electronic Journal of Statistics 10*(1), 352–379.

Wu, W. B. and Z. Zhao (2007). Inference of trends in time series. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69*(3), 391–410.

Yao, Y.-C. (1988). Estimating the number of change-points via schwarz'criterion. *Statistics & Probability Letters 6*(3), 181–189.

Yao, Y.-C. and S.-T. Au (1989). Least-squares estimation of a step function. *Sankhyā: The Indian Journal of Statistics, Series A*, 370–381.

Zaciu, R., C. Lamba, C. Burlacu, and G. Nicula (1996). Image compression using an overcomplete discrete wavelet transform. *IEEE Transactions on Consumer Electronics 42*(3), 800–807.

Zeng, K. and E. F. E. Atta Mills (2023). Can economic links explain lead–lag relations across firms? *International Journal of Finance & Economics 28*(2), 1338–1363.

Zhang, N. R. and D. O. Siegmund (2007). A modified bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics 63*(1), 22–32.

Zhao, Z., F. Jiang, and X. Shao (2022). Segmenting time series via self-normalisation. *Journal of the Royal Statistical Society Series B: Statistical Methodology 84*(5), 1699–1725.

Zhong, M., A. F. Darrat, and R. Otero (2004). Price discovery and volatility spillovers in index futures markets: Some evidence from mexico. *Journal of Banking & Finance 28*(12), 3037–3054.