Methods for Change-Point Detection with Additional Interpretability

London School of Economics and Political Sciences

Anna Louise M. M. Schröder

A thesis submitted to the Department of Statistics of the London School of Economics for the degree of Doctor of Philosophy, London, March 2016

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 57,322 words.

Statement of conjoint work

I confirm that Chapter 3 was jointly co-authored with Professor Piotr Fryzlewicz and I contributed 75% of this work. I confirm that Chapter 5 was jointly co-authored with Professor Hernando Ombao and I contributed 75% of this work.

Abstract

The main purpose of this dissertation is to introduce and critically assess some novel statistical methods for change-point detection that help better understand the nature of processes underlying observable time series.

First, we advocate the use of change-point detection for local trend estimation in financial return data and propose a new approach developed to capture the oscillatory behaviour of financial returns around piecewise-constant trend functions. Core of the method is a data-adaptive hierarchically-ordered basis of Unbalanced Haar vectors which decomposes the piecewise-constant trend underlying observed daily returns into a binary-tree structure of one-step constant functions. We illustrate how this framework can provide a new perspective for the interpretation of change points in financial returns. Moreover, the approach yields a family of forecasting operators for financial return series which can be adjusted flexibly depending on the forecast horizon or the loss function.

Second, we discuss change-point detection under model misspecification, focusing in particular on normally distributed data with changing mean and variance. We argue that ignoring the presence of changes in mean or variance when testing for changes in, respectively, variance or mean, can affect the application of statistical methods negatively. After illustrating the difficulties arising from this kind of model misspecification we propose a new method to address these using sequential testing on intervals with varying length and show in a simulation study how this approach compares to competitors in mixed-change situations.

The third contribution of this thesis is a data-adaptive procedure to evaluate EEG data, which can improve the understanding of an epileptic seizure recording. This change-point detection method characterizes the evolution of frequency-specific energy as measured on the human scalp. It provides new insights to this high dimensional high frequency data and has attractive computational and scalability features. In addition to contrasting our method with existing approaches, we analyse and interpret the method's output in the application to a seizure data set.

Dedication

To my family - für meine Familie: meine Eltern und meine Schwester Julia, die immer für mich da waren, wenn ich Ansporn oder Rat brauchte. Und für meine lieben Großeltern, die mich inspiriert haben, diese Herausforderung zu meistern.

Acknowledgements

I would like to thank many people who have helped me through the completion of this dissertation. Firstly, I thank my advisor, Piotr Fryzlewicz, and Hernando Ombao at University of California, Irvine, who I had the luck to collaborate with. Both gave me excellent guidance and feedback for my research over the last years, mentored and advised me honestly and shaped my perspective on the academic world. I consider myself fortunate to have been part of the LSE Department of Statistics, which offered me a great intellectual and work environment. I want to thank especially Ian Marshall, who supported academic staff and students beyond all measure.

I am deeply grateful for my friends, who continue to impress me with their intelligence, positivity and wit. I would not be the person I am without them and this dissertation has greatly benefited from our many motivating and inspiring moments in the last years. Karin Kaiser, Maria Carvalho, Kevin K'Bidi, Anja Schanbacher and Jennifer Petzold - you make London my home. Philipp Kurbel, Karol Campaña, José Ramírez Bucheli, Alex Pucola and Hanna Hoffmann - having you in my life, no matter how far, will always be great source of my energy. Guillaume Henry, your optimism and esprit made a big difference to my life in the past two years and I am deeply grateful for them.

Contents

4 Detection of Changes in Mean and/or Variance using Sequential

List of Figures

List of Tables

Chapter 1

Introduction

The systematic understanding of our past has been in the focus of scientific research since its inception: it allows us to draw conclusions which affect future behaviour. Driven by the desire to grasp the unobservable structure underlying evolutionary time series, an increasing systematization of recorded measurements can be observed throughout history in a variety of areas - from environmental data such as temperature and rainfall, via socioeconomic measures of health, poverty and demographic trends, to the increasingly precise measurement of time itself. Over the last century, digitalisation and technological advances in data management and storage have enabled scientists to systematically analyse evolution in such time series recordings with the general goal of better understanding unobserved underlying processes.

Many evolutionary processes, albeit not all, can be viewed as undergoing occasional, sudden transitions (Brodsky and Darkhovsky, 2013). For example, economists and policy makers classify the state of an economy into recession and recovery and are interested in identifying points in time when a transition between these states can be attested to decide on the appropriate policy measures at any such time (Diebold and Rudebusch, 1996). Similarly, in the field of seismology, researchers are concerned with identifying earthquake predictors by monitoring velocity changes in

seismic waves (Ogata, 2005). Medical doctors are monitoring physiological time se-
ries to identify sudden abnormalities that may relate to a medical condition (Bélisle
et al., 1998; Strens et al., 2003). From a statistical point of view, if the process
underlying such a time series has a time-invariant nature between any two neigh-
bouring points of transition, it can be described as piecewise-stationary process.
The estimation of the number and location of time points where a transition takes
place is subject of this work and will be referred to as change-point detection or
time-series segmentation. Moreover, we show how certain designs of change-point
detection procedures can provide insights in the data beyond change-point locali-
sation.

The main purpose of this thesis is to introduce and critically assess some novel
statistical methods for change-point detection that help better understand the na-
ture of processes underlying observable time series. While these methods can be
used in a range of different applications, they share the objective of detecting sud-
den changes in piecewise-stationary time series to segment these data-adaptively
into approximately stationary intervals of varying length. The ultimate goal of
this work is to fill some gaps in the change-point literature and to add to the in-
terpretability of certain types of observed time series, thereby allowing for a more
insightful analysis and meaningful interpretation of data. Prior to our review of
existing literature we illustrate the usefulness of change-point estimation by means
of a few real-world applications.

It is worth emphasizing that the following are simple examples chosen from
a plethora of fields where change-point detection is applied. In recent years, the
immense growth of digital storage and data processing capacities has moved the

analysis of time series data to a new stage. Data recordings from, to name a few, climate (Reeves et al., 2007; Naveau et al., 2014), speech (Chen and Gopalakrishnan, 1998; Zhang and Hansen, 2008), financial markets (Andreou and Ghysels, 2002; Christoffersen and Diebold, 2006), physiological processes (Bélisle et al., 1998; Strens et al., 2003), internet traffic (Kwon et al., 2006; Kim and Reddy, 2008) or astronomy (Friedman, 1996; Scargle et al., 2013) have increased in quantity and quality, giving researchers room for critical reassessment and possible enhancement of existing methods and development of new tools to analyse, understand and interpret this data.

We describe here two specific data recordings taken from the areas of finance and neuroscience. This data is interesting in the context of change-point detection as we illustrate in Chapters 3 and 5, respectively. First, Figure 1.1 shows the evolution of the S&P 500 equity index between January 1990 and June 2013. The index reflects the market-capitalization weighted price evolution of the 500 largest listed companies in the USA. Its evolution serves as gauge on the general state of one of the world's largest economies and thus plays an important role for economists, policy makers and financial market agents globally. As can be seen, the price index reached a number of local peaks and lows in the last two decades, some of which initiate sharp trend reversals. While the trends might appear fairly obvious at the time resolution shown here, there is no apparent pattern in the corresponding points of trend reversal. In this application, change-point locations can be estimated and this information can be exploited subsequently in forecasting, for instance by constructing a recursive forecast estimator based on the estimated change points (Pesaran and Timmermann, 2002).
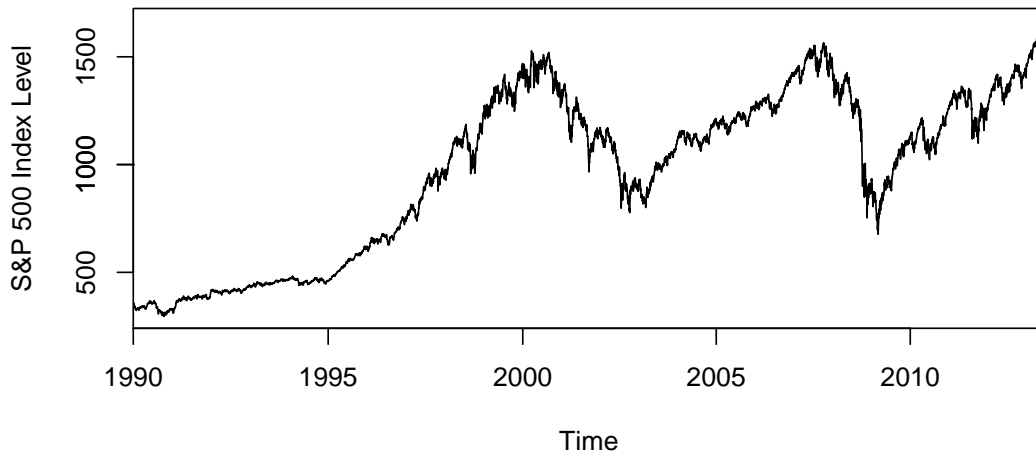
Figure 1.1: S&P500 Price Index, end-of-day level, daily observations between January 1990 and June 2013

A data recording from a second field of application is shown in Figure 1.2, which displays a measurement of electrical activity in the human brain. In a clinical environment, the subject was connected to an electroencephalogram (EEG) while she experienced an epileptic seizure. The data is collected via electrodes placed on the human scalp, at a high frequency of 100 observations per second. This EEG recording contains data collected at 21 electrode channels, but for the purpose of illustration we display a single time series showing the left temporal lobe activity only, i.e. an electrode placed above the left ear. This data set is particularly interesting to better understand epileptic seizures because it was collected starting minutes before the seizure onset. Using tools like change-point analysis, researchers can improve their understanding of abnormalities in the energy distribution and connectivity of brain regions in the minutes leading up to the seizure, to potentially develop new early warning systems for epilepsy patients. Other relevant conclusions that can be drawn from change-point analysis of this data set are the identification of the seizure onset and the spatial seizure origin. Both characteristics are
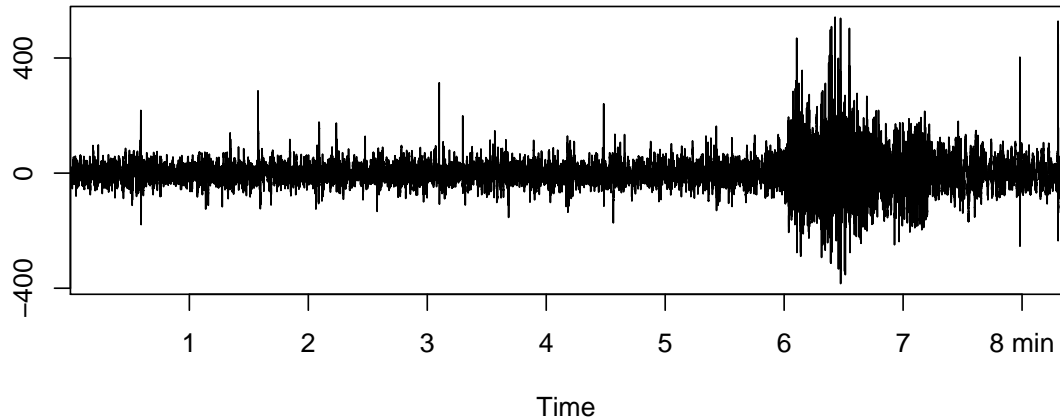
Figure 1.2: Recording before and during a spontaneous epileptic seizure via electroencephalogram. Bipolar electrode measuring electrical activity at the left-temporal channel (T3, located approximately above the left ear)

often estimated by the physician via eye-balling, while automated methods can be valuable tools for verification of such heuristic findings (Tzallas et al., 2012).

We now briefly lay out the structure of this thesis and discuss the contribution of each chapter.

Chapter 2 provides a systematic overview of the existing literature on change-point analysis and various subclasses of the topic. It introduces the reader to the relevant elementary concepts and foundations on which we build in the following chapters.

Chapter 3 discusses a formal stochastic time series model that is used to model local trends in financial time series, including the one depicted in Figure 1.1. The method detects change points in log-return series, the logarithm of daily changes in the observed price level, and takes the current trend estimate to be the aver-

age return between the most recent estimated change point and the current time. The contribution of this work goes beyond advocating change-point detection to estimate local trends in financial returns. Our main objective is to propose an alternative approach to statistical time series analysis, whereby the time series is spanned by an orthonormal oscillatory basis induced by change points in the series' conditional mean value. In this context, the representation basis is assumed to be unknown and needs to be estimated data-adaptively using a change-point detection procedure. Our approach yields a family of forecasting operators for the financial return series, parameterised by a single threshold parameter, which can be adjusted flexibly depending on forecast horizon or loss function.

Chapter 4 discusses a question related to that of Chapter 3: how to detect possibly many change points in a time series when changes can take place in the series' conditional mean, variance or both quantities simultaneously. This more general problem of estimating number, location and type of change points has not received much attention in the literature. It comes with a number of challenges which we address using a sequential testing method. Our core contributions are to systematically describe the difficulties for change-point detection and to propose a novel method to overcome these. Moreover, we discuss computational aspects of the method and compare it to competing approaches in a simulation study.

Chapter 5 introduces a new approach for the analysis of EEG data, a data-adaptive way to evaluate EEG recordings of brain activity as the one shown in Figure 1.2. The method proves to be very useful in increasing the understanding of brain processes before and during an epileptic seizure. Our proposed approach

characterizes energy evolution measured on the human scalp through change-point detection. In addition to discussing the features of the method and contrasting it with existing approaches, we analyse and interpret the method's output in the application to the seizure data set which is partly depicted in Figure 1.2.

Chapter 6 concludes.

As the different approaches covered in Chapters 3-5 require different mathematical frameworks and therewith notation, a glossary containing the relevant symbols is provided for reference at the end of each chapter. General mathematical conventions and our strong desire to use simple and short notation make it necessary to redefine some notation in different chapters, but we aim at maximizing consistency. A list of abbreviations repeatedly used in this work is provided in Section D in the Appendix. In the form of co-authored papers, parts of this thesis have been submitted to or printed by peer-reviewed statistical journals. Wherever this is the case, we will declare it explicitly.

Chapter 2

Background and Related Work

2.1 Piecewise Stationary Processes

The change-point detection problem has received much attention by the academic community. A gauge of the development of this research area is presented in Figure 2.1, which contains anecdotal evidence on the number of search results for the term 'change point' in Google Scholar. This chapter provides an introduction and overview to this growing field of statistical research and discusses different existing approaches to the problem.

Time series processes can be classified as being stationary or nonstationary. There are different definitions of stationarity, meaning in the most narrow sense (strict statonarity) time-invariance of the distribution underlying the process, or time-invariance of mean and covariance in the wide-sense (weak) stationarity. In many real-world applications stationarity is arguably a strong assumptions as process characteristics evolve over time.

In this thesis only sudden changes are considered as opposed to e.g. smooth transitioning. We refer the reader to general introductory readings on time series for an overview on other types of nonstationarity, such as the discussions in
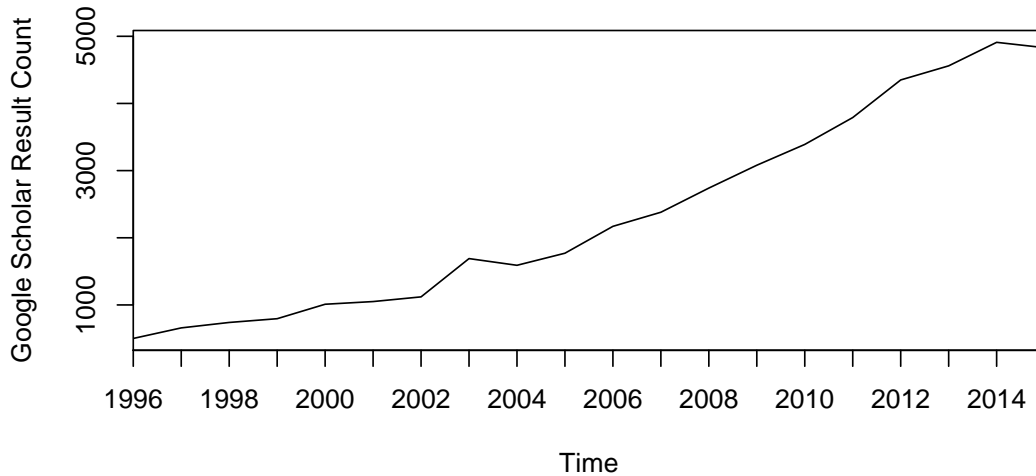
Figure 2.1: Number of search results of the term 'change point' in Google Scholar, by year. Approximate number for values exceeding 1000. Source: scholar.google.co.uk; retrieved 22 January 2016

Priestley (1988) and Nason (2006). In the presence of sudden changes we can view a univariate process $\mathbf{x} = \{x(t), t = 1, \ldots, T\}$ as following some distribution $\mathcal{F}_i$ between the $(i-1)$th and the ith change point,

$$x(t) \sim \mathcal{F}_i \leftrightarrow \eta_{i-1} < t \leq \eta_i, \ i = \{1, \ldots, N+1\} \tag{2.1}$$

where we define the set of change points as an ordered collection of points in $\{1, \ldots, T\}$, $\mathbb{N} = \{\eta_i : \ 1 < \eta_1 < \eta_2 < \cdots < \eta_N < T\}$ with the convention $\eta_0 = 0, \eta_{N+1} = T$.

The process $\mathbf{x}$ is locally piecewise-stationary and can be approximated by a sequence of stationary processes that may share certain features such as the general functional form of the distributions $\mathcal{F}_i$ or characteristics such as an expected value of zero. The focus of this work is the general univariate change-point process of Equation (2.1) and to some extent also a multivariate generalization. However,

there exist a number of process subclasses we mention briefly below, for completeness and because they can be of interest in specific applications.

Firstly, the most simple case of $N = 1$ change point has received most attention, often for the reason of better analytical traceability. In the next section we discuss methods of change-point detection for this simple setting. Secondly, we mention here the situation where a process transitions into an abnormal state and returns to its initial state, i.e. $N = 2$ and $\mathcal{F}_1 = \mathcal{F}_3 \neq \mathcal{F}_2$. This process describes an epidemic change and is of interest in many quality control and medical applications, see e.g. Levin and Kline (1985); Yao (1993); Chen and Gupta (2012) or Kirch et al. (2015).

Thirdly, building on the same principle assumption of a specific number of states is the wider family of regime-switching models. The process jumps between, say, K different distributions $\mathcal{F}_i, i = \{1, \ldots, K\}$, $K \leq N$. The goal is to estimate the number of change points N, their locations η_i, and, if not assumed to be known, also the number of regimes $K \in \{1, \ldots, N\}$. In practice we frequently have $K \ll N$, which makes this estimation problem simpler than the unrestricted change-point problem where generally $\mathcal{F}_i \neq \mathcal{F}_j, \ i, j = \{1, \ldots, N+1\}, i \neq j$.

The simplification can be easily justified if there exists a priori information about the (small) number of possible states. In particular applications that require the estimation of a large number of unknowns with comparably small number of observations T can strongly benefit from assuming the process to be regime-switching instead of having a general piecewise-stationary form. Moreover, transitions between regimes provide the analyst with a new dimension to the interpretation of results - the estimation of transition probabilities that may or may not be time-varying. These probabilities describe the chance that at any time point the process

transitions into a certain regime, given its history (or, assuming a Markov-type transition process, the only most recent regime). Popular fields of application are economic and financial time series, but also image or DNA segmentation (see e.g. Hamilton, 1990; Green, 1995; Braun and Muller, 1998; Pelletier, 2006; Battaglia and Protopapas, 2012).

2.2   Approaches to Change-Point Detection

The methods for the detection of change points in the general setting of Equation (2.1) can be classified as online (sequential) or offline (retrospective).

In online change-point detection the analysis is performed sequentially as more data becomes available. The goal is typically to be fast in identifying change points near the most recent observation, while controlling the rate of false positive detections. We refer the reader to Basseville and Nikiforov (1993) and Chakraborti et al. (2001) for a more thorough discussion of the problem of online change-point detection and its early development. Adams and MacKay (2007) introduce a detection procedure with a Bayesian formulation, while Choi et al. (2008) discuss a spectral method. Turner et al. (2009) and Caron et al. (2012) are, respectively, examples of recent maximum likelihood and nonparametric approaches to the problem. All these works contain an overview of the respective strands of literature.

The focus of this work is offline detection, in which information on the full time series $\mathbf{x} = \{x(t), t = 1, \ldots, T\}$ is available for analysis. Here, the researcher is usually interested in identifying the correct number of change points and their locations as precisely as possible. The following two sections discuss, respectively,

offline detection of single and multiple change points.

2.2.1 Offline Testing and Detection Procedures for a Single Change Point

We discuss here statistical tests and detection methods for a single change point, $\mathbb{N} = \{\eta^*\}$, in a (fully observed) time series $\mathbf{x}$.

Testing for a change point and detecting it are two different problems, but they often go hand in hand. In applications, the research question can mingle both as they may be equally important, asking *is there a change point* and *if so, where is it most likely located* simultaneously. Testing the existence of a change point at an undetermined location, or estimating its most likely location while assuming the existence of a change point, i.e. the situation where only one of the questions matters, are usually of limited relevance in practice.

As noted by Basseville (1988), change-point estimation differs from classical hypothesis testing in that a multiple testing problem is implied: every point (except near boundaries, see Section 2.3) is a priori a candidate change point. For an appropriately constructed test statistic, if there is evidence of a change, the candidate point providing strongest evidence becomes the change point estimate.

Likelihood- and Cumulative Sum-Based Detection

Frequently encountered in parametric change-point detection are proposals using ***likelihood ratio (LR)*** tests. Here, the analyst assumes a common functional

form, say $\mathcal{F}_i = \bar{\mathcal{F}}(\boldsymbol{\theta}_i) \; \forall \; i$, where $\boldsymbol{\theta}_i = (\theta_{i,1}, \ldots, \theta_{i,Q})'$ is a parameter vector fully defining the distribution $\mathcal{F}_i$. In general the principle of likelihood estimation is apparent in the multiplicative partitioning

$$\mathcal{L}(\boldsymbol{\theta}|\mathbf{x}) = \mathcal{L}(\theta_1, \ldots, \theta_Q|x(1), \ldots, x(T)) = \prod_{t=1}^{T} \boldsymbol{f}(x(t)|\theta_1, \ldots, \theta_Q)$$

where $\boldsymbol{f}(\cdot|\boldsymbol{\theta})$ denotes the pdf or pmf of $\bar{\mathcal{F}}$ conditional on the parameter vector $\boldsymbol{\theta}$. The likelihood function is maximized to estimate the parameter vector $\boldsymbol{\theta}$ given the time series $\mathbf{x}$ and assuming some distributional form. To compare a model containing a change point at point b to one that does not contain a change point, the likelihood function suprema over the parameter spaces are compared. When the location of the change point is unknown, the double-supremum of the total likelihood function over all candidate points b and the corresponding parameter spaces are compared to the supremum of the no-change likelihood over the parameter space. This defines the likelihood ratio for the estimation of a change point,

$$\mathcal{LR}(\mathbf{x}) = \frac{\sup_b \sup_{\boldsymbol{\theta}_i \in \boldsymbol{\Theta}_i, i=\{1,2\}} \mathcal{L}(\boldsymbol{\theta_1}|x(t), t = 1, \ldots, b) \mathcal{L}(\boldsymbol{\theta_2}|x(t), t = b+1, \ldots, T)}{\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\theta}|x(t), t = 1, \ldots, T)}$$

where $\boldsymbol{\Theta}$ and $\boldsymbol{\Theta}_i, i = \{1, 2\}$ denote, respectively, the parameter space for the no-change and change cases.

One of the first works considering likelihood theory is the Sup-F or Sup-Wald test (Quandt, 1958, 1960) for a change in the conditional mean of normally distributed data. Since these early publications, much has been contributed to the understanding of the limiting and exact distribution of this type of test statistic for various distributional forms. The cases of normally distributed data and of pro-

cesses where only a single parameter varies have received most attention (e.g. Sen and Srivastava, 1975; Horváth, 1993; Chen and Gupta, 1995; Gombay and Horváth, 1996; Chen and Gupta, 1997), but numerous other situations such as binomial and exponential processes (Worsley, 1983, 1986), linear regression (Kim and Siegmund, 1989), autoregressive moving-average (ARMA) models (Robbins et al., 2015) and copula models (Bouzebda and Keziou, 2013) have also been discussed. We consider LR tests for normally distributed data with changes in the mean, the variance or both parameters in Chapter 4. See also Chen and Gupta (2012) for an overview of change-point analysis for various parametric models.

Another strand of literature is concerned with so-called **cumulative sum (CUSUM)** tests for change-point detection. This approach goes back to the seminal work of Page (1955), where the CUSUM test is derived for a change in one distribution parameter in a sequential setting. The academic community has introduced a plethora of modifications and extensions since, but the principle remains that inference is made on the cumulative sum of the data or a transformation thereof, weighted in some way. The point at which the (absolute) value of this cumulative sum is maximized is considered the most likely change-point location. To assess the significance of the test, this maximum is compared to a threshold. If exceeded, the candidate location is regarded a change point.

To give an intuition of this type of test, we adapt a systematization taken from Brodsky and Darkhovsky (1993). This formulation describes a family of retrospective CUSUM statistics for a change in mean, but one can substitute transformations of $x(t)$ to test for other types of change (for example the squared deviation from the sample mean to test for a change in the variance). For the change in mean, we

write

$$\mathcal{C}(\mathbf{x}|b) = \left(\frac{b(T-b)}{T}\right)^{\delta} \left(\frac{1}{b}\sum_{t=1}^{b} x(t) - \frac{1}{T-b}\sum_{t=b+1}^{T} x(t)\right) \tag{2.2}$$

with the change-point candidate $b \in \{1, \ldots, T\}$ and $\delta \in [0, 1]$. The function is maximized in absolute terms where the weighted contrast between the segments before and after the change-point candidate b is maximized. The case of $\delta = 0$ is discussed in Brodsky and Darkhovsky (1993, Section 3.3), the case of $\delta = 1$ in Deshayes and Picard (1986, see also the link to trend detection we mention in Section 2.4). We discuss some interesting properties of the CUSUM statistic for $\delta = 1/2$ in detail in Chapter 3; for a Gaussian sequence the CUSUM test corresponds in this case to the maximum likelihood function. For a process following $\bar{\mathcal{F}}(\boldsymbol{\theta}_1) = N(\mu_1, \sigma)$ before a change in mean and $\bar{\mathcal{F}}(\boldsymbol{\theta}_2) = N(\mu_2, \sigma)$ afterwards, the level of δ determines the balance between size and power of the CUSUM test of Equation (2.2). Brodsky and Darkhovsky (1993) show that under certain conditions and as $T \to \infty$, $\delta = 1$ offers the minimum false positive rate, while $\delta = 0$ yields the lowest false negative rate. $\delta = 1/2$ is optimal in terms of change-point estimation accuracy, i.e. the absolute distance between estimated and true change point, $|\hat{\eta} - \eta|$.

We conclude this discussion mentioning some developments related to the family of CUSUM tests: Inclan and Tiao (1994) discuss the detection of a change in the variance; Kokoszka and Leipus (1999) the detection of changes in the parameters of an autoregressive conditional heteroscedasticity (ARCH) process and Whitcher et al. (2002) the detection of changes under strong dependence. McGilchrist and Woodyer (1975) derive a distribution-free CUSUM, Pettitt (1979) introduces a robust CUSUM statistic related to the Mann-Whitney sign test and Bhattacharya

and Zhou (1994) propose a robust version of the CUSUM statistic of Equation (2.2) that operates on the ranks of $\mathbf{x} = \{x(t), t = 1, \ldots, T\}$.

Remark on Thresholding In the hypothesis testing using LR we explicitly compare the test statistic under the null hypothesis to the one under the alternative evaluated at the most likely change-point location. This ratio is compared to a critical value that is derived under the null hypothesis and reflects a significance level. In CUSUM tests, the statistic maximizes the contrast between two segments. To decide about the significance of a change point, the absolute value of this statistic is also compared to a threshold derived under the null of no change. Alternatively to using a theoretically derived threshold, the analyst can estimate the threshold via **permutation**. To illustrate, in a simple setting this can be done as follows.

Randomly permute the original time series $\mathbf{x} = \{x(t), t = 1, \ldots, T\}$ to generate $\mathbf{x}^{(p)} = \{x(s), s = r_1^{(p)}, \ldots, r_T^{(p)}\}$ where the set $\{r_i^{(p)}\}$ is a permutation of $\{1, \ldots, T\}, \forall p$. Compute the maximum test statistic $\mathcal{T}(\mathbf{x}^{(p)})$ over all change-point candidates. The empirical distribution of this statistic found by repeating this procedure P times then allows for change-point inference: its $(1 - \alpha)$-quantile defines the permutation-based critical value $\tau_p = \mathrm{Quant}^{\alpha\%}(\mathcal{T}(\mathbf{x}^{(p)}), p = 1, \ldots, P)$. This simple approach can be extended in various directions, for instance to preserve the dependency structure using blockwise permutation. In this context Hušková et al. (2008) discuss the regression and the pair bootstrap for autoregressive processes with lag L. A CUSUM-type test statistic is constructed using a permutation of estimated residuals from an $AR(L)$ model containing a change point. The authors show that the resulting test procedure yields consistent results and compare it to

asymptotic threshold results of Hušková et al. (2007) for the same test statistic. Related literature on permutation-based estimation includes Antoch and Huskova (2001); Kirch (2007); Zeileis and Hothorn (2013); Matteson and James (2014) and Arias-Castro et al. (2015).

Detection using Bayesian Methods

Bayesian approaches to change-point detection require the specification of priors on the number and position of change points and the parameter vectors $\boldsymbol{\theta}_i$ between any two change points; the latter can be specified to allow for a specific type of change, e.g. in only one element of $\boldsymbol{\theta}_i$. As with many areas of statistics, we claim that it is a matter of application and the analyst's choice whether to use frequentist or Bayesian analysis; it follows that further advancements are desirable in both methodological families, see also the discussion in Bayarri and Berger (2004). The following chapters employ frequentist-type methods, but for completeness we provide a short overview of existing Bayesian approaches to offline change-point detection here.

Early Bayesian formulations of the single change-point problem can be found in Chernoff and Zacks (1964); Broemeling (1972, 1974) and Smith (1975). The authors cover the general situation where the distributional forms on the segments before and after the change point are unknown, or only one of them is known. The product partitioning model, formally introduced by Barry and Hartigan (1992), decomposes the total likelihood of $\mathbf{x} = \{x(t), t = 1, \ldots, T\}$ under the alternative into a product of the segmentwise likelihood functions

$$\prod_{i=1}^{N+1} \prod_{t=\eta_{i-1}+1}^{\eta_i} \boldsymbol{f}_i(x(t)|\boldsymbol{\theta}_i)$$

where $\boldsymbol{f}_i(\cdot|\boldsymbol{\theta}_i)$ denotes the probability density function of $\mathcal{F}_i$ conditional on the parameter vector $\boldsymbol{\theta}_i$. For the single change-point problem, $N = 1$, Chernoff and Zacks (1964) cover the change-in-mean case for small samples. Broemeling (1972, 1974) provide results on the detection of a change point for the cases of Bernoulli, exponential and normal sequences, the latter with known or unknown variance and a focus on estimating the change size. Smith (1975) discusses the case of normal distributions with a change in mean in more detail and derive the estimator of the probability of a change at every point $t \in \{1, \ldots, T\}$ using maximum likelihood. Booth and Smith (1982) discuss the corresponding multivariate case, as well as regression and ARMA models. Carlin et al. (1992) approach the problem using Gibbs sampling within a hierarchical Bayesian formulation and illustrate the principles on poisson and Markov-chain processes and in a regression context. In a series of papers, Perreault and coauthors use Gibbs sampling to detect changes in the mean or variance of univariate data and in the mean vector of multivariate data (Perreault et al., 1999, 2000a,b,c). For a recent review of (primarily) offline Bayesian and maximum likelihood methods see also Jandhyala et al. (2013).

Nonparametric Detection

Harchaoui et al. (2009) is an example of a completely nonparametric change-point detection method. The authors propose kernel estimation and a discriminant ratio

to test for existence and to locate a single change point. Other authors have considered direct density-ratio estimation for change-point identification (Kawahara and Sugiyama, 2009; Kanamori et al., 2010; Liu et al., 2013). Here, inference is made on the ratio between the densities of the data before and after a change-point candidate. The central argument in this strand of literature is that by considering a density ratio, these approaches do not require knowledge about the densities themselves. Other examples for nonparametric approaches are CUSUM-type statistics based on the empirical distribution functions, see Giraitis et al. (1996) and, within a bootstrap approach, Inoue (2001).

Detection in the Frequency Domain

Change-point detection methods have also been developed in the frequency domain. Consider for example the piecewise stationary Fourier process $\mathbf{y} = \{y(t), t = 1, \ldots, T\}$ that goes back to the work of Priestley (1965),

$$y(t) = \int_{(-\pi,\pi]} A_i(\omega) \exp(\imath \omega t) dZ(\omega), \quad \eta_{i-1} < t \leq \eta_i \tag{2.3}$$

where $\exp(\imath \omega t)$ is a complex exponential oscillating at frequency ω, and $A_i(\omega)$ and $dZ(\omega)$ are, respectively, the associated local amplitude and the corresponding orthonormal infinitesimal increment of a stochastic process $Z(\cdot)$. The support of the complex exponential $\exp(\imath \omega t)$ is the set of all integers t and its oscillations exhibit a homogeneous behaviour within each segment $\{\eta_{i-1} + 1, \ldots, \eta_i\}$, for all ω. In this representation change-point detection is typically considered equivalent to identifying changes in the distribution of the amplitude coefficients $A_i(\omega)$ and detection

methods have been developed for example in Deshayes and Picard (1980); Picard (1985); Giraitis and Leipus (1992); Wang (1995); Lavielle and Ludeña (2000); Choi et al. (2008) and Preuß et al. (2015).

Since detection of change points in the frequency domain may seem less intuitive to some, we provide two examples in which this approach is used. In Chapter 5 we analyse EEG data as shown in Figure 1.2 and assume the data to follow a process similar to the one in Equation (2.3). EEG data is commonly analysed in the frequency domain by considering changes in the distribution of the frequency-dependent squared amplitude coefficients, the so-called spectral energy distribution. Hence the analyst benefits from a detection procedure that directly provides insights into the nature of a change in terms of this energy distribution, such as whether the change can be identified at only a narrow band of frequencies or is widely apparent over the frequency range.

Another application where change points in the frequency domain are of interest can be found in the financial world: in financial theory, links are derived between multifrequency models and market participants with different investment horizons (Hasbrouck and Sofianos, 1993; Gençay et al., 2001; Calvet et al., 2006; Baron et al., 2012). Any information about the type of market participant who induces a change in the overall trade pattern are valuable to other market participants and the regulator. For example, the controversial theories that link increased market volatility to high frequency or algorithmic traders can be either further supported (e.g. Zhang, 2010) or rebutted (e.g. Brogaard, 2010).

Change-Type Specific Detection

We discriminate above between different estimation methods to detecting a single change point. Now we want to briefly draw attention to an essential underlying question: what type of change do we aim at detecting? Arguably most change-point detection tests focus on univariate processes with changes in the mean only, assuming the conditional distribution to be unchanged otherwise, see e.g. Chernoff and Zacks (1964); Sen and Srivastava (1975); Gupta and Chen (1996); Vogelsang (1998); Horváth et al. (1999); Harchaoui et al. (2009); Arlot and Celisse (2011); Ning et al. (2012) or Arlot et al. (2012). As we discuss in Chapter 4, if unfounded the assumption of changes occurring in the mean only can yield severe distortions, even in (from a detection point of view) 'simple' situations.

Other methods focus on detecting changes in the variance, while assuming the mean to remain constant or known (Bhattacharyya and Johnson, 1968; Hsu, 1977; Davis, 1979; Inclan and Tiao, 1994; Chen and Gupta, 1997; Lee et al., 2003; Sansó et al., 2003; Berkes et al., 2004; Gombay, 2008; Casas and Gijbels, 2009). As with the change in mean-only case, if the assumption of changes occurring only in the variance is inappropriate the consequences can be substantial.

Again other authors consider detection of changes in other characteristics, e.g. the covariance (Chen and Gupta, 2004; Lavielle and Teyssiere, 2006; Aue et al., 2009), extremes (Raimondo and Tajvidi, 2004; Naveau et al., 2014) or quantile-based copula spectra (Birr et al., 2015). In principle, a change can be detected in different characteristics of the underlying process by transforming the data, as argued in Brodsky and Darkhovsky (2013). However, as we show in Chapter 4,

in the presence of an unknown number of possibly coinciding changes in different characteristics a powerful change-point detection method cannot rely solely on this theoretical argument.

2.2.2 Estimation of Multiple Change-Point Locations

This section presents a review of selected multiple change-point detection approaches. Assume there are $N > 1$ change points present in $\mathbf{x} = \{x(t), t = 1, \ldots, T\}$. The analyst is concerned with the identification of the number of change points and their locations $\eta_i, i \in \{1, \ldots, N\}$. With this goal, one would ideally want to consider all possible combinations of number and locations of change points on $\{1, \ldots, T\}$. The resulting models have to be compared using some goodness-of-fit measure while also taking the model parsimony into account.

However, the comparison of all possible models is not feasible for T larger than a few hundred. For the purpose of illustration, consider a situation where change points are separated by some small integer δ, $0 < \delta \ll T$. Then one would have to estimate the model fit for T models with exactly one change point, $T(T - 2\delta - 1)$ models with two change points, etc. As we explain below, existing procedures for multiple change-point detection use smart arguments to exclude those solutions that are inferior to at least one competitor solution. The more efficient this exclusion process is implemented, the faster the optimal solution can be estimated. If a near-optimal solution is acceptable to the analyst, the computational speed of the estimation can be further improved, but potentially at the expense of accuracy.

In the following, we first discuss some global optimization approaches that solve

the detection problem considering all possible candidates at once. Later we discuss recursive optimization, which identifies locally optimal estimates and thus accepts a possible small inaccuracy in terms of global optimality in exchange for, generally speaking, higher computational speed. For global optimization, two parts are of interest: a) estimating an optimal fit for a fixed number of change points, $N^* = \{0, \ldots, N^{\max}\}$, and b) selecting the best model based on goodness of fit and a specified cost function or penalty that reflects model complexity.

Dynamic Programming Algorithms and Pruning

The goal of global optimization algorithms is to alleviate the computational difficulties described in the previous paragraph. Hawkins (2001) and Maboudou-Tchao and Hawkins (2013), for instance, discuss multiple change-point estimation via dynamic programming using a likelihood-based segmentation. Within this framework, the separability of the likelihood over different segments facilitates the application of Bellman's principle of optimality, i.e. the total model fit can be directly decomposed into the model fit on the individual segments. This allows for the construction of an recursive procedure with complexity $O(N^{\max}T^2)$, where $N^{\max}$ is the user-defined maximum number of permitted change points. Hawkins (2001) proposes a sequence of generalized LR tests to compare the respectively best model with N^* change points versus a competitor with $(N^* + 1)$ change points. Maboudou-Tchao and Hawkins (2013) propose the use of an information criteria-based penalty for model selection, on which we expand below. The works of Bai (1997); Prandom et al. (1997); Bai and Perron (1998, 2003); Zeileis et al. (2003) and de Castro and Leonardi (2015) are examples using similar dynamic programming arguments in

various applications, such as regression.

Jackson et al. (2005) and Killick et al. (2012) also develop dynamic programming algorithms, respectively, without and with a pruning step to improve computational speed. Both approaches are exact and operate with a worst-case complexity of $O(T^2)$, and both employ a penalty function to account for the number of segments. This function can be decomposed as the sum of segment-specific costs, which allows the use of fundamental dynamic programming arguments. However, in the pruning algorithm of Killick et al. (2012), if there are change points in the data, under certain assumptions on the parameters and change-point spacings and using a linear penalty the practical complexity can be near-linear. Rigaill (2015) also propose an algorithm containing a pruning step, but this is based on the functional cost of segmentation and is linear in the single parameter case. Maidstone et al. (2016) develop the pruning idea further by combining the last two approaches conceptually and, inter alia, propose a penalty-based pruning algorithm that is as good or better than the algorithm of Killick et al. (2012), in the sense that it prunes more.

Genetic Algorithms

Davis et al. (2006) propose the use of a genetic algorithm instead. The term 'genetic' is founded in a link to the Darwinian evolution theory: the best segmentation is recovered by successively combining parent segmentations into (in terms of some specified optimality criterion) better fitting children segmentations, while discarding any combinations that do not improve the fit. Ultimately, by so-called crossover and mutation steps and by discarding segmentations with relatively poor fit, if used with the appropriate specifications the algorithm converges after several

generations to a best possible offspring. The concept of genetic algorithms has also been applied by e.g. Battaglia and Protopapas (2011) for the segmentation of a regime-switching process and Hu et al. (2011) for a nonparametric model. In Davis et al. (2006), optimality is defined by the minimum description length principle, which we discuss briefly below.

Model Selection Approaches

One of the most popular approaches to model selection for multiple change-point detection is via classical information criteria (IC) such as the Schwarz criterion (commonly abbreviated SIC or BIC, Yao, 1988). See e.g. Yao and Au (1989) for a testing procedure using maximum likelihood, Chen and Gupta (1997) using a CUSUM test, Arlot et al. (2012) using a kernel-based test, and Braun et al. (2000) and Bardet et al. (2012) for quasi maximum likelihood approaches. Typically, a given test statistic is optimized for each of a range of change-point numbers $N^* = \{0, \ldots, N^{\max}\}$ and then the resulting models are compared based on their goodness of fit while penalizing the model complexity by adding an IC term. For example, for likelihood-based testing the SIC penalty $Q(N+1) \ln T$ is added to twice the negative log-likelihood, where $Q(N+1)$ is the number of model parameters, Q per segment.

There are numerous alternative penalty functions to the SIC criterion. We mention here Pan and Chen (2006), who advocate the modified information criterion (MIC) for the application to change-point models, which contains an additional term to balance out effects that occur if a change point is very close to the start or end of the interval. The MIC penality is equivalent to the SIC under the null

hypothesis. Under the alternative, it can be expressed as

$$\text{SIC} + C \sum_{i=1}^{N+1} \left((\eta_i - \eta_{i-1})/T - 1/(N+1) \right)^2 \ln T$$

where $C > 0$ is a constant. Pan and Chen (2006) show in a simulation study that SIC and MIC are both converging, but MIC is generally more powerful if changes occur at the extremes. Another series of papers proposes a simple penalization by the scaled number of segments $N + 1$ (Lavielle and Moulines, 2000; Lavielle and Ludeña, 2000; Lavielle and Teyssiere, 2007). This is closely related to works adapting the least absolute shrinkage and selection operator to the change-point detection problem, such as Harchaoui and Levy-Leduc (2010); Bleakley and Vert (2011) and Shen et al. (2014).

As argued e.g. in Chen and Gupta (2012), the IC approach and similar penalization techniques for model selection do not require the analyst to specify a threshold or significance level for a change-point test. However, if the difference between penalized goodness of fit measures of different model specifications is small, any conclusion based on a particular model selection penalty can be purely a result of disturbances. Hence a certain minimum difference may be required which, following Chen and Gupta (2012), can be interpreted in a similar way as thresholds. Arlot and Celisse (2011) propose an alternative for model selection, a cross validation procedure for change-point detection which has the objective of minimizing the quadratic risk. Their two-step method proceeds similarly to the dynamic programming methods by first choosing a segmentation for a range of change-point numbers and then selecting the optimal number of change points, but it uses cross selection in both steps. Hocking et al. (2013) propose the data-adaptive estimation

of an appropriate threshold in an application-specific training set using interval regression.

Fryzlewicz (2012) and Haynes et al. (2015) point out that, in general, a flexible detection procedure might be of interest to the analyst where the number of change points is estimated as function of the threshold level. Fryzlewicz (2012) makes a case for the use of a range of threshold values to gain additional insights in the data. Haynes et al. (2015) point out that while theoretical thresholds are appropriate under the correct assumptions about the underlying process, their performance under misspecification can be poor - an observation we can confirm in Chapter 4. The authors propose a new detection method that relies on a dynamic algorithm but finds the optimal segmentation for a range of threshold values.

Binary Segmentation Algorithm

Binary segmentation (BS, Vostrikova, 1981) is an recursive optimization approach, frequently described as greedy algorithm, which tests for the existence of one change point at each stage, i.e. it maximizes the fit successively change point by change point and thus achieves optimality at each stage. A generic description of the recursive algorithm underlying binary segmentation is provided in Figure 2.2. It is initiated by setting the interval boundaries to $s = 1$ and $e = T$ and requires the specification of a threshold ζ and a test statistic $\mathcal{T}_{s^*,b^*,e^*}(x) = f(x(s^*), \ldots, x(e^*))$, evaluated on $x(t), t = \{s^*, \ldots, e^*\}$ with change-point candidate b^*. Standard BS has an empirical computation complexity of $O(T \log T)$ and is easily implemented. Vostrikova (1981) shows that the algorithm yields consistent estimates of the change-point number and locations and many authors have used this approach

Initialise $\hat{\mathbb{N}} = \emptyset$
function BINARY SEGMENTATION($\mathcal{T}_{s^*,b^*,e^*}(x)$, s, e, ζ)
    if $e - s > 1$ **then**
        $b_0 := \arg\max_b \mathcal{T}_{s,b,e}(x)$
        if $\mathcal{T}_{s,b,e}(x) > \zeta$ **then**
            add b_0 to the set of estimated change-points $\hat{\mathbb{N}}$
            BINARY SEGMENTATION($\mathcal{T}_{s^*,b^*,e^*}(x)$, s, b_0, ζ)
            BINARY SEGMENTATION($\mathcal{T}_{s^*,b^*,e^*}(x)$, $b_0 + 1$, e, ζ)
        end if
    end if
end function

Figure 2.2: Binary segmentation algorithm

to detect multiple change points in time series, for example Vostrikova (1981); Venkatraman (1992); Bai (1997); Cho and Fryzlewicz (2012) and Fryzlewicz and Subba Rao (2014). Variations of BS are for example Circular Binary Segmentation (Olshen et al., 2004; Venkatraman and Olshen, 2007) and Wild Binary Segmentation (WBS, Fryzlewicz, 2014) which offer more accurate estimation at the expense of some of the computational efficiency.

We discuss the principles of BS and WBS in more details in the following chapters. Here we conclude by noting three substantial differences to global optimization-type algorithms. First, as pointed out e.g. by Hawkins (2001) and Killick et al. (2012), while BS maximizes the fit at each step locally, the end result is not necessarily identical to an optimal segmentation in the global sense. In practice, however, this difference also represents the trade-off with computational speed: in the worst case, even pruning algorithms have a computational complexity of O(T^2); e.g. for the penalty-based algorithm of Maidstone et al. (2016) such a case is present when the number of change points is high relative to T, because the cost function of each previously detected change point has to be updated at every time point. One can

also construct situations where BS has a computational complexity of $O(T^2)$, but these are pathological cases: this result can only be achieved by continued false detection of change points at the interval boundaries, meaning that the detection method is in any case inadequate for the data. Second, the recursive identification of change points can be interpreted as in a test-statistic specific importance ranking, i.e. the hierarchical structure of the identified change points may help to interpret data. We illustrate the usefulness of this additional information by means of a few simple examples in financial time series in Section 3.5.2. For the family of dynamic algorithms (with and without pruning), this additional interpretability is generally not available as direct output. Finally, dynamic programmes typically require the specification of some penalty function, and the results are sensitive to this choice as becomes clear in Chapter 4. While BS requires the specification of a threshold, the number of change point estimates is monotonically decreasing as the threshold increases and the change-point location estimates are not changing if the threshold is lowered, in the sense that if for some threshold $\zeta > 0$, $\hat{\eta} \in \hat{\mathbb{N}}_\zeta$, then this $\hat{\eta}$ will also be in the set $\hat{\mathbb{N}}_{\zeta - c}$, for any constant c with $0 \leq c \leq \zeta$.

Bottom-Up Algorithms

We mention another strand of literature in change-point detection, which basically inverts the principle of BS, starting with a fine partition and subsequently merging adjacent segments by removing the change point that divides them. The initial partitioning restricts the potential change-point locations to coincide with dyadic partitions of the full time interval, i.e. locations that are multiples of 2^j with j being a positive integer. An example for this dyadic partitioning is Ombao et al. (2002)

who introduce the so-called SLEX transformation: the smooth localised complex exponentials transformation describes a family of orthogonal Fourier-type transformations that are localised in time. Via the best-basis selection algorithm, a bottom-up approach as described above, the best segmentation is chosen in $O(T \log T)$ by minimizing a cost function that accounts for model complexity.

The preferred optimality criterion of Coifman and Wickerhauser (1992) in the context of best basis selection is Shannon entropy, which emphasizes the preservation of information. However, other authors have put forward application-specific alterations, e.g. Brooks et al. (1996) who suggest to use an additional smoothness measure to avoid too fine segmentation. See also Rao and Kreutz-Delgado (1999) for a discussion. A similar goal of representing the time series in an compressed way is pursued by Davis et al. (2006, 2008) within their genetic algorithm. The authors use the minimum description length (MDL) principle, which defines the best-fitting model as the model that enables maximum compression of the data (Rissanen, 1989).

Bayesian Approaches to Multiple Change-Point Detection

Bayesian methods have been proposed in various formulations and with various estimation methods. For instance, Chib (1998) formulates the problem using a latent Markov-process state variable to specify the segment to which a particular observation belongs. Inclan (1993) and Stephens (1994) formulate the problem assuming that the joint distribution of the parameters $\boldsymbol{\theta}_i$ is exchangeable and independent of the change points. Barry and Hartigan (1993) discuss the product partitioning model formulation mentioned in Section 2.2.1, while Loschi et al. (2010) propose

a modified version of their algorithm to sample from the posteriors of the parameters in a linear regression model, thus obtaining information beyond the posterior means. Chopin (2007) discuss a state space formulation using filtering and smoothing.

Estimation within a Bayesian framework can be computationally demanding if the number of change points is unknown, and it has deteriorating convergence properties as the length of the time series grows (Chib, 1998; Young Yang and Kuo, 2001; Chopin, 2007), but most approaches readily provide the user with an estimate of point-wise probabilities of a change occurring, $\Pr(t \in \mathbb{N})$. This is an attractive side-product from a practitioner's point of view, as it gives a gauge of confidence surrounding a change-point estimate. Markov-chain Monte Carlo (MCMC) is a popular tool for estimation, see e.g. Stephens (1994); Green (1995); Chib (1998); Lavielle and Lebarbier (2001) and Rosen et al. (2012). Stephens and Smith (1993) and Wang and Zivot (2000) discuss Gibbs sampling for change-point detection. For specific process families it is possible to derive fast recursive methods, as shown for instance in Fearnhead (2006) for processes with either a specified prior on number and, conditionally, location of change points, or a special case of the product partitioning model where a joint prior on number and locations of changes is specified; both processes assume independence of the parameters across segments and the detection method runs in $\mathrm{O}(T^2)$, or even $\mathrm{O}(T)$ in an approximate version. See also Xuan and Murphy (2007) for a multivariate version and Seidou and Ouarda (2007) for an adaptation to the multiple linear model. Ko et al. (2015) propose a MCMC algorithm to estimate their Dirichlet process hidden Markov model which, opposed to related earlier methods like Chib (1998), does not

require the specification of the number of change points in the central algorithm and thus does not require a second model-selection stage (see also Quintana and Iglesias, 2003, for the connection between Dirichlet process and the product partitioning model).

2.3 Brief General Remarks

2.3.1 Detection in Multivatiate Time Series

With the increased availability of large data sets and computational power to process them, multivariate time series have become more frequently subject of change-point analysis (Srivastava and Worsley, 1986; Ombao et al., 2005; Aue et al., 2009; Vert and Bleakley, 2010; Chen and Gupta, 2012; Horváth and Hušková, 2012; Bardwell and Fearnhead, 2014; Cho and Fryzlewicz, 2015a; Lung-Yut-Fong et al., 2015). The analysis of multivariate data with possibly many time series faces many well-known challenges, such as the curse of dimensionality for covariance estimation, or time displacement, i.e. a misalignment of time points of observation over different time series. The latter is relevant e.g. in asset price data traded in different time zones or functional magnetic resonance imaging in neuroscience.

For change-point detection, an additional concern is whether changes are assumed to occur in all time series at once; depending on the application at hand this may be overly restrictive. However, if this assumption is not made and a change is observed in a subset of the time series, questions arise such as how to consistently estimate the (partially time-varying) covariance structure. Moreover, with increas-

ing dimensionality the analyst can consider new forms of classifying changes in an importance ranking, e.g. by the number of time series they affect, or, inversely, classifying the time series based on their homogeneity of change points, which can be interpreted as a dependency measure.

2.3.2   Standard Assumptions

Change-point detection procedures require some assumptions on the change-point process, which generally depend on the detection problem at hand. Intuitively, to ensure identifiability of change points in time series at least two assumptions have to be made: firstly change points have to be sufficiently distant from their immediate neighbours and the boundary points $\eta_0 = 0, \eta_{N+1} = T$. Intuitively, two directly adjacent change points cannot be both identified as the segment between them has length zero. Secondly, changes have to be sufficiently pronounced, i.e. the minimum change size requires some positive lower bound.

These assumptions are necessary in any change-point detection problem. They are related, as illustrated e.g. in Chan and Walther (2013) for a simple piecewise-constant signal plus Gaussian noise situation, $x(t) = f(t) + \epsilon(t)$, $t = 1, \ldots, T$, $f(t) = \sum_{i=1}^{N+1} \mu_i \mathbb{I}(\eta_{i-1} < t \leq \eta_i)$. Here $\mathbb{I}(\cdot)$ denotes the indicator function and $\epsilon(t) \sim N(0, 1)$ is i.i.d. noise. For this process the conditions reduce to the relation between minimum change-point spacing $s_{\min} = \min_{i=1,\ldots,N+1}(\eta_i - \eta_{i-1})$ and minimum absolute change size $\delta_{\min} = \min_{i=2,\ldots,N+1} |\mu_i - \mu_{i-1}|$. For consistent detectability of changes the proportion of change size to spacing must be bounded

by

$$\delta_{\min}\sqrt{s_{\min}} \geq \frac{\sqrt{2\log 1/s_{\min}} + b_T}{\sqrt{T}}, \quad b_T \to \infty$$

This means that if we consider absolute change-point locations, i.e. $\eta_i/T \to 0$ and thus $s_{\min} \to 0$ as $T \to \infty$, change-point detection requires $\delta_{\min}\sqrt{s_{\min}} \geq (\sqrt{2} + \varepsilon_T)\sqrt{(\log 1/s_{\min})/T}$ with $\varepsilon_T \to 0$ provided that $\varepsilon_T\sqrt{\log 1/s_{\min}} \to \infty$.

If change-point locations are defined as fraction of the total time length T, i.e. $\eta_i/T \to c_i$ with c_i being constants in $(0,1)$, then $\lim_{T\to\infty} s_{\min} > 0$. It follows that any point with non-zero change can be estimated provided that $\delta_{\min} \geq b_T/\sqrt{T}$ with $b_T \to \infty$ as $T \to \infty$. Jandhyala et al. (2013) provides a comparison between the two definitions, absolute and rescaled change points, and an extensive literature review. In the following chapters, we use the rescaled-time definition unless stated otherwise.

2.3.3   Comparability of Change-Point Detection Methods

While there is no single best method for all change-point formulations and applications, methods can be compared based on general characteristics or with a specific detection problem in mind. This comparison can naturally be made in terms of properties such as test size or power or the rate of convergence to estimate the correct number of change points N and the change-point locations $\eta_i, i = \{1, \ldots, N\}$.

Other relevant dimensions of comparison are computational speed in terms of average and worst-case complexity, scalability of a method and the number and type of tuning parameters the user has to specify. Whether a small or large number of tuning parameters is preferable is application- and user-dependent, but

in general when a frequently deployed method requires time-intense calibration of the tuning parameter(s) this can be considered unattractive. Moreover, in the absence of clear guidance on the calibration it can raise suspicion that the end result suffers from an overfitting bias. Comparisons via empirical exercises can add to the relative assessment of a method's ability to detect change points in finite sample applications. These can include particularly application-relevant or realistic process specifications that are in some way challenging, e.g. with a high density of change points or very small change sizes.

We claim that the empirical success of any change-point detection method, including the ones discussed in this thesis, depends strongly on its appropriateness for a given application. Factors determining this success vary with the research question and data at hand, as much as with the analyst's ultimate goal, e.g. whether to detect change points in an online or retrospective analysis, whether to focus on time or frequency domain or whether additional information beyond change-point number and location is of interest. Among other things, the nature of data, most crucially the distributional form, the length of a time series and the sampling frequency as well as the dimensionality are of importance. Moreover, beliefs on the type and number of change points can be determinants of a successful change-point analysis.

2.4   Related Areas of Research

We already mentioned various subclasses of the change-point detection problem, such as the regime-switching processes which also receive extensive attention in

the econometrics literature (Gray, 1996; Hardy, 2001; Hamilton and Raj, 2013). However, change-point detection in itself can also be considered a subclass of other areas of research, notedly the signal recovery and representation literature (Rao and Kreutz-Delgado, 1999) and feature extraction (Guyon et al., 2008). Considering changes in the mean, a particularly strong link can be established with problems of trend detection, trend reversal estimation and trend filtering (Johansen and Sornette, 1999, 2000; Davies and Kovac, 2001; Wu et al., 2001; Kim et al., 2009; Tibshirani, 2014). Johansen and Sornette (1999, 2000), for instance, propose a Bayesian algorithm for applications to equity indices to capture trend reversal in time and evaluate the performance of their model successively at the out-of-sample financial crash in Japan in 1999. Davies and Kovac (2001) advertise the taut-string algorithm and multiscale decomposition to identify local extremes in a time series, which corresponds in their application to estimating piecewise linear trends between local peaks and lows using a total variation penalty. In Chapter 3 we draw parallels between this and related trend filtering methods and our change-point detection method for changes in the mean.

Change-point detection can be related to the broad field of feature extraction by interpreting it as specific form of a classification or clustering problem where the time-ordered or sequential structure of the data is accounted for. This becomes obvious in applications outside time series, such as image compression or DNA segmentation (Pal and Pal, 1993; Keogh et al., 2001; Lee and Lewicki, 2002). Chen et al. (2011) discuss a linear regression cluster detection when a change point can exist at a cluster-specific location. Arias-Castro et al. (2005) propose a method to detect geometric objects in noisy image data and discuss interesting features of the

detection problem in two dimensions, such as the concept of effective dimension which captures the multiplicity arising from testing in two or more dimensions.

Many approaches developed for network community detection also mirror ideas that we find in change-point detection. Basically, one can interpret a time series as network or graph where every observation is a node and has edges connecting it to the observations immediately before and after in time. The required segmentation can be based on similar technical arguments as in the time-ordered case, such as hierarchical (bottom-up or top-down) clustering (Friedman et al., 2001) or genetic algorithms (Liu et al., 2007). Parallels can also be drawn between community detection in networks and Bayesian change-point detection procedures, see e.g. Fortunato (2010). The author as well as Akoglu et al. (2015) provide overviews of literature on dynamic networks, which include networks undergoing sudden structural changes over time. For instance, Peel and Clauset (2014) define an online probabilistic learning framework for network change-point detection and introduce a detection approach. Marangoni-Simonsen and Xie (2015) propose change-point detection approaches for online community detection.

Chapter 3

Adaptive Trend Estimation in Financial Time Series via Multiscale Change-Point-Induced Basis Recovery

Low-frequency financial returns can be modelled as centered around piecewise-constant trend functions which change at certain points in time. We propose a new stochastic time series framework which captures this feature. The main ingredient of our model is a hierarchically-ordered oscillatory basis of simple piecewise-constant functions that is determined by change points, and hence needs to be estimated from the data. The resulting model enables easy simulation and provides interpretable decomposition of nonstationarity into short- and long-term components. The model permits consistent estimation of the multiscale change-point-induced basis via binary segmentation, which results in a variable-span moving-average estimator of the current trend, and allows for short-term forecasting of the average return.

Declaration   This chapter is in parts based on joint work with Piotr Fryzlewicz as published in Statistics and Its Interface (Schröder and Fryzlewicz, 2013).

3.1   Introduction

In this chapter, we consider the problem of statistical modelling and forecasting of daily financial returns based on past observations, but the methodology we propose will also be of relevance to financial data at other frequencies. More generally, it leads to a new generic approach to statistical time series analysis, via adaptive oscillatory bases induced by change points, which can be of interest in other fields of application beyond finance.

Given a time series $\mathbf{p} = \{p(t), t = 1, \ldots, T\}$ of daily speculative prices on risky financial instruments, such as equities, equity indices, commodities, or currency exchange rates, their daily returns $\mathbf{x} = \{x(t), t = 1, \ldots, T\}$ are defined by $x(t) = \ln(p(t)) - \ln(p(t-1))$. Forecasting future values of $\mathbf{x}$ based on its own past is of major interest to quantitative finance practitioners, but presents an extremely challenging task due to the perceived low predictive content of past returns with respect to the future. The importance and difficulty of the problem have led to the use of a large number of statistical and data analytic techniques to tackle it. In particular, we mention return forecasting based on time-varying regression parameters (Pesaran and Timmermann, 2002), traditional ARMA time series modelling with heteroscedastic innovations (Berkowitz and O'Brien, 2002; Garcia et al., 2005), methods stemming from the technical analysis of price series such as those based on moving average cross-overs, breakout and other technical signals (Katz and Mc-

Cormick, 2000), as well as various machine-learning techniques such as those based on support vector machines (Kim, 2003; Camci and Chinnam, 2008) and neural networks for return forecasting (Catalão et al., 2007; Kim and Shin, 2007).

Our approach rests on the observation that the logarithmic price $\ln(p(t))$ can meaningfully and interpretably be modelled as fluctuating around a trend which started at a certain unknown time in the past, having a positive or negative linear slope. The points in time at which the slope changes will be referred to as change points. The movements of $\ln(p(t))$ around the trend resemble random walk with heteroscedastic innovations. After differencing, this pattern translates to a piecewise-constant trend function in the return domain, plus serially uncorrelated deviations from it. Change points in the slope of the linear trend in $\ln(p(t))$, or alternatively in the magnitude of the piecewise-constant trend in $x(t)$, can be related to structural changes, coinciding for example with regulatory alterations, macroeconomic announcements, technological innovation or an economy's transition from recovery to recession or vice versa.

Trend detection in financial returns is a much-studied topic and a range of methods are widely applied in practice; in Section 2.4 we mention among others Davies and Kovac (2001) who advertise the taut-string algorithm and multiscale decomposition to identify local extremes in a time series. Basically, the taut-string algorithm identifies a piecewise linear function between local peaks and lows that minimizes the squared deviation with the data while accounting for the scaled sum of one-step absolute differences in the function. Cho and Fryzlewicz (2011) show that this corresponds to $\delta = 1$ in the CUSUM statistic of Equation (2.2) and is therefore not optimal in terms of power and precision of detection in the setting of

Brodsky and Darkhovsky (1993, ch. 3). Kim et al. (2009) introduce the L_1-trend filtering approach, which is similar to the taut-string algorithm in the sense that it seeks to minimize the squared deviation between data and a piecewise linear trend subject to an L_1-penalty. The authors consider a linear penalty based on the total absolute second-order difference of the trend function scaled by a user-specified penalty coefficient. Other approaches to trend detection are simple moving-average and other one-sided kernel smoothing of the returns, moving-average cross-overs at the level of logarithmic prices, local polynomial smoothing, spline smoothing and nonlinear wavelet shrinkage. These and other techniques are reviewed from a practitioner's perspective in Bruder et al. (2008).

One contribution of this chapter is to advocate a trend-detection methodology for financial returns that works by detecting change points in the return series and taking the current trend estimate to be the average return between the most recent estimated change point and the current time. This amounts to averaging over the current estimated interval of stationarity in the conditional mean; related but different adaptive procedures for volatility (as opposed to trend) estimation appeared e.g. in Fryzlewicz et al. (2006), Spokoiny (2009) and Čížek et al. (2009). The first stage of our procedure is the segmentation of the returns series. Although many of the available techniques for time series segmentation could be used, we advocate the use of binary segmentation and justify this choice below. An example of our model fit using one particular value of the threshold parameter is shown in Figure 3.1 based on daily closing values of the S&P 500 index.

However, the contribution of this chapter goes beyond merely advocating change-point detection as a useful approach to local trend estimation in finan-
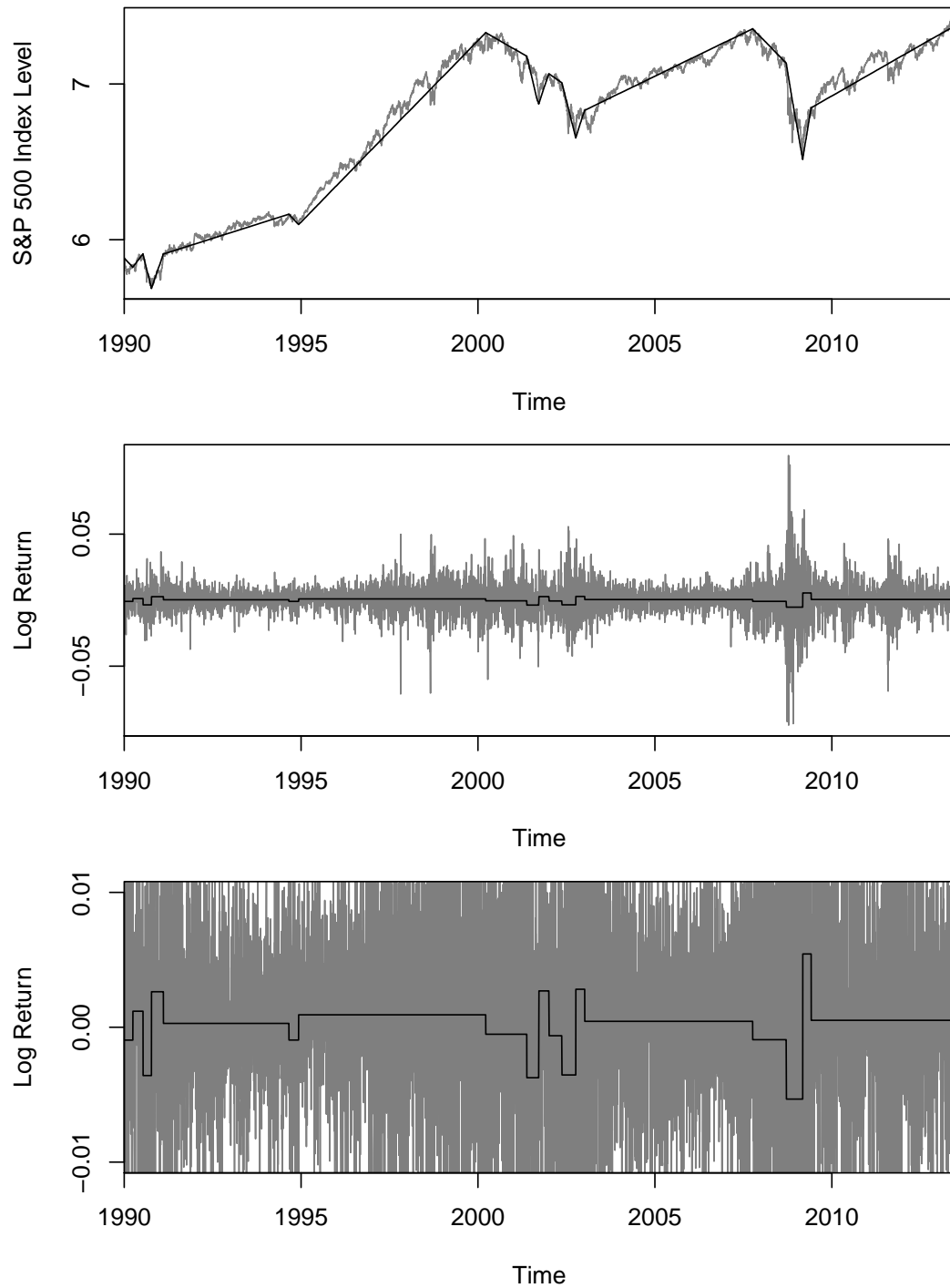
Figure 3.1: Daily closing values of the S&P 500 equity index between January 1990 and June 2013; top: log-price (grey) and the cumulatively integrated fit $\sum_{s=1}^{t} \hat{f}(s)$ from our model (black), middle: log-return (grey) and our model fit $\hat{f}(t)$ (black), bottom: the same as middle but with a shorter range on the y-axis. Model fit shown for threshold parameter $C = 0.3$ and imposing a minimum change-point distance of 60 days.

cial returns. Our main objective is to propose a new approach to statistical time series analysis, whereby the time series, generically denoted here by $\mathbf{x} = \{x(t), t = 1, \ldots, T\}$, is spanned by an orthonormal oscillatory basis induced by change points in the conditional mean value of $\mathbf{x}$. In this chapter, the representation basis is assumed to be unknown to the analyst and needs to be estimated from $\mathbf{x}$ using a change-point detection procedure; hence we will occasionally refer to such a basis as 'data-driven' or 'adaptive'. This in contrast to classical spectral approaches to time series analysis, which use a particular fixed basis that is known to the user, e.g. the Fourier basis in the classical spectral theory (Priestley, 1983), or a fixed wavelet system as in e.g. Nason et al. (2000) and Sharifzadeh et al. (2005). The SLEX method of Ombao et al. (2002), although also using a data-driven basis principle, is fundamentally different from ours in that it models the second-, not the first-order structure of $\mathbf{x}$ and is limited to change points occurring at dyadic locations. In financial applications, spectral analysis has occasionally been related to agents trading at different time horizons (see also Section 2.2.1, p. 41), although this point of view is of no primary relevance to us.

Our new adaptive basis approach to time series modelling opens up many interesting avenues. It leads to a formal stochastic time series model, used here to model local trends in financial returns but also applicable more widely, with interpretable nonstationarities in the autocovariance structure and in the conditional mean of $\mathbf{x}$. It allows for a decomposition of the nonstationarity in the variance of $\mathbf{x}$ into longer-term trends and short-term outbursts of volatility. Finally, it yields a family of forecasting operators for $\mathbf{x}$, parameterised by a single threshold parameter, which can be adjusted flexibly depending on the forecast horizon or the error

criterion.

This chapter is structured as follows. Section 3.2 motivates and defines the model and its building blocks, and studies its probabilistic properties. Section 3.3 describes the methodology and theory of change point detection and basis recovery, as well as the implied methodology for current trend estimation and forecasting. Section 3.4 illustrates basis recovery in a numerical study. Section 3.5 shows the estimated bases for data examples from various asset classes and performs a forecasting competition between our method and the benchmark moving window approach. We conclude with a brief discussion of a multivariate extension in Section 3.6. Proofs are in the appendix. For ease of reference, we provide a glossary of important terms at the end of this chapter (p. 103).

3.2 The Model

3.2.1 Motivation and Basic Ingredients

As illustrated in Figure 3.1, piecewise-linear modelling of trends in $\{\ln(p(t)), t = 1, \ldots, T\}$ results in the average of the returns series $\mathbf{x} = \{x(t), t = 1, \ldots, T\}$ oscillating around zero in a piecewise-constant fashion. We wish to embed this feature into a rigorous stochastic framework by formulating a time series model for $\mathbf{x}$ that captures this oscillatory behaviour.

Time series modelling using oscillatory building blocks is a well-established technique. In this chapter we do not consider the traditional Fourier basis decomposition as described in Equation (2.3), but rather employ a wavelet decomposition

technique. The building blocks used in our basis construction are the Unbalanced Haar (UH) wavelet vectors (Girardi and Sweldens, 1997; Delouille et al., 2001; Fryzlewicz, 2007; Baek and Pipiras, 2009; Timmermans et al., 2012), which have the advantage of being particularly simple, well-suited to the task of change-point modelling, and hierarchically organised into a multi-scale system, which is useful for the interpretability of the estimated change-point locations and basis vectors, and facilitates their arrangement according to their importance. We use the UH basis vectors to define the Unbalanced Haar time series model in Section 3.2.2.

Consider first the locally stationary wavelet model (Nason et al., 2000),

$$y(t) = \sum_{j=1}^{\infty} \sum_{k \in \mathbb{Z}} w_{j,k} \psi^j(t-k) \xi_{j,k}, \quad t \in [1, T], \tag{3.1}$$

where j is the scale parameter, analogous to frequency in (2.3), and $\boldsymbol{\psi}^j$ are compactly-supported wavelet vectors with elements $\{\psi^j(t), t = 1, \ldots, T\}$, oscillatory in the sense that $\sum_u \psi^j(u) = 0$, and such that the length of their support increases, but the speed of their oscillation decreases, with j. The parameters $w_{j,k}$ are amplitudes, localised over time-location k, and $\xi_{j,k}$ are mutually uncorrelated increments. We refer the reader to Vidakovic (2009) and Nason (2008) for overviews of the use of wavelets in statistical modelling.

The above approach could be used to model returns processes $\mathbf{x}$ such as that illustrated in Figure 3.1, but is not ideal. Our signal of interest, the piecewise-constant average return, is blocky, and oscillates around zero in an inhomogeneous fashion in the sense that there is significant variation in the lengths of the constant intervals. Therefore, similar to the complex exponentials in the Fourier representa-

tion, wavelet vectors arising from continuous functions are ill-suited for our purpose of modelling piecewise constancy. The exception that can reflect this blocky oscillatory behaviour is the Haar wavelet, as we discuss below. Also, crucially, both classical Fourier and wavelet decomposition techniques use bases which are not data-adaptive in the sense that they are fixed before the analysis rather than being tailored to, or estimated from, the data. In particular, one could possibly entertain the thought of using the piecewise-constant Haar wavelets for our purpose, but they would only permit change points at dyadic locations $kT2^{-j}$, $j = 1, 2, \ldots$, $k = 1, \ldots, 2^j - 1$, where T is the sample size.

In our data, change points occur at arbitrary locations and we hope to be able to capture this feature by the use of a suitably flexible oscillatory basis that permits adaptive choice of change points in the basis vectors, allowing for a sparse representation of the piecewise-constant trend. Arguably the simplest such construction is furnished by the Unbalanced Haar wavelets. With $\mathbb{I}(\cdot)$ denoting the indicator function, the elements of a generic UH vector $\boldsymbol{\psi}^{s,b,e}$ are

$$
\begin{aligned}
\psi^{s,b,e}(t) \quad = \quad & \left\{ \frac{1}{b - s + 1} - \frac{1}{e - s + 1} \right\}^{1/2} \mathbb{I}(s \le t \le b) \\
& - \left\{ \frac{1}{e - b} - \frac{1}{e - s + 1} \right\}^{1/2} \mathbb{I}(b + 1 \le t \le e),
\end{aligned}
$$

where s and e are, respectively, the start- and end-point of its support, and b is the location of a change point. $\boldsymbol{\psi}^{s,b,e}$ is constant and positive before the change point, constant and negative after the change point, and such that $\sum_t \psi^{s,b,e}(t) = 0$ and $\sum_t (\psi^{s,b,e}(t))^2 = 1$. If constructed as follows, a set of $(T - 1)$ UH vectors plus one constant vector constitutes an orthonormal basis of $\mathbb{R}^T$: define the first UH

basis vector $\boldsymbol{\psi}^{b_{0,1}} = \boldsymbol{\psi}^{1,b_{0,1},T}$; the change point $b_{0,1}$ needs to be chosen and we later say how. Then, repeat this construction following the binary segmentation logic on the two parts of the domain determined by $b_{0,1}$: provided that $b_{0,1} \geq 2$, define $\boldsymbol{\psi}^{b_{1,1}} = \boldsymbol{\psi}^{1,b_{1,1},b_{0,1}}$, and provided that $T - b_{0,1} \geq 2$, define $\boldsymbol{\psi}^{b_{1,2}} = \boldsymbol{\psi}^{b_{0,1}+1,b_{1,2},T}$. The recursion then continues in the same manner for as long as feasible, with each vector $\boldsymbol{\psi}^{b_{j,k}}$ having at most two children vectors $\boldsymbol{\psi}^{b_{j+1,2k-1}}$ and $\boldsymbol{\psi}^{b_{j+1,2k}}$. Additionally, we define a vector $\boldsymbol{\psi}^{b_{-1,0}}$ with elements $\psi^{b_{-1,0}}(t) = T^{-1/2}\mathbb{I}(1 \leq t \leq T)$. The indices j and k are scale and location parameters, respectively. The larger the value of j, the finer the scale, as in the classical wavelet theory.

Example. We consider an example of a set of UH vectors for $T = 6$. The rows of the matrix $\mathbf{W}$ defined below contain (from top to bottom) vectors $\boldsymbol{\psi}^{b_{-1,0}}$, $\boldsymbol{\psi}^{b_{0,1}}$, $\boldsymbol{\psi}^{b_{1,2}}$, $\boldsymbol{\psi}^{b_{2,3}}$, $\boldsymbol{\psi}^{b_{2,4}}$ and $\boldsymbol{\psi}^{b_{3,7}}$ determined by the following array of change points: $(b_{0,1}, b_{1,2}, b_{2,3}, b_{2,4}, b_{3,7})' = (1, 3, 2, 5, 4)'$.

$$
\mathbf{W} = \begin{pmatrix}
1/\sqrt{6} & 1/\sqrt{6} & 1/\sqrt{6} & 1/\sqrt{6} & 1/\sqrt{6} & 1/\sqrt{6} \\
\sqrt{5/6} & -1/\sqrt{30} & -1/\sqrt{30} & -1/\sqrt{30} & -1/\sqrt{30} & -1/\sqrt{30} \\
0 & \sqrt{3/10} & \sqrt{3/10} & -\sqrt{2/15} & -\sqrt{2/15} & -\sqrt{2/15} \\
0 & 1/\sqrt{2} & -1/\sqrt{2} & 0 & 0 & 0 \\
0 & 0 & 0 & 1/\sqrt{6} & 1/\sqrt{6} & -\sqrt{2/3} \\
0 & 0 & 0 & 1/\sqrt{2} & -1/\sqrt{2} & 0
\end{pmatrix}
$$

In the above example it is not possible to create further vectors $\boldsymbol{\psi}^{b_{j,k}}$; $\mathbf{W}$ constitutes an orthonormal basis of $\mathbb{R}^6$.

3.2.2 Definition and Examples

Motivated by the above discussion, our model for the returns series $\mathbf{x}$ is defined as follows.

Definition 3.2.1 *A stochastic process* $\mathbf{x} = \{x(t), t = 1, \ldots, T\}$ *is called* Unbalanced Haar *process if it has a representation*

$$x(t) = T^{1/2} \sum_{(j,k) \in \mathcal{I}} A_{j,k} \psi^{b_{j,k}}(t) + \sigma(t)\varepsilon(t), \quad t = \{1, \ldots, T\}, \qquad (3.2)$$

where $\mathcal{I}$ *is a set of indices, of finite dimensionality* $|\mathcal{I}| = N + 1 < \infty$, *such that* $(-1, 0) \in \mathcal{I}$, *and connected in the sense that if a child index* $(j + 1, 2k - 1)$ *or* $(j + 1, 2k)$ *is in* $\mathcal{I}$, *then so is their parent* (j, k). *The random variables* $\{A_{j,k}, (j, k) \in \mathcal{I}\}$ *are mutually independent, drawn from continuous distributions and satisfy* $E(A_{j,k}) = 0$ *and* $E(A_{j,k}^2) < \infty$. *The vectors* $\boldsymbol{\psi}^{b_{j,k}}$ *are UH vectors defined in Section 3.2.1. The constants* $\sigma(t)$ *are such that* $0 < \underline{\sigma} < \sigma(t) < \overline{\sigma} < \infty$, *and* $\{\varepsilon(t), t = 1, \ldots, T\}$ *is a sequence of independent standard normal variables, also independent of* $A_{j,k}$.

Denote the elements of the set $\{b_{j,k}, (j, k) \in \mathcal{I}\}$, sorted in increasing order, by $\mathbb{N} = \{\eta_i, i = 1, \ldots, N\}$. For completeness, denote $\eta_0 = 0$, $\eta_{N+1} = T$. We assume that the points η_i are fixed in rescaled time in the sense that for each i, we have $\eta_i = \lfloor T v_i \rfloor$, $\mathbb{V} = \{v_i, i = 1, \ldots, N\}$ is an increasing sequence of constants in $(0, 1)$ and $v_0 = 0$, $v_{N+1} = 1$.

The Unbalanced Haar process (UHP) defined in 3.2.1 contains two additive parts, interpretable, respectively, as signal (or trend) and noise. The signal,

$T^{1/2} \sum_{(j,k) \in \mathcal{I}} A_{j,k} \psi^{b_{j,k}}(t)$, is designed to model the piecewise-constant average return and provides a multi-scale representation of this quantity in terms of the basis functions $\boldsymbol{\psi}^{b_{j,k}}$. Heuristically speaking, it is composed of a constant vector $\boldsymbol{\psi}^{b-1,1}$ plus N UH vectors $\boldsymbol{\psi}^{b_{j,k}}$, each multiplied by its own independent amplitude $A_{j,k}$. This mimics the construction of the piecewise stationary Fourier process in Equation (2.3) and the locally stationary wavelet process in Equation (3.1), where the process in question is also composed of oscillatory vectors at different frequencies or scales, with random amplitudes. The fact that $\mathcal{I}$ is connected leads to $T^{1/2} \sum_{(j,k) \in \mathcal{I}} A_{j,k} \psi^{b_{j,k}}(t)$ being a random, piecewise-constant signal with N change points located at η_i, $i = 1, \ldots, N$. As in (2.3), the basis vectors $\boldsymbol{\psi}^{b_{j,k}}$ also change with the sample size, but our notation does not reflect this for simplicity. The factor $T^{1/2}$ is required to keep the scale of the amplitudes $A_{j,k}$ constant with respect to T.

The noise, $\sigma(t)\varepsilon(t)$, models the random movements of $x(t)$ around the trend and, for technical simplicity, is assumed to be Gaussian. In this work, we do not dwell on the issue of estimating the volatility parameters $\sigma(t)$, but treat them as constant in our theoretical considerations. Naturally, in practice, they need to be estimated from the data; we later specify what estimators we propose.

The UHP of Equation (3.2) enables easy simulation of sample paths of $x(t)$ as we illustrate in Figure 3.2. The simulated sample paths of $x(t)$ from the UHP use the canonical basis (see Section 3.3.2 for details). $\sigma(t)$ is estimated from the real data example from Figure 3.1, and the coefficients $A_{j,k}$ are drawn from the normal distribution with mean zero and variances matching the corresponding empirical variances from the data.
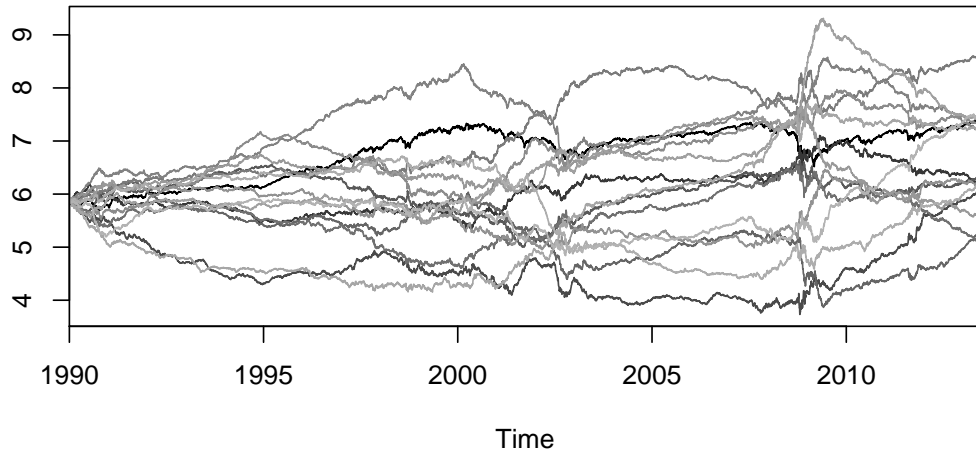
Figure 3.2: Daily closing values of the S&P 500 equity index (log-price, black line) between January 1990 and June 2013 and simulated sample paths (grey lines) from the estimated canonical basis $\{\hat{b}_{j,k}, (j,k) \in \hat{\mathcal{I}}\}$ (see Section 3.3); the coefficients $A_{j,k}$ are drawn from the normal distribution with mean zero and variances matching the corresponding empirical variances $\sigma(t)$ as estimated from the data. The model is estimated with a threshold parameter $C = 0.3$ and a minimum change-point distance of 60 days.

3.2.3 Unconditional Properties of the Model

We start by discussing some simple unconditional probabilistic properties of the Unbalanced Haar process. We use the term unconditional to mean that we do not condition on particular values of the random coefficients $A_{j,k}$. Were we to do this, our analysis would amount to considering a particular, rather than random, piecewise-constant signal plus noise, a set-up which we investigate in Section 3.2.4 below.

$\mathrm{E}(A_{j,k}) = 0$ and $\mathrm{E}(\varepsilon(t)) = 0$ imply $\mathrm{E}(x(t)) = 0$. The variance and autocovari-

ance of $x(t)$ admit the following decompositions:

$$\mathrm{Var}(x(t)) \quad = \quad T \sum_{(j,k)\in\mathcal{I}} \mathrm{E}(A_{j,k}^2)(\psi^{b_{j,k}}(t))^2 + (\sigma(t))^2 \qquad (3.3)$$

$$\mathrm{Cov}(x(t), x(t+\tau)) \quad = \quad T \sum_{(j,k)\in\mathcal{I}} \mathrm{E}(A_{j,k}^2)\psi^{b_{j,k}}(t)\psi^{b_{j,k}}(t+\tau), \ \tau \neq 0. \qquad (3.4)$$

A few remarks are in order.

Stationarity. Clearly, $x(t)$ is stationary in the mean, but is variance- and covariance-nonstationary. The nonstationarity in the variance arises not just because of the $\sigma(t)$ term, but also because of the term $T\sum_{(j,k)\in\mathcal{I}} \mathrm{E}(A_{j,k}^2)(\psi^{b_{j,k}}(t))^2$, which captures the variability of longer-term trends, as opposed to the daily variability captured by $\sigma(t)$. The daily variability $\sigma(t)$ could also, in principle, be modelled as a GARCH-type process with a conditional, rather than unconditional, heteroscedasticity, but we do not pursue this option in this work because of the technical requirement of the boundedness of $\sigma(t)$ from above.

Variance. Firstly, we observe that $\psi^{b_{j,k}}(t) = 0$ if t is outside the support of $\boldsymbol{\psi}^{b_{j,k}}$, and the values of $\psi^{b_{j,k}}(t)$ differ depending on whether t falls within the positive or the negative part of the support of $\boldsymbol{\psi}^{b_{j,k}}$; the shorter the relevant part of the support, the higher the value of $(\psi^{b_{j,k}}(t))^2$, and hence Equation (3.3) indicates that shorter trends contribute more to the variability of $x(t)$. In simple heuristic terms, this can be interpreted as shorter-term trends being more variable. Secondly, due to the fact that $\sum_t(\psi^{b_{j,k}}(t))^2 = 1$ the average variance admits the following

decomposition:

$$T^{-1}\sum_{t=1}^{T}\mathrm{Var}(x(t)) = \sum_{(j,k)\in\mathcal{I}}\mathrm{E}(A_{j,k}^2) + T^{-1}\sum_{t=1}^{T}(\sigma(t))^2,$$

Therefore, the average variance of the signal has a simple representation in terms of the variances of the amplitudes $A_{j,k}$.

Autocovariance.   If both t and $t+\tau$ are within the same, positive or negative, part of the support of $\boldsymbol{\psi}^{b_{j,k}}$, then $\psi^{b_{j,k}}(t)\psi^{b_{j,k}}(t+\tau)$ reduces to $(\psi^{b_{j,k}}(t))^2$, and by Equations (3.3) and (3.4), the contribution of the term indexed (j,k) to $\mathrm{Var}(x(t))$ and $\mathrm{Cov}(x(t),x(t+\tau))$ is the same and equal to $\mathrm{E}(A_{j,k}^2)(\psi^{b_{j,k}}(t))^2$. On the other hand, if t and $t+\tau$ are within the support of $\boldsymbol{\psi}^{b_{j,k}}$ but on two different sides of the change point $b_{j,k}$, then the term $\mathrm{E}(A_{j,k}^2)\psi^{b_{j,k}}(t)\psi^{b_{j,k}}(t+\tau)$ contributes negatively to $\mathrm{Cov}(x(t),x(t+\tau))$, which reflects the fact that t and $t+\tau$ belong to two opposing trends at scale j.

Unbalanced Haar spectrum.   By Equations (3.3) and (3.4), the term $T\sum_{(j,k)\in\mathcal{I}}\mathrm{E}(A_{j,k}^2)(\psi^{b_{j,k}}(t))^2$ in the variance of $\mathbf{x}$ as well as the autocovariance of $\mathbf{x}$ have a representation in terms of $\mathrm{E}(A_{j,k}^2),(j,k)\in\mathcal{I}$. Therefore, it is natural to introduce a separate definition for this quantity.

Definition 3.2.2 *Let* $\mathbf{x} = \{x(t), t = 1,\ldots,T\}$ *follow model (3.2). The sequence* $E(A_{j,k}^2), (j,k) \in \mathcal{I}$ *is referred to as the* Unbalanced Haar spectrum *of* $\mathbf{x}$ *with respect to the basis* $\{\boldsymbol{\psi}^{b_{j,k}}, (j,k) \in \mathcal{I}\}$.

We emphasise that the Unbalanced Haar spectrum is defined with respect to a fixed basis $\{\boldsymbol{\psi}^{b_{j,k}}, (j,k) \in \mathcal{I}\}$; a different basis for the same process $\mathbf{x}$ would result in a different Unbalanced Haar spectrum. In particular, if the Unbalanced Haar spectral approach is used to compare two or more time series, a common basis should be used in order for the comparison to be meaningful, unless it is of interest to study feature misalignment between the different series. Section 3.6 discusses a multivariate extension of the Unbalanced Haar process and how a common basis can be chosen. For univariate Unbalanced Haar processes, we discuss the important issue of basis selection below and propose a data-adaptive approach to discovering a unique basis for an observed time series $\mathbf{x}$.

The Unbalanced Haar spectrum is an analogue of the spectral density in the Fourier representation of Equation (2.3) and the wavelet spectrum in the LSW model of Equation (3.1) in the sense that it also arises as the sequence of variances of the random amplitudes associated with the oscillatory building blocks used in the construction of the process. However, the role of the UH building blocks in Equation (3.2) is subtly different from those of the Fourier exponentials in Equation (2.3) and wavelets in Equation (3.1): the latter two directly model the contribution of the corresponding oscillations to the autocovariance structures of the respective processes, while our primary aim in using the UH building blocks is to model the piecewise-constant trend in the process $\mathbf{x}$ rather than represent the autocovariance of $\mathbf{x}$. The fact that the autocovariance structure of $\mathbf{x}$ also has a representation in terms of $\boldsymbol{\psi}^{b_{j,k}}$ is nothing but a useful and interpretable by-product of this modelling approach.

If the basis $\{\boldsymbol{\psi}^{b_{j,k}}, (j,k) \in \mathcal{I}\}$ is known to the analyst, then the Unbalanced Haar periodogram, defined below, provides an asymptotically unbiased but inconsistent estimate of the Unbalanced Haar spectrum, in the same way as the classical periodogram is asymptotically unbiased but inconsistent for the spectral density in the classical Fourier theory. The case of an unknown basis is more delicate and will be discussed in more detail in Sections 3.2.4 and 3.3.2, as will be the issue of restoring, in a certain sense, the consistency of spectral estimation in Definition 3.2.1. In the remainder of this work, $\langle \cdot, \cdot \rangle$ denotes inner product between two vectors.

Definition 3.2.3 *Let* $\mathbf{x} = \{x(t), t = 1, \ldots, T\}$ *follow model (3.2). The sequence of statistics defined by*

$$I_{j,k} = T^{-1} \langle \mathbf{x}, \boldsymbol{\psi}^{b_{j,k}} \rangle^2 = T^{-1} \left(\sum_{t=1}^{T} x(t) \psi^{b_{j,k}}(t) \right)^2 \tag{3.5}$$

is called the Unbalanced Haar periodogram *of* $\mathbf{x}$ *with respect to the basis* $\{\boldsymbol{\psi}^{b_{j,k}}, (j,k) \in \mathcal{I}\}$.

The following result quantifies the asymptotic unbiasedness of $I_{j,k}$ for $\mathrm{E}(A_{j,k}^2)$.

Proposition 3.2.1 *We have*

$$| E(I_{j,k}) - E(A_{j,k}^2)| \leq \overline{\sigma}^2 T^{-1}.$$

3.2.4 Properties of the Model Conditional on $A_{j,k}$

Having observed a sample path of $\mathbf{x}$, it is of interest to establish the number and locations of change points in the observed signal $T^{1/2} \sum_{(j,k) \in \mathcal{I}} A_{j,k} \psi^{b_{j,k}}(t)$. From the

point of view of this task, it is helpful to treat $A_{j,k}$ as already observed, and therefore carry out the change-point analysis conditioning on their values $A_{j,k} = a_{j,k}$, for all $(j,k) \in \mathcal{I}$. Define $f(t) = T^{1/2} \sum_{(j,k) \in \mathcal{I}} a_{j,k} \psi^{b_{j,k}}(t)$, suppressing the dependence of $f(t)$ on $a_{j,k}$ and $b_{j,k}$, for simplicity of notation. In fact, neither $a_{j,k}$ nor $b_{j,k}$ are observed directly, but instead we observe a noisy version of $\mathbf{f} = \{f(t), t = 1, \ldots, T\}$, that is $\mathbf{x}$ with

$$x(t) = f(t) + \sigma(t)\varepsilon(t), \tag{3.6}$$

as in Equation (3.2) conditional on $A_{j,k} = a_{j,k}$. We now briefly discuss some properties of $\mathbf{f}$. Let P^A be the probability measure induced by the random variables $A_{j,k}, (j,k) \in \mathcal{I}$. The signal $\mathbf{f}$ satisfies the following properties.

Property 3.2.1   (i) $\mathbf{f}$ is piecewise-constant with at most N change points, and bounded.

(ii) The magnitude of the ith change point in $\mathbf{f}$ is of the form $|\sum_{j,k} \alpha_{j,k,i} a_{j,k}|$, where $\alpha_{j,k,i}$ are scalars.

(iii) Let A_0 be the set of $(N+1)$-tuples of those values of $a_{j,k}, (j,k) \in \mathcal{I}$, for which $\mathbf{f}$ has fewer than N change points. We have $P^A(A_0) = 0$.

(iv) With probability one with respect to P^A, change points in $\mathbf{f}$ are located at $b_{j,k}$; sorted in increasing order, their locations are $\{\eta_i\}_{i=1}^N$.

Naturally, conditioning on $A_{j,k}$ changes the first- and second-order properties of $\mathbf{x}$. From (3.6), we have $\mathrm{E}(x(t)|A_{j,k} = a_{j,k}) = f(t)$, $\mathrm{Var}(x(t)|A_{j,k} = a_{j,k}) = (\sigma(t))^2$ and $\mathrm{Cov}(x(t), x(t+\tau)|A_{j,k} = a_{j,k}) = 0$ for $\tau \neq 0$.

The following section discusses the estimation of various aspects of the conditional signal $\mathbf{f}$.

3.3 Estimation and Forecasting

3.3.1 Change-Point Detection

Given observations from the conditional model (3.6), it is of interest to estimate the number N of change points in $\mathbf{f}$ and the set of their locations, $\mathbb{N} = \{\eta_i, i = 1, \ldots, N\}$. Multiple change-point detection has been widely studied in literature, see Section 2.2.2 for examples. We propose the use of the binary segmentation (BS) method in this chapter because it is a natural tool for the estimation of UH basis vectors from the data due to the hierarchical structure of this procedure, cf. Fryzlewicz (2007). Via BS, we simultaneously estimate the canonical basis, which is discussed in Section 3.3.2 below. Another benefit of binary segmentation is rate optimality for the estimators of change-point locations in the particular setting of the spacing between change points being of order $\mathrm{O}(T)$, as is the case in our model. However, we emphasise that other change-point detection methods could also be used for this purpose.

For clarity, we describe below in pseudocode how BS is used to recursively

discover change points in the Unbalanced Haar process of Equation (3.2). Let

$$
\begin{aligned}
\tilde{X}_{s,e}^{b} &= \langle \mathbf{x}, \boldsymbol{\psi}^{s,b,e} \rangle \\
\tilde{\sigma}_{s}^{e} &= \left(\frac{1}{e-s+1} \sum_{t=s}^{e} (\sigma(t))^2 \right)^{1/2} .
\end{aligned}
$$

The main function is defined in Figure 3.3. The standard UHP-BS procedure is launched by the call UHP-BS($\mathbf{x}$, 1, T, ζ_T), where ζ_T is a threshold parameter. We have the following result regarding the consistency of the UHP-BS procedure for the number and locations of change points.

function UHP-BS($\mathbf{x}$, s, e, ζ_T)
    if $e - s < 1$ **then**
        STOP
    else
        $b_0 := \mathrm{argmax}_b |\tilde{X}_{s,e}^{b}|$
        if $|\tilde{X}_{s,e}^{b_0}|/\tilde{\sigma}_{s}^{e} > \zeta_T$ **then**
            add b_0 to the set of estimated change points
            UHP-BS($\mathbf{x}$, s, b_0, ζ_T)
            UHP-BS($\mathbf{x}$, $b_0 + 1$, e, ζ_T)
        else
            STOP
        end if
    end if
end function

Figure 3.3: UHP-BS algorithm

Theorem 3.3.1 *Let* $\mathbf{f} = \{f(t), t = 1, \ldots, T\}$ *in Equation (3.6) be constructed with parameters* $a_{j,k}$ *lying outside set* A_0 *from Property 3.2.1. Let* N *and* $\eta_1, \ldots, \eta_N$ *denote, respectively, the number and locations of change points in* $\mathbf{f}$*. Let* $\hat{N}$ *denote the number, and* $\hat{\eta}_1, \ldots, \hat{\eta}_N$ *the locations, sorted in increasing order, of the change-point estimates obtained by the UHP-BS algorithm. Let the threshold parameter satisfy* $\zeta_T \geq c_1 \log^p T$ *(p > 1/2) and* $\zeta_T \leq c_2 T^\theta$ *($\theta < 1/2$), for any positive constants* c_1, c_2*. Then there exists a positive constant* C_1 *such that* $P(\mathcal{A}_T) \to 1$*, where*

$$\mathcal{A}_T = \{\hat{N} = N; \quad \max_{i=1,\ldots,N} |\hat{\eta}_i - \eta_i| \leq C_1 \epsilon_T\}$$

with $\epsilon_T = O(\log T)$*, where* $P(\cdot)$ *is the probability measure induced by* $\{\varepsilon(t), t = 1, \ldots, T\}$*.*

3.3.2 Basis Recovery

Given any piecewise-constant signal $\mathbf{f} = \{f(t), t = 1, \ldots, T\}$ with $N > 1$ change points in $\mathbb{N}$, there are multiple ways to represent it in a UH basis $\{\boldsymbol{\psi}^{b_{j,k}}, (j, k) \in \mathcal{I}\}$, as any of the N change points can be assigned to scale $j = 0$; that is, there are N mutually exclusive possibilities: $b_{0,1} = \eta_1, b_{0,1} = \eta_2, \ldots, b_{0,1} = \eta_N$. If $N > 2$, then there is similar choice at the following scales $j > 0$. Some of the possible UH bases are more interpretable and useful than others. Here, we define and focus on one particular type of UH basis, termed the *canonical UH basis*, which enables the partial ordering of the set of change points, $\mathbb{N}$, according to what can be interpreted as their importance.

Recall the construction of any UH basis, described in Section 3.2.1. Given a

current interval of interest $\{s, \ldots, e\}$, a basis selection procedure will be completely specified if one specifies how to choose $b_{j,k}$ on that interval. The canonical basis is defined as follows.

Definition 3.3.1 *Given a piecewise-constant signal* $\mathbf{f} = \{f(t), t = 1, \ldots, T\}$ *with* $N > 0$ *change points at* $\eta_1, \ldots, \eta_N$, *a* canonical Unbalanced Haar basis *is a basis for which* $b_{j,k}$ *on the current interval* $[s, e]$ *(on which* $\mathbf{f}$ *is non-constant) is assigned as*

$$b_{j,k} = \arg\max_{s < \eta_i < e, i = 1, \ldots, N} |\langle \mathbf{f}, \boldsymbol{\psi}^{s, \eta_i, e} \rangle|,$$

If there is more than one such bases, any of them is referred to as canonical.

In other words, applying Definition 3.3.1 sequentially, a canonical basis for $\mathbf{f}$ is chosen as follows. At scale 0, $b_{0,1}$ is chosen as the η_i that maximises $|\langle \mathbf{f}, \boldsymbol{\psi}^{1, \eta_i, T} \rangle|$; that is, the η_i which defines the step function with one change point that fits $\mathbf{f}$ best in the least-squares sense. Such η_i is not necessarily unique; however, ties are not an issue for us as we clarify further below. Having identified $b_{0,1}$, canonical basis selection then proceeds sequentially as described in Section 3.2.1, at each scale fitting the best approximation in the L_2-sense to $\mathbf{f}$ on the relevant interval $\{s, \ldots, e\}$, by means of a step function with one change point, until all change points in $\mathbf{f}$ have been accounted for.

We now give a result that specifies the uniqueness of a canonical basis for $\mathbf{f}$ from model (3.6).

Proposition 3.3.1 *There is a set* B_0 *of* $(N + 1)$-*tuples* $\{a_{j,k}, (j, k) \in \mathcal{I}\}$ *with* $P^A(B_0) = 0$ *such that for* $\{a_{j,k}\}$ *from outside* B_0, $\mathbf{f}$ *from Equation (3.6) has a*

unique canonical basis.

Summarising the above discussion, for a signal $\mathbf{f}$ defined by model (3.6), there exists a unique canonical UH basis for it with probability 1 with respect to P^A. The unique canonical basis partially orders the change points in $\mathbf{f}$ according to their importance in terms of explaining $\mathbf{f}$ in the L_2-sense. The change point $b_{0,1}$ can be interpreted as the most important, with change points as subsequent finer scales being interpretable as gradually less important.

In this setting, we now state that it is possible to use binary segmentation to reconstruct the canonical basis $\{b_{j,k}\}$ of $\mathbf{f}$ from the noisy observations defined by Equation (3.6), as well as the corresponding canonical basis coefficients $\bar{a}_{j,k}$. The latter fact can be interpreted as the conditional consistency of spectral estimation in Equation (3.2) with respect to each unique canonical basis.

Theorem 3.3.2 *Let $\mathbf{f}$ in Equation (3.6) be constructed with parameters $a_{j,k}$ lying outside set $A_0 \cup B_0$. Let N and $\{\eta_1, \ldots, \eta_N\}$ denote, respectively, the number and locations of change points in $\mathbf{f}$, and let $b_{j,k}$ define the canonical basis of $\mathbf{f}$, ordered according to increasing j, and with the $b_{j,k}$'s within each scale j sorted in increasing order. Let $\bar{a}_{j,k}$ be the UH coefficients with respect to $b_{j,k}$; that is, $\bar{a}_{j,k} = T^{-1/2}\langle \mathbf{f}, \boldsymbol{\psi}^{b_{j,k}}\rangle$. Let $\hat{N}$ denote the number, and $\{\hat{\eta}_1, \ldots, \hat{\eta}_N\}$ the locations, sorted in increasing order, of the change-point estimates obtained by the UHP-BS algorithm, and let $\hat{b}_{j,k}$ be the estimated change points in the order returned by the UHP-BS algorithm. Let $\hat{a}_{j,k} = T^{-1/2}\langle \mathbf{x}, \boldsymbol{\psi}^{\hat{b}_{j,k}}\rangle$. Let the threshold parameter satisfy $\zeta_T \geq c_1 \log^p T$ ($p > 1/2$) and $\zeta_T \leq c_2 T^\theta$ ($\theta < 1/2$), for any positive constants c_1, c_2.*

Then there exist positive constants C_1, C_2 such that $P(\mathcal{A}_T) \to 1$, where

$$
\begin{aligned}
\mathcal{A}_T \;=\; \{\hat{N} = N; \quad &\max_{(j,k)\in\mathcal{I}} |\hat{b}_{j,k} - b_{j,k}| \leq C_1 \epsilon_T; \\
&\max_{(j,k)\in\mathcal{I}} |\hat{a}_{j,k} - \bar{a}_{j,k}| \leq C_2 T^{-1/2} \log^{1/2} T\}
\end{aligned}
$$

with $\epsilon_T = O(\log T)$.

The result of Theorem 3.3.2 goes one step further than the change-point detection result of Theorem 3.3.1. It states that it is not only possible to detect the number and location of the change points, but also their importance as defined by the partial ordering specified by the canonical basis, as well as the coefficient values with respect to that basis. Another way of viewing this result is that the canonical basis $\{\boldsymbol{\psi}^{b_{j,k}}, (j,k) \in \mathcal{I}\}$ can be estimated from the noisy observations $\mathbf{x}$. The implication of this result is that binary segmentation can be used to estimate not only the change points in $\mathbf{f}$, but also their relative importance with respect to each other. This is attractive from the point of view of both the interpretability of the detected change points, and their use in forecasting, as we demonstrate further below.

We end this section by remarking that although it is tempting to attempt to define a "canonical UH spectrum" in Equation (3.2), this does not appear to be a straightforward task. The reason for this is that the canonical basis is defined for each signal $\mathbf{f}$ separately, and therefore two different sets of $\{a_{j,k}\}$, leading to two different realisations of $\mathbf{f}$, can result in two different canonical bases. Therefore, there is no such thing as a single specific canonical basis for the random generator $T^{1/2} \sum_{(j,k)\in\mathcal{I}} A_{j,k} \psi^{b_{j,k}}(t)$ of signal $\mathbf{f}$. Hence, defining a canonical UH spectrum,

being the set of variances of $A_{j,k}$ with respect to a canonical basis, is not particularly obvious here.

3.3.3 Forecasting

We define our forecasting task in Equation (3.2) as follows. Having observed $\mathbf{x}$, we are interested in predicting the value of $f(T+h)$, the conditional mean of $x(T+h)$, where $h \geq 1$ is the forecasting horizon. However, we note that if h is small with respect to T, which we formally quantify as $h = \mathrm{o}(T)$, then our model guarantees that $f(T+h) = f(T)$. This is because the change points locations η_i in our model satisfy $\eta_i = \lfloor Tv_i \rfloor$ where the v_i's are constants, so the next change point after η_N is not expected until $T(1+\delta) > T+h$, with δ being a positive constant. Therefore, the task of forecasting $f(T+h)$ in model (3.2) is equivalent to the task of estimating $f(T)$.

The condition $f(T) = f(T+h)$ for $h = \mathrm{o}(T)$ carries an implicit assumption that the asset modelled by $\mathbf{x} = \{x(t), t = 1, \ldots, T\}$ is of a *trend following* type: we assume that the current average return $f(T)$ will not change in the near future. However, our framework may also be a useful starting point for the modelling and forecasting of *mean-reverting* assets, in which $f(T+h)$ is likely to be different from $f(T)$; negative if $f(T) > 0$, or vice versa. In the trend-reversal situation, one possibility is to assume that $f(T+h) = g(f(T))$ where $g(\cdot)$ is possibly different from identity and needs to be estimated from the data. Although rigorous treatment of this case is beyond the scope of this work as it is not technically covered by our modelling framework, we discuss it from the practical point of view in Section 3.5.

Representing $\mathbf{f} = \{f(t), t = 1, \ldots, T\}$ in its canonical basis, we obtain

$$f(t) = T^{1/2} \sum_{(j,k)\in\mathcal{I}} \bar{a}_{j,k}\psi^{b_{j,k}}(t) = \sum_{(j,k)\in\mathcal{I}} \langle \mathbf{f}, \boldsymbol{\psi}^{b_{j,k}}\rangle\psi^{b_{j,k}}(t).$$

A natural estimator for $f(t)$ is

$$\hat{f}(t) = T^{1/2} \sum_{(j,k)\in\hat{\mathcal{I}}} \hat{a}_{j,k}\psi^{\hat{b}_{j,k}}(t) = \sum_{(j,k)\in\hat{\mathcal{I}}} \langle \mathbf{x}, \boldsymbol{\psi}^{\hat{b}_{j,k}}\rangle\psi^{\hat{b}_{j,k}}(t),$$

where $\hat{\mathcal{I}}$ is the set of estimated indices of the estimated change points; note that by Theorem 3.3.2, we have $\hat{\mathcal{I}} = \mathcal{I}$ with high probability. Therefore, $\hat{\mathbf{f}}$ is the orthogonal projection of the data $\mathbf{x}$ onto the space spanned by the estimated canonical UH vectors $\boldsymbol{\psi}^{\hat{b}_{j,k}}$. Hence, $\hat{f}(T)$ reduces to

$$\hat{f}(T) = \frac{1}{T - \hat{\eta}_N} \sum_{t=\hat{\eta}_N+1}^{T} x(t),$$

where $\hat{\eta}_N$ is the most recent estimated change point. Thus, $\hat{f}(T)$ can be interpreted as an adaptive, as opposed to fixed-span, moving average of the recent values of $\mathbf{x}$, where the adaptation is with respect to the estimated change-point structure in the data.

Our estimation procedure is parameterised by the threshold parameter ζ_T. The permitted theoretical range of ζ_T is specified in Theorem 3.3.2, and the practical choice of the constants in ζ_T is discussed in Section 3.5. The lower the value of ζ_T, the later the UHP-BS procedure is likely to stop, and therefore the closer to T the final detected change point $\hat{\eta}_N$ is likely to lie. Therefore, lower (higher) values of ζ_T are likely to lead to shorter (longer) average spans $T - \hat{\eta}_N$. In other words,

recalling the importance interpretation of the detected change points discussed earlier, higher values of ζ_T lead to forecasts based on more important detected change points, whereas the lower the value of ζ_T, the higher the chance of basing the forecasts on less important detected change points.

3.4 Simulation Study

In this section, we briefly exhibit the change-point detection and canonical basis recovery capabilities of the UHP-BS algorithm. We generate trends $\mathbf{f} = \{f(t), t = 1, \ldots, T\}$, with $T = 1000$, as follows. For each trend, the number N of change points is drawn from the Poisson(5) distribution and their locations η_i are drawn uniformly on $\{1, \ldots, T\}$. The jump sizes are simulated as independent normal variables with zero mean and variance V_{η_i}. We repeat the trend generation 100 times for $V_{\eta_i} = 1$ and 100 times for $V_{\eta_i} = 2$. We define the canonical basis of $\mathbf{f}$ by $\{b_{j,k}, (j,k) \in \mathcal{I}\}$. For each trend $\mathbf{f}$, we simulate 1000 sample paths of $x(t) = f(t) + \sigma(t)\varepsilon(t)$ with $\sigma(t) = 1$. The resulting sample paths tend to have low signal-to-noise ratios and are challenging from the point of view of change-point detection.

In the UHP-BS algorithm, we use the threshold $\zeta_T = \sqrt{C} \log^p T$ with $p = 1/2$, which is the lower end of the permitted theoretical range from Theorem 3.3.2. The study is repeated for $C = \{0.5, 0.75, 1, 1.25, 1.5, 2\}$. For robustness, we estimate the local volatility $\tilde{\sigma}_s^e$ using the Median Absolute Deviation estimator for the Gaussian distribution, both here and in the remainder of this chapter. The estimated number of change points is denoted by $\hat{N}$.

To judge the quality of change-point detection and basis recovery, we use three

statistics, one of which is $N - \hat{N}$. To gauge the distance between the estimated and true canonical bases, we take all those sample paths for which $\hat{N} = N$. For those sample paths, we define the 'Not Assigned' (NA) and 'Scale Difference' (SD) statistics, which measure, respectively, the number of estimated change points that cannot be assigned to any true one in terms of their location, and the sum of differences in scales between the estimated and true change points if assignment is possible.

The statistics are estimated via the following procedure. Define $\hat{\mathcal{B}}_{j,k} = \{\hat{b}_{i,l} : \hat{b}_{i,l} \in (b_{j,k} - \Delta_T, b_{j,k} + \Delta_T)\}$, where $\Delta_T = 5$, and let Υ denote the set of all estimated change points $\hat{b}_{j,k}$. Set SD $= 0$. For all (j, k), from coarser to finer scales and from left to right, if $|\hat{\mathcal{B}}_{j,k}| = 1$, then the matching for that (j, k) is completed. If $|\hat{\mathcal{B}}_{j,k}| > 1$, choose as the closest match the $\hat{b}_{i,l} \in \hat{\mathcal{B}}_{j,k}$ that minimises $|i - j|$. If there are multiple such $\hat{b}_{i,l}$'s, choose the one closest to $b_{j,k}$. Delete the matched estimated change point from Υ. Add $|i - j|$ to SD. After considering all (j, k)'s, set NA $= |\Upsilon|$.

Table 3.1 summarizes the results over various parameter specifications. As expected, $\hat{N}$ is closer to N if jump sizes have a larger variance V_{η_i}. If ζ_T is too small, many spurious change points are detected, while if it is too large, too few are identified. The average NA and SD measures are small in value, which shows the closeness of the true and estimated canonical bases, provided $\hat{N}$ is estimated correctly.

V_{η_i}	C	Prop $\hat{N} = N$	Prop $\hat{N}$ in $[N-1, N+1]$	Prop $\hat{N}$ in $[N-2, N+2]$	Avg $(N - \hat{N})$	Avg $(|N - \hat{N}|)$	Avg NA	Avg SD
	0.50	0.02	0.05	0.08	-15.47	15.52	2.29	0.29
	0.75	0.13	0.35	0.51	-3.15	3.75	2.21	0.40
1	1.00	0.22	0.56	0.76	-0.14	1.76	1.85	0.42
	1.25	0.22	0.57	0.78	0.96	1.57	1.49	0.38
	1.50	0.20	0.51	0.74	1.47	1.71	1.18	0.33
	2.00	0.16	0.42	0.66	1.97	2.02	0.83	0.25
	0.50	0.01	0.03	0.05	-17.15	17.17	1.20	0.15
	0.75	0.12	0.29	0.43	-4.25	4.48	1.25	0.25
2	1.00	0.26	0.60	0.77	-1.09	1.79	1.11	0.31
	1.25	0.32	0.71	0.88	0.05	1.18	0.91	0.32
	1.50	0.33	0.73	0.90	0.58	1.08	0.74	0.33
	2.00	0.30	0.67	0.87	1.05	1.20	0.51	0.31

Table 3.1: Simulation results for various V_{η_i} and C; shown are, respectively, proportions or averages, over all signals and sample paths. Columns 3–5 show the proportion of the number of estimated change points $\hat{N}$ that equal or fall within a small range of N. Columns 6 and 7 show $N - \hat{N}$ and $|N - \hat{N}|$, respectively. Columns 8 and 9 show the NA and SD measures, respectively.

3.5 Data Analysis

3.5.1 Data

We analyse the performance of our model in terms of basis recovery and forecasting in an application to 16 financial time series, four each from of these four asset classes: equity indices, single-name stocks, foreign exchange rates and commodity futures. We consider their daily closing prices available between 1 January 1990 and 21 June 2013, obtained from Bloomberg. Details of the assets are provided in Table A.1. We ignore days on which no price data are available. Correspondingly, we obtain $T_a^* \in \{4774, \ldots, 6124\}$ data points per each asset a.

3.5.2 Interpretation of Change-Point Importance

Figure 3.4 shows log-price series and our integrated model fit $\sum_{s=1}^{t} \hat{f}(s)$ for the General Electric (GE) equity, the GBP-USD exchange rate and the WTI crude oil commodity future. For the S&P 500 index we refer back to Figure 3.1. While the patterns exhibited in the three series from Figure 3.4 differ, two of the more pronounced trend changes take place for all three series at the start of the global financial crisis in early 2008 and following the dot-com bubble burst in 2000/2001. However, the estimated canonical bases reveal that both of these events were more important in the evolution of the General Electric share price than in those of the GBP-USD exchange rate or the crude oil: the earliest change point in the General Electric price series that can be attributed to the recent global financial crises is estimated at scale $j = 2$ of the canonical basis, while corresponding change points for GBP-USD and crude oil only appear at scales $j = 7$ and $j = 8$, respectively. The dot-com burst corresponds to the change point at scale $j = 1$ for the GE data and hence this event can be interpreted as the most important for this series.

For the GBP-USD exchange rate, other events are more important: the change point at scale $j = 1$ is detected in 1990, corresponding to the British currency's joining of the European Exchange Rate Mechanism. For the oil price, the most important change point is detected early in time and can be related to the price shock due to the Iraqi invasion of Kuwait in the summer of 1990. Outside of the period 1990/1991, the Asian financial crisis in 1997/1998 triggered some important changes in the oil price trend. Given these examples, we concur that the interpretation of the hierarchical order of the canonical UH basis is coherent with historic events that had an impact on world financial markets in the period at hand.
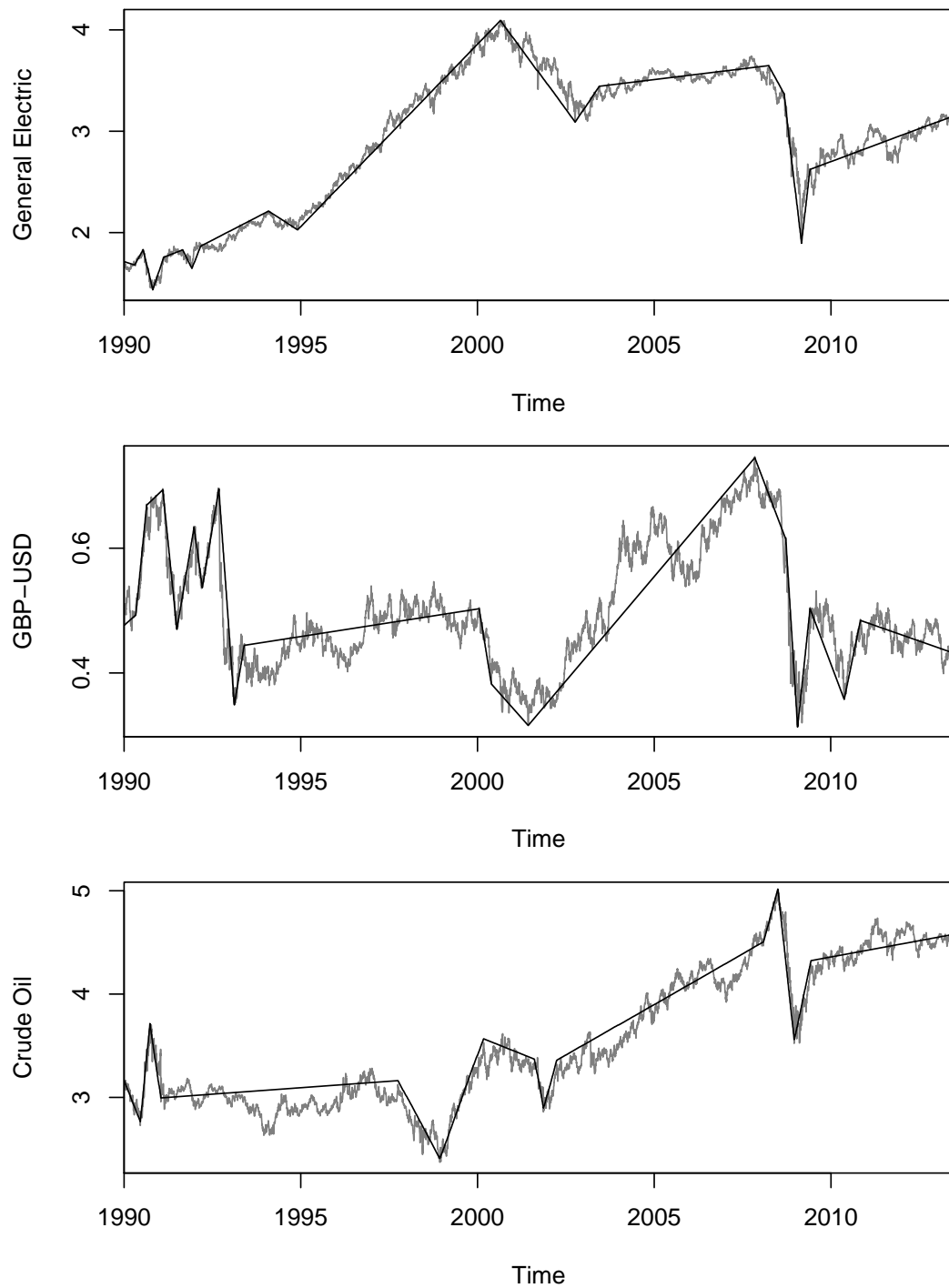
Figure 3.4: Daily closing values of the General Electric share price, the GBP-USD exchange rate and WTI crude oil between January 1990 and June 2013; log-price (grey) and the cumulatively integrated fit $\sum_{s=1}^{t} \hat{f}(s)$ from our model (black), shown for threshold parameter $C = 0.3$ and imposing a minimum distance between change points of 60 trading days.

3.5.3 Forecast Evaluation

Our model's ability to predict returns is evaluated in a rolling window forecast analysis conducted for each of the 16 assets. Formally, for observed returns on each asset a ending at time $\bar{T} + s$, we evaluate the h-day cumulative return forecast $\hat{f}^{a,h}(s, \bar{T}) = \sum_{t=1}^{h} \hat{f}^a(s + \bar{T} + t) = h\hat{f}^a(s + \bar{T} + 1)$, and compare it to the observed return $x^{a,h}(s, \bar{T}) = \sum_{t=1}^{h} x^{a,h}(s + \bar{T} + t)$. To avoid overlapping forecast windows, the forecasts are evaluated for $s = \{1, 1 + h, 1 + 2h, \ldots, 1 + (H_{a,h} - 1)h\}$, where $H_{a,h} = \lfloor (T_a^* - \bar{T})/h \rfloor$ is a function of the forecast horizon $h = \{1, 2, 5, 10, 15, 20, 40\}$. The length of the estimation period is set to $\bar{T} = 1,750$ days, or around 7 trading years, which roughly corresponds to the typical length of a US business cycle since 1980 (NBER, 2013). The parameters of the UHP-BS procedure are as in Section 3.4, with the threshold constant C taking values in the set $\{0.1, 0.2, \ldots 3.0\}$.

In the following, the index a is suppressed and it is understood that the forecast performance is evaluated for each asset a. The forecast $\hat{f}^h(s, \bar{T})$ equals h-times the moving average of the most recent observations of $\mathbf{x}$, with a span $\bar{T} + s - \hat{\eta}_{\hat{N}}$ chosen data-adaptively from the data as described in Section 3.3.3. Hence, a natural benchmark is a non-adaptive moving-average estimator that forecasts $\tilde{f}^h(s, \bar{T}, w) = h/w \sum_{t=s+\bar{T}-w+1}^{s+\bar{T}} x(t)$. As the choice of w will clearly affect the performance, we compare our model's forecast $\hat{f}^h(s, \bar{T})$ with an optimised threshold constant C against the best forecast from a range of moving-average models with $w = \{h, 2h, \ldots 10h\}$ for each forecast horizon h.

Our success criterion is sign predictability, defined as the proportion of correctly predicted signs. Some authors, e.g. Leitch and Tanner (1991), argue that this

approach provides more robust results than statistics based on the level of predicted returns and provide evidence that in financial applications under the objective of profit maximization, the proportion of correctly predicted signs beats, amongst others, the mean square prediction error as a criterion for choosing forecasts. We define the 'Relative Success Ratio' for a model with forecast $\dot{f}^h$ at forecast horizon h as $\text{RSR}_h = 1/H_h \sum_{i=0}^{H_h-1} \mathbb{I}(x^h(1+ih,\bar{T})\dot{f}^h(1+ih,\bar{T} > 0) - 0.5$.

A large positive RSR_h provides evidence for predictability and can be expected if an asset behaves in a trend-following way. However, as mentioned in Section 3.3.3, a large negative RSR_h can be related to mean reverting behaviour and also provides a meaningful trading signal. We only compare the two models in terms of their RSR_h if both provide the same directional signal, that is if both suggest trend-following or mean-reversion behaviour. Table 3.2 summarizes the general results as well as results for forecast horizons grouped into short, medium and long term. The results are individually optimised in terms of the threshold constant C (for our model) and the moving average span w (for the benchmark) for each forecast horizon to yield the largest absolute RSR_h, i.e. the model fit with the proportion of predicted signs deviating furthest from 0.5.

Over all forecast horizons combined, our model and the benchmark are performing comparably, with our model offering better performance for seven assets, ahead of the benchmark, which performs better for five assets. The difference in performance appears to be due to the commodities asset class, where the sign predictability of our model is clearly more pronounced for gold and oil, which show trend-following behaviour. Differentiating over forecast horizons, our model tends to outperform the benchmark in short- and medium-term prediction, but performs

worse in long-run predictions.

Table 3.3 summarizes the RSRs during times of strong movements which often reflect periods of financial market distress. We classify time periods as containing strong movements if their h-day cumulative return lies in the top or bottom 10% of their historic distribution. Formally, we consider only those forecasts for which $x^h(s, \bar{T} - h + 1) < \text{Quant}^{10\%}_{s,\bar{T},h}$ or $x^h(s, \bar{T} - h + 1) > \text{Quant}^{90\%}_{s,\bar{T},h}$ with $\text{Quant}^{\alpha\%}_{s,\bar{T},h}$ denoting the α-percentile of non-overlapping h-day returns in the period $[s, s + \bar{T}]$. In this subset our model does well particularly in currencies, where the apparent mean-reverting behaviour can be captured better by our adaptive predictor. Otherwise the pattern is similar to that in Table 3.2, but even more favourable to our adaptive predictor. In particular for the long-term forecasts, our model now performs well compared to the benchmark.

Out-of-Sample Evaluation

We end this empirical analysis with a brief out-of-sample evaluation of the UHP-BS procedure. Using a rolling-window approach similar to the one discussed in the previous section with a seven-year estimation period, we now split the available data in two parts: training and testing set. The number of forecasts resulting from this set-up is given in Table A.2 in the appendix (p. 198). After estimating our model and the benchmark model on the training set, the criterion to determine empirical optimality of the model parameter, the threshold value C for our model or moving average factor MA for the benchmark, respectively, is defined as the in absolute terms largest Relative Success Ratio over the in-sample period per forecast horizon h. This implies that no discretionary decision is imposed as to whether an

| | $h = 1, 2, 5, 10, 15, 20, 40$ | | $h = 1, 2$ | | $h = 5, 10, 15$ | | $h = 20, 40$ | |
	UHP	Benchmark	UHP	Benchmark	UHP	Benchmark	UHP	Benchmark
DAX	6.8	**6.8**	**3.0**	2.2	**7.1**	6.1	10.0	**12.6**
FTSE	**5.6**	2.6	1.7	-1.9	5.0	-0.4	10.4	**11.6**
HSI	2.5	**4.6**	1.6	**3.0**	4.2	4.2	0.7	**6.9**
S&P	**7.2**	6.2	**2.4**	0.2	**7.7**	6.6	11.3	**11.5**
LHA	-0.1	4.4	-2.0	0.6	0.1	**5.9**	1.4	**6.0**
GE	-1.1	**-3.4**	**-2.3**	-2.0	**-1.2**	-0.8	0.2	-8.9
JNJ	-1.2	0.9	-1.7	**-1.7**	-1.4	0.4	-0.2	4.1
MSFT	**-1.8**	-1.7	**-2.1**	-0.3	-2.2	0.1	-1.0	**-5.8**
GBPUSD	**-4.6**	-3.2	**-1.3**	-0.1	**-3.9**	-3.3	**-8.9**	-6.0
USDJPY	0.2	**1.7**	**-1.5**	-0.1	1.8	1.7	-0.5	3.4
AUDUSD	3.9	**4.8**	2.1	-0.3	3.3	**4.5**	6.6	**10.3**
USDMXN	**-4.1**	-0.4	-2.3	1.7	-5.4	1.0	-4.2	**-4.6**
Oil	**6.6**	1.8	**2.9**	0.1	**6.2**	1.7	**10.8**	3.5
Gold	**2.5**	1.6	3.3	-2.2	**4.1**	0.1	-0.6	7.4
Live Cattle	2.0	-6.1	-1.7	**-2.9**	0.8	-5.8	7.5	-9.7
Sugar	-2.8	3.4	-2.0	0.3	-1.5	4.1	-5.5	5.5

Table 3.2: Relative success ratio in percent for our model with optimised threshold value and the benchmark model with optimised moving average window length. If the relative success ratio has the same sign for both models, the larger absolute value is in bold. Forecast horizons are grouped with $h = 1, 2$, $h = 5, 10, 15$ and $h = 20, 40$ representing short-, medium- and long-term forecasts, respectively.

asset should be trend-following or mean-reverting. However, we implicitly assume that this asset-specific characteristic is time invariant.

We evaluate the h-day return forecasts using the optimised parameters, respectively, C and MA, from the corresponding h-day forecast in-sample estimation. Tables A.3 and A.4 in the appendix show these parameter values. As in the previous section, the analysis focuses on the RSR_h statistics which are considered comparable if they have the same sign. Our model beats the benchmark for the horizons one, five, ten and 15 days, as shown in Table 3.4, but performs less well for longer horizons. Looking at an asset-class level, we find evidence that our model's success is mainly visible for equity indices and commodities and less so for currencies and single stocks. A possible reason for the difference in the performance for

| | $h = 1, 2, 5, 10, 15, 20, 40$ | | $h = 1, 2$ | | $h = 5, 10, 15$ | | $h = 20, 40$ | |
	UHP	Benchmark	UHP	Benchmark	UHP	Benchmark	UHP	Benchmark
DAX	**7.0**	3.0	-0.1	2.6	**7.7**	4.7	**13.1**	0.9
FTSE	2.4	-3.4	**-2.7**	-2.4	3.0	-7.6	**6.8**	1.9
HSI	3.0	**9.5**	**4.0**	3.2	2.9	**9.1**	2.1	**16.4**
S&P	0.9	**4.3**	-4.1	**-4.4**	2.2	**4.3**	3.8	**12.9**
LHA	-4.2	0.1	-3.5	3.4	**-10.1**	-6.0	3.8	**5.8**
GE	**-3.1**	-2.3	**-0.7**	-0.3	**-5.4**	-5.4	-2.1	0.5
JNJ	-0.3	2.1	-4.4	**-6.2**	1.4	-1.0	1.2	**15.0**
MSFT	-3.4	2.2	-3.4	4.4	-2.8	2.1	-4.4	0.3
GBPUSD	**-7.4**	-2.0	-3.4	0.2	-4.2	2.4	**-16.4**	-11.0
USDJPY	**-2.1**	-0.4	**-0.4**	0.0	**-1.9**	-0.6	**-3.9**	-0.7
AUDUSD	1.3	-2.3	-0.6	**-4.3**	3.4	-1.3	0.0	**-1.8**
USDMXN	**-3.1**	-2.1	-0.5	3.8	-3.0	0.3	-5.7	**-11.4**
Oil	4.5	4.4	0.7	-1.4	5.3	**10.4**	**6.9**	1.3
Gold	**-5.6**	-5.3	3.3	-0.7	**-6.9**	-0.6	-12.8	**-16.8**
Live Cattle	-5.4	**-6.1**	-2.1	**-2.5**	**-9.1**	-4.1	-3.0	**-12.8**
Sugar	-1.9	5.8	**-3.8**	-3.2	3.4	**8.8**	-8.0	10.4

Table 3.3: Relative success ratio RSR during times of strong movements in percent for our model with optimised threshold value and the benchmark model with optimised moving average window length. If the relative success ratio has the same sign for both models, the larger absolute value is in bold. Results for forecasts taking place when the most recent h-day cumulative return is in the top or bottom decile of its historical distribution. Forecast horizons are grouped with $h = 1, 2$, $h = 5, 10, 15$ and $h = 20, 40$ representing short-, medium- and long-term forecasts, respectively.

equity indices versus single stocks is that the latter are more driven by idiosyncratic shocks.

An analysis with a two-year estimation period instead of the seven-years window used above yields similar results, with our model clearly beating the benchmark for most assets where a comparison is possible at the horizons two, five, ten and 15 years. Details on this analysis are provided in Section A.2.2 in the appendix.

Forecast horizon	1		2		5		10		15		20		40	
	UHP	BM	UHP	BM	UHP	BM	UHP	BM	UHP	BM	UHP	BM	UHP	BM
DAX	**2.27**	1.84	**2.15**	1.00	**4.89**	0.60	**5.02**	1.67	**10.71**	9.29	3.33	**10.00**	7.69	7.69
FTSE	0.89	-2.23	-0.19	**-0.57**	**-3.48**	-0.36	**6.94**	3.11	**11.87**	2.52	**9.05**	8.10	5.77	-1.92
HSI	**1.53**	0.84	-0.44	1.73	**2.59**	0.86	**3.20**	1.72	4.81	-0.37	3.47	**7.43**	**8.82**	4.90
S&P	-0.91	0.50	2.02	**2.50**	0.12	**6.12**	**10.58**	4.81	1.08	**6.12**	**10.58**	3.85	13.46	9.62
LHA	0.44	-0.39	-1.80	0.44	-0.24	1.71	0.73	**3.69**	2.17	-1.45	-2.88	10.58	7.69	**13.46**
GE	**-2.65**	-2.25	-1.07	0.59	2.43	-0.85	1.92	-1.44	0.36	0.36	4.81	-5.77	-3.85	**-5.77**
JNJ	**0.44**	0.19	1.01	-0.77	**-3.62**	-3.28	**4.11**	3.62	**-5.80**	-1.45	1.92	-1.92	-1.92	**-3.85**
MSFT	0.29	-1.31	0.39	**0.83**	-6.04	0.72	-0.96	4.33	**-5.40**	-1.80	-0.96	2.88	-	-5.77
GBPUSD	0.39	**0.81**	-0.60	**-1.74**	0.34	-2.52	-1.83	**-5.50**	**-5.48**	-4.11	-1.38	**-5.05**	-6.36	6.36
USDJPY	-0.05	**-0.99**	-0.83	0.18	-0.80	1.26	-0.68	**-2.05**	1.37	**2.74**	0.46	-5.05	4.55	**13.64**
AUDUSD	2.70	-1.09	**2.57**	1.66	1.95	-3.55	2.05	**2.75**	-3.42	-	0.46	**9.63**	0.91	2.73
USDMXN	-2.86	1.67	0.09	**0.91**	-4.69	0.57	-1.14	0.68	-0.34	3.79	1.38	-4.13	-1.85	**-5.56**
Oil	2.35	-0.97	**3.24**	0.48	**5.56**	1.93	-6.04	4.59	6.52	-1.45	4.81	-0.96	**15.38**	3.85
Gold	**-0.97**	-0.46	**5.90**	0.39	6.28	-0.48	**6.04**	2.66	-	-4.35	-1.92	0.96	-3.85	3.85
Live Cattle	-1.10	**-2.26**	-1.79	**-2.68**	**3.49**	1.81	-3.14	4.15	2.17	-9.42	6.86	-8.82	9.62	-0.98
Sugar	**1.13**	0.17	0.05	**0.69**	1.21	**1.71**	0.97	**3.17**	1.09	**1.82**	-2.43	1.46	5.77	**7.69**

Table 3.4: Relative success ratio RSR in percent in the out-of-sample test; for our model with in-sample optimised threshold value and the benchmark model (BM) with in-sample optimised moving average window length, for a seven-year estimation period. If the relative success ratio has the same sign for both models, the larger absolute value is in bold.

3.6   Extension into a Multivariate Setting

Multivariate extensions of the nonstationary piecewise constant process are interesting from both a theoretical and a practical perspective. They open the door for dimension reduction techniques such as nonstationary dynamic factor models (Fan et al., 2008; Motta and Ombao, 2012) or, in the context of change-point detection, sparsifying methods such as Cho and Fryzlewicz (2015a). Multivariate extensions also can lead to adaptations focussing on specific characteristics, such as time-varying correlations (Pelletier, 2006). A natural extension of the adaptive change-point induced basis recovery model into a multivariate setting can be defined as follows

Definition 3.6.1 *A D-variate stochastic process $\mathbf{X}(t) = (\mathbf{x}_1(t), \ldots, \mathbf{x}_D(t))'$, $D > 1$, $t = \{1, \ldots, T\}$ is called the* Multivariate Unbalanced Haar *process if it has a representation*

$$
\begin{aligned}
\mathbf{X}(t) &= T^{1/2} \sum_{(j,k)\in\mathcal{I}} \psi^{b_{j,k}}(t)\mathbf{A}_{j,k} + \mathbf{\Sigma}(t)\boldsymbol{\varepsilon}(t), \quad t = \{1, \ldots, T\}, \quad where \quad (3.7) \\
\mathbf{A}_{j,k} &= (A_{j,k}^{(1)}, \ldots, A_{j,k}^{(D)})' &(3.8) \\
\mathbf{\Sigma}(t) &= diag(\sqrt{\sigma_1(t)}, \ldots, \sqrt{\sigma_D(t)}\, \boldsymbol{R}\, diag(\sqrt{\sigma_1(t)}, \ldots, \sqrt{\sigma_D(t)} &(3.9) \\
\boldsymbol{\varepsilon}(t) &= (\varepsilon_1(t), \ldots, \varepsilon_D(t))' &(3.10)
\end{aligned}
$$

$\mathcal{I}$ and $\psi^{b_{j,k}}$ are as in Definition 3.2.1. The D-dimensional random variables $\{\mathbf{A}_{j,k}, (j,k) \in \mathcal{I}\}$ are independent of the noise term $\boldsymbol{\varepsilon}(t)$ for all t,j,k. $\{\mathbf{A}_{j,k}, (j,k) \in \mathcal{I}\}$ are drawn from a continuous distribution, satisfy $E(\mathbf{A}_{j,k}) = \mathbf{0}$ and have covariance $Cov(A_{j,k}^{(d)}, A_{j,k}^{(d')}) = \sigma_d \sigma_{d'} \rho_{j,k}^{(d,d')} < \infty\; \forall d, d'$. The volatility of the noise is

bounded, $0 < \underline{\sigma} < \sigma_d(t) < \overline{\sigma} < \infty \; \forall d, t$, *and the correlation between the noise terms is constant, i.e. the correlation matrix* $\boldsymbol{R}$ *has off-diagonal entries* $\boldsymbol{R}_{d,d'} = \rho_\varepsilon^{(d,d')}$ $\forall d \neq d'$. $\varepsilon_d(t)$ *are Gaussian random variables.*

The variance and auto-covariance terms correspond to those discussed in Section 3.2.3. The Unbalanced Haar cross-spectrum can be defined similarly to Definition 3.2.2. The concept of change-point induced basis recovery is core to this model.

As discussed above, if the Unbalanced Haar spectral approach is used to compare two or more time series, a common basis should be used in order for the comparison to be meaningful, unless it is of interest to study feature misalignment between the different series. A natural question is therefore how to discover an uniquely identifiable change-point hierarchy in the multivariate framework. Cho and Fryzlewicz (2015a) discuss a thresholded-sum approach to aggregate CUSUM-based test statistics over multiple time series, and compare this to the approaches of Groen et al. (2013), a maximum and an average over the univariate CUSUM statistics per point in time. We discuss the three versions in Chapter 5 which focuses on multivariate data. All variants can be embedded directly within a binary segmentation algorithm and yield a unique hierarchical structure of change points for a given multivariate time series. Therefore we can construct a basis of Unbalanced Haar vectors by the same logic as in the univariate case.

The unconditional cross-covariance between the two time series is the sum of the covariance between the scaled wavelet coefficients per scale j and location k

and the error covariance term,

$$\text{Cov}(\mathbf{x}_d(t), \mathbf{x}_{d'}(t)) \;=\; T \sum_{(j,k)\in\mathcal{I}} \text{Cov}(A_{j,k}^{(d)}, A_{j,k}^{(d')}) + \sigma_d(t)\sigma_{d'}(t)\rho_\varepsilon^{(d,d')} \quad (3.11)$$

The covariance between wavelet coefficients per scale j and location k is the product of asset volatilities and scale- and location specific correlation coefficients, $\text{Cov}(A_{j,k}^{(d)}, A_{j,k}^{(d')}) = \sigma_d \sigma_{d'} \rho_{j,k}^{(d,d')}$

The correlation between error terms $\rho_\varepsilon^{(d,d')}$ is assumed to be constant over time. Hence, if changes in the correlation between asset returns are not induced by time-varying volatility, they can be attributed to the correlation terms in the signal part. One of the crucial points in the multivariate framework is to make reasonable assumptions regarding the cross-correlation $\rho_{j,k}^{(d,d')}$ and to define an estimator. Clearly, correlation should vary over time, and a natural approach would be to assume similar correlations per scale j. However, in practice this is of limited usefulness given the data-adaptive nature of our model and hence possibly strongly varying vector lengths at lower scales. A more sensible approach is to assume that correlation is similar for intervals of similar length or to estimate it over groups of wavelets of 'similar shape', e.g. conditioning on segment length, change-point location and change direction or height.

Figure 3.5 illustrates how the multivariate UH model can provide new insights to data. The model is fitted to a panel of time series containing the four equity indices of our previous analysis, DAX, FTSE, HSI and S&P, in the same period and using the same parameters as in the univariate setting. We are using the average CUSUM approach of Groen et al. (2013) and it appears that idiosyncratic

variation, by which we mean changes in the mean return of single indices, is not reflected by a change point at scales $j = \{1, \ldots, 5\}$. The HSI, which lies above the other indices in this representation, contains various changes in its trend in the period 1993-1998 that are not apparent in the other three indices and are detected as change points only at finer scales than the ones shown here. We see that the first change identified corresponds to the burst of the dot com bubble at the beginning of this century, while the turning points induced by the financial crisis appear only at scales four and five.

3.7 Concluding Remarks

The UHP-BS procedure is a new change-point detection method, developed to capture the oscillatory behaviour of financial returns around piecewise-constant trend functions. Core of the method is a data-adaptive hierarchically-ordered basis of Unbalanced Haar vectors which decomposes the piecewise-constant trend underlying observed daily returns into a binary-tree structure of one-step constant functions. The UH process provides a new perspective for the interpretation of change points in financial returns: for instance, the unconditional variance can be decomposed into location- and scale-specific variances of the UH amplitudes $A_{j,k}$, plus the error variance. A similar interpretation exists for the autocovariance, which is decomposed into positive and negative parts, with the sign depending on whether at a given scale two points t, $t + \tau$ are on the same segment of the scale-specific UH vector or separated by a change point. Moreover, UHP-BS provides the analyst with a family of forecast operators. We compare the performance of UHP-BS to a
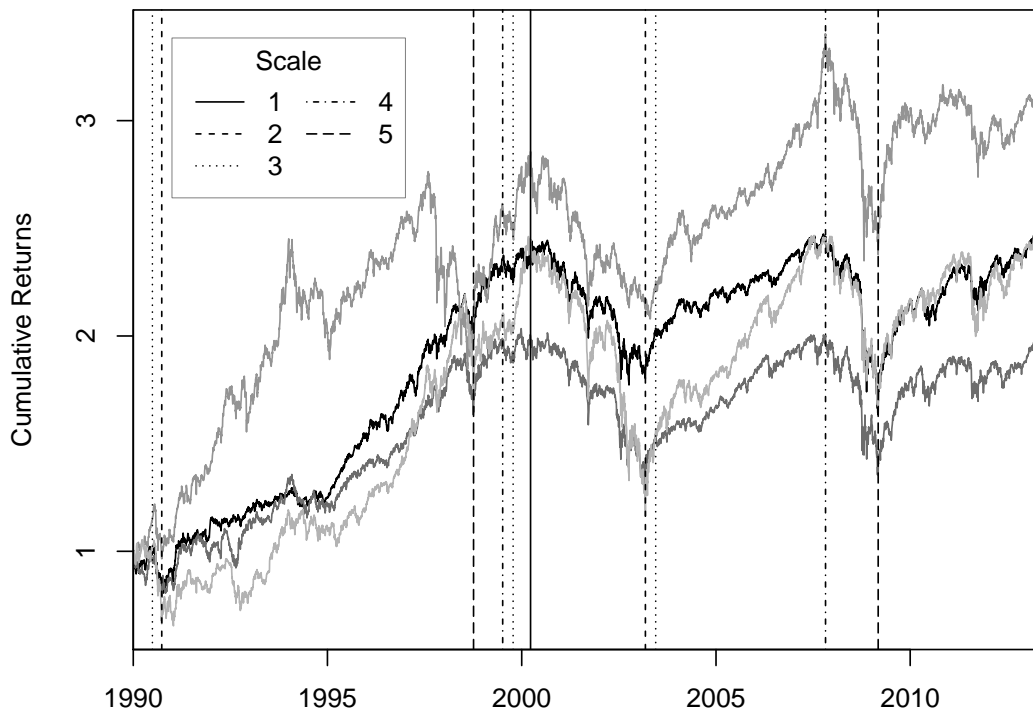
Figure 3.5: Cumulative log-returns of equity indices between January 1990 and June 2013; from dark to light grey: S&P 500, FTSE, HSI, DAX. The change-point locations identified by the multivariate UH model using an average CUSUM approach and a minimum distance between change points of 60 trading days are indicated as vertical lines, with different line types corresponding to different scales j.

benchmark set of moving average forecasts using a data set of 16 assets from four asset classes and a period of over 20 years of daily observations. We also briefly discuss an extension of the UHP into a multivariate setting and illustrate how a data-adaptive basis can be determined and how the hierarchical structure of this basis can help interpret change-point importance in a multivariate setting.

There are various interesting paths going forward. For instance in the development of a multivariate framework for the Unbalanced Haar process it would be interesting to analyse under which assumptions common empirical observations (also termed 'stylised facts', Cont, 2001) of asset returns can be captured, e.g. extremal correlation between assets or what we label macrostructure noise, the characteristic upward trend of absolute correlation between return series when it is computed on decreasing frequency. Another interesting project would be to evaluate a trading model build on the estimated coefficients $A_{j,k}^{(d)}$. We would expect that for time series estimated on the same basis, a function of the difference $A_{j,k}^{(d)} - A_{j,k}^{(d')}, d \neq d'$, could be a valuable indicator for pair trading.

It is also interesting to explore alternative basis definitions. Core to UHP-BS is a data-adaptive orthonormal oscillatory basis which is defined by a hierarchical structure from Unbalanced Haar vectors using binary segmentation. The resulting implicit partial ordering has attractive theoretical properties, is computationally fast compared to most competing approaches and has the convenient feature of making the change-point hierarchy interpretable. However, the canonical UH basis depends on the jump sizes at the discovered change points. This makes it difficult to define a data generating process in a simulation study and hence allows for empirical evaluation only conditional on a signal **f**. Furthermore, updating the model in an

online estimation setting is equivalent to re-estimating the model over the entire time interval. Given a set of change points $\mathbb{N}$, alternative basis definitions could be based on a deterministic rule, e.g. assign coarser scales to change points that are more distant from their neighbours. One could also sample bases at random and average over the resulting estimates. Finally, one could consider an alternative procedure for change-point estimation that naturally defines a hierarchical basis, e.g. via WBS (Fryzlewicz, 2014).

The UHP-BS procedure provides interesting insights in both univariate and multivariate financial return data. The approach can be applied to other data that can be described as piecewise-constant signal plus noise, yielding interpretable, data-driven segmentation. However, as we illustrate in the following chapter, there are certain limitations arising from the assumption of changes occuring only in the mean of the underlying process.

3.8 Glossary: Most Essential Notation

Expression	Meaning
Functions	
$O(\cdot)$	Upper bound in the sense that $f(x) = O(g(x)) \leftrightarrow |f(x)| \leq C|g(x)|$ for all $x \geq x_0$ and some $C > 0$
$o(\cdot)$	Lower bound in the sense that $f(x) = o(g(x)) \leftrightarrow |f(x)| \leq \epsilon|g(x)|$ for all $x \geq x_0$ and all $\epsilon > 0$
$\mathbb{I}(\cdot)$	Indicator function
$\langle \cdot, \cdot \rangle$	Inner product
$g(\cdot)$	Unknown function determining the forecasting relation: $f(T + h) = g(f(T))$
$\text{Quant}_{s,\bar{T},h}^{\alpha\%}$	$\alpha\%$ quantile of non-overlapping h-day returns in the period $[s, s + \bar{T}]$
Objects	
$\mathbb{N} = \{\eta_i, i = 1, \ldots, N\}, (\hat{\mathbb{N}})$	True (estimated) set of change points in the UH process $\mathbf{x}$
$\mathbb{V} = \{v_i, i = 1, \ldots, N\}$	Set of change points in rescaled time
$\mathbf{p} = \{p(t), t = 1, \ldots, T\}$	Time series array of daily asset prices
$\mathbf{x} = \{x(t), t = 1, \ldots, T\}$	Time series array of daily asset returns, s.t. $x(t) = \ln(p(t)) - \ln(p(t-1))$
$\mathbf{f} = \{f(t), t = 1, \ldots, T\}$	Time series array of conditional mean of $\mathbf{x}$
$\boldsymbol{\psi}^{b_{j,k}} = \{\psi^{b_{j,k}}(t), t = 1, \ldots, T\}$	Unbalanced Haar array at scale j and location k, corresponding to change point $b_{j,k}$ as defined on page 67
$\bar{a}_{j,k}, \hat{a}_{j,k}$	UH coefficients at scale j and location k w.r.t. $\mathbf{f}$ and $\mathbf{x}$, respectively
$\mathcal{I} = \{(j, k)\}$	Set of indices with cardinality $|\mathcal{I}| = N + 1 < \infty$; with $(-1, 0) \in \mathcal{I}$, and connected in the sense that if a child index $(j + 1, 2k - 1)$ or $(j + 1, 2k)$ is in $\mathcal{I}$, then so is their parent (j, k).
ζ_T	Threshold parameter

Chapter 4

Detection of Changes in Mean and/or Variance using Sequential Testing on Intervals

We consider the situation in which different types of sudden changes are present in an univariate time series: changes in the mean only, simultaneously in mean and variance, and in the variance only. We illustrate that this is not a straight-forward generalization of the change-in-mean detection problem discussed in the previous chapter and that change-point detection procedures face specific non-trivial challenges. To summarize our findings, detection procedures that only consider changes in mean or variance while assuming that there are no sudden changes present in the other quantity can yield misleading conclusions regarding the data. Procedures that assume both quantities to change simultaneously have relatively poor power properties when only mean or only variance changes. This is particularly apparent when the number of change points is relatively large. Our main contribution in this chapter is to propose a novel method for detection of changes in mean and/or variance. The method is developed to account for certain effects from model misspecification that classical change-point detection methods face. For this purpose, it

105

uses Sequential Testing on varying-length Intervals and is thus abbreviated SeqToI.
SeqToI is computationally demanding compared to existing procedures that have the
same purpose of detect change points in mean and variance, but its structure al-
lows us to tackle even difficult change-point detection settings with many change
points of different types and relatively small magnitudes. We illustrate the perfor-
mance of SeqToI in a simulation study in which we compare it with state-of-the-art
competitor methods.

4.1   Introduction

This chapter is concerned with a problem related to but more general than that
of change-in-mean detection discussed previously. We focus here on the case of
an unknown number of sudden changes that can appear in mean and variance,
possibly but not necessarily coinciding. We first introduce a process that describes
this situation and then provide examples where the research question is relevant in
practice. As it turns out, assuming only changes in the mean or only changes in the
variance when both quantities change can have substantial negative effects on the
accuracy of a change-point detection procedure, while assuming that both quan-
tities change only simultaneously can impact the power of a detection procedure
negatively.

The mean and/or variance change process is piecewise-stationary, and at every
change point in $\mathbb{N} = \{\eta_i, i = 1, \ldots, N\}$ at least either mean or variance changes,
but both may also change simultaneously. Between change points, the process
is stationary, i.e. mean and variance are constant. The arguably most simple

representation of this process $\mathbf{x} = \{x(t), t = 1, \ldots, T\}$ is

$$x(t) = f(t) + s(t)\epsilon(t), \quad t = 1, \ldots, T$$
$$f(t) = \mu_i, \ s^2(t) = \sigma_i^2 \quad \text{if } \eta_{i-1} < t \leq \eta_i, \ i = 1, \ldots, N+1 \quad (4.1)$$

where we assume the sequence of means to be bounded, $\{\mu_i : |\mu_i| < \infty, i = 1, \ldots, N+1\}$, and the sequence of variances to be finite and positive, $\{\sigma_i^2 : 0 < \sigma_i < \infty, i = 1, \ldots, N+1\}$. $\epsilon(t) \sim N(0, 1)$ is i.i.d. Gaussian noise, but this can be generalized as discussed in Section 4.2. At any change point $\eta_i \in \mathbb{N}$, at least one of the following holds: a) the mean changes by some nonzero quantity, $|\mu_i - \mu_{i-1}| > 0$, or b) the variance changes by some nonzero quantity, $|\sigma_i - \sigma_{i-1}| > 0$. Moreover, the distance between neighbouring change points is bounded from below, $\eta_i - \eta_{i-1} > \delta_T > 0$. Our goal is to estimate the number of change points N and their locations $\mathbb{N} = \{\eta_i, i = 1 \ldots, N\}$.

In this mixed change-type situation, at each change point η_i either the mean changes, or the variance changes, or both quantities change simultaneously. The resulting change-point detection problem thus has a new dimension compared to previously discussed approaches, the change type, in addition to change-point number and change-point locations. As we illustrate below, ignoring this new dimension yields at best a loss of power in the detection method, but can also severely distort the change-point detection results if inappropriate estimation procedures are used.

The detection of possibly coinciding changes in mean and variance is important in a wide range of applications, for example in climatology. The question of global warming has received much attention in the past decades and a strand of litera-

ture is concerned with the analysis of evidence for changes in temperature mean and variance. For instance Parey et al. (2013) argue that accounting for changes in the temperature mean and variance is crucial for the analysis of temperature extremes. Vasseur et al. (2014) provide evidence that increases in temperature variation poses a greater risk to species than climate warming, i.e. changes to the mean temperature. The upper panel of Figure 4.1 shows the temporal evolution of monthly average temperature in Berlin, Germany, for the month of July. From mere eye-balling one could suspect that both mean and variance (approximated by squared deviation from the sample mean and shown in the lower panel) change over time and that their change points do not necessarily coincide.

We return to one of the examples introduced in Chapter 1 for a second illustration of a time series with changes in mean and variance. Figure 4.2 shows the S&P 500 for the period of 2006 to 2008. The index underwent sudden changes in mean return level (corresponding to changing linear trends in the price level) and in the return variance, estimated by the squared deviation from the sample mean return. In asset price data the correct identification of change points in mean returns and asset price variation is an important component for investment and risk management strategies. Understanding changes in periods leading up to a financial market turmoil are of particular interest. For instance, Ang and Timmermann (2012) point out that the presence of sudden changes in mean and variance of asset returns can severely impact investors' optimal portfolio choice. Guidolin and Hyde (2012) find evidence that over the period 1953 to 2009, accounting for sudden changes in asset returns is strongly beneficial for investors concerned with long-run strategic multi-asset allocation when compared to models assuming smooth dynamics.
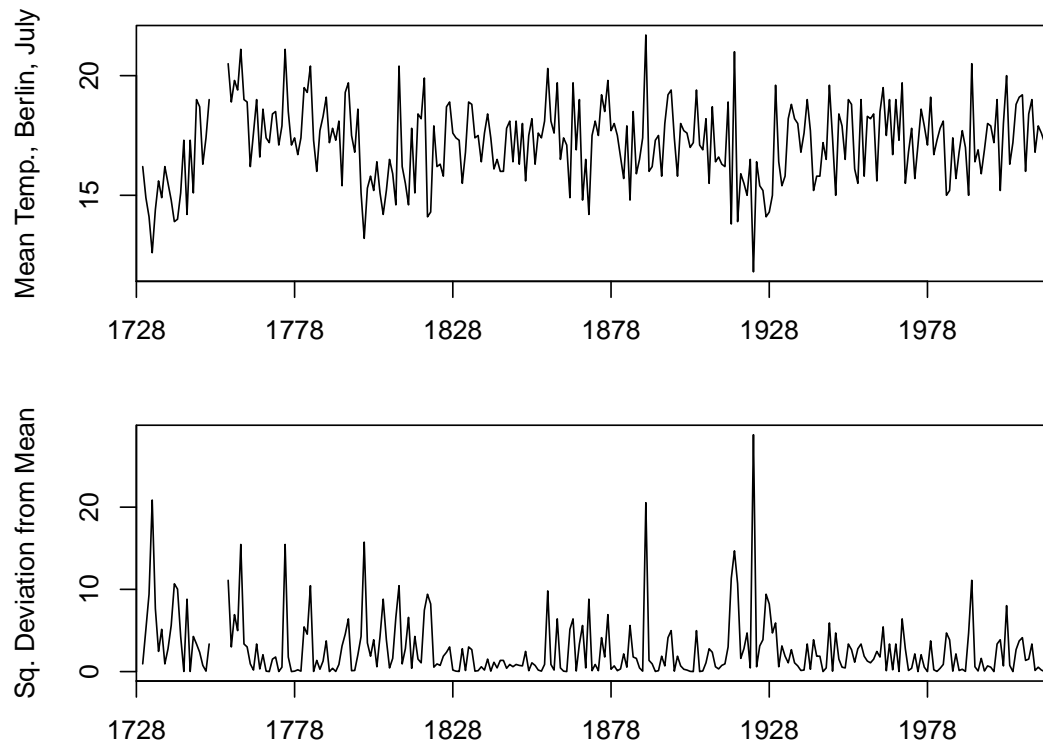
Figure 4.1: Average temperature in the month of July in Berlin, Germany. Upper panel: temperature in Celsius. Lower panel: Mean squared deviation of temperature. Source: Global Historical Climatology Network (www.ncdc.noaa.gov, retrieved 10/01/2016); 6 missing data points in the period 1728-2015
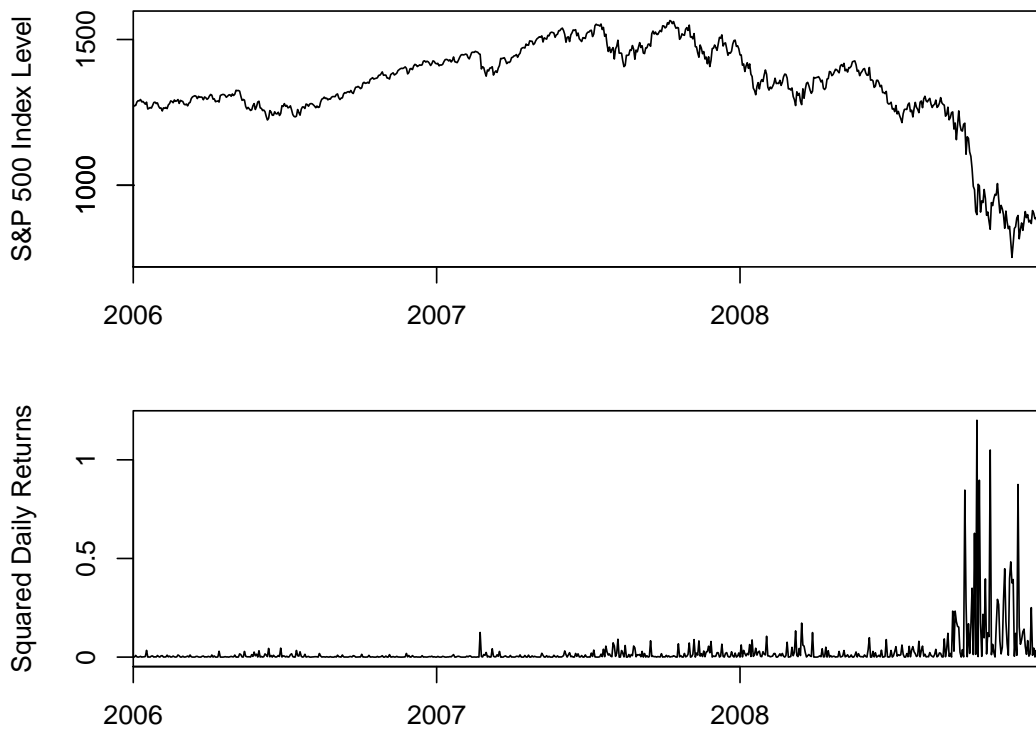
Figure 4.2: S&P500 Price Index, daily observations between January 2006 and December 2008. Upper panel: end-of-day level. Lower panel: squared log-returns

Other areas where research has been concerned with the identification of changes in mean and variance are for instance DNA segmentation (Braun et al., 2000; Chen and Wang, 2009), economics (Pitarakis, 2004) and navigation engineering (Spangenberg et al., 2008).

Because of the unknown mixture of change-point types, the process of Equation (4.1) is a challenging change-point detection problem. We discuss the difficulties arising from the mixed-change situation in detail in the next section but point out here that the performance of single-change-type detection methods, i.e. methods detecting changes in the mean only or in the variance only, can be substantially affected in the presence of a mixture of change types. In particular, in specific situations single change-type detection methods are at best less powerful and at worst biased when they operate under this kind of model misspecification. On the one side, detection procedures that only consider changes in mean *or* variance while assuming the other quantity to be constant can yield misleading conclusions regarding the data. On the other side, principally, changes that occur only in the mean or only in the variance are detectable by procedures that assume both quantities to change simultaneously. However, it can be shown that these procedures have inferior power properties when only mean or only variance changes (cf. Perron and Zhou, 2008; Zhou and Perron, 2008, for analysis and discussion of changes in coefficients and/or error variance in a regression framework), especially when the number of change points is relatively large. This has not received much attention from researchers concerned with change points in time series, possibly because in some situations the loss of power might be acceptable in light of faster computation and easier analytical traceability. We therefore proceed by reviewing previous at-

tempts to identify change points under the assumption of simultaneously changing mean and variance.

Work of Wang and Zivot (2000) proposes a Bayesian method for the detection of changes in mean, trend and variance of a time series using the Gibbs sampler; as often encountered in multiple change-point detection, their method requires knowledge about the total number of change points in the time series. Similarly, using an information criteria approach to estimate the correct number of change points, Bai (2000) develops a quasi maximum likelihood method to detect changes in mean and variance of a vector autoregressive process. For any admissible, user-defined number of change points this method is applied in an iterative manner to estimate change-point locations and model parameters in $O(T^2)$. Tsay (1988) propose an iterative heuristic procedure for the specific case of an ARMA process. Jong et al. (2003) discuss a heuristic procedure for DNA segmentation assuming changes in mean and variance by combining genetic and local search algorithms with maximum likelihood to identify the optimal number of change-points. In a similar setting, Picard et al. (2005) use penalized likelihood and dynamic programming for change-point detection. For a biomedical data set, the authors compare segmentation results testing for changes in mean (assuming constant variance) to those testing for simultaneous changes in mean and variance. However, they neither address the possibility of a mixed-change situation nor discuss systematic issues related to potential model misspecification. We return to this example of testing for different change types at the end of Section 4.2.2.

Hawkins (2001) discusses a maximum likelihood method for change-point detection that requires either prior knowledge about the number of change points or a

fixed maximum number of change points and a supplemental analysis using a gen-eralized likelihood ratio test for changes in the mean and variance of normal data. In a series of papers, Perron and Zhou discuss the estimation of regression models that contain change points in mean or variance, possibly coinciding (Perron, 2006; Perron and Zhou, 2008; Zhou and Perron, 2008). The authors discuss a quasi max-imum likelihood approach to testing for a specific number of change points in the mean and/or a specific number of change points in the variance, which corresponds to a global change-point algorithm optimized over two dimensions. This procedure is then combined with a double-maximization algorithm (Bai and Perron, 1998) to estimate the correct number of change points, which requires a maximum number of allowed change points to be specified by the user. The method is shown to perform well in applications with few change points, at most two in mean and two in variance. Other global optimization methods such as the PELT algorithm of Killick et al. (2012) are able to detect change points in much more general settings, but require the specification of the change-point test, which can be e.g. penalized likelihood for changes in mean, in variance, or (coinciding) in both quantities.

A different strand of literature is concerned with the explicit formulation of the relation between changes in mean and in variance. For example, Braun et al. (2000) circumvent the difficulties encountered when trying to detect changes in mean and/or variance by assuming an explicit dependence structure between mean and variance in applications to DNA segmentation. Within the financial and econo-metrics literature, much work focused on modelling an explicit relation between changing mean and variance (Haugen et al., 1991; Glosten et al., 1993; Christof-fersen and Diebold, 2006; Buraschi et al., 2014), often termed leverage effect or

asymmetric volatility. Put in simple terms, these concepts describe the observed negative correlation between past returns and future return variance. While this effect can also be modelled as smoothly varying conditional mean and variance, existing literature on threshold- and Markov-switching models aims at capturing the empirically observed sudden changes in the mean and variance of financial returns, see. e.g. Rabemananjara and Zakoian (1993); Gray (1996); Ardia (2009) and Rohan and Ramanathan (2012). However, our goal here is more generally the detection of change points without assuming an explicit relation between changes in mean and in variance.

Nonparametric change-point detection methods that do not require the assumption of a distributional form fully described by a set of parameters (see Section 2.2.1) are an alternative suitable for the unknown change-type situation in the process of Equation (4.1). These methods are appropriate for the change-point problem we are interested in as long as they do not assume first or second moment of the data to be fixed (as is the case e.g. for the nonparametric method of Preuß et al., 2015). However, by ignoring information regarding the distribution of data these nonparametric approaches can be expected to have poor power properties compared to parametric approaches that exploit this information. This will be apparent in Section 4.4 of this chapter where we include the distribution-free approach of Matteson and James (2014) in our comparison of methods. The authors' e-divisive algorithm uses a nonparametric test for change-point detection based on differences in empirical distributions nested within binary segmentation. The authors claim that e-divisive consistently estimates change points while it does "not make any assumptions regarding the nature of the change in distribution".

In the case of piecewise-stationary data following a normal distribution on each segment, this nonparametric method yields poor results compared to the PELT algorithm used for detection of (simultaneous) changes in mean and variance and the sequential testing procedure we propose here.

The contribution of this chapter is as follows. First, we illustrate the difficulties in detecting changes in mean and/or variance if different change types are present in a time series. After observing the shortcomings of existing change-point tests under model misspecification, we derive a sequential testing approach that addresses these issues. The procedure is implemented within a binary segmentation-type algorithm that evaluates test statistics on intervals with boundaries $\{(s, e) : 1 \leq s < e \leq T\}$ of varying length, which allows us to detect change points even if they are relatively closely spaced. While the proposed method is computationally demanding, we show in a numerical study that the results can compete with two state-of-the-art approaches, the e-divisive method and the PELT method for detection of simultaneous changes in mean and variance. Our simulation study focuses on the relative performance in what can be summarized as difficult situations, i.e. processes with many change points and generally small change sizes.

4.2 Motivation and Toolbox

4.2.1 Statistical Framework for Known Change Types

Maximum likelihood (ML) is the arguably most common approach to change-point detection in the mean and/or the variance of normal data, and we will use three

types of ML estimators in this chapter, for change in mean, change in variance and simultaneous change in mean and variance. The ML change-point estimation methods are well-established for a range of situations. For changes in a univariate normal process, see e.g. Csörgö and Horváth (1997); Jandhyala and Fotopoulos (1999) or Jandhyala et al. (2002). We state the main results for the single change-point case below.

As outlined previously (see p. 33), for $\mathbf{x}$ following a parametric distribution $\mathcal{F} = \bar{\mathcal{F}}(\boldsymbol{\theta})$ on an interval $\{s, \ldots, e\}$ the principle of likelihood estimation is apparent in the multiplicative partitioning

$$\mathcal{L}(\boldsymbol{\theta}|x(s), \ldots, x(e)) = \mathcal{L}(\theta_1, \ldots, \theta_Q|x(s), \ldots, x(e)) = \prod_{t=s}^{e} \boldsymbol{f}(x(t)|\boldsymbol{\theta})$$

where $\boldsymbol{f}(\cdot|\boldsymbol{\theta})$ denotes the pdf or pmf of $\bar{\mathcal{F}}$ conditional on the parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_Q)'$. The likelihood function is maximized to estimate the parameter vector $\boldsymbol{\theta}$ given the time series on the interval $\{s, \ldots, e\}$ and assuming some distributional form. If testing for a single change point, one is interested in comparing the likelihood for a model assuming $N = 1$ to that of $N = 0$ given the time series $\mathbf{x} = \{x(t), t = 1, \ldots, T\}$. The model for $N = 1$ is chosen to maximize the likelihood over all change-point candidates b, defining the likelihood ratio

$$\mathcal{LR}(\mathbf{x}) = \frac{\sup_b \sup_{\boldsymbol{\theta}_i \in \boldsymbol{\Theta}_i, i=\{1,2\}} \mathcal{L}(\boldsymbol{\theta_1}|x(t), t = 1, \ldots, b)\mathcal{L}(\boldsymbol{\theta_2}|x(t), t = b+1, \ldots, T)}{\sup_{\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\theta}_0|x(t), t = 1, \ldots, T)}$$

where $\boldsymbol{\Theta}$ and $\boldsymbol{\Theta}_i, i = \{1, 2\}$, denote, respectively, the parameter space for the 'no-change' and 'change' cases. We use the notation $\mathcal{LR}(\mathbf{x}|b)$ to indicate conditionality on a change-point candidate, i.e. $\mathcal{LR}(\mathbf{x}) = \sup_b \mathcal{LR}(\mathbf{x}|b)$. The likelihood function

of random variables $\mathbf{x} = \{x(t), t = s, \ldots, e\}$ for the normal density $\boldsymbol{f}_N(\cdot)$ with parameters μ and σ^2 is

$$\mathcal{L}(\mu, \sigma | x(s), \ldots, x(e)) = \Pi_{t=s}^e \boldsymbol{f}_N(x(t)|\mu, \sigma) = \Pi_{t=s}^e 1/(\sqrt{2\pi}\sigma) \exp\{-\frac{(x(t) - \mu)^2}{(2\sigma^2)}\}.$$

Given a change-point candidate b, we are interested in three different change-point tests:

1. simultaneous change in mean and variance, i.e. at point b the mean changes from μ_1 to μ_2 and the variance changes from σ_1^2 to σ_2^2

2. change in variance from σ_1^2 to σ_2^2 (under constant mean μ_0)

3. change in mean from μ_1 to μ_2 (under constant variance σ_0^2)

The resulting log-likelihood ratio test statistics comparing either of the three change-point models for a given b against the no-change model are, scaled for

ease of notation,

$$\frac{1}{2}\mathcal{LLR}^{(k)}(\mathbf{x}|b) = -b\ln\left(\hat{\sigma}^2_{(k),1}(b)\right) - (T-b)\ln\left(\hat{\sigma}^2_{(k),2}(b)\right) + T\ln\left(\hat{\sigma}^2_0\right), \quad k = 1, 2,$$

(4.2)

$$\frac{1}{2}\mathcal{LLR}^{(3)}(\mathbf{x}|b) = -T\ln\left(\hat{\sigma}^2_{(3)}(b)\right) + T\ln\left(\hat{\sigma}^2_0\right) \qquad \text{with}$$

(4.3)

$$\hat{\sigma}^2_0 = 1/T\sum_{t=1}^{T}(x(t)-\hat{\mu}_0)^2$$

(4.4)

$$\hat{\sigma}^2_{(1),1}(b) = 1/b\sum_{t=1}^{b}(x(t)-\hat{\mu}_1)^2 \qquad \hat{\sigma}^2_{(1),2}(b) = 1/(T-b)\sum_{t=b+1}^{T}(x(t)-\hat{\mu}_2)^2$$

(4.5)

$$\hat{\sigma}^2_{(2),1}(b) = 1/b\sum_{t=1}^{b}(x(t)-\hat{\mu}_0)^2 \qquad \hat{\sigma}^2_{(2),2}(b) = 1/(T-b)\sum_{t=b+1}^{T}(x(t)-\hat{\mu}_0)^2$$

(4.6)

$$\hat{\sigma}^2_{(3)}(b) = 1/T\left(\sum_{t=1}^{b}(x(t)-\hat{\mu}_1)^2 + \sum_{t=b+1}^{T}(x(t)-\hat{\mu}_2)^2\right)$$

(4.7)

$$\hat{\mu}_0 = 1/T\sum_{t=1}^{T}x(t) \qquad \hat{\mu}_1 = 1/b\sum_{t=1}^{b}x(t) \qquad \hat{\mu}_2 = 1/(T-b)\sum_{t=b+1}^{T}x(t)$$

(4.8)

where the subscripts $(k), k = \{1, 2, 3\}$ denote, respectively, the models for change in mean and variance, change in variance and change in mean. We compare the test statistics at the supremum over candidates b, i.e. $\mathcal{LLR}^{(k)}(\mathbf{x}) = \sup_b \mathcal{LLR}^{(k)}(\mathbf{x}|b), k = \{1, 2, 3\}$. In classical LR testing, these statistics are standardized, see e.g. Csörgö and Horváth (1997) for details and a further discussion. The intuition behind this family of statistics is straight-forward: all are maximized where, under the respective assumptions of changes in mean and variance, variance only or mean only, the weighted log-squared deviation $\hat{\sigma}^2_{(k)}(b), k = \{1, 2, 3\}$ from the sample mean $\hat{\mu}_j, j = \{0, 1, 2\}$ is minimized. The differences between the three

alternatives lie in the weighting scheme and in whether the local sample mean $\hat{\mu}_j$ is assumed to be constant ($j = 0$ corresponding to the change in variance only, $k = 2$) or to change at the single change point ($j = 1$ up to point b and $j = 2$ thereafter).

We use the LR testing approach because the principle is established in the literature and provides explicit tests for the three change types that can occur in an i.i.d. normally distributed process. Under normality and assuming that the type of change is known, the likelihood is well-defined and LR testing is preferable to nonparametric methods in the sense that it exploits the distributional information. This also holds for multiple change-point detection, where one may deploy binary segmentation (Jandhyala et al., 2013; Fryzlewicz, 2014) or dynamic programming approaches (Killick et al., 2012; Maboudou-Tchao and Hawkins, 2013; de Castro and Leonardi, 2015). The following section illustrates the challenges for using the LR procedures outlined above when there are different types of change points present. Similar issues arise from any change-point test for a given change type if applied under wrong assumptions.

4.2.2 Estimation of a Change-Point Location under Model Misspecification

Assume we observe a realization of the mixed-change type process in Equation (4.1). Without knowledge on the types of changes contained in the time series, the question arises which testing procedure to use. We consider first the change-point location estimate, i.e. the point $b^* = \text{argmax}_b \mathcal{LLR}^{(k)}(\mathbf{x}|b)$ for $k = \{1, 2, 3\}$, and

illustrate the effect of model misspecification (in the sense of not accounting for the presence of different change-point types) by means of simple examples with a small number of change points; c.f. Pitarakis (2004) who illustrates in a series of simple simulation studies the difficulties arising when testing for a change in mean in the presence of a change in variance, and vice versa, for least-squares estimation. The next section discusses the effect of model misspecification on the significance level or threshold to which a test statistic is compared.

Example 1:   Figure 4.3 shows a realization of the mixed-change process of Equation (4.1). In the upper panel, a time series with a change in mean and constant variance is displayed. The mean changes at $\eta_1 = 300$ from 0 to 1. The figure also shows the change-in-mean test statistic $\mathcal{LLR}^{(3)}(\mathbf{x}|b)$, rescaled to fit the y-axis range. The lower panel shows a realization generated using the same random seed but containing an additional change in the variance, which increases from 1 to 3^2 at $\eta_2 = 400$. If applied in the appropriate context, i.e. under constant variance, the change in mean can be detected with high precision. However, under the presence of a change in variance, the change-in-mean test statistic can fail in providing a consistent estimate; in the process realization shown, its maximum is far from the true change-point location of the change in mean. Note also that it does not fall near the location of the change in variance.

Example 2:   Figure 4.4 shows a similar example. Here, in the upper panel the variance changes at $t = 700$ from 1.75^2 to 3^2 and the mean is constant. In this case, the log-LR statistic for a change in variance, $\mathcal{LLR}^{(2)}(\mathbf{x}|b)$, achieves its maximum near the single change point. However, in the bottom panel, where we show a
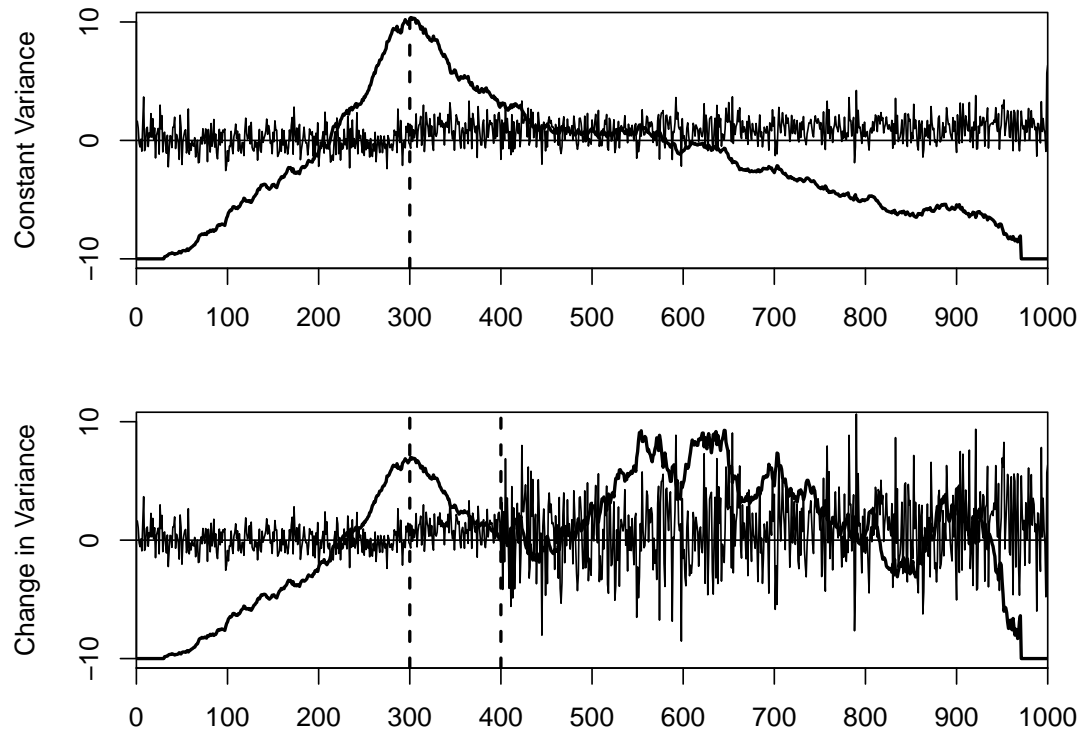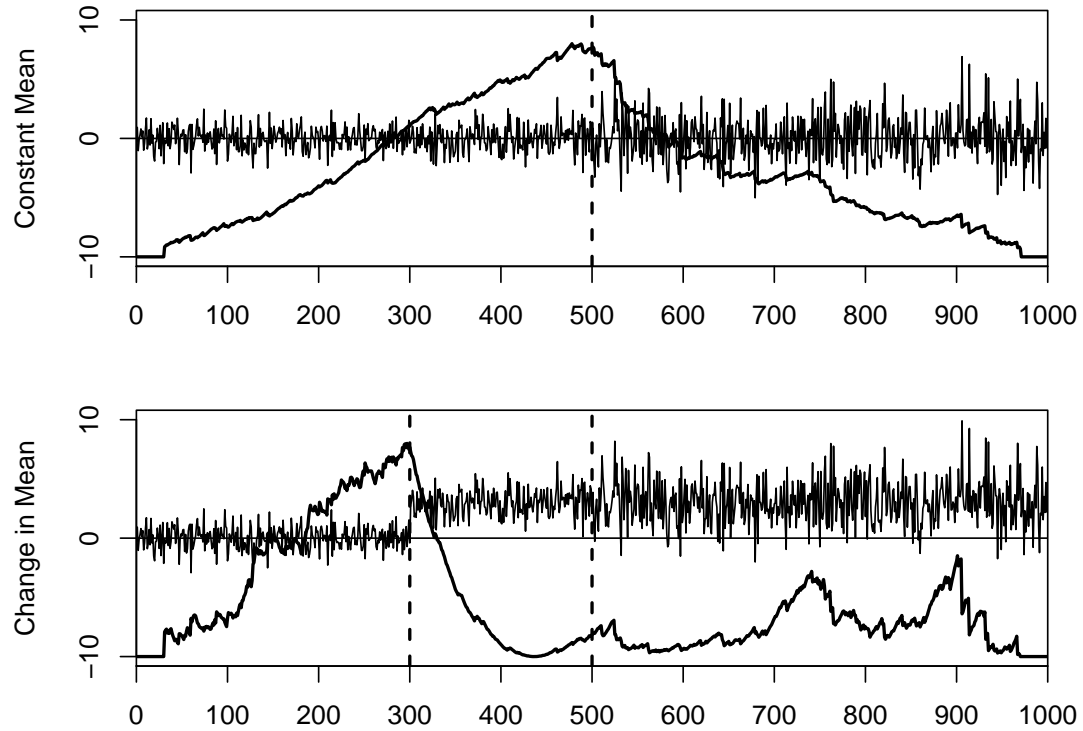
Figure 4.3: Realization of the process in Equation (4.1) and change-in-mean test statistic $\mathcal{LLR}^{(3)}(\mathbf{x}|b)$, rescaled to fit. Upper panel: change in mean at $\eta_1 = 300$ from 0 to 1, under constant variance. Lower panel: change in mean at $\eta_1 = 300$ from 0 to 1 and change in variance at $\eta_2 = 400$ from 1 to 3^2
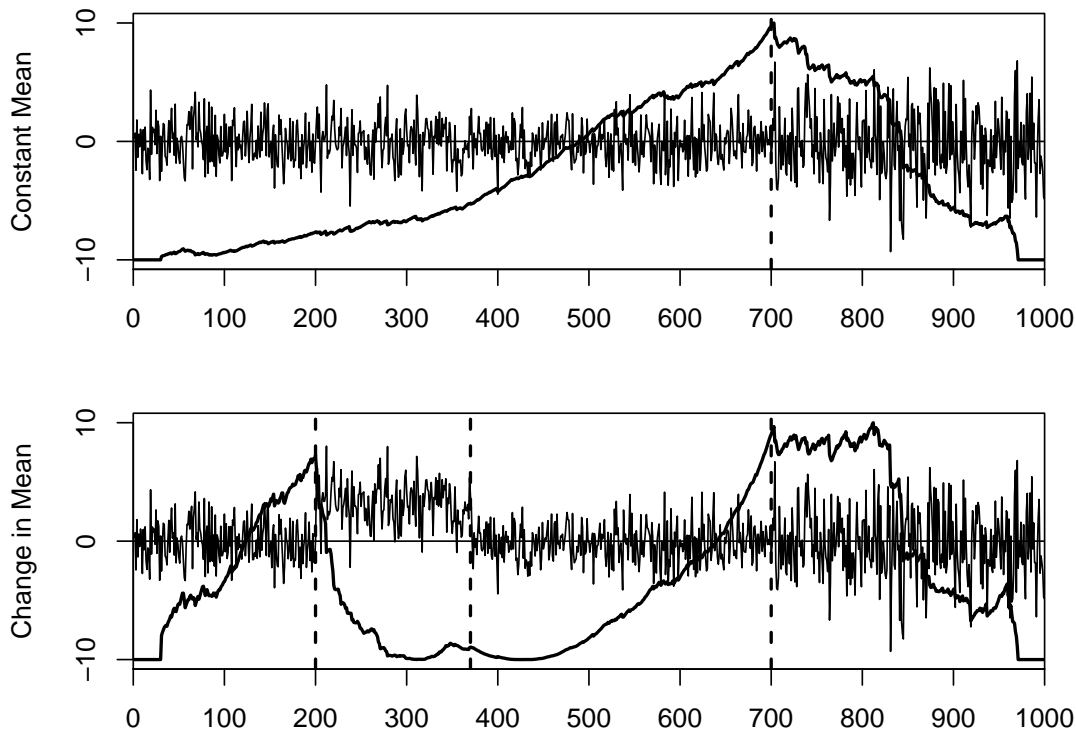
Figure 4.4: Realization of the process in Equation (4.1) and change-in-variance test statistic $\mathcal{LLR}^{(2)}(\mathbf{x}|b)$, rescaled to fit. Upper panel: change in variance at $\eta_1 = 700$ from 1.75^2 to 3^2, constant mean. Lower panel: change in variance at $\eta_3 = 700$ from from 1.75^2 to 3^2 and changes in mean from zero to 3.25 and back at, respectively, $\eta_1 = 200$ and $\eta_2 = 370$.

realization of the same process but containing two changes in the mean, from 0 to 3.25 at 200 and back to zero at 370, the statistic $\mathcal{LLR}^{(2)}(\mathbf{x}|b)$ peaks at the point $t = 812$, i.e. not near to any of the change points.

Example 3:   We include the third example to point out that changes in the mean generate changes in the variance. This implicit relation can yield the identification of changes in mean instead of changes in variance, if both are present. Figure 4.5 shows a similar example: in the upper panel the variance changes variance

Figure 4.5: Realization of the process in Equation (4.1) and change-in-variance test statistic $\mathcal{LLR}^{(2)}(\mathbf{x}|b)$, rescaled to fit. Upper panel: change in variance at $\eta_1 = 500$ from 1 to 2^2, constant mean. Lower panel: change in variance at $\eta_2 = 500$ from 1 to 2^2 and change in mean at $\eta_1 = 300$ from 1 to 3

changes half-way in time from 1 to 2^2 and the mean is constant. The log-LR statistic for a change in variance, $\mathcal{LLR}^{(2)}(\mathbf{x}|b)$, achieves its maximum near the single change point. However, in the bottom panel, where we show a realization of the same process but containing a change in mean from 0 to 3 at 300, the statistic $\mathcal{LLR}^{(2)}(\mathbf{x}|b)$ peaks near the change in mean, not the change in variance.

Example 4: Finally, we turn to the case of simultaneously changing mean and variance. Figure 4.6 shows a process realization with single change point

$\eta_1 = 100$, at which the mean changes from 0 to 1 and the variance changes from 1 to 1.5^2. The upper panel shows the time series, the lower panel the test statistics $\{\mathcal{LLR}^{(k)}(\mathbf{x}|b), k = 1, 2, 3\}$, with, respectively, bold continuous, thin continuous and bold dotted lines indicating test statistics for change in mean and variance, change variance and change in mean. In this illustration, neither the test for a change in mean (peaking around 550) nor that for a change in variance (peaking around 230) would indicate a change near the true change-point location. In this setting and for the same random seed a change only in mean or only in variance can be localised by the respective test statistic shown Figures B.1 and B.2 in the appendix.

In conclusion, the simple examples above show how challenging change-point detection under unknown change types can be. Specifically, we find that

- under changing variance, estimates for changes in the mean can be severely biased;

- in the presence of mean changes and variance changes, test statistics for change in variance can be biased or identify changes in mean instead of those in the variance;

- there exist situations where a simultaneous change in mean and variance cannot be identified by single-parameter test statistics, but can be identified by a test for simultaneous changes.

Returning to the example of Picard et al. (2005) who compare segmentation results testing for changes in mean (assuming constant variance) to those testing for simultaneous changes in mean and variance for a biomedical data set, it becomes
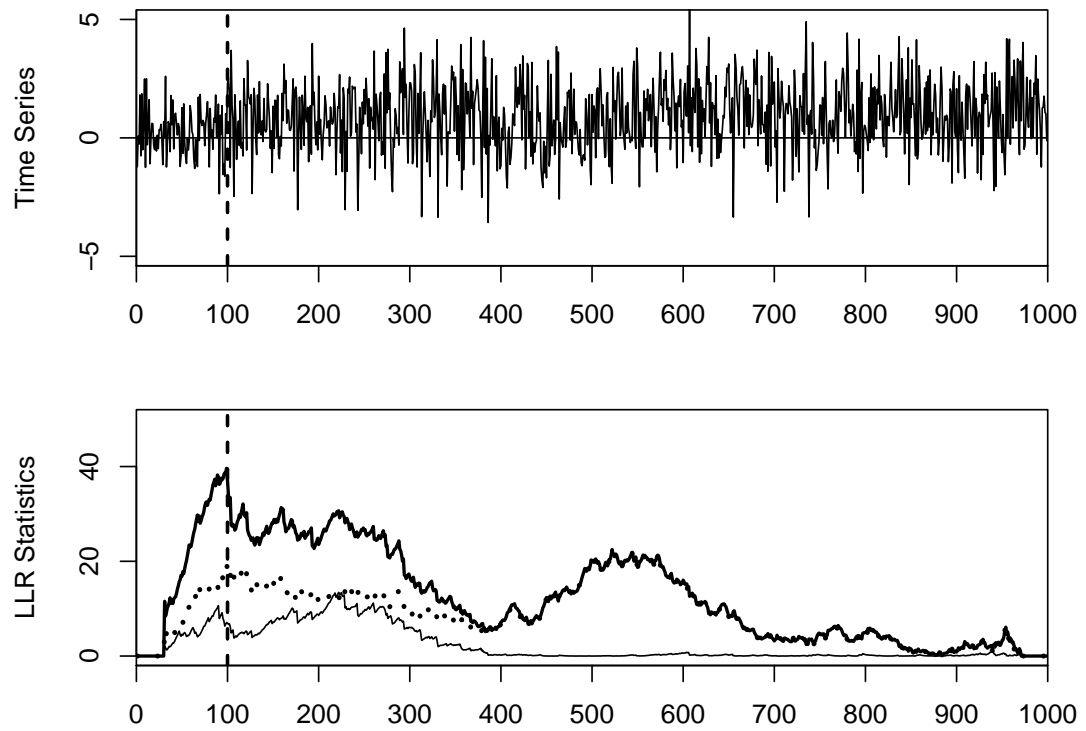
Figure 4.6: Realization of the process in Equation (4.1) with simultaneously changing mean and variance. At $\eta_1 = 100$ the mean increases from 0 to 1 and the variance from 1 to 1.5^2. Upper panel: process realization. Lower panel: test statistics $\mathcal{LLR}^{(k)}(\mathbf{x}|b), k = \{1, 2, 3\}$. Continuous thick line: change in mean and variance, $k = 1$. Continuous thin line: change in variance, $k = 2$. Dotted thick line: change in mean, $k = 3$

clear that this kind of comparison can result in misleading interpretation of data. In fact, if the presence of a sudden change in variance can be suspected in the data, the results of tests for a change in mean only are questionable.

While, based on the above, testing for simultaneous change points may appear sufficient, we illustrate in a simulation study in Section 4.4 that the mean-and-variance test statistic is less powerful in detecting changes of one parameter only.

For processes containing an unknown number of change points with unknown type, we propose a testing procedure where the log-LR statistics are used sequentially. The full SeqToI method is described formally in Section 4.3. The intuition behind sequential testing is as follows: within a binary-segmentation type algorithm, the procedure first tests for simultaneous changes in mean and variance. When no more change points of this first type can be detected (by a criterion we discuss in the next section), it tests the segments between adjacent change-point estimates for changes in the variance only. Finally, accounting for both simultaneous and variance-only change-point estimates, the segments between adjacent change-point estimates are tested for change points in the mean only.

Through this sequence of testing, SeqToI overcomes the three issues enlisted above: first, it tests for changes in mean and variance. At this stage all simultaneous change points are detected with high probability, thus addressing the risk of missing a simultaneous change by using only change-in-mean or change-in-variance tests, as illustrated in Figure 4.6. At the next stage, SeqToI tests for changes in variance only accounting for previously detected change points (in a way we specify in the next section). This means that at the third stage with high probability the set of previously detected change points includes all simultaneous changes in mean

and variance and all changes in variance only. Accounting for these, testing for a change in mean only can proceed without risking a bias as described in Figure 4.3. The next section introduces the tool of using varying-length intervals for the computation of the test statistics, which reduces the issue arising from a bias in the variance test statistic due to the presence of changes in the mean, as illustrated in Example 2. We do not claim that SeqToI identifies specific change types at the respective stage; in fact the procedure can detect, for instance, a change in mean only at any of the three stages. However, by accounting for change points detected at previous stages SeqToI avoids double-counting a change point, as we describe in the next section.

The top panel of Figure 4.7 shows a time series containing five change points: two simultaneous changes of mean and variance, one of the variance only and two in the mean only. We illustrate here the sequential approach to change-point detection using SeqToI: the second panel shows two steps in the estimation using the change in mean and variance statistic, $\mathcal{LLR}^{(1)}(\mathbf{x}|b)$. The third panel shows the $\mathcal{LLR}^{(2)}(\mathbf{x}|b)$ statistic achieving its maximum at the variance change point. The bottom panel shows the $\mathcal{LLR}^{(3)}(\mathbf{x}|b)$ statistic accounting for previously detected change points. Note that this illustration only shows the sequential testing component of SeqToI without considering varying-length intervals discussed in Section 4.3, as the full algorithm does. Instead, for ease of illustration the logic of sequential testing is illustrated in Figure 4.7 combined with standard binary segmentation as in Chapter 3.
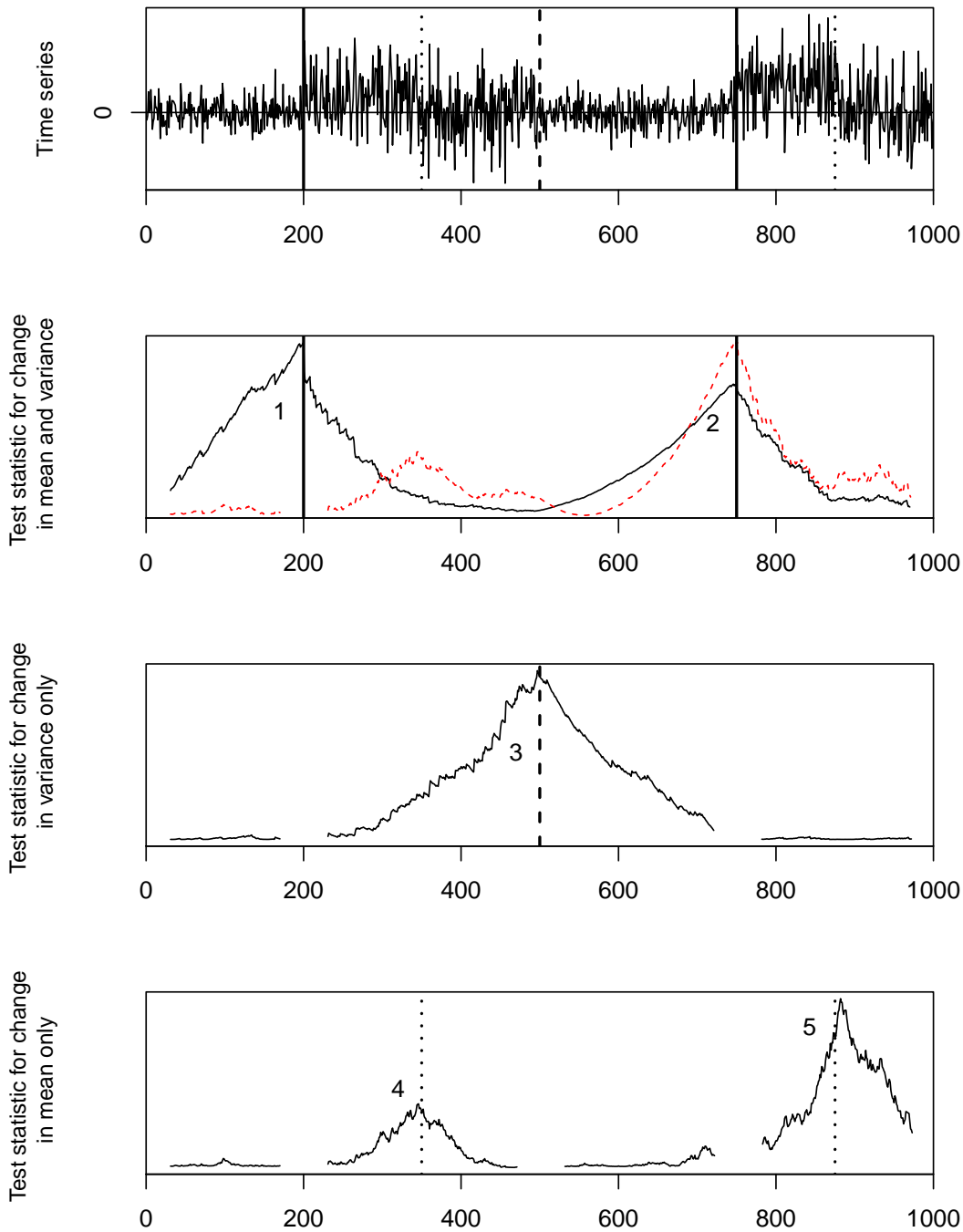
Figure 4.7: Illustration of sequential testing approach. Top panel shows data with various change point locations as vertical lines (continuous: mean and variance, broken: variance only, dotted: mean only). The lower three panels show the resulting test statistics assuming $\delta_T = 30$, rescaled for the purpose of this illustration. The sequential approach proceeds as follows: first, all significant change points in mean and variance are detected (2nd panel). Conditional on these changes, the method tests for changes in the variance only (3rd panel). Finally, accounting for all previously detected change points, the method tests for changes in the mean only (bottom panel). Note that the full SeqToI method formally introduced on the following pages extends the approach shown here to testing on varying-length intervals using a permutation-based approach to estimate thresholds, which is not reflected in this illustration.

4.2.3 Thresholding under Model Misspecification

In classical change-point detection using likelihood ratio test statistics, the distribution of the test statistic under the null hypothesis of no change is used to derive a conclusion about the significance of a candidate change point. As we illustrated in the previous section, misspecification of this null hypothesis can lead to substantial estimation bias. As alternative, we use an data-adaptive approach to threshold estimation. Threshold permutation has received much attention in the literature and we mention a few developments in Section 2.2.1. Below we describe a simple permutation approach based on results of Arias-Castro et al. (2015) who develop a permutation-based thresholding procedure to detect subsets of abnormalities in an observed time series with unknown distribution. This data-adaptive approach also allows for the application of SeqToI to a setting with non-Gaussian noise within a quasi maximum likelihood (QML) framework. The principles of QML for change-point detection have been discussed in, e.g., Nunes et al. (1995); Braun et al. (2000); Bai (2000). Bai (2010) propose QML for the detection of a single change in mean and/or variance in panel data with serially uncorrelated noise. Bardet et al. (2012) use a data-adaptive procedure to estimate the appropriate thresholds for change-point detection in a general class of ARCH models using QML.

4.3   Sequential Testing on Intervals for Change-Point Detection

In this section, the Sequential Testing on Intervals (SeqToI) procedure for change-point detection is formally introduced. We begin by describing its components.

Varying-length Intervals.   We employ a tool for change-point detection introduced recently by Fryzlewicz (2014) as Wild Binary Segmentation (WBS) to strengthen our sequential testing method's performance. The idea is to randomly draw R intervals contained in $\{1, \ldots, T\}$, $\mathbb{RI}_0 = \{(s_r, e_r) :\ 1 \leq s_r < e_r \leq T, r = 1, \ldots, R\}$ and to compute the log-LR test statistics on each of these intervals, $\mathcal{LLR}^{(k)}(x(s_r), \ldots, x(e_r)) \equiv \mathcal{LLR}_r^{(k)}(\mathbf{x}), k = \{1, 2, 3\}, r = \{1, \ldots, R\}$. As illustrated in Fryzlewicz (2014) for the case of detection of changes in the mean only, this approach is superior to standard BS when change points are close to another. We can extend the argument here be noting that if R is large, with high probability we have for any change point some interval containing only this point. In this case, the test statistics are applied under the correct assumptions and can with high probability detect the respective change point. In principle, a deterministic rule to select the R intervals can be equally valid provided R is large enough.

Data-Adaptive Thresholding via Permutation.   As previously discussed, the appropriate thresholds of the test statistics $\{\mathcal{LLR}_r^{(k)}(\mathbf{x}),\ k = 1, 2, 3,\ r = 1, \ldots, R\}$ depend on the presence and type of other change points in an interval $\{s_r, \ldots, e_r\}$. We therefore centre our attention here on the data-adaptive estimation of thresh-

olds via permutation. This approach has been previously developed in the context of change-point detection by Antoch and Huskova (2001), who also argue in favour of permutation-based thresholding for possibly multiple change points in small and moderate sample sizes, as those resulting from randomly drawing intervals on $\{1, \ldots, T\}$, even if T is large. The authors discuss asymptotic properties for tests of changes in mean and/or variance, as well as situations where multiple change points are present. We construct the permutation-based thresholds as follows.

On each interval defined by $(s_r, e_r) \in \mathbb{RI}_0$, permute the data P times at random, generating $\{x^{(p)}(t_r^*)\}$ where the set $\{t_r^*\}$ is a permutation of $\{s_r, \ldots, e_r\}, \forall p$. Then compute the test statistics $\{\mathcal{LLR}_r^{(k,p)}(\mathbf{x}), \ k = 1, 2, 3, \ r = 1, \ldots, R\}$. For each test statistic k and interval r we use the $(1 - \alpha)$-quantile of the empirical distribution of the statistic over permutations $q_{r,\alpha}^{(k)} = \mathrm{Quant}^{\alpha\%}(\mathcal{LLR}_r^{(k,p)}(\mathbf{x}), p = 1, \ldots, P)$ to evaluate significance of a change-point candidate. This is done by computing per test type k the differences between the test statistic $\mathcal{LLR}_{r^*}^{(k)}(\mathbf{x})$ and the quantile $q_{r^*,\alpha}^{(k)}$ for all admissible intervals r^*, where admissibility depends on the stage of the procedure as we explain below. To decide on the significance of a change-point candidate at any stage of the algorithm, the maximum of such a set of differences $d_\alpha^{(k)} = \max_{r^*} \mathcal{LLR}_{r^*}^{(k)}(\mathbf{x}) - q_{r^*,\alpha}^{(k)}$ is then compared to a threshold $\zeta^{(k)}$.

Sequential Testing. We recap the sequential testing approach outlined in the previous section. For ease of notation, let $\hat{\mathbb{N}}^{(0)} = \{0, T\}$. First, the procedure estimates change points using the mean-variance log-LR statistics $\mathcal{LLR}_r^{(1)}(\mathbf{x})$. Denoting the set of these change-point estimates $\hat{\mathbb{N}}^{(1)}$, at the second stage SeqToI estimates variance change points on all segments $\{\hat{\eta}_i + 1, \ldots \hat{\eta}_{i+1}\}$ between any

two adjacent points in $\cup_{k=0}^{1}\hat{\mathbb{N}}^{(k)}$ considering only intervals r^* contained in a given segment. Once no more change points can be estimated using the variance statistic $\mathcal{LLR}_r^{(2)}(\mathbf{x})$, the procedure reaches its third and final stage where it searches for change points in the mean using $\mathcal{LLR}_r^{(3)}(\mathbf{x})$ on all segments between adjacent points in $\cup_{k=0}^{2}\hat{\mathbb{N}}^{(k)}$, again considering for any segment only intervals r^* that are contained in the segment. With high probability, this sequence of testing avoids the potential issues arising from model misspecification as illustrated in Section 4.2. As in Chapter 3 and illustrated in Figure 4.7, we do not consider points within δ_T-distance of the interval boundaries of any interval r as candidate change points.

The SeqToI procedure combines the components of sequential testing, random intervals and permutation-based thresholding as follows. Initiate the set of change point estimates $\hat{\mathbb{N}} = \emptyset$ and the set of random intervals $\mathbb{RI}_{1,T} = \mathbb{RI}_0$.

I. In this stage, $k = 1$. Consider the test statistics $\{\mathcal{LLR}_r^{(k)}(\mathbf{x}), r = 1, \ldots, R\}$.

   i. Over all r, let $d_\alpha^{(k)} = \max_r \mathcal{LLR}_r^{(k)}(\mathbf{x}) - q_{r,\alpha}^{(k)}$.

   ii. If $d_\alpha^{(k)} > \zeta^{(k)}$, add $\hat{\eta}_1 = \operatorname{argmax}_b \mathcal{LLR}_{\bar{r}}^{(k)}(\mathbf{x}|b)$ to $\hat{\mathbb{N}}$, where $\bar{r}$ is the index corresponding to $d_\alpha^{(k)}$. Proceed to step (I.iii). Else proceed to the next stage.

   iii. Repeat the above in a binary-segmentation system: do step (I.i) only considering intervals in $\mathbb{RI}_{1,\hat{\eta}_1} = \{(s_r, e_r) \in \mathbb{RI}_0 : 1 \le s_r < e_r \le \hat{\eta}_1\}$, i.e. fully contained in $\{1, \ldots, \hat{\eta}_1\}$, and only considering intervals in $\mathbb{RI}_{\hat{\eta}_1+1,T} = \{(s_r, e_r) \in \mathbb{RI}_0 : \hat{\eta}_1 + 1 \le s_r < e_r \le T\}$.

II. Update the set of random intervals accounting for previously detected change points, $\mathbb{RI}_{1,T} = \{(s_r, e_r) \in \mathbb{RI}_0 : (s_r, e_r < \hat{\eta}) \text{ OR } (s_r, e_r > \hat{\eta}), \forall \hat{\eta} \in \hat{\mathbb{N}}; \ 1 \le$

$s_r < e_r \leq T\}$. Repeat (I.i-iii) for $k = 2$.

III. Repeat (II.) for $k = 3$. Stop when no more change points are detected.

Some remarks are in order.

Remark. SeqToI does not necessarily identify only changes in mean and variance at the first stage, or only changes in the variance at the second stage. For instance, in a simple setting with a single change point corresponding to a change in mean the procedure could detect the change point at any of the three stages: a) if sufficiently pronounced, it could be detected when testing for simultaneous changes in mean and variance using $\mathcal{LLR}_r^{(1)}(\mathbf{x})$, b) it could be detected by $\mathcal{LLR}_r^{(2)}(\mathbf{x})$, as illustrated in Section 4.2.2, and c) if not detected at a previous stage, with high probability it will be detected by $\mathcal{LLR}_r^{(3)}(\mathbf{x})$ (Fryzlewicz, 2014). Moreover, the exclusion of intervals containing previously detected change points, consistently applied over all stages, prevents SeqToI from spuriously detecting a change point twice: if a change point was detected at the some stage, at any later stage no interval is considered that contains this point at any location other than its end point.

Remark. The proposed procedure allows us to overcome the problems described in Section 4.2 which could not be addressed by a simultaneous detection method that, say, applies the three types of change-point tests in parallel and then merges the resulting sets of change points. Such a simultaneous-testing approach raises two questions. First, whether and how to merge change points if they were detected by different tests and are close to another. Second, how to prevent biases from testing under model misspecification, for instance biased estimation of the location of a

change in mean because changes in the variance are not accounted for. Therefore it is not clear how such parallel detection could be implemented.

Remark. The total number of intervals between neighbouring change-point estimates $R_{\hat{\mathbb{N}}} = \sum_i |\mathbb{RI}_{\hat{\eta}_i, \hat{\eta}_{i+1}}|$ (with $|\cdot|$ denoting the cardinality of a set), decreases monotonically with every newly detected change point. If $R_{\hat{\mathbb{N}}}$ is small we can expect that the performance of SeqToI is negatively affected. However, the decrease is data-adaptive in the sense that, for instance, if a time series contains no change in the variance, no further decline of $R_{\hat{\mathbb{N}}}$ can be expected at Stage (II) of SeqToI (other than by detecting a change in mean only that would otherwise be detected in the following stage).

4.4 Simulation Study

4.4.1 Data Generating Process

SeqToI is compared to two competitor methods mentioned in Section 4.1, the PELT algorithm and the e-divisive algorithm, and to a simplified version of itself where the algorithm terminates after Stage (I) and thus tests only for simultaneous changes in mean and variance using $\mathcal{LLR}^{(1)}(\mathbf{x})$, abbreviated MV in the following. Our goal is to show strengths and limitations of SeqToI in situations with varying density of change points and relatively small change sizes, i.e. changes that appear to be generally difficult to detect. In particular, we choose a mixed-change type setting that allows us to trace back the source of particular poor or strong overall

performance with respect to the change type it detects poorly or very well.

We generate piecewise-stationary processes with length $T = 1000$, normal errors and varying change-point characteristics. We consider the cases of $N = \{3, 6, 9, 12\}$ change points with equal spacing on $\{1, \ldots, T\}$, corresponding to low, medium-low, medium-high and high change-point density. For ease of interpretability, the changes in the first third of the time series are changes in the mean only, those in the second third are simultaneous changes in mean and variance and those in the final third are changes in the variance only. We sample from a function-plus-noise process as in Equation (4.1) with fixed change sizes and block structure, i.e. a positive change in a quantity is followed by a negative change of identical magnitude. Table 4.1 summarizes the change sizes for mean and variance. These are chosen so that at any change-point density level changes are difficult to detect, with the most obvious appearing at in the high-density setting ($N = 12$) where the close spacing between change points poses a challenge to detection methods. To illustrate, we sample an ($N = 3$)-process with a change in mean at $t = 250$ from 1.0 to 1.5, a change in mean and variance at $t = 500$, where the mean changes from 1.5 to 1 and the variance changes from 1 to 1.5^2, and a change in the variance back to unity at $t = 750$. We sample from each process 500 times.

N	$|\mu_i - \mu_{i-1}|$	$|\sigma_i - \sigma_{i-1}|$
3	0.5	0.5
6	0.8	0.8
9	1.0	1.0
12	2.0	2.0

Table 4.1: Parameter settings for the simulation study; base level is 1.0

4.4.2   Competitors

We compare the performance of SeqToI to a parametric and a nonparametric competitor method.

PELT

The PELT method of Killick et al. (2012) is devised as computationally efficient dynamic program to detect change points in a global optimization framework. At its core, the algorithm requires the specification of a cost function that is the sum of a measure of goodness of fit (typically twice the negative log-likelihood) and a penalty term against overfitting. The algorithm is based on optimal partitioning but uses a pruning step to improve computational performance and indeed achieves in practice usually a speed of $O(T)$.

PELT can be used with any of the three tests discussed in Section 4.2 that test, respectively, for simultaneous changes in mean and variance, changes in the variance only and changes in the mean only. It is not clear how a combination into a sequential procedure could be possible within the framework of global optimization. In the mixed-change setting we are concerned with here, SeqToI is compared to the output of the function `cpt.meanvar` in the R package `changepoint` (version 2.2) that finds the optimal segmentation assuming simultaneous changes in mean and variance. The PELT method is developed for linear penalty functions and we evaluate it with three readily available choices: a) the standard Schwarz information criterion (SIC) as used in Killick et al. (2012) (see also Chen and Gupta (1999) for a discussion in the context of changes in mean and variance), b) the SIC0 which

is equivalent to the SIC minus the number of change-point estimates $\hat{N}$, i.e. it does not count a change point as additional parameter, and c) the default penalty function, the Modified Bayes Information Criterion (MBIC, Zhang and Siegmund, 2007). The minimum change-point spacing parameter is set to $0.03T$, as for SeqToI. Otherwise the default parameter settings as specified in the R package are used.

E-Divisive

As second competitor considered here is the nonparametric e-divisive (ED) method of Matteson and James (2014). The ED method combines binary segmentation with a nonparametric divergence measure based on the weighted difference of local characteristic functions (Szekely and Rizzo, 2005). They derive the following test statistic for a change at $b \in \{s, \ldots, e\}$.

$$
\begin{aligned}
\mathcal{ED}_{s,b,e}(x|\beta) \;=\; & \frac{(b-s+1)(e-b)}{e-s+1}\left(\frac{2}{(b-s+1)(e-b)}\sum_{t_1=s}^{b}\sum_{t_2=b+1}^{e}|x(t_1)-x(t_2)|^\beta\right.\\
& -\binom{b-s+1}{2}^{-1}\sum_{s\le t_1<k\le b}|x(t_1)-x(k)|^\beta-\binom{e-b}{2}^{-1}\sum_{b<t_2<k\le e}|x(t_2)-x(k)|^\beta\Bigg)
\end{aligned}
$$

with $\beta \in (0,2)$. This measure of difference compares the distribution of $\mathbf{x}$ on the interval $\{s, \ldots, b\}$ to that on $\{b+1, \ldots, e\}$ using the L_β-norm. Arguing that allowing e to vary increases the power of a detection procedure, the change-point estimate of ED on a set of intervals with boundaries $\{(s, e^*), e^* \le T\}$ becomes

$$
(\hat{\eta}_{ED}, \hat{e}) = \operatorname*{argmax}_{(b,e^*)} \mathcal{ED}_{s,b,e^*}(x|\beta). \tag{4.9}
$$

Matteson and James (2014) propose a hierarchical algorithm similar to classical binary segmentation. Given a set of $\tilde{N}$ change points, they estimate the resulting non-overlapping segments $\{\eta_{i-1}+1,\ldots,\eta_i\}$, $\forall i = 1,\ldots,\tilde{N}+1$. Within each of the $(\tilde{N}+1)$ segments they identify a change-point candidate according to Equation (4.9) and then estimate significance using a permutation test. If the candidate is deemed significant, it is included in the set of change-point estimates and the binary segmentation continues. Otherwise, the procedure terminates. The ED method is readily available in the R package `ecp`. We use the default parameter settings and analyse the performance for 1% and 5% significance levels and the parameter β in the centre and at the upper end of its permitted range, $\beta = \{1, 1.99\}$.

4.4.3   Main Results

We discuss here the performance of SeqToI compared to PELT with three different penalty function, ED with a total of four different parameter combinations and the MV-variant of SeqToI that stops after Stage (I) and thus only tests for simultaneous changes in mean and variance within the varying-length interval framework. For SeqToI and MV we set $P = 150$, $R = 20000$, $\alpha = 0.05$ and the threshold vector $\boldsymbol{\zeta} = (\zeta^{(1)}, \zeta^{(2)}, \zeta^{(3)})' = (1.0, 1.5, 1.5)'$, against which the maximum differences $d_\alpha^{(k)}$ are evaluated at every step.

The choice of varying thresholds is motivated as follows. Firstly, within the framework of multiple unknown change types considered here theoretical results are challenging to derive in general, although in principle results from specific problems could be considered for guidance. However, we rely on a data-adaptive

approach where the threshold parameters are chosen within a simpler simulation setting, where we also made the empirical observation that within the sequential testing framework, tests at later stages require a slightly more conservative threshold. The thresholds were chosen data-adaptively as follows: within a framework corresponding to the one described above but letting the number of change points vary between one and twelve and the change sizes increase with the number of change points in a similar pattern as described in Table 4.1, we first identified $\zeta^{(1)}$ close to one to be optimal if there are solely simultaneous changes present in the data. Repeating this but considering changes in mean and variance as well as changes in variance only, we identified that the later-stage threshold should be higher and set it to 1.5, which is close to the optimum. Finally, we found similar results in the general mixed-change situation for $\zeta^{(3)}$. We note that due to the permutation approach deployed here, the exact choice of $\boldsymbol{\zeta}$ does not appear to have a large impact on the overall results: for every stage k of SeqToI, $\zeta^{(k)}$ is not a threshold in the sense of Chapter 3 but instead indicates the required level of exceedance over the data-adaptively chosen critical value $d_\alpha^{(k)}$. We impose a minimum change-point distance $\delta_T = 0.03T$, which means that at every stage of the algorithm we consider only those change-point candidates on (sufficiently long) intervals $\{s_r, \ldots, e_r\}$ that are at least δ_T distant from the boundaries. This is in-line with the corresponding minimum spacings between change points in Killick et al. (2012) and Matteson and James (2014).

Tables 4.2 and 4.3 summarize the results for $N = \{3, 6\}$ and $N = \{9, 12\}$, respectively. We report the proportion of simulation runs where the number of change points is under-, over- and correctly estimated ($\hat{N} <, >, = N$) and, con-

ditional on $\hat{N} = N$, the mean absolute distance between change-point estimate
and change point, $1/N \sum_{i=1}^{N} |\hat{\eta}_i - \eta_i|$, averaged over the number of simulation runs
where $\hat{N} = N$. Moreover, the tables show combined results and results on individ-
ual change types. For instance, with $N = 6$ in the lower part of Table 4.2 the first
block (columns 3-6) summarizes performance over the full generated time series of
length $T = 1000$; the second block summarizes the results on the first third of the
time series to point $t = 333$, an interval on which our simulation generates $N/3 = 2$
change points in the mean only. Similarly, the third block summarizes the meth-
ods' performances on the interval $\{334, \ldots, 667\}$, which contains for $N = 6$ another
two change points, both with simultaneous changing mean and variance. The final
block then contains the performance results of changes in the variance only. For
the $N = 3$ case we find that SeqToI and its simplified version MV perform bet-
ter than the ED method with various parameter choices and also outperform the
PELT-MBIC method (abbreviated MBIC in the following, and similarly for PELT
with SIC and SIC0), in terms of estimating the correct number and location of
change points. In both cases, the competitor method underestimates N; the poor
performance of ED can be traced back to the method's inability to detect changes
in the variance, which is apparent throughout all simulation settings. The changes
in mean only are better detected by ED, while this turns out to be a setting in
which MBIC0 breaks down. However, SIC and SIC0 show a higher percentage of
simulations with overall $N = \hat{N}$ than SeqToI, which can be attributed to higher de-
tection power in the change-in-mean segment. Moreover, we observe that SeqToI
has a slight tendency to overestimate change points on the simultaneous-change
segment. As this is not the case for the MV variation we can attribute this to the
second and third stages of the algorithm.

		Overall				Change in mean only				Change in mean and variance				Change in variance only			
		$\hat{N}<N$	$\hat{N}=N$	$\hat{N}>N$	MAD	$\hat{N}<N$	$\hat{N}=N$	$\hat{N}>N$	MAD	$\hat{N}<N$	$\hat{N}=N$	$\hat{N}>N$	MAD	$\hat{N}<N$	$\hat{N}=N$	$\hat{N}>N$	MAD
Low Change-Point Density, N = 3																	
PELT	SIC	9.0	90.4	0.6	11.3	8.0	91.6	0.4	13.3	0.2	98.2	1.6	6.8	3.0	96.6	0.4	11.2
	SIC0	0.6	90.4	9.0	11.5	1.4	95.0	3.6	13.7	0.0	96.6	3.4	6.7	1.2	95.0	3.8	11.7
	MBIC	52.4	47.6	0.0	10.1	44.8	55.2	0.0	11.3	7.8	91.2	1.0	7.5	19.0	81.0	0.0	11.1
ED	1%,$\beta = 1.00$	92.4	6.4	1.2	35.6	12.4	87.6	0.0	13.5	5.8	90.8	3.4	16.8	90.8	8.2	1.0	23.7
	5%,$\beta = 1.00$	74.4	20.6	5.0	33.4	5.8	93.8	0.4	13.9	2.6	90.2	7.2	17.0	76.4	20.0	3.6	23.4
	1%,$\beta = 1.99$	94.8	4.2	1.0	62.2	19.6	80.4	0.0	11.8	15.2	81.0	3.8	27.2	95.2	3.8	1.0	49.9
	5%,$\beta = 1.99$	85.0	10.2	4.8	64.2	7.2	92.8	0.0	12.7	5.2	84.8	10.0	26.2	89.8	7.8	2.4	53.9
SeqToI		11.8	76.0	12.2	12.8	12.2	87.2	0.6	12.7	0.4	86.2	13.4	7.6	3.2	95.0	1.8	11.8
MV		23.6	76.0	0.4	11.0	19.6	80.4	0.0	12.6	0.4	98.2	1.4	7.8	6.0	93.8	0.2	11.8
Medium-Low Change-Point Density, N = 6																	
PELT	SIC	6.0	93.8	0.2	5.2	5.4	94.4	0.2	6.0	0.0	99.4	0.6	3.8	1.0	99.0	0.0	5.6
	SIC0	0.0	90.0	10.0	5.2	0.4	94.6	5.0	6.0	0.0	96.8	3.2	3.7	0.0	97.4	2.6	5.6
	MBIC	61.2	38.8	0.0	4.8	49.4	50.6	0.0	5.4	3.8	95.4	0.8	4.1	23.6	76.4	0.0	5.4
ED	1%,$\beta = 1.00$	90.8	9.2	0.0	13.7	6.6	93.4	0.0	5.6	3.8	93.4	2.8	8.6	90.4	9.6	0.0	27.0
	5%,$\beta = 1.00$	68.2	29.6	2.2	13.4	2.0	98.0	0.0	5.8	1.0	94.2	4.8	8.5	67.8	31.0	1.2	23.3
	1%,$\beta = 1.99$	98.2	1.6	0.2	23.0	23.6	76.4	0.0	5.5	18.8	79.0	2.2	14.5	97.2	2.6	0.2	49.9
	5%,$\beta = 1.99$	93.0	6.4	0.6	29.0	7.8	92.2	0.0	5.6	8.0	86.8	5.2	13.8	91.8	7.8	0.4	48.9
SeqToI		22.8	65.4	11.8	5.9	14.6	84.8	0.6	6.0	0.8	85.8	13.4	4.0	12.2	87.0	0.8	5.5
MV		41.2	58.8	0.0	5.3	24.6	75.4	0.0	5.9	1.6	97.6	0.8	4.3	21.0	79.0	0.0	5.5

Table 4.2: Performance summary for sparse change-point numbers $N = \{3, 6\}$: proportion of simulation runs where the estimated number of change points $\hat{N}$ is less than, equal to or larger than N, and the mean absolute distance between the estimated and true change-point locations, conditional on $\hat{N} = N$ (MAD). We show overall results and results on intervals for each of the three change types. The parameters of the ED method are significance level and β of Equation (4.9)

		Overall				Change in mean only				Change in mean and variance				Change in variance only			
		$\hat{N}<N$	$\hat{N}=N$	$\hat{N}>N$	MAD	$\hat{N}<N$	$\hat{N}=N$	$\hat{N}>N$	MAD	$\hat{N}<N$	$\hat{N}=N$	$\hat{N}>N$	MAD	$\hat{N}<N$	$\hat{N}=N$	$\hat{N}>N$	MAD
Medium-High Change-Point Density, $N=9$																	
	SIC	8.4	91.2	0.4	3.4	2.6	97.4	0.0	3.7	0.0	99.8	0.2	2.7	6.2	93.4	0.4	3.9
PELT	SIC0	0.2	89.0	10.8	3.5	0.0	96.8	3.2	3.6	0.0	95.4	4.6	2.7	0.4	96.0	3.6	4.0
	MBIC	84.8	15.2	0.0	3.1	51.2	48.8	0.0	3.2	9.0	90.8	0.2	2.8	62.2	37.8	0.0	3.8
	$1\%,\beta=1.00$	99.0	1.0	0.0	22.3	5.0	95.0	0.0	3.5	7.2	90.6	2.2	7.1	99.2	0.8	0.0	11.4
ED	$5\%,\beta=1.00$	91.0	7.6	1.4	11.6	0.6	99.4	0.0	3.6	1.2	92.6	6.2	6.8	92.0	7.8	0.2	13.3
	$1\%,\beta=1.99$	100.0	0.0	0.0	NA	41.0	59.0	0.0	3.0	37.4	60.2	2.4	12.0	99.8	0.2	0.0	29.0
	$5\%,\beta=1.99$	99.2	0.8	0.0	34.0	12.4	87.6	0.0	3.2	16.6	77.0	6.4	11.6	99.6	0.4	0.0	42.8
SeqToI		18.6	42.4	39.0	9.4	4.8	91.0	4.2	3.8	2.2	54.2	43.6	4.8	21.0	74.6	4.4	4.2
MV		90.2	9.8	0.0	3.3	65.0	35.0	0.0	3.2	18.8	81.2	0.0	2.7	66.8	33.2	0.0	3.9
High Change-Point Density, $N=12$																	
	SIC	0.0	100.0	0.0	1.6	0.0	100.0	0.0	1.3	0.0	100.0	0.0	1.5	0.0	100.0	0.0	1.9
PELT	SIC0	0.0	90.4	9.6	1.6	0.0	97.0	3.0	1.3	0.0	95.8	4.2	1.5	0.0	97.4	2.6	1.9
	MBIC	0.8	99.2	0.0	1.6	0.0	100.0	0.0	1.3	0.0	100.0	0.0	1.5	0.8	99.2	0.0	1.9
	$1\%,\beta=1.00$	76.0	23.6	0.4	3.1	0.0	100.0	0.0	0.7	0.0	99.2	0.8	2.7	76.2	23.6	0.2	4.3
ED	$5\%,\beta=1.00$	27.2	70.6	2.2	2.7	0.0	100.0	0.0	0.7	0.0	98.2	1.8	2.7	27.6	71.4	1.0	4.3
	$1\%,\beta=1.99$	100.0	0.0	0.0	NA	2.0	98.0	0.0	0.7	9.6	89.6	0.8	5.3	100.0	0.0	0.0	NA
	$5\%,\beta=1.99$	99.6	0.2	0.2	39.3	0.0	100.0	0.0	0.7	2.0	94.2	3.8	5.2	99.8	0.2	0.0	23.5
SeqToI		6.0	87.6	6.4	1.7	0.0	100.0	0.0	1.3	2.2	92.6	5.2	1.8	3.8	95.0	1.2	1.9
MV		76.4	23.6	0.0	1.7	14.2	85.8	0.0	1.3	13.2	86.8	0.0	1.7	71.0	29.0	0.0	1.9

Table 4.3: Performance summary for dense change-point numbers $N = \{9, 12\}$: proportion of simulation runs where the estimated number of change points $\hat{N}$ is less than, equal to or larger than N, and the mean absolute distance between the estimated and true change-point locations, conditional on $\hat{N} = N$ (MAD). We show overall results and results on intervals for each of the three change types. The parameters of the ED method are significance level and β of Equation (4.9)

Our medium-low density setting with $N = 6$ and change magnitudes of size 0.8 (in the sense of Table 4.1) appears to be a bigger challenge for most methods, and again SIC and SIC0 perform best in terms of estimating the correct number of change points $\hat{N} = N$, followed by SeqToI. Comparing over individual change-types, the three methods SeqToI are the only ones consistently identifying the correct number of change points in more than 80% of simulation runs. In terms of mean absolute distance between change points and their estimates, our method is similarly accurate as the PELT variants.

At the $N = 9$ with change magnitudes of size 1, we find further deterioration of overall performance measures. In this setting, SIC and SIC0 perform clearly better than SeqToI and we attribute the latter's underperformance to a substantial over-estimation of the number of change points in the segment containing simultaneous changes in mean and variance, coupled with a tendency to underestimate the number of change points in the change in variance segment. Finally, for the setting with $N = 12$ and much more pronounced change size of 2, we see all methods except ED with $\beta = 1.99$ performing better than in the $N = 9$ setting, and this is visible for almost every method at each of the individual change types. It is remarkable how all PELT-type methods performs good overall, followed by SeqToI. MV fails to detect the correct number of change points in the variance-only segment, where the benefit of sequential testing appears most obvious, but also underperforms relative to SeqToI in the other two change-type segments.

In summary, we find that SeqToI performs as good or almost as good as the state-of-the-art PELT method for detection of changes in mean and variance with appropriately specified model selection criterion: SIC or SIC0, except for the dense

change-point setting where SIC and MBIC perform best. Just as SeqToI requires the specification of thresholds, PELT requires the choice of a model selection criterion and the optimality of the choice appears to depend on change type, change-point density and change magnitudes. Overall, PELT-SIC appears to be the best method in this comparative study, but this does not hold in specific situations such as the change-in-mean segment in the sparse change-point setting. The ED method misses change points in the variance-only segment in almost all situations and using various parameter setting. Moreover, the ED method shows a tendency to underestimate the number of change points for the change in mean only when there are sparse change points with small change magnitudes.

The simulation settings considered here are designed to pose a challenge to change-point detection methods, with relatively small change magnitudes for a range of change-point density levels. SeqToI does not beat the top state-of-the-art competitor procedures, but performs relatively well in many settings and shows a comparably stable performance. This can be attributed to two drivers: sequential testing that overcomes the misspecification issues discussed in Section 4.2 and testing on intervals, some of which contain only a single change point which means the test statistics are computed without misspecification. To be able to differentiate these two drivers, the MV-variant is considered separately in the simulation study. This variant of SeqToI tests only for simultaneous changes in mean and variance, i.e. it stops after Stage (I) of the SeqToI algorithm.

The results indicate that MV tends to perform better than SeqToI in the segment where mean and variance change simultaneously. Indeed, in this segment SeqToI overestimates the number of change points. Comparing SeqToI to SIC and

SIC0, we can also identify this tendency as cause for underperformance in the sense that the accuracy in estimating change locations is similar and one can argue that if SeqToI was be less prone to overestimation on the mean-and-variance segment, the simulation results would be much closer. However, the added value of sequential testing becomes obvious on segments where only mean or only variance changes, particularly with increasing change-point density. It follows that SeqToI performs better than MV in three of our four simulation settings in terms of identifying the total number of change points.

4.4.4   Computational Considerations

The SeqToI procedure is a computationally elaborate method, requiring the calculation of the test statistics $\mathcal{LLR}_r^{(k)}(\mathbf{x}), k = \{1, 2, 3\}, (P+1)$ times on each interval r, on the original data and the P permutations. However, we note that the computation of test statistics can be parallelised easily and conducted separately from the binary segmentation, making this approach potentially faster than standard binary segmentation. Among the competitors, PELT is by far the fastest method, while ED has a comparable computation time to SeqToI due to its permutation test, which is however nested in a classical binary segmentation algorithm that limits the possibilities for parallelisation.

4.5   Concluding Remarks

This chapter sheds light on an issue that has not been accounted for in Chapter 3 and is generally often overlooked in the development and application of change-point detection methods. Yet, the problem of detection of changes in mean and/or variance posed by the process of Equation (4.1) is relevant in many applications to time series such as the climate and financial data shown in Section 4.1. It is more challenging than detection of changes in mean or variance only and the existing methods we are aware of do not explicitly address this problem in its general formulation in a satisfactory way. Indeed, the mixed change-type situation with no prior assumption on the number, type or location of changes has not received much attention in the change-point detection literature. We provide various examples to show that inappropriate assumptions (such as constant variance when testing for a change in mean) can lead to undesirable outcomes. Consequently, such assumptions have to be formulated carefully and the user of any method should be fully aware of potential risks from model misspecification.

In the situation with changes in mean and/or variance, our analysis suggests that testing for simultaneous changes in mean and variance is preferable, albeit it is less powerful in single change-type (mean or variance) situations. A test for changes in the mean can result in biased estimates if changes in the variance are present, while testing for a change in the variance can be biased or detect a change in mean instead, if both are present and do not coincide. Both single change-type tests can fail to detect a simultaneous change of mean and variance.

We propose an intuitive approach to the problem and illustrate it using Gaus-

sian data with sudden changes. Using ML test statistics within a sequential procedure that first tests for simultaneous changes in mean and variance, then for changes in the variance and finally for changes in the mean and accounting consistently for all previously detected change points the procedure can handle the model misspecification inherent to a mixed-change situation. To the best of our knowledge, this is the first work where the concept of binary segmentation using varying-length intervals (Fryzlewicz, 2014) is applied within a sequential testing framework to detect change points in different characteristics of a nonstationary time series. The procedure proposed here is relatively computationally intense, but as outlined above the calculations can be parallelised. Moreover, it requires the choice of a set of thresholds. We provide the reader with a suggestion for this choice and show how this performs in a range of settings in our simulation study; the derivation of more extensive guidance is left for future work. While we focus on the case of normally distributed data where the process is fully characterised by the sequence $\{(\mu_i, \sigma_i^2), i = 1, \ldots, N + 1\}$, our observation carries over to other, more complex situations. Consequently, in applications to observed time series, the analyst should take these findings into consideration when choosing an appropriate change-point detection method in any application.

4.6   Glossary: Most Essential Notation

Expression	Meaning
Functions	
$O(\cdot)$	Upper bound in the sense that $f(x) = O(g(x)) \leftrightarrow$ $\lvert f(x) \rvert \leq C \lvert g(x) \rvert$ for all $x \geq x_0$ and some $C > 0$
$\text{Quant}^{\alpha\%}(\mathcal{LLR}_r^{(k,p)}(\mathbf{x}), p = 1, \ldots, P)$	$(1 - \alpha)\%$ quantile of the $\mathcal{LLR}^{(k)}$ statistic over all P permutations of $\mathbf{x}$ on interval r
Objects	
$\mathbb{N} = \{\eta_i, i = 1, \ldots, N\}, (\hat{\mathbb{N}}, \hat{\mathbb{N}}^{(k)})$	True (estimated) set of change points in the process $\mathbf{x}$; superscripts (k) indicate stage of change-point detection procedure
$\mathbb{RI}_0 = \{(s_r, e_r) : \ 1 \leq s_r < e_r \leq T\}$	Set of randomly drawn intervals $r = \{1, \ldots, R\}$ on $\{1, \ldots, T\}$
$\mathbf{x} = \{x(t), t = 1, \ldots, T\}$	Time series with changes in the local mean μ_i and/or variance σ_i^2, $i = \{1, \ldots, N + 1\}$
$\mathbf{x}^{(p)} = \{x^{(p)}(t_r^*)\}$	Permutation of $\mathbf{x}$ where the set $\{t_r^*\}$ is a permutation of $\{s_r, \ldots, e_r\}$, for all permutations p and intervals r
$\boldsymbol{\zeta} = (\zeta^{(1)}, \zeta^{(2)}, \zeta^{(3)})$	Threshold parameter array
$\mathcal{LLR}^{(k)}(\mathbf{x})$	Maximized log-likelihood ratio test statistic; $\mathcal{LLR}(\mathbf{x}\lvert b)$ indicates conditionality on a change-point candidate b and the subscripts $(k), k = \{1, 2, 3\}$ denote, respectively, the models for change in mean and variance, change in variance and change in mean
$d_\alpha^{(k)} = \max_{r^*} \mathcal{LLR}_{r^*}^{(k)}(\mathbf{x}) - q_{r^*,\alpha}^{(k)}$	Maximum difference between log-likelihood statistic for the kth change-type and the permutation-based quantile, over all intervals r^*

Chapter 5

Frequency-Specific Change-Point Detection in EEG Data

The goal in this chapter is to develop a practical tool that identifies changes in the brain activity as recorded in electroencephalograms (EEG) by applying a multiple change-point detection algorithm to estimates of spectral features. Our method is devised to detect possibly subtle disruptions in normal brain functioning that precede the onset of an epileptic seizure. In particular, we develop a procedure that detects changes in the frequency bands of autospectra and cross-coherences in multi-channel EEG data. The proposed frequency-specific change-point detection method (FreSpeD) employs a multivariate cumulative sum-type test statistic within a binary-segmentation type algorithm. We demonstrate its advantages against state-of-the-art competitor methods in a simulation study and moreover show the robustness over a range of parameter choices in a sensitivity analysis. When applied to epileptic seizure EEG data, FreSpeD identifies the correct brain region as the focal point of seizure, as well as the timing of the seizure onset. Moreover, FreSpeD detects changes in cross-coherence immediately before seizure onset which indicate an evolution leading up to the seizure onset. These changes are subtle and were not captured by the methods that previously analysed the same EEG data. Thus, the

directly interpretable output of FreSpeD gives additional insights into the complex nature of an epileptic seizure and can support the identification of seizure precursors.

Declaration   This chapter is in parts based on joint work with Hernando Ombao, submitted to the Journal of the American Statistical Association, Applications and Case Studies.

5.1   Introduction

Epileptic seizures have received much attention by the neuroscience community. The analysis of these episodes of abnormal brain activity helps researchers understand disruptions in normal brain functioning with the ultimate goal of developing more precise diagnosis, improved therapy and effective early-warning systems for onset of seizure activity (Tzallas et al., 2012; Ramgopal et al., 2014). We analyse here electroencephalographic (EEG) recording of a spontaneous (non-induced) seizure and focus on the temporal evolution of energy distributions at individual EEG channels and interactions between channel pairs. We aim at characterizing the seizure and evaluate its spatial and temporal evolution. Moreover, we are able to identify a pre-seizure build up of subtle changes in the cross-coherence between channels, which has not been identified by other methods for change-point detection and cannot be detected by visual inspection. Our method offers new insights to the temporal and spatial location of abrupt changes in the energy spectrum during the seizure, with the feature of identifying the specific frequency bands that

drive these changes.

During seizure the spectral profile of brain activity displays sudden and abrupt changes. To capture this in a parsimonious but effective way, EEG recordings are frequently modelled as piecewise stationary process (see e.g. Saab and Gotman, 2005; Terrien et al., 2013; Kumar et al., 2014). As a direct implication, a multivariate change-point detection method is required to identify change points and thus produce a segmentation of the multi-channel EEG data into quasi-stationary blocks. The method introduced here allows not only this segmentation, but also direct interpretation in the sense of assigning a change point to one or multiple EEG channels and frequency bands.

In this chapter, we are not concerned with estimating parameters of a classical model such as the change in mean process of Chapter 3 or the change in mean and/or variance process of Chapter 4. However, much effort has been dedicated in the past to developing segmentation methods for multivariate data that rely on detection of change points in model parameters, where we mention the vector autoregressive (VAR) models in Kirch et al. (2015), Davis et al. (2006) and the works referenced in Chen and Gupta (2012). Parametric change-point detection methods are efficient if the underlying model is correct, but one can expect that possible model misspecification severely affects the performance. Another limitation of the parametric approach is the difficulty in the interpretation of detecting a change in a parameter. For example, a change in a high order lag parameter of a VAR model is not straight-forward to explain to a neuroscientist or clinician. Most neuroscientists have a better intuition for the conventional approach of monitoring the energy distribution over different oscillations compared to the parameters of a

complex multivariate model.

Another class of change-point methods focus on detecting changes in the spectral characteristics of EEG data. See, for example, Adak (1998), Ombao et al. (2005), Terrien et al. (2013) and references cited therein. One feature common to many of these methods is the use of a dyadic segmentation which restricts both the permitted data length and the change-point locations to being multiples of 2^j, j some positive integer. Another shared feature is that change detection lacks specificity in frequency which is a serious limitation because, for interpretability, methods for EEG seizure must be able to identify not only the change points but also the frequency bands to which these changes can be attributed. Such frequency-specific change detection can help to link the segmentation results to neuroscientific findings on event-related changes in brain activity at different frequency bands (Alarcon et al., 1995; Blondin and Greer, 2011; Schmitt et al., 2012).

Our contribution in this chapter is a method that identifies frequency-specific temporal changes in both the spectral energy distribution at each channel and in the cross-coherence between pairs of channels. The detection of frequency-specific changes for a subset of frequency bands has been discussed in the literature, e.g. by Kaplan et al. (2001) and Saab and Gotman (2005). However, detection of frequency-specific change points in coherence has received limited attention to date. Coherence is a frequency-specific measure of linear relation between a pair of channels. More formally, coherence at one frequency band is asymptotically equivalent to the square of the linear correlation between this frequency band's oscillations (obtained by filtering) at a pair of channels, as shown in Ombao and van Bellegem (2008). The evolution of coherence between EEG activity at different channels

or neuronal activity between regions on the cortical surface can provide interesting insights. For example, Sun et al. (2004) showed time-invariant differences in coherence for different motor tasks. In a learning experiment, where a subject determines associations between two sets of pictures, Gorrostieta et al. (2012) showed that coherence between the nucleus acumbens and the hippocampus evolves during the experiment. The nucleus acumbens, being a part of the reward system, plays a significant role in reward and reinforcement learning. The hippocampus is engaged in the memory formation and retention. Thus, the nature of the interaction between nucleus acumbens and the hippocampus also evolves while the subject learns the association and retains memory throughout the experiment. More generally, the interactions between brain regions is increasingly receiving attention in biostatistics and neuroscience. A good example are recent works on the detection of temporal change points in functional magnetic resonance imaging (fMRI, see e.g. Aston and Kirch, 2012; Cribben et al., 2012, 2013). However, fMRI data generally is collected at much lower temporal resolution and therefore inherently cannot be used to analyse the fast-progressing evolution of events like epileptic seizure.

To detect the number and location of possibly many change points of the piecewise-stationary EEG process we use a multivariate CUSUM-type procedure that consistently detects change points in autospectra and coherences of the epileptic seizure data. To give an intuition, the central test statistic identifies at any stage of the iterative algorithm the point that gives the maximum aggregated contrast between adjoining blocks in a given interval, over a set of time series. If the test statistic exceeds a certain threshold for which we provide practical guidance we save the thus detected change point and proceed in a binary-segmentation logic

by evaluating the two subintervals resulting from dividing the original interval at this change point. The proposed procedure is motivated by the sparsified binary segmentation algorithm (Cho and Fryzlewicz, 2015a). The authors apply it to a multivariate locally stationary wavelet process to simultaneously detect change points in Haar wavelet periodograms and cross-periodograms and derive some relevant theoretical results. We adapt the sparsified binary segmentation concept with the goal of detecting frequency-specific change points in both autospectra and cross-coherences of multivariate locally stationary time series. Crucially, our approach acknowledges that changes may not be synchronized over EEG channels, leading us to distinguish between channel-specific change points and those detectable between channel pairs. Moreover, our method provides insight in terms of the conventionally used frequency-band interpretation of EEG dynamics and thus improves understanding of the process underlying multi-channel EEG data.

The analysis of brain activity immediately prior to and during epileptic seizure requires a number of characteristics from a statistical method for change-point detection. In light of the complex nature of the observed time series, nonparametric methods have the obvious advantage of avoiding the risk of model misspecification. The method developed in Preuß et al. (2015) is a good example, as it furthermore allows for direct assignment of a change point to channels or brain regions, which is not necessarily the case for parametric methods that fit a sequence of (V)AR processes to data, such as proposed in Kirch et al. (2015) and Davis et al. (2006).

However, another important feature for EEG analysis is that the segmentation should be flexible in two respects, (a) change-point candidates should not be restricted to dyadic locations and (b) the method should be applicable to multi-

channel data, i.e. we should be able to simultaneously analyse EEG data recorded at different (possibly many) channels, and their interactions. The methods of Cho and Fryzlewicz (2015a), Kirch et al. (2015) and Preuß et al. (2015) are shown to handle data with up to 100, 12 and 5 components, respectively, while Davis et al. (2006) only provides results to the two-component case. The data from the epileptic seizure we are analysing here is a 21 channel recording and frequently experimental data is recorded at 64 or 256 channels simultaneously.

In terms of interpretability, neither Kirch et al. (2015) nor Davis et al. (2006) nor Cho and Fryzlewicz (2015a) are able to assign change points to components - at least not as immediate output - and thus neither method allows for direct interpretation of changes in the dependency between channels. Preuß et al. (2015) emphasize that their method can attribute changes to specific components of the spectral matrix, i.e. autospectra or cross-spectra. However, they illustrate this only for financial time series with length up to 2000 and as this method relies on global optimization it cannot handle long time series such as our data set of over eight minutes (500sec) sampled at 100Hz (100 points per second). Large data is generally a challenge for global segmentation methods, except for dynamic programming approaches. Approaches as in Killick et al. (2012) and Maidstone et al. (2016) are in this respect noteworthy, as the algorithms use a pruning step to detect multiple change points in univariate data by minimizing a cost function, typically in linear time. However, it is not clear how these methods scale to high dimensional data or how they could take the cross-dependence between components of a multivariate time series into account.

The method of Kirch et al. (2015), placed within a binary segmentation al-

gorithm as suggested by the authors, can detect multiple change points in long multivariate time series, but the required bootstrapping of critical values is computationally demanding. Another nonparametric method for change-point detection is the multivariate SLEX method of Ombao et al. (2005), who apply their method to a subset of channels from our data set. However, the SLEX method suffers from the restriction that both total time length and change-point locations must be dyadic. Finally, none of the here-discussed multivariate approaches to time series segmentation provides interpretation of change points in the Fourier frequency domain. This stands in contrast to the way physicians and neuroscientists read EEG data, namely in terms of energy distribution over different Fourier frequency bands.

To summarize, the method proposed in this work is the first to address the wide range of requirements for the analysis of EEG data that evolves e.g. before and during an epileptic seizure: it offers direct interpretability of changes both with respect to location on the scalp and in terms of frequency bands, which make the output relevant to neuroscientists. It is also computationally fast and scalable, by which we mean that it can be applied to long recordings over many channels. Change points are not restricted to have dyadic locations and are detected in both autospectra and cross-coherences. Our method shows a high power in simulation studies with long time series, where it proves to be sensitive to changes even if those occur only in a small subset of time series components.

The application of the proposed method to EEG seizure data illustrates the easy, direct interpretation of the identified changes and provides insights to spectral energy evolution in pre-seizure brain activity. We can identify interesting

pattern in the detected changes that point to the existence of pre-seizure biomarkers. Furthermore, we show how the method can support the analysis of epileptic seizure EEG by the neurologist.

5.2 Modeling Epileptic Seizure EEG as Piecewise-Stationary Process

Epileptic seizure signals can be considered realizations of an underlying brain process containing abrupt changes between different quasi-stationary regimes. The process can be characterised by the spectral energy distribution, i.e. how energy is distributed over different frequencies: as the seizure process unfolds, the energy concentration at channels undergoes sharp changes at specific frequencies (or frequency bands). Moreover, the strength of the interactions between oscillations at different brain regions also changes. We propose the following formalization of this process. Denote the zero-mean $D-$channel EEG signal to be $\mathbf{X}(t^*) = (\mathbf{x}_1(t^*), \ldots, \mathbf{x}_D(t^*))'$, $t^* = \{1, \ldots, T^*\}$. The dimension D is known and, due to physical constraints, must be finite. For the entire observation period, denote the number of stationary segments to be $N + 1$ and define the set of change-point locations to be $\mathbb{N} = \{\eta_i, i = 1, 2, \ldots, N\}$. By convention, $\eta_0 = 0$ and $\eta_{N+1} = T^*$. The i-th stationary segment is denoted $\mathcal{I}_i = \{\eta_{i-1} + 1, \ldots, \eta_i\}$. The EEG signal has the piecewise-stationary representation

$$\mathbf{X}(t^*) = \sum_{i=1}^{N+1} \mathbb{I}(t^* \in \mathcal{I}_i) \int_{-0.5}^{0.5} \exp(i2\pi\omega t^*) d\mathbf{Y}^i(\omega) \tag{5.1}$$

where the indicator function $\mathbb{I}(t^* \in \mathcal{I}_i) = 1$ if $t^* \in \mathcal{I}_i$ and 0 otherwise and $d\mathbf{Y}^i(\omega) = (dY_1^i(\omega), \dots dY_D^i(\omega))'$ is a zero-mean uncorrelated random increment process on the i-th stationary segment with covariance matrix

$$\text{Cov}(d\mathbf{Y}^i(\omega)) = \mathbf{F}^i(\omega)d\omega = \begin{bmatrix} f_{1,1}^i(\omega) & \dots & f_{1,D}^i(\omega) \\ \dots & \dots & \dots \\ f_{D,1}^i(\omega) & \dots & f_{D,D}^i(\omega) \end{bmatrix} d\omega. \tag{5.2}$$

This process, which contains finite jumps (i.e., jumps of finite size bounded away from zero) at a finite number of change points, belongs to the family of locally stationary processes (Dahlhaus, 2012). Moreover, under Gaussianity, it is fully characterized by the localised spectral matrix

$$\mathbf{F}(t^*, \omega) = \sum_{i=1}^{N+1} \mathbb{I}(t^* \in \mathcal{I}_i)\mathbf{F}^i(\omega), \ \omega \in [-0.5, 0.5].$$

The diagonal elements of $\mathbf{F}(t^*, \omega)$ represent the univariate localised autospectral densities $f_{d,d}(t^*, \omega)$ for $d = \{1, \dots, D\}$ which take the value $f_{d,d}^i(\omega)$ when $t^* \in \mathcal{I}_i$. The off-diagonal entries of $f_{d,d'}(t^*, \omega)$, where $d \neq d'$, are localised cross-spectral densities which take the value $f_{d,d'}^i(\omega)$ when t^* is in the i-th stationary segment.

To facilitate scale-free comparison of the dependence between different pairs of EEG channels across time, we shall study the dependence between channels d and d' via the time-varying coherence defined to be

$$\rho_{d,d'}(t^*, \omega) = |f_{d,d'}(t^*, \omega)|^2 / (|f_{d,d}(t^*, \omega)||f_{d',d'}(t^*, \omega)|),$$

with $d, d' \in \{1, \dots, D\}, d \neq d', t^* \in \{1, \dots, T^*\}$. This metric can be interpreted

as localised frequency-specific (squared) correlation between a pair of time series components. It measures the strength of linear association between two stationary processes at a particular frequency (or frequency band) and admits values in the range $[0, 1]$ indicating weak to strong linear dependence.

The process defined in Equations (5.1)-(5.2) does not require that all channels and all pairs of channels exhibit a change at each of the time points in the set $\mathbb{N}$. Instead, the set $\mathbb{N}$ is a collection of time points at which the autospectrum of at least one channel or the coherence of at least one pair of channels exhibits a change at some frequency. We denote the component-specific subsets $\mathbb{N}^{d,d'} = \{\eta_i^{d,d'}, i = 1, 2, \ldots, N^{d,d'}\} \subseteq \mathbb{N}$ to be a collection of change points in the autospectrum (when $d = d'$) or in the cross-coherence (when $d \neq d'$) where $N^{d,d'} \leq N$.

To identify change points in the piecewise-stationary process $\mathbf{X}(t^*)$ following Equations (5.1)-(5.2), we require some assumptions about the underlying process. Firstly, the process must be bounded, i.e. for all $d, d' : d \neq d'$, $f_{d,d}(t^*, \omega) \leq f^* < \infty$ and $\rho_{d,d'}(t^*, \omega) < 1 \ \forall t^*, \omega$. Secondly, the changes have to be nonzero, i.e. for all $d, d' : d \neq d'$ and all i, $\exists \omega^*$ such that $|f_{d,d}(\eta_i^{d,d}, \omega^*) - f_{d,d}(\eta_{i-1}^{d,d}, \omega^*)| \geq f_* > 0$ and similarly, $\exists \omega^{**}$ such that $|\rho_{d,d'}(\eta_i^{d,d'}, \omega^{**}) - \rho_{d,d'}(\eta_{i-1}^{d,d'}, \omega^{**})| \geq \rho_* > 0$. Thirdly, changes have to be sufficiently distant, i.e. $|\eta_i^{d,d'} - \eta_{i-1}^{d,d'}| \geq \delta_T > 0$ for all d, d' and all i. We allow the minimum change-point distance δ_T to grow with T^*, but at a lower rate as explained in the following section.

The process defined in Equations (5.1)-(5.2) is nonparametric and thus general while still sufficiently powerful in capturing structural changes. Such changes may relate to events such as the presentation of an external stimuli or the build-up to an epileptic seizure. Any change in brain activity will be reflected by a discontinuity

in the spectral matrix at the points $\eta_i, i \in \{1, \ldots, N\}$, and therefore shows in at least a change in the spectral energy of a component $\mathbf{x}_d(t^*)$ or a change in the coherence between two components $\mathbf{x}_d(t^*)$ and $\mathbf{x}_{d'}(t^*)$.

The representation of the time-varying spectral density depends on the frequency ω. That is, the autospectrum or coherence may remain constant for some frequencies but may change over time for others. Hence change points can be attributed to specific frequencies, for both spectral energy and coherence. This provides additional specific and interpretable results giving a competitive edge over the majority of existing change-point detection methods which compare the full spectral distribution over time, such as Preuß et al. (2015).

In the following section, we develop the FreSpeD method and provide insights into how it allows for interpretation of detected change points in EEG data. While the change-point detection theory behind FreSpeD is related to the theoretical findings in Cho and Fryzlewicz (2015b,a), for a number of reasons it is more complicated and beyond the scope of this application-focused chapter. However, we discuss the results in a simplified framework in Section 5.3.3.

5.3   Frequency-Specific Change-Point Detection

The overall aim in this chapter is to study the evolution of brain activity in a multi-channel epileptic seizure EEG recording $\mathbf{X}(t^*)$ with D channels. We characterize the underlying dynamics via change points in time and frequency and for all D autospectra and $D(D-1)/2$ different cross-coherences in $\mathbf{X}(t^*)$. The first stage of our method is to estimate these unknown spectral quantities.

5.3.1   Estimation of the spectral quantities

To obtain information about the spectral properties of $\mathbf{X}(t^*)$, it is well-known that some trade-off with the series' resolution in time is required. Our approach to this is to partition the entire time index $\{1, \ldots, T^*\}$ into short localised intervals of length ν, which are defined by the boundaries $\mathbb{T} = \{t : t = \nu, 2\nu, \ldots, T\nu\}$ where $T = T^*/\nu$ (see Figure 5.1). As we describe below, we aggregate information on the short intervals $\{t - \nu + 1, \ldots, t\}$, $t \in \mathbb{T}$, using standard Fourier estimation principles and index this by the elements of $\mathbb{T} = \{t : t = \nu, 2\nu, \ldots, T\nu\}$. The set $\mathbb{T}$ contains all points that we consider as change-point candidates and therefore ideally the set of true change points $\mathbb{N}$ is a subset of $\mathbb{T}$. We assume that $T = \mathrm{O}((T^*)^\Theta)$ with $0.5 < \Theta < 1$: as $T^* \to \infty$, $T = T^*/\nu \to \infty$, but at a smaller rate. Moreover, the minimum change-point distance δ_T grows proportional to T, $\delta_T \propto T$. This implies that as $T^* \to \infty$, the partitioning $\mathbb{T}$ becomes finer but at a lower rate, meaning that the window length ν also grows, but slowly. Under this assumption, intuitively the chances of $\mathbb{N} \not\subset \mathbb{T}$ decrease.

The true spectral quantities we are concerned with here are the averages over an interval up to point t, $\{t - \nu + 1, \ldots, t\}$, denoted for, respectively, autospectra and cross-coherences as

$$
\begin{aligned}
\bar{f}_{\nu,d,d'}(t,\omega) &= \frac{1}{\nu} \sum_{t^*=t-\nu+1}^{t} f_{d,d'}(t^*,\omega), \\
\bar{\rho}_{\nu,d,d'}(t,\omega) &= |\bar{f}_{\nu,d,d'}(t,\omega)|^2 / (|\bar{f}_{\nu,d,d}(t,\omega)||\bar{f}_{\nu,d',d'}(t,\omega)|), \qquad \forall t \in \mathbb{T}.
\end{aligned}
$$

If an interval up to point t, $\{t - \nu + 1, \ldots, t\}$ does not contain a change point, the average quantities are identical to the true autospectra and coherences at point

t. To derive mean-squared consistent estimators of the autospectra and coherence on an interval of length ν under stationarity, we compute the discrete Fourier transform on each of M non-overlapping blocks of equal length that together cover the whole interval of length ν. Following Welch (1967), we then average over the M periodograms to obtain a consistent estimator. Formally, the estimator of the local spectral matrix on the interval $\{t - \nu + 1, \ldots, t\}$ is

$$\widehat{\mathbf{F}}_\nu(t, \omega_l) = \frac{1}{M} \sum_{m=1}^{M} \widehat{\mathbf{F}}_\nu(m, t, \omega_l) = \frac{1}{M} \sum_{m=1}^{M} \boldsymbol{d}_\nu^m(t, \omega_l) \boldsymbol{d}_\nu^{m,\star}(t, \omega_l), \quad t \in \mathbb{T} \qquad (5.3)$$

where $\boldsymbol{d}_\nu^m(t, \omega_l)$ is the discrete Fourier transform of $\mathbf{X}(t)$ of the m-th segment of the interval indicated by t, $\{t - (M - m + 1)\nu/M + 1, \ldots, t - (M - m)\nu/M\}$, and $\boldsymbol{d}_\nu^{m,\star}$ denotes the complex conjugate transpose of $\boldsymbol{d}_\nu^m$. In the following, we drop the subscript ν for ease of notation.

In reality, we do not search for change points at ('singleton') frequency ω. This is motivated by the standard practice in EEG analysis were conclusions are made based on averages over traditional frequency *bands* delta, theta, alpha, beta and gamma. In this work, the number of frequency bands L in a local time block, controlled by the window length ν and the number of blocks M, $L = \lceil \nu/(2M) \rceil$, is assumed to be a constant. We note in the theoretical framework developed in Cho and Fryzlewicz (2015b,a) for the combined detection of change points in multiple time series (corresponding to L here) an extension letting L grow slowly is possible, i.e. $LT^{-\log T} \to 0$ as $L, T \to \infty$. However, because the interpretation of results based on growing L is of limited practical use, we do not pursue this here.

Remark.   There are practical trade-offs for using small interval lengths ν and a large number of blocks M in the Welch estimator. On the one hand, the frequency resolution will be poor. On the other hand, the advantage is that the estimator will have a lower variance (see e.g. Fiecas et al., 2010). In our application here with data sampled at the relatively low rate of 100Hz, we choose a balanced approach setting $\nu = 200$ (2sec) and $M = 10$, which provides us with 10 frequency bands covering $[0, 50]$Hz. We analyse the sensitivity of our method to the choice of ν in Section 5.4 for simulated data. As it turns out, a wide range of ν provides reasonable change-point detection results. This is consistent with previous results on the dependence of the quality of spectral estimates on local window length (Qin and Wang, 2008; Fiecas and Ombao, 2016). In this work, the fact that the approximation of the spectral estimates defined below are close to the true quantities for reasonable interval lengths ν under stationarity suggests to choose Θ to be close to one: as T grows at $\mathrm{O}((T^*)^\Theta)$ and $\nu = T^*/T$, $\nu = \mathrm{O}((T^*)^{1-\Theta})$. So as long as $\Theta < 1$, ν is growing with T^* and the relation between true and estimated spectral quantities as defined below holds under stationarity.

The estimators for time-varying autospectra, cross-spectra and cross-coherences as extracted from the spectral matrix $\widehat{\mathbf{F}}(\mathbf{t}, \omega_{\mathbf{l}})$ are denoted by, respectively, $\widehat{f}_{d,d}(t, \omega_l)$; the off-diagonal entries $\widehat{f}_{d,d'}(t, \omega_l)$; and the coherence estimator $\widehat{\rho}_{d,d'}(t, \omega_l) = |\widehat{f}_{d,d'}(t, \omega_l)|^2/|\widehat{f}_{d,d}(t, \omega_l)||\widehat{f}_{d',d'}(t, \omega_l)|$, $\forall t \in \mathbb{T}$. To be able to apply the same testing concept to the autospectra and cross-coherences while achieving consistency, we consider the Fisher-z transforms $\bar{\rho}^*_{d,d'}(t, \omega_l) = 1/2 \log \frac{1+\bar{\rho}_{d,d'}(t,\omega_l)}{1-\bar{\rho}_{d,d'}(t,\omega_l)}$ and $\widehat{\rho}^*_{d,d'}(t, \omega_l) = 1/2 \log \frac{1+\widehat{\rho}_{d,d'}(t,\omega_l)}{1-\widehat{\rho}_{d,d'}(t,\omega_l)}$. If the set of change points is indeed a subset of the set of change-point candidates, $\mathbb{N} \subset \mathbb{T}$, the estimators and the estimands (true but
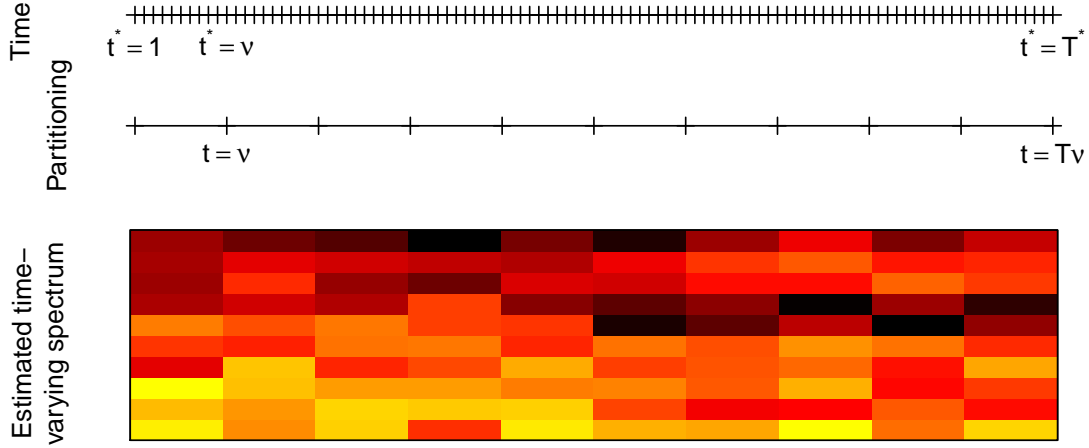
Figure 5.1: Partitioning of $\{1, \ldots, T^*\}$ into nonoverlapping intervals of length ν and example of an estimated spectrum (x-axis: time, y-axis: frequency, colour indicates level of energy: higher energy corresponds to darker colouring)

unknown spectral quantities) are related by the following approximations:

$$\widehat{f}_{d,d}(t, \omega_l) \approx \frac{1}{2M} \bar{f}_{d,d}(t, \omega_l) \chi^2_{2M}(t) \tag{5.4}$$

$$\widehat{\rho}^*_{d,d'}(t, \omega_l) \approx \bar{\rho}^*_{d,d'}(t, \omega_l) + \sqrt{M}\varepsilon(t), \qquad t \in \mathbb{T}, \tag{5.5}$$

where $\{\chi^2_{2M}(t),\ t \in \mathbb{T}\}$ are i.i.d. chi-square distributed random variables with $2M$ degrees of freedom and $\{\varepsilon(t),\ t \in \mathbb{T}\}$ are standard normal random variables. The random variables are i.i.d. over frequency bands ω_l. For ease of notation we suppress the subscripts (d, d) and (d, d'), respectively. The results for the autospectra follow directly from applying Welch's method to compute the periodogram, while the results for the transformed cross-coherence follow from the asymptotic distribution properties of the Fisher-z transform (Brockwell and Davis, 1991).

5.3.2 Estimation of number and locations of change points

We present now the heart of the FreSpeD method, the test statistic and its algorithmic framework that allow estimation of change points in time. For a specific channel d or pair of channels (d, d'), the method simultaneously detects change points and provides information about the frequency bands at which a change point is detectable. The test statistic and algorithm are specified in a way to account for multiple testing over frequencies. The FreSpeD method produces output that is directly interpretable and thus is an attractive tool for neuroscientists and physicians.

We formulate the algorithm for a generic panel of time series $\widehat{\mathbf{Z}}(t)$ with members $\widehat{\mathbf{z}}_l = \{\widehat{z}_l(t), t \in \mathbb{T}\}$, $l \in \{1, \dots, L\}$. The algorithm is applied to estimate change points in the autospectrum of channel d by setting $\widehat{z}_l(t) := \widehat{f}_{d,d}(t, \omega_l) \; \forall l$, and for changes in the coherence between channels d and d' to $\widehat{z}_l(t) := \widehat{\rho}^*_{d,d}(t, \omega_l) \; \forall l$. Consider the thresholded sum of CUSUM test statistic, defined on the interval $\{(s, e) \in \mathbb{T} \times \mathbb{T} : \nu \leq s < e \leq T\}$ of length $n = e - s + 1$ as

$$\mathfrak{C}_{s,b,e}(\widehat{\mathbf{Z}}) = \sum_{l=1}^{L} \mathcal{C}^*_{s,b,e}(\widehat{\mathbf{z}}_l)\mathbb{I}(\mathcal{C}^*_{s,b,e}(\widehat{\mathbf{z}}_l) > \zeta_T) \tag{5.6}$$

$$\mathcal{C}^*_{s,b,e}(\widehat{\mathbf{z}}_l) = \left| \sqrt{\frac{e-b}{n(b-s+1)}} \sum_{t=s}^{b} \widehat{z}_l(t) - \sqrt{\frac{b-s+1}{n(e-b)}} \sum_{t=b+1}^{e} \widehat{z}_l(t) \right| \Bigg/ \widehat{\sigma}_{s,e}(\widehat{\mathbf{z}}_l) \tag{5.7}$$

Here, ζ_T is a global threshold and $\mathbb{I}(\mathcal{C}^*_{s,b,e}(\widehat{\mathbf{z}}_l) > \zeta_T) = 1$ if $\mathcal{C}^*_{s,b,e}(\widehat{\mathbf{z}}_l) > \zeta_T$, zero otherwise. $\widehat{\sigma}_{s,e}(\widehat{\mathbf{z}}_l)$ denotes a scaling factor which accounts for the variation of $\widehat{\mathbf{z}}_l$

on $\{s, \ldots, e\}$. For the autospectra and cross-coherences, respectively,

$$
\begin{aligned}
\widehat{\sigma}_{s,e}(\widehat{\mathbf{f}}(\omega_l)) &= \frac{1}{n\sqrt{M}} \sum_{t=s}^{e} \widehat{f}(t, \omega_l) \\
\widehat{\sigma}_{s,e}(\widehat{\boldsymbol{\rho}}^*(\omega_l)) &= \sqrt{(n-1)^{-1} \sum_{t=s}^{e} \left(\widehat{\rho}^*(t, \omega_l) - n^{-1} \sum_{u=s}^{e} \widehat{\rho}^*(u, \omega_l) \right)^2}.
\end{aligned}
$$

For a single time series $\widehat{\mathbf{z}}_l$, $\mathcal{C}^*_{s,b,e}(\widehat{\mathbf{z}}_l)$ is maximized where the contrast between two adjacent blocks $\{\widehat{z}_l(s), \ldots, \widehat{z}_l(b)\}$ and $\{\widehat{z}_l(b+1), \ldots, \widehat{z}_l(e)\}$ is maximized. If this contrast, scaled by the process variability, exceeds a threshold ζ_T, then we consider b a change-point candidate. For any panel of processes $\widehat{\mathbf{Z}} = \{\widehat{\mathbf{z}}_l, l = 1, \ldots, L\}$ that can be described as (approximately) piecewise-constant functions plus noise the statistic $\mathfrak{C}_{s,b,e}(\widehat{\mathbf{Z}})$ is maximized at one of their change points, and even for large L this statistic can consistently detect change points that occur only in a small subset of the group while being robust to false positives (Cho and Fryzlewicz, 2015a). Competitor approaches have considered averaging or taking the local pointwise maximum CUSUM statistic (Groen et al., 2013),

$$
\mathfrak{C}^{avg}_{s,b,e}(\widehat{\mathbf{Z}}) = \frac{1}{L} \sum_{l=1}^{L} \mathcal{C}^*_{s,b,e}(\widehat{\mathbf{z}}_l) \qquad \mathfrak{C}^{max}_{s,b,e}(\widehat{\mathbf{Z}}) = \max_{l=1,\ldots,L} \mathcal{C}^*_{s,b,e}(\widehat{\mathbf{z}}_l)
$$

In a sense, the thresholded-sum CUSUM lies in-between the maximum CUSUM and the average CUSUM (scaled by L). The thresholded-sum CUSUM is preferable in settings with large L: here, averaging over CUSUMs can yield to false negatives, i.e. undetected sparse change points. At the same time, the maximum CUSUM can lead to false positives, i.e. spurious detection of change points by chance. An interesting property of the thresholded-sum CUSUM of Equation (5.6) is the

assignment of frequency bands to change points. If a point $\hat{\eta}_i$ is detected as change point, we can consider the exceedance of the threshold for specific frequency bands $l^* \in \{1, \ldots, L\}$ as evidence that the change occurs at these frequency bands and is not visible in other frequency bands.

Within the FreSpeD method, all D autospectra and $D(D-1)/2$ different coherence pairs of the time series $\mathbf{X}(t^*)$ can be analysed separately. This is a fast approach to analysing EEG data with potentially many channels, because the method can be parallelized. For each autospectrum and cross-coherence, we use the thresholded CUSUM statistic within a binary Segmentation algorithm. In this framework, if we identify a change point on the interval $\{1, \ldots, T\}$ we split the interval at this point, say b_0, and repeat the procedure on the resulting subintervals $\{1, \ldots, b_0\}$ and $\{b_0 + 1, \ldots, T\}$.

The algorithm in a pseudo code function is displayed in Figure 5.2. This function is called for $\widehat{z}_l(t) = \widehat{f}_{d,d}(t, \omega_l)$ and $\widehat{z}_l(t) = \widehat{\rho}^*_{d,d'}(t, \omega_l)$, $\forall d, \ d' \neq d$.

The algorithm includes an extra step of evaluating the thresholded CUSUM statistic on the interval $\mathcal{B}_0$ of length $\Delta_T + 1$ around a change-point candidate b_0 to avoid spurious detections.

5.3.3 Heuristic Justification in a Simplified Framework

The focus of this chapter is on the application of a new and flexible method to epilepsy data to generate new insights for the neuroscience community, mainly in terms of interpretability of change points both spatially over the brain and with respect to their appearance in frequency bands. While we illustrate the performance

function FRESPED($\widehat{\mathbf{Z}}$, s, e, ζ_T)
    if $e - s \geq 2\delta_T + 1$
        $\mathcal{B} := \{s + \delta_T, s + \delta_T + 1, \ldots, e - \delta_T\}$
        $b_0 := \arg\max_{b \in \mathcal{B}} \mathfrak{C}_{s,b,e}(\widehat{\mathbf{Z}})$, $\mathcal{B}_0 := \{b_0 - \Delta_T, \ldots, b_0 + \Delta_T\}$
        if $\mathfrak{C}_{s,b_0,e}(\widehat{\mathbf{Z}}) > \zeta_T$ and $\min_{t \in \mathcal{B}_0} \mathfrak{C}_{s,t,e}(\widehat{\mathbf{Z}}) > 0$
            add b_0 to the set of estimated change points $\mathbb{N}$
            FRESPED($\widehat{\mathbf{Z}}$, s, b_0, ζ_T)
            FRESPED($\widehat{\mathbf{Z}}$, $b_0 + 1$, e, ζ_T)
        else
            if $|\mathcal{B}| > 1$
                set $\mathcal{B} := \mathcal{B} \setminus b_0$, go to line 4: $b_0 := \arg\max_{b \in \mathcal{B}} \mathfrak{C}_{s,b,e}(\widehat{\mathbf{Z}})...$
            else
                STOP
            end if
        end if
    end if
end function

Figure 5.2: FreSpeD algorithm; initiated by setting $\widehat{z}_l(t) = \widehat{f}_{d,d}(t, \omega_l)$ and $\widehat{z}_l(t) = \widehat{\rho}^*_{d,d'}(t, \omega_l), t \in \mathbb{T}$ and $s = \nu$, $e = T$; $\forall d$, $d' \neq d$

of our method compared to state-of-the-art competitor methods in Section 5.4 and discuss findings from the application to an epileptic seizure recording in Section 5.5, the aim of the following discussion is to provide an intuition to the theoretical properties of our method in a simplified framework. The proof belonging to this discussion is provided in the appendix.

As pointed out before, the complex nature of the observed EEG time series requires a change-point detection procedure in a flexible process framework. The non-parametric method proposed here has the advantage of avoiding the risk of model misspecification as it does not assume much beyond piecewise stationarity (see page 159). However, the estimation of spectral quantities is not straight-forward in this general framework and only allows for stating approximations instead of

equalities in Equations (5.4) and (5.5). Moreover, independence (over time blocks and frequency bands) is not granted in general. Finally, in the general framework the set of change points is not guaranteed to be a subset of the set of change-point candidates, i.e. it is possible that $\mathbb{N}^{d,d'} \not\subset \mathbb{T}$ for some $(d, d') \in \{1, \ldots, D\}^2$, which can lead to a bias in the change-point statistic.

We now consider a simplified framework where the approximations of Equations (5.4) and (5.5) are equalities, with independent random variables and where $\mathbb{N} \subset \mathbb{T}$. Moreover, we require a balancedness between change points, i.e. $\max\left(\frac{\eta_i^{d,d'} - \eta_{i-1}^{d,d'} + 1}{\eta_{i+1}^{d,d'} - \eta_{i-1}^{d,d'} + 1}, \frac{\eta_{i+1}^{d,d'} - \eta_i^{d,d'}}{\eta_{i+1}^{d,d'} - \eta_{i-1}^{d,d'} + 1}\right) \leq c_* \in [1/2, 1)$. This is a technical assumption on the change-point locations that is necessary to ensure consistency of our CUSUM statistic which can exhibit erratic behaviour close at the edges of the interval on which it is computed. Recalling the assumptions of boundedness, minimum separability, identifiability stated on page 159 and the rate requirement $T = \mathrm{O}\left((T^*)^\Theta\right)$, $\Theta \in (0.5, 1)$ we can then show that there exists $C > 0$ and $\kappa > 0$ s.t. for $\epsilon_T = \log^{2+\vartheta} T$, $\zeta_T = \kappa \log^{1+\varpi} T$ and $\Delta_T \propto \epsilon_T$ with $\vartheta > 0, \varpi > \vartheta/2$, as $T^* \to \infty$

1. the set of change points $\widehat{\mathbb{N}}^{d,d}$ detected by FreSpeD in the autospectrum of channel d satisfies

$$\Pr\left(\widehat{N}^{d,d} = N^{d,d}; |\widehat{\eta}_i^{d,d} - \eta_i^{d,d}| < C\epsilon_T \text{ for } i = 1, \ldots, N^{d,d}\right) \to 1 \text{ as } T^* \to \infty$$

2. the set of change points $\widehat{\mathbb{N}}^{d,d'}$ detected by FreSpeD in the cross-coherence between channels d and d' satisfies

$$\Pr\left(\widehat{N}^{d,d'} = N^{d,d'}; \ |\widehat{\eta}_i^{d,d'} - \eta_i^{d,d'}| < C\epsilon_T \text{ for } i = 1, \ldots, N^{d,d'}\right) \to 1 \text{ as } T^* \to \infty.$$

The theoretical derivation of the threshold ζ_T and the change-point accuracy bound ϵ_T in this simplified framework are shown in the appendix on page 209 ff. We note here that for ease of notation the quantities are expressed in terms of T rather than T^*; the dependency on ν follows directly from the equality $T = T^*/\nu$.

The concept and theory of thresholded-sum CUSUM statistics for change-point detection were to the best of our knowledge first developed in Cho and Fryzlewicz (2015b,a) and used in combination with a wavelet-type decomposition of multivariate data. The FreSpeD method was developed specifically for the application to EEG data and differs substantially in two ways. First, FreSpeD uses estimates of local Fourier autospectra and coherences. While some authors apply wavelet decomposition to EEG signals (Schiff et al., 1994; Adeli et al., 2003; Wang et al., 2011), the convention is interpretation of spectral energy over Fourier frequency bands and we tailor our method following this convention. Second, FreSpeD does not analyse the panel of EEG recordings over channels as a single time series, but rather considers recordings at individual channels and channel pairs separately. This allows the spatial localisation of change points in the sense that we can observe a first change, say, in the seizure focal point, followed by changes in neighbouring areas. Opposed to this, the algorithm of Cho and Fryzlewicz (2015a) considers the aggregate of Haar wavelet periodograms at all channels and channel pairs simultaneously.

5.3.4   Implementation in `R`

The `R` package `FreSpeD` developed by the first author of this chapter contains an efficient implementation of the FreSpeD method. *A first, blinded ver-*

sion of this package is made available as supplementary material on Github (`https://github.com/almms/FreSpeD.git`). Prior to publication on CRAN the package can be installed on Windows from the *.zip-file and requires the installation of dependencies `doParallel` *and* `foreach`. In addition to the standard code and a set of default parameter values, prior information can be included by adjusting a range of parameters. For example, depending on the available time resolution (sampling rate) and frequency-band resolution of interest, parameter adjustments may be desirable to answer a specific research question. For instance, if the user's interest is focused on activity at the gamma band (32-50Hz), more coarse resolution can be allowed for by decreasing the window length ν, at the benefit of a finer resolution in time. We also point out that the package offers parallelization of the computations over a user-specified number of cores. More details can be found in the package description.

5.4 Simulation Study

Simulation set up. To investigate the performance of the proposed methodology for finite samples, we conduct a simulation study using parametric models under the following settings:

- The total time series length is $T^* = 50000$, in line with the EEG data analysed in Section 5.5.

- The number of components is $D = 2$ or 20.

- The number of components that exhibit a change is denoted D^c. When

$D = 2$, we investigate the situation where $D^c = 1$ or 2; when $D = 20$, $D^c = 1, 10$ or 20.

- The number of change points is $N = 1$ or 5, which are equidistant between themselves and to start and end of the time series.

We consider four parametric processes, VAR(2), VAR(6), VAR(10) and VARMA(2,2). Here, we only vary one or two parameters of the process components that undergo a change, letting them switch between parameter sets A and B, while the remaining $D^{nc} = D - D^c$ components are fully described by parameter set A. The cross-correlations of these processes are captured by the covariance matrix of the innovations. To illustrate, a general form of the D-variate VAR(p) process used here is

$$\mathbf{X}(t^*) = \begin{bmatrix} \mathbf{x}_1(t^*) \\ \mathbf{x}_2(t^*) \\ \dots \\ \mathbf{x}_D(t^*) \end{bmatrix} = \sum_{k=1}^{p} \left(\begin{bmatrix} \phi_{k,1} & 0 & \cdots & 0 \\ 0 & \phi_{k,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \phi_{k,D} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1(t^* - k) \\ \mathbf{x}_2(t^* - k) \\ \dots \\ \mathbf{x}_D(t^* - k) \end{bmatrix} \right) + \begin{bmatrix} \epsilon_1(t^*) \\ \epsilon_2(t^*) \\ \dots \\ \epsilon_D(t^*) \end{bmatrix},$$

where $(\epsilon_1(t^*), \dots, \epsilon_D(t^*))'$ are i.i.d. Gaussian white noise with covariance Σ. Thus, we summarize the process parameter set by the vectors $\Phi_k = (\phi_{k,1}, \dots, \phi_{k,D})'$. The first D^c elements of the parameter vector take the values $\phi_{k,j} = \phi_k^A \; \forall k \in \{1, \dots, p\}$ before the first change point and $\phi_{k,j} = \phi_k^B$ afterwards, and alternate between these two values at each subsequent change point. The remaining D^{nc} vector elements of Φ_k take the value ϕ_k^A throughout time. For the VARMA(2,2), the corresponding MA parameter vectors are denoted in this section $\Theta_k = (\theta_{k,1}, \dots, \theta_{k,D})', \; k = \{1, 2\}$ and their elements take the values θ_k^A or θ_k^B in the same manner as described for

the VAR parameters above.

1. VAR(2): $(\phi_1^A, \phi_2^A)' = (-0.15, 0.53)'$, $(\phi_1^B, \phi_2^B)' = (-\phi_1^A, \phi_2^A)'$

2. VAR(6): $(\phi_1^A, \ldots, \phi_6^A)' = (0.10, -0.10, 0.00, 0.10, 0.20, -0.35)'$ and $(\phi_1^B, \ldots, \phi_6^B)' = (\phi_1^A, \ldots, \phi_4^A, -\phi_5^A, -\phi_6^A)'$

3. VAR(10): $(\phi_1^A, \ldots, \phi_{10}^A)' = (0.10, -0.10, 0.00, 0.10, 0.20, -0.35, 0.00, 0.00, 0.20, -0.35)'$ and $(\phi_1^B, \ldots, \phi_{10}^B)' = (\phi_1^A, \ldots, \phi_8^A, -\phi_9^A, -\phi_{10}^A)'$

4. VARMA(2,2): $(\phi_1^A, \phi_2^A)' = (\phi_1^B, \phi_2^B)' = (-0.15, 0.53)'$, $(\theta_1^A, \theta_2^A)' = (0.10, -0.10)'$ and $(\theta_1^B, \theta_2^B)' = (2\theta_1^A, 3\theta_2^A)'$

One realization of a changing component of each process with $N = 5$ change points is displayed in Figure 5.3, where we show a window of size 400 around the first change point to illustrate that changes in the processes are difficult to identify by eye-balling. The autospectrum of a changing component of the VAR(6) process is given in Figure 5.4. The figures suggest that while it is difficult to visually spot the change points in the time series, these changes are more pronounced in the spectra, hereby illustrating the potential of using the spectral data features in a change-point detection algorithm for EEG data.

Remarks on competitor methods. The FreSpeD method is compared with approach of Cho and Fryzlewicz (2015a,b) using wavelet decompositions and the VAR-based method of Kirch, Muhsal, and Ombao (2015) (abbreviated KMO). Cho and Fryzlewicz (2015a,b) apply the principle of change-point detection via thresholded-sum CUSUM statistics to their multivariate Locally Stationary Wavelet model
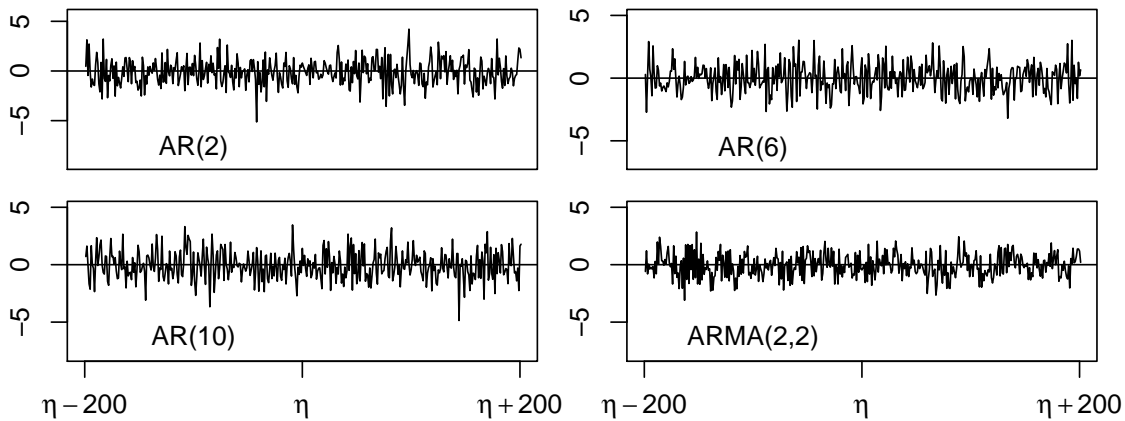
Figure 5.3: Realizations of single components of the processes used in the simulation study, with a total of $N = 5$ equidistant change points; processes are shown on a window of length 400 around the first change point

(LSW). In a similar way as FreSpeD analyses estimates of spectral energy and coherence at individual frequency bands, LSW considers Haar wavelet periodograms and a transformation of the wavelet cross-periodograms at different scales. The LSW method is introduced together with practical guidance regarding the choice of parameters and the estimation of scale-specific thresholds. We use an implementation of the method and a bootstrap procedure for threshold estimation kindly provided by the authors, and the guidance provided in Cho and Fryzlewicz (2015a) for the choice of the parameter $\gamma = 0.499$, which is at the conservative end of its scale.

The KMO method is implemented using orders $p = 2$ and $p = 6$. The choice of these orders is somewhat arbitrary since KMO does not provide objective measures for selecting the optimal order for change-point estimation. Kirch et al. (2015) claim that the choice of the order should not adversely impact the result but the simulation studies suggest otherwise, with the most severe negative result occurring
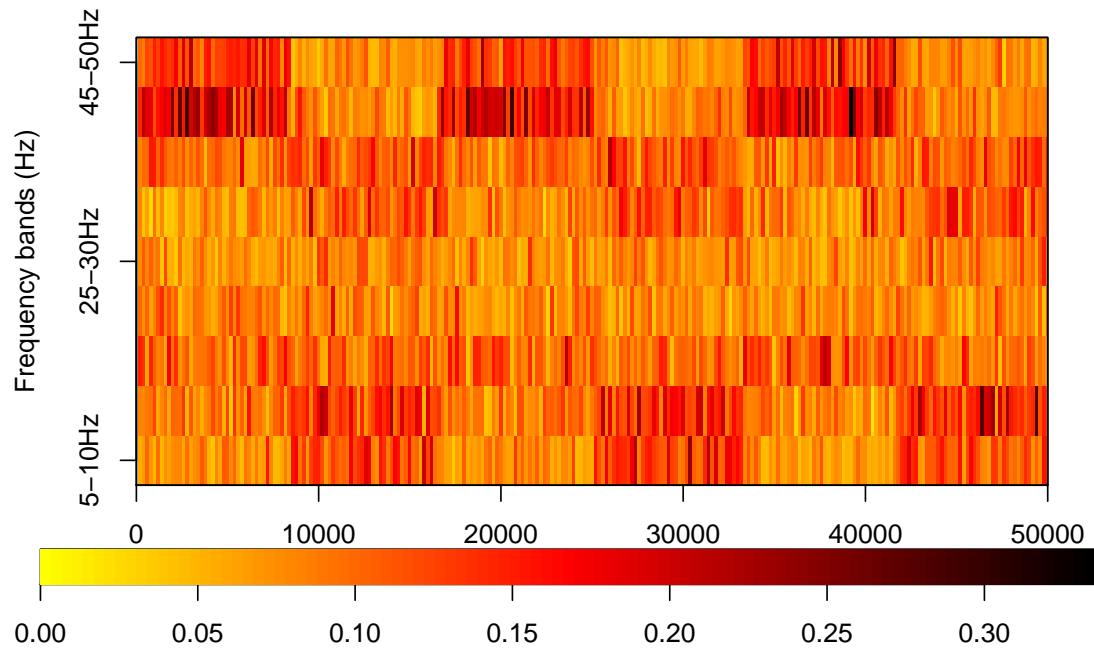
Figure 5.4: Estimated time-varying autospectrum of one component of the VAR(2) process depicted in Figure 5.3, assuming sampling rate of 100Hz

when the method uses an order that is much smaller than that of the true underlying process. When KMO uses an order (here $p = 2$) that is lower than that of the true process (e.g., $p = 6$ or 10), it appears that one loses information in the dimension reduction step that could be crucial to change-point detection. Because KMO with order $p = 2$ underperforms compared to the higher order parametrization $p = 6$ and compared to FreSpeD, we provide the results for the $p = 2$ parametrization only in the appendix (Table C.5 on page 227). In the following, unless explicitly stated we shall discuss only the results of KMO with order $p = 6$. The KMO method is not originally designed to identify more than two change points but, as stated in the paper, can be applied within a BS algorithm.

The LSW and the KMO methods are good benchmarks to which FreSpeD can be compared because they are among the very few methods that can analyse

multi-channel EEG data of length exceeding a few thousand and at the same time provide flexible segmentation. While the spectral method in Preuß et al. (2015) is relevant and comparable to FreSpeD it is not feasible for analysing big time series data such as the epileptic seizure data introduced in the following section, with $D = 21$ and $T^* = 50000$. In terms of computational speed, in practice KMO requires the estimation of critical values which depend on D and the interval length $e - s + 1 \leq T^*$ and thus have to be estimated for a range of interval lengths within the binary Segmentation framework. For the LSW method, Cho and Fryzlewicz (2015a) suggest a data-adaptive simulation to estimate the scale-specific thresholds in the multiscale wavelet model by fitting a simple AR(1) model and bootstrapping the errors. This turns out to be computationally intensive and we use here the work-around the authors develop for their simulation study, which is based on a set of thresholds from AR(1) models with varying parameter values. In their implementation of the change-point detection algorithm, the AR(1) parameter of each component is estimated and the appropriate threshold for each scale is selected based on this estimate.

In contrast to LSW and KMO, the FreSpeD method uses a single critical value that we discuss below. In this numerical evaluation we chose as threshold for the FreSpeD method the value $\zeta_T = 0.8 \log^{1.1}(T)$, a theoretically valid choice within the simplified framework of Section 5.3.3 and Appendix C.1. If we choose a too conservative threshold, we risk not detection true change points. If this threshold is chosen too small, we risk detecting spurious change points. However, as the FreSpeD algorithm shows (see Figure 5.2), the step of testing the immediate neighbourhood of a change-point candidate generally reduces this risk. This value

provided good results in a small initial simulation for an AR(2) process with $T^* = 50000$, $D = 2$, $D^c = 2$, $N = 1$, $\nu = 200$ and it appears to be robust for a range of window widths ν parameter settings. Further we set the minimum change-point distance $\delta_T = 1400$ (14sec). The evaluation of seizure data in the following section uses identical parameters. We provide simulation results for FreSpeD with varying window length $\nu \in \{60, 100, 400, 600\}$ in Section C.2 the appendix; the results are similar to the ones for $\nu = 200$, but as argued above the parametrization discussed in detail was chosen as it offers a good balance between frequency and time resolution for the data set with sampling rate 100Hz. We note that the results are based on the same threshold value.

Discussion of results. A total of 500 multivariate time series were generated from each setting. The results of the simulation studies are reported in Tables 5.1 to 5.4 where the total number of change points are $N = 1$ and $N = 5$, respectively, and the number of components is $D = 2$ and $D = 20$. The columns denoted $\hat{N} < N$, $\hat{N} = N$ and $\hat{N} > N$ give the percentage among 500 simulated time series in which the estimated number of change points is less than, equal to or larger than the true number of change points. The column MAD (mean absolute deviation) gives the average absolute distance between true and actual change points for those simulation runs where $\hat{N} = N$.

D^c	Process	FreSpeD				LSW				KMO, $p=6$			
		$\hat{N}<N$	$\hat{N}=N$	$\hat{N}>N$	MAD	$\hat{N}<N$	$\hat{N}=N$	$\hat{N}>N$	MAD	$\hat{N}<N$	$\hat{N}=N$	$\hat{N}>N$	MAD
1	VAR(2)	0.0	100.0	0.0	43.6	0.0	99.6	0.4	95.8	0.0	100.0	0.0	38.7
1	VAR(6)	0.0	99.6	0.4	30.9	0.0	87.6	12.4	524.6	0.0	96.8	3.2	364.0
1	VAR(10)	0.0	96.0	4.0	43.8	0.0	0.0	100.0	0.0	0.0	44.4	55.6	99.1
1	VARMA(2,2)	0.0	100.0	0.0	57.6	0.0	95.6	4.4	152.2	0.0	100.0	0.0	72.1
2	VAR(2)	0.0	99.9	0.1	43.0	0.0	100.0	0.0	46.4	0.0	100.0	0.0	19.7
2	VAR(6)	0.0	99.2	0.8	29.4	0.0	86.4	13.6	244.7	0.0	97.4	2.6	175.8
2	VAR(10)	0.0	95.9	4.1	32.7	0.0	0.0	100.0	0.0	0.0	47.6	52.4	41.3
2	VARMA(2,2)	0.0	100.0	0.0	62.4	0.0	96.4	3.6	81.3	0.0	100.0	0.0	33.1

Table 5.1: Single change point case for $D=2$. MAD: mean absolute distance between estimated and true change point if $\hat{N}=N$

D^c	Process	FreSpeD				LSW				KMO, $p=6$			
		$\hat{N}<N$	$\hat{N}=N$	$\hat{N}>N$	MAD	$\hat{N}<N$	$\hat{N}=N$	$\hat{N}>N$	MAD	$\hat{N}<N$	$\hat{N}=N$	$\hat{N}>N$	MAD
1	VAR(2)	0.0	100.0	0.0	157.1	0.0	76.6	23.4	70.2	0.0	98.0	2.0	69.1
1	VAR(6)	0.0	98.2	1.8	107.5	3.6	4.2	92.2	1701.8	90.6	9.0	0.4	467.7
1	VAR(10)	1.6	89.4	9.0	234.7	0.0	0.0	100.0	NA	14.8	7.0	78.2	3806.6
1	VARMA(2,2)	62.2	37.8	0.0	433.6	0.0	50.0	50.0	91.2	0.0	94.8	5.2	103.9
2	VAR(2)	0.0	100.0	0.0	159.2	0.0	78.4	21.6	38.3	0.0	98.8	1.2	37.4
2	VAR(6)	0.0	98.7	1.3	116.1	0.0	4.6	95.4	225.1	6.6	68.4	25.0	252.9
2	VAR(10)	1.9	88.6	9.5	234.8	0.0	0.0	100.0	NA	1.0	3.0	96.0	76.9
2	VARMA(2,2)	61.4	38.6	0.0	432.1	0.0	50.2	49.8	47.6	0.0	96.8	3.2	55.7

Table 5.2: Five change-points case for $D=2$. MAD: mean absolute distance between estimated and true change point if $\hat{N}=N$

D^c	Process	FreSpeD				LSW				KMO, $p=6$			
		$\hat{N}<N$	$\hat{N}=N$	$\hat{N}>N$	MAD	$\hat{N}<N$	$\hat{N}=N$	$\hat{N}>N$	MAD	$\hat{N}<N$	$\hat{N}=N$	$\hat{N}>N$	MAD
1	VAR(2)	0.0	99.8	0.2	39.3	0.0	87.0	13.0	92.1	0.0	100.0	0.0	41.1
1	VAR(6)	0.0	98.6	1.4	34.9	0.0	3.6	96.4	176.5	30.8	69.2	0.0	440.5
1	VAR(10)	0.0	96.6	3.4	45.5	0.0	0.0	100.0	0.0	0.0	59.2	40.8	179.3
1	VARMA(2,2)	0.0	100.0	0.0	82.4	0.0	10.3	89.7	52.0	0.0	100.0	0.0	72.1
10	VAR(2)	0.0	99.9	0.1	33.7	0.0	91.4	8.6	9.5	0.0	100.0	0.0	6.5
10	VAR(6)	0.0	99.4	0.6	42.4	0.0	7.0	93.0	24.3	0.0	100.0	0.0	41.7
10	VAR(10)	0.0	95.5	4.5	29.7	0.0	0.0	100.0	NA	0.0	86.2	13.8	17.9
10	VARMA(2,2)	0.0	100.0	0.0	69.7	0.0	16.0	84.0	8.2	0.0	100.0	0.0	8.9
20	VAR(2)	0.0	99.9	0.1	31.8	0.0	91.6	8.4	5.0	0.0	100.0	0.0	5.7
20	VAR(6)	0.0	99.3	0.7	40.2	0.0	6.7	93.3	1.7	0.0	100.0	0.0	21.4
20	VAR(10)	0.0	95.7	4.3	31.2	0.0	0.0	100.0	NA	0.0	93.2	6.8	12.0
20	VARMA(2,2)	0.0	100.0	0.0	71.2	0.0	21.0	79.0	2.8	0.0	100.0	0.0	6.9

Table 5.3: Single change point case for $D = 20$. MAD: mean absolute distance between estimated and true change point if $\hat{N} = N$

D^c	Process	FreSpeD				LSW				KMO, $p=6$			
		$\hat{N}<N$	$\hat{N}=N$	$\hat{N}>N$	MAD	$\hat{N}<N$	$\hat{N}=N$	$\hat{N}>N$	MAD	$\hat{N}<N$	$\hat{N}=N$	$\hat{N}>N$	MAD
1	VAR(2)	0.0	99.6	0.4	158.8	10.0	83.0	7.0	874.2	67.0	33.0	0.0	85.8
1	VAR(6)	0.0	98.4	1.6	115.1	20.0	9.0	71.0	3947.8	100.0	0.0	0.0	NA
1	VAR(10)	0.6	90.4	9.0	235.9	0.0	0.0	100.0	NA	100.0	0.0	0.0	NA
1	VARMA(2,2)	59.6	40.4	0.0	405.6	3.0	6.0	91.0	1079.0	100.0	0.0	0.0	NA
10	VAR(2)	0.0	99.8	0.2	158.0	1.4	88.4	10.2	83.7	0.0	100.0	0.0	10.5
10	VAR(6)	0.1	98.7	1.2	113.9	0.0	2.0	98.0	274.4	0.0	99.4	0.6	69.5
10	VAR(10)	1.5	87.0	11.5	209.3	0.0	0.0	100.0	NA	0.0	21.2	78.8	40.1
10	VARMA(2,2)	59.0	41.0	0.0	425.6	0.0	6.4	93.6	16.7	0.0	100.0	0.0	14.9
20	VAR(2)	0.1	99.7	0.2	160.4	0.6	84.6	14.8	50.1	0.0	100.0	0.0	7.3
20	VAR(6)	0.0	98.5	1.5	113.9	0.0	2.0	98.0	349.9	0.0	99.8	0.2	35.5
20	VAR(10)	1.7	87.3	11.1	218.8	0.0	0.0	100.0	NA	0.0	22.6	77.4	30.1
20	VARMA(2,2)	59.1	40.9	0.0	424.8	0.0	11.6	88.4	12.7	0.0	100.0	0.0	9.8

Table 5.4: Five change-points case for $D = 20$. MAD: mean absolute distance between estimated and true change point if $\hat{N} = N$

For the setting with only a single change point ($N = 1$) with a total of $D = 2$ components and of which there is a change in only one ($D^c = 1$) or in both ($D^c = 2$) components (Table 5.1), we find that all methods perform well for the most simple VAR process with lag $p = 2$ and for the VARMA(2,2) in terms of identifying the correct number of change points (from the column $\hat{N} = N$). For these cases, KMO is most precise in terms of MAD, while LSW is least accurate. For VAR(6) and VAR(10), FreSpeD outperforms both LSW and KMO which show a tendency to overestimate the number of change points. The poor performance of the parametric KMO method for the VAR(10) case indicates that change-point detection in the reduced subspace can be problematic, especially when the changes are happening at VAR coefficient parameters at higher lags.

For the situation where $N = 5$ change points and $D = 2$ components (Table 5.2), FreSpeD outperforms LSW and KMO in the three VAR processes in terms of estimating the correct number of change points. An exception is the VARMA(2,2) process where the change is subtle and takes place only in a very narrow frequency interval at the upper end of the frequency range. Both nonparametric methods cannot isolate this change. For the FreSpeD method we attribute this to it attenuating this difference by smoothing within frequency bands which are broader than the narrow interval on which the difference exists. The LSW method performs slightly better than FreSpeD for this VARMA(2,2) case, but shows generally a high proportion of times where it overestimates the number of change points, for all processes and independent of the number of changing components D^c. While we can only speculate about the cause, it might be related to the way the threshold is simulated following Cho and Fryzlewicz (2015a), by bootstrapping under the

assumption that the observed time series follows an AR(1) process.

Conditional on correctly estimating the correct number of change points $N = 5$, the mean absolute distance between true and estimated change points is overall larger for FreSpeD but, compared to KMO, displays a much less erratic behaviour over processes. Note that this is a natural result from the way the simulation study is designed: four of the $N = 5$ equidistant change points do not coincide with any candidate change points from the partitioning $\mathbb{T}$ using $\nu = 200$. This effect is also apparent in the supplemental analysis for varying ν provided in the appendix. The accuracy of LSW is about the same order as KMO for VAR(2) and VARMA(2,2).

Next, we examine the case where there are $D = 20$ components and $N = 1$ change points shown in Table 5.3. Regardless of how many components ($D^c = 1, 10$ or 20) change, FreSpeD displays a stable behaviour in terms of correctly identifying the number of change points. As the single change point is exactly on the border between two estimation-interval windows of length ν, the accuracy is generally good (lower than 0.5ν). In particular for $D^c = 1$ KMO shows weaknesses as it tends to underestimate the number of change points in the VAR(6) case and tends to overestimate the number of change points in the VAR(10) case, but this improves as D^c increases. LSW again overestimates the number of change points but, conditional on $\hat{N} = N$, is accurate in the change-point localization.

Finally, for the case $D = 20$ components, $N = 5$ change points (Table 5.4), KMO beats FreSpeD when the true process is VARMA(2,2). A plausible explanation for the somewhat better performance of the KMO is that the VARMA(2,2) spectrum of the parametrization used here can be well approximated by a VAR(2) spectrum and, as discussed earlier, the change is only pronounced in a narrow

frequency interval. However, it is interesting to note that KMO underestimates the number of change points for the simple VAR(2) process when the changes are subtle in only one component, while FreSpeD and LSW show stable performance over D^c. We attribute this stability to the use of the thresholded-sum CUSUM by both methods. The failure of a method to detect subtle changes could have adverse consequences in monitoring EEGs of epilepsy patients because, as we demonstrate in the EEG analysis section, some changes involve only a small subset of the channels. While FreSpeD is sensitive to these changes, it appears that the test statistic in the KMO method is dominated by the majority of the channels that exhibited no change. Overall, FreSpeD performs equally well or better than KMO and LSW when the true process is any of the three VAR processes in terms of detecting the correct number of change points.

5.5   Analysis of Seizure EEG

We now discuss a number of interesting findings from the application of the Fre-SpeD method to a seizure recording. This recording captured brain activity of a subject who suffered a spontaneous epileptic seizure while being monitored at the epilepsy center at the University of Michigan. The EEG was sampled at 100Hz (100 observations per second) and lasted for about 8.3 minutes (500sec). The total length is $T^* = 50000$. The EEG was recorded at 21 channels, 19 bipolar scalp electrodes placed according to the 10-20 system and two sphenoidal electrodes placed at the base of the temporal lobe. Figure 5.5 illustrates the placement of the scalp electrodes. In the 10-20 system, even numbers refer to right and uneven to left
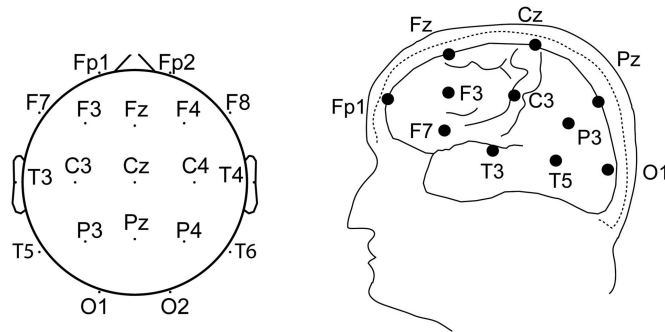
Figure 5.5: EEG scalp topography of a 10-20 electrode system

hemisphere electrodes. The abbreviations indicate the location on the scalp, from front to back: frontal polar (Fp), frontal (F), central (C), temporal (T), parietal (P) and occipital (O). Figure 5.6 gives an example of the recording, the time series of channel F3 and the estimated time-varying autospectrum with interval length $\nu = 200$.

The goal of this paper is to capture the dynamic structure of the epileptic seizure process and to identify even the subtle changes in the electrophysiology that precede seizure onset. The FreSpeD method identifies 413 change points over all 21 channels and 210 channel pairs. In total, 105 change points are identified in the channel-specific autospectra and the remaining in pairwise cross-coherences. Figure 5.7 shows the cumulative sum of detected change points over time, respectively for autospectra, coherences and total. We see that there are barely any changes detected in the preictal period, with the total number of change points rising sharply at around 5.67min (roughly at 340 seconds or at $t^* = 34,000$). This sudden increase corresponds with what the attending neurologist identified as seizure onset in this recording. This demonstrates that FreSpeD is able to data-adaptively identify the
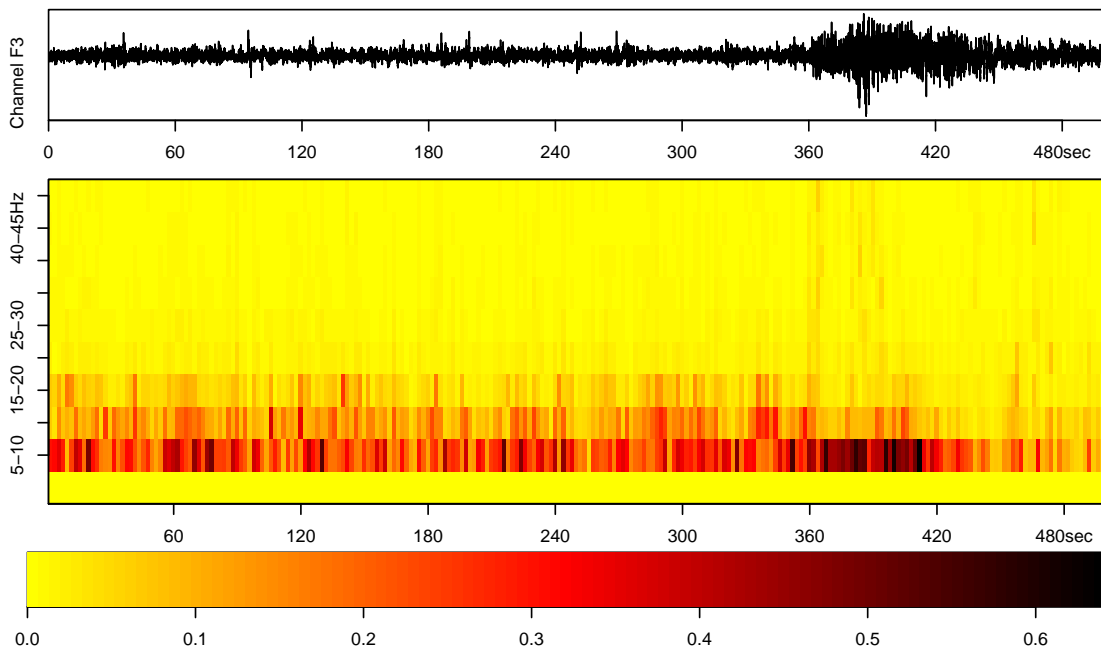
Figure 5.6: EEG recording at channel F3 (front left); top: time series; bottom: estimated autospectrum with $\nu = 200$

starting point of an epileptic seizure.

In addition to estimating the time of seizure onset, the estimated change-point distribution can be used to identify the focal point (spatial center) of this particular seizure episode. From the full recording, most change points can be identified in the autospectra of channels T3 and T5, each of which contains eight change points. Figure 5.8 illustrates where these changes are located in time and frequency. Again, this is consistent with the neurologist's diagnosis that this patient has left temporal lobe epilepsy and thus electrical activity on this area is projected to the the field on the scalp that covers the T3 and T5 channels.

For a given change point the interpretation of the value of thresholded CUSUM statistics is not straightforward. The number of frequency bands where a change is
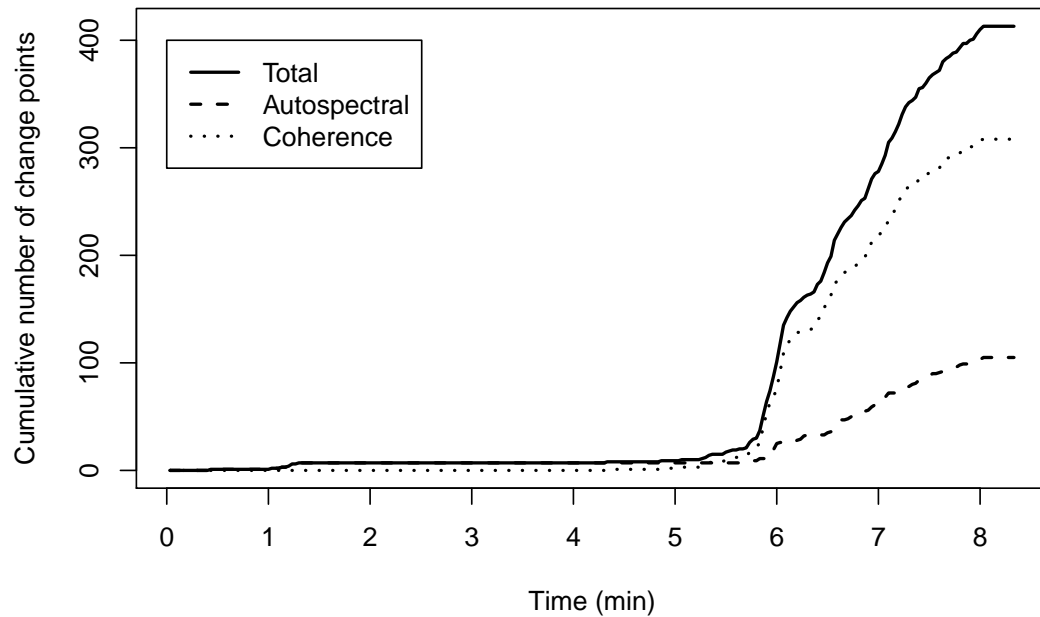
Figure 5.7: Cumulative sum of detected change points over all channels and/or channel pairs

detectable is a more robust and easily interpretable measure which illustrates the intensity or spread of a change. Both channels, T3 and T5, have more and less spread change points around the seizure onset. However, overall, there are more frequency bands showing a change at channel T3 - the primary focal point according to the neurologist. We note that this data-adaptive seizure localisation by the FreSpeD method can be of great value to neuroscientists: automated mechanisms to support and validate a neurologist's judgement diminish the risk of human error in the visual inspection of EEG traces and are more time efficient (Tzallas et al., 2012).

We now examine the occurrence of change points across frequency bands shown in Table 5.5. Most often the method was able to identify changes in the autospectra and coherence at low frequency ranges, most pronounced in the theta band. While
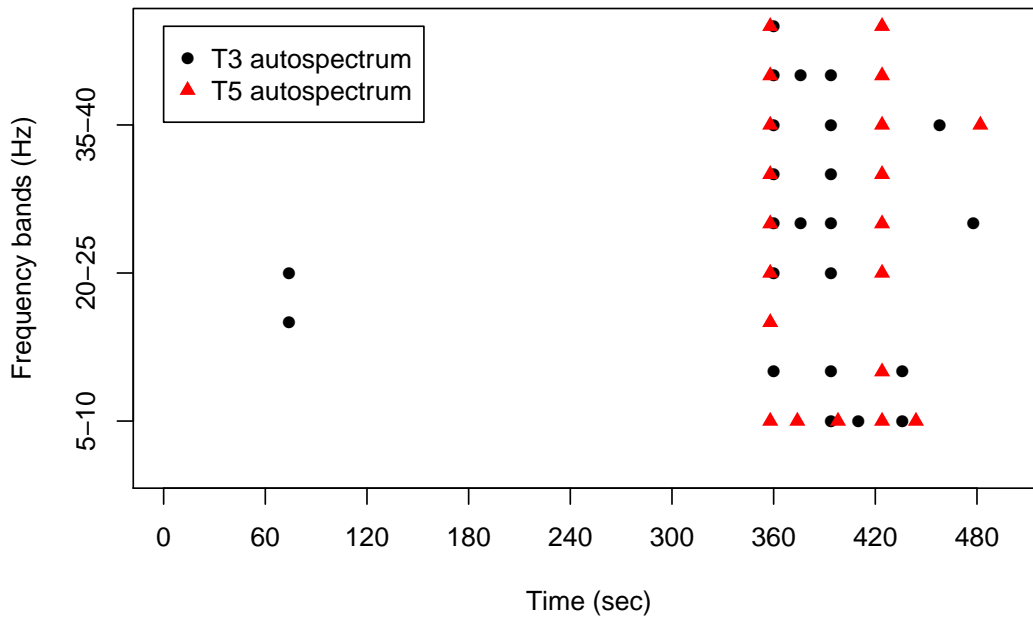
Figure 5.8: Change points in the autospectra of channels T3 and T5 in time (x-axis) vs frequency (y-axis)

we identify comparably few changes at higher frequencies, the cumulative absolute magnitude is larger and thus more emphatic. This is in-line with visual inspection and literature on seizure data analysis (Worrell et al., 2004; Jiruska et al., 2010). During normal brain states, energy concentration at high frequencies is low, but we observe a pronounced sudden increase at seizure onset. However, the observation of greater number but smaller magnitude (and therefore possibly not visually detectable) of change points at lower frequencies has not received much attention to date. This suggests that low-frequency energy and coherence vary more frequently immediately before and during epileptic seizure.

The evolution of brain processes leading up to a seizure onset is of high interest in the context of seizure warning systems. With the general aim of an improved understanding of changes in the spectral features of the brain process we now

Frequency band	Start freq.	End freq.	Prop. of total change points		Prop. of change magnitude	
			autospectra	coherences	autospectra	coherences
theta	5	10	58.1	32.2	54.8	51.3
alpha	10	15	33.3	24.0	23.1	32.4
beta	15	20	24.8	23.5	15.0	30.6
beta	20	25	24.8	18.6	28.3	28.6
beta	25	30	28.6	16.7	41.2	31.2
gamma	30	35	30.5	18.4	50.5	36.9
gamma	35	40	38.1	23.2	73.5	50.3
gamma	40	45	41.0	24.0	80.6	55.6
gamma	45	50	35.2	23.7	83.9	55.3

Table 5.5: Frequency-specific proportion of change points and change magnitude. Change magnitude is measured as sum over thresholded CUSUM statistics, over time, frequency and components or component pairs

focus on highlighting a number of findings regarding preictal EEG. However, it should be noted that seizure precursors can vary between different seizures of one patient and between patients. For this particular data, the FreSpeD method was able to detect very subtle changes that are not visually obvious to the neurologist. Firstly, a cluster of preictal change points is identified at the very start of the recording (Figure 5.9). This cluster consists of seven channel-specific changes in the autospectrum and contains no coherence change. Six of the detected changes occur between 1.00min and 1.30min (60-78sec) from the start of the recording.

There is also a very first change within the first 30 seconds of recording. In light of the instability of the CUSUM statistic at the very borders of the interval and the small magnitude of the detected change, it can be debated if this first change point at 30 seconds is spurious. However, just as each of the following six change points it is detected in the lower end of the beta-band. Within this cluster of six autospectra changing shortly after another, all channels are located in the back half of the brain and none in the frontal lobe. The largest change in terms of

magnitude is detected in channel T5, whereas channel T3 is the only channel with a change detected also in the mid-range of the beta-band, i.e. here the change is more spread over frequencies. While the temporal lag to the seizure onset makes it difficult to argue in favour of a direct link, the fact that FreSpeD independently picks up changes at six channels, all early in the second minute of the recording, reflects event-related global brain activity changes.

A second cluster can be identified at approximately 1.30min (78sec) immediately prior to seizure onset. Here, the FreSpeD method detects 13 changes in the coherence of several channel pairs and, quite interestingly and different to the fist observed cluster, none in channel-specific autospectra. As illustrated in Figure 5.9, FreSpeD captures these changes in coherence that can be described as slowly building up, culminating into the full-blown seizure onset. Furthermore, the changes are not confined at a local spatial region. Rather these were evenly distributed over the brain with 15 channels being involved, two centrally located, seven on the left and six on the right hemisphere. The changes are mostly visible in the theta/low alpha and gamma frequency bands. At the time of writing, such an observation has not received attention by researchers. In fact, these changes were not detected in other works that previously analysed this same data set such as Ombao et al. (2002, 2005) and Davis et al. (2006). The FreSpeD method demonstrates that, even prior to seizure onset, dependencies between brain regions already change, both within the seizure area and with more distant regions. As already noted, these changes are not visually identifiable. In fact, many epileptologists determine seizure onset by visual inspection of EEG traces and hence can only identify changes at single channel variance and spectral energy. The observation made here could potentially
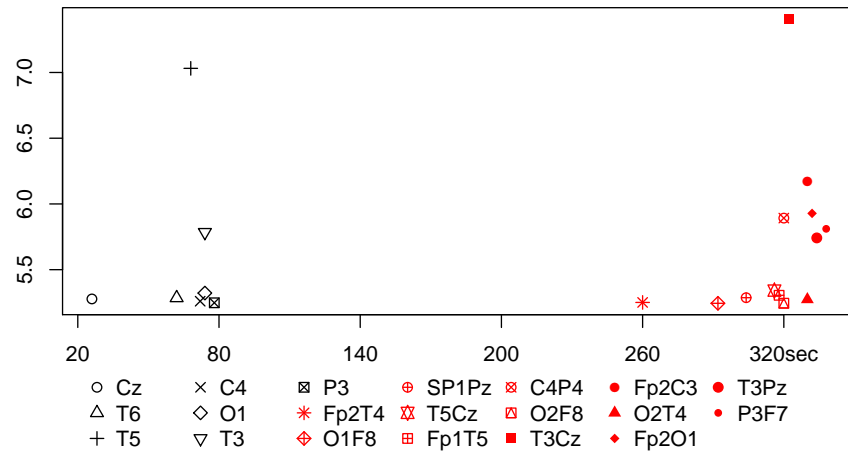
Figure 5.9: Preictal changes at all autospectra and cross-coherences. The x-axis displays time (with the seizure onset at around 340sec), the y-axis shows the level of the test statistics at the change points, averaged by number of frequencies in which a change is detected.

change the paradigm for seizure characterization that it is beyond abnormal local changes. It is worthwhile to point out that this observation is an immediate output of out method, as the detected changes have a direct interpretation with respect to frequency band and topological location.

To illustrate the detection of change points at channels and channel pairs, we provide a short video clip as supplementary material. The *.avi-file is available for download on Github (`https://github.com/almms/FreSpeD.git`) and shows a schematic 3D brain with points indicating electrodes. When a change point is detected in an autospectrum, the corresponding electrode flashes briefly. If a change is detected in the pairwise coherence between two channels, a line appears briefly connecting these electrodes. A panel below this illustration shows the time progression relative to the total EEG recording time.

5.6   Concluding Remarks

This chapter demonstrates that the FreSpeD method effectively detects change points in multi-channel EEG traces and hereby can help improve our understanding of complex brain activity such as epileptic seizure. In contrast to other methods, changes detected by FreSpeD can be directly attributed to specific frequency bands; autospectra of specific EEG channels; and cross-coherences of specific pairs of channels. To the best of our knowledge, this direct and detailed interpretability is unique to the FreSpeD method and thus makes it a particularly attractive tool to neurologists.

When applied to the multi-channel EEG recording of a spontaneous epileptic seizure, the FreSpeD method identifies changes in cross-coherence immediately before seizure onset. We emphasize that these changes are subtle and were, in fact, not detected by the methods that previously analysed the same EEG data. These changes are not sufficiently obvious to be detected through mere eyeballing. Thus the FreSpeD method gives additional insights into the complex nature of an epileptic seizure and is a new approach that can potentially identify seizure precursors. Furthermore, the FreSpeD method estimates the timing of seizure onset and the spatial focal point consistently with the neurologist's diagnosis. This data-adaptive seizure analysis via FreSpeD can support physicians' diagnoses, is time-efficient and can reduce the risk of human error (Tzallas et al., 2012; Ramgopal et al., 2014).

The comparison of FreSpeD against a parametric and a nonparametric change-point detection method for multi-channel EEG data in a simulation study shows the robustness of our approach to model misspecification. It also underlines that the

FreSpeD method works well even when the dimensionality increases. In addition, due to possible parallelization of the method and a computationally attractive algorithmic structure, FreSpeD is fast even when the number of observations in time and the number of channels is large.

Based on the approach presented here, there are a number future research directions that can be pursued. First, multi-channel EEG data with a high number of channels is typically highly collinear. This calls for transforming the data into a lower dimensional set of signals via some appropriate transformation that effectively captures the change-points dynamics of the original data. One could then apply the FreSpeD method to the summarized lower dimensional signal to discover the change-points in the original data. This would be in analogous to the recent work of Aston and Kirch (2012) but the emphasis of our work would be on detecting changes on the spectral, rather than time domain, quantities.

In the current implementation, FreSpeD is designed to detect change points in the autospectra and in (classical) coherence. Coherence at the same frequency may not completely characterize the dependence in complex data such as multi-channel EEG. A more general measure of dependence is evolutionary dual-frequency coherence where one can examine dependence between, say, the alpha oscillation in one channel and the theta oscillation in another (Gorrostieta et al., 2012). Thus, building on FreSpeD one can develop a procedure for detecting change points in dual-frequency coherence to understand the evolutionary dependence structure between EEG channels or brain regions.

5.7   Glossary: Most Essential Notation

Expression	Meaning
Functions	
$O(\cdot)$	Upper bound in the sense that $f(x) = O(g(x)) \leftrightarrow |f(x)| \leq C|g(x)|$ for all $x \geq x_0$ and some $C > 0$
$\mathbb{I}(\cdot)$	Indicator function
Objects	
$\mathbb{N} = \{\eta_i, i = 1, \ldots, N\}$ $(\hat{\mathbb{N}})$, $\mathbb{N}^{d,d'}$ $(\hat{\mathbb{N}}^{d,d'})$	True (estimated) set of change points in the process $\mathbf{X}(t^*)$; superscripts d, d' indicate channel $(d = d')$ or channel-pair $d \neq d'$ specific change points
$\mathbb{T} = \{t : t = \nu, 2\nu, \ldots, T\nu\}$	Set of partitioning of $\{1, \ldots, T^*\}$ with window length ν
$\mathfrak{C}_{s,b,e}(\mathbf{Z}) = \sum_{l=1}^{L} \mathcal{C}_{s,b,e}^*(\mathbf{z}_l)\mathbb{I}(\mathcal{C}_{s,b,e}^*(\mathbf{z}_l) > \zeta_T)$	Thresholded sum of CUSUM statistics $\mathcal{C}_{s,b,e}^*$ of the generic time series $\mathbf{Z}$ with frequency-specific components $\mathbf{z}_l, l = \{1, \ldots, L\}$
$\mathbf{X}(t^*) = (\mathbf{x}_1(t^*), \ldots, \mathbf{x}_D(t^*))$	D-channel EEG signal
$\mathbf{F}(t^*, \omega)$	Localised spectral matrix
$f_{d,d}(t^*, \omega)$	Localised autospectral density of channel d
$\rho_{d,d'}(t^*, \omega) = \frac{|f_{d,d'}(t^*,\omega)|^2}{(|f_{d,d}(t^*,\omega)||f_{d',d'}(t^*,\omega)|)}$	Localised coherence between channels d and d' $(d \neq d')$
$\bar{f}_{\nu,d,d'}(t, \omega), \bar{\rho}_{\nu,d,d'}(t, \omega)$	Averages over any window with length ν defined on the set $\mathbb{T}$
$\widehat{f}_{d,d'}(t, \omega), \widehat{\rho}_{d,d'}(t, \omega)$	Corresponding estimated quantities
$\bar{\rho}_{\nu,d,d'}^*(t, \omega), \widehat{\rho}_{d,d'}^*(t, \omega)$	Fisher-z transforms
ζ_T	Threshold parameter

Chapter 6

Conclusion

The previous chapters propose three methods to solve change-point detection problems that are relevant for different types of data. The methods are devised to estimate the number and locations of change points in time series and moreover offer insights to data beyond these estimates - this is what we call *additional interpretability* in the title of this work. One avenue we explore is interpretability via hierarchical ordering which results from binary segmentation. This can give an interpretation to change points in the context of relative importance, as illustrated in Chapter 3. The concept can be extended into a more general setting in Chapter 4, where interpretability of hierarchical orderings is possible within stages of the sequential procedure. The contribution of Chapter 5 proceeds in a different direction, enabling direct interpretation of change points in EEG data with respect to location in space (on the human scalp) and frequency band.

Chapter 3 discusses a method for local trend detection in financial time series. Within a formal statistical framework, we define a data-adaptive basis decomposition which allows for a hierarchical interpretation of change points as we illustrate with some examples. The approach also yields a family of forecasting operators which we further analyse in a comparative performance evaluation on a set of 16 financial time series from four different asset classes.

Chapter 4 draws attention to the importance of appropriate model specification in the context of changes in mean and/or variance. Based on our literature review, there is limited awareness of potential issues arising from model misspecification in the sense of not accounting for changes in mean or variance when testing for changes in, respectively, variance or mean, and there are few authors who explicitly address this issue. We argue that this limited awareness can affect the application of statistical methods negatively. After illustrating the challenges that can arise from not accounting for the presence of a mix of change types we propose a new method to address these challenges using sequential testing on intervals with varying length and show in a simulation study how this approach compares to competitors in mixed-change situations.

Finally, Chapter 5 discusses a new approach to change-point detection in EEG data with application to an epileptic seizure recording. We illustrate how this method, which is tailored to a specific scientific research question, can provide new insights that are valuable to an interdisciplinary scientific community. In particular, our approach offers a directly interpretable output which is in-line with the general convention of EEG analysis. A first implementation of the method is readily available in `R`.

At their core, all methods introduced in this work offer additional means of direct interpretability of change-point estimates. In light of the increased availability of data and rising demand for analytical tools, change-point detection methods that open avenues for data analysis beyond the estimation of change-point number and locations can be valuable in many applications.

Appendix A

Appendix of Chapter 3

A.1 Proofs

Proof of Proposition 3.2.1. Using the orthonormality of the basis $\{\boldsymbol{\psi}^{b_{j,k}}, (j,k) \in \mathcal{I}\}$, we have

$$
\begin{aligned}
\mathrm{E}(I_{j,k}) &= T^{-1}\mathrm{E}\left(T^{1/2} \sum_{(i,l)\in\mathcal{I}} A_{i,l} \sum_{t=1}^{T} \psi^{b_{i,l}}(t)\psi^{b_{j,k}}(t) \right. \\
&\quad \left. + \sum_{t=1}^{T} \sigma(t)\varepsilon(t)\psi^{b_{j,k}}(t) \right)^2 \\
&= T^{-1}\mathrm{E}\left(T^{1/2}A_{j,k} + \sum_{t=1}^{T} \sigma(t)\varepsilon(t)\psi^{b_{j,k}}(t) \right)^2 \\
&= \mathrm{E}(A_{j,k}^2) + T^{-1}\mathrm{E}\left(\sum_{t=1}^{T} \sigma(t)\varepsilon(t)\psi^{b_{j,k}}(t) \right)^2,
\end{aligned}
$$

which gives

$$
|\mathrm{E}(I_{j,k}) - \mathrm{E}(A_{j,k}^2)| \leq T^{-1}\overline{\sigma}^2 \sum_{t=1}^{T} (\psi^{b_{j,k}}(t))^2 = \overline{\sigma}^2 T^{-1}.
$$

Proof of Property 3.2.1. (i) results from the fact that $\mathbf{f}$ is a realisation of

195

a piecewise-constant random process with at most N change points and that by (ii), their magnitudes are $|\sum_{j,k} \alpha_{j,k,i} a_{j,k}|$, which is a finite quantity. For (ii), note that $f(t) - f(t-1) = T^{1/2} \sum_{(j,k) \in \mathcal{I}} a_{j,k}(\psi^{b_{j,k}}(t) - \psi^{b_{j,k}}(t-1))$ and that $T^{1/2}|\psi^{b_{j,k}}(t) - \psi^{b_{j,k}}(t-1)|$ is bounded in T as $\psi^{b_{j,k}}(t) = \mathrm{O}(T^{-1/2})$ due to the fact that the spacings between change points satisfy $\min_{i=1,\dots,N+1} |\eta_i - \eta_{i-1}| = \mathrm{O}(T)$. (iii) is implied by the fact that $P^A(|\sum_{j,k} \alpha_{j,k,i} A_{j,k}| = 0) = 0$ since the distributions of $A_{j,k}$ are continuous and mutually independent. Outside of the set A_0, (iv) holds because $\mathbf{f}$ can only have change points at $b_{j,k}$ as this is where the change points in the basis vectors $\boldsymbol{\psi}^{b_{j,k}}$ are located and the ends of their supports coincide with their parents' change points as the set $\mathcal{I}$ is connected.

Proof of Theorem 3.3.1. The proof is based on the proof of Theorem 3.1 from Fryzlewicz (2014). Firstly, we observe that our conditional signal $\mathbf{f}$ satisfies the assumptions of that Theorem, since $\{a_{j,k}, (j,k)_{\in \mathcal{I}}\} \notin A_0$. Further, Lemmas A.1 and A.2 from Fryzlewicz (2014) hold if λ_1 in those Lemmas is replaced by $\overline{\sigma}\lambda_1$. Lemma A.3 holds with $\lambda_2 = \mathrm{O}(\log T)$ and $\epsilon_T = \mathrm{O}(\log T)$ since in our case, $\delta_T = \mathrm{O}(T)$. Lemmas A.4 and A.5 hold with the respective changes to λ_1, λ_2 and ϵ_T, as above. Hence, the proof of Theorem 3.3.1 proceeds in the same way as the proof of Theorem 3.1 in Fryzlewicz (2014).

Proof of Proposition 3.3.1. The proof proceeds similarly to that of Property 3.2.1 (iii) by noting that $\langle \mathbf{f}, \boldsymbol{\psi}^{s,\eta_i,e} \rangle$ is a linear combination of $\{a_{j,k}, (j,k) \in \mathcal{I}\}$, different for each i such that $\eta_i \in [s, e]$.

Proof of Theorem 3.3.2. The proof is a straightforward modification of the proof of Theorem 3.3.1, which itself uses the proof of Theorem 3.1 from Fryzlewicz (2014). It is sufficient to observe that in our context, a stronger version of Lemma A.2 in the

latter work holds, whereby η_{p_0+r} achieves the unique maximum of $|\langle \mathbf{f}, \boldsymbol{\psi}^{s,\eta,e}\rangle|$ over $\eta \in [s, e]$. This is because if the unique maximum were achieved by $\eta_{p_0+q} \neq \eta_{p_0+r}$, then $|\langle \mathbf{f}, \boldsymbol{\psi}^{s,\eta_{p_0+q},e}\rangle| > |\langle \mathbf{f}, \boldsymbol{\psi}^{s,\eta_{p_0+r},e}\rangle| = \mathrm{O}(T^{1/2})$ by Lemma 1 of Cho and Fryzlewicz (2012), and therefore the b maximising $|\langle \mathbf{x}, \boldsymbol{\psi}^{s,b,e}\rangle|$ would have to fall near η_{p_0+q}, rather than η_{p_0+r} by Lemma A.1 from Fryzlewicz (2014).

For the estimation of $\bar{a}_{j,k}$, we have

$$
\begin{aligned}
|\hat{a}_{j,k} - \bar{a}_{j,k}| &= T^{-1/2}|\langle \mathbf{f}, \boldsymbol{\psi}^{b_{j,k}}\rangle - \langle \mathbf{x}, \boldsymbol{\psi}^{\hat{b}_{j,k}}\rangle| \\
&\leq T^{-1/2}\left\{ |\langle \mathbf{f}, \boldsymbol{\psi}^{b_{j,k}}\rangle - \langle \mathbf{f}, \boldsymbol{\psi}^{\hat{b}_{j,k}}\rangle| \right. \\
&\qquad \left. + |\langle \mathbf{x}, \boldsymbol{\psi}^{\hat{b}_{j,k}}\rangle - \langle \mathbf{f}, \boldsymbol{\psi}^{\hat{b}_{j,k}}\rangle| \right\} \\
&\leq T^{-1/2}\left\{ \mathrm{O}(\epsilon(t)T^{-1/2}) + \mathrm{O}(\log^{1/2} T) \right\} \\
&= \mathrm{O}(T^{-1/2}\log^{1/2} T),
\end{aligned}
$$

where we use, respectively, the triangle inequality, a technique as in Lemma 2 of Cho and Fryzlewicz (2012) and Lemma A.1 of Fryzlewicz (2014).

A.2 Data

Details of the financial return data considered in this chapter are provided in Table A.1.

A.2.1 Details on the Out-of-Sample Analysis with a Seven-Year Estimation Period

Asset Class	Asset	Bloomberg Ticker	Primary Quote/Trading Venue	Obs
Equity Index	Dax Index	DAX Index	Germany - DB	5,936
Equity Index	FTSE 100 Index	UKX Index	United Kingdom - LSE	5,930
Equity Index	Hang Seng Index	HSI Index	Hong Kong - HKSE	5,802
Equity Index	S&P 500 Index	SPX Index	United States - NYSE	5,915
Equity	Deutsche Lufthansa	LHA GR Equity	Germany - Xetra	5,891
Equity	General Electric	GE US Equity	United States - NYSE	5,915
Equity	Johnson & Johnson	JNJ US Equity	United States - NYSE	5,915
Equity	Microsoft	MSFT US Equity	United States - Nasdaq	5,915
Currency	AUD/USD Spot rate	GBPUSD Curncy	London Composite	6,124
Currency	GBP/USD Spot rate	USDJPY Curncy	London Composite	6,124
Currency	USD/JPY Spot rate	AUDUSD Curncy	London Composite	6,124
Currency	USD/MXN Spot rate	USDMXN Curncy	London Composite	6,124
Commodity Future	Crude Oil	CL1 Comdty	New York Mercantile Exchange	5,895
Commodity Future	Gold	GC1 Comdty	CMX-Commodity Exchange	5,895
Commodity Future	Live Cattle	LC1 Comdty	Chicago Mercantile Exchange	5,923
Commodity Future	Sugar	SB1 Comdty	NYB-ICE Futures US Softs	5,880

Table A.1: Data series used in the empirical evaluation; data provider: Bloomberg; number of observations (Obs) corresponds to number of days for which a quote is available between 1 January 1990 and 21 June 2013.

Forecast horizon	1	2	5	10	15	20	40
DAX	2093	1046	418	209	139	104	52
FTSE	2090	1045	418	209	139	104	52
HSI	2026	1013	405	202	135	101	50
S&P	2082	1041	416	208	138	104	52
LHA	2070	1035	414	207	138	103	51
GE	2082	1041	416	208	138	104	52
JNJ	2082	1041	416	208	138	104	52
MSFT	2082	1041	416	208	138	104	52
GBPUSD	2187	1093	437	218	145	109	54
USDJPY	2187	1093	437	218	145	109	54
AUDUSD	2187	1093	437	218	145	109	54
USDMXN	2187	1093	437	218	145	109	54
Oil	2072	1036	414	207	138	103	51
Gold	2072	1036	414	207	138	103	51
Live Cattle	2086	1043	417	208	139	104	52
Sugar	2064	1032	412	206	137	103	51

Table A.2: Number of forecasts of the out-of-sample analysis: based on original data of Table A.1, using a seven-year estimation period, non-overlapping forecast windows and, respectively, half of the remaining as training and test period

Forecast horizon	Overall best	1	2	5	10	15	20	40
DAX	0.8	0.3	1	0.2	0.6	0.8	0.3	0.4
FTSE	0.2	0.5	1.4	0.1	0.6	0.8	0.6	0.5
HSI	0.8	0.9	0.1	0.1	0.8	0.8	0.8	0.4
S&P	0.1	0.2	2.9	0.2	0.5	0.1	0.5	0.9
LHA	0.5	0.3	0.9	0.5	0.6	0.5	0.6	0.6
GE	0.1	0.1	0.1	0.1	0.1	0.1	0.8	0.5
JNJ	0.1	0.2	0.2	0.1	1	0.2	0.7	1
MSFT	0.1	0.1	0.1	0.9	0.8	1.4	1.1	0.2
GBPUSD	1	0.1	2.1	0.1	0.6	1.4	0.6	1
USDJPY	0.4	0.3	0.1	0.5	1.5	1.4	1.4	0.4
AUDUSD	0.8	1.4	0.8	0.6	0.8	0.3	0.8	0.8
USDMXN	0.1	2.7	0.1	0.1	0.6	0.8	0.1	1.7
Oil	0.1	1.3	0.7	1.3	0.1	0.6	1.3	1.3
Gold	0.1	0.1	0.7	1	1	0.1	0.1	0.1
Live Cattle	0.6	0.4	2.8	0.8	0.3	0.4	0.2	0.6
Sugar	0.2	0.3	3	3	2.7	2.4	3	0.2

Table A.3: Optimal in-sample C parameter in our model, per forecast horizon, seven-year estimation period

Forecast horizon	Overall best	1	2	5	10	15	20	40
DAX	6	8	9	10	3	9	6	6
FTSE	1	3	3	6	8	9	9	2
HSI	2	3	10	3	1	1	4	2
S&P	1	3	9	3	10	7	9	5
LHA	4	3	5	4	1	4	8	4
GE	1	2	8	8	6	9	10	5
JNJ	1	2	8	1	9	10	6	6
MSFT	1	1	2	9	3	6	1	1
GBPUSD	2	2	10	5	7	6	1	2
USDJPY	6	3	7	8	6	3	10	6
AUDUSD	3	4	4	6	10	10	7	3
USDMXN	9	6	1	10	7	9	7	9
Oil	3	1	1	3	7	2	9	3
Gold	1	7	1	6	5	3	9	1
Live Cattle	3	10	1	10	1	7	9	3
Sugar	1	2	7	3	4	2	4	8

Table A.4: Optimal in-sample MA parameter in the benchmark model, per forecast horizon, seven-year estimation period

A.2.2 Results of the Out-of-Sample Analysis with a Two-Year Estimation Period

Forecast horizon	1	2	5	10	15	20	40
DAX	2718	1359	543	271	181	135	67
FTSE	2715	1357	543	271	181	135	67
HSI	2651	1325	530	265	176	132	66
S&P	2707	1353	541	270	180	135	67
LHA	2695	1347	539	269	179	134	67
GE	2707	1353	541	270	180	135	67
JNJ	2707	1353	541	270	180	135	67
MSFT	2707	1353	541	270	180	135	67
GBPUSD	2812	1406	562	281	187	140	70
USDJPY	2812	1406	562	281	187	140	70
AUDUSD	2812	1406	562	281	187	140	70
USDMXN	2812	1406	562	281	187	140	70
Oil	2697	1348	539	269	179	134	67
Gold	2697	1348	539	269	179	134	67
Live Cattle	2711	1355	542	271	180	135	67
Sugar	2689	1344	537	268	179	134	67

Table A.5: Number of forecasts of the out-of-sample analysis: based on original data of Table A.1, using a two-year estimation period, non-overlapping forecast windows and, respectively, half of the remaining as training and test period

Forecast horizon	1		2		5		10		15		20		40	
	UHP	BM	UHP	BM	UHP	BM	UHP	BM	UHP	BM	UHP	BM	UHP	BM
DAX	**3.00**	0.34	2.45	-0.14	3.06	-3.06	2.38	-0.34	7.14	**8.16**	1.35	**4.05**	-1.35	-1.35
FTSE	0.72	-2.43	**1.16**	0.75	2.22	-3.92	**4.42**	0.34	10.82	-1.55	**8.11**	4.05	-1.35	1.35
HSI	-0.89	0.68	**1.07**	0.07	0.36	-0.36	-	0.71	-4.26	1.06	**-2.86**	-1.43	4.29	4.29
S&P	-0.51	**-2.20**	3.16	3.16	2.05	**8.90**	**9.59**	5.48	**13.92**	8.76	2.05	**3.42**	**17.57**	6.76
LHA	-0.70	0.31	**-2.57**	-0.21	-1.04	3.63	-0.35	5.24	1.55	**9.79**	-0.68	**-2.05**	8.33	**16.67**
GE	-0.17	**-1.66**	-0.90	0.35	**5.02**	0.52	**3.42**	1.37	**6.70**	3.61	3.42	-10.27	-1.35	**-4.05**
JNJ	0.80	**0.84**	**1.80**	0.07	**-3.98**	-2.10	8.62	-4.48	-2.08	5.21	**-7.53**	-3.42	1.35	-1.35
MSFT	0.21	-1.33	0.34	-0.42	**-1.56**	-1.05	**0.68**	0.34	**-4.64**	-2.58	6.16	-0.68	1.35	**4.05**
GBPUSD	**0.58**	0.32	**-1.41**	-0.71	-0.64	2.88	-5.48	3.90	**-5.77**	-3.85	-12.82	5.13	-3.85	6.41
USDJPY	0.42	-1.07	1.23	-0.71	2.73	-3.23	-0.64	-0.64	**1.92**	0.96	5.13	-2.56	8.97	**14.10**
AUDUSD	**1.51**	0.13	**1.41**	0.77	-2.24	2.56	5.13	**6.41**	-5.77	0.96	-2.56	8.97	-1.28	1.28
USDMXN	0.83	**1.76**	0.96	**2.88**	-3.85	2.88	1.28	-1.28	-8.25	8.25	3.85	-6.41	-2.63	**-10.53**
Oil	1.87	-1.01	3.25	-0.21	**5.71**	0.17	**5.86**	1.72	3.61	-1.55	3.42	-2.05	**13.89**	2.78
Gold	-0.62	**-1.43**	5.94	-1.38	-2.25	1.21	**9.31**	4.48	-0.52	0.52	-2.05	2.05	2.78	**8.33**
Live Cattle	-1.95	**-2.80**	-0.21	**-2.02**	3.45	-2.08	**-5.17**	-2.41	4.64	-6.70	7.75	-3.52	**12.16**	9.46
Sugar	**-1.58**	-0.21	0.35	**1.32**	1.39	**3.66**	2.08	-1.75	**3.13**	1.04	-5.56	0.00	-5.56	8.33

Table A.6: Relative success ratio RSR in percent in the out-of-sample test; for our model with in-sample optimized threshold value and the benchmark model (BM) with in-sample optimized moving average window length, for a *two-year* estimation period. If the relative success ratio has the same sign for both models, the larger absolute value is in bold.

Forecast horizon	Overall best	1	2	5	10	15	20	40
DAX	0.8	0.7	0.7	0.5	0.6	0.8	0.8	0.7
FTSE	0.5	0.5	1.6	0.5	0.6	0.8	0.5	0.5
HSI	0.8	0.5	0.5	0.1	0.1	0.4	0.8	1.3
S&P	0.1	0.1	3	0.2	0.9	0.5	0.1	0.4
LHA	0.1	2.8	1.3	2.8	0.6	0.5	1.4	0.5
GE	0.3	0.2	0.1	0.1	0.2	0.1	0.8	0.3
JNJ	0.2	0.2	0.2	0.1	0.9	0.2	0.2	1
MSFT	0.1	0.1	0.1	0.6	1.1	1.4	0.2	1.1
GBPUSD	1	0.1	2.1	0.1	0.2	2.5	0.2	1
USDJPY	0.4	1.5	0.5	0.5	0.1	1.6	0.4	0.4
AUDUSD	0.3	0.8	0.8	0.2	0.8	0.3	0.8	0.8
USDMXN	1.7	0.6	0.1	0.1	0.6	3	0.1	1.7
Oil	0.1	1.3	0.7	1.3	1.3	1.3	1.3	1.3
Gold	0.1	0.1	0.7	0.1	2.1	0.1	0.1	2.1
Live Cattle	0.6	0.3	0.1	0.8	0.3	0.8	0.2	0.6
Sugar	0.7	1.2	3	3	2.7	2.4	0.7	0.1

Table A.7: Optimal in-sample C parameter in our model, per forecast horizon for the two-year estimation period analysis

Forecast horizon	Overall best	1	2	5	10	15	20	40
DAX	6	9	10	10	3	9	6	6
FTSE	5	3	3	4	8	9	9	5
HSI	1	3	7	3	1	2	8	4
S&P	1	1	9	3	10	10	9	5
LHA	5	3	5	3	7	5	1	4
GE	7	2	1	8	7	4	2	7
JNJ	1	2	8	1	2	6	6	3
MSFT	1	1	2	1	10	6	1	3
GBPUSD	2	1	8	3	1	5	9	2
USDJPY	6	3	7	5	6	6	1	6
AUDUSD	4	5	7	5	9	10	7	4
USDMXN	1	6	1	10	7	5	7	9
Oil	3	1	7	3	7	7	9	3
Gold	3	1	7	1	6	3	9	1
Live Cattle	3	10	1	4	8	7	7	3
Sugar	4	1	7	8	4	7	7	10

Table A.8: Optimal in-sample MA parameter in the benchmark model, per forecast horizon for the two-year estimation period analysis

Appendix B

Appendix of Chapter 4

B.1   Additional Figures
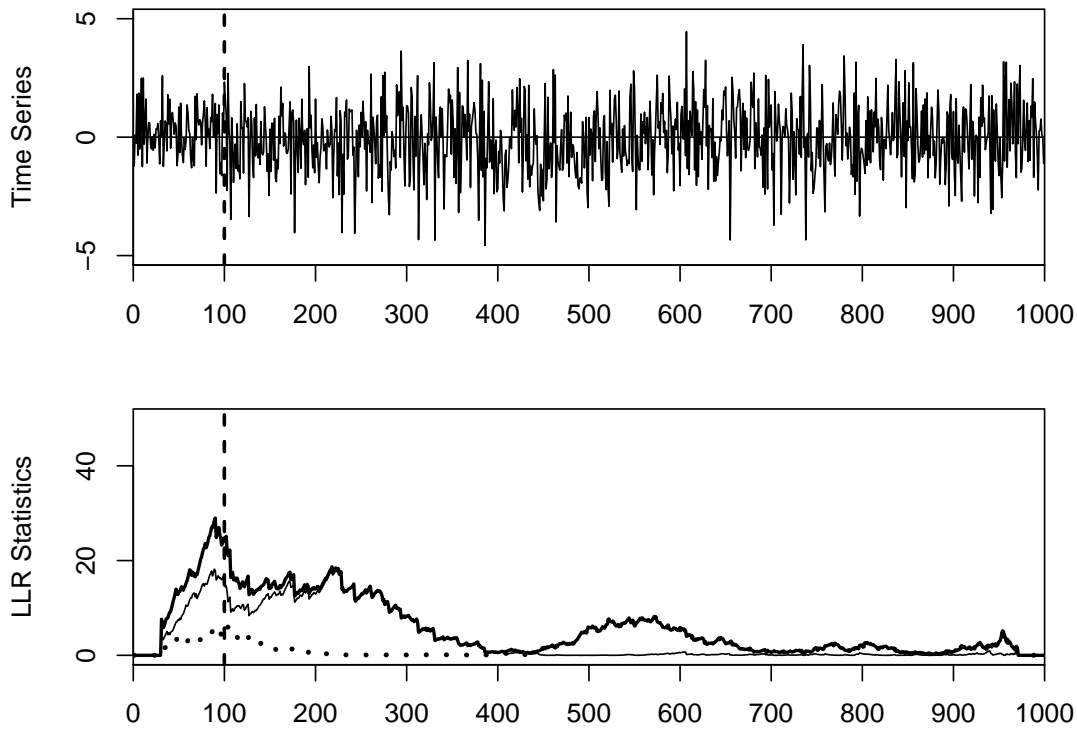
Figure B.1: Realization of the process of Figure 4.6 with changing variance only. At $\eta_1 = 100$ the variance from 1 to 1.5^2. Upper panel: process realization. Lower panel: test statistics $\mathcal{LLR}^{(k)}(\mathbf{x}|b), k = \{1, 2, 3\}$. Continuous thick line: change in mean and variance, $k = 1$. Continuous thin line: change in variance, $k = 2$. Dotted thick line: change in mean, $k = 3$
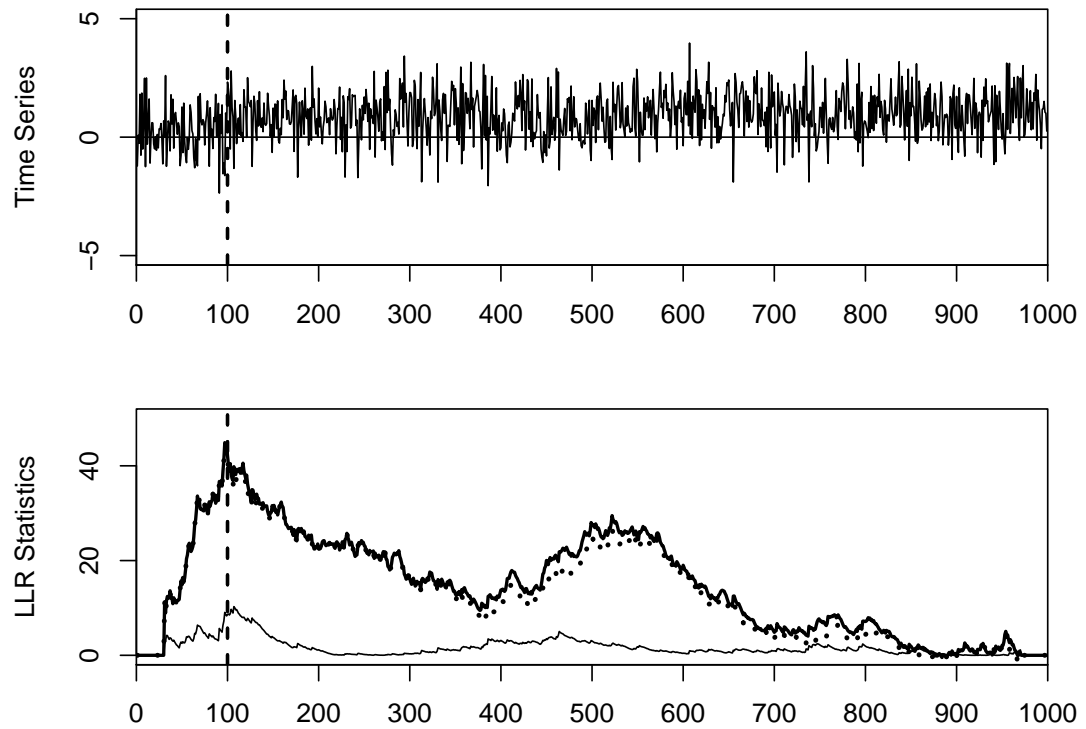
Figure B.2: Realization of the process of Figure 4.6 with changing mean only. At $\eta_1 = 100$ the mean increases from 0 to 1. Upper panel: process realization. Lower panel: test statistics $\mathcal{LLR}^{(k)}(\mathbf{x}|b), k = \{1, 2, 3\}$. Continuous thick line: change in mean and variance, $k = 1$. Continuous thin line: change in variance, $k = 2$. Dotted thick line: change in mean, $k = 3$

Appendix C

Appendix of Chapter 5

C.1 Proof in the Simplified Framework

The structure of the proof is as follows. We first show within the simplified framework of Section 5.3.3 the consistency of change-point detection for estimated spectral energy at a single frequency band ω_l. Then we extend this to the thresholded sum statistic of Equation (5.6). Where possible we drop the channel-indicating subscripts (d, d'). All of the following is applied to the partitioning $\mathbb{T} = \{t : t = \nu, 2\nu, \ldots, T\nu\}$. The proof can be compressed by noting that the core principles equally apply to autospectra and cross-coherences. Whenever this is the case we use notation consistent with that of Section 5.3.2 and denote the estimated time series as $\widehat{\mathbf{z}}_l = \{\widehat{z}_l(t), t \in \mathbb{T}\}$, $l \in \{1, \ldots, L\}$ with $\widehat{z}_l(t) := \widehat{f}(t, \omega_l)$ or $\widehat{z}_l(t) := \widehat{\rho}^*(t, \omega_l)$, and the true average quantities correspondingly as $\bar{z}_l(t), t \in \mathbb{T}$, where the subscript l refers to a frequency band.

C.1.1   CUSUM for a Single Frequency Band ω_l

Let us fist introduce some notation. On an interval $\{s, \ldots, e\}$ with $\{(s, e) \in \mathbb{T} \times \mathbb{T} : \nu \leq s < e \leq T$ we define the statistics

$$\mathcal{C}_{s,b,e}(\widehat{z}_l) = \left| \sqrt{\frac{e-b}{n(b-s+1)}} \sum_{t=s}^{b} \widehat{z}_l(t) - \sqrt{\frac{b-s+1}{n(e-b)}} \sum_{t=b+1}^{e} \widehat{z}_l(t) \right| \tag{C.1}$$

for the time-varying estimator of the spectral quantity and, correspondingly, $\mathcal{C}_{s,b,e}(z_l)$ for the true quantities, with $n = e - s + 1$. Note that this is the CUSUM statistic of Equation (5.7) up to scaling: $\mathcal{C}_{s,b,e}(x) = \sigma_{s,e}(x)\mathcal{C}^*_{s,b,e}(x)$.

At any stage of the algorithm, for $1 \leq i_1 < i_2 \leq N_{d,d'}$ $\eta^{d,d'}_{i_1} \leq s < \eta^{d,d'}_{i_1+1} < \cdots < \eta^{d,d'}_{i_2} < e \leq \eta^{d,d'}_{i_2+1}$. In the following all results hold for subsets $\mathbb{N}^{d,d}, \mathbb{N}^{d,d'} \subseteq \mathbb{N}$. If, at any stage, there are change points on $\{s, \ldots, e\}$ then both of the following conditions hold. This will become obvious on the next pages.

$$s < \eta_{i_1+i} - c_1\delta_T < \eta_{i_1+i} + c_1\delta_T < e \text{ for some } 1 \leq i \leq i_2 - i_1 \tag{C.2}$$

$$\max\left\{\min(\eta_{i_1+1} - s, s - \eta_{i_1}), \min(\eta_{i_2+1} - e, e - \eta_{i_2})\right\} \leq c_2\epsilon_T \tag{C.3}$$

Here, c_1, c_2 are positive constants and ϵ_T is as defined in Section 5.3.3. The following Lemmata are adapted from Cho and Fryzlewicz (2015a) and the corrections available in Cho and Fryzlewicz (2015a).

Lemma 1 *Let $\{s, \ldots, e\}$ satisfy Equation (C.2). Then there exists a i^* where $1 \leq i^* \leq i_2 - i_1$ such that for a positive constant C,*

$$|\mathcal{C}_{s,\eta_{i_1+i^*},e}(\bar{z}_l)| \geq C\delta_T/\sqrt{T} \tag{C.4}$$

Proof For the single change-point case, $i^* = 1$, $\bar{z}_l(t)$ is piecewise-constant before and after $\eta_{i_1+i^*}$ on $\{s, \ldots, e\}$. Thus

$$
\begin{aligned}
|\mathcal{C}_{s,\eta_{i_1+i^*},e}(\bar{z}_l)| &= \sqrt{(\eta_{i_1+i^*} - s + 1)(e - \eta_{i_1+i^*})}/\sqrt{n}\,|\bar{z}_l(\eta_{i_1+i^*} + 1) - \bar{z}_l(\eta_{i_1+i^*})| \\
&\geq z_* c_1 \delta_T/\sqrt{T},
\end{aligned}
$$

where z_* is the lower bound imposed on change sizes in the corresponding spectral quantities and the inequality follows from the assumptions in Section 5.3.3. For the case of multiple change points, by the assumptions of boundedness and minimum separability for any interval satisfying the condition of Equation (C.2) there exists at least one i^* s.t.

$$
\frac{1}{\eta_{i_1+i^*} - s + 1} \sum_{t=s}^{\eta_{i_1+i^*}} \bar{z}_l(t) - \frac{1}{e - \eta_{i_1+i^*}} \sum_{t=\eta_{i_1+i^*}+1}^{e} \bar{z}_l(t)
$$

is bounded away from zero. This generates the same situation as for the single change-point case and thus Lemma 1 is shown.

Lemma 2 *Under the condition of Equation (C.2), there exists a $C > 0$ such that for points $b \in \{s, \ldots, e\}$ with $|\eta_{i_1+i} - b| > c_0 \epsilon_T$ for some k and $\mathcal{C}_{s,\eta_{i_1+i},e}(\bar{z}_l) > \mathcal{C}_{s,b,e}(\bar{z}_l)$,*

$$
\mathcal{C}_{s,\eta_{i_1+i},e}(\bar{z}_l) > \mathcal{C}_{s,b,e}(\bar{z}_l) + C\epsilon_T \delta_T/T^2 \mathcal{C}_{s,\eta_{i_1+i},e}(\bar{z}_l) \tag{C.5}
$$

Proof Without loss of generality, let $\eta \equiv \eta_{i_1+i} < b$. The proof follows directly from the proof of Lemma 2.6, Case 2 in Venkatraman (1992), which is at three full pages rather lengthy. Essentially, one can show that as $T \to \infty$, the difference

$\mathcal{C}_{s,\eta_{i_1+i},e}(\bar{z}_l) - \mathcal{C}_{s,b,e}(\bar{z}_l)$ is bounded below by a term is dominated by what Venka-traman (1992) calls E_{1l}. In our notation and noting that $\delta_T > 2c_0\epsilon_T$ for large T,

$$
\begin{aligned}
&E_{1l} \\
&= \frac{(\delta_T - c_0\epsilon_T)c_0\epsilon_T\sqrt{n_\eta + n_e + \delta_T}\left[(\eta_{i_1+i} - s + 1)^{-1}\sum_{t=s}^{\eta_{i_1+i}}\bar{z}_l(t) - (e - \eta_{i_1+i})^{-1}\sum_{t=\eta_{i_1+i}+1}^{e}\bar{z}_l(t)\right]}{\sqrt{n_\eta(n_e + \delta_T)(n_\eta + c_0\epsilon_T)(n_e + \delta_T - c_0\epsilon_T)}(\sqrt{(n_\eta + c_0\epsilon_T)(n_e + \delta_T - c_0\epsilon_T)} + \sqrt{n_\eta(n_e + c_0\epsilon_T)})} \\
&\geq C\epsilon_T\delta_T/T^2\mathcal{C}_{s,\eta_{i_1+i},e}(\bar{z}_l)
\end{aligned}
$$

where $n_\eta = \eta_{i_1+i} - s + 1$ and $n_e = e - \eta_{i_1+i} - \delta_T$.

Lemma 3 *Let* $\mathcal{D} = \{(s,b,e) : 1 \leq s < e \leq T; n = e - s + 1 \geq \delta_T; \max(b - s + 1, e - b) \leq nc_*\}$, *where* c_* *as on page 169. Then as* $T \to \infty$,

$$
\Pr\left(\max_{(s,b,e)\in\mathcal{D}}|\mathcal{C}_{s,b,e}(\hat{f}(\omega_l)) - \mathcal{C}_{s,b,e}((2M)^{-1}\bar{f}(\omega_l))| > \log T\right) \to 0 \qquad \text{(C.6)}
$$

Proof We can drop at this point non-relevant subscripts and, using Equation (5.4), consider

$$
\begin{aligned}
&\Pr\left(|\sum_{t=s}^{e} c(t)\bar{f}(t,\omega_l)(\chi^2_{2M}(t) - 2M)/(2M)| > \sqrt{n}\log T\right) \\
&= \Pr\left(|\sum_{t=s}^{e} c(t)\bar{f}(t,\omega_l)\sum_{m=1}^{2M}(U_m^2(t) - 1)/(2M)| > \sqrt{n}\log T\right)
\end{aligned}
$$

with $c(t) = \sqrt{e - b}/\sqrt{b - s + 1}$ if $s \leq t \leq b$ and $c(t) = -\sqrt{b - s + 1}/\sqrt{e - b}$ if $b < t \leq e$, which are finite on $\mathcal{D}$ as $|c(t)| \leq c^* \equiv \sqrt{c_*/(1 - c_*)} < \infty$. $\chi^2_{2M}(t) - 2M = \sum_{m=1}^{2M}(U_m^2(t) - 1)$ where $U_m(t)$ are independent standard normal random variables. By standard results there exists $C > 0$ s.t. $\mathrm{E}(|U_m^2 - 1|)^j \leq C^{j-2}j!\,\mathrm{E}(|U_m^2 - 1|)^2$ (see

e.g. Johnson and Kotz, 1970). Thus by Bernstein's inequality (Bosq, 1998),

$$
\Pr \left(|\sum_{t=s}^{e} \sum_{m=1}^{2M} \frac{c(t)\bar{f}(t,\omega_l)}{(2M)^2}(U_m^2(t)-1)| > \sqrt{n}\log T \right)
$$
$$
\leq \quad \exp\left\{ -\frac{1/2n\log^2 T}{2\sum_{t=s}^{e}\frac{(c(t)\bar{f}(t,\omega_l))^2}{(2M)^3} + C\max_t \frac{c(t)\bar{f}(t,\omega_l)}{(2M)^2}\sqrt{n}\log T} \right\}
$$

As $|c(t)| \leq c^*$ and $|\bar{f}(t,\omega_l)| \leq f^*$, it follows that

$$
\Pr \left(\max_{(s,b,e)\in\mathcal{D}} |\mathcal{C}_{s,b,e}(\widehat{f}(\omega_l)) - \mathcal{C}_{s,b,e}((2M)^{-1}\bar{f}(\omega_l))| > \sqrt{n}\log T \right)
$$
$$
\leq \quad \sum_{(s,b,e)\in\mathcal{D}} \exp\left\{ -\frac{1/2n\log^2 T}{2n\frac{(c^*f^*)^2}{(2M)^3} + C\frac{c^*f^*}{(2M)^2}\sqrt{n}\log T} \right\}
$$
$$
\leq \quad T^3 \exp(-C\log^2 T)
$$

The last converges to zero as we assume $n \geq \delta_T \geq \log T$ and that all constants are finite.

Lemma 4 *Under $\mathcal{D}$ as in Lemma 3 and the assumptions of Section 5.3.3, as $T \to \infty$*

$$
\Pr \left(\max_{(s,b,e)\in\mathcal{D}} |\mathcal{C}_{s,b,e}(\widehat{\rho}^*(\omega_l)) - \mathcal{C}_{s,b,e}(\bar{\rho}^*(\omega_l))| > \log T \right) \to 0 \tag{C.7}
$$

Proof From Equation (5.5) we see that the difference between estimated and true coherence is i.i.d. normal random variables. Thus, using the Bernoulli inequality,

for a standard normal variable ε with pdf $\phi_N(\cdot)$ and $\lambda_1 \geq \log T$,

$$\Pr\left(\max_{(s,b,e)\in\mathcal{D}} |\mathcal{C}_{s,b,e}(\widehat{\rho}^*(\omega_l)) - \mathcal{C}_{s,b,e}(\bar{\rho}^*(\omega_l))| > \lambda_1\right)$$

$$\leq \sum_{(s,b,e)\in\mathcal{D}} \Pr\left(|\varepsilon| > \lambda_1\right) \leq \frac{T^3\phi_N(\lambda_1)}{\lambda_1} \leq \frac{C}{T}$$

Lemma 5 *Under Equations (C.2) and (C.3), for $\mathcal{D}_{s,e} = \{s < t < e; n = e-s+1 \geq \delta_T; \max(t-s+1, e-t) \leq nc_*\}$, with c_* as on page 169, there exists $1 \leq i^* \leq i_2 - i_1$ s.t. $\eta_{i_1+i^*} \in \mathcal{D}_{s,e}$ and $|\eta_{i_1+i^*} - b| < c_0\epsilon_T$ for $b = argmax_{t\in\mathcal{D}_{s,e}} |\mathcal{C}_{s,t,e}(\widehat{f}(\omega_l))|$, or $b = argmax_{t\in\mathcal{D}_{s,e}} |\mathcal{C}_{s,t,e}(\widehat{\rho}^*(\omega_l))|$, respectively.*

Proof   The following covers the general discussion of the CUSUM statistic in the additive function plus noise setting. For the multiplicative model for the autospectra in Equation (5.4) we rewrite

$$\widehat{f}(t,\omega_l) = \frac{1}{2M}\bar{f}(t,\omega_l)\chi^2_{2M}(t) = \bar{f}(t,\omega_l) + \frac{1}{2M}\bar{f}(t,\omega_l)(\chi^2_{2M}(t) - 2M)$$

Thus, we can formulate a generic model $\widehat{z}_l(t) = g(t) + e(t)$ with piecewise-constant function $g(t)$ (with possibly many change points) that covers the autospectral case (Equation (5.4)) and the coherence case (Equation (5.5)) for energy at a frequency band ω_l. Following Venkatraman (1992, Lemmata 2.2-2.3), under the assumption of balancedness in Section 5.3.3 and the conditions specified in Equations (C.2) and (C.3) for $g^0(t)$ denoting a one-step piecewise-constant function on $\{s, \ldots, e\}$ that has its change point η on $\mathcal{D}_{s,e}$ and minimizes the mean-square distance between $g(t)$ and any one-step piecewise-constant function on $\{s, \ldots, e\}$, η corresponds to one of the change points of $g(t)$ on $\{s, \ldots, e\}$.

In the framework of piecewise-constant function plus noise, detecting a change point on an interval $\{s, \ldots, e\}$ implies finding the mean-square optimal fit of a step function with a single step at b, denoted as $\widehat{g}_b(t)$. We know that for $b^* = \mathrm{argmin}_b \sum_{t \in \{s, \ldots, e\}} (\widehat{z}(t) - \widehat{g}_b(t))^2$,

$$\sum_{t=s}^{e} (\widehat{z}(t) - g^0(t))^2 > \sum_{t=s}^{e} (\widehat{z}(t) - \widehat{g}_{b^*}(t))^2.$$

Therefore, if we can show that for some $c_0 \epsilon_T$,

$$\sum_{t=s}^{e} (\widehat{z}(t) - g^0(t))^2 < \sum_{t=s}^{e} (\widehat{z}(t) - \widehat{g}_b(t))^2$$

with $c_0 \epsilon_T < |b - \eta|$, this would imply that $|b^* - \eta| \le c_0 \epsilon_T$. Subtracting the right part and expanding the above,

$$\sum_{t=s}^{e} (g(t) + e(t) - g^0(t))^2 - \sum_{t=s}^{e} (g(t) + e(t) - \widehat{g}_b(t))^2$$

$$= \sum_{t=s}^{e} \{ (g(t) - g^0(t))^2 - (g(t) - \widehat{g}_b(t))^2 \} + 2 \sum_{t=s}^{e} e(t)(\widehat{g}_b(t) - g^0(t)) \equiv I + II$$

$I < 0$ as $g^0(t)$ is defined to minimize the least squares fit to $g(t)$, over all step functions with a single step. We now show that I absolutely exceeds II as $T \to \infty$. Let $\psi \in \Psi$ be those vectors defined on $\{s, \ldots, e\}$ with components that are first constant and positive and then constant and negative, such that these components are zero on average and squared sum up to one. Denoting as $\bar{g}$ the mean of $g(t)$ on $\{s, \ldots, e\}$ and $\psi^0 \in \Psi$ s.t. $g^0(t) = \bar{g} + \langle g, \psi^0 \rangle \psi^0(t)$ with $\langle \cdot \rangle$ denoting the inner

product. Then

$$\sum_{t=s}^{e}(g(t) - g^0(t))^2 \tag{C.8}$$

$$= \sum_{t=s}^{e}(g(t) - \bar{g})^2 - 2\langle g, \psi^0 \rangle \sum_{t=s}^{e}(g(t) - \bar{g})\psi^0(t) + \langle g, \psi^0 \rangle^2 \sum_{t=s}^{e}(\psi^0(t))^2 \tag{C.9}$$

$$= \sum_{t=s}^{e}(g(t) - \bar{g})^2 - \langle g, \psi^0 \rangle^2 \tag{C.10}$$

Let $\tilde{g}(t)$ be a step function that changes at b and minimizes the squared distance to $g(t)$, then

$$\sum_{t=s}^{e}(g(t) - \tilde{g}(t))^2 \leq \sum_{t=s}^{e}(g(t) + \widehat{g}_b(t))^2 \tag{C.11}$$

Using Equations (C.8) and (C.11), for $\tilde{g}(t) = \bar{g} + \langle g, \tilde{\psi} \rangle \tilde{\psi}(t)$, I can be bounded as follows

$$\sum_{t=s}^{e}\{(g(t) - g^0(t))^2 - (g(t) - \widehat{g}_b(t))^2\}$$

$$\leq \sum_{t=s}^{e}\{(g(t) - g^0(t))^2 - (g(t) - \tilde{g}(t))^2\} = \langle g, \psi^0 \rangle^2 - \langle g, \tilde{\psi} \rangle^2$$

$$= (|\langle g, \psi^0 \rangle| - |\langle g, \tilde{\psi} \rangle|)(|\langle g, \psi^0 \rangle| + |\langle g, \tilde{\psi} \rangle|) \leq (|\langle g, \psi^0 \rangle| - |\langle g, \tilde{\psi} \rangle|)|\langle g, \psi^0 \rangle|$$

Note that $|\langle g, \psi^0 \rangle| = |\mathcal{C}_{s,\eta,e}(g)|$ and $|\langle g, \tilde{\psi} \rangle| = |\mathcal{C}_{s,b,e}(g)|$. Thus, with the distance $c_0 \epsilon_T$ between η and b the above is bounded by $-C\delta_T^3 \epsilon_T / T^3$, by Lemmata 1 and 2.

Consider now term II divided by 2,

$$\sum_{t=s}^{e} e(t)(\widehat{g}_b(t) - g^0(t))$$

$$= \sum_{t=s}^{e} e(t)(\tilde{g}_b(t) - g^0(t)) + \sum_{t=s}^{e} e(t)(\widehat{g}_b(t) - \tilde{g}(t)) = II.i + II.ii$$

For each of these sums we consider separately the following three constant intervals: assuming wlog $\eta < b$, $II.i$ can be divided into

$$\sum_{t=s}^{e} e(t)(\tilde{g}_b(t) - g^0(t))$$

$$= \sum_{t=s}^{\eta} e(t)(\tilde{g}_b(t) - g^0(t)) + \sum_{t=\eta+1}^{b} e(t)(\tilde{g}_b(t) - g^0(t)) + \sum_{t=b+1}^{e} e(t)(\tilde{g}_b(t) - g^0(t))$$

$$= II.i.1 + II.i.2 + II.i.3$$

By Lemmata 2 - 4, as $T \to \infty$

$$|II.i.1| = \left| \frac{1}{\sqrt{\eta - s + 1}} \sum_{t=s}^{\eta} e(t) \right| \left| \frac{\sqrt{\eta - s + 1}}{b - s + 1} \sum_{t=s}^{b} g(t) - \frac{1}{\sqrt{\eta - s + 1}} \sum_{t=s}^{\eta} g(t) \right|$$

$$\leq C \log T \frac{\sqrt{\eta - s + 1}\sqrt{e - \eta}}{e - s + 1} \epsilon_T/\delta_T \mathcal{C}_{s,\eta,e}(\bar{z}_l) \leq C' \epsilon_T/\delta_T \log T \mathcal{C}_{s,\eta,e}(\bar{z}_l)$$

Similar bounds can be derived for $|II.i.3|$ (of the same order as the above) and $|II.i.2|$ (of order $C'' \sqrt{\epsilon_T/\delta_T} \log T \mathcal{C}_{s,\eta,e}(\bar{z}_l)$). Furthermore, $II.ii$ can be decomposed as

$$\sum_{t=s}^{e} e(t)(\widehat{g}_b(t) - \tilde{g}(t)) = \sum_{t=s}^{b} e(t)(\widehat{g}_b(t) - \tilde{g}(t)) + \sum_{t=b+1}^{e} e(t)(\widehat{g}_b(t) - \tilde{g}(t))$$

$$= II.ii.1 + II.ii.2$$

II.ii.2 is of same order as *II.ii*.1, and, as $T \to \infty$,

$$|II.ii.1| = \frac{1}{b-s+1} \left(\sum_{t=1}^{b} e(t) \right)^2 = \log^2 T \tag{C.12}$$

Thus for I to absolutely exceed II as $T \to \infty$, we require $\frac{\epsilon_T \delta_T^3}{T^3} > \max \left(\frac{\epsilon_T \log T}{\sqrt{T}}, \sqrt{\epsilon_T \delta_T / T} \log T, \log^2 T \right)$. This implies that $\epsilon_T > T^2 \log^2 T / \delta_T^3$, i.e. if $\epsilon_T = \max(T, \log^{2+\vartheta} T)$ and $\vartheta > 0$ suffice to meet the requirement derived above. Now we turn to the last Lemma needed to show consistent detection of change points for autospectra and cross-coherences, before discussing the consistency of the stopping mechanism of the algorithm when there are no more change points to be detected. For the following, we require $\delta_T^{-5/2} T^{5/2} \log T < \sqrt{\epsilon_T} < \zeta_T < \delta_T / \sqrt{T}$, which implies $\zeta_T = \kappa \log^{1+\varpi} T$ with any $\varpi > \vartheta/2$ and $\kappa > 0$.

Lemma 6 *Under the conditions stated in Equations (C.2) and (C.3) as* $T \to \infty$, $\Pr(\mathcal{A}) \to 0$ *for* $\mathcal{A} = \{|\mathcal{C}_{s,b,e}(\widehat{f}(\omega_l))| < \zeta_T \sigma_{s,e}(\widehat{f}(\omega_l))\}$ *with* $b = \text{argmax}_{t \in \mathcal{D}_{s,e}} |\mathcal{C}_{s,t,e}(\widehat{f}(t, \omega_l))|$.

Proof Define the event $\mathcal{B}_f = \{n^{-1}|\sum_{t=s}^{e} \widehat{f}(t, \omega_l) - (2M)^{-1} \sum_{t=s}^{e} \bar{f}(t, \omega_l)| < \check{f}\}$ with $\check{f} \equiv (4nM)^{-1} \sum_{t=s}^{e} \bar{f}(t, \omega_l)$. Using the Bernstein inequality (c.f. Lemma 3), it can be shown that $\Pr(\mathcal{B}_f) \to 1$ as $T \to \infty$, faster than for Equation (C.6) in Lemma 3. Thus $\Pr(n^{-1} \sum_{t=s}^{e} \widehat{f}(t, \omega_l) \in (\check{f}/2, 3\check{f}/2)) \to 1$ and, since $\Pr(\mathcal{A}) \leq \Pr(\mathcal{A} \cap \mathcal{B}_f) + \Pr(\mathcal{B}_f^c)$, it suffices to show that $\Pr(\mathcal{A} \cap \mathcal{B}_f) \to 0$. From Lemma 5, we know that there exists $1 \leq i^* \leq i_2 - i_1$ s.t. $\eta_{i_1+i^*} \in \mathcal{D}_{s,e}$ and $|\eta_{i_1+i^*} - b| < c_0 \epsilon_T$ for $b = \text{argmax}_{t \in \mathcal{D}_{s,e}} |\mathcal{C}_{s,b,e}(\widehat{f}(\omega_l))|$. Wlog let $\eta_{i_1+i^*} < b$ and denote $\bar{f}_1(\omega_l) \equiv$

$\bar{f}(\eta_{i_1+i^*}, \omega_l) \neq \bar{f}(\eta_{i_1+i^*}+1, \omega_l) \equiv \bar{f}_2(\omega_l)$. Then using Lemma 3 and the assumption of balancedness of Section 5.3.3 in probability

$$
\begin{aligned}
|\mathcal{C}_{s,b,e}(\widehat{f}(\omega_l))| \;\geq\;& |\mathcal{C}_{s,b,e}((2M)^{-1}\bar{f}(\omega_l))| - \log T \\
\geq\;& \sqrt{\frac{(b-s+1)(e-b)}{n}} \\
& \left| \frac{((\eta_{i_1+i^*}-s+1)\bar{f}_1(\omega_l) + (b-(\eta_{i_1+i^*}))\bar{f}_2(\omega_l)}{b-s+1} - \bar{f}_2(\omega_l) \right| - \log T \\
\geq\;& \sqrt{\frac{e-b}{n(b-s+1)}} f_*(\eta_{i_1+i^*}-s+1) - \log T \\
\geq\;& \sqrt{\frac{1-c_*}{nc_*}} f_*(\eta_{i_1+i^*}-s+1) - \log T \\
\geq\;& \frac{C\delta_T}{c^*\sqrt{T}} - \log T > \frac{3\breve{f}\zeta_T}{2}
\end{aligned}
$$

Lemma 7 *For some $C, C' > 0$ let $\{s, e\}$ be such that either*

1. $\exists 1 \leq k \leq N$ *with* $s \leq \eta_i \leq e$ *and* $\min(\eta_i - s + 1, e - \eta_i) \leq C\epsilon_T$ *or*

2. $\exists 1 \leq k \leq N$ *with* $s \leq \eta_i < \eta_{i+1} \leq e$ *and* $\max(\eta_i - s + 1, e - \eta_{+1k}) \leq C'\epsilon_T$

Then for $\mathcal{A}' = \{|\mathcal{C}_{s,b,e}(\widehat{f}(\omega_l))| > \zeta_T\sigma_{s,e}(\widehat{f}(t,\omega_l))\}$ with $b = argmax_{t \in \mathcal{D}_{s,e}} |\mathcal{C}_{s,t,e}(\widehat{f}(t,\omega_l))|$, $\Pr(\mathcal{A}') \to 0$ as $T \to \infty$.

Proof We apply the same strategy as for the Proof of Lemma 6 using the event $\mathcal{B}_f$. We will show below that $\Pr(\mathcal{A} \cap \mathcal{B}_f) \to 0$. Assuming $\eta_i - s + 1 \leq C\epsilon_T$ implies

$b > \eta_i$, so

$$
\begin{aligned}
|\mathcal{C}_{s,b,e}(\widehat{f}(\omega_l))| \quad &\leq \quad |\mathcal{C}_{s,b,e}((2M)^{-1}\bar{f}(\omega_l))| + \log T \\
&\leq \quad \sqrt{\frac{(b-s+1)(e-b)}{n}} \\
&\qquad \left| \frac{((\eta_i - s + 1)\bar{f}_1(\omega_l) + (b - (\eta_i)\bar{f}_2(\omega_l)}{b - s + 1} - \bar{f}_2(\omega_l) \right| + \log T \\
&\leq \quad \sqrt{\frac{e-b}{n(b-s+1)}} 2f^*(\eta_i - s + 1) + \log T \\
&\leq \quad \sqrt{\frac{e-\eta_i}{n(\eta_i - s + 1)}} 2f^*(\eta_i - s + 1) + \log T \\
&\leq \quad 2f^*\sqrt{C\epsilon_T} + \log T < \frac{\breve{f}\zeta_T}{2}
\end{aligned}
$$

The proof of the second case uses the same argument.

Functioning of the FreSpeD method for a single frequency band ω_l
through Binary Segmentation   The algorithm is initiated with $s = 1, e = T$
and assuming there are $N > 0$ sufficiently central change points (by the balanced-
ness assumption in Section 5.3.3) and the conditions of Equations C.2 and C.3 are
met, Lemma 6 holds. Then the method detects a change point within the distance
$c_0\epsilon_T$ from a true change point by Lemma 5. The resulting two subsegments are
such that Lemma 6 continues to hold. The algorithm continues until all change
points are detected, and the segments resulting from the detected N change points
fulfill one of the criteria of Lemma 7, so the algorithm terminates.

C.1.2 Thresholded CUSUM on the Set of Frequency Bands $\omega_l, l = \{1, \dots, L\}$

To ultimately prove the statements in Section 5.3.3 we need to show consistency of the thresholded sum of CUSUM statistics over all L frequency bands. Let $\mathcal{L} = \{1, \dots, L\}$ and $\mathcal{L}_{s,e} \subseteq \mathcal{L}$ the set of frequency bands with at least one change point in $z_l(t), t \in \{s, \dots, e\}$, $s, t, e \in \mathbb{T}$. The following is based on the arguments in Cho and Fryzlewicz (2015a) on thresholded CUSUMs. By Lemmata 3 and 4, respectively, we have $\max_l \max_{\{s,t,e\} \in \mathcal{D}} |\mathcal{C}_{s,t,e}(\widehat{f}(\omega_l)) - \mathcal{C}_{s,t,e}((2M)^{-1}\bar{f}(\omega_l))| \leq \log T$ with probability bounded from below by $1 - (CLT^3 \exp(-C \log^2 T)) \to 1$ and $\max_l \max_{\{s,t,e\} \in \mathcal{D}} |\mathcal{C}_{s,b,e}(\widehat{\rho}^*(\omega_l)) - \mathcal{C}_{s,b,e}(\bar{\rho}^*(\omega_l))| > \sqrt{8M \log T}$ with probability bounded by $1 - (CL/T) \to 1$ as $T \to \infty$. Based on these events, on the one hand by Lemma 7 we know that the absolute CUSUMs of any $l \in \mathcal{L} \setminus \mathcal{L}_{s,e}$ will not exceed the threshold ζ_T on the interval $\{s, \dots, e\}$. On the other hand, by Lemma 6 the CUSUMs of frequency bands containing true change points will exceed the threshold around these change points with high probability.

The following is an observation on the behaviour of the CUSUMs of the true spectral quantities, which are piecewise-constant functions, based on Venkatraman (1992, Lemma 2.2). By Lemmata 3 and 4 the CUSUMs of the estimated quantities are close. Consider a generic piecewise-constant function $h_l(x) = (\alpha_{l,x}x + \beta_{l,x})/\sqrt{(x(1-x))}$ with $x = (t - s + 1)/n \in (0, 1)$, where $\alpha_{l,x}$, $\beta_{l,x}$ depend on the change size and change-point locations and are constant between change points. The scaling of the CUSUM by the locally estimated variability affects the magnitudes of these constants but does not affect the functional form

of $h_l(x)$. These functions are either monotonic or v-shaped between two adjacent change points. They achieve their maximum over $\{s, \ldots, e\}$ at some change point contained in this interval.

The pointwise summation of the CUSUM statistics belonging to set $\mathcal{L}_{s,e}$, $h(x)$, has the same functional form as the individual components $h_l(x)$. We thus have on an interval $\{s, \ldots, e\}$,

$$
\begin{aligned}
\frac{\mathfrak{C}_{s,b,e}(\widehat{z})}{|\mathcal{L}_{s,e}|} &= \frac{\sum_{l=1}^{L} \mathcal{C}_{s,b,e}(\widehat{z}_l)/\sigma_{s,e}(\widehat{z}_l)\mathbb{I}(\mathcal{C}_{s,b,e}(\widehat{z}_l) > \zeta_T)}{|\mathcal{L}_{s,e}|} \\
&= \frac{\sum_{l=1}^{L} \mathcal{C}_{s,b,e}(z_l)/\sigma_{s,e}(\widehat{z}_l)\mathbb{I}(\mathcal{C}_{s,b,e}(\widehat{z}_l) > \zeta_T)}{|\mathcal{L}_{s,e}|} \\
&\quad + \frac{\sum_{l=1}^{L} (\mathcal{C}_{s,b,e}(\widehat{z}_l) - \mathcal{C}_{s,b,e}(z_l))/\sigma_{s,e}(z_l)\mathbb{I}(\mathcal{C}_{s,b,e}(\widehat{z}_l) > \zeta_T)}{|\mathcal{L}_{s,e}|}
\end{aligned}
$$

where $|\mathcal{L}_{s,e}|$ denotes the cardinality of the set $\mathcal{L}_{s,e}$. By Lemma 3 and Lemma 4, respectively, the second part of the summation is bounded by $C \log T$. The first part of the summation can be seen as average of CUSUMs of piecewise-constant functions (from the set $\mathcal{L}_{s,e}$), rescaled by constants. Because thresholding has no effect in the neighbourhood $C\epsilon_T$ around a true change point, by Lemma 5 $b = \text{argmax}_{t \in \{s,e\}} \mathfrak{C}_{s,t,e}(\widehat{z})$ satisfies $|b - \eta_i| < c_0 \epsilon_T$ for some i. The algorithm proceeds iteratively and detects all N change points. It stops when all segments between change points satisfy the conditions of Lemma 7 for all $l = \{1, \ldots, L\}$.

We conclude with an outline of the consistency of our detection mechanism using the extra step of evaluating the thresholded CUSUM statistic on the interval $\mathcal{B}_0$ of length $\Delta_T + 1$ around a change-point candidate b_0, which is shown in the pseudo-code representation of the FreSpeD method in Figure 5.2. Using the arguments of Lemma 2, for any $l \in \mathcal{L}_{s,e}$ there exists $C > 0$ s.t. $\mathcal{C}_{s,t,e}(z_l) > \zeta_T$ for $|t - \eta_i| < C\epsilon_T$

for some change point η_i of z_l on $\{s, \ldots, e\}$. Then $\mathcal{C}_{s,b,e}(\widehat{z}_l) > \zeta_T$ within a distance

$\Delta_T \propto \epsilon_T$ for $b = \operatorname{argmax}_{t \in \{s, \ldots, e\}} \mathcal{C}_{s,t,e}(\widehat{z}_l)$ and thus this extra step is consistent.

C.2   Sensitivity to the Interval Length ν

We summarize the results of the simulation study specified in Section 5.4 below for

varying interval length $\nu = \{60, 100, 400, 600\}$.

D	D^c	Process	$N = 1, \nu = 60$				$N = 1, \nu = 400$			
			$\hat{N} < N$	$\hat{N} = N$	$\hat{N} > N$	MAD	$\hat{N} < N$	$\hat{N} = N$	$\hat{N} > N$	MAD
2	1	AR(2)	0.0	100.0	0.0	55.3	0.0	100.0	0.0	200.8
2	1	AR(6)	0.0	100.0	0.0	42.4	0.0	99.8	0.2	200.0
2	1	AR(10)	0.0	87.4	12.6	46.9	0.0	100.0	0.0	208.0
2	1	ARMA(2,2)	0.0	100.0	0.0	111.4	0.0	100.0	0.0	226.4
2	2	AR(2)	0.0	100.0	0.0	57.7	0.0	100.0	0.0	202.4
2	2	AR(6)	0.0	100.0	0.0	41.5	0.0	99.9	0.1	202.8
2	2	AR(10)	0.0	89.0	11.0	45.1	0.0	99.7	0.3	204.8
2	2	ARMA(2,2)	0.0	100.0	0.0	101.0	0.0	100.0	0.0	228.4
D	D^c	Process	$N = 1, \nu = 100$				$N = 1, \nu = 600$			
			$\hat{N} < N$	$\hat{N} = N$	$\hat{N} > N$	MAD	$\hat{N} < N$	$\hat{N} = N$	$\hat{N} > N$	MAD
2	1	AR(2)	0.0	100.0	0.0	27.8	0.0	99.8	0.2	235.7
2	1	AR(6)	0.0	100.0	0.0	15.8	0.0	100.0	0.0	216.0
2	1	AR(10)	0.0	98.6	1.4	22.9	0.0	99.6	0.4	220.1
2	1	ARMA(2,2)	0.0	100.0	0.0	91.6	0.0	100.0	0.0	260.8
2	2	AR(2)	0.0	100.0	0.0	29.9	0.0	99.9	0.1	232.2
2	2	AR(6)	0.0	100.0	0.0	14.3	0.0	100.0	0.0	219.6
2	2	AR(10)	0.0	98.2	1.8	23.1	0.0	99.8	0.2	214.4
2	2	ARMA(2,2)	0.0	100.0	0.0	77.9	0.0	100.0	0.0	254.8

Table C.1: Simulation results for varying interval length ν with $D = 2$ components and $N = 1$ change points.

D	D^c	Process	$N = 5, \nu = 60$				$N = 5, \nu = 400$			
			$\hat{N} < N$	$\hat{N} = N$	$\hat{N} > N$	MAD	$\hat{N} < N$	$\hat{N} = N$	$\hat{N} > N$	MAD
2	1	AR(2)	0.0	100.0	0.0	115.0	2.6	97.0	0.4	319.1
2	1	AR(6)	0.0	99.6	0.4	70.6	1.2	98.6	0.2	266.9
2	1	AR(10)	0.2	86.0	13.8	122.9	2.2	97.4	0.4	277.9
2	1	ARMA(2,2)	22.8	77.2	0.0	290.9	81.6	18.4	0.0	495.9
2	2	AR(2)	0.0	100.0	0.0	114.4	2.1	97.6	0.3	324.9
2	2	AR(6)	0.0	99.8	0.2	70.8	1.9	97.9	0.2	277.6
2	2	AR(10)	0.1	88.5	11.4	130.2	2.5	97.0	0.5	261.5
2	2	ARMA(2,2)	23.6	76.4	0.0	289.6	80.2	19.8	0.0	510.7
D	D^c	Process	$N = 5, \nu = 100$				$N = 5, \nu = 600$			
			$\hat{N} < N$	$\hat{N} = N$	$\hat{N} > N$	MAD	$\hat{N} < N$	$\hat{N} = N$	$\hat{N} > N$	MAD
2	1	AR(2)	0.0	99.8	0.2	110.0	26.4	73.6	0.0	474.8
2	1	AR(6)	0.0	100.0	0.0	75.1	19.2	80.8	0.0	446.2
2	1	AR(10)	0.2	96.0	3.8	129.6	8.4	91.6	0.0	387.8
2	1	ARMA(2,2)	29.6	70.4	0.0	338.6	93.4	6.6	0.0	601.6
2	2	AR(2)	0.0	99.7	0.3	117.5	25.2	74.8	0.0	484.2
2	2	AR(6)	0.0	99.9	0.1	74.8	17.4	82.6	0.0	457.9
2	2	AR(10)	0.1	95.3	4.6	131.3	9.4	90.5	0.1	378.0
2	2	ARMA(2,2)	31.1	68.9	0.0	333.2	92.8	7.2	0.0	596.1

Table C.2: Simulation results for varying interval length ν with $D = 2$ components and $N = 5$ change points.

D	D^c	Process	$N = 1, \nu = 60$				$N = 1, \nu = 400$			
			$\hat{N} < N$	$\hat{N} = N$	$\hat{N} > N$	MAD	$\hat{N} < N$	$\hat{N} = N$	$\hat{N} > N$	MAD
20	1	AR(2)	0.0	100.0	0.0	55.8	0.0	99.8	0.2	201.6
20	1	AR(6)	0.0	100.0	0.0	45.9	0.0	100.0	0.0	200.0
20	1	AR(10)	0.0	89.6	10.4	46.7	0.0	100.0	0.0	200.8
20	1	ARMA(2,2)	0.0	100.0	0.0	96.3	0.0	100.0	0.0	222.4
20	10	AR(2)	0.0	100.0	0.0	58.3	0.0	100.0	0.0	203.8
20	10	AR(6)	0.0	100.0	0.0	42.4	0.0	100.0	0.0	201.8
20	10	AR(10)	0.0	89.1	10.9	45.6	0.0	99.9	0.1	201.4
20	10	ARMA(2,2)	0.0	100.0	0.0	99.2	0.0	100.0	0.0	221.0
20	20	AR(2)	0.0	100.0	0.0	56.8	0.0	100.0	0.0	204.2
20	20	AR(6)	0.0	100.0	0.0	42.4	0.0	99.9	0.1	202.2
20	20	AR(10)	0.0	89.2	10.8	45.4	0.0	99.9	0.1	202.2
20	20	ARMA(2,2)	0.0	100.0	0.0	99.0	0.0	100.0	0.0	221.4
D	D^c	Process	$N = 1, \nu = 100$				$N = 1, \nu = 600$			
			$\hat{N} < N$	$\hat{N} = N$	$\hat{N} > N$	MAD	$\hat{N} < N$	$\hat{N} = N$	$\hat{N} > N$	MAD
20	1	AR(2)	0.0	100.0	0.0	33.6	0.0	100.0	0.0	227.6
20	1	AR(6)	0.0	100.0	0.0	16.2	0.0	99.8	0.2	219.2
20	1	AR(10)	0.0	97.4	2.6	26.3	0.0	100.0	0.0	210.0
20	1	ARMA(2,2)	0.0	100.0	0.0	73.2	0.0	100.0	0.0	250.8
20	10	AR(2)	0.0	100.0	0.0	32.2	0.0	99.8	0.2	230.4
20	10	AR(6)	0.0	100.0	0.0	15.9	0.0	99.9	0.1	228.4
20	10	AR(10)	0.0	97.1	2.9	24.2	0.0	99.9	0.1	209.5
20	10	ARMA(2,2)	0.0	100.0	0.0	75.1	0.0	100.0	0.0	254.3
20	20	AR(2)	0.0	100.0	0.0	32.1	0.0	99.9	0.1	231.7
20	20	AR(6)	0.0	100.0	0.0	15.5	0.0	99.9	0.1	225.6
20	20	AR(10)	0.0	97.1	3.0	24.1	0.0	99.8	0.2	211.4
20	20	ARMA(2,2)	0.0	100.0	0.0	75.2	0.0	100.0	0.0	253.6

Table C.3: Simulation results for varying interval length ν with $D = 20$ components and $N = 1$ change point.

D	D^c	Process	$N=5, \nu=60$				$N=5, \nu=400$			
			$\hat{N}<N$	$\hat{N}=N$	$\hat{N}>N$	MAD	$\hat{N}<N$	$\hat{N}=N$	$\hat{N}>N$	MAD
20	1	AR(2)	0.0	100.0	0.0	114.6	1.0	99.0	0.0	312.6
20	1	AR(6)	0.0	100.0	0.0	71.6	1.8	97.8	0.4	271.4
20	1	AR(10)	0.0	89.6	10.4	132.0	3.2	96.6	0.2	266.5
20	1	ARMA(2,2)	21.2	78.8	0.0	296.5	76.2	23.8	0.0	523.8
20	10	AR(2)	0.0	100.0	0.0	114.6	2.3	97.7	0.0	315.7
20	10	AR(6)	0.0	99.9	0.1	72.6	1.6	98.2	0.2	269.7
20	10	AR(10)	0.0	89.7	10.3	134.2	2.6	97.3	0.2	256.3
20	10	ARMA(2,2)	23.6	76.4	0.0	301.7	78.9	21.1	0.0	526.8
20	20	AR(2)	0.0	100.0	0.1	114.1	2.7	97.2	0.1	319.0
20	20	AR(6)	0.0	99.8	0.2	72.2	1.7	98.1	0.2	267.6
20	20	AR(10)	0.1	89.7	10.3	137.5	2.5	97.3	0.3	257.5
20	20	ARMA(2,2)	24.0	76.0	0.0	299.6	78.3	21.7	0.0	521.7

D	D^c	Process	$N=5, \nu=100$				$N=5, \nu=600$			
			$\hat{N}<N$	$\hat{N}=N$	$\hat{N}>N$	MAD	$\hat{N}<N$	$\hat{N}=N$	$\hat{N}>N$	MAD
20	1	AR(2)	0.0	100.0	0.0	121.0	27.0	73.0	0.0	451.0
20	1	AR(6)	0.0	99.8	0.2	76.3	17.0	83.0	0.0	442.3
20	1	AR(10)	0.2	93.4	6.4	129.8	11.0	89.0	0.0	382.6
20	1	ARMA(2,2)	30.0	70.0	0.0	330.2	93.0	7.0	0.0	596.5
20	10	AR(2)	0.0	100.0	0.0	122.0	23.2	76.6	0.1	488.9
20	10	AR(6)	0.0	99.9	0.1	78.1	16.4	83.5	0.1	457.0
20	10	AR(10)	0.1	95.2	4.8	133.6	10.3	89.6	0.1	375.7
20	10	ARMA(2,2)	31.2	68.8	0.0	335.1	93.7	6.3	0.0	610.3
20	20	AR(2)	0.0	100.0	0.1	121.1	24.5	75.4	0.1	487.7
20	20	AR(6)	0.0	99.9	0.1	78.1	16.4	83.4	0.1	456.8
20	20	AR(10)	0.2	95.1	4.7	135.5	9.8	90.0	0.2	374.5
20	20	ARMA(2,2)	30.5	69.5	0.0	336.7	93.7	6.3	0.0	617.7

Table C.4: Simulation results for varying interval length ν with $D = 20$ components and $N = 5$ change points.

C.3 Results of the KMO method with lag $p = 2$

D	D^c	Process	$N = 1$				$N = 5$			
			$\hat{N} < N$	$\hat{N} = N$	$\hat{N} > N$	MAD	$\hat{N} < N$	$\hat{N} = N$	$\hat{N} > N$	MAD
2	1	AR(2)	0	92	8	39.7	0	76.6	23.4	70.2
2	1	AR(6)	0	29.4	70.6	470.6	3.6	4.2	92.2	1701.8
2	1	AR(10)	0	0	100	NA	0	0	100	NA
2	1	ARMA(2,2)	0	80.6	19.4	74.1	0	50	50	91.2
2	2	AR(2)	0	91.8	8.2	19.3	0	78.4	21.6	38.3
2	2	AR(6)	0	34.8	65.2	175.7	0	4.6	95.4	225.1
2	2	AR(10)	0	0	100	NA	0	0	100	NA
2	2	ARMA(2,2)	0	80	20	31.2	0	50.2	49.8	47.6
20	1	AR(2)	0	98.6	1.4	42.3	0	90.6	9.4	89.7
20	1	AR(6)	0	0	100	NA	0.4	0	99.6	NA
20	1	AR(10)	0	0	100	NA	0	0	100	NA
20	1	ARMA(2,2)	0	67.6	32.4	60.5	0	33	67	103.6
20	10	AR(2)	0	98.4	1.6	4.4	0	92.2	7.8	9.1
20	10	AR(6)	0	0	100	NA	0	0	100	NA
20	10	AR(10)	0	0	100	NA	0	0	100	NA
20	10	ARMA(2,2)	0	64.6	35.4	6.1	0	30.8	69.2	12.6
20	20	AR(2)	0	96.8	3.2	2.7	0	88.2	11.8	5.2
20	20	AR(6)	0	0.6	99.4	6.3	0	0	100	NA
20	20	AR(10)	0	0	100	NA	0	0	100	NA
20	20	ARMA(2,2)	0	63.2	36.8	3.3	0	28.2	71.8	6.9

Table C.5: Simulation results for the KMO method with lag parameter $p = 2$.

Appendix D

List of Frequently Used Abbreviations

Below we list the most abbreviations that are most frequently referred to in this work.

Expression	Meaning
AR	Autoregressive
ARCH	Autoregressive conditional heteroskedasticity
ARMA	Autoregressive moving-average
BS	Binary segmentation
CUSUM	Cumulative sum
EEG	Electroencephalography
FreSpeD	Frequency-specific change-point detection
IC	Information criterion
LR	Likelihood Ratio
ML	Maximum likelihood
SeqToI	Sequential testing on intervals
UH	Unbalanced Haar

Bibliography

S. Adak. Time-dependent spectral analysis of nonstationary time series. *Journal of the American Statistical Association*, 93(444):1488–1501, 1998.

R. P. Adams and D. J. MacKay. Bayesian online changepoint detection. *ArXiv*, 2007.

H. Adeli, Z. Zhou, and N. Dadmehr. Analysis of EEG records in an epileptic patient using wavelet transform. *Journal of Neuroscience Methods*, 123(1):69–87, 2003.

L. Akoglu, H. Tong, and D. Koutra. Graph based anomaly detection and description: a survey. *Data Mining and Knowledge Discovery*, 29(3):626–688, 2015.

G. Alarcon, C. Binnie, R. Elwes, and C. Polkey. Power spectrum and intracranial EEG patterns at seizure onset in partial epilepsy. *Electroencephalography and Clinical Neurophysiology*, 94(5):326–337, 1995.

E. Andreou and E. Ghysels. Detecting multiple breaks in financial market volatility dynamics. *Journal of Applied Econometrics*, 17(5):579–600, 2002.

A. Ang and A. Timmermann. Regime Changes and Financial Markets. *Annual Review of Financial Economics*, 4(1):313–337, 2012.

J. Antoch and M. Huskova. Permutation tests in change point analysis. *Statistics and Probability Letters*, 53:37–46, 2001.

D. Ardia. Bayesian estimation of a markov-switching threshold asymmetric

GARCH model with student-t innovations. *The Econometrics Journal*, 12(1): 105–126, 2009.

E. Arias-Castro, D. L. Donoho, and X. Huo. Near-optimal detection of geometric objects by fast multiscale methods. *Information Theory, IEEE Transactions on*, 51(7):2402–2425, 2005.

E. Arias-Castro, R. M. Castro, E. Tánczos, and M. Wang. Distribution-free detection of structured anomalies: Permutation and rank-based scans. *ArXiv*, 2015.

S. Arlot and A. Celisse. Segmentation of the mean of heteroscedastic data via cross-validation. *Statistics and Computing*, 21:613–632, 2011.

S. Arlot, A. Celisse, and Z. Harchaoui. Kernel change-point detection. *ArXiv*, 2012.

J. A. Aston and C. Kirch. Evaluating stationarity via change-point alternatives with applications to fMRI data. *The Annals of Applied Statistics*, 6(4):1906–1948, 2012.

A. Aue, S. Hörmann, L. Horváth, and M. Reimherr. Break detection in the covariance structure of multivariate time series models. *Annals of Statistics*, 37(6): 4046–4087, 2009.

C. Baek and V. Pipiras. Long range dependence, unbalanced Haar wavelet transformation and changes in local mean level. *International Journal of Wavelets, Multiresolution and Information Processing*, 7:23–58, 2009.

J. Bai. Estimating multiple breaks one at a time. *Econometric Theory*, 13:315–352, 1997.

J. Bai. Vector autoregressive models with structural changes in regression coefficients and in variance-covariance matrices. *Annals of Economics and Finance*, 1 (2):303–339, 2000.

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

J. Bai and P. Perron. Estimating and testing linear models with multiple structural changes. *Econometrica*, 66(1):47–78, 1998.

J. Bai and P. Perron. Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, 18(1):1–22, 2003.

J.-M. Bardet, W. Kengne, and O. Wintenberger. Multiple breaks detection in general causal time series using penalized quasi-likelihood. *Electronic Journal of Statistics*, 6:435–477, 2012.

L. Bardwell and P. Fearnhead. Bayesian detection of abnormal segments in multiple time series. *ArXiv*, 2014.

M. Baron, J. Brogaard, and A. Kirilenko. The trading profits of high frequency traders. Technical report, Foster School of Business, University of Washington, 2012.

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

D. Barry and J. A. Hartigan. Product partition models for change point problems. *Annals of Statistics*, 20(1):260–279, 1992.

M. Basseville. Detecting changes in signals and systems - A survey. *Automatica*, 24:309–326, 1988.

M. Basseville and I. V. Nikiforov. *Detection of abrupt changes: theory and application*, volume 104. Prentice Hall Englewood Cliffs, 1993.

F. Battaglia and M. Protopapas. Time-varying multi-regime models fitting by genetic algorithms. *Journal of Time Series Analysis*, 32(3):237–252, 2011.

F. Battaglia and M. Protopapas. Multiregime models for nonlinear nonstationary time series. *Computational Statistics*, 27:319–341, 2012.

M. J. Bayarri and J. O. Berger. The interplay of Bayesian and frequentist analysis. *Statistical Science*, pages 58–80, 2004.

P. Bélisle, L. Joseph, B. MacGibbon, D. B. Wolfson, and R. Du Berger. Change-point analysis of neuron spike train data. *Biometrics*, pages 113–123, 1998.

I. Berkes, L. Horvath, and P. Kokoszka. Testing for parameter constancy in GARCH models. *Statistics & Probability Letters*, 70(4):263 – 273, 2004.

J. Berkowitz and J. O'Brien. How accurate are value-at-risk models at commercial banks? *Journal of Finance*, 57:1093–1111, 2002.

P. Bhattacharya and H. Zhou. A rank-cusum procedure for detecting small changes in a symmetric distribution. *Lecture Notes-Monograph Series*, pages 57–65, 1994.

G. K. Bhattacharyya and R. A. Johnson. Nonparametric tests for shift at an unknown time point. *The Annals of Mathematical Statistics*, 39(5):1731–1743, 1968.

S. Birr, S. Volgushev, T. Kley, H. Dette, and M. Hallin. Quantile spectral analysis for locally stationary time series. *ECARES Working Papers*, 2015.

K. Bleakley and J.-P. Vert. The group fused lasso for multiple change-point detection. *ArXiv*, 2011.

N. Blondin and D. Greer. Neurologic prognosis in cardiac arrest patients treated with therapeutic hypothermia. *The Neurologist*, 17(5):241–248, 2011.

N. Booth and A. Smith. A Bayesian approach to retrospective identification of change-points. *Journal of Econometrics*, 19(1):7–22, 1982.

D. Bosq. Nonparametric statistics for stochastic processes: Estimation and prediction. 1998.

S. Bouzebda and A. Keziou. A semiparametric maximum likelihood ratio test for the change point in copula models. *Statistical Methodology*, 14:39–61, 2013.

J. V. Braun and H.-G. Muller. Statistical methods for DNA sequence segmentation. *Statistical Science*, pages 142–162, 1998.

J. V. Braun, R. Braun, and H.-G. Müller. Multiple changepoint fitting via quasi-likelihood, with application to DNA sequence segmentation. *Biometrika*, 87(2): 301–314, 2000.

P. Brockwell and R. Davis. Time series theory and methods. 1991.

E. Brodsky and B. Darkhovsky. *Nonparametric Methods in Change Point Problems*, volume 243. Springer Science & Business Media, 1993.

E. Brodsky and B. S. Darkhovsky. *Nonparametric methods in change point problems*, volume 243. Springer Science & Business Media, 2013.

L. D. Broemeling. Bayesian procedures for detecting a change in a sequence of random variables. *Metron*, 30:1–14, 1972.

L. D. Broemeling. Bayesian inferences about a changing sequence of random variables. *Communications in Statistics*, 3(3):243–255, 1974.

J. Brogaard. High frequency trading and its impact on market quality. *Northwestern University Kellogg School of Management Working Paper 66*, 2010.

D. H. Brooks, H. Krim, J.-C. Pesquet, and R. S. MacLeod. Best basis segmentation of ecg signals using novel optimality criteria. In *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, volume 5, pages 2750–2753, 1996.

B. Bruder, T.-L. Dao, J.-C. Richard, and T. Roncalli. Trend filtering methods for momentum strategies. Technical report, Lyxor Asset Management, 2008.

A. Buraschi, F. Trojani, and A. Vedolin. When uncertainty blows in the orchard: Comovement and equilibrium volatility risk premia. *The Journal of Finance*, 69 (1):101–137, 2014.

L. E. Calvet, A. J. Fisher, and S. B. Thompson. Volatility comovement: a multi-frequency approach. *Journal of Econometrics*, 131(1):179–215, 2006.

F. Camci and R. B. Chinnam. General support vector representation machine for one-class classification of non-stationary classes. *Pattern Recognition*, 41(10): 3021–3034, 2008.

B. P. Carlin, A. E. Gelfand, and A. F. Smith. Hierarchical Bayesian analysis of changepoint problems. *Applied statistics*, pages 389–405, 1992.

F. Caron, A. Doucet, and R. Gottardo. On-line changepoint detection and parameter estimation with application to genomic data. *Statistics and Computing*, 22 (2):579–595, 2012.

I. Casas and I. Gijbels. Unstable volatility functions: the break preserving local linear estimator. CREATES Research Papers 2009-48, School of Economics and Management, University of Aarhus, 2009.

J. P. S. Catalão, S. J. P. S. Mariano, V. M. F. Mendes, and L. A. F. M. Ferreira. Short-term electricity prices forecasting in a competitive market: a neural network approach. *Electric Power Systems Research*, 77:1297–1304, 2007.

S. Chakraborti, D. L. Van, and S. T. Bakir. Nonparametric control charts: An overview and some results. *Journal of Quality Technology*, 33(3):304–315, 2001.

H. P. Chan and G. Walther. Detection with the scan and the average likelihood ratio. *Statistica Sinica*, 23:409–428, 2013.

C. W. S. Chen, J. S. K. Chan, M. K. P. So, and K. K. M. Lee. Classification in segmented regression problems. *Computational Statistical Data Analysis*, 55(7): 2276–2287, 2011.

J. Chen and A. Gupta. Likelihood procedure for testing change point hypothesis for multivariate Gaussian model. *Random Operators and Stochastic Equations*, 3:235–244, 1995.

J. Chen and A. Gupta. Testing and locating variance changepoints with application to stock prices. *Journal of the American Statistical Association*, 92(438):739–747, 1997.

J. Chen and A. Gupta. Change point analysis of a Gaussian model. *Statistical Papers*, 40(3):323–333, 1999.

J. Chen and A. Gupta. Statistical inference of covariance change points in Gaussian model. *Statistics*, 38(1):17–28, 2004.

J. Chen and K. Gupta. *Parametric Statistical Change Point Analysis: With Applications to Genetics, Medicine, and Finance*. Boston: Birkhäuser Boston, 2012.

J. Chen and Y.-P. Wang. A statistical change point model approach for the detection of DNA copy number variations in array CGH data. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 6(4):529–541, 2009.

S. Chen and P. Gopalakrishnan. Speaker, environment and channel change detection and clustering via the Bayesian information criterion. In *Proceedings DARPA Broadcast News Transcription and Understanding Workshop*, volume 8, 1998.

H. Chernoff and S. Zacks. Estimating the current mean of a normal distribution which is subjected to changes in time. *The Annals of Mathematical Statistics*, 35(3):999–1018, 1964.

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

H. Cho and P. Fryzlewicz. Multiscale interpretation of taut string estimation and its connection to Unbalanced Haar wavelets. *Statistics and Computing*, 21(4): 671–681, 2011.

H. Cho and P. Fryzlewicz. Multiscale and multilevel technique for consistent segmentation of nonstationary time series. *Statistica Sinica*, 22:207–229, 2012.

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

H. Cho and P. Fryzlewicz. Corrections on multiple-change-point detection for high dimensional time series via sparsified binary segmentation. 2015b.

H. Choi, H. Ombao, and B. Ray. Sequential change-point detection methods for nonstationary time series. *Technometrics*, 50(1):40–52, 2008.

N. Chopin. Dynamic detection of change points in long time series. *Annals of the Institute of Statistical Mathematics*, 59(2):349–366, 2007.

P. F. Christoffersen and F. X. Diebold. Financial asset returns, direction-of-change forecasting, and volatility dynamics. *Management Science*, 52(8):1273–1287, 2006.

R. R. Coifman and M. V. Wickerhauser. Entropy-based algorithms for best basis selection. *Information Theory, IEEE Transactions on*, 38(2):713–718, 1992.

R. Cont. Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*, 1(2):223–236, 2001.

I. Cribben, R. Haraldsdottir, L. Atlas, T. Wager, and M. Lindquist. Dynamic connectivity regression: determining state-related changes in brain connectivity. *Neuroimage*, 61(4):907–920, 2012.

I. Cribben, T. Wager, and M. Lindquist. Detecting functional connectivity change points for single-subject fMRI data. *Frontiers in Computational Neuroscience*, 7, 2013.

M. Csörgö and L. Horváth. *Limit Theorems in Change-Point Analysis*. Wiley, 1997.

R. Dahlhaus. Locally stationary processes. *Handbook of Statistics, Time Series Analysis: Methods and Applications*, pages 351–408, 2012.

P. Davies and A. Kovac. Local extremes, runs, strings and multiresolution. *Annals of Statistics*, pages 1–48, 2001.

R. Davis, T. Lee, and G. Rodriguez-Yam. Structural break estimation for nonstationary time series models. *Journal of the American Statistical Association*, 101: 223–239, 2006.

R. A. Davis, T. Lee, and G. A. Rodriguez-Yam. Break detection for a class of nonlinear time series models. *Journal of Time Series Analysis*, 29(5):834–867, 2008.

W. W. Davis. Robust methods for detection of shifts of the innovation variance of a time series. *Technometrics*, 21(3):313–320, 1979.

B. M. de Castro and F. Leonardi. A model selection approach for multiple sequence segmentation and dimensionality reduction. *ArXiv*, 2015.

V. Delouille, J. Franke, and R. von Sachs. Nonparametric stochastic regression with design-adapted wavelets. *Sankhyā: The Indian Journal of Statistics Series A*, 63:328–366, 2001.

J. Deshayes and D. Picard. *Testing for a change-point in statistical models*. Université de Paris-Sud. Département de Mathématique, 1980.

J. Deshayes and D. Picard. Off-line statistical analysis of change-point models using non parametric and likelihood methods. In M. Basseville and A. Benveniste, editors, *Detection of Abrupt Changes in Signals and Dynamical Systems*, volume 77 of *Lecture Notes in Control and Information Sciences*, pages 103–168. Springer Berlin Heidelberg, 1986.

F. X. Diebold and G. D. Rudebusch. Measuring business cycles: A modern perspective. *The Review of Economics and Statistics*, pages 67–77, 1996.

J. Fan, Y. Fan, and J. Lv. High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147(1):186–197, 2008.

P. Fearnhead. Exact and efficient Bayesian inference for multiple changepoint problems. *Statistics and computing*, 16(2):203–213, 2006.

M. Fiecas and H. Ombao. Modeling the evolution of dynamic brain processes during an associative learning experiment. *Journal of the American Statistical Association*, (just-accepted), 2016.

M. Fiecas, H. Ombao, C. Linkletter, W. Thompson, and J. Sanes. Functional connectivity: Shrinkage estimation and randomization test. *Neuroimage*, 49(4): 3005–3014, 2010.

S. Fortunato. Community detection in graphs. *Physics reports*, 486(3):75–174, 2010.

J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

P. Friedman. A change point detection method for elimination of industrial interference in radio astronomy receivers. In *Statistical Signal and Array Processing, 1996. Proceedings., 8th IEEE Signal Processing Workshop on*, pages 264–266, 1996.

P. Fryzlewicz. Unbalanced Haar technique for nonparametric function estimation. *Journal of the American Statistical Association*, 102:1318–1327, 2007.

P. Fryzlewicz. Time–threshold maps: Using information from wavelet reconstructions with all threshold values simultaneously. *Journal of the Korean Statistical Society*, 41(2):145–159, 2012.

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

P. Fryzlewicz and S. Subba Rao. Multiple-change-point detection for autoregressive conditional heteroscedastic processes. *Journal of the Royal Statistical Society: series B (statistical methodology)*, 76(5):903–924, 2014.

P. Fryzlewicz, T. Sapatinas, and S. S. Rao. A Haar-Fisz technique for locally stationary volatility estimation. *Biometrika*, 93:687–704, 2006.

R. C. Garcia, J. Contreras, M. van Akkeren, and J. B. C. Garcia. A GARCH

forecasting model to predict day-ahead electricity prices. *IEEE Transactions on Power Systems*, 20:867–874, 2005.

R. Gençay, F. Selçuk, and B. J. Whitcher. *An Introduction to Wavelets and Other Filtering Methods in Finance and Economics.* Academic Press, 2001.

L. Giraitis and R. Leipus. Testing and estimating in the change-point problem of the spectral function. *Lithuanian Mathematical Journal*, 32(1):15–29, 1992.

L. Giraitis, R. Leipus, and D. Surgailis. The change-point problem for dependent observations. *Journal of Statistical Planning and Inference*, 53(3):297 – 310, 1996.

M. Girardi and W. Sweldens. A new class of unbalanced Haar wavelets that form an unconditional basis for L_p on general measure space. *Journal of Fourier Analysis and Applications*, 7:457–474, 1997.

L. R. Glosten, R. Jagannathan, and D. E. Runkle. On the relation between the expected value and the volatility of the nominal excess return on stocks. *The Journal of finance*, 48(5):1779–1801, 1993.

E. Gombay. Change detection in autoregressive time series. *Journal of Multivariate Analysis*, 99(3):451–464, 2008.

E. Gombay and L. Horváth. On the rate of approximations for maximum likelihood tests in change-point models. *Journal of Multivariate Analysis*, 56(1):120–152, 1996.

C. Gorrostieta, H. Ombao, R. Prado, S. Patel, and E. Eskandar. Exploring de-

pendence between brain signals in a monkey during learning. *Journal of Time Series Analysis*, 33(5):771–778, 2012.

S. F. Gray. Modeling the conditional distribution of interest rates as a regime-switching process. *Journal of Financial Economics*, 42(1):27–62, 1996.

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

J. Groen, G. Kapetanios, and S. Price. Multivariate methods for monitoring structural change. *Journal of Applied Econometrics*, 28(2):250–274, 2013.

M. Guidolin and S. Hyde. Can VAR models capture regime shifts in asset returns? A long-horizon strategic asset allocation perspective. *Journal of Banking & Finance*, 36(3):695–716, 2012.

A. Gupta and J. Chen. Detecting changes of mean in multidimensional normal sequences with applications to literature and geology. *Computational Statistics*, 11(3):211–221, 1996.

I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh. *Feature extraction: foundations and applications*, volume 207. Springer, 2008.

J. D. Hamilton. Analysis of time series subject to changes in regime. *Journal of Econometrics*, 45(1):39–70, 1990.

J. D. Hamilton and B. Raj. *Advances in Markov-Switching Models: Applications in Business Cycle Research and Finance*. Springer Science & Business Media, 2013.

Z. Harchaoui and C. Levy-Leduc. Multiple change-point estimation with a total variation penalty. *Journal of the American Statistical Association*, 105(492): 1480–1493, 2010.

Z. Harchaoui, E. Moulines, and F. R. Bach. Kernel change-point analysis. In *Advances in Neural Information Processing Systems*, pages 609–616, 2009.

M. R. Hardy. A regime-switching model of long-term stock returns. *North American Actuarial Journal*, 5(2):41–53, 2001.

J. Hasbrouck and G. Sofianos. The trades of market makers: An empirical analysis of NYSE specialists. *Journal of Finance*, 48:1565–1593, 1993.

R. Haugen, E. Talmor, and W. Torous. The effect of volatility changes on the level of stock prices and subsequent expected returns. *The Journal of Finance*, 46(3): 985–1007, 1991.

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

K. Haynes, I. A. Eckley, and P. Fearnhead. Computationally efficient change-point detection for a range of penalties. *Journal of Computational and Graphical Statistics*, (just-accepted), 2015.

T. Hocking, G. Rigaill, J.-P. Vert, and F. Bach. Learning sparse penalties for change-point detection using max margin interval regression. In *Proceedings of The 30th International Conference on Machine Learning*, pages 172–180, 2013.

L. Horváth. The maximum likelihood method for testing changes in the parameters of normal observations. *Annals of Statistics*, pages 671–680, 1993.

L. Horváth and M. Hušková. Change-point detection in panel data. *Journal of Time Series Analysis*, 33(4):631–648, 2012.

L. Horváth, P. Kokoszka, and J. Steinebach. Testing for changes in multivariate dependent observations with an application to temperature changes. *Journal of Multivariate Analysis*, 68(1):96 – 119, 1999.

D. A. Hsu. Tests for variance shift at an unknown time point. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 26(3):279–284, 1977.

B. Hu, T. Rakthanmanon, Y. Hao, S. Evans, S. Lonardi, and E. Keogh. Discovering the intrinsic cardinality and dimensionality of time series using MDL. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, pages 1086–1091, 2011.

M. Hušková, Z. Prášková, and J. Steinebach. On the detection of changes in autoregressive time series i. asymptotics. *Journal of Statistical Planning and Inference*, 137(4):1243–1259, 2007.

M. Hušková, C. Kirch, Z. Prášková, and J. Steinebach. On the detection of changes in autoregressive time series, ii. resampling procedures. *Journal of Statistical Planning and Inference*, 138(6):1697–1721, 2008.

C. Inclan. Detection of multiple changes of variance using posterior odds. *Journal of Business & Economic Statistics*, 11(3):289–300, 1993.

C. Inclan and G. C. Tiao. Use of cumulative sums of squares for retrospective detection of changes of variance. *Journal of the American Statistical Association*, 89(427):913–923, 1994.

A. Inoue. Testing for distributional change in time series. *Econometric Theory*, 17 (1):156–187, 2001.

B. Jackson, J. D. Scargle, D. Barnes, S. Arabhi, A. Alt, P. Gioumousis, E. Gwin, P. Sangtrakulcharoen, L. Tan, and T. T. Tsai. An algorithm for optimal partitioning of data on an interval. *Signal Processing Letters, IEEE*, 12(2):105–108, 2005.

B. Jandhyala and S. B. Fotopoulos. Capturing the distributional behaviour of the maximum likelihood estimator of a changepoint. *Biometrika*, 86(1):129–140, 1999.

V. Jandhyala, S. Fotopoulos, I. MacNeill, and P. Liu. Inference for single and multiple change-points in time series. *Journal of Time Series Analysis*, (just-accepted), 2013.

V. K. Jandhyala, S. B. Fotopoulos, and D. M. Hawkins. Detection and estimation of abrupt changes in the variability of a process. *Computational statistics & data analysis*, 40(1):1–19, 2002.

P. Jiruska, J. Csicsvari, A. Powell, J. Fox, W.-C. Chang, M. Vreugdenhil, X. Li, M. Palus, A. Bujan, R. Dearden, and J. Jefferys. High-frequency network activity, global increase in neuronal activity, and synchrony expansion precede epileptic seizures in vitro. *The Journal of Neuroscience*, 30(16):5690–5701, 2010.

A. Johansen and D. Sornette. Modeling the stock market prior to large crashes. *The European Physical Journal B-Condensed Matter and Complex Systems*, 9 (1):167–174, 1999.

A. Johansen and D. Sornette. Evaluation of the quantitative prediction of a trend reversal on the japanese stock market in 1999. *International Journal of Modern Physics C*, 11(02):359–364, 2000.

N. Johnson and S. Kotz. *Distributions in Statistics: Continuous Univariate Distributions: Vol.: 1.* Houghton Mifflin, 1970.

K. Jong, E. Marchiori, A. Van Der Vaart, B. Ylstra, M. Weiss, and G. Meijer. Chromosomal breakpoint detection in human cancer. In *Applications of Evolutionary Computing*, pages 54–65. Springer, 2003.

T. Kanamori, T. Suzuki, and M. Sugiyama. Theoretical analysis of density ratio estimation. *IEICE transactions on fundamentals of electronics, communications and computer sciences*, 93(4):787–798, 2010.

A. Kaplan, J. Röschke, B. Darkhovsky, and J. Fell. Macrostructural EEG characterization based on nonparametric change point segmentation: application to sleep analysis. *Journal of Neuroscience Methods*, 106(1):81–90, 2001.

J. O. Katz and D. McCormick. *The Encyclopedia of Trading Strategies.* McGraw-Hill New York, 2000.

Y. Kawahara and M. Sugiyama. Change-point detection in time-series data by direct density-ratio estimation. In *SDM*, volume 9, pages 389–400. SIAM, 2009.

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

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

H.-J. Kim and K.-S. Shin. A hybrid approach based on neural networks and genetic algorithms for detecting temporal patterns in stock markets. *Applied Soft Computing*, 7:569–576, 2007.

H.-J. Kim and D. Siegmund. The likelihood ratio test for a change-point in simple linear regression. *Biometrika*, 76(3):409–423, 1989.

K.-J. Kim. Financial time series forecasting using support vector machines. *Neurocomputing*, 55:307–319, 2003.

S. Kim and A. L. N. Reddy. Statistical techniques for detecting traffic anomalies through packet header data. *IEEE: Transactions on Networking*, 16(3):562–575, 2008.

S.-J. Kim, K. Koh, S. Boyd, and D. Gorinevsky. L1 trend filtering. *SIAM review*, 51(2):339–360, 2009.

C. Kirch. Block permutation principles for the change analysis of dependent data. *Journal of Statistical Planning and Inference*, 137(7):2453 – 2474, 2007.

C. Kirch, B. Muhsal, and H. Ombao. Detection of changes in multivariate time series with application to EEG data. *Journal of the American Statistical Association*, 110(511):1197–1216, 2015.

S. I. Ko, T. T. Chong, and P. Ghosh. Dirichlet process hidden markov multiple change-point model. *Bayesian Analysis*, 10(2):275–296, 2015.

P. Kokoszka and R. Leipus. Testing for parameter changes in ARCH models. *Lithuanian Mathematical Journal*, 39(2):182–195, 1999.

T. Kumar, V. Kanhangad, and R. Pachori. Classification of seizure and seizure-free EEG signals using multi-level local patterns. In *Digital Signal Processing, IEEE 2014 19th International Conference on*, pages 646–650, 2014.

D. Kwon, K. Ko, M. Vannucci, A. Reddy, and S. Kim. Wavelet methods for the detection of anomalies and their application to network traffic analysis. *Quality and Reliability Engineering International*, 22(8):953–969, 2006.

M. Lavielle and E. Lebarbier. An application of mcmc methods for the multiple change-points problem. *Signal Processing*, 81(1):39–53, 2001.

M. Lavielle and C. Ludeña. The multiple change-points problem for the spectral distribution. *Bernoulli*, 6(5):845–869, 2000.

M. Lavielle and E. Moulines. Least-squares estimation of an unknown number of shifts in a time series. *Journal of Time Series Analysis*, 21:33–59, 2000.

M. Lavielle and G. Teyssiere. Detection of multiple change-points in multivariate time series. *Lithuanian Mathematical Journal*, 46(3):287–306, 2006.

M. Lavielle and G. Teyssiere. Adaptive detection of multiple change-points in asset price volatility. In *Long memory in economics*, pages 129–156. Springer, 2007.

S. Lee, O. Na, and S. Na. On the cusum of squares test for variance change in nonstationary and nonparametric time series models. *Annals of the Institute of Statistical Mathematics*, 55(3):467–485, 2003.

T.-W. Lee and M. S. Lewicki. Unsupervised image classification, segmentation, and enhancement using ica mixture models. *Image Processing, IEEE Transactions on*, 11(3):270–279, 2002.

G. Leitch and J. E. Tanner. Economic forecast evaluation: profits versus the conventional error measures. *American Economic Review*, 81:580–590, 1991.

B. Levin and J. Kline. The cusum test of homogeneity with an application in spontaneous abortion epidemiology. *Statistics in Medicine*, 4(4):469–488, 1985.

S. Liu, M. Yamada, N. Collier, and M. Sugiyama. Change-point detection in time-series data by relative density-ratio estimation. *Neural Networks*, 43:72–83, 2013.

X. Liu, D. Li, S. Wang, and Z. Tao. Effective algorithm for detecting community structure in complex networks based on GA and clustering. In *Computational Science–ICCS 2007*, pages 657–664. Springer, 2007.

R. H. Loschi, J. G. Pontel, and F. R. Cruz. Multiple change-point analysis for linear regression models. *Chilean Journal of Statistics*, 1(2):93–112, 2010.

A. Lung-Yut-Fong, C. Lévy-Leduc, and O. Cappé. Homogeneity and change-point detection tests for multivariate data using rank statistics. *Journal de la Société Française de Statistique*, 156(4):133–162, 2015.

E. M. Maboudou-Tchao and D. M. Hawkins. Detection of multiple change-points in multivariate data. *Journal of Applied Statistics*, 40(9):1979–1995, 2013.

R. Maidstone, T. Hocking, G. Rigaill, and P. Fearnhead. On optimal multiple changepoint algorithms for large data. *Statistics and Computing*, pages 1–15, 2016.

D. Marangoni-Simonsen and Y. Xie. Sequential changepoint approach for online community detection. *Signal Processing Letters, IEEE*, 22(8):1035–1039, 2015.

D. S. Matteson and N. A. James. A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association*, 109(505):334–345, 2014.

C. McGilchrist and K. Woodyer. Note on a distribution-free cusum technique. *Technometrics*, 17(3):321–325, 1975.

G. Motta and H. Ombao. Evolutionary factor analysis of replicated time series. *Biometrics*, 68(3):825–836, 2012.

G. Nason. Stationary and non-stationary times series. *Statistics in Volcanology. Special Publications of IAVCEI*, 2006.

G. P. Nason. *Wavelet Methods in Statistics with R.* Springer, 2008.

G. P. Nason, R. von Sachs, and G. Kroisandt. Wavelet processes and adaptive estimation of the evolutionary wavelet spectrum. *Journal of the Royal Statistical Society Series B*, 62:271–292, 2000.

P. Naveau, A. Guillou, and T. Rietsch. A non-parametric entropy-based approach to detect changes in climate extremes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(5):861–884, 2014.

NBER. US business cycle expansions and contractions. National Bureau of Economic Research, 2013. URL `http://www.nber.org/cycles/`.

W. Ning, J. Pailden, and A. Gupta. Empirical likelihood ratio test for the epidemic change model. *Journal of Data Science*, 10:107–127, 2012.

L. C. Nunes, C.-M. Kuan, and P. Newbold. Spurious break. *Econometric Theory*, 11(4):736–749, 1995.

Y. Ogata. Detection of anomalous seismicity as a stress change sensor. *Journal of Geophysical Research: Solid Earth (1978–2012)*, 110(B5), 2005.

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

H. Ombao and S. van Bellegem. Evolutionary coherence of nonstationary signals. *Signal Processing, IEEE Transactions on*, 56(6):2259–2266, 2008.

H. Ombao, J. Raz, R. von Sachs, and W. Guo. The SLEX model of a non-stationary random process. *Annals of the Institute of Statistical Mathematics*, 54:171–200, 2002.

H. Ombao, R. von Sachs, and W. Guo. SLEX analysis of multivariate nonstationary time series. *Journal of the American Statistical Association*, 100(470):519–531, 2005.

E. S. Page. A test for a change in a parameter occurring at an unknown point. *Biometrika*, 42(3/4):523–527, 1955.

N. R. Pal and S. K. Pal. A review on image segmentation techniques. *Pattern Recognition*, 26(9):1277–1294, 1993.

J. Pan and J. Chen. Application of modified information criterion to multiple change point problems. *Journal of Multivariate Analysis*, 97(10):2221–2241, 2006.

S. Parey, T. T. H. Hoang, and D. Dacunha-Castelle. The importance of mean and variance in predicting changes in temperature extremes. *Journal of Geophysical Research: Atmospheres*, 118(15):8285–8296, 2013.

L. Peel and A. Clauset. Detecting change points in the large-scale structure of evolving networks. *ArXiv*, 2014.

D. Pelletier. Regime switching for dynamic correlations. *Journal of Econometrics*, 131(1):445–473, 2006.

L. Perreault, M. Haché, M. Slivitzky, and B. Bobée. Detection of changes in precipitation and runoff over eastern Canada and US using a Bayesian approach. *Stochastic Environmental Research and Risk Assessment*, 13(3):201–216, 1999.

L. Perreault, J. Bernier, B. Bobée, and E. Parent. Bayesian change-point analysis in hydrometeorological time series. part 1. the normal model revisited. *Journal of Hydrology*, 235(3):221–241, 2000a.

L. Perreault, J. Bernier, B. Bobée, and E. Parent. Bayesian change-point analysis in hydrometeorological time series. part 2. comparison of change-point models and forecasting. *Journal of Hydrology*, 235(3):242–263, 2000b.

L. Perreault, E. Parent, J. Bernier, B. Bobee, and M. Slivitzky. Retrospective multivariate Bayesian change-point analysis: a simultaneous single change in the mean of several hydrological sequences. *Stochastic Environmental Research and Risk Assessment*, 14(4-5):243–261, 2000c.

P. Perron. Dealing with structural breaks. *Palgrave handbook of econometrics*, 1: 278–352, 2006.

P. Perron and J. Zhou. Testing jointly for structural changes in the error variance and coefficients of a linear regression model. Technical report, Boston University-Department of Economics, 2008.

M. H. Pesaran and A. Timmermann. Market timing and return prediction under model instability. *Journal of Empirical Finance*, 9(5):495–510, 2002.

A. N. Pettitt. A non-parametric approach to the change-point problem. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(2):126, 1979.

D. Picard. Testing and estimating change-points in time series. *Advances in Applied Probability*, 17(4):841–867, 1985.

F. Picard, S. Robin, M. Lavielle, C. Vaisse, and J.-J. Daudin. A statistical approach for array CGH data analysis. *BMC bioinformatics*, 6(1):27, 2005.

J.-Y. Pitarakis. Least squares estimation and tests of breaks in mean and variance under misspecification. *The Econometrics Journal*, 7(1):32–54, 2004.

P. Prandom, M. Goodwin, and M. Vetterli. Optimal time segmentation for signal modeling and compression. *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, 3:2029–2032, 1997.

P. Preuß, R. Puchstein, and H. Dette. Detection of multiple structural breaks in multivariate time series. *Journal of the American Statistical Association*, 110 (510):654–668, 2015.

M. B. Priestley. Evolutionary spectra and non-stationary processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, 27(2):204–237, 1965.

M. B. Priestley. *Spectral Analysis and Time Series*. Academic Press, 1983.

M. B. Priestley. *Non-Linear and Non-Stationary Time Series*. Academic Pr, 1988. ISBN 012564910X.

L. Qin and Y. Wang. Nonparametric spectral analysis with applications to seizure characterization using EEG time series. *The Annals of Applied Statistics*, pages 1432–1451, 2008.

R. E. Quandt. The estimation of the parameters of a linear regression system obeying two separate regimes. *Journal of the American Statistical Association*, 53(284):873–880, 1958.

R. E. Quandt. Tests of the hypothesis that a linear regression system obeys two separate regimes. *Journal of the American Statistical Association*, 55(290):324–330, 1960.

F. A. Quintana and P. L. Iglesias. Bayesian clustering and product partition models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65 (2):557–574, 2003.

R. Rabemananjara and J.-M. Zakoian. Threshold arch models and asymmetries in volatility. *Journal of Applied Econometrics*, 8(1):31–49, 1993.

M. Raimondo and N. Tajvidi. A peaks over threshold model for change-point detection by wavelets. *Statistica Sinica*, 14(2):395–412, 2004.

S. Ramgopal, S. Thome-Souza, M. Jackson, N. E. Kadish, I. S. Fernández, J. Klehm, W. Bosl, C. Reinsberger, S. Schachter, and T. Loddenkemper.

Seizure detection, seizure prediction, and closed-loop warning systems in epilepsy. *Epilepsy & Behavior*, 37:291–307, 2014.

B. D. Rao and K. Kreutz-Delgado. An affine scaling methodology for best basis selection. *Signal Processing, IEEE Transactions on*, 47(1):187–200, 1999.

J. Reeves, J. Chen, X. L. Wang, R. Lund, and L. QiQi. A review and comparison of changepoint detection techniques for climate data. *Journal of Applied Meteorology and Climatology*, 46(6):900–915, 2007.

G. Rigaill. A pruned dynamic programming algorithm to recover the best segmentations with 1 to K change-points. *Journal de la Société Française de Statistique*, 156(4), 2015.

J. Rissanen. *Stochastic complexity in statistical inquiry*. World scientific, 1989.

M. W. Robbins, C. M. Gallagher, and R. B. Lund. A general regression changepoint test for time series data. *Journal of the American Statistical Association*, (just-accepted), 2015.

N. Rohan and T. Ramanathan. Asymmetric volatility models with structural breaks. *Communications in Statistics-Simulation and Computation*, 41(9):1519–1543, 2012.

O. Rosen, S. Wood, and D. S. Stoffer. Adaptspec: Adaptive spectral estimation for nonstationary time series. *Journal of the American Statistical Association*, 107(500):1575–1589, 2012.

M. Saab and J. Gotman. A system to detect the onset of epileptic seizures in scalp EEG. *Clinical Neurophysiology*, 116(2):427–442, 2005.

A. Sansó, V. Aragó, and J. L. Carrion. Testing for changes in the unconditional variance of financial time series. DEA Working Papers 5, Universitat de les Illes Balears, Departament d'Economia Aplicada, 2003.

J. D. Scargle, J. P. Norris, B. Jackson, and J. Chiang. Studies in astronomical time series analysis. vi. Bayesian block representations. *The Astrophysical Journal*, 764(2):167, 2013.

S. J. Schiff, A. Aldroubi, M. Unser, and S. Sato. Fast wavelet transformation of EEG. *Electroencephalography and Clinical Neurophysiology*, 91(6):442–455, 1994.

S. Schmitt, K. Pargeon, E. Frechette, L. Hirsch, J. Dalmau, and D. Friedman. Extreme delta brush: A unique EEG pattern in adults with anti-NMDA receptor encephalitis. *Neurology*, 79(11):1094–1100, 2012.

A. L. Schröder and P. Fryzlewicz. Adaptive trend estimation in financial time series via multiscale change-point-induced basis recovery. *Statistics and its Interface*, 6(4):449–461, 2013.

O. Seidou and T. B. Ouarda. Recursion-based multiple changepoint detection in multiple linear regression and application to river streamflows. *Water Resources Research*, 43(7), 2007.

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

M. Sharifzadeh, F. Azmoodeh, and C. Shahabi. Change detection in time series data using wavelet footprints. In *Advances in Spatial and Temporal Databases*, pages 127–144. Springer, 2005.

J. Shen, C. M. Gallagher, and Q. Lu. Detection of multiple undocumented change-points using adaptive lasso. *Journal of Applied Statistics*, 41(6):1161–1173, 2014.

A. F. M. Smith. A Bayesian approach to inference about a change-point in a sequence of random variables. *Biometrika*, 62(2):407–416, 1975.

M. Spangenberg, J. Tourneret, V. Calmettes, and G. Duchateau. Detection of variance changes and mean value jumps in measurement noise for multipath mitigation in urban navigation. In *Signals, Systems and Computers, 2008 42nd Asilomar Conference on*, pages 1193–1197, 2008.

V. Spokoiny. Multiscale local change point detection with applications to value-at-risk. *Annals of Statistics*, 37:1405–1436, 2009.

M. Srivastava and K. Worsley. Likelihood ratio tests for a change in the multivariate normal mean. *Journal of the American Statistical Association*, 81(393):199–204, 1986.

D. Stephens and A. Smith. Bayesian edge-detection in images via changepoint methods. In *Computer Intensive Methods in Statistics*, pages 1–29. Springer, 1993.

D. A. Stephens. Bayesian retrospective multiple-changepoint identification. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 43(1):159–178, 1994.

L. H. Strens, N. Fogelson, P. Shanahan, J. C. Rothwell, and P. Brown. The ipsilateral human motor cortex can functionally compensate for acute contralateral motor cortex dysfunction. *Current Biology*, 13(14):1201–1205, 2003.

F. Sun, L. Miller, and M. D'Esposito. Measuring interregional functional connectivity using coherence and partial coherence analyses of fMRI data. *Neuroimage*, 21(2):647–658, 2004.

G. J. Szekely and M. L. Rizzo. Hierarchical clustering via joint between-within distances: Extending Ward's minimum variance method. *Journal of Classification*, 22(2):151–183, 2005.

J. Terrien, G. Germain, C. Marque, and B. Karlsson. Bivariate piecewise stationary segmentation; improved pre-treatment for synchronization measures used on non-stationary biological signals. *Medical Engineering & Physics*, 35(8):1188–1196, 2013.

R. J. Tibshirani. Adaptive piecewise polynomial estimation via trend filtering. *Annals of Statistics*, 42(1):285–323, 2014.

C. Timmermans, L. Delsol, and R. von Sachs. Using Bagidis in nonparametric functional data analysis: predicting from curves with sharp local features. *Journal of Multivariate Analysis*, 115:421–444, 2012.

R. S. Tsay. Outliers, level shifts, and variance changes in time series. *Journal of Forecasting*, 7(1):1–20, 1988.

R. Turner, Y. Saatci, and C. E. Rasmussen. Adaptive sequential Bayesian change point detection. *Temporal Segmentation Workshop at NIPS*, 2009.

A. T. Tzallas, D. G. Tsalikakis, E. C. Karvounis, L. Astrakas, M. Tzaphlidou, M. G. Tsipouras, and S. Konitsiotis. *Automated epileptic seizure detection methods: a review study*. InTech, 2012.

D. A. Vasseur, J. P. DeLong, B. Gilbert, H. S. Greig, C. D. Harley, K. S. McCann, V. Savage, T. D. Tunney, and M. I. O'Connor. Increased temperature variation poses a greater risk to species than climate warming. *Proceedings of the Royal Society of London B: Biological Sciences*, 281(1779):20132612, 2014.

P. Čížek, W. Härdle, and V. Spokoiny. Adaptive pointwise estimation in time-inhomogeneous conditional heteroscedasticity models. *Econometrics Journal*, 12:248–271, 2009.

E. Venkatraman and A. B. Olshen. A faster circular binary segmentation algorithm for the analysis of array cgh data. *Bioinformatics*, 23(6):657–663, 2007.

E. S. Venkatraman. *Consistency results in multiple change-point problems*. PhD thesis, Department of Statistics, Stanford University, 1992.

J.-P. Vert and K. Bleakley. Fast detection of multiple change-points shared by many signals using group lars. In *Advances in Neural Information Processing Systems*, pages 2343–2351, 2010.

B. Vidakovic. *Statistical Modeling by Wavelets*. John Wiley & Sons, 2009.

T. J. Vogelsang. Testing for a shift in mean without having to estimate serial-correlation parameters. *Journal of Business & Economic Statistics*, 16(1):73–80, 1998.

L. Vostrikova. Detecting 'disorder' in multidimensional random processes. *Soviet Mathematics Doklady*, 24:55–59, 1981.

D. Wang, D. Miao, and C. Xie. Best basis-based wavelet packet entropy feature

extraction and hierarchical EEG classification for epileptic detection. *Expert Systems with Applications*, 38(11):14314–14320, 2011.

J. Wang and E. Zivot. A Bayesian time series model of multiple structural changes in level, trend, and variance. *Journal of Business & Economic Statistics*, 18(3): 374–386, 2000.

Y. Wang. Jump and sharp cusp detection by wavelets. *Biometrika*, 82(2):385–397, 1995.

P. Welch. The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Transactions on audio and electroacoustics*, pages 70–73, 1967.

B. Whitcher, S. D. Byers, P. Guttorp, and D. B. Percival. Testing for homogeneity of variance in time series: Long memory, wavelets, and the nile river. *Water Resources Research*, 38(5):12–1–12–16, 2002.

G. Worrell, L. Parish, S. Cranstoun, R. Jonas, G. Baltuch, and B. Litt. High-frequency oscillations and seizure generation in neocortical epilepsy. *Brain*, 127 (7):1496–1506, 2004.

K. Worsley. The power of likelihood ratio and cumulative sum tests for a change in a binomial probability. *Biometrika*, 70(2):455–464, 1983.

K. J. Worsley. Confidence regions and tests for a change-point in a sequence of exponential family random variables. *Biometrika*, 73(1):91–104, 1986.

W. B. Wu, M. Woodroofe, and G. Mentz. Isotonic regression: Another look at the changepoint problem. *Biometrika*, 88(3):793–804, 2001.

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

Q. Yao. Tests for change-points with epidemic alternatives. *Biometrika*, 80(1): 179–191, 1993.

Y.-C. Yao. Estimating the number of change-points via Schwarz' criterion. *Statistics & Probability Letters*, 6(3):181–189, 1988.

Y.-C. Yao and S. T. Au. Least-squares estimation of a step function. *Sankhyā: The Indian Journal of Statistics Series A*, 51:370–381, 1989.

T. Young Yang and L. Kuo. Bayesian binary segmentation procedure for a poisson process with multiple changepoints. *Journal of Computational and Graphical Statistics*, 10(4):772–785, 2001.

A. Zeileis and T. Hothorn. A toolbox of permutation tests for structural change. *Statistical Papers*, 54(4):931–954, 2013.

A. Zeileis, C. Kleiber, W. Krämer, and K. Hornik. Testing and dating of structural changes in practice. *Computational Statistics & Data Analysis*, 44(1):109–123, 2003.

C. Zhang and J. H. Hansen. Effective segmentation based on vocal effort change point detection. *Speech Analysis and Processing for Knowledge Discovery, Aalborg, Denmark*, 2008.

F. Zhang. High-frequency trading, stock volatility, and price discovery. *Available at SSRN 1691679*, 2010.

N. R. Zhang and D. O. Siegmund. A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, 63(1):22–32, 2007.

J. Zhou and P. Perron. Testing for breaks in coefficients and error variance: Simulations and applications. Technical report, Citeseer, 2008.