Randomised and L_1-penalty approaches to segmentation in time series and regression models

Karolos K. Korkas

A thesis submitted to the Department of Statistics of the

London School of Economics

for the degree of Doctor of Philosophy, London, August 2014

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).

I declare that my thesis consists of about 50,000 words.

Abstract

It is a common approach in statistics to assume that the parameters of a stochastic model change. The simplest model involves parameters than can be exactly or approximately piecewise constant. In such a model, the aim is the posteriori detection of the number and location in time of the changes in the parameters. This thesis develops segmentation methods for non-stationary time series and regression models using randomised methods or methods that involve L_1 penalties which force the coefficients in a regression model to be exactly zero. Randomised techniques are not commonly found in nonparametric statistics, whereas L_1 methods draw heavily from the variable selection literature. Considering these two categories together, apart from other contributions, enables a comparison between them by pointing out strengths and weaknesses. This is achieved by organising the thesis into three main parts.

First, we propose a new technique for detecting the number and locations of the change-points in the second-order structure of a time series. The core of the segmentation procedure is the Wild Binary Segmentation method (WBS) of Fryzlewicz (2014), a technique which involves a certain randomised mechanism. The advantage of WBS over the standard Binary Segmentation lies in its localisation feature, thanks to which it works in cases where the spacings between change-points are short. Our main change-point detection statistic is the wavelet periodogram which allows a rigorous estimation of the local autocovariance of a piecewise-stationary process. We provide a proof of consistency and examine the performance of the method on simulated and real data sets.

Second, we study the fused lasso estimator which, in its simplest form, deals with the estimation of a piecewise constant function contaminated with Gaussian noise (Friedman et al. (2007)). We show a fast way of implementing the solution path algorithm of Tibshirani and Taylor (2011) and we make a connection between their algorithm and the taut-string method of Davies and Kovac (2001). In addition, a theoretical result and a simulation study indicate that the fused lasso estimator is suboptimal in detecting the location of a change-point.

Finally, we propose a method to estimate regression models in which the coefficients vary with respect to some covariate such as time. In particular, we present a path algorithm based on Tibshirani and Taylor (2011) and the fused lasso method of Tibshirani et al. (2005). Thanks to the adaptability of the fused lasso penalty, our proposed method goes beyond the estimation of piecewise constant models to models where the underlying coefficient function can be piecewise linear, quadratic or cubic. Our simulation studies show that in most cases the method outperforms smoothing splines, a common approach in estimating this class of models.

Acknowledgments

First and foremost, I am deeply grateful to my supervisor Prof. Piotr Fryzlewicz. The joy and enthusiasm for his research was contagious and inspirational for me. I appreciate all his contributions of time and ideas to make my Ph.D experience productive and stimulating.

I am also grateful to Prof. Ryan Tibshirani from the Department of Statistics, Carnegie Mellon University for his warm hospitality during my visit to the department to carry out part of the research.

I gratefully acknowledge the generous financial support from the Engineering and Physical Sciences Research Council (EPSRC) and the Doctoral Training Centre in Financial Computing based in the University College London (UCL). I am also thankful to Prof. Philip Treleaven, Director of the Financial Computing Centre at UCL, for his support and encouragement during my doctoral studies.

I would like to thank my family. My mother for showing me how to be pragmatic. My father for teaching me that every problem is more complicated than it initially looks and that there is no right or wrong in almost every aspect of life. And my (big) brother for all the good moments in the past.

And most of all, I am grateful to Dionysia for all her support all these years; for giving me love she has so gratefully received from her parents and grandparents; and for consistently reminding me that *"...there is no passion to be found playing small - in settling for a life that is less than the one you are capable of living" (Nelson Mandela).*

Dedicated to Dionysia

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In many practical applications it is often more realistic to assume that the parameters of a stochastic model do not remain constant. For instance, the market volatility observed in many financial time series is unlikely to remain constant through time. A model that considers the varying parameter will probably result in a better forecasting performance and, therefore it is important to estimate it accurately. This issue has attracted considerable attention within the statistical and econometric literature mainly due to the wide range of applicability of these models. On top of that the technological advancement has generated a tremendous amount of data (now commonly referred to as *big data*). All these leave much space for the development of new estimation methods which need to be faster and more accurate.

This thesis deals with the problem of estimating a model that possibly has a varying structure. The main estimation tools are based on randomised algorithms or methods with L_1 penalties. We consider the segmentation of non-stationary time series as well as more general regression models where the coefficients are allowed to vary with respect to some variable. In Chapter 2, we review the literature in

the relevant areas including non-stationary time series, variable selection methods and non-parametric regression. The rest of the thesis consists of three parts and is organised as follows.

Chapter 3: Multiple change-point detection for non-stationary time series using Wild Binary Segmentation

In this chapter, which has been submitted to a journal and is currently under consideration for publication, we propose a new technique for consistent estimation of the number and locations of the change-points in the second-order structure of a time series. The core of the segmentation procedure is the Wild Binary Segmentation method (WBS) proposed by Fryzlewicz (2014), a technique which involves a certain randomised mechanism. The advantage of WBS over the standard Binary Segmentation lies in its localisation feature, thanks to which it works in cases where the spacings between change-points are short. In addition, we do not restrict the total number of change-points a time series can have. We also ameliorate the performance of our method by combining the CUSUM statistics obtained at different scales of the wavelet periodogram, our main change-point detection statistic, which allows a rigorous estimation of the local autocovariance of a piecewise-stationary process. We provide an extensive simulation study to examine the performance of our method for different types of scenarios. Finally, we examine the practical performance of our method by applying it to the US Gross National Product (GNP) data with the purpose to detect peaks and troughs in the growth of the US economy; and the infant electrocardiogram data (ECG) with the purpose to identify the sleep states.

Chapter 4: A fast implementation and a criticism of the fused lasso estimator

In this chapter, we build upon the solution path algorithm of Tibshirani and Taylor (2011), a method developed to solve lasso-type problems. We are particularly interested in the fused lasso estimator which, in its simplest form, deals with the estimation of a piecewise constant function contaminated with Gaussian noise (Friedman et al. (2007)). We show a faster way of implementing this method by exploiting the special structure of the matrix multiplications embedded in this algorithm. In addition, we make a connection between the solution path algorithm of Tibshirani and Taylor (2011) and the taut-string method of Davies and Kovac (2001) which also solves problems with total variation penalties. Further, we show that their algorithm has a "top-down" approach resembling other methods such as the Binary Segmentation method, which was also shown in Cho and Fryzlewicz (2011) for the taut-string method. As such we are able to compare the two methods both theoretically and practically through an extensive simulation study. In addition, Brodsky and Darkhovsky (1993) and Cho and Fryzlewicz (2011) argue that estimators with total variation penalties are suboptimal in detecting the number and locations of the change-points. We prove an exact rate of convergence for an estimated change-point using the fused lasso method and provide numerical evidence to support this claim.

Chapter 5: Adaptive Estimation of Time-Varying Models

Regression models in which the coefficients vary with respect to some covariate, such as time, arise naturally in many practical studies. This is due to the fact that the assumption of constant coefficients can possibly reduce the forecasting accuracy of a model if the coefficients exhibit smooth transitions, present in many aspects of science. This chapter proposes a path algorithm based on Tibshirani and Taylor (2011) and the fused lasso method of Tibshirani et al. (2005). The latter is an

extension of the lasso, a variable selection tool widely used in the context of high dimensional linear regression problems, i.e. cases where the number of the variables is larger than the sample size. Thanks to the adaptability of the fused lasso penalty, our proposed method goes beyond the estimation of piecewise constant models (the main contribution of Chapters 3 and 4) to models where the underlying coefficient function can be piecewise linear, quadratic or cubic. Our method draws from a new adaptive technique in nonparametric regression of Tibshirani (2014). The examples considered in the simulation study show that in most cases our method outperforms smoothing splines, a common approach in estimating this class of models.

It is noteworthy that Chapter 4, among other side contributions, serves as a comparison study between L_1 methods and the binary segmentation search (and to an extent the randomised binary segmentation search). It justifies the use of the latter in non-parametric regression, but it points out the flexibility, adaptability and the extensive coverage in the literature of methods with L_1 penalties. It is mainly for these reasons we choose to develop a lasso-type algorithm in Chapter 5 without, of course, arguing that other methods could not potentially perform better.

Finally, Chapter 6 summarises the contributions of this thesis and proposes directions for future research.

Chapter 2

Literature Review

2.1 Non-stationary time series

2.1.1 Stationary and locally stationary models

A time series is a collection of random variables measured at successive points in time. They are found in different aspects of science, technology, economics, medicine etc and the demand for effective tools for analysing and modelling them is strong. The literature on time series is vast and we refer the reader to some standard monographs, i.e. Priestley (1981), Brockwell and Davis (2002) or Hamilton (1994). The main challenge of a typical time series is that the observations are not independent, but rather they posses a degree of a stochastic dependence.

The statistical literature is mainly focused around stationary time series. We say that X_t is a stationary time series when its statistical properties remain unchanged through time. We say that X_t is a strictly stationary time series where the joint distribution of $(X_{t_1}, ..., X_{t_n})$ is the same as $(X_{t_1+\tau}, ..., X_{t_n+\tau})$ for all $n,\ t_1, ..., t_n$ and τ. Strict stationarity is often restrictive for practical purposes and the following form

is used. We say that a univariate, zero-mean time series X_t is weakly or second-order stationary if the autocovariance function $c_X(\tau) = \text{cov}(X_t, X_{t+\tau})$ is a function of τ only. Finally, another way of examining a time series is the spectral density function (or *spectrum*), $f_X(\omega)$, which determines how much energy is contained in a time series at different frequencies $\omega \in (-\pi, \pi)$.

A zero-mean stationary process admits the following Crámer representation

$$X_t = \int_{-\pi}^{\pi} A(\omega) \exp(i\omega t) \mathrm{d}\xi(\omega), \quad t \in \mathbb{Z} \tag{2.1}$$

where $A(\omega)$ is the amplitude of X_t and $\mathrm{d}\xi(\omega)$ is an orthonormal increment process. Simply, the Crámer representation says that X_t is the (weighted) sum of Fourier exponentials oscillating at different frequencies. Under mild conditions, the covariance structure of X_t can be expressed as

$$c_X(\tau) = \int_{-\pi}^{\pi} f_X(\omega) \exp(i\omega\tau) \mathrm{d}\omega$$

where $f_X(\omega) := |A(\omega)|^2$.

In practice, it is rare to find time series that are stationary (even in the weakest form) and it can have important implications when fitting models developed for stationary time series to real data (Mercurio and Spokoiny (2004)). It is therefore necessary to focus on non-stationary modelling and avoid all the restrictions imposed by assuming stationarity. One way of doing this is to introduce time dependence into the Crámer representation by replacing the constant $A(\omega)$ with a time-varying amplitude function $A_t(\omega)$. Dahlhaus (1997) proposes a class of *locally stationary processes* where X is modelled as a triangular stochastic array $\{X_{t,T}\}_{t=1}^{T}$ for $T = 1, 2, ...$ such that (for simplicity assume that there is no trend present in the process)

$$X_t = \int_{-\pi}^{\pi} A^0_{t,T}(\omega) \exp(i\omega t) \mathrm{d}\xi(\omega)$$

and there exists a function $A : [0,1] \times \mathbb{R} \to \mathbb{C}$, which is 2π-periodic in ω with $A(u,-\omega) = \overline{A(u,\omega)}$ such that

$$\sup_{t,\omega} \left| A^0_{t,T}(\omega) - A(\frac{t}{T},\omega) \right| \leq KT^{-1} \;\; \forall\, T$$

for some $K > 0$.

An alternative to the Fourier based approach for modeling time series whose spectral characteristics change over time is the class of locally stationary wavelet processes (LSW) introduced by Nason et al. (2000) where the difference lies in the use of non-decimated wavelets instead of Fourier exponentials. The use of wavelets means that the LSW model is localised both in time and in frequency and it has been embraced or adapted to model many types of non-stationary time series, e.g. financial (Fryzlewicz (2005)), image texture (Eckley et al. (2010)), experimental neuroscience (Sanderson et al. (2010)). We present the LSW model in Section 2.1.2, after a brief introduction to wavelets in Section 2.1.2.1.

2.1.2 Wavelets and the locally stationary wavelet model

2.1.2.1 Introduction to wavelets

Wavelets are localised, compactly supported oscillating functions which integrate to zero. This is in contrast to sine and cosine functions in Fourier analysis, which also oscillate, but the amplitude of their oscillation always remain unchanged. This is why when plotted resemble "little waves". Wavelets have received significant attention from the mathematical (Daubechies (1992), Mallat (1999)) and applied sciences community such as signal processing (Rioul and Vetterli (1991), Shapiro (1993) among others) or image and audio compression (Salomon (2004)). For wavelets in

statistics we refer the reader to Nason (2008), Vidakovic (2009), Percival and Walden (2000) for applications in time series analysis, Antoniadis (2007) for a review article and Abramovich et al. (2000) for an introductory paper. In Section 2.4.5 we discuss the application of wavelets in non-parametric regression which deals with extracting the signal from a noisy series.

We now formally describe wavelets. A function $\psi(x) \in \mathbb{L}_2(\mathbb{R})$ (i.e. a function that belongs to the space of all square-integrable functions) is defined to be a wavelet function (or *mother wavelet*) if it satisfies the admissibility condition

$$C_\psi = \int_{-\infty}^{\infty} \frac{\Psi(\omega)^2}{|\omega|} \, \mathrm{d}\omega < \infty \tag{2.2}$$

where $\Psi(\omega)$ is the Fourier transform of $\psi(x)$. Condition (2.2) implies that

$$\int \psi(x) \mathrm{d}x = \Psi(0) = 0. \tag{2.3}$$

A family of functions $\psi_{a,b}, a \in \mathbb{R}\backslash\{0\}, b \in \mathbb{R}$ are generated from the mother wavelet as translated (shifts) and dilated (stretches) versions of ψ, i.e.

$$\psi_{a,b} = \frac{1}{\sqrt{a}} \psi \left(\frac{x-b}{a} \right).$$

Condition (2.2) means that $\psi(x)$ has an exponential decay over $\mathbb{L}_2(\mathbb{R})$ and, hence, it should be localised in frequency. Condition (2.3) ensures that $\psi(x)$ has an oscillatory behaviour and it is also localised in time.

Continuous and discrete wavelet transform

The continuous wavelet transform of any function $f \in \mathbb{L}_2(\mathbb{R})$ is defined as a function of two variables

$$\mathrm{CWT}_f(a,b) = \langle f, \psi_{a,b} \rangle = \int f(x) \psi_{a,b}(x) \mathrm{d}x.$$

Under Condition (2.2), the original function f is recovered through the following inverse transform ("resolution of identity")

$$f(x) = C_{\psi}^{-1} \int_{\mathbb{R}^2} \mathrm{CWT}_f(a,b) f(x) \psi_{a,b}(x) a^{-2} \mathrm{d}a\mathrm{d}b.$$

$\mathrm{CWT}_f(a,b)$ is a function of two real variables and, hence, it is a redundant transform. To reduce this redundancy the values of a and b can be discretised so that the invertibility of the transform is maintained. To preserve all the information such a discretisation cannot be coarser than the *critical sampling*. The critical sampling will produce a basis for $\mathbb{L}_2(\mathbb{R})$ for $a = 2^{-j}$ and $b = r2^{-j}$ and under mild conditions on ψ, the basis

$$\psi_{j,r}(x) = 2^{j/2} \psi \left(2^j x - r\right) \quad j, r \in \mathbb{Z}$$

will be orthonormal. Index j is referred to as scale and r as location while large (small) values of j denote finer (coarser) scales. A theoretical framework for this discretisation is the *multiresolution analysis* which we do not cover here but we refer the reader to Mallat (1989).

Haar wavelets

The simplest and best-known example of wavelets are Haar wavelets given by

$$\psi(x) = \begin{cases} 1 \text{ if } 0 \leq x \leq 1/2 \\ -1 \text{ if } 1/2 < x \leq 1 \\ 0 \text{ otherwise.} \end{cases}$$

We note that the Haar wavelet belongs to the compactly supported Daubechies wavelets. Daubechies (1992) identifies the Extremal Phase family of wavelet systems which are compactly supported wavelets, possessing different degrees of freedom. Many other wavelets or families of wavelets exist, for example Daubechies Least

Asymmetric family of wavelets, Meyer's wavelets, Shannon's wavelets, see Vidakovic (2009) for some examples of these. In this thesis we only make use of the piecewise constant Haar wavelets which are a natural choice given that we are interested in processes whose second-order structure evolves over time in a piecewise constant manner (Chapter 3).

2.1.2.2 Locally stationary wavelet model

For the LSW model Nason et al. (2000) apply the pyramid algorithm to construct compactly supported discrete wavelets $\psi_j = (\psi_{j,0}, ..., \psi_{j,(N_j)-1})$ of length $\mathcal{N}_j$ for scale $j < 0$ such that

- $\psi_{-1,n} = \sum_r g_{n-2r}\delta_{0,r} = g_n$ for $n = 0, ..., N_{-1} - 1$
- $\psi_{(j-1),n} = \sum_r h_{n-2r}\psi_{j,r} = g_n$ for $n = 0, ..., N_{j-1} - 1$
- $\mathcal{N}_j = (2^{-j} - 1)(N_h - 1) + 1$

where $\delta_{0,r}$ is the Kronecker delta and N_h denotes the number of the elements of $\{h_r\}$ that are non-zero.

The key difference now is that *non-decimated wavelets* rather than wavelet functions as in Section 2.1.2.1 are used which can be shifted to any location at each scale and not by shifts by 2^{-j} so that $\psi_{j,r}(\tau) = \psi_{j,(r-\tau)}$, $\tau \in \mathbb{Z}$.

We now proceed with the definition of the LSW model.

Definition 2.1. *A triangular stochastic array* $\{X_{t,T}\}_{t=0}^{T-1}$ *for* $T = 1, 2, ...,$ *is in the class of Locally Stationary Wavelet (LSW) processes if there exists a mean-square representation*

$$X_{t,T} = \sum_{j=-J(T)}^{-1} \sum_{r=-\infty}^{\infty} \omega_{j,r;T}\psi_{j,r}(t)\xi_{j,r} \tag{2.4}$$

where $\psi_{j,r}(t)$ *are the non-decimated discrete wavelets vectors,* $J(T) = -\min\{j : \mathcal{N}_j \leq T\}$*,* $\omega_{j,r;T}$ *are real constants and* $\xi_{j,r}$ *are zero-mean, orthonormal, identically distributed random variables. In addition, for each* j *there exists a Lipschitz-continuous function* $W_j(z) : [0,1] \to \mathbb{R}$ *such that*

- $\sum_{j=-\infty}^{-1} W_j^2(z) < \infty$ *uniformly in* z*,*
- *the Lipschitz constant* L_j *are uniformly bounded in* j *and satisfy* $\sum_{j=-\infty}^{-1} 2^{-j} L_j < \infty$*, and*
- *there exists a sequence of constants* C_j *which satisfy* $\sum_{j=-\infty}^{-1} 2^{-j} C_j < \infty$ *and*

$$sup_{0 \leq r \leq T-1} \left| \omega_{j,r;T} - W_j\left(\frac{r}{T}\right) \right| \leq \frac{C_j}{T}$$

for each T *and* $j = -1, ..., -J(T)$*.*

The usual summary statistic in the general time series is the *spectrum* and a similar quantity can be defined within the LSW framework. The *evolutionary wavelet spectrum* (EWS) is defined in rescaled time as

$$S_j(z) = W_j^2(z) = \lim_{T \to \infty} \omega_{j,\lfloor zT \rfloor;T}.$$

The LSW model implies that $X_{t,T}$ is a linear combination of oscillatory functions $(\psi_{j,r})$ and the autocovariance function will depend on time if $X_{t,T}$ is locally stationary. Analogous to the stationary time series where the spectral density is related to the autocovariance function (one being the Fourier transform of the other) a similar relationship can also be shown for the LSW. First, let $c_T(z, \tau)$ denote the finite-sample covariance function of $X_{t,T}$ at lag τ and rescaled time location z

$$c_T = \mathbb{E}(X_{\lfloor zT \rfloor,T}, X_{\lfloor zT \rfloor+\tau,T}).$$

Now, define the *autocorrelation wavelet* $\Psi_j(\tau)$ (Nason et al. (2000))

$$\Psi_j(\tau) = \sum_k \psi_{j,r}\psi_{j,r+\tau}.$$

Further, let $c(z,\tau)$ be the *asymptotic local autocovariance* function of $X_{t,T}$ at lag τ and rescaled time location z, defined as a transform of $S_j(z)$ with respect to the set of autocorrelation wavelets

$$c(z,\tau) = \sum_{j=-\infty}^{-1} S_j(z)\Psi_j(\tau). \tag{2.5}$$

Nason et al. (2000) show that under the assumptions of Definition 2.1 the asymptotic local autocovariance $c(z,\tau)$ is a good approximation to the sample covariance $c_T(z,\tau)$, i.e. $|c_T(z,\tau) - c(z,\tau)| = \mathcal{O}(T^{-1})$. This is an interesting link between the autocovariance of $X_{t,T}$ and the EWS, an analogue of the usual formula, that is, the autocovariance of stationary process is the Fourier transformation of the spectrum. This one-to-one correspondence can be also seen from the invertibility of (2.5), i.e.

$$S_j(z) = \sum_\tau \left(\sum_r \Psi_r(\tau) A^{-1}_{j,j'} c(z,\tau) \right)$$

where $A_{j,j'} = \sum_\tau \Psi_j(\tau)\Psi_{j'}(\tau)$.

Estimation of the LSW model

For a time series at hand it is important to have a means of estimating the EWS. Nason et al. (2000) define and propose to use the *raw wavelet periodogram* as a method of estimating the EWS. Since (2.4) indicates that the time series $X_{t,T}$ is the inverse wavelet transform of the coefficients $\omega_{j,r;T}\psi_{j,r}(t)\xi_{j,r}$, then the EWS can be estimated from the squares of the non-decimated wavelet coefficients.

We now provide the definition of the wavelet periodogram from Nason et al. (2000).

Definition 2.2. *Let $X_{t,T}$ be an LSW process constructed using the wavelet system ψ. The triangular stochastic array*

$$I_{t,T}^{(j)} = \left| \sum_{s} X_{s,T} \psi_{i,s-t} \right|^2 \tag{2.6}$$

is called the wavelet periodogram of $X_{t,T}$ at scale j.

Nason et al. (2000) show that the wavelet periodogram (3.8) is not an asymptotically unbiased estimator of the wavelet spectrum. Indeed, they prove the following result for all $j \leq -1$

$$\mathbb{E} I_{t,T}^{(j)} = \sum_{j'=-\infty}^{-1} S_{j'} \left(\frac{t}{T} \right) A_{j,j'} + \mathcal{O}(2^{-j}/T).$$

To deal with the inconsistency of the EWS estimator Nason et al. (2000) choose to smooth the wavelet periodogram by the use of wavelet shrinkage. Smoothing a wavelet periodogram is not an easy task, mainly due to the fact that $I_{t,T}^{(j)}$ is a correlated series. Neumann and Von Sachs (1995), in a similar setting, use a non-linear estimation technique, however, it involves a pre-estimate of the local variance of the observations and can reduce the performance of the method, see e.g. Fryzlewicz (2005).

For a different approach in wavelet smoothing we refer the reader to Fryzlewicz and Nason (2006) and Fryzlewicz et al. (2006) who propose a device that stabilises the variance with the purpose to bring the data closer to Gaussianity with constant variance. The authors propose a transformation of data, called "Fisz-transform", but combined with wavelet (Haar) coefficients and, hence, termed "Haar-Fisz" technique.

Finally, we note that in this thesis we adopt a modified version of the LSW model in Definition 2.1 by Fryzlewicz and Nason (2006). In this version, the authors

assume that the transfer function $W_j(z)$ is piecewise constant with a finite number of change-points (jumps) and not Lipschitz-continuous as in Nason et al. (2000). This enables us to model a non-stationary process whose autocovariance structure evolves over time in a piecewise constant manner with the purpose to locate the areas of discontinuities which is the topic of Chapter 3.

2.2 Model selection methods using penalised least squares estimation

Variable selection in a high-dimensional statistical problem has attracted significant attention from researchers in different fields such as science, humanities, genomics, finance and machine learning. Donoho (2000) has stressed the importance for new developments in high-dimensional statistics.

One of the main challenges in dimension reduction is the estimation of the coefficients $\beta \in \mathbb{R}^{p\times 1}$ in the following model

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

where $\mathbf{y} \in \mathbb{R}^n$ is the response vector, $\mathbf{X} \in \mathbb{R}^{n\times p}$ is the design matrix, $\epsilon \in \mathbb{R}^n$ are iid random errors and p (the dimensionality of the data) is very large, possibly, $n << p$. The main challenge therefore is to select a subset $\mathcal{B}$ of variables that contribute to the response y, i.e.

$$\mathcal{B} = \{1 \leq j \leq p : \beta_j \neq 0\}.$$

There are many studies in the literature that deal with this high-dimensional problem. Some of the well-known classes of approaches include *greedy* methods (e.g.

forward and backward-stepwise selection) and methods that add a penalty to the minimisation of the loss function. By adding a penalty it is expected that a method will lead to a sparse solution with the hope that all the irrelevant variables will have coefficients close or equal to zero.

This section reviews certain model selection techniques in the context of L_q penalised least squares estimation that are most relevant to this thesis with the aim to build a pillar for Chapters 4 and 5 which deal with non-parametric regression and time-varying estimation respectively.

2.2.1 Ridge regression

Ridge regression aims to shrink the regression coefficients by minimising the following penalised cost function

$$f(\beta) = \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 \tag{2.7}$$

where $\lambda \geq 0$ controls the amount of shrinkage. In a regression setting where many correlated variables are present it is possibly that the estimated coefficients can exhibit high variance. By imposing the penalty constraint this problem is relieved.

Due to the different scaling of the predictors we can standardise $x_{i,j}$ such that $\sum_{i=1}^{N} x_{ij} = 0$ and $\sum_{i=1}^{N} x_{ij}^2 = 1$ (note that the intercept is left out from the penalty term so to avoid the procedure depending on the origin). We can now find the estimated coefficients by

$$\hat{\beta}^{\text{ridge}} = (X^T X + \lambda I)^{-1} X^T y$$

where I is the $p \times p$ identity matrix. One can see that that λI adds a positive

constant to the diagonal of X^TX and hence making the problem nonsingular if λ is chosen appropriately. In the case where $x_{i,j}$ are orthonormal then the estimated ridge coefficients can be obtained from $\hat{\beta}^{\text{ridge}} = \hat{\beta}^{\text{ols}}/(1+\lambda)$.

2.2.2 Least absolute shrinkage and selection operator (lasso)

Lasso performs a similar task to ridge regression and the aim of this technique is to shrink coefficients towards zero (Tibshirani (1996)). Similarly with ridge regression we can re-parameterise the constant β_0 by standardising $x_{i,j}$. Lasso finds those β_js that minimise the following Lagrange function

$$f(\beta) = \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| \tag{2.8}$$

where λ is a tuning parameter - large values means more coefficients are set to zero and hence the selected model is more sparse. Using the L_q norm notation the above problem takes the following matrix form

$$\min_{\beta\in\mathbb{R}^n} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda||\beta||_1$$

where $||v||_1 = |v_1| + |v_2| + ... + |v_n|$.

Typically, the solution of a lasso problem is carried out using a quadratic programming algorithm (see, e.g., Boyd and Vandenberghe (2004)), but other efficient algorithms are available such as the Least Angle Regression of Efron et al. (2004). It is worth mentioning that lasso can be seen as an iterated reweighted ridge regression and, hence, admits an (approximate) closed form solution (Tibshirani (1996)). In the simplest scenario where the predictors are uncorrelated with each other the solution to a lasso problem can be obtained by simple thresholding i.e. $\text{sign}(\hat{\beta}^{\text{ols}})(|\hat{\beta}^{\text{ols}}| - \lambda)_+$.

In a more general context, lasso can be seen as a variable selection method by setting coefficients exactly equal to zero. Hence, unlike ridge regression, lasso produces interpretable submodels. However, the former does better when variables are highly correlated for usual $n > p$ situations (Tibshirani (1996)).

2.2.3 The elastic net

The elastic net method of Zou and Hastie (2005) can be seen as a combined method of lasso and ridge regression. It adds a second constraint $||\beta||_2^2$ to the lasso problem, that is,

$$\min_{\beta \in \mathbb{R}^n} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda_1||\beta||_1 + \lambda_2||\beta||_2^2.$$

This method inherits features from both the ridge and lasso estimation in that it simultaneously does continuous shrinkage and variable selection. An important feature is that it allows more than n variables to be selected since lasso selects up to n variables under the $n << p$ paradigm. Another advantage over lasso is that it encourages a grouping effect, where predictors with high pairwise correlation tend to be in or out of the model together. By contrast, the lasso tends to select only one variable from such a group without any preference.

2.2.4 Fused lasso

Many extensions of the lasso have been proposed, e.g. the adaptive lasso (Zou (2006)), the elastic net (Zou and Hastie (2005)), the randomised lasso (Meinshausen and Bühlmann (2010)), the random lasso (Wang et al. (2011)). These methods mainly focus on improving the performance of lasso. A particularly interesting extension is when some prior information about the model is known that could be incorporated

into it. The fused lasso of Tibshirani et al. (2005) (see also Tibshirani and Wang (2008)) takes advantage of this information by using simultaneously the lasso penalty and an L_1 (total variation) penalty on the differences of neighbouring coefficients. Hence, it favours solutions that are both sparse and blocky. The loss function takes the following form

$$f(\beta) = \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda_1 \sum_{j=1}^{p}|\beta_j| + \lambda_2 \sum_{j=2}^{p}|\beta_j - \beta_{j-1}| \tag{2.9}$$

where λ_2 controls the smoothness of the resulting solution. There are no clear directions for how the regularisation parameters λ_1 and λ_2 are simultaneously tuned, at least in the context of variable selection. Hence, we can select the tuning parameters by cross-validation which is a commonly used method in penalised regression problems (see Hastie et al. (2009)).

It is not necessary to impose a penalty on neighbouring coefficients, but rather one can penalise differences of coefficients that correspond to an edge in a graph. This permits more flexibility and it has found applications in e.g. biostatistics and genetics where the purpose is to find associations between phenotypes (outputs) and a few single neucleotide polymorphisms (SNPs) out of millions SNPs (inputs) where inputs are closely related to each other (see Kim et al. (2009b) among others).

A special case of the fused lasso is when the predictor matrix $X = I \in \mathbb{R}^{n\times n}$ and it is termed the Fused Lasso Signal Approximator (FLSA), see Tibshirani and Wang (2008). The loss functions has the following form

$$f(\beta) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \beta_i)^2 + \lambda_1 \sum_{i=1}^{n}|\beta_i| + \lambda_2 \sum_{i=2}^{n}|\beta_i - \beta_{i-1}|. \tag{2.10}$$

The FLSA is related to the non-parametric regression which we discuss in Section 2.4. In Chapter 4, among other things, we explore its estimation performance. The

FLSA should be categorised as a denoising method, that is, extracting the signal μ from a noisy series y_i

$$y_i = \mu_i + \epsilon_i \text{ for } i \in 1, ..., n \text{ and } \epsilon_i \text{ are iid r.v.} \tag{2.11}$$

and not as a variable selection method. The reason we report it here is that the algorithms found in the literature focus on the minimisation of (2.10) mainly due to its conceptual simplicity. These algorithms are then extended to other settings, such as (2.9), see Hoefling (2010) or Tibshirani and Taylor (2011). Another important feature of the FLSA is the fact that the penalty parameter λ_1 which controls the lasso term can be set equal to 0. This is thanks to the following theorem

Theorem 2.1. ***(Friedman et al. (2007))*** *The solution for any value of* (λ_1, λ_2) *can be found by simple soft-thresholding of the solution obtained for* $(0, \lambda_2)$. *More precisely, if* $X = I$ *and the solution for* $\lambda_1 = 0$ *and* $\lambda_2 > 0$ *is a known quantity* $\beta(0, \lambda_2)$ *then the solution for* λ_1 *is* $\beta_i(\lambda_1, \lambda_2) = sign(\beta_i(0, \lambda_2))(|\beta_i(0, \lambda_2)| - \lambda_1)^+$.

Figure 2.1 shows an example of an application of FLSA on the Blocks signal, first examined by Donoho and Johnstone (1994), for different values of λ_2 and for $\lambda_1 = 0$. For $\lambda_2 \to \infty$, i.e. no regularisation, the estimated signal is a straight line (red), the mean of the series y_i. The other two lines are for $\lambda_2 = 5, 50$ and one can see that smoothness increases for smaller values by removing noise and improving estimation.

2.2.5 A note on subgradient theory

In the case where the loss function is convex but non-differentiable then the gradient is not defined. This is common with lasso-type optimisation problems where a loss function has a minimum, but it is not differentiable, so the standard gradient method

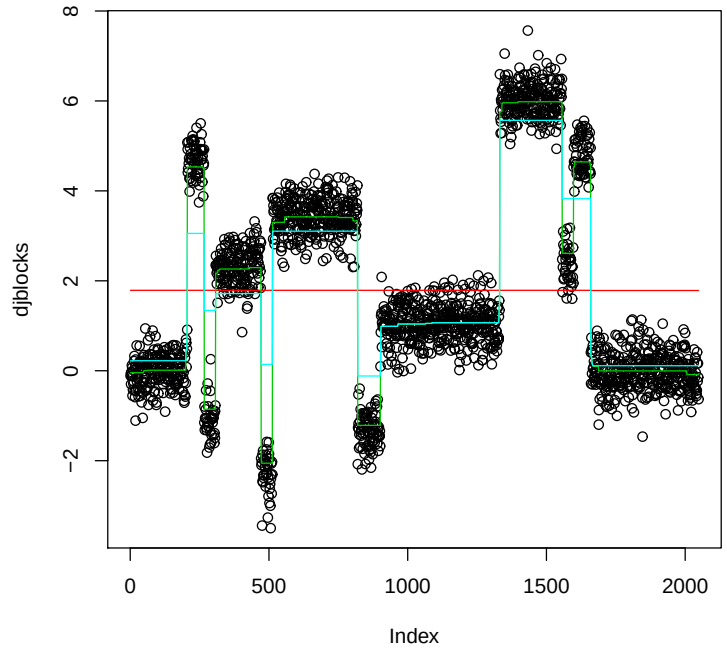

Figure 2.1: FLSA on the Blocks signal (Donoho and Johnstone (1994)), obtained from the R *package* wavethresh *(Nason (2013b)). Green, blue and red line is for* $\lambda_2 = 5$, $\lambda_2 = 50$ *and* $\lambda_2 = 700$ *respectively.*

cannot be used, see Figure 2.2 for an illustration of this. Hence, we need to introduce the notion of a *subgradient*. We say a vector $g \in \mathbb{R}^n$ is a subgradient of $f : \mathbb{R}^n \to \mathbb{R}$ at $x \in \text{dom} f$ if for all $y \in \text{dom} f$, $f(y) \geq f(x) + g^T(y - x)$. This is equivalent to the first-order condition when f is differentiable $f(y) \geq f(x) + \nabla f(x)^T(y - x)$ where $\nabla f(x)$ denotes the gradient of function f.

A function f is called subdifferentiable at x if there exists at least one subgradient at x. The set of all the subgradients at x of function f is called the subdifferential of f at x and is denoted by $\partial f(x)$.

A point $x^\star$ is said to be a minimiser of the convex function f if f is subdifferen-

tiable at that point and a subgradient $g \in \partial f(x)$ exists such that

$$g = 0.$$

If f is differentiable then the above optimality condition reduces to

$$\nabla f(x) = 0.$$

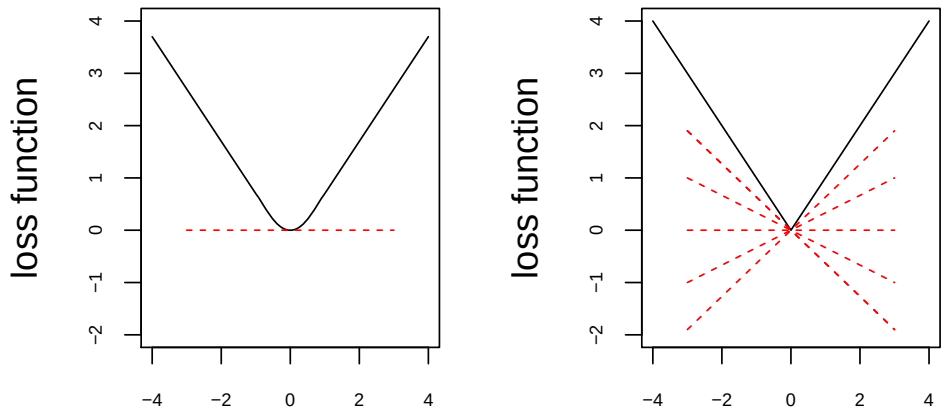

Figure 2.2: Left panel: a convex and differentiable loss function with a minimum at $x_0 = 0$. Right panel: a convex and non-differentiable loss function (right panel) with a minimum at $x_0 = 0$.

Calculus of subgradient

There are two rules that apply in subgradient calculus, i.e. the "weak" and the "strong". The aim of the former is to produce one subgradient, arbitrarily chosen even if more subgradients exist. It is sufficient in practice since most algorithms require only one subgradient at each stage. On the other hand, the "strong" calculus describe the complete set of subgradients $\partial f(x)$ as a function of f. We do not elaborate more on the calculus of subgradients, but we refer the reader to Bertsekas (1999).

2.3 Review of Estimation Methodologies

We now present different methods that have been proposed to solve the fused lasso problem. These can be categorised into two groups: the first includes path-wise algorithms which find the whole solution path for an increasing or decreasing regularisation parameter. The other includes algorithms that employ gradient descent optimisation methods that solve the fused lasso problem at a fixed value of the regularisation parameter, often determined by the user. The advantage of the first group over the second is that the user can obtain the whole path of solutions for all values of the regularisation parameter. On the other hand, the second group is not restricted by the rank of the predictor matrix X (when $n < p$ then X does not have full rank), a typical limitation of most methods of the first group.

2.3.1 Path algorithms

2.3.1.1 Pathwise coordinate optimisation

Friedman et al. (2007) explore "one-at-a-time" coordinate-wise algorithms which according to the authors is faster than the LARS algorithm of Efron et al. (2004) when trying to compute the lasso solution on a range of values of λ_1. Coordinate-wise algorithms apply an iterative soft-thresholding with a partial residual as a response variable. By rewriting (2.8) as

$$f(\tilde{\beta}) = \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{k\neq j} x_{ik}\tilde{\beta}_k - x_{ij}\beta_j\right)^2 + \lambda_1 \sum_{k\neq j}|\tilde{\beta}_j| + \lambda_1|\beta_j|$$

where the values of β_k for $k \neq j$ are fixed at values $\tilde{\beta}_k(\lambda_1)$ and minimising $f(\tilde{\beta})$ w.r.t β_j, we have

$$\tilde{\beta}_j(\lambda_1) \leftarrow S\left(\sum_{i=1}^{n} x_{ij}(y_i - \tilde{y}_i^{(j)}), \lambda_1\right)$$

where $S(\alpha, \beta) = \text{sign}(\alpha)(|\alpha| - \beta)_+$ is the threshold function. The update is repeated until the algorithm converges.

The coordinate-wise descent algorithm works for a range of penalised least squares problems, such as the elastic net (Zou and Hastie (2005)), the least absolute deviation regression (Li and Arce (2004)), the grouped lasso (Yuan and Lin (2006)) or the negative garrotte (Breiman (1995)). However, the coordinate-wise descent procedure needs to be substantially modified in order to be applied to the FLSA case because (2.9) is not continuously differentiable despite the fact that it is convex (Tseng (2001)). Hence, the algorithm can get stuck in a corner of the loss function $f(\beta)$. To advance to the minimum, we have to move coefficients together. Friedman et al. (2007) generalise the algorithm for the FLSA case in order to deal with this issue and their algorithm is still faster than a typical quadratic optimisation solver. Briefly, for $\lambda_1 = 0$ their algorithm is summarised into three nested cycles

- *Descent cycle:* Coordinate-wise descent is run for each parameter β_j, while all the others are held fixed.

- *Fusion cycle:* Neighbouring pairs of parameters are fused, followed by coordinate-wise descent.

- *Smoothing cycle:* Penalty λ_2 is increased by a small amount δ, and the two previous cycles are rerun.

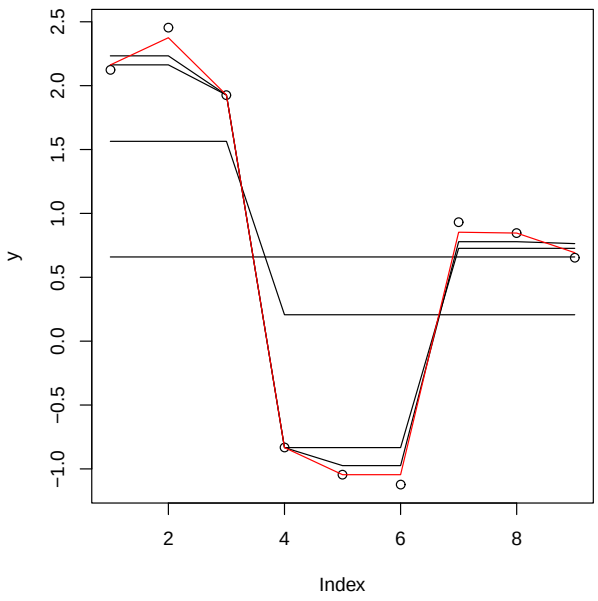

Figure 2.3: The path algorithm of Hoefling (2010) applied on a piecewise constant function contaminated with Gaussian noise for different values of λ_2. Points 5 and 6 have the closest value and thus are the first to be fused (red line). These points form the set $F_5 \in \{5, 6\}$. The next points to be fused are 7 and 8.

2.3.1.2 Path algorithm for the FLSA

Hoefling (2010) presents a faster algorithm than that of Friedman et al. (2007) which gives a solution for all values of λ_2, applies to the 2-dimensional FLSA problem (denoising an image) and to the general fused lasso when $\text{rank}(X) = p$. The path algorithm proposed from Hoefling (2010) is based on the idea of fused sets. We present the method for the 1-dimensional FLSA problem (2.10). The algorithm starts by setting $\lambda_2 = 0$ and then increase it until all coefficients are equal. A pair of sets F_i, F_{i+1} of coefficients that are "close" in values are merged (fused), and they

form a new set $F_{i'}$. This works as follows. The quantity

$$h_{i,i+1}(\lambda_2) = \frac{\beta_{F_i} - \beta_{F_{i+1}}}{\frac{\partial \beta_{F_{i+1}}}{\partial \lambda_2} - \frac{\partial \beta_{F_i}}{\partial \lambda_2}} + \lambda_2 \quad \text{for} \quad i = 1, ..., n_F(\lambda_2) - 1$$

where $n_F(\lambda_2)$ is the number of fused sets, will determine which two neighbouring sets of fused coefficients can be fused and have the same value by finding $i' = \arg\min_{h(\lambda_2)>\lambda_2} h_{i,i+1}$. Now, the coefficients $\beta_{F_{i'}}$ and $\beta_{F_{i'+1}}$ are fused and form the set $F_{i'}$ (for an illustration see Figure 2.3) with $\beta_{F_{i'}} = \beta_{F_i} + (h_{i,i+1}(\lambda_2) - \lambda_2)\left(\frac{\partial \beta_{F_{i+1}}}{\partial \lambda_2} - \frac{\partial \beta_{F_i}}{\partial \lambda_2}\right)$. The iteration continues until all coefficients are equal to each other and to the mean of y_i. Briefly, the method works by fusing together adjacent coordinates with similar values, which produces a blocky estimate. This "bottom-up" method is an opposite approach to the solution path algorithm of Tibshirani and Taylor (2011) where coefficients that are most different are identified first. This gives a computational advantage of the former since Hoefling (2010) calculates the computational complexity to be $\mathcal{O}(n \log n)$ (due to the tree structure of the algorithm) compared with the solution path algorithm of Tibshirani and Taylor (2011) (see Section 2.3.1.3) which, as we show in Chapter 4, is $\mathcal{O}(n^2)$ where n is the sample size. However, the latter solves the generalised lasso and, hence, it can be used in a range of different problems. Finally, we note that a similar "bottom-up" approach has been suggested by Fryzlewicz (2007) in the context of estimating in the model (2.11).

2.3.1.3 Solution path of the generalised lasso

Tibshirani and Taylor (2011) propose an algorithm to calculate the full path for the generalised lasso problem which includes the fused lasso. The authors consider an

argument from Kim et al. (2009a) who transform the following (primal) problem

$$\min_{\beta \in \mathbb{R}^n} \frac{1}{2}||y - \beta||_2^2 + \lambda_2 ||D\beta||_1$$

where $D \in \mathbb{R}^{(n-1)\times n}$ is the penalty matrix, into a simpler one with no linear transformations (the dual problem)

$$\min_{u \in R^m} \frac{1}{2}||y - D^T u||_2^2 \quad \text{subject to} \quad \|u\|_\infty \leq \lambda_2 \tag{2.12}$$

where $\|\Delta\|_\infty$ denotes the maximum absolute element of a matrix or vector Δ. The reason for this is that the L_1 penalty is composed with a linear transformation of β. It is easier to work with the dual (2.12): a regression with a simple constraint set. Starting from $\lambda_2 = \infty$ and moving towards $\lambda_2 = 0$, one can find the dual coordinates that hit the boundary (the constraint $\|u\|_\infty \leq \lambda_2$) in an one-by-one manner.

The solution path algorithm is based on the fact that the active set $\mathcal{B}$ (which contains the hitting coordinates) does not change as $\lambda_2 \to 0$ thanks to the following lemma

Lemma 2.1. ***(Tibshirani and Taylor (2011)):*** *For the 1-dimensional fused lasso (FLSA) we have that for any coordinate i, the solution $\hat{u}_\lambda$ of (2.12) satisfies*

$$\hat{u}_{\lambda_0,i} = \lambda_0 \Rightarrow \hat{u}_{\lambda,i} = \lambda \ \ for \ \ all \ \ \lambda \ \ \in [0, \lambda_0]$$

and

$$\hat{u}_{\lambda_0,i} = -\lambda_0 \Rightarrow \hat{u}_{\lambda,i} = -\lambda \ \ for \ \ all \ \ \lambda \ \ \in [0, \lambda_0].$$

Simply, the lemma states that for decreasing λ_2 the coordinate u_i stays within the boundary i.e. $u_i = \lambda_2$ and thus at every iteration we need to solve only for the interior coordinates. The boundary lemma is the equivalent of Proposition 2 of

Friedman et al (2007) which states that when two values $\hat{\beta}$ fuse then for increasing λ_2 those values remain always fused. The boundary lemma is about the fusion of the dual solutions $\hat{u}$ for decreasing λ_2.

The solution path algorithm can be extended to other fused lasso problems by making a transformation of the design matrix X, the response vector y and the penalty matrix D. However, Lemma 2.1 does not hold anymore and the algorithm needs to keep track of the dual coefficients that leave the active set $\mathcal{B}$.

In Chapter 4 we show, among other things, a fast version of this algorithm which does not involve matrix multiplications. In addition, in Chapter 5 we adapt the solution path algorithm to estimate time-varying models and hence propose a new path algorithm which is shown to perform well on the examples we consider in our simulation study.

A note on duality theory

Duality theory shows a way of constructing an alternative problem (the *dual problem*) from the original optimisation problem (the *primal problem*) and the optimisation problem is viewed from two different perspectives. Its purpose is e.g. to obtain easily a lower bound on the optimal value of the objective function for the original problem or because it is easier to computationally solve the dual than the primal problem.

Consider the (primal) problem

$$\min_{x \in \mathbb{R}^n} \sum_{i=1}^{n} f_i(x_i)$$

$$\text{subject to } a^T x \leq b$$

where a is a vector, b a scalar and $f_i : \mathbb{R} \to \mathbb{R}$ is a convex continuously differentiable

function. The main idea behind duality is to take the constraints in the problem above into account by adding to the objective function the constraint functions, i.e. form the Lagrangian function

$$\mathcal{L}(x, \lambda) = \sum_{i=1}^{n} f_i(x_i) + \lambda(b - a^T x).$$

The dual function is defined by

$$q(\lambda) = \inf_{x \in \mathbb{R}^n} (\mathcal{L}(x, \lambda)) = \inf_{x \in \mathbb{R}^n} \left(\sum_{i=1}^{n} f_i(x_i) + \lambda(b - a^T x) \right).$$

Hence, the dual problem is

$$\max q(\lambda) \text{ subject to } \lambda \in \mathbb{R}.$$

In general, the optimal solution to the primal problem is not necessarily equal to that of the dual problem and, hence, a *duality gap* exists. When the objective function is convex and strictly feasible (in the sense that the inequality constraints are strictly inequalities), then the optimal duality gap is zero and we say that *strong duality holds* (see Boyd and Vandenberghe (2004)). In the case of the solution path algorithm (Section 2.3.1.3) strong duality holds (there is only an equality constraint) and it is preferred to derive the dual problem since it is easier to work with.

2.3.2 Convex optimisation techniques

2.3.2.1 Proximal-gradient method for optimisation with smooth penalty term

Chen et al. (2010) propose a proximal-gradient method using an auxiliary matrix that smoothes the loss function. Recall from Section 2.2.5 that the loss function has a unique minimum, but it is not differentiable and, therefore standard proximal

gradient methods cannot be adopted. However, according to the authors, the approximation of the smoothed function is sufficiently close to the original objective function. Their algorithm can work efficiently under the $n << p$ paradigm while it is simple in the implementation. Notably, the method can be easily adapted to deal with many cases such as the (fused) lasso case, the 1-dimensional or 2-dimensional case or other structures such as grids or graphs. In a graph structure we are not restricted to sparsity in differences of neighbour coefficients, i.e.

$$f(\beta) = \frac{1}{2}\|Y - X\beta\|_2^2 + \lambda_1 \sum_{j=1}^{p} \beta_j + \lambda_2 \sum_{(\kappa,l)\in E, \kappa<l} |\beta_\kappa - \beta_l| \tag{2.13}$$

where E are the edges in a graph $G = (V, E)$ with $V = \{1, ..., p\}$ representing the variables. By restricting $\kappa < l$ we ensure that any two coefficients are penalised only once since the edges are assumed to be undirected.

Define the penalty matrix $C = (\lambda_1 I, \lambda_2 D)$ where $|D\beta| = \sum_{(k,l)\in E, \kappa<l} |\beta_\kappa - \beta_l|$. Then, (2.13) can be rewritten as follows

$$f(\beta) = \frac{1}{2}||Y - X\beta||_2^2 + ||\beta^T C||_1. \tag{2.14}$$

Now, an auxiliary vector α is defined with domain $Q = \{||a||_1 \leq 1\}$ such that

$$||\beta^T C||_1 = \max_{\alpha\in Q} \alpha\beta^T C. \tag{2.15}$$

The reformulated penalty term (2.15) can be seen as the inner product of the auxiliary vector α and the linear mapping of β via a linear operator $\Gamma(\beta) = \beta^T C$. Yet, it remains a non-smooth function of β, hence, optimisation is still not feasible. To deal with this, an auxiliary convex function $d(a)$ is defined on Q such that

$$f_\mu(\beta) = \max_{\alpha\in Q} \alpha\beta^T C - \mu d(a) \tag{2.16}$$

where μ is a smoothness parameter. The algorithm of Chen et al. (2010) utilises the optimal solution of (2.16) and propose $d(\alpha) = \frac{1}{2}||\alpha||_2^2$. To achieve efficient convergence they set $\mu = e/2G$ where e is the desired accuracy and G depends on α. By smoothing the objective function, the problem can be solved efficiently using a standard proximal gradient method such as the FISTA method (fast iterative shrinkage-thresholding algorithm) of Beck and Teboulle (2009).

2.3.2.2 Other methods

Alternative techniques are that of Ye and Xie (2011) who use a Split Bregman method to solve the fused lasso problem. The authors augment equation (2.9) by adding two terms that penalise the two linear constraints (the lasso and the fusion penalty). Then, the solution is found by solving a system of linear equations. Main limitation of this method is the choice of two extra parameters that can affect the rate of convergence of the algorithm. Wang et al. (2013) also augment the Lagrangian loss function with squares of the constraint functions. The attractive feature of this technique is the simplicity in implementation, yet this method applies only to the FLSA case. Lin et al. (2011) approach the minimisation by alternately solving two subproblems, the squared error function and the fusion penalty function which are both linearised (and thus termed *alternating linearisation*). Finally, Liu et al. (2010) propose a transformation of the general fused lasso problem into a standard FLSA problem and the use of a gradient descent method on its dual.

2.4 Non-parametric regression

Non-parametric regression focuses on the estimation of a function f_0 given the observations $y_1, ..., y_n \in \mathbb{R}$ from the following model

$$y_i = f_0(x_i) + \epsilon_i, \quad \text{for} \quad i = 1, ..., n \tag{2.17}$$

where $x_1, ..., x_n \in \mathbb{R}$ are input points and $\epsilon_1, ..., \epsilon_n \in \mathbb{R}$ are independent errors. In addition, it is assumed that the inputs $x_1, ..., x_n$ are evenly spaced over the unit interval $[0, 1]$, i.e. $x_i = i/n$ for $i = 1, ..., n$.

The non-parametric regression toolbox is highly-developed with plenty of methods based on kernels, polynomials, splines or wavelets. We review some of these in the next sections.

2.4.1 Kernel smoothing

Define a kernel K as a weighted mean function $K : \mathbb{R} \to \mathbb{R}$ such that

$$\int K(x)\mathrm{d}x = 1, \quad \int xK(x)\mathrm{d}x = 0, \quad \int x^2K(x)\mathrm{d}x < \infty.$$

Two well-known kernels are the Gaussian

$$K(x) = \frac{1}{\sqrt{2\pi}}\exp(-x^2/2),$$

and the Epanechnikov kernel

$$K(x) = \begin{cases} \frac{3}{4}(1 - x^2) \ \text{ if } \ |x| \leq 1 \\ \\ 0 \ \text{ otherwise.} \end{cases}$$

Now, a *kernel-smoother* is defined as

$$\hat{f}(x_0) = \frac{\sum_{i=1}^n K\left(\frac{x_0 - x_i}{h}\right) y_i}{\sum_{s=1}^n K\left(\frac{x_0 - x_s}{h}\right)}$$

where $h > 0$ is the bandwidth and determines the width of the local neighbourhoud at x_0 and it controls the "roughness" of the estimated function $\hat{f}(x_0)$ as, e.g., high values of h averages more observations reducing the variance, but increasing the bias. Notice that the kernel estimator is a linear smoother with weights

$$w_i(x_0) = \frac{K(\frac{x_0 - x_i}{h})}{\sum_{j=1}^{n} K(\frac{x_0 - x_i}{h})}.$$

A noticeable shortcoming of the kernel smoothing is that it suffers from poor bias at the boundaries of the domain of $x_1, ..., x_n$ arising from the asymmetry of w_i in these regions. To overcome this limitation we can move from a local constant fit to a local higher-order fit. This can be done by local polynomials presented in the next section.

2.4.2 Local polynomials

Due to the bias present in the boundaries of $x_1, ..., x_n$ using the kernel smoother a first-order correction can alleviate the issue by employing the estimate $\hat{f}(x) = \hat{\alpha} + \hat{\beta}x$, where $\hat{\alpha}$ and $\hat{\beta}$ are such that

$$\min_{\alpha(x_0), \beta(x_0)} \sum_{i=1}^{n} K\left(\frac{x_0 - x_i}{h}\right) (y_i - \alpha(x_0) - \beta(x_0)x_i)^2.$$

This is the local linear regression and it can be shown that it is also linear in the observations $\{y_i\}_{i=1}^n$. In addition, we do not necessarily need to stop at linear fits, but we can move to higher orders and fit $\hat{f}(x_0) = \hat{\beta}_0(x_0) + \sum_{j=1}^{p} \hat{\beta}_j(x_0)x^j$.

2.4.3 Smoothing splines

Smoothing splines are a popular tool and have been studied both in computational and theoretical terms, see de Boor (1978), Wahba (1990) or Green and Silverman

(1994). These estimators perform a regularised regression over the natural spline basis without the need to select knots, but rather they place them at all inputs $x_1, ..., x_n$. A natural spline of order k with knots at $t_1 < ... < t_m$ is a piecewise polynomial function f such that i. f reduces to a polynomial of degree k on each of $[t_1, t_2], ..., [t_{m-1}, t_m]$ ii. f reduces to a polynomial of degree $(k-1)/2$ on $[-\infty, t_1]$ and $[t_m, \infty]$ (and, hence, natural splines are only defined for odd order k) iii. f is continuous and has continuous derivatives of orders $1, ..., k-1$ at its knot points.

For a given order k the smoothing spline estimate $\hat{f}$ is defined as

$$\hat{f} = \arg\min_f \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int \{f''(x)\}^2 \mathrm{d}x$$

where λ is a fixed smoothing parameter which controls the curvature in the function $f(x_i)$. A noticeable result here is that smoothing splines are also linear smoothers since the problem above can be re-parameterised and can be written as

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^n} \|y - \mathbf{N}\beta\|_2^2 + \lambda \beta^T \mathbf{\Omega} \beta$$

where $\{\mathbf{N}\}_{ij} = N_j(x_i)$, $N_j(x)$ are a set of basis functions for natural splines with knots over $x_1, ..., x_n$ and $\{\mathbf{\Omega}\}_{jk} = \int N_j''(t) N_k''(t) \mathrm{d}t$. Since this is a generalised ridge regression the solution is given by

$$\hat{\beta} = (\mathbf{N}^T \mathbf{N} + \lambda \mathbf{\Omega})^{-1} \mathbf{N}^T y$$

which is linear in the observations $\{y\}_{i=1}^n$.

Hastie and Tibshirani (1993) extend smoothing splines to regression models with varying coefficients which are the focus of Chapter 5. In the same chapter, we compare the performance of a new estimator with a total variation penalty with smoothing splines and we show that the former outperforms the latter in many cases using

both simulated and real data. In the next section, we review two methods in non-parametric regression that adopt total variation penalties.

2.4.4 Trend filtering and locally adaptive regression splines

Tibshirani (2014) proposes a new class of estimators for non-parametric regression, termed trend filtering. This term was first used by Kim et al. (2009a), but the authors focus mainly on piecewise linear estimation. For a given integer $k \geq 0$, Tibshirani (2014) considers the following penalised least squares optimisation problem

$$\hat{\beta} \in \arg\min_{\beta \in \mathbb{R}^n} \|y - \beta\|_2^2 + \lambda ||D^{(k+1)}\beta||_1 \tag{2.18}$$

where λ is a tuning parameter and $D^{(k+1)} \in \mathbb{R}^{(n-k)\times n}$ is the discrete difference operator of order $k+1$. When $k = 0$ then $\left\|D^{(1)}\beta\right\|_1 = \sum_{t=1}^{n} |\beta_t - \beta_{t-1}|$ which is the 1-dimensional total variation denoising (see Rudin et al. (1992)) or the 1-dimensional FLSA of Tibshirani et al. (2005), already mentioned in the context of variable selection in Section 2.2.4. For $k \geq 0$, the operator $D^{(k+1)} \in \mathbb{R}^{(n-k)\times n}$ is recursively defined

$$D^{(k+1)} = D^{(1)} D^{(k)}.$$

Hence, the matrix $D^{(k+1)}$ can be seen as the discrete analogy to the $(k+1)$st order derivative operator and the penalty term in (2.18) penalises the changes in the discrete kth derivative of β.

In addition, Tibshirani (2014) shows that the trend filtering achieves the same minimax rate with the locally adaptive regression splines of Mammen and van de Geer (1997), a total variation type of estimator. He shows that the two methods are equivalent when $k = 0$ (piecewise constant) or $k = 1$ (piecewise linear), but trend

filtering has a computational advantage over locally adaptive regression splines when $k \geq 2$.

In Chapter 4, we show a fast way of implementing the solution path algorithm of Tibshirani and Taylor (2011) for the case $k = 0$. Further, we provide a theoretical result for the consistency of the trend filtering (equivalently, the locally adaptive regression splines) when $k = 0$. Finally, in Chapter 5 we extend trend filtering to estimating regression models with varying coefficients as an alternative to the smoothing splines.

2.4.5 Wavelet smoothing

Wavelet methods have been widely employed for non-parametric regression and they perform well in cases where the signal has spatially heterogeneous degree of smoothness, for example, it can be "wiggly" in some regions of a signal and piecewise constant in some others. The local adaptivity of wavelet smoothing is attributed to the fact that it selects a sparse wavelet coefficient vector by shrinking coefficients that are zero or close to zero. This is termed wavelet shrinkage, first introduced to the statistical literature by Donoho (1993), Donoho (1995), Donoho and Johnstone (1994) and Donoho et al. (1995). Wavelet shrinkage has been shown to work well with correlated data (see Figure 2.4 for an illustration of the whitening property of wavelets), non-Gaussian error (see Averkamp and Houdré (2003) and references therein) or irregularly spaced data in the sense that $x_i \neq i/n$ (see Nunes et al. (2006) and references therein).

The first step is to form a wavelet basis matrix $\mathcal{W} \in \mathbb{R}^{n \times n}$ and then perform a

Discrete Wavelet Transform (DWT) of the outputs $\{y_i\}_{i=1}^n$ from the model (2.17)

$$d_{j,r} = \theta_{j,r} + \epsilon_{j,r}$$

where $d = \mathcal{W}y$. Note that $\epsilon_{j,r}$ are still iid Gaussian due to the orthonormality of the DWT.

The next step is to threshold the vector d

$$\hat{d} = \mathcal{T}_\lambda(d)$$

for some $\lambda > 0$, such that some of the coefficients $d_{j,r}$ are shrunk towards zero. The hope is that with an appropriate threshold some of the coefficients of the vector $\hat{d}$ will be more significant indicating an irregularity in the function f. Those $d_{j,r}$ that are zero or close to zero correspond to regions where is f smooth and are set equal to zero.

Finally, the method involves an inverse wavelet transform of $\hat{d}$

$$\hat{f} = \mathcal{W}^T\hat{d}$$

to obtain an estimate of the function f.

Donoho and Johnstone (1994) propose the hard-thresholding

$$\mathcal{T}_\lambda(d_{j,r}) = d_{j,r}\mathbb{I}(|d_{j,r}| > \lambda)$$

and the soft-thresholding

$$\mathcal{T}_\lambda(d_{j,r}) = \text{sign}(d_{j,r})\max(|d_{j,r}| - \lambda, 0),$$

making use of the universal threshold $\lambda^\star = \sigma\sqrt{2\log(n)}$ where σ is unknown, but it can be estimated through the Median Absolute Deviation of the sequence $\left|\frac{X_{i+1}-X_i}{\sqrt{2}}\right|_{i=1}^{n-1}$.

It is particularly interesting to see that the wavelet shrinkage estimate is the solution to the lasso problem

$$\min_{d\in\mathbb{R}^n} ||y - \mathcal{W}d|| + \lambda||d||_1$$

and since $\mathcal{W}$ is orthonormal then the lasso estimates are obtained from soft-thresholding.

Finally, we note that the DWT (Mallat (1989)) is a fast decomposition and reconstruction algorithm for discrete data, analogous to the Fast Fourier Transform of Cooley and Tukey (1965). It produces a vector of wavelet coefficients at dyadic scales and locations without involving matrix multiplication hence its complexity is $\mathcal{O}(n)$ and not $\mathcal{O}(n^2)$ where n is the length of the input vector.

It is important to notice that in Chapter 4 we use similar ideas to overcome the matrix multiplications in the algorithm by Tibshirani and Taylor (2011) resulting in lower complexity.

2.4.6 Methods for piecewise constant estimation

An important class of non-linear estimators are the piecewise constant estimators which have been shown to approximate a wide range of function spaces (DeVore (1998)) well. This means that the underlying function f_0 in the model (2.17) may belong to different smoothness classes, including the case where f_0 is discontinuous, e.g. piecewise constant. In this thesis we mainly focus on piecewise stationarity and we review estimation methods in the next lines.

We consider the following model

$$X_t = f_t + \epsilon_t \quad \text{for} \quad t = 1, ..., n \tag{2.19}$$

where f_t is a piecewise constant and deterministic function with N change-points at

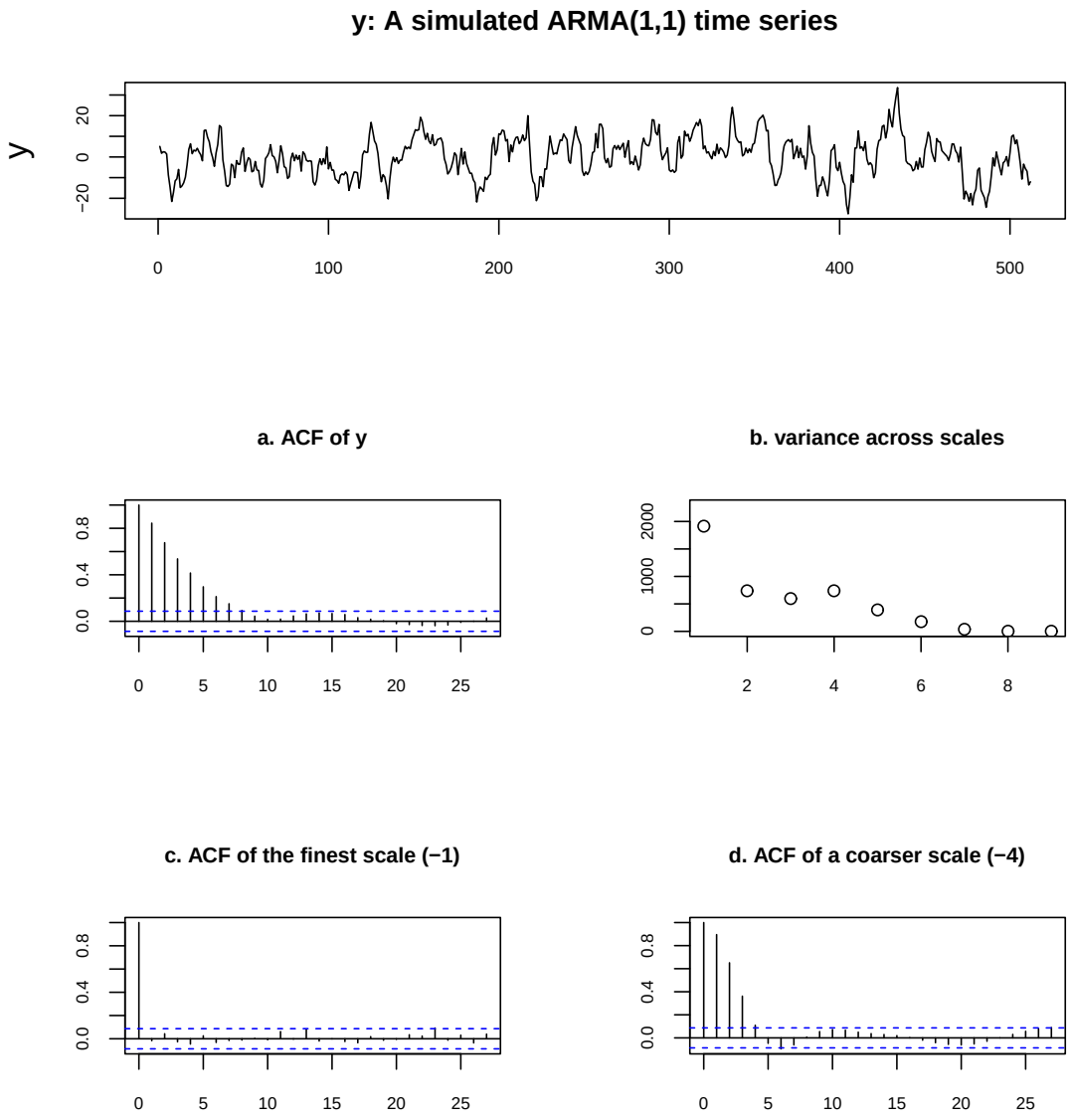

Figure 2.4: Whitening property of the wavelet transformation for an ARMA(1,1) process. The acf of the ARMA(1,1) process (panel a.) indicates high autocorrelation which decays slowly. Panels c. and d. are the acf of the finest and a coarser scale DWT showing significantly reduced autocorrelations.

locations $\eta = \{\eta_1, ..., \eta_N\}$. Both N and the locations of change-points are unknown to the user and need to be estimated.

One branch of change-point estimators are formulated as multivariate optimisa-

tion problems, i.e.

$$\min_{\eta} J(\eta, X_t) + \lambda \text{pen}(\eta)$$

where $J(\eta, X_t)$ is a measure of fit, also termed cost or contrast function. For the cost function $J(\eta, X_t)$, the least squares or twice the negative log-likelihood (Chen and Gupta (2011)) are typically used. In addition, the penalty function $\text{pen}(\eta)$ depends on the number N of change-points and a linear relation, i.e. $\text{pen}(\eta) = N$, is commonly used. This, in practice, is similar to using an information criterion such as AIC (Akaike's IC) where $\lambda = 2\bar{N}$ and $\bar{N}$ are the additional change-points. Another example of $\text{pen}(\eta)$ is when $\lambda = \bar{N} \log n$ which coincides with the Schwarz Information Criterion (SIC or BIC). Hence, the purpose of the penalty $\lambda \text{pen}(\eta)$ is to control for over-fitting. Techniques that involve minimisation of penalised functions have been proposed by Yao (1988), Braun et al. (2000), Auger and Lawrence (1989) and Killick et al. (2012) among others.

A different route in the estimation of the model (2.19) using penalised regression is through the use of L_1 penalties. In Section 2.2.4 we discussed the fused lasso method. A different approach to solving the problem (2.10) is to transform it into a lasso one. Harchaoui and Lévy-Leduc (2010) choose to solve the following problem

$$\arg\min_{\beta \in \mathbb{R}^n} \frac{1}{n} (Y_i - (X_n \beta)_i)^2 \quad \text{s.t.} \quad \sum_{i=1}^{n} |\beta_i| \leq N J^{\star}_{max} \tag{2.20}$$

where X_n is a $n \times n$ lower triangular matrix with nonzero elements equal to one, N is the true number of change-points, and $J^{\star}_{max}$ is the maximum distance between two consecutive change-points. Since the number of change-points is not known, the authors impose an upper bound for N.

On the other hand, the estimation of change-points can be formulated as a prob-

lem of minimising a series of univariate cost functions i.e. detecting a single change-point and then progressively moving to identify more. The Binary Segmentation method belongs to this category and we discuss it in the next section.

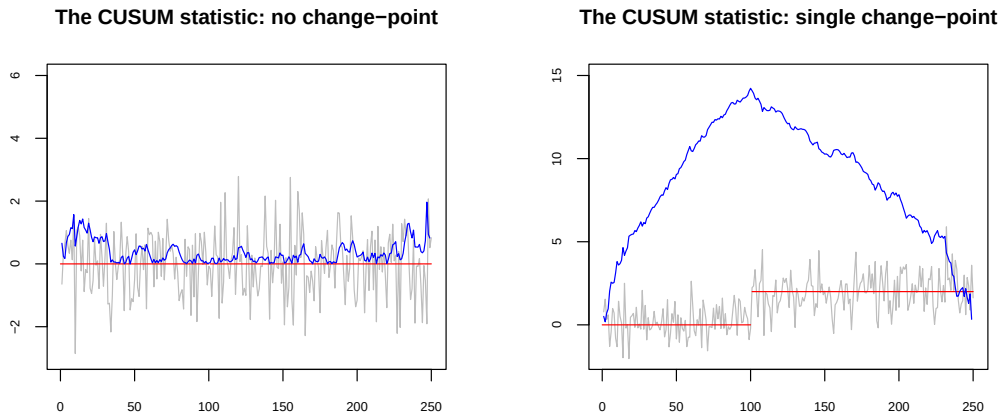

Figure 2.5: The CUSUM statistic (blue lines) applied to a noisy signal with no change-points (left) and a single change-point (right). The underlying signals are shown in red. When there is no change-point the CUSUM looks "flatter" while a peak is formed in the case of a change.

2.4.6.1 Binary Segmentation

The Binary Segmentation (BS) method (Vostrikova (1981)), is a generic technique where the change-point detection starts with a single change-point b, using, for example, the following *Cumulative Sum* statistic (henceforth, CUSUM)

$$\tilde{X}^b_{s,e} = \sqrt{\frac{e-b}{\bar{n}(b-s+1)}} \sum_{t=s}^{b} X_t - \sqrt{\frac{b-s+1}{\bar{n}(e-b)}} \sum_{t=b+1}^{e} X_t \tag{2.21}$$

where $s = 1$, $e = n$ and $\bar{n} = e - s + 1$.

The intuition of the CUSUM is that it computes a statistic sequentially as a difference of two weighted sums (the left and right segment with varying size). At

the point of change, say b, the CUSUM statistic takes its maximum value in absolute terms, see also Figure 2.5. If the obtained statistic $\tilde{X}^b_{s,e}$ is larger than a threshold ζ_n then we conclude that a change-point has occurred.

Now, BS continues on the left and on the right of b until no further change-points are detected. This "greedy" approach is, perhaps, the most widely used change-point search method (Killick et al. (2012)) and the main reasons are the simplicity in implementing it and its low complexity $\mathcal{O}(n \log n)$. In addition, it has found many applications in other setting such as in the multiple detection of change-point in variance (Inclan and Tiao (1994)), in autocovariance (Cho and Fryzlewicz (2012)), or in the conditional variance (Fryzlewicz and Subba Rao (2013)).

The BS method may be unsuitable in cases where the change-points occur close to each other and particularly if the minimum spacing between them is of order $\mathcal{O}(n^{3/4})$ only then BS is consistent in the number and locations of the change-points (Fryzlewicz (2014)). In particular, Venkatraman (1992) shows that the BS method fails to detect change-points that are not separated by at least $n^{1/2}$ observations, see Figure 2.6 for an illustration of this argument (BS estimates are shown by a blue dotted line).

Fryzlewicz (2014) attempts to eliminate this weakness by proposing a randomised binary segmentation, termed Wild Binary Segmentation (WBS), where the search for change-points proceeds by using the CUSUM statistic in smaller segments. To put it simply, at the initiation of the search the CUSUM (2.21) is not calculated globally ($s = 1$ and $e = n$), but rather over multiple sub-samples such that $1 \leq s < e \leq n$. It is therefore expected that CUSUMs with starting and ending points within a short distance from a certain change-point will be more alert in identifying it. In order to

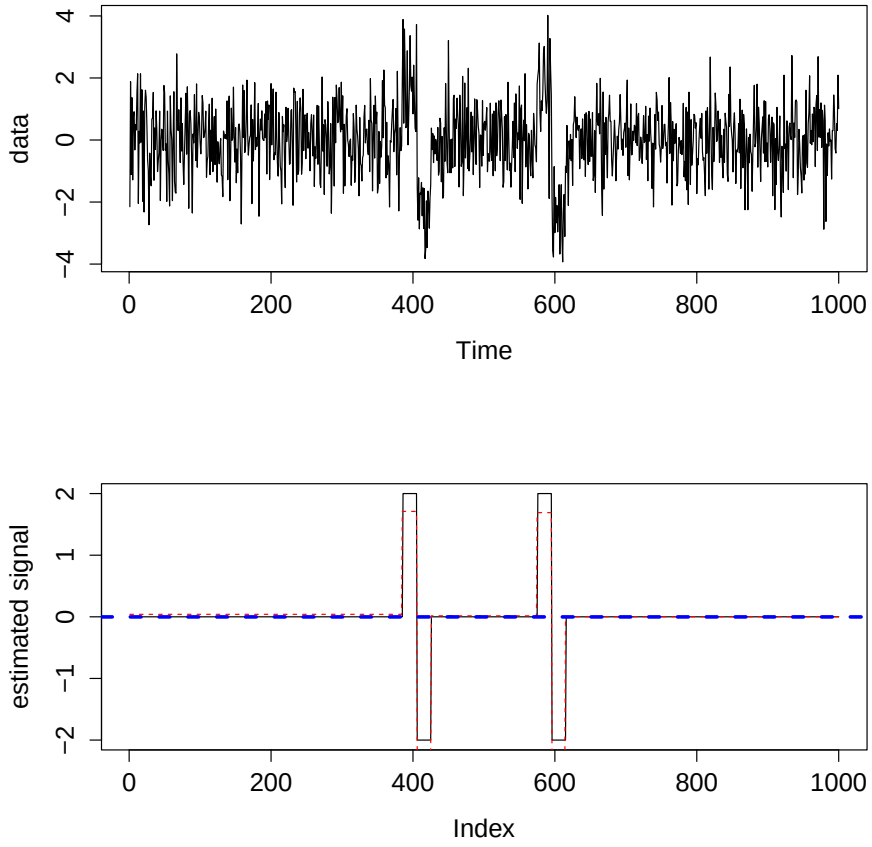

Figure 2.6: A typical example of the BS (blue dotted line) method failing to detect change-points within short distance. The WBS method (red dotted line) detects all six change-points.

avoid the restriction from choosing a window or span parameter the author randomly selects the starting and ending points with the hope that with a high probability a favourable interval with a single change-point will be found. Finally, the method inherits the main feature of the BS search, i.e. after identifying a change-point the problem is divided into to two sub-problems where for each segment we again test for further change-points.

We note that another attempt to improve the performance of BS is found in Olshen et al. (2004). The authors suggest the Circular Binary Segmentation (CBS)

which requires a choice of window for larger data sets and thus making it less user-friendly. In addition, the CBS method involves a permutation approach making it computationally prohibitive for large samples. A faster CBS is proposed by Venkatraman and Olshen (2007), however, the authors notice that it comes with a loss in accuracy. This is due to the approximation of the P-value, used in deciding on the existence of a change-point, which does not affect the estimated locations of the change-points but can result in fewer detected change-points.

In Chapter 3, we adopt the WBS method in order to estimate the number and locations of the change-points in a non-stationary time series motivated by its good practical performance in the simplest model (2.19).

Chapter 3

Multiple change-point detection for non-stationary time series using Wild Binary Segmentation

Introduction

The assumption of stationarity has been the dominant framework for the analysis of many real data. However, in practice, time series entail changes in their dependence structure and therefore modelling non-stationary processes using stationary methods to capture their time-evolving dependence aspects will most likely result in a crude approximation. As pointed out by Mercurio and Spokoiny (2004) the risk of fitting a stationary model to non-stationary data can be high in terms of prediction and forecasting. Many examples of non-stationary data exist; for example, in biomedical signal processing of electroencephalograms (EEG) see Ombao et al. (2001); in audio signal processing see Davies and Bland (2010); in finance see Stărică and Granger

(2005); in oceanography see Killick et al. (2013), to name but a few. In this chapter we deal with piecewise stationarity, arguably the simplest type of deviation from stationarity. This implies a time-varying process where its parameters evolve through time but remain constant for a specific period of time.

The problem of change-point estimation has attracted significant attention. A branch of the literature deals with the estimation of a single change-point (for a change in mean see e.g. Sen and Srivastava (1975); for time series see Davis et al. (1995), Gombay (2008), Gombay and Serban (2009) and references therein) while another extends it to multiple change-points with many changing parameters such as Ombao et al. (2001) who divide a time series into dyadic segments and choose the one with the minimum cost. The latter branch can be further categorised. On the one hand, the multiple change-point estimation can be formulated through an optimisation task i.e. minimising a multivariate cost function (or criterion). When the number of change-points N is unknown then a penalty is typically added e.g. the Schwarz criterion (see Yao (1988)). In addition, the user can adopt certain cost functions to deal with the estimation of specific models: the least-squares for change in the mean of a series (Yao and Au (1989) or Lavielle and Moulines (2000)), the Minimum Description Length criterion (MDL) for non-stationary time series (Davis et al. (2006)), the Gaussian log-likelihood function for changes in the volatility (Lavielle and Teyssiere (2007)) or the covariance structure of a multivariate time series (Lavielle and Teyssiere (2006)).

Several algorithms for minimising a cost function are based on dynamic programming (Bellman and Dreyfus (1966) and Kay (1998)) and they are often used in solving change-point problems, see e.g. Perron (2006) and references therein. Auger

and Lawrence (1989) propose the Segment Neighbourhood method with complexity $\mathcal{O}(QT^2)$ where Q is the maximum number of change-points. An alternative method is the exact method of Optimal Partitioning by Jackson et al. (2005), but its complexity of $\mathcal{O}(T^2)$ makes it suitable for smaller samples.

Change-point estimators that adopt a multivariate cost function often come with a high computational cost. An attempt to reduce the computational burden is found in Killick et al. (2012) who extend the Optimal Partitioning method of Jackson et al. (2005) (termed PELT) and show that the computational cost is $\mathcal{O}(T)$ when the number of change-points increases linearly with T. Another attempt is found in Davis et al. (2006) and Davis et al. (2008) who suggest a genetic algorithm to detect change-points in a piecewise-constant AR model or non-linear processes, respectively, where the MDL criterion is used.

On the other hand, the estimation of change-points can be formulated as a problem of minimising a series of univariate cost functions i.e. detecting a single change-point and then progressively moving to identify more. The Binary Segmentation method (BS) belongs to this category and uses a certain test statistic (such as the CUSUM) to reject the null hypothesis of no change-point. The BS has been widely used and the main reasons are its low computational complexity and the fact that it is conceptually easy to implement: after identifying a change-point the detection of further change-points continues to the left and to the right of the initial change-point until no further changes are found.

The BS method has been adopted to solve different types of problems. Inclan and Tiao (1994) detect breaks in the variance of a sequence of independent observations; Berkes et al. (2009) use a weighted CUSUM to reveal changes in the mean or the

covariance structure of a linear process; Lee et al. (2003) apply the test in the residuals obtained from a least squares estimator; and Kim et al. (2000) and Lee and Park (2001) extend Inclan and Tiao (1994) method to a GARCH(1,1) model and linear processes, respectively. A common factor of most of these methods is the estimation of the long-term variance or autocovariance; a rather difficult task when the observations are dependent. Cho and Fryzlewicz (2012) apply the binary segmentation method on the wavelet periodograms with the purpose to detect change-points in the second-order structure of a non-stationary process. Using the wavelet periodogram, Killick et al. (2013) propose a likelihood ratio test under the null and alternative hypotheses. The authors apply the binary segmentation algorithm but assume an upper bound for the number of change-points. Fryzlewicz and Subba Rao (2013) adopt the binary segmentation search to test for multiple change-points in a piecewise constant ARCH model. BS is also used for multivariate (possibly high-dimensional) time series segmentation in Cho and Fryzlewicz (2013) and in Schröder and Fryzlewicz (2013) in the context of trend detection for financial time series.

In this chapter we develop a detection method to estimate the number and locations of change-points for a piecewise stationary time series model using the non-parametric Locally Stationary Wavelet (LSW) process of Nason et al. (2000). The LSW model provides a complete description of the second-order structure of a stochastic process and, hence, it permits a fast estimation of the local autocovariance through the evolutionary wavelet spectrum. This choice, however, should not be seen as a restriction and potentially other models can form the basis for our algorithm.

In order to implement the change-point detection we adopt the Wild Binary Segmentation (WBS) method proposed in the signal+iid Gaussian noise set-up by

Fryzlewicz (2014) which attempts to overcome the limitations of the BS method. Under specific setups where many change-points are present the BS search may be inefficient in detecting them. This stems from the fact that the BS starts its search assuming a single change-point. To correct this limitation, Fryzlewicz (2014) proposes the WBS algorithm that involves a "certain random localisation mechanism". His method can be summarised as follows. At the beginning of the algorithm the CUSUM statistic is not calculated over the entire set $\{1, ..., T\}$ where T is the sample size but only over M local segments $[s, e]$. The starting s and ending e points are randomly drawn from a uniform distribution $U(1, T)$ and the hope is that for a large enough M a specific $[s, e]$ will contain a single change-point. The method then proceeds similarly to BS: if the obtained CUSUM statistic exceeds a threshold then it is deemed to be a change-point and the procedure continues to its left and right.

To summarise, our contribution in this work is twofold: i. to adopt the WBS technique to the segmentation of a piecewise stationary time series and ii. to propose a method of combining the estimated change-points across wavelet periodogram scales. The chapter is structured as follows: in Section 3.1 we present and review the WBS algorithm in the context of time series. The reasons for selecting the LSW model as the core of our detection algorithm are given in Section 3.2. The main algorithm is presented in Section 3.3 along with its theoretical consistency in estimating the number and locations of change-points. In addition, we conduct an extensive simulation study to examine the performance of the algorithm; the results are given in Section 3.4. In Section 3.5 we apply our method to two real datasets. Finally, proofs of the theorems related to our method are in Section 3.6.

3.1 The Wild Binary Segmentation Algorithm

The BS algorithm for a stochastic process was first introduced by Vostrikova (1981) who showed its consistency for the number and locations of change-points for a fixed N. A proof of its consistency is also given by Venkatraman (1992) for the Gaussian function+noise model, though the rates for the locations of the change-points are suboptimal. Improved rates of convergence of the locations of the change-points for the BS method are given by Fryzlewicz (2014).

As a preparatory exercise before considering segmentation in the full time series model (3.7) we first examine the following multiplicative model

$$Y_{t,T}^2 = \sigma_{t,T}^2 Z_{t,T}^2, \ t = 0, ..., T-1 \tag{3.2}$$

where $\sigma_{t,T}^2$ is a piecewise constant function and the series $Z_{t,T}$ are possibly autocorrelated standard normal variables. This generic set-up is of interest to us because the wavelet periodogram, used later in the segmentation of (3.7), follows model (3.2).

A potential change-point b_0 on a segment $[s, e]$ is given by

$$b_0 = \arg\max_b \left| \tilde{Y}_{s,e}^b / q_{s,e} \right|$$

where $\tilde{Y}_{s,e}^b$ is the CUSUM statistic

$$\tilde{Y}_{s,e}^b = \sqrt{\frac{e-b}{n(b-s+1)}} \sum_{t=s}^{b} Y_t^2 - \sqrt{\frac{b-s+1}{n(e-b)}} \sum_{t=b+1}^{e} Y_t^2, \tag{3.3}$$

$q_{s,e} = \sum_{t=s}^{e} Y_t^2/n$ and $n = e - s + 1$.

The value $|\tilde{Y}_{s,e}^{b_0}/q_{s,e}| = \max_b |\tilde{Y}_{s,e}^b/q_{s,e}|$ will be tested against a threshold ω_T in order to decide whether the null hypothesis of no change-point is rejected or not. The BS proceeds by recursively applying the above CUSUM on the two, newly-created segments defined by the already detected b_0, i.e $[s, b_0]$ and $[b_0 + 1, e]$. The

algorithm stops in each current interval when no further change-points are detected, that is, the obtained CUSUM values fall below threshold ω_T.

The BS method has the disadvantage of possibly fitting the wrong model when multiple change-points are present as it searches the whole series. The CUSUM formula (3.3) can result in spurious change-points when e.g. the true change-points occur close to each other. This is due to the fact that the BS method begins by assuming a single change-point exists in the series and, hence, the CUSUM statistic looks flatter. Especially, the BS method can fail to detect a small change in the middle of a large segment (Olshen et al. (2004)) which is illustrated in Fryzlewicz (2014).

Fryzlewicz (2014) proposes a randomised binary segmentation (termed Wild Binary Segmentation - WBS) where the search for change-points proceeds by calculating the CUSUM statistic in smaller segments whose length is random. By doing so, it is expected that CUSUMs with starting and ending points within a short distance from the change-points will be more alert in identifying them. Ideally, an interval over which the CUSUM statistic is maximised over a large collection of random intervals should contain a single change-point. Since the number and location of the change-points are unknown, Fryzlewicz (2014) suggests to take multiple random intervals such that with high probability a favourable interval with a single change-point is found (see Figure 3.1). The binary segmentation procedure is not altered, meaning that after identifying a change-point the problem is divided into two sub-problems where for each segment we again test for further change-points. Finally, the computational complexity of the method can be reduced by noticing that the randomly drawn intervals and their corresponding CUSUM statistics can be calculated once

at the start of the algorithm. Then, as the algorithm proceeds at a generic segment $[s, e]$, the obtained statistics can be retrieved making sure the random starting and end points fall within $[s, e]$.

The main steps of the WBS algorithm modified for the model (3.2) are outlined below.

- Calculate the CUSUM statistics over a collection of random intervals $[s_m, e_m]$. The starting and ending points are not fixed but are sampled from a uniform distribution with replacement making sure that

$$e_m \geq s_m + \Delta_T \tag{3.4}$$

where $\Delta_T > 0$ defines the minimum size of the interval drawn.

Denote with $\mathcal{M}_{s,e}$ the set of all random intervals $[s_m, e_m]$ where $m = 1, ..., M$ such that $[s_m, e_m] \subseteq [s, e]$; then the likely location of a change-point is

$$(m_0, b_0) = \underset{(m \in \mathcal{M}_{s,e}, b \in s_m, ..., e_m - 1)}{\arg\max} \left| \tilde{Y}^b_{s_m, e_m} / q_{s_m, e_m} \right| \tag{3.5}$$

such that

$$\max \left(\frac{e_{m_0} - b_0}{e_{m_0} - s_{m_0} + 1}, \frac{b_0 - s_{m_0} + 1}{e_{m_0} - s_{m_0} + 1} \right) \leq c_\star \tag{3.6}$$

where $c_\star$ is a constant satisfying $c_\star \in [2/3, 1)$. The conditions (3.4) and (3.5) do not appear in the original work by Fryzlewicz (2014), but they are necessary since the assumption of an iid Gaussian process does not hold for the model (3.2).

- The obtained CUSUM values are rescaled and tested against a threshold ω_T. This will ensure that with a high probability only the significant change-points will survive. The choice of the threshold ω_T is discussed in Section 3.3. If the

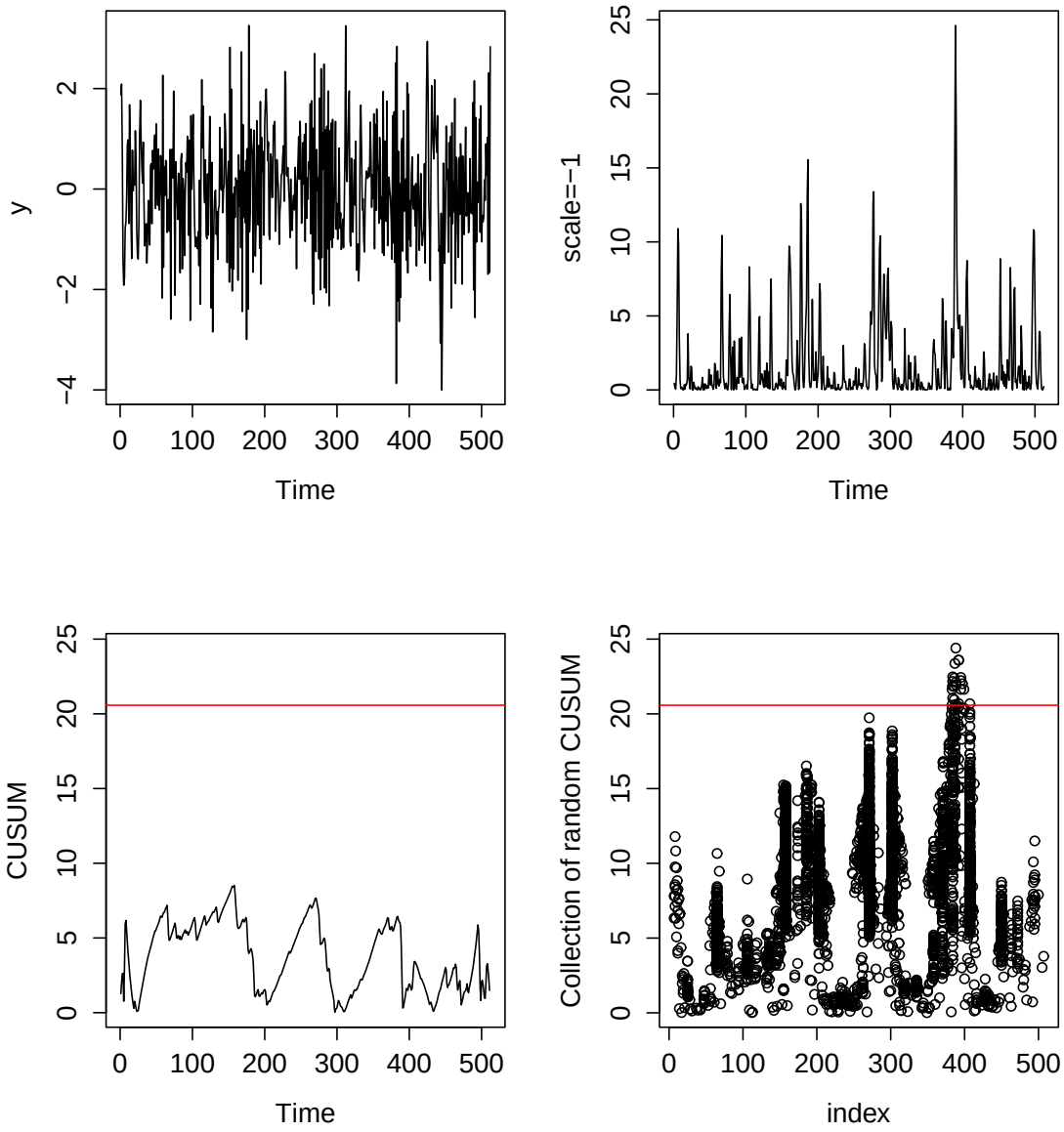

Figure 3.1: A simulated series (top-left) of an AR(1) model $y_t = \phi_t y_{t-1} + \varepsilon_t$ *with* $\phi_t = (0.5, 0.0)$ *and change-points at* $\{50, 100, ..., 450\}$. *The Wavelet Periodogram at scale* -1 *(top-right). The CUSUM statistic of scale* -1 *(bottom-left) as in the BS method; the red line is the threshold defined in the main algorithm, i.e.* $C\log(T)$. *The CUSUM statistics with random sample sizes (bottom-right) as in the WBS method; the red line is the same threshold.*

obtained CUSUM statistic is significant then the search is continued to the left and to the right of b_0; otherwise the algorithm stops. This step differs from the original WBS method of Fryzlewicz (2014) in that the CUSUM statistics are rescaled using q_{s_m,e_m} so that ω_T not to depend on $\sigma^2_{t,T}$.

3.2 Locally Stationary Wavelets and the Multiplicative Model

The LSW process enables a time-scale decomposition of a process and thus permits a rigorous estimation of the evolutionary wavelet spectrum and the local autocovariance and can be seen as an alternative to the Fourier based approach for modelling time series. We refer the reader to Definition 2.4 in Chapter 2 for more discussion on LSW since here we are interested in non-stationary processes whose second-order structure is piecewise constant and therefore, we use the definition of the LSW from Cho and Fryzlewicz (2012): a triangular stochastic array $\{X_{t,T}\}_{t=0}^{T-1}$ for $T = 1, 2, ...$, is in a class of Locally Stationary Wavelet (LSW) processes if there exists a mean-square representation

$$X_{t,T} = \sum_{i=-\infty}^{-1} \sum_{k=-\infty}^{\infty} W_i(k/T)\psi_{i,t-k}\xi_{i,k} \tag{3.7}$$

with $i \in -1, -2, ...$ and $k \in \mathbb{Z}$ are, respectively, scale and location parameters, $(\psi_{i,0}, ..., \psi_{i,\mathcal{L}-1})$ are discrete, real-valued, compactly supported, non-decimated wavelet vectors with support length $\mathcal{L} = O(2^{-i})$, and the $\xi_{i,k}$ are zero-mean, orthonormal, identically distributed random variables. In this set-up we replace the Lipschitz-continuity constraint on $W_i(z)$ by the piecewise constant constraint, which allows us to model a process whose second-order structure evolves in a piecewise constant

manner over time with a finite but unknown number of change-points. Let L_i be the total magnitude of change-points in $W_i^2(z)$, then the functions $W_i(z)$ satisfy

- $\sum_{i=-\infty}^{-1} W_i^2 < \infty$ uniformly in z
- $\sum_{i=-I}^{-1} 2^{-i} L_i = \mathcal{O}(\log T)$ where $I = \log_2 T$.

The simplest type of a wavelet system that can be used in formula (3.7) are the Haar wavelets. Specifically,

$$\psi_{i,k} = 2^{i/2}\mathbb{I}_{0,\ldots,2^{-j-1}-1}(k) - 2^{i/2}\mathbb{I}_{2^{-j-1},\ldots,2^{-i}-1}(k)$$

for $i = -1, -2, ..., k$ $k \in \mathbb{Z}$ where $\mathbb{I}_A(k)$ is 1 if $k \in A$ and 0 otherwise. Further, small absolute values of the scale parameter i denote "fine" scales, while large absolute values denote "coarser" scales. In fine scales the wavelet vectors are most oscillatory and localised. On the contrary, coarser scales have longer, less oscillatory wavelet vectors.

Throughout this chapter, we assume that $\xi_{i,k}$ are distributed as $\mathcal{N}(0, 1)$ and this leads to $X_{t,T}$ being Gaussian itself. In addition, the choice of the Haar wavelets is natural given that the second-order structure of the non-stationary processes we consider in this chapter evolves over time in a piecewise constant manner.

Of main interest in the LSW set-up is the Evolutionary Wavelet Spectrum (EWS) $S_i(z) = W_i^2(z)$, $i = -1, -2, ...$, defined on the rescaled-time interval $z \in [0, 1]$. The estimation of the EWS is done through the wavelet periodogram (Nason et al. (2000)) and its definition is given below:

Definition: Let $X_{t,T}$ be an LSW process constructed using the wavelet system ψ. The triangular stochastic array

$$I_{t,T}^{(i)} = \left| \sum_s X_{s,T} \psi_{i,s-t} \right|^2 \tag{3.8}$$

is called the wavelet periodogram of $X_{t,T}$ at scale i.

We also recall two further definitions from Nason et al. (2000): the autocorrelation wavelets $\Psi_i(\tau) = \sum_k \psi_{i,k}\psi_{i,k-\tau}$ and the autocorrelation wavelet inner product matrix $A_{i,k} = \sum_\tau \Psi_i(\tau)\Psi_k(\tau)$. Fryzlewicz and Nason (2006) show that $\mathbb{E}I_{t,T}^{(i)}$ is "close" (in the sense that the integrated squared bias converges to zero) to the function $\beta_i(z) = \sum_{j=-\infty}^{-1} S_j(z)A_{i,j}$, a piecewise constant function with at most N change-points, whose set is denoted by $\mathcal{N}$. Every change-point in the autocovariance structure of the time series results in a change-point in at least one of the $\beta_i(z)$; therefore, detecting a change-point in the wavelet periodogram implies a change-point in the autocovariance structure of the process.

In addition, note that each wavelet periodogram ordinate is a squared wavelet coefficient of a standard Gaussian time series and it satisfies

$$I_{t,T}^{(i)} = \mathbb{E}I_{t,T}^{(i)} Z_{t,T}^2 \tag{3.9}$$

where $\{Z_{t,T}\}_{t=0}^{T-1}$ are autocorrelated standard normal variables (or equivalently the distribution of the squared wavelet coefficient $I_{t,T}^{(i)}$ is that of a scaled χ_1^2 variable). Then, the quantities $I_{t,T}^{(i)}$ and $\mathbb{E}I_{t,T}^{(i)}$ can be seen as special cases of $Y_{t,T}^2$ and $\sigma_{t,T}^2$ respectively of the multiplicative model (3.2). To enable the application of the model (3.9) in this context, we assume the following condition:

(A0): $\sigma_{t,T}^2$ is deterministic and "close" to a piecewise constant function $\sigma^2(t/T)$ (apart from intervals around the discontinuities in $\sigma^2(t/T)$ which have length at most

$K2^{-i}$) in the sense that $T^{-1}\sum_{t=0}^{T-1}|\sigma^2_{t,T}-\sigma^2(t/T)|^2 = o(\log^{-1}T)$ where the rate of convergence comes from the integrated squared bias between $\beta_i(t/T)$ and $\mathbb{E}I^{(i)}_{t,T}$ (see Fryzlewicz and Nason (2006)).

3.3 The Algorithm

In this section we present the WBS algorithm within the framework of the LSW model. First, we form the following CUSUM-type statistic

$$\mathbb{Y}^{b(i)}_{s_m,e_m} = \sqrt{\frac{e_m-b}{n(b-s_m+1)}}\sum_{t=s_m}^{b} I^{(i)}_{t,T} - \sqrt{\frac{b-s_m+1}{n(e_m-b)}}\sum_{t=b+1}^{e_m} I^{(i)}_{t,T} \tag{3.10}$$

where the subscript $(.)_m$ denotes an element chosen randomly from the set $\{1,...,T\}$ as in (3.4), $n = e_m - s_m + 1$ and $I^{(i)}_{t,T}$ are the wavelet periodogram ordinates at scale i that form the multiplicative model $I^{(i)}_{t,T} = \mathbb{E}I^{(i)}_{t,T}Z^2_{t,T}$ discussed in Section 3.2. The likely location of a change-point b_0 is then given by (3.5).

The following stages summarise the recursive procedure:

Stage I: Start with $s = 1$ and $e = T$.

Stage II: Examine whether $h_{m_0} = |\mathbb{Y}^{b_0}_{s_{m_0},e_{m_0}}|/q_{s_{m_0},e_{m_0}} > \omega_T = C\log(T)$ where $q_{s_{m_0},e_{m_0}} = \sum_{t=s_{m_0}}^{e_{m_0}} I^{(i)}_{t,T}/n_{m_0}$, $n_{m_0} = e_{m_0} - s_{m_0} + 1$ and m_0, b_0 as in (3.5); C is a parameter that remains constant and only varies between scales. In other words, perform hard-thresholding on h_{m_0}, i.e. $h'_{m_0} = h_{m_0}\mathbb{I}(h_{m_0} > \omega_T)$ where $\mathbb{I}(.)$ is 1 if the inequality is satisfied and 0 otherwise.

Stage III: If $h'_{m_0} > 0$, then add b_0 to the set of estimated change-points; otherwise if $h'_{m_0} = 0$ stop the algorithm.

Stage IV: Repeat stages II-III to each of the two segments $(s,e) = (1,b_0)$ and $(s,e) = (b_0+1,T)$ if their length is more than Δ_T.

The choice of parameters C and Δ_T is described in Section 3.3.4. We note that in addition to the random intervals $[s_m, e_m]$ we also include into $\mathcal{M}_{s,e}$ the segment $[s, e]$. This implies that the BS method is also taken into consideration when calculating the CUSUM statistic and it improves the method in two directions i. even with a small value of M the hope is that the performance of the BS will improve and ii. the BS has better performance when no or only one change-point is present in the current interval.

Further, we expect that finer scales will be more useful in detecting the number and locations of the change-points in $\mathbb{E}I_{t,T}^{(i)}$. This is because as we move to coarser scales the autocorrelation within $I_{t,T}^{(i)}$ becomes stronger and the intervals on which a wavelet periodogram sequence is not piecewise constant become longer. Hence, we select the scale $i < -I^\star$ where $I^\star = \lfloor \alpha \log\log T \rfloor$ and $\alpha \in (0, 3\lambda]$ for $\lambda > 0$ such that the consistency of our method is retained.

In stage II, we rescale the statistic h_{m_0} before we test it against the threshold. This division plays the role of stabilising the variance and, therefore threshold ω_T does not depend on $\sigma^2(z)$ and can be selected more easily.

Finally, we notice that Horváth et al. (2008) propose a similar type of CUSUM statistic which does not require an estimate of the variance of a stochastic process by using the ratio of the maximum of two local means. However, the authors apply the method to detect a single change-point in the mean of a stochastic process under independent, correlated or heteroscedastic error settings.

3.3.1 Technical assumptions and consistency

In this section we present the consistency theorem of the WBS algorithm for the total number N and locations of the change-points $0 < \eta_1 < ... < \eta_N < T-1$ with $\eta_0 = 0$ and $\eta_{N+1} = T$. To achieve consistency, we impose the following assumptions:

(A1): $\sigma^2(t/T)$ is bounded from above and away from zero, i.e. $0 < \sigma^2(t/T) < \sigma^\star < \infty$ where $\sigma^\star \leq \max_{t,T} \sigma^2(t/T)$. Further, the number of change-points N in (3.2) is unknown and allowed to increase with T i.e. only the minimum distance between the change-points can restrict the maximum number of N.

(A2): $\{Z_{t,T}\}_{t=0}^{T-1}$ is a sequence of standard Gaussian variables and the autocorrelation function $\rho(\tau) = \sup_{t,T} |\text{cor}(Z_{t,T}, Z_{t+\tau,T})|$ is absolutely summable, that is it satisfies $\rho_\infty^1 < \infty$ where $\rho_\infty^p = \sum_\tau |\rho(\tau)|^p$.

(A3): The distance between any two adjacent change-points satisfies $\min_{r=1,...,N+1} |\eta_r - \eta_{r-1}| \geq \delta_T$, where $\delta_T \geq C \log^2 T$ for a large enough C.

(A4): The magnitude of the change-points satisfy $\inf_{1\leq r\leq N} |\sigma((\eta_r + 1)/T) - \sigma(\eta_r/T)| \geq \sigma_\star$ where $\sigma_\star > 0$.

(A5): $\Delta_T \asymp \delta_T$ where Δ_T as defined in (3.4).

Theorem 1 *Let $Y_{t,T}^2$ follow model (3.2), and suppose that Assumptions (A1)-(A5) hold. Denote the number of change-points in $\sigma^2(t/T)$ as N and the locations of those change-points as $\eta_1, ..., \eta_N$. Let $\hat{N}$ and $\hat{\eta}_1, ..., \hat{\eta}_N$ be the number and locations of the change-points (in ascending order), respectively, estimated by the Wild Binary Segmentation algorithm. There exist two constants C_1 and C_2 such that if $C_1 \log T \leq \omega_T \leq C_2\sqrt{\delta_T}$, then $\mathbb{P}(\mathcal{Z}_T) \to 1$, where*

$$\mathcal{Z}_T = \{\hat{N} = N; \quad \max_{r=1,...,N} |\hat{\eta}_r - \eta_r| \leq C \log^2 T\}$$

for a certain $C > 0$*, where the guaranteed speed of convergence of* $\mathbb{P}(\mathcal{Z}_T)$ *to* 1 *is no faster than* $T\delta_T^{-1}(1 - \delta_T^2(1-\bar{c})^2T^{-2}/9)^M$ *where* M *is the number of random draws and* $\bar{c} = 3 - 2/c_\star$ *for* $c_\star$ *as in (3.6).*

For the purpose of comparison we note that the rate of convergence for the estimated change-points obtained for the BS method by Cho and Fryzlewicz (2013) is of order $\mathcal{O}(\sqrt{T}\log^{(2+\vartheta)} T)$ and $\mathcal{O}(\log^{(2+\vartheta)} T)$ for any positive constant ϑ when δ_T is of order $T^{3/4}$ and T respectively. In the WBS setting, the rate is square logarithmic when δ_T is of order $\log^2 T$, hence the improvement is significant. In addition, the lower threshold is always of order $\log T$ regardless of the minimum space between the change-points.

A natural question that arises at this point is whether improved consistency can be achieved by reconsidering the output of the BS algorithm. To be more specific, let us assume that the BS algorithm identifies $\hat{N}$ change-points instead of N where $\hat{N} < N$. With the reduced set, the BS algorithm can be re-applied to each of the $\hat{N}+1$ segments. However, it is not guaranteed that the change-points will be recovered with high probability. To see that, consider, for example, the occasion where Assumption (A3) is satisfied. Then, at the start of the BS algorithm no change-points will be detected and, hence, a further improvement is not feasible at all. This is where WBS achieves consistency over BS.

We also elaborate on the choice of the minimum number of random intervals M required to ensure consistency. From Fryzlewicz (2014) and by taking into consideration the "balanceness" parameter $\bar{c}$ we have that

$$M \geq \frac{9T^2}{\delta_T^2(1-\bar{c})}\log(T^2\delta_T^{-1}).$$

Hence, when $\delta_T = \mathcal{O}(T)$ only a small (logarithmic) number of random draws is necessary. However, a larger M is needed when δ_T is e.g. square logarithmic. In addition, to avoid the restriction of balanceness between the change-points, as in the BS method (see Cho and Fryzlewicz (2012) and Cho and Fryzlewicz (2013)), we assume that s_m and e_m are randomly drawn symmetrically around a certain change-point. To accomplish this a balanced draw is required so if, for example, we choose $c_\star \approx 1$ (i.e. an unbalanced draw) then M increases very fast.

3.3.2 Simultaneous across-scale post-processing

Theorem 1 covers the case of the multiplicative model (3.2). We now consider change-point detection for the function W_i^2 of the full model (3.7). Recall that any change-points in the piecewise constant functions $W_i(z)$ correspond to change-points in the autocovariance function $c_T(z,\tau) = \mathrm{cov}(X_{\lfloor zT \rfloor,T}, X_{\lfloor zT \rfloor+\tau,T}), \tau = 0, 1, ...$ of $X_{t,T}$ which in turn correspond to the change-points in $\mathbb{E}I_{t,T}^{(i)}$. Therefore, we are required to examine $I_{t,T}^{(i)}$ across scales $i = -1, -2, ..., -I^\star$ in order to detect the change-points and to accomplish this we propose two methods.

Method 1: The search for further change-points in each interval (s_m, e_m) proceeds to the next scale $i-1$ only if no change-points are detected at scale i on that interval. It therefore ensures that the finest scales are preferred (since change-points detected at the finest scales are likely to be more accurate) and only moves to coarser if necessary. Cho and Fryzlewicz (2012) use a similar technique to combine across scales change-points, but involving an extra parameter. The role of this parameter is to create groups of estimated change-points which are close to each other. Then, only one change-point (detected at the finest scale) from each of these groups will

survive the post-processing. Hence, their method will be used as a benchmark for our first type of across-scale post-processing.

Method 2: Alternatively, we suggest a method that simultaneously joins the estimated change-points across all the scales such that all the information from every scale is combined making it more likely for the true change-points and not spurious ones to exceed the threshold. Namely, motivated by Cho and Fryzlewicz (2013) who propose an alternative aggregation method to these of Groen et al. (2011) in order to detect change-points in the second order structure of a high-dimensional time series we define the following statistic

$$\mathbb{Y}_t^{thr} = \sum_{i=-I^\star}^{-1} \mathcal{Y}_t^{(i)} \mathbb{I}(\mathcal{Y}_t^{(i)} > \omega_T^{(i)}) \text{ for } i = -1, ..., -I^\star \tag{3.11}$$

where $\mathcal{Y}_t^{(i)} = |\mathbb{Y}_{s_m,e_m}^{b(i)}|/q_{s_m,e_m}^{(i)}$. This statistic differs from that of Cho and Fryzlewicz (2013) in that it applies across the scales $i = -1, -2, ..., -I^\star$ of a univariate time series, whereas Cho and Fryzlewicz (2013) calculate their statistic on the scales across many time series.

The algorithm is identical to the algorithm in Section 3.3 except for replacing (3.10) with (3.11). In addition, if the obtained $\mathbb{Y}_t^{thr} > 0$ there is no need to test further for the significance of b_0.

Below, we present the consistency theorem for the across-scale post-processing algorithm:

Theorem 2 *Let X_t follow model (3.7), and suppose that Assumptions (A1)-(A5) for $\sigma^2(t/T)$ hold for each $\beta_i(z)$ and $i = -1, -2, ..., I^\star$. Denote the number of change-points in $\beta_i(z)$ as N and the locations of those change-points as $\theta_1, ..., \theta_N$. Let $\hat{N}$ and $\hat{\theta}_1, ..., \hat{\theta}_N$ be the number and locations of the change-points (in ascending order), respectively, estimated by the across-scale post-processing method 1 or 2. There exist*

two constants C_3 *and* C_4 *such that if* $C_3 \log T \leq \omega_T \leq C_4 \delta_T$, *then* $\mathbb{P}(\mathcal{U}_T) \to 1$, *where*

$$\mathcal{U}_T = \{\hat{N} = N; \quad \max_{r=1,...,N} |\hat{\theta}_r - \theta_r| \leq C' \log^2 T\}$$

for a certain $C' > 0$, *where the guaranteed speed of convergence is the same as in Theorem 1.*

3.3.3 Post-processing

In order to control the number of change-points estimated from the WBS algorithm and to reduce the risk of over-segmentation we propose a post-processing method similar to Cho and Fryzlewicz (2012) and Inclan and Tiao (1994). More specifically, we compare every change-point against the adjacent ones using the CUSUM statistic making sure that (3.6) is satisfied. That is, for a set $\hat{\mathcal{N}} = \{\hat{\theta}_0, ..., \hat{\theta}_{N+1}\}$ where $\hat{\theta}_0 = 0$ and $\hat{\theta}_{N+1} = T$ we test whether $\hat{\theta}_r$ satisfies

$$\mathbb{Y}_t^{thr} = \sum_{i=-I^\star}^{-1} \mathcal{Y}_t^{(i)} \mathbb{I}(\mathcal{Y}_t^{(i)} > \omega_T^{(i)}) > 0 \text{ for } i = -1, ..., -I^\star$$

where $\mathcal{Y}_t^{(i)} = |\mathbb{Y}^{\hat{\theta}_r(i)}_{\hat{\theta}_{r-1},\hat{\theta}_{r+1}}|/|q^{(i)}_{\hat{\theta}_{r-1},\hat{\theta}_{r+1}}|$ and

$$\max\left(\frac{\hat{\theta}_{r+1} - \hat{\theta}_r}{\hat{\theta}_{r+1} - \hat{\theta}_{r-1} + 1}, \frac{\hat{\theta}_r - \hat{\theta}_{r-1} + 1}{\hat{\theta}_{r+1} - \hat{\theta}_{r-1} + 1}\right) \leq c_\star. \tag{3.12}$$

If $\mathbb{Y}_t^{thr} = 0$ then change-point $\hat{\theta}_r$ is temporarily eliminated from set $\hat{\mathcal{N}}$. In the next run, when considering change-point $\hat{\theta}_{r+1}$, the adjacent change-points are $\hat{\theta}_{r-1}$ and $\hat{\theta}_{r+2}$. When the post-processing finishes its cycle all temporarily eliminated change-points are reconsidered using as adjacent change-points those that have survived the first cycle. It is necessary for $\hat{\theta}_r$ to satisfy (3.12) with its adjacent estimated change-points $\hat{\theta}_{r-1}$ and $\hat{\theta}_{r+1}$, otherwise it is never eliminated. The algorithm is terminated when the set of change-points does not change.

3.3.4 Choice of threshold and parameters

In this section we present the choices of the parameters involved in the algorithms. From Theorems 1 and 2 we have that the threshold ω_T includes the constant $C^{(i)}$ which varies between the scales. The values of $C^{(i)}$ will be the same for all the methods presented, either BS/WBS or the Methods 1 and 2 in Section 3.3.2. Therefore, we can use the thresholds by Cho and Fryzlewicz (2012) who conduct experiments to establish the value of the threshold parameter under the null hypothesis of no change-points such that when the obtained statistic exceeds the threshold the null hypothesis is rejected. However, in that work the threshold is of the form $\tau_0 T^{\vartheta_0}\sqrt{\log T}$ where $\vartheta_0 \in (1/4, 1/2)$ and $\tau_0 > 0$ is the parameter that changes across scales. For that reason, we repeat the experiments which are described below.

We generate a vector $\mathbf{X} \sim \mathcal{N}(0, \Sigma)$ where the covariance matrix $\Sigma = (\sigma_{\kappa,\kappa'})_{\kappa,\kappa'=1}^{T}$ and $\sigma_{\kappa,\kappa'} = \rho^{|\kappa-\kappa'|}$. Then we find v that maximises (3.10). The following ratio

$$C_T^{(i)} = \mathbb{Y}_v^{(i)} (\log T)^{-1} \left(\sum_{t=1}^{T} I_{t,T}^{(i)} \right)^{-1} T$$

gives us an insight into the magnitude of parameter $C^{(i)}$. We repeat the experiment for different values of ρ and for every scale i we select $C^{(i)}$ as the 95% quantile. The same values are used for the post-processing method explained in Section 3.3.3. Our results indicate that $C^{(i)}$ tends to increase as we move to coarser scales due to the increasing dependence in the wavelet periodogram sequences.

Further, based on empirical evidence we select the scale $I^\star$ by setting $\lambda = 0.7$. In stage III of the algorithm we mentioned that the procedure is terminated when either the CUSUM statistic does not exceed a certain threshold or the length of the respective segment is Δ_T. This also defines the minimum length of a favourable

draw from (3.4). We choose Δ_T to be of the same order as δ_T since this is the lowest permissible order of magnitude according to (A5). Practically, we find that the choice $\Delta_T = \lfloor \log^2 T/3 \rfloor$ works well. Finally, we set $c_\star = 0.75$.

3.4 Simulation study

We present a set of simulation studies to assess the performance of our methods. In all the simulations we assume sample sizes to be either 256 or 1024 over 100 iterations. For comparison we also report the performance of the method by Cho and Fryzlewicz (2012) - henceforth CF - using the default values specified in their paper. BS1 and BS2 refer to the Method 1 and Method 2 of aggregation (as described in Section 3.3.2) using the BS technique, respectively. WBS1 and WBS2 refer to the Method 1 and Method 2 of aggregation (as in Section 3.3.2) using the Wild Binary Segmentation technique, respectively.

3.4.1 Models with no change-points

We simulate stationary time series with innovations $\varepsilon_t \sim \mathcal{N}(0, 1)$ and we report the number of occasions (out of 100) the methods incorrectly rejected the null hypothesis of no change-points. The models S1-S7 (Table 3.1) we consider here are taken from Nason (2013a).

The results of Table 3.1 indicate our methods' good performance over that of Cho and Fryzlewicz (2012) apart from models $S3$ and $S7$ where all methods incorrectly reject the null hypothesis frequently in many occasions. A visual inspection of an AR(1) process with $\phi = -0.9$ could confirm that this type of process exhibits a "clus-

Table 3.1: Stationary processes results. For all the models the sample size is 1024 and there are no change-points. Figures show the number of occasions the methods detected change-points with the universal thresholds $C^{(i)}$ obtained as described in Section 3.3.4. Figures in brackets are the number of occasions the methods detected change-points with the thresholds $C^{(i)}$ obtained as described in Section 3.4.1.

Model	BS1	WBS1	BS2	WBS2	CF
S1: iid standard normal	1 [0]	3 [2]	0 [0]	1 [0]	4
S2: AR(1) with parameter 0.9	3 [1]	5 [1]	1 [1]	5 [1]	9
S3: AR(1) with parameter -0.9	58 [0]	93 [0]	46 [0]	48 [5]	79
S4: MA(1) with parameter 0.8	2 [3]	7 [4]	3 [3]	1 [0]	7
S5: MA(1) with parameter -0.8	2 [0]	4 [2]	4 [0]	0 [0]	7
S6: ARMA(1,0,2) with AR= $\{-0.4\}$ and MA= $\{-0.8, 0.4\}$	8 [0]	27 [0]	8 [0]	8 [0]	25
S7: AR(2) with parameters 1.385929 and -0.9604	88 [3]	99 [4]	88 [3]	88 [5]	96

tering behaviour" which mimics changing variance. Hence, the process is interpreted as non-stationary by the wavelet periodogram resulting in erroneous outcomes. A similar argument is valid for S7 model. To correct that limitation, parameter $C^{(i)}$ should be chosen with care. Higher values will ensure that the null hypothesis is not rejected frequently. This is achieved by not using universal thresholds (as shown in Section 3.3.4) but calculating them for every instance. Specifically, given a time series y_t we fit an AR(p) model. Then we generate 100 instances of the same length and with the same AR(p) coefficients. Similarly with Section 3.3.4 we select $C^{(i)}$ as the 95% quantile. This procedure is more computationally intensive and imposes a parametric assumption about the underlying processes but improves the method significantly; see the figures in brackets (Table 3.1).

3.4.2 Non-stationary models

We now examine the performance of our method for a set of non-stationary models by using and extending the examples from Cho and Fryzlewicz (2012). Since the WBS method has improved rates of convergence new simulation results are presented which assess how close to the real change-points the estimated ones are. For this reason we report the total number of change-points identified within $\lfloor 5\% \cdot T \rfloor$ from the real ones. The results favour WBS methods even more when $\lfloor 2.5\% \cdot T \rfloor$ distances are considered and, hence, omitted for brevity.

The accuracy of a method should be also judged in parallel with the total number of change-points identified. We propose a test that tries to accomplish this. Assuming that we define the maximum distance from a real change-point η as $d_{\max}$, an estimated change-point $\hat{\eta}$ is correctly identified if $|\eta - \hat{\eta}| \leq d_{\max}$ (here within 5% of the sample size). If two (or more) estimated change-points are within this distance then only one change-point which is the closest to the real change-point is classified as correct. The rest are deemed to be false, except if any of these are close to another change-point. An estimator performs well when the hit ratio HR is closer to 1

$$HR = \frac{\#\text{correct change-points identified}}{\max(N, \hat{N})}.$$

By using the term $\max(N, \hat{N})$ we aim to penalise cases where, for example, the estimator correctly identifies a certain number of change-points all within the distance $d_{\max}$ but $\hat{N} < N$. It also penalises the estimator when $\hat{N} > N$ and all $\hat{N}$ change-points are within distance $d_{\max}$.

3.4.2.1 Large sample size simulation study

We proceed by assessing the performance of the methods when $T = 1024$ using the following models. Models A and C are taken from Davis et al. (2006) and models B, E and F from Cho and Fryzlewicz (2012). All the results are shown in Table 3.2.

Model A: *A non-stationary process that includes one AR(1) and two AR(2) processes with two clearly observable change-points*

$$y_t = \begin{cases} 0.9y_{t-1} + \varepsilon_t,\ \varepsilon_t \sim \mathcal{N}(0,1) & \text{for } 1 \leq t \leq 512 \\ 1.68y_{t-1} - 0.81y_{t-2} + \varepsilon_t,\ \varepsilon_t \sim \mathcal{N}(0,1) & \text{for } 513 \leq t \leq 768 \\ 1.32y_{t-1} - 0.81y_{t-2} + \varepsilon_t,\ \varepsilon_t \sim \mathcal{N}(0,1) & \text{for } 769 \leq t \leq 1024 \end{cases}$$

Both BS and WBS detect change-points with high accuracy. The two procedures over-segmented the process less than 30% of the time. CF tended to detect spurious change-points mainly towards the end of the series.

Model B: *A non-stationary process with two less clearly observable change-points*

$$y_t = \begin{cases} 0.4y_{t-1} + \varepsilon_t,\ \varepsilon_t \sim \mathcal{N}(0,1) & \text{for } 1 \leq t \leq 400 \\ -0.6y_{t-1} + \varepsilon_t,\ \varepsilon_t \sim \mathcal{N}(0,1) & \text{for } 401 \leq t \leq 612 \\ 0.5y_{t-1} + \varepsilon_t,\ \varepsilon_t \sim \mathcal{N}(0,1) & \text{for } 613 \leq t \leq 1024 \end{cases}$$

All methods do well in the estimation of this type of model. Approximately the same number of change-points within 5% are detected even though BS and WBS were more conservative in the total number of change-points. This results in the improved hit ratio.

Model C: *A non-stationary process with a short segment at the start*

$$y_t = \begin{cases} 0.75y_{t-1} + \varepsilon_t,\ \varepsilon_t \sim \mathcal{N}(0,1) & \text{for } 1 \leq t \leq 50 \\ -0.5y_{t-1} + \varepsilon_t,\ \varepsilon_t \sim \mathcal{N}(0,1) & \text{for } 51 \leq t \leq 1024 \end{cases}$$

In this type of model both BS2 and CF perform well compared with the BS1, WBS1 and WBS2 methods. Over WBS it is expected that binary segmentation methods will perform better due to the fact that the latter starts its search assuming a single change-point. Hence, the CUSUM statistic will take its maximum value when the starting and ending point is $s = 1$ and $e = 1024$ respectively, which we observed to happen less frequently for the WBS methods.

Model D: *A non-stationary process similar to model B but with the two change-points at a short distance from each other.*

In this model, the two change-points occur very close to each other i.e. $(400, 470)$ instead of $(400, 612)$. The CF method, BS1 and BS2 do not perform well as in half of the cases the two change-points were detected. On the contrary, the WBS1 and WBS2 methods achieved high hit ratio (almost double of that of the BS methods) and in less than 8% of the cases did not detect any change-point.

Model E: *A highly persistent non-stationary process with time-varying variance*

$$y_t = \begin{cases} 1.399y_{t-1} - 0.4y_{t-2} + \varepsilon_t,\ \varepsilon_t \sim \mathcal{N}(0,0.8) & \text{for } 1 \leq t \leq 400 \\ 0.999y_{t-1} + \epsilon_t,\ \varepsilon_t \sim \mathcal{N}(0,1.2^2) & \text{for } 401 \leq t \leq 750 \\ 0.699y_{t-1} + 0.3y_{t-1} + \varepsilon_t,\ \varepsilon_t \sim \mathcal{N}(0,1) & \text{for } 751 \leq t \leq 1024 \end{cases}$$

The CF and BS1 methods perform well since they detect most of the change-points within 5% distance from the real ones. From our simulations we noticed that in most cases the two change-points were found in the finest scale ($i = -1$). The aggregation Method 2 does not improve the estimation since its purpose is to simultaneously combine the information from different scales not just from a single one. On the other hand, the CF method and Method 1 favour change-points detected in the finest scales and this is the reason for their good performance.

Model F: *A piecewise constant ARMA(1,1) process*

$$y_t = \begin{cases} 0.7y_{t-1} + \varepsilon_t + 0.6\varepsilon_{t-1}, & \text{for } 1 \leq t \leq 125 \\ 0.3y_{t-1} + \varepsilon_t + 0.3\varepsilon_{t-1}, & \text{for } 126 \leq t \leq 532 \\ 0.9y_{t-1} + \varepsilon_t, & \text{for } 533 \leq t \leq 704 \\ 0.1y_{t-1} + \varepsilon_t - 0.5\varepsilon_{t-1}, & \text{for } 704 \leq t \leq 1024 \end{cases}$$

The first change-point is the least apparent and is left undetected in most cases when applying the CF method. Our methods are capable of capturing this point more frequently while in almost double of the cases they find the correct number of change-points within 5% of their real positions.

Model G: *A near-unit-root non-stationary process with time-varying variance*

$$y_t = \begin{cases} 0.999y_{t-1} + \varepsilon_t,\ \varepsilon_t \sim \mathcal{N}(0,1) & \text{for } 1 \leq t \leq 200,\ 401 \leq t \leq 600 \text{ and } 801 \leq t \leq 1024 \\ 0.999y_{t-1} + \varepsilon_t,\ \varepsilon_t \sim \mathcal{N}(0,1.5^2) & \text{for } 201 \leq t \leq 400 \text{ and } 601 \leq t \leq 800 \end{cases}$$

In this near-unit-root process there are 4 change-points in its variance. All binary

segmentation methods do not perform well as they often miss the middle change-points. Both WBS1 and WBS2 manage to detect most of the change-points achieving a hit ratio almost three times higher than BS2. In almost 70% of the occasions WBS2 detects at least 4 change-points.

Model H: *A non-stationary process similar to model F but with the three change-points at a short distance from each other.*

In this model the three change-points occur close to each other, i.e. $\mathcal{N} = (125, 325, 550)$. The first two change-points fail to be detected by the CF and BS methods in many instances. By contrast, WBS2 performs better in this case by identifying them more often. This results in a higher hit ratio.

Model I: *A non-stationary AR process with many changes within close distances.*

We simulate instances with 5 change-points occurring at uniformly distributed positions. We allow the distances to be as small as 30 and not larger than 100.

In this scenario, CF correctly identifies more than 4 change-points in 15% instances while BS1 and BS2 in 15% and 16% respectively. Again, the WBS methods do well in revealing the majority of the change-points and in many cases close to the real ones.

3.4.2.2 Small sample size simulation study

We proceed by assessing the performance of the methods when $T = 256$ using the following models. These models are modifications of the models discussed above except Cs which is taken from Killick et al. (2013). All the results are shown in Table 3.3.

Model As: *A non-stationary process similar to model A*

Table 3.2: Non-stationary processes results for $T = 1024$. Panel I shows the number of occasions a method detected that number of change-points within a distance of 5% from the real ones. Bold: the method with the highest hit ratio or within 10% from the highest. Panel II shows the percentage of occasions a method detected that number of change-points. True number of change-points is in bold.

Panel I

Number of Change-points within 5% (Panel I)

Model	A					B					C				
	BS1	BS2	WBS1	WBS2	CF	BS1	BS2	WBS1	WBS2	CF	BS1	BS2	WBS1	WBS2	CF
0	2	0	1	0	3	0	0	0	0	0	39	12	35	21	6
1	29	15	16	21	29	11	8	4	9	7	61	88	65	79	94
2	69	85	83	79	68	89	92	96	91	93	-	-	-	-	-
Hit ratio	**0.768**	**0.850**	**0.817**	**0.808**	0.712	**0.928**	**0.921**	**0.966**	**0.928**	0.865	0.580	**0.860**	0.600	0.746	**0.853**

Number of Change-points within 5%

Model	D					E					F				
	BS1	BS2	WBS1	WBS2	CF	BS1	BS2	WBS1	WBS2	CF	BS1	BS2	WBS1	WBS2	CF
0	36	52	12	11	48	6	12	8	11	1	2	0	0	0	1
1	58	14	9	11	12	40	42	59	53	40	18	6	5	3	7
2	6	34	79	78	40	54	46	33	36	59	32	32	22	24	45
3	-	-	-	-	-	-	-	-	-	-	48	62	73	73	47
Hit ratio	0.428	0.403	**0.835**	**0.835**	0.436	**0.712**	0.649	0.610	0.611	**0.743**	0.744	**0.847**	**0.890**	**0.894**	0.765

Number of Change-points within 5%

Model	G					H					I				
	BS1	BS2	WBS1	WBS2	CF	BS1	BS2	WBS1	WBS2	CF	BS1	BS2	WBS1	WBS2	CF
0	58	60	9	11	39	0	0	2	2	0	0	2	1	0	0
1	11	11	13	6	20	40	33	23	16	29	39	33	8	8	39
2	20	21	20	20	30	38	37	38	40	57	16	15	8	7	27
3	6	5	15	22	5	22	30	37	42	14	23	27	20	18	25
4	5	3	43	41	6	-	-	-	-	-	14	11	22	18	3
5	-	-	-	-	-	-	-	-	-	-	8	12	41	49	6
Hit ratio	0.222	0.200	**0.671**	**0.686**	0.297	0.605	0.654	**0.693**	**0.732**	0.603	0.472	0.496	**0.745**	**0.779**	0.419

Panel II

Number of Change-points

Model	A					B					C				
	BS1	BS2	WBS1	WBS2	CF	BS1	BS2	WBS1	WBS2	CF	BS1	BS2	WBS1	WBS2	CF
0	0	0	0	0	0	0	0	0	0	0	20	9	23	10	2
1	8	0	0	0	0	0	0	0	0	1	**67**	**86**	**66**	**80**	**81**
2	**68**	**76**	**74**	**70**	**65**	**95**	**88**	**96**	**91**	**70**	12	5	7	7	16
>2	24	24	26	30	35	5	12	4	9	29	1	0	4	3	1
Total	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100

Number of Change-points

Model	D					E					F				
	BS1	BS2	WBS1	WBS2	CF	BS1	BS2	WBS1	WBS2	CF	BS1	BS2	WBS1	WBS2	CF
0	44	42	8	7	38	1	0	0	0	0	0	0	0	0	0
1	15	18	5	4	17	19	22	28	26	19	15	2	0	1	1
2	**39**	**38**	**87**	**89**	**38**	**69**	**69**	**64**	**66**	**65**	15	12	15	13	19
3	2	2	0	0	7	10	8	7	8	15	**65**	**83**	**83**	**83**	**65**
>3	0	0	0	0	0	1	1	1	0	1	5	3	2	3	15
Total	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100

Number of Change-points

Model	G					H					I				
	BS1	BS2	WBS1	WBS2	CF	BS1	BS2	WBS1	WBS2	CF	BS1	BS2	WBS1	WBS2	CF
0	58	59	9	11	38	0	0	0	0	0	0	0	0	0	0
1	7	7	5	2	16	24	21	14	13	12	33	30	7	7	22
2	24	23	25	18	32	36	28	28	29	51	12	10	5	5	28
3	2	2	0	1	3	**39**	**50**	54	**55**	**30**	24	24	15	16	24
4	**9**	**9**	**59**	**66**	**11**	1	1	4	3	7	16	20	13	9	11
>4	0	0	2	2	0	0	0	0	0	0	**15**	**16**	**60**	**63**	**15**
Total	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100

In this model the change-points occur in positions $(128, 188)$. All methods perform similarly.

Model Bs: *A non-stationary process similar to model B*

In this model the change-points occur in positions $(100, 153)$. WBS1 and WBS2 do well in this example achieving a hit ratio almost double than that of the rest methods. In more than 67% of the occasions they detected two change-points without over-segmenting the series.

Model Cs: *A piecewise constant MA process*

$$y_t = \begin{cases} \varepsilon_t + 0.8\varepsilon_{t-1},\ \varepsilon_t \sim \mathcal{N}(0,1) & \text{for } 1 \leq t \leq 128 \\ \varepsilon_t + 1.68\varepsilon_{t-1} - 0.81\varepsilon_{t-2},\ \varepsilon_t \sim \mathcal{N}(0,1) & \text{for } 129 \leq t \leq 256 \end{cases}$$

All our methods outperform the CF method and, in particular, WBS2 is more accurate in detecting the single change-point in 87 occasions. In addition, even though a single change-point is present in the time series WBS methods do better in this example.

Model Ds: *A non-stationary process similar to model Bs but with the two change-points at a short distance from each other.*

In this model, the two change-points occur very close to each other i.e. $(100, 135)$ and, hence, is a harder version of model Bs. Due to the short distance between the change-points the WBS methods do well here detecting both change-points in at least 50% of the occasions and in most cases within 5% from the real ones. On the contrary, CF, BS1 and BS2 fail to detect any change-points in more than 70% of the occasions even though the former performed slightly better.

Model Es: *A non-stationary process similar to model Cs but with two change-points and at a short distance from each other.*

In this model, the two change-points occur very close to each other i.e. $(85, 120)$. The WBS methods achieve a high hit ratio and WBS1 detected both change-points correctly in 85% of the occasions, more than double from BS1 which performed the worst.

Model Fs: *A non-stationary AR process with many changes within close distances.*

We simulate instances with 4 change-points occurring at uniformly distributed positions. We allow the distances to be as small as 15 and not larger than 80.

Again, the WBS methods do well in revealing most of the change-points and with a good accuracy. This resulted in the higher hit ratio.

3.5 Applications

3.5.1 US Gross National Product series (GNP)

We obtain the GNP from the Federal Reserve Bank of St. Louis web page[1]. The seasonally adjusted and quarterly data is expressed in billions of dollars and spans the period from 1947:1 to 2013:1 but we only use the last 256 observations, i.e. from 1949:4. On the left panel of Figure 3.2 one can see the logarithm of the GNP series. As in Shumway and Stoffer (2011) we only examine the first difference of the logarithm of the GNP (also called the growth rate) since there is an obvious linear trend. From the right panel of Figure 3.2 which illustrates the growth rate it is visually clear that

[1]See http://research.stlouisfed.org/fred2/series/GNP

Table 3.3: Non-stationary processes results for $T = 256$. Panel I shows the number of occasions a method detected that number of change-points within a distance of 5% from the real ones. Bold: the method with the highest hit ratio or within 10% from the highest. Panel II shows the percentage of occasions a method detected that number of change-points. True number of change-points is in bold.

Panel I

Number of Change-points within 5%

Model	As					Bs					Cs				
	BS1	BS2	WBS1	WBS2	CF	BS1	BS2	WBS1	WBS2	CF	BS1	BS2	WBS1	WBS2	CF
0	30	22	26	25	28	64	65	31	33	67	17	19	16	13	17
1	67	73	68	70	61	14	14	24	24	12	83	81	84	87	83
2	3	5	6	5	11	22	21	45	43	21	-	-	-	-	-
Hit ratio	0.363	**0.413**	**0.396**	**0.400**	**0.411**	0.286	0.280	**0.570**	**0.550**	0.265	**0.810**	**0.790**	**0.835**	**0.865**	0.775

Number of Change-points within 5%

Model	Ds					Es					Fs				
	BS1	BS2	WBS1	WBS2	CF	BS1	BS2	WBS1	WBS2	CF	BS1	BS2	WBS1	WBS2	CF
0	85	85	43	47	82	45	44	5	7	44	47	46	24	25	53
1	4	4	6	5	6	17	16	10	16	17	36	36	34	40	36
2	11	11	51	48	12	38	40	85	77	39	16	16	19	19	11
3	-	-	-	-	-	-	-	-	-	-	1	2	11	7	0
4	-	-	-	-	-	-	-	-	-	-	0	0	12	9	0
Hit ratio	0.126	0.130	**0.540**	**0.505**	0.150	0.458	0.473	**0.896**	**0.843**	0.475	0.177	0.185	**0.382**	**0.337**	0.145

Panel II

Number of Change-points

Model	As					Bs					Cs				
	BS1	BS2	WBS1	WBS2	CF	BS1	BS2	WBS1	WBS2	CF	BS1	BS2	WBS1	WBS2	CF
0	2	2	3	2	2	57	58	27	29	57	0	0	0	0	0
1	82	76	80	75	69	7	7	2	3	11	**96**	**96**	**99**	**99**	**88**
2	**15**	**21**	**16**	**23**	**28**	**35**	**35**	**71**	**67**	**30**	4	4	1	1	12
>2	1	1	1	0	1	1	0	0	1	2	0	0	0	0	0
Total	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100

Number of Change-points

Model	Ds					Es					Fs				
	BS1	BS2	WBS1	WBS2	CF	BS1	BS2	WBS1	WBS2	CF	BS1	BS2	WBS1	WBS2	CF
0	75	75	41	43	70	44	44	5	6	43	27	26	17	20	23
1	13	13	4	5	18	10	8	4	4	11	44	44	35	34	46
2	**11**	**12**	**55**	**52**	**12**	**44**	**46**	**90**	**88**	46	26	26	21	27	30
3	1	0	0	0	0	2	2	1	2	**0**	2	2	12	8	0
>3	0	0	0	0	0	0	0	0	0	0	**1**	**2**	**15**	**11**	**1**
Total	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100

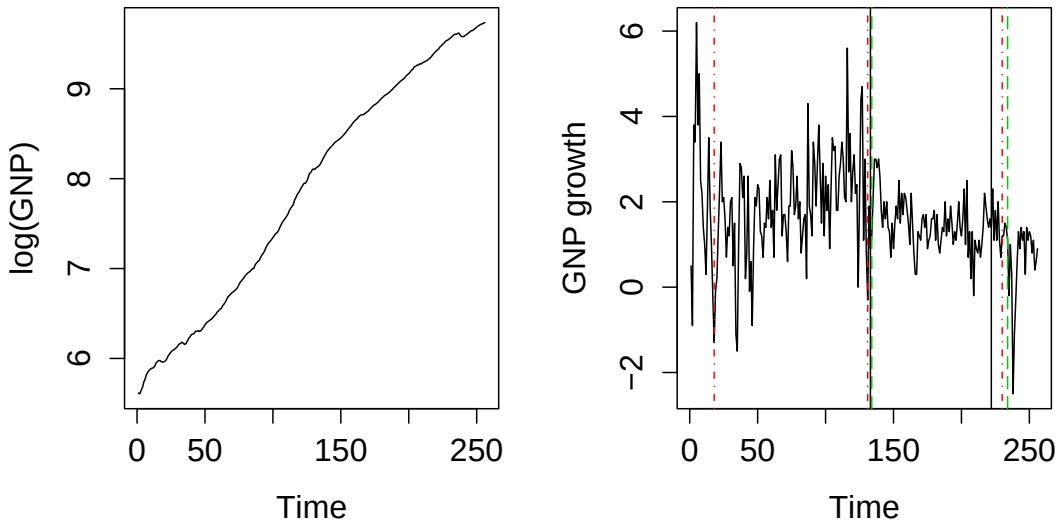

Figure 3.2: Natural logarithm of the GNP series (left) and its first difference (right). The black, green and red vertical lines are the change-points as estimated by BS2, CF and WBS2 respectively.

the GNP series exhibits less variability in the right side. We are interested in finding whether our method is capable of spotting this change and/or possibly others.

Applying our method i.e. BS2 and WBS2 (BS1 and WBS1 produced identical results) we find that BS2 detects two change-points $\hat{\eta} = \{133, 222\}$ while the WBS2 detects three at positions $\{18, 131, 230\}$. For the sake of comparison, CF detects two possible change-points i.e. $\hat{\eta} = \{134, 234\}$. The acf graphs in Figure 3.3 confirm that there are changes in the autocovariance structure for all the possible sets of change-points.

Change-point 18 i.e. 1953(3) almost exactly coincides with a peak of the GNP growth as decided by the Business Cycle Dating Committee of the National Bureau of Economic Research where the official date is July 1953 (note that cycles do not

necessarily overlap with the quarterly publications of the GNP). In addition, change-points 131, 133 and 134 lie within a cycle that peaks in January 1981 and has a trough in November 1982. This cycle corresponds to the start of the Great Moderation (around 1980s), a period that experienced more efficient monetary policy and shocks of small magnitude, see Blanchard and Simon (2001), Stock and Watson (2003), Bernanke (2004) and Clark (2009) among others. Finally, all three methods detected a change-point towards the end of the series - $222, 230, 234$ which are dated 2004(3), 2006(3) and 2007(3) respectively. According to Clark (2009) the Great Moderation had reversed and the decline is offset by negative growth rates due to the recent economic recession[2].

3.5.2 Infant Electrocardiogram Data (ECG)

We apply the three methods (CF, BS2, WBS2) to the ECG data of an infant found in the *R* package *wavethresh* (Nason (2013b)). This is a popular example of a non-stationary time series and it has been analysed in e.g. Nason et al. (2000). The local segments of possible stationarity indicate the sleep state of the infant and it is classified on a scale from 1 to 4, see the caption to Figure 3.4. The same figure plots the time series with the respective estimated change-points (the methods were applied on the first difference so that its mean is approximately zero). All methods identify most of the sleep states and, notably, WBS2 detects the abrupt change of short duration (quite sleep-awake-quiet sleep) towards the end of the series.

[2]It should be mentioned that other econometric techniques return multiple change-points, see Hamilton (1989) for an early attempt to examine GNP for the identification of "contraction" and "expansion" states. However, our findings are most related to the studies mentioned in the text.

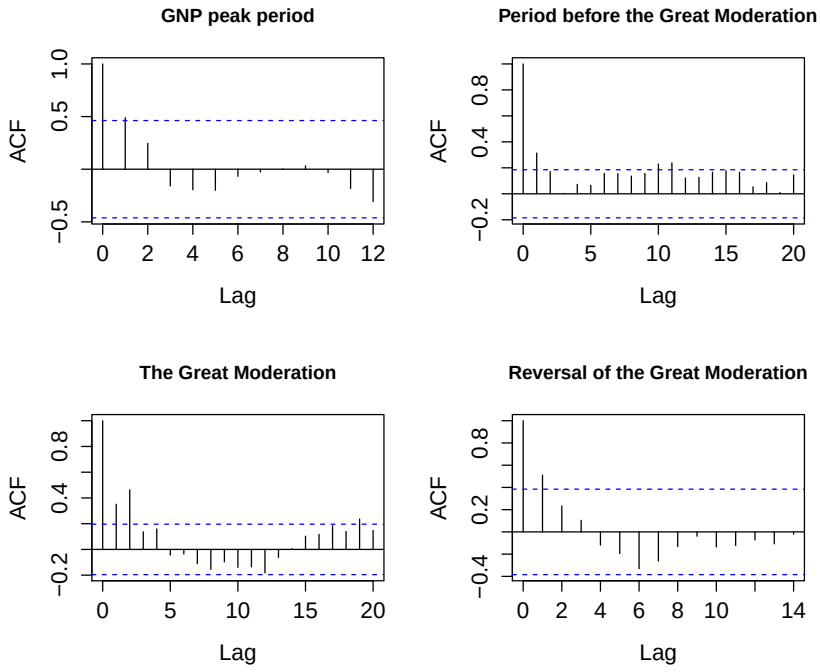

Figure 3.3: The graphs are the acfs for the four periods discussed in the text for the change-points estimated by WBS2.

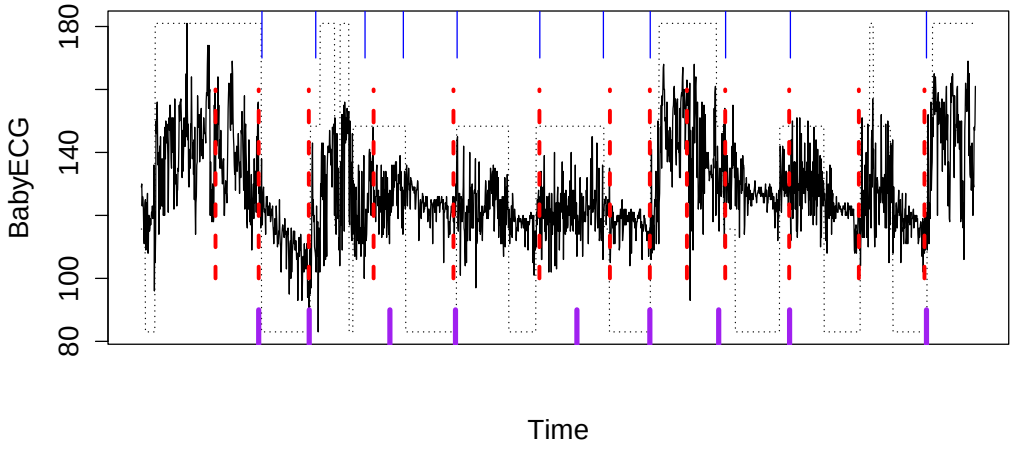

Figure 3.4: Plot of BabyECG data. The top blue, middle red and bottom purple vertical lines are the change-points as estimated by CF, WBS2 and BS2 respectively. The horizontal dotted line represents the sleep states i.e. 1 = quiet sleep, 2 = quiet-to-active sleep, 3 = active sleep, 4 =awake.

3.6 Proofs

Proof of Theorem 1

We notice that the proof of consistency is based on the following multiplicative model

$$\tilde{Y}_{t,T} = \sigma(t/T)^2 Z_{t,T}^2 \quad t = 0, ..., T-1.$$

We define the following two CUSUM statistics

$$\mathbb{Y}_{s,e}^b = \sqrt{\frac{e-b}{n(b-s+1)}} \sum_{t=s}^{b} \tilde{Y}_{t,T} - \sqrt{\frac{b-s+1}{n(e-b)}} \sum_{t=b+1}^{e} \tilde{Y}_{t,T}$$

and

$$\mathbb{S}_{s,e}^b = \sqrt{\frac{e-b}{n(b-s+1)}} \sum_{t=s}^{b} \sigma^2(t/T) - \sqrt{\frac{b-s+1}{n(e-b)}} \sum_{t=b+1}^{e} \sigma^2(t/T)$$

where $n = e - s + 1$, the size of the segment defined by (s, e).

$\mathbb{Y}_{s,e}^b$ can be seen as the inner product between sequence $\{\tilde{Y}_{t,T}\}_{t=s,...,e}$ and a vector $\psi_{s,e}^b$ whose elements $\psi_{s,e,t}^b$ are constant and positive for $t \leq b$ and constant and negative for $t > b$ such that they sum to zero and sum to one when squared. Similarly for $\mathbb{S}_{s,e}^b$.

Let s, e satisfy $\eta_{p_0} \leq s < \eta_{p_0+1} < ... < \eta_{p_0+q} < e \leq \eta_{p_0+q+1}$ for $0 \leq p_0 \leq N - q$. The inequality will hold at all stages of the algorithm until no undetected change-points are remained. We impose at least one of the following conditions

$$s < \eta_{p_0+r'} - C\delta_T < \eta_{p_0+r'} + C\delta_T < e, \quad \text{for some } 1 \leq r' \leq q \tag{3.13}$$

$$\{(\eta_{p_0+1} - s) \wedge (s - \eta_{p_0})\} \vee \{(\eta_{p_0+q+1} - e) \wedge (e - \eta_{p_0+q})\} \leq C\epsilon_T \tag{3.14}$$

where $\wedge$ and $\vee$ denote the minimum and maximum operators, respectively. These inequalities will hold throughout the algorithm until no further change-points are detected.

We define symmetric intervals $\mathcal{I}_r^L$ and $\mathcal{I}_r^R$ around change-points such that for every triplet $\{\eta_{r-1}, \eta_r, \eta_{r+1}\}$

$$\mathcal{I}_r^L = \left[\eta_r - \frac{2}{3}\delta_{\min}^r, \eta_r - \frac{1}{3}\delta_{\min}^r (1+\bar{c})\right]$$

and

$$\mathcal{I}_r^R = \left[\eta_r + \frac{1}{3}\delta_{\min}^r (1+\bar{c}), \eta_r + \frac{2}{3}\delta_{\min}^r\right] \quad \text{for } r = 1, ..., N+1$$

where $\delta_{\min}^r = \min\{\eta_r - \eta_{r-1}, \eta_{r+1} - \eta_r\}$ and $\bar{c} = 3 - \frac{2}{c_\star}$ for $c_\star$ as in (3.6). We recall that at every stage of the WBS algorithm M intervals (s_m, e_m), $m = 1, ..., M$ are drawn from a discrete uniform distribution over the set $\{(s, e) : s < e, 1 \le s \le T-1, 2 \le e \le T\}$.

We define the event D_T^M as

$$D_T^M = \{\forall r = 1, ..., N \;\; \exists\, m = 1, ..., M \;\; (s_m, e_m) \in \mathcal{I}_r^L \times \mathcal{I}_r^R\}.$$

Also, note that

$$\mathbb{P}((D_T^M)^c) \le \sum_{r=1}^{N} \prod_{m-1}^{M} (1 - \mathbb{P}((s_m, e_m) \in \mathcal{I}_r^L \times \mathcal{I}_r^R)) \le \frac{T}{\delta_T}(1 - \delta_T^2(1-\bar{c})^2 T^{-2}/9)^M.$$

Similarly with Fryzlewicz (2014), for M large enough we have that the interval (s_m, e_m) is such that it contains only one change-point. On a generic interval satisfying (3.13) and (3.14) we consider

$$(m_0, b) = \arg \max_{(m,t): m \in \mathcal{M}_{s,e}, s_m \le t \le e_m} |\tilde{Y}_{s_m, e_m}^t| \tag{3.15}$$

where $\mathcal{M}_{s,e} = \{m : (s_m, e_m) \subseteq (s, e), 1 \le m \le M\}$.

Lemma 3.1.

$$\mathbb{P}\left(\max_{(s_{m_0}, b, e_{m_0}) \in \mathcal{M}_{s,e}} \left| \mathbb{Y}_{s_{m_0}, e_{m_0}}^b - \mathbb{S}_{s_{m_0}, e_{m_0}}^b \right| > \lambda_1 \right) \to 0 \tag{3.16}$$

$$\lambda_1 \ge \log T$$

Proof. We start by studying the following event

$$\left| \sum_{t=s_{m_0}}^{e_{m_0}} c_t \sigma(t/T)^2 (Z_{t,T}^2 - 1) \right| > \sqrt{n_{m_0}} \lambda_1$$

where $c_t = \sqrt{(e_{m_0} - b_{m_0})/(b - s_{m_0} + 1)}$ and $c_t = \sqrt{(b - s_{m_0} + 1)/(e_{m_0} - b_{m_0})}$ for $t \leq b$ and $b + 1 \leq t$ respectively. From (3.6), we have that $c_t \leq c^\star \equiv \sqrt{\frac{c_\star}{1 - c_\star}} < \infty$. The proof proceeds as in Cho and Fryzlewicz (2013) and we have that (3.16) is bounded by

$$\sum_{(s_{m_0}, b, e_{m_0}) \in \mathcal{M}_{s,e}} 2 \exp\left(- \frac{n_{m_0} \lambda_1^2}{4 c_\star^2 \max_z \sigma^2(z) n_{m_0} \rho_\infty^2 + 2 c_\star \max_z \sigma(z) \sqrt{n_{m_0}} \lambda_1 \rho_\infty^1} \right)$$
$$\leq 2T^3 \exp\left(-C_1'(c^{\star^{-2}}) \log^2 T \right)$$

which converges to 0 since $n_{m_0} \geq \delta_T = \mathcal{O}(\log^2 T)$ and $\rho_\infty^1 < \infty$ from (A2). □

Lemma 3.2. *Assuming that (3.13) holds, then there exists $C_2 > 0$ such that for b satisfying $|b - \eta_{p_0 + r'}| = C_2 \gamma_T$ for some r', we have $|\mathbb{S}_{s_{m_0}, e_{m_0}}^{\eta_{p_0+r'}}| \geq |\mathbb{S}_{s_{m_0}, e_{m_0}}^{b}| + C\gamma_T \delta_T^{-1/2} \geq |\mathbb{S}_{s_{m_0}, e_{m_0}}^{b}| + 2\lambda_1$, where $\gamma_T = \sqrt{\delta_T} \lambda_1$.*

Proof. From the proof of Theorem 3.2 in Fryzlewicz (2014) and Lemma 1 in Cho and Fryzlewicz (2012) we have the following result

$$|\mathbb{S}_{s_{m_0}, e_{m_0}}^{b}| \geq |\mathbb{Y}_{s_{m_0}, e_{m_0}}^{b}| - \lambda_1 \geq C_3 \sqrt{\delta_T} \tag{3.17}$$

provided that $\delta_T \geq C_4 \lambda_1^2$.

By Lemma 2.2 in Venkatraman (1992) there exists a change-point $\eta_{p_0+r'}$ immediately to the left or right of b such that

$$|\mathbb{S}_{s_{m_0}, e_{m_0}}^{\eta_{p_0+r'}}| > |\mathbb{S}_{s_{m_0}, e_{m_0}}^{b}| \geq C_3 \sqrt{\delta_T}$$

Now, the following three cases are not possible:

1. (s_{m_0}, e_{m_0}) contains a single change-point, $\eta_{p_0+r'}$, and both $\eta_{p_0+r'} - s_{m_0}$ and $e_{m_0} - \eta_{p_0+r'}$ are not bounded from below by $c_1\delta_T$.

2. (s_{m_0}, e_{m_0}) contains a single change-point, $\eta_{p_0+r'}$, and either $\eta_{p_0+r'} - s_{m_0}$ or $e_{m_0} - \eta_{p_0+r'}$ are not bounded from below by $c_1\delta_T$.

3. (s_{m_0}, e_{m_0}) contains two change-point, $\eta_{p_0+r'}$ and $\eta_{p_0+r'+1}$, and both $\eta_{p_0+r'} - s_{m_0}$ and $e_{m_0} - \eta_{p_0+r'+1}$ are not bounded from below by $c_1\delta_T$.

The first case is not permitted by (A5). For the last two, if either case were true, then following the arguments as in Lemma A.5 of Fryzlewicz (2014), we would obtain $\max_{t:s_{m_0}\le t\le e_{m_0}} |\mathbb{S}^t_{s_{m_0},e_{m_0}}|$ were not bounded from below by $C_3\sqrt{\delta_T}$ which contradicts (3.17). Hence, interval (s_{m_0}, e_{m_0}) satisfies condition (3.13) and following a similar argument with the proof of Lemma 2 in Cho and Fryzlewicz (2012) we can show that for any b satisfying $|b - \eta_{p_0+r'}| = C_2\gamma_T$, then $|\mathbb{S}^{\eta_{p_0}+r'}_{s_{m_0},e_{m_0}}| \ge |\mathbb{S}^b_{s_{m_0},e_{m_0}}| + C\gamma_T\delta_T^{-1/2}$. □

Lemma 3.3. *Under conditions (3.13) and (3.14) there exists* $1 \le r' \le q$ *such that* $|b - \eta_{p_0+r'}| \le \epsilon_T$, *where* b *is given in (3.15) and* $\epsilon_T = C\log^2 T$ *for a positive constant* C.

Proof. First, we mention that the model (3.2) can be written as $\tilde{Y}_{t,T} = \sigma(t/T)^2 + \sigma(t/T)^2(Z^2_{t,T}-1)$ which has the form of a signal+noise model i.e. $Y_t = f_t+\varepsilon_t$. Now, let $\bar{f}^d_{s_{m_0},e_{m_0}}$ define the best function approximation to f_t such that $\arg\max_d |\langle\psi^d_{s_{m_0},e_{m_0}}, f\rangle| = \arg\min_d \sum_{t=s_{m_0}}^{e_{m_0}} (f_t - \bar{f}^d_{s_{m_0},e_{m_0}})$ where $\bar{f}^d_{s_{m_0},e_{m_0}} = \bar{f} + \langle f, \psi^d_{s_{m_0},e_{m_0}}\rangle\psi^d_{s_{m_0},e_{m_0}}$, $\bar{f}$ is the mean of f and $\psi^d_{s_{m_0},e_{m_0}}$ is a set of vectors that are constant and positive until d and then constant and negative from $d+1$ until e_{m_0}.

If it can be shown that for a certain $\epsilon_T < C_2\gamma_T$, we have

$$\sum_{t=s_{m_0}}^{e_{m_0}} (Y_t - \bar{Y}^d_{s_{m_0},e_{m_0},t})^2 > \sum_{t=s_{m_0}}^{e_{m_0}} (Y_t - \bar{f}^{\eta_{p_0+r'}}_{s_{m_0},e_{m_0},t})^2 \tag{3.18}$$

as long as

$$\epsilon_T \leq |d - \eta_{p_0+r'}|$$

then this would prove necessarily that $|b - \eta_{p_0+r'}| \leq \epsilon_T$.

By Lemma 3.2 and Lemma A.3 in Fryzlewicz (2014), we have the same triplet of inequalities with the argument in the proof of Theorem 3.2 in Fryzlewicz (2014) i.e.

$$|d - \eta_{p_0+r'}| \geq C(\lambda_2|d - \eta_{p_0+r'}|\delta_T^{-1/2}) \vee (\lambda_2|d - \eta_{p_0+r'}|^{-1/2}) \vee (\lambda_2^2). \tag{3.19}$$

Hence, with the requirement that $|d - \eta_{p_0+r'}| \leq C_2\gamma_T = C_2\lambda_1\sqrt{\delta_T}$ we obtain

$$\delta_T > C^2\lambda_2^2 \max(C^2C_2^{-2}\lambda_1^{-2}\lambda_2^2, 1)$$

and $\epsilon_T = \max(1, C^2)\lambda_2^2$. From Lemma 3.1 λ_1 is of order $\mathcal{O}(\log T)$. For λ_2, which appears in the following two terms of the decomposition of (3.18)

$$I = \frac{1}{d - s_{m_0} + 1}\left(\sum_{t=s_{m_0}}^{d} \varepsilon_t\right)^2 \quad \text{and} \quad II = \frac{1}{e_{m_0} - d + 1}\left(\sum_{t=d+1}^{e_{m_0}} \varepsilon_t\right)^2$$

we show below that with probability tending to 1, $I \leq \lambda_2^2 = \log^2 T$. From Lemma 3.1 we have that $c_t = 1$ for $t = s_{m_0}, ..., d$ and thus

$$\mathbb{P}\left(\frac{1}{\sqrt{d - s_{m_0} + 1}}\left|\sum_{t=s_{m_0}}^{d} \varepsilon_t\right| > \lambda_2\right) \to 0$$

since by the Bernstein inequality the probability is bounded by

$$2T^2 \exp\left(-\frac{(d - s_{m_0} + 1)\lambda_2^2}{4\max_z \sigma^2(z)(d - s_{m_0} + 1)\rho_\infty^2 + 2c' \max_z \sigma(z)\sqrt{d - s_{m_0} + 1}\lambda_2\rho_\infty^1}\right)$$
$$\leq 2T^2 \exp\left(-C_3'\lambda_2^2\right)$$

which converges to 0 due to $(d - s_{m_0} + 1) = \mathcal{O}(\delta_T)$ from (3.6). Note that II has similar order and we omit the details. This concludes the lemma. □

Lemma 3.4. *Under conditions (3.13) and (3.14)*

$$\mathbb{P}\left(|\mathbb{Y}^b_{s_{m_0},e_{m_0}}| > \omega_T \frac{\sum_{t=s_{m_0}}^{e_{m_0}} \tilde{Y}_t}{n_{m_0}}\right) \to 1$$

where b is given in (3.15).

Proof. We define the following two events $\mathcal{A} = \left\{|\mathbb{Y}^b_{s_{m_0},e_{m_0}}| < \omega_T \frac{1}{n_{m_0}} \sum_{t=s_{m_0}}^{e_{m_0}} \tilde{Y}_{t,T}\right\}$ and $\mathcal{B} = \left\{\frac{1}{n_{m_0}} \left|\sum_{t=s_{m_0}}^{e_{m_0}} \tilde{Y}_{t,T} - \sum_{t=s_{m_0}}^{e_{m_0}} \sigma(t/T)^2\right| < \bar{\sigma} = \frac{1}{2n_{m_0}} \sum_{t=s_{m_0}}^{e_{m_0}} \sigma^2(t/T)\right\}$.

Since $\mathbb{P}(\mathcal{A}) \leq \mathbb{P}(\mathcal{A} \cap \mathcal{B}) + \mathbb{P}(\mathcal{B}^c)$ we need to show that $\mathbb{P}(\mathcal{B}) \to 1$ and $\mathbb{P}(\mathcal{A} \cap \mathcal{B}) \to 1$. To show that $\mathbb{P}(\mathcal{B}) = \mathbb{P}\left(\frac{1}{n_{m_0}} \sum_{t=s_{m_0}}^{e_{m_0}} \tilde{Y}_{t,T} \in (\bar{\sigma}/2, 3\bar{\sigma}/2)\right) \to 1$ we apply the Bernstein inequality as in Lemma 3.1 and we have that

$$\begin{aligned}
\mathbb{P}(\mathcal{B}^\downarrow) &= \mathbb{P}\left(\frac{1}{n_{m_0}} \left|\sum_{t=s_{m_0}}^{e_{m_0}} \tilde{Y}_{t,T} - \sum_{t=s_{m_0}}^{e_{m_0}} \sigma(t/T)^2\right| > \bar{\sigma}\right) \\
&= \mathbb{P}\left(\left|\sum_{t=s_{m_0}}^{e_{m_0}} \sigma(t/T)^2 (Z^2_{t,T} - 1)\right| > n_{m_0}\bar{\sigma}\right).
\end{aligned}$$

Hence,

$$\mathbb{P}(\mathcal{B}^\downarrow) \leq 2\exp\left(-\frac{n^2_{m_0}\bar{\sigma}^2}{4\max_z \sigma^2(z) n_{m_0} \rho^2_\infty + 2c' \max_z \sigma(z) n_{m_0} \bar{\sigma} \rho^1_\infty}\right) \leq 2T^2 \exp\left(-C'_4 \log^2 T\right)$$

which converges to 0 since $n_{m_0} \geq \delta_T = \mathcal{O}(\log^2 T)$ and $\rho^1_\infty < \infty$ from (A2). Now, from Lemma (3.3), we have some $\eta \equiv \eta_{p_0+r'}$ satisfying $|b - \eta| \leq C\epsilon_T$. Turning to $\mathbb{P}(\mathcal{A} \cap \mathcal{B})$ we have from conditions (3.13) and (3.14)

$$\begin{aligned}
|\mathbb{Y}^b_{s_{m_0},e_{m_0}}| \geq |\mathbb{Y}^\eta_{s_{m_0},e_{m_0}}| &\geq |\mathbb{S}^\eta_{s_{m_0},e_{m_0}}| - \log T \\
&= \left|\sqrt{\frac{(\eta - s_{m_0} + 1)(e_{m_0} - \eta)}{n_{m_0}}} \left(\sigma\left(\frac{\eta}{T}\right)^2 - \sigma\left(\frac{\eta+1}{T}\right)^2\right)\right| - \log T \\
&= \sqrt{\frac{e_{m_0} - \eta}{n_{m_0}(\eta - s_{m_0} + 1)}} (\eta - s_{m_0} + 1)\sigma_\star - \log T \\
&\geq C\sqrt{\delta_T} - \log T > \omega_T 3\bar{\sigma}/2.
\end{aligned}$$

□

Lemma 3.5. *For some positive constants C, C', let s, e satisfy either*

- $\exists 1 \leq p \leq N$ *such that* $s \leq \eta_p \leq e$ *and* $(\eta_p - s + 1) \wedge (e - \eta_p) \leq C\epsilon_T$ *or*
- $\exists 1 \leq p \leq N$ *such that* $s \leq \eta_{p+1} \leq e$ *and* $(\eta_p - s + 1) \vee (e - \eta_{p+1}) \leq C'\epsilon_T$.

Then,

$$\mathbb{P}\left(|\mathbb{Y}^b_{s_{m_0},e_{m_0}}| < \omega_T \frac{\sum_{t=s_{m_0}}^{e_{m_0}} Y_t}{n_{m_0}}\right) \to 1$$

where b is given in (3.15).

Proof. A similar argument with the proof of Lemma 3.5 is applied here. We only need to show that $\mathbb{P}(\mathcal{A} \cap \mathcal{B}) \to 0$ where now event $\mathcal{A} = \left\{|\mathbb{Y}_{s_{m_0},b,e_{m_0}}| > \omega_T \frac{1}{n_{m_0}} \sum_{t=s_{m_0}}^{e_{m_0}} \tilde{Y}_{t,T}\right\}$. Using condition (i) or (ii) we have that

$$\begin{aligned} |\mathbb{Y}^b_{s_{m_0},e_{m_0}}| &\leq |\mathbb{S}^b_{s_{m_0},e_{m_0}}| + \log T \\ &= \left|\frac{\sqrt{b - s_{m_0} + 1}\sqrt{e_{m_0} - b}}{\sqrt{n_{m_0}}} \left(\sigma^2(b/T) - \sigma^2((b+1)/T)\right)\right| + \log T \\ &\leq \sigma^* C\sqrt{\epsilon_T} + \log T < \omega_T \bar{\sigma}/2. \end{aligned}$$

□

The proof of Theorem 1 proceeds as follows: at the start of the algorithm when $s = 0$ and $e = T - 1$ all the conditions of (3.13) & (3.14) required by Lemma 3.4 are met and thus it detects a change-point on that interval defined by formula (3.15) within the distance of $C\epsilon_T$ (by Lemma 3.3). The conditions of Lemma (3.4) are satisfied until all change-points have been identified. Then, every random interval (s_m, e_m) does not contain a change-point or the conditions of Lemma 3.5 are met; hence no more change-points are detected and the algorithm stops.

Proof of Theorem 2

We start by the first method of aggregation. From the invertibility of the autocorrelation wavelet inner product matrix A, there exists at least one ordinate of wavelet periodogram in which a change-point θ_r is detected. From Theorem 1 it holds that $|\theta_r - \hat{\theta}_r| \leq C\epsilon_T$ with probability converging to 1 regardless of the scale i. Since the algorithm begins its search from the finest scale and only proceeds to the next one if no change-point is detected (until scale $I^\star$) then consistency is preserved.

We now turn to the second method of aggregation. We note that $\mathbb{Y}_t^{thr}$ has the same functional form with each of $\mathcal{Y}_t^{(i)}$ i.e. $h^{(i)}(x) = (x(1-x))^{-1/2}(c_x^{(i)}x + d_x^{(i)}x)$ for $x = (t - s_m + 1)/n \in (0, 1)$, where $c_x^{(i)}$, $d_x^{(i)}$ are determined by the location and the magnitude of the change-points of $I_{t,T}^{(i)}$. Let $b = \arg\max_{s_{m_0}<t<e_{m_0}} \mathbb{Y}_t^{thr}$; then following a similar argument with Lemma 2 of Fryzlewicz (2014) we can show that $\mathbb{Y}_t^{thr}$ must have a local maximum at $t = \theta_{p_0+r'}$ and that $|b - \theta_{p_0+r'}| \leq C_5\gamma_T$. With this result, we can show that $|b - \theta_{p_0+r}| \leq C'\epsilon_T$ for some $1 \leq r' \leq q$ as in Lemma 3.3 above by constructing a signal+noise model $y_t = f_t + \varepsilon_t$ and substituting f_t with $\sum_{i=-I^\star}^{-1} \mathbb{E}I_{t,T}^{(i)}\mathbb{I}(\mathcal{Y}_t^{(i)} > \omega_T^{(i)})/q_{s_m,e_m}^{(i)}$. Then, conditions (3.13) and (3.14) are satisfied within each segment for at least one scale $i \in \{-1, ..., -I^\star\}$. When all change-points have been detected every subsequent random interval (s_m, e_m) will satisfy the conditions of Lemma 3.5 for every $i \in \{-1, ..., -I^\star\}$ and the algorithm stops.

Finally, we examine whether condition A0, i.e. the bias present in $\mathbb{E}I_{t,T}^{(i)}$, will affect the consistency of the proofs above. Fryzlewicz and Nason (2006) - see Proposition 2.1 - show that the integrated bias between $\mathbb{E}I_{t,T}^{(i)}$ and $\beta_i(t/T)$ converges to zero.

We now define $\tilde{\mathbb{S}}_{s,e}^t$ similarly with $\mathbb{S}_{s,e}^t$ by replacing $\sigma(t/T)^2$ with $\sigma_{t,T}^2$. Assume that η_r is a change-point within the interval $[s_{m_0}, e_{m_0}]$ and $b = \arg\max_{t\in(s_{m_0},e_{m_0})} |\mathbb{S}_{s_{m_0},e_{m_0}}^b|$ and $\hat{b} = \arg\max_{t\in(s_{m_0},e_{m_0})} |\tilde{\mathbb{S}}_{s_{m_0},e_{m_0}}^b|$. Recall that $\mathbb{E}I_{t,T}^{(i)}$ is constant within each seg-

ment apart from short intervals around true change-point η_r i.e. $[\eta_r - K2^{-i}, \eta_r + K2^{-i}]$. In addition, from Theorem 2 in Cho and Fryzlewicz (2013) the finest scale should satisfy $i \geq I^\star = -\lfloor \alpha \log\log T \rfloor$ in order for (A4) to hold. Then, $|\hat{b} - b| \leq K2^{I^\star} < \epsilon_T$ holds since $I^\star = \mathcal{O}(\log\log T)$. Therefore, bias does not affect the results of the lemmas above and consistency is preserved.

Chapter 4

A fast implementation and a criticism of the fused lasso estimator

Introduction

An important problem in statistics is the estimation of a parameter, such as the mean, of a stochastic model that does not remain constant. In its simplest form, it entails removing the noise from a piecewise constant signal i.e. estimating a one-dimensional function μ from the noisy observations y_i in the following model

$$y_i = \mu_i + \varepsilon_i \tag{4.1}$$

where $\mu \in \mathbb{R}^n$ is the unknown vector of mean values with change-points whose number $\mathcal{N}$ and their locations $\mathcal{J} = \{\eta_1, ..., \eta_{\mathcal{N}}\}$ are unknown. Further, the noise ε_i is assumed to be iid Gaussian.

The problem of estimating the underlying function μ has attracted considerable

attention, mainly because piecewise stationarity is easier to interpret in the sense that the parameters of the process in every segment remain constant. In this chapter we are interested in methods where the estimation procedure has a "top-down" approach, i.e. starting from a single change-point and then progressively continuing to identify more. The Binary Segmentation method (BS) belongs to this category and it has been shown to perform well, both theoretically and practically, see Vostrikova (1981), Venkatraman (1992), Fryzlewicz (2007) or Fryzlewicz (2014). BS also has an interpretation in terms of "Unbalanced Haar" wavelets (Fryzlewicz (2007)) and inherits many features from the "multiscale" wavelet methods for which a representative example is the work by Donoho and Johnstone (1994). The authors propose the wavelet thresholding to estimate the model (4.1) by using the simplest form of wavelets, i.e. the Haar wavelets, and they show that the thresholded estimation is theoretically tractable. Kolaczyk and Nowak (2004) develop a recursive partitioning estimator noticing that there is a link to Unbalanced Haar wavelets and, hence, it is multiscale in nature. Another method with a top-down approach is the CART methodology of Breiman et al. (1983), an adaptive recursive partitioning which produces a piecewise constant reconstruction where the pieces are terminal nodes of the partition. The CART method is also used by Gey and Lebarbier (2008) who then prune the output change-points using an exhaustive search algorithm.

A different approach is to see the estimation of (4.1) as a problem where the purpose is to minimise a cost function such as the likelihood ratio. Methods based on this approach date back to Chernoff and Zacks (1964) and Kander and Zacks (1966), and have received significant attention afterwards by Worsley (1986), Siegmund (1988), Siegmund and Venkatraman (1995), Antoch and Hušková (2003), to name but a

few. However, these methods require a predefined maximal number of change-points. Since in most of the cases the true number of change-points is unknown, a penalty is typically added in order to control the total number of change-points. This penalty acts as a model selection criterion and prevents overfitting. Yao (1988) introduced BIC-type penalties, but other more sophisticated penalties have been proposed by e.g. Yao and Au (1989) and Lavielle and Moulines (2000); Birgé and Massart (2001) who use a generalisation of the C_p criterion (Mallows (1973)); Davis et al. (2006) who propose the Minimum Description Length Criterion (MDL), but in the context of change-point detection for non-stationary time series.

In order to solve these optimisation problems dynamic programming techniques are often adopted, see Bellman and Dreyfus (1966), Kay (1998), Jackson et al. (2005). Given that the complexity of $\mathcal{O}(n^2)$ is prohibitive for large samples Rigaill (2010), Killick et al. (2012) and Frick et al. (2014) include pruning steps into the dynamic program with the aim to reduce the computational burden under certain assumptions.

Another notable penalisation method is the method introduced by Mammen and van de Geer (1997), which uses a linear combination of the total variation and the L_1 penalty. It is of importance to notice that this method was later discovered by Friedman et al. (2007) and named fused lasso signal approximator (FLSA), but without references to the work by Mammen and van de Geer (1997). An algorithm for solving the total variation problem, termed taut string, already existed before FLSA and it was proposed by Davies and Kovac (2001). In Cho and Fryzlewicz (2011) the taut string method is shown to have a multiscale nature and from that perspective it can be also categorised as a top-down method. Another algorithm for solving the FLSA is developed by Tibshirani and Taylor (2011), which is also

the main topic of interest in this chapter. We do not argue that other algorithms should not be preferred. Davies and Kovac (2001), Friedman et al. (2007), Hoefling (2010) and Harchaoui and Lévy-Leduc (2010) all propose methods that solve the FLSA problem. However, the algorithm of Tibshirani and Taylor (2011) is designed to solve many other lasso-type problems and with this in mind we believe that we can shed more light into other set-ups.

One of the contributions in this chapter is to show a faster implementation of the algorithm by Tibshirani and Taylor (2011). This is achieved by replacing the matrix multiplications involved in their algorithm with simple cumulative summations in the spirit of "Mallat" pyramids (Mallat (1989)). In addition, we establish a link between their algorithm and the taut string technique of Davies and Kovac (2001). By doing so we are able to exploit the multiscale structure of the algorithm by Tibshirani and Taylor (2011) and to argue that trend filtering - a total variation technique which goes beyond the model (4.1) to assume that μ is piecewise linear (Kim et al. (2009a)), piecewise quadratic, piecewise cubic etc (Tibshirani (2014)) - can be also categorised as a multiscale method. Another contribution of this chapter is a result about the suboptimality of lasso-type estimators in change-point detection, an argument that has been made earlier by Brodsky and Darkhovsky (1993) and Cho and Fryzlewicz (2011). Here, we prove an exact rate of convergence for an estimated change-point and to support our argument we also provide a detailed simulation study by comparing the fused lasso estimator with the BS method.

This chapter is organised as follows. In Section 4.1 we present a fast version of the algorithm by Tibshirani and Taylor (2011). Then, we make a connection between the taut string method of Davies and Kovac (2001) and the algorithm by Tibshirani

and Taylor (2011) (Section 4.2). This is followed by a consistency result about the FLSA estimator for the model (4.1) with a single change-point (Section 4.3). In Section 4.4 we discuss model selection for the SPA method by taking advantage of its multiscale nature. A simulation study to assess the performance of the FLSA method in comparison with the Binary Segmentation method is presented in Section 4.5. In Section 4.6 we argue that faster versions to other set-ups using the algorithm by Tibshirani and Taylor (2011) are possible, but technically challenging. The penultimate section contains proofs related to the consistency theorem of the FLSA (Section 4.7). Finally, Section 4.8 establishes a bridge between this chapter and Chapter 5.

4.1 The solution path algorithm

4.1.1 The fused lasso estimator

Considering the model (4.1) we are now interested in estimating μ and, hence, producing estimates of the unknown partition $\mathcal{F}_j,\ j = 1, ..., \mathcal{N}$ by finding the number $\mathcal{N}$ and locations $\mathcal{J}$ of the change-points. One way of doing this is to minimise the following penalised cost function

$$\hat{\mu}^{FL} = \arg\min_{\mu \in \mathbb{R}^n} \frac{1}{2} \sum_{i=1}^{n} (y_i - \mu_i)^2 + \lambda_1 ||\mu||_1 + \lambda_2 ||\mu||_{TV} \tag{4.2}$$

where λ_1 and λ_2 are tuning parameters, and $||\mu||_1 = |\mu_1| + |\mu_2| + ... + |\mu_n|$. The total variation norm $||\mu||_{TV} = \sum_{i=2}^{n} |\mu_i - \mu_{i-1}|$ is particularly important for the recovery of the change-points. This type of penalty for signal estimation is found in Mammen and van de Geer (1997) and Davies and Kovac (2001) (but with $\lambda_1 = 0$), while Friedman et al. (2007) call it the FLSA and treat it as a special case of the fused lasso method of Tibshirani et al. (2005) used in the context of variable selection

by penalising neighbouring coefficients (in addition to the coefficients themselves). Further, it is common to examine the cost function without the λ_1 penalty and find the piecewise estimates $\hat{\mu}_i$ for some values of λ_2. Then, from Proposition 1 in Friedman et al. (2007), the fused lasso estimator $\hat{\mu}^{FL}$ can be obtained by soft-thresholding the individual coordinates $\hat{\mu}_i$ for a given value of λ_1 and, hence, we take $\lambda_1 = 0$ for the rest of this chapter.

Davies and Kovac (2001), Friedman et al. (2007) and Hoefling (2010) propose methods that solve problem (4.2). Here, we focus on the solution path algorithm (henceforth, SPA) of Tibshirani and Taylor (2011) which provides an exact solution to problems with the following form

$$\hat{\mu} \in \arg\min_{\mu \in \mathbb{R}^p} \frac{1}{2} \|y - X\mu\|_2^2 + \lambda_2 \|D\mu\|_1 . \tag{4.3}$$

For problem (4.2), $X = I \in \mathbb{R}^{n \times n}$ and

$$D = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix} \in \mathbb{R}^{(n-1) \times n}. \tag{4.4}$$

By considering the corresponding Lagrange dual problem of (4.3), which is conceptually clearer in that the L_1 penalty does not involve a linear transformation of μ, Tibshirani and Taylor (2011) devise the SPA method. We recall the details of the dual path algorithm, a "top-down" approach in estimating the knots of a signal (we use the term "knot" interchangeably with "change-point"). Tibshirani and Taylor (2011) (note that since $X = I$, X has full column rank, i.e. $\text{rank}(X) = n$) re-write the primal problem (4.3) into its Lagrangian form (taking $\lambda = \lambda_2$ for notational simplicity)

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

and the dual problem is derived by minimising $\mathcal{L}(\mu, z, u)$ over μ and z, i.e.

$$\min_{u \in \mathbb{R}^m} \frac{1}{2}||y - D^T u||_2^2 \quad \text{subject to} \quad \|u\|_\infty \leq \lambda \tag{4.5}$$

which is a regression problem with a simple constraint set. The algorithm starts from $\lambda_{\max} = (DD^T)^{-1}Dy$ (an unconstrained least squares estimate) and progressively identifies further knots until $\hat{\mu}_i = y_i \; \forall i = 1, ..., n$ and $\lambda = 0$. At the qth iteration the dual solution is given by

$$\hat{u}_{\lambda,\mathcal{B}} = \lambda_q \mathcal{S} \;\text{ for all }\; \lambda \in [0, \lambda_q]$$

where set $\mathcal{B}$ contains the coordinates (knots) that are currently on the boundary (called *boundary coordinates*), or the active set of the constraint $\|u\|_\infty \leq \lambda$. Alternatively, we can interpret $\mathcal{B}$ as the active set, the set which contains the estimated change-points. Finally, we denote with $\mathcal{S}$ the vector that contains the signs of $\hat{u}_{\lambda,\mathcal{B}}$. Now, since these do not change for decreasing λ_q, then we only need to find the "interior coordinates" $i \in -\mathcal{B}$, i.e. those dual coordinates that do not belong to the set $\mathcal{B}$ and lie strictly between $-\lambda_q$ and λ_q. These are found by

$$\hat{u}_{\lambda_q,-\mathcal{B}} = (D_{-\mathcal{B}}(D_{-\mathcal{B}})^T)^{-1} D_{-\mathcal{B}}(y - \lambda_q (D_{-\mathcal{B}})^T \mathcal{S}) \tag{4.6}$$

where $D_{-\mathcal{B}}$ denotes the penalty matrix which does not contain the row corresponding to point $i \in \mathcal{B}$. An important step of this algorithm is to determine the next point i that will be included in the set $\mathcal{B}$

$$i_{q+1} = \arg\max_i h_i \tag{4.7}$$

where

$$h_i = \frac{[D_{-\mathcal{B}}(D_{-\mathcal{B}})^T)^{-1} D_{-\mathcal{B}} y]_i}{[D_{-\mathcal{B}}(D_{-\mathcal{B}})^T)^{-1} D_{-\mathcal{B}}(D_{\mathcal{B}})^T \mathcal{S}]_i \pm 1} \tag{4.8}$$

and either -1 or 1 in (4.8) will yield a value in $[0, \lambda_q]$. Finally, the primal solutions are obtained by

$$\hat{\mu} = y - D^T \hat{u}. \tag{4.9}$$

The steps of the algorithm are summarised below:

Solution Path Algorithm for FLSA

- Set $\mathcal{B} = \emptyset$, $\mathcal{S} = \emptyset$, $\lambda = \infty$.
- For $q = 0, 1, ..., n-2$,
  1. Compute the solution at λ_q using (4.6).
  2. Find $\lambda_{q+1} = \max_i(h_i)$ where h_i as in (4.8).
  3. Locate the next knot i_{q+1} using (4.7).
  4. Add i_{q+1} to $\mathcal{B}$ and its sign to $\mathcal{S}$.
- Compute $\hat{\mu}$ using (4.9).

4.1.2 Fast implementation of the solution path algorithm

In this section, we introduce a more efficient implementation of the SPA method. The SPA method involves a heavy use of matrix multiplication which increases its complexity. Formula (4.8) contains the $D \in \mathbb{R}^{(m-1)\times n}$ matrix and, therefore the numerator requires $\mathcal{O}(n^2)$ operations. The same number of operations applies to the denominator. Tibshirani and Taylor (2011) report that the total complexity of their algorithm (presented here for the FLSA case) is $\mathcal{O}(qn^2)$ where q is the number of

iterations (notice that $q = n$ for the full path). Even if we assume that $q << n$, the complexity is again very high. Here, we show a way to reduce the complexity. However, regardless of the issue of speed we believe that a deeper understanding of the SPA algorithm will provide a better insight into not only the signal+noise problem but other more complex settings.

Our suggestion for a faster implementation of (4.8) is based on the fact that matrix multiplications can be replaced with simpler calculations if the matrices have special structures. For example, the pyramid algorithm of Mallat (1989) is widely used in the Discrete Wavelet Transformation (DWT) and it is preferred due to its low computational cost. The DWT of an input vector $y = \{y_i\}_{i=1}^n$ is the vector of inner products between y and $\psi^{j,r}$ for all j and r i.e.

$$\text{DWT}(y)^{j,r} = \left\langle y, \psi^{j,r} \right\rangle$$

where $\langle ., . \rangle$ denotes the inner product operation and $\psi^{j,r}$ are the wavelet vectors. If we define matrix $\mathcal{W}$ such that the $J + 1$ rows contain the wavelet vectors $\psi^{j,r}$ where $J = \log_2 N$, then the DWT can be conducted through the following matrix multiplication

$$\text{DWT}(y) = \mathcal{W}y$$

which typically requires $\mathcal{O}(n^2)$ operations, but thanks to pyramidal multiplication schemes and the way wavelets are constructed it only takes $\mathcal{O}(n)$ operations. However, it is not necessary for $\mathcal{W}$ to be a wavelet basis (the simplest one is the *Haar wavelet*) to take advantage of these fast multiplications. For example, Fryzlewicz (2007) proposes a method to construct an orthonormal Haar-like basis which, unlike the Haar wavelets, avoids the restriction of jumps in the basis function to occur in

the middle of their support. The basis $\mathcal{W}$ for a toy example when the sample size of a series y is 6 has the following form

$$\mathcal{W} = \begin{pmatrix} 6^{-1/2} & 6^{-1/2} & 6^{-1/2} & 6^{-1/2} & 6^{-1/2} & 6^{-1/2} \\ (5/6)^{1/2} & -30^{-1/2} & -30^{-1/2} & -30^{-1/2} & -30^{-1/2} & -30^{-1/2} \\ 0 & (3/10)^{1/2} & (3/10)^{1/2} & -(2/15)^{-1/2} & -(2/15)^{-1/2} & -(2/15)^{-1/2} \\ 0 & 2^{-1/2} & -2^{-1/2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 6^{-1/2} & 6^{-1/2} & -(2/3)^{1/2} \\ 0 & 0 & 0 & 2^{-1/2} & -2^{-1/2} & 0 \end{pmatrix}$$

where vectors $\psi^{j,r}$ are defined by the set of the change-points $(1, 3, 2, 5, 4)$. Of utmost importance is to select a suitable UH basis which amounts to choosing change-point $b^{j,r}$ for each vector $\psi^{j,r}$. One way of doing this is described in Fryzlewicz (2007): the first change-point is detected as the one that maximises $b^{0,1} = \arg\max_i |\langle y, \psi^{1,i,n}\rangle|$ where

$$\psi^{1,i,n} = \begin{pmatrix} (5/6)^{1/2} & -30^{-1/2} & -30^{-1/2} & -30^{-1/2} & -30^{-1/2} & -30^{-1/2} \\ 3^{-1/2} & 3^{-1/2} & -12^{-1/2} & -12^{-1/2} & -12^{-1/2} & -12^{-1/2} \\ 6^{-1/2} & 6^{-1/2} & 6^{-1/2} & -6^{-1/2} & -6^{-1/2} & -6^{-1/2} \\ 12^{-1/2} & 12^{-1/2} & 12^{-1/2} & 12^{-1/2} & -3^{-1/2} & -3^{-1/2} \\ 30^{-1/2} & 30^{-1/2} & 30^{-1/2} & 30^{-1/2} & 30^{-1/2} & -(5/6)^{1/2} \end{pmatrix}.$$

As mentioned above, the matrix multiplication would require $\mathcal{O}(n^2)$ operations, but due to the specific form of the UH vectors it can be computed in $\mathcal{O}(n)$ (in a similar manner with the computation of cumulative means of a vector with length n which also takes time $\mathcal{O}(n)$).

The question now is how we can utilise the above arguments to formula (4.8) with the aim to reduce the computational cost significantly. We notice that (4.8) contains inner products of the form $(D_{-\mathcal{B}}(D_{-\mathcal{B}})^T)^{-1}D_{-\mathcal{B}}y$. For example, at the initiation of the algorithm the first knot b is found by

$$b = |\arg\max_{i\in\{1,\ldots,n\}} (DD^T)^{-1}Dy|$$

since $\mathcal{B} = \emptyset$ and, hence, $D_{\mathcal{B}}$ in (4.8) contains only zeros.

We denote $\xi^{1,i,n}$ the quantity $(DD^T)^{-1}D$. To visualise its form and compare it

against $\psi^{1,i,n}$, we illustrate it through the toy example for $n = 6$; hence,

$$\xi^{1,i,n} = \begin{pmatrix} -5/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ -2/3 & -2/3 & 1/3 & 1/3 & 1/3 & 1/3 \\ -1/2 & -1/2 & -1/2 & 1/2 & 1/2 & 1/2 \\ -1/3 & -1/3 & -1/3 & -1/3 & 2/3 & 2/3 \\ -1/6 & -1/6 & -1/6 & -1/6 & -1/6 & 5/6 \end{pmatrix}.$$

The function $\xi^{1,i,n}$ differs from $\psi^{1,i,n}$ in that the former is fixed and only depends on the sample size n; whereas $\psi^{1,i,n}$ is only one way to construct an UH basis i.e. one can choose any $b_{j,k}$ and still obtain a UH basis. However, we can still use the same argument: the inner product between y and $\xi^{1,i,n}$ can be re-written as the cumulative mean of a vector of length n which only requires $\mathcal{O}(n)$ operations, i.e.

$$\left\langle y, \xi^{1,i,n} \right\rangle = \frac{i-n}{n} \sum_{u=1}^{i} y_u + \frac{i}{n} \sum_{u=i+1}^{n} y_u. \tag{4.10}$$

The SPA method proceeds repeatedly to identify the next knot i. Previous knots determine the segment in which the formula (4.8) is applied. The multiscale nature of the algorithm allows us to divide the problem into two sub-problems like in the binary segmentation algorithm. Hence, the locating function (4.11) is calculated on a smaller segment, i.e. only a specific segment needs to be updated every time. To be more precise, let us assume that at iteration q the knots that have been added to the set $\mathcal{B}$ are $\{b^1, b^2, b^3, ..., b^\kappa\}$ where $\kappa \in \{1, ..., K\}$, $K < n$ and n is the sample size. Also assume that the last knot added to $\mathcal{B}$ is b^h. Normally, to find the next knot we need to calculate formula (4.8). However, this can be avoided by noticing that inner products of every cycle will remain identical except from ξ_{b^l+1,i,b^r}, where b^l and b^r are the knots to the left and to the right of b^h respectively.

In particular, for a generic interval $[s, e]$ (at the initiation of the procedure, $s = 1$

and $e = n$) we calculate the locating function

$$h_i = \frac{[\langle y, \xi^{s,i,e} \rangle]_i}{[\langle \mathcal{G}, \xi^{s,i,e} \rangle]_i + g_i} \quad \text{for} \quad i \in [s, e] \tag{4.11}$$

where $\mathcal{G}$ is a vector whose elements $\mathcal{G}_{i_q} = [\text{sign}(\langle y, \xi^{s,i,e} \rangle)]_{i_q}$ and zero otherwise. This locating function replaces (4.8) in SPA. Finally, (4.6) also contains an expression of the form $(DD^T)^{-1}D\alpha$ where $\alpha \in \mathbb{R}^{n \times 1}$ and the same cumulative sum technique is applied to it. If our interest is the detection of the knots and not the estimation of μ itself, then the calculation of (4.6) can be omitted.

The steps of our implementation of the SPA algorithm are given below.

Solution Path Algorithm for FLSA without matrix operations

- Set $\mathcal{B} = \emptyset$.
- Find the first knot $b = \arg\max_{i \in \{1,\dots,n\}} | \langle y, \xi^{1,i,n} \rangle |$ where $\langle y, \xi^{1,i,n} \rangle$ as in (4.10).

1. Set $s = 1$ and $e = b$.
2. Locate the next knot $b = \arg_{i \in (s,e)} \max_{g_i = \pm 1} h_i$ where h_i as in (4.11).
3. Add b to $\mathcal{B}$.
4. Repeat steps 2 and 3 to the segment $s = b + 1$ to $e = n$.

- Repeat steps 1-4 until $\forall i \in \mathcal{B}$

4.1.3 Computational Complexity

We notice that the main computational burden stems from the matrix multiplication involved in the main formula. Taking the simple example of the piecewise constant signal+noise problem then the operations required to find a knot that joins $\mathcal{B}$ are of order $\mathcal{O}(n)$ and therefore, to find the first q knots $\mathcal{O}(qn)$ operations are required. If we are interested in obtaining the solution at all values of the tuning parameter λ_2 (the solution path), the calculation from the first knot to join the active set $\mathcal{B}$ until all n of them join $\mathcal{B}$ will increase the complexity of our method to $\mathcal{O}(n^2)$. Finally, the action of dividing the problem into two sub-problems reduces the computational time in practice for both our implementation of SPA and the original algorithm, but not the overall complexity.

4.2 The solution path algorithm and its connection with the multiscale taut string method

Penalised least squares methods where the penalty term uses the total variation norm have been widely used previously. We refer the reader to Mammen and van de Geer (1997) who propose the locally adaptive regression splines. These estimators penalise the total variation of the kth derivative and, hence, when $k = 0$ this gives the fused lasso estimator (4.2). A practically very similar method (termed trend filtering) to the locally adaptive regression splines has been introduced by Tibshirani (2014).

The taut string method (TS) is an alternative technique to solve the optimisation problem (4.2) for $\lambda_1 = 0$. Davies and Kovac (2001) develop a method that works as follows. Define the integrated process $\mathcal{Y}(i/n) = \sum_{v=1}^{i} y_v$ (with $\mathcal{Y}(0) = 0$) and a tube with lower and upper bounds $L^n = \mathcal{Y}(i/n) - \gamma$ and $U^n = \mathcal{Y}(i/n) + \gamma$ respectively

for $\gamma > 0$ (where $\gamma = \lambda_2$). A piece of string $f : [0,1] \to \mathbb{R}$ is attached such that it connects $(0, \mathcal{Y}(0))$ and $(1, \mathcal{Y}(1))$. Now, the string is pulled until it is taut, while being constrained to lie between the boundaries of the tube, touching either side at possibly multiple knots. The taut string f_n has the smallest length and also minimises the total variation

$$TV(f_n) = \int_0^1 |f_n(t)^{(1)}| \mathrm{d}t$$

where $f_n(t)^{(1)}$ denotes the derivative of $f_n(t)$, such that $f_0 = \mathcal{Y}(0), f_n = \mathcal{Y}(n), L^n \leq f_t \leq U^n$. Starting from left to right the TS algorithm simultaneously computes the greatest convex minorant of U_n (between two knots at which the string only touches U_n) and the least concave majorant of L_n (between two knots at which the string only touches L_n). Finally, at points where the string switches from touching U_n (L_n) to touching L_n (U_n) the derivative $f_n(t)^{(1)}$ has a local maximum (minimum).

Since SPA and TS are two methods that solve the same problem (4.2) it is of interest to examine their connection. We start from SPA and consider the simple case where we look at the first knot joining the active set. From the dual problem (4.5) we see that u can be estimated from least squares under the constraint set $\{u : ||u||_\infty \leq \lambda_2\}$. This means that the estimated dual variable u must satisfy element-wise

$$-\lambda_2 \leq \hat{u} \leq \lambda_2$$

or

$$-\lambda_2 \leq (DD^T)^{-1} Dy \leq \lambda_2$$

and using (4.10)

$$-\lambda_2 \leq \frac{i-n}{n} \sum_{v=1}^{i} y_v + \frac{i}{n} \sum_{v=i+1}^{n} y_v \leq \lambda_2$$

$$\sum_{v=1}^{i} y_v - \lambda_2 \leq \frac{i}{n} \sum_{v=1}^{n} y_v \leq \sum_{v=1}^{i} y_v + \lambda_2 \quad \text{for} \quad i = 1, ..., n. \tag{4.12}$$

The limits of the inequality (4.12) define the tube while the term $\frac{i}{n} \sum_{v=1}^{n} y_v$ is the "taut string", see also Figure 4.1 which illustrates this for a model with a single change-point at $\frac{i}{n} = 2/3$. However, for small values of λ_2 we cannot solve (4.8) by defining a tube and its associated string. This is because in order to solve this problem for any given value of λ_2 the previous knots must be given in hand. Hence, for a certain value λ_2^0 a solution always requires $\mathcal{O}(qn)$ for SPA, while TS is a faster algorithm since a solution is obtained in linear time.

We note that Cho and Fryzlewicz (2011) show that the TS algorithm also has a multiscale nature where the "parent" segment is split into two "children" subsegments. The authors therefore refer to their version of this algorithm as "multiscale TS". They define the "locating function" used to find the location of change-points (knots) in a given segment. The locating function is equivalent to (4.11) [and (4.10) for the first knot only] and, therefore the SPA algorithm is the same with the "multiscale TS".

4.3 Lack of sign consistency of the FLSA estimator

Cho and Fryzlewicz (2011) argue that the total variation method is suboptimal in detecting both the number and the locations of the change-points in the model (4.1). Their argument is based on a theorem about a family of test statistics for change-point detection by Brodsky and Darkhovsky (1993). In this section, we aim to find an exact rate of convergence for the estimated location of a change-point and, hence, claim that the FLSA estimator cannot recover its exact location. Even though sign

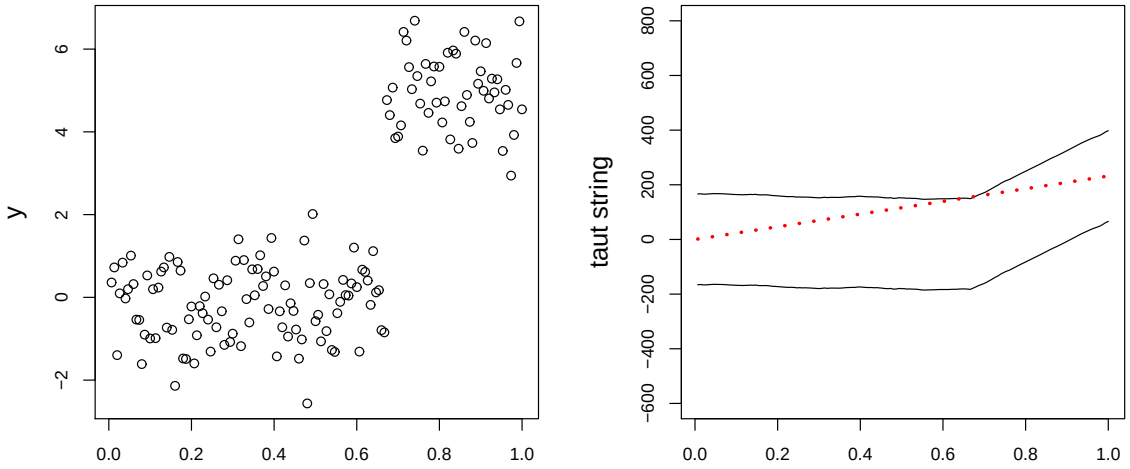

Figure 4.1: Left panel: a simulated data set with a change-point at $\frac{i}{n} = 2/3$. *Right panel: at the initiation of the procedure the solution path algorithm defines a tube (the black, symmetrical lines) and a string (red line), pulled until it is taut. The dotted red line coincides with the greatest convex minorant of* U_n *which in this case is a linear function because the tube is "squeezed" until it touches the first knot.*

consistency can hold under e.g. the "irrepresentable condition" of Meinshausen and Yu (2009) for the lasso problem in the context of variable selection, here we argue that for FLSA it does not, i.e. $\mathbb{P}(\{\hat{\mathcal{N}} = \mathcal{N}\} \cap \{\text{sign}(\hat{\mu}_i - \hat{\mu}_{i-1}) = \text{sign}(\mu_i - \mu_{i-1})\})$ does not tend to 1 as $n \to \infty$. This assertion invalidates Theorem 2.5 of Rinaldo (2009) who claims that the exact recovery of the change-points in the model (4.1) is feasible with high probability. This erroneous result has also been noted out by Rojas and Wahlberg (2014).

We consider a model with a single change-point η i.e.

$$y_i = \begin{cases} \mu_0 + \varepsilon_i, & \text{for } 1 \leq i \leq \eta \\ \mu_1 + \varepsilon_i, & \text{for } \eta + 1 \leq i \leq n \end{cases} \tag{4.13}$$

The following lemma gives the FLSA estimates for the two segments in the model (4.13).

Lemma 4.1. *For the model (4.13) and $\mu_0 < \mu_1$ the estimated μ_0 and μ_1 are given respectively by*

$$\hat{\mu}_0 = \frac{\sum_{i=1}^{\hat{\eta}} y_i}{\hat{\eta}} + \frac{\lambda_2}{\hat{\eta}}, \text{ for } i = 1, ..., \hat{\eta} \tag{4.14}$$

$$\hat{\mu}_1 = \frac{\sum_{i=\hat{\eta}+1}^{n} y_i}{n - \hat{\eta}} - \frac{\lambda_2}{n - \hat{\eta}}, \text{ for } i = \hat{\eta} + 1, ..., n. \tag{4.15}$$

Proof: See Section 4.7.

From these two equations we see that the FLSA estimator introduces a bias into the mean of segment $\hat{\mathcal{F}}_1$ ($\hat{\mathcal{F}}_2$) equal to $\lambda_2/\hat{\eta}$ ($\lambda_2/(n - \hat{\eta})$) where its sign depends on whether μ_0 (or μ_1) is a local maximum or minimum. As it is apparent from the formulation, the bias increases as $\lambda_2 \to \infty$ and $\hat{\eta} \to 0$.

We make the following assumptions

Assumptions 4.1.

1. *The random sequence $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, 1)$.*

2. *The sequence $\{\mu_i\}_{i=1}^n$ is bounded, i.e. $|\mu_0|, |\mu_1| < \infty$ for $i = 1, ..., n$.*

3. *The magnitude of the jump between μ_0 and μ_1 satisfies $|\mu_0 - \mu_1| \geq \underline{\mu}$ where $\underline{\mu} \geq C_0/n^\omega$, with $\omega \geq 0$.*

4. *The distance between η and either $\eta_0 = 0$ or $\eta_{\mathcal{N}} = n$ is at least $\delta_n \geq C_1 n^\theta$ and $\theta \leq 1$.*

5. *The parameters satisfy* $\theta - \omega > 1/2$.

The reason we choose $\text{var}(\varepsilon_i) = 1$ is purely for simplicity and in practice it can be accurately estimated, for example, by the Median Absolute Deviation (Hampel (1974))

$$\sigma_n = \frac{1.48}{\sqrt{2}}\text{Median}\{|y_2 - y_1)|, ..., |y_n - y_{n-1}|\}. \tag{4.16}$$

In addition, Assumption 4.1(4) ensures that the change-point η is not too close to the start or the end of the series. In the multiple change-point setting this assumption establishes the minimum distance between successive change-points.

The following consistency theorem holds.

Theorem 4.1. *Let* y_i *follow model (4.13), and suppose that Assumptions 4.1 hold. Let* $\hat{\eta}$ *be the estimated change-point by the FLSA and* $\lambda_2 \geq \sqrt{2n \log n}$ *where* λ_2 *as in (4.2). Then there exists a positive constant* C *such that* $\mathbb{P}(\mathcal{U}_n) \to 1$, *where*

$$\mathcal{U}_n = \{|\hat{\eta} - \eta| \leq C\epsilon_n\}$$

with $\epsilon_n = n^{3/2}\sqrt{\log n}\underline{\mu}^{-1}\delta_n^{-1}$.

The proof of Theorem 4.1 is given in Section 4.7. We elaborate on the rate of convergence. When δ_n is of $\mathcal{O}(n)$, then ϵ_n is of $\mathcal{O}(\sqrt{n})$ whereas in rescaled time ϵ_n/n is $\mathcal{O}(1/\sqrt{n})$ and, hence, the rate is suboptimal. On the contrary, from Fryzlewicz (2014) the BS method achieves a near-optimal rate of $\mathcal{O}(\log n/n)$ when δ_n is again $\mathcal{O}(n)$.

In Section 4.5 we discuss the performance of the FLSA estimator and we provide numerical evidence through finite sample size examples that the FLSA is suboptimal in detecting the locations of change-points. In fact, the simulation study indicates that BS does better in locating change-points.

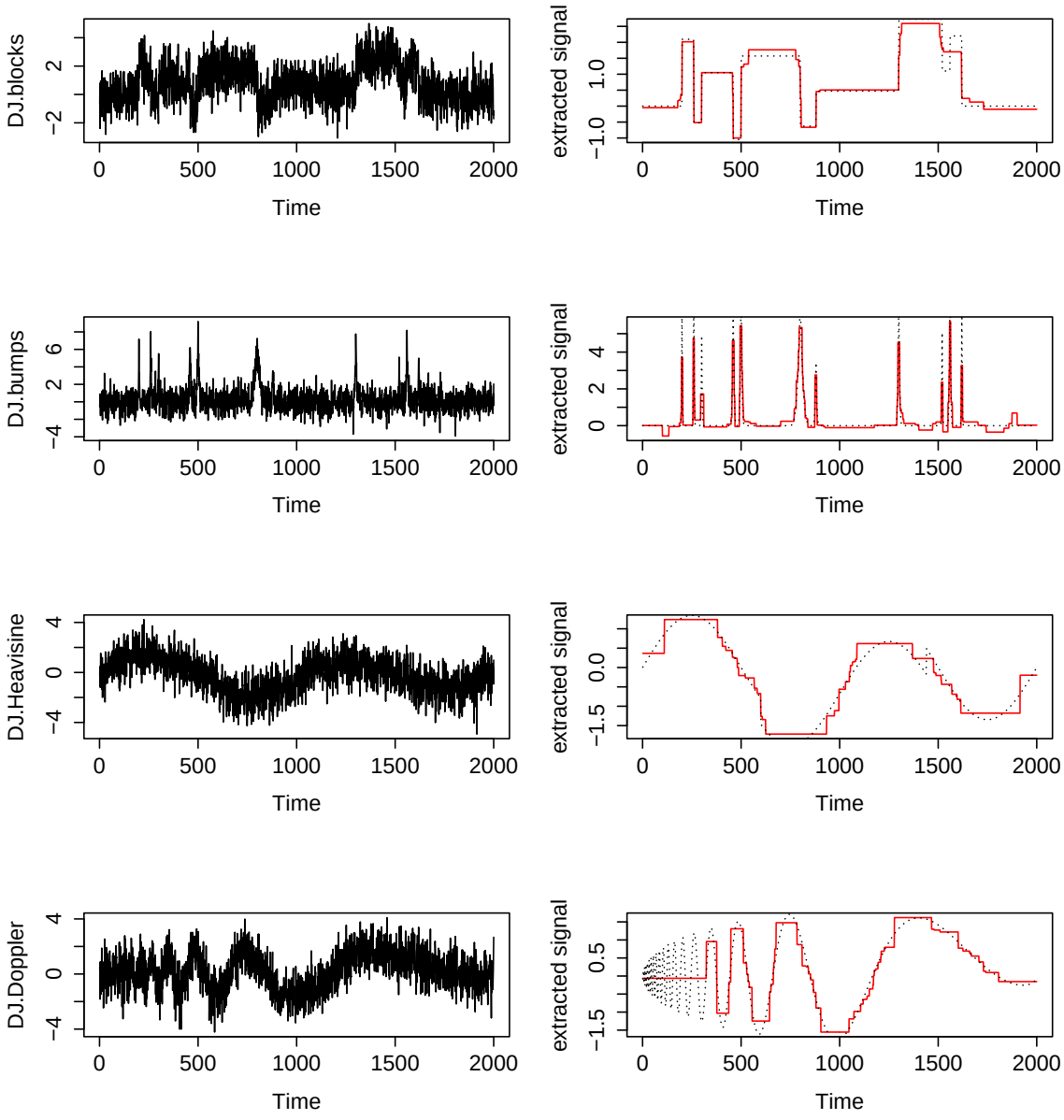

Figure 4.2: Extracting signal using the Fused Lasso estimator with a multiresolution criterion (right - red line). The real signal (right - black, dotted line) is the DJblocks data contaminated with white noise with $\sigma = 3$ (left).

4.4 Model selection

In this section we briefly discuss optimal stopping for the SPA method. Tibshirani and Taylor (2011) propose the use of information criteria that account for error reduction and penalise for over-segmentation. Since the SPA method progressively identifies the knots in the model (4.1) an information criterion is monotonically related with the decrease of λ_2. Therefore, at every iteration the algorithm may include an extra step for calculating a value for the information criterion. The algorithm stops as soon as it gets a value larger or smaller compared with the previous cycle.

An alternative approach is to take advantage of the multiscale nature of SPA and stop the path as soon as the obtained residuals "look" like white noise. Davies and Kovac (2001) utilise the multiresolution criterion by estimating the multiresolution coefficients $w_{j,r}$

$$w_{j,r} = 2^{-j/2} \sum_{i=r2^j+1}^{(r+1)2^j} \hat{\varepsilon}_i$$

if n is a power of two; otherwise the interval $[(r2^j+1)/n, (k+1)2^j/n]$ can be replaced by $[(r2^j+1)/n, \min\{(k+1)2^j/n, 1\}]$ (Davies and Kovac (2001)).

Then, the residuals can be adequately approximated by Gaussian white noise (see e.g. Lemma 4.2 in Section 4.7) if

$$|w_{j,r}| \leq \sigma_n \sqrt{\tau \log n} \tag{4.17}$$

where σ_n is some measure of the scale of the residuals, such as (4.16).

Figure 4.2 shows the estimated signals for the DJ data found in the *R* package *wavethresh* and contaminated with iid Gaussian noise with $\sigma = 3$. We set $\tau = 2.5$ as Davies and Kovac (2001) suggest that it gives good results.

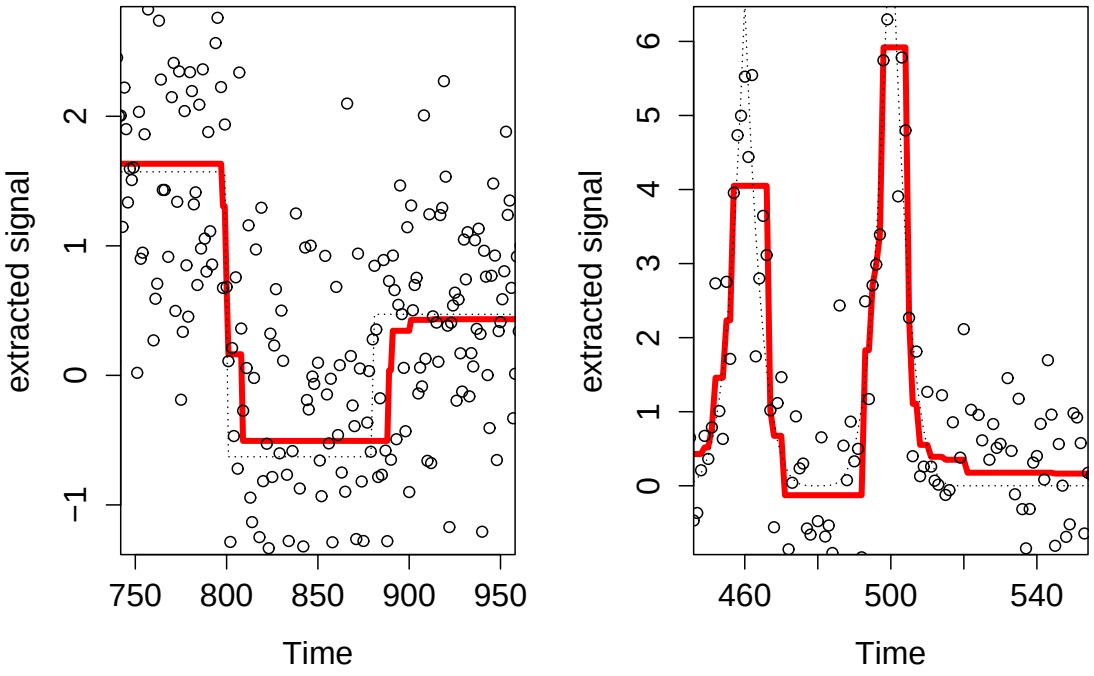

Figure 4.3: Zoomed in version of the blocks *(left) and* bumps *(right) series. The signals are estimated using the SPA method (red line). The real signal is shown by a black dotted line.*

4.5 Simulation study

Introduction

A careful inspection of the estimated signals of Figure 4.2 indicates that the FLSA method introduces a bias in the estimated piecewise intervals, see also Figure 4.3. Recall from Section 4.3 that the magnitude of the bias depends on the size and the location of the segments. To circumvent this one can estimate the mean of y_i between the change-points. To some extent this can improve the performance in the ℓ_2 sense, i.e. $\|\mu - \hat{\mu}\|_2$, as we discuss in Section 4.5.2. We refer to this estimator as mFLSA.

From Figure 4.3 it is also evident that the FL estimator tends to multi-segment

the series around the true change-points. This has also been noticed by Davies and Kovac (2001) (in the context of the taut-string method) and Harchaoui and Lévy-Leduc (2010) who treat the estimation as a lasso problem. The authors use the Least Angle Regression algorithm (LARS) of Efron et al. (2004) and they impose an upper bound $\bar{\mathcal{N}}$ for the number of change-points due to the fact that the true number $\mathcal{N}$ is not known. Their algorithm has a complexity of $\mathcal{O}(\bar{\mathcal{N}} n \log(n))$ which is slower than our approach by a logarithmic term (assuming that we also impose an upper bound of the maximum number of change-points; see also Section 4.1.3). A post-selection method is suggested by the authors which selects those change-points that have the most significant reduction in the variance of the error. To perform this post-selection they use a dynamic programming algorithm, originally proposed by Fisher (1958) and Bellman (1961). With the reduced set of change-points, the complexity of the post-selection procedure is $\mathcal{O}(\bar{\mathcal{N}}^3)$ resulting in a total complexity of $\mathcal{O}(\bar{\mathcal{N}}^3 + \bar{N} n \log(n))$. However, the authors do not provide a consistency result for this hybrid method, whereas the post-processing will increase the computational complexity if $\mathcal{N}$ is allowed to increase with the sample size.

For this reason, we choose not to evaluate the performance of the FLSA estimator based on the estimated change-point vis-á-vis the real ones. Instead, we conduct a study that examines the performance in terms of the estimated location of a single change-point, i.e. how far an estimated change-point is from the real one (see Section 4.5.1), as well as in the ℓ_2 sense (see Section 4.5.2).

Finally, to enable comparison with other methods we choose the Binary Segmentation method which is computationally fast, theoretically consistent and has good performance in numerical simulation studies, see Fryzlewicz (2014).

4.5.1 Location accuracy performance

In Section 4.3 we showed that the FLSA estimator is suboptimal in detecting the location of a change-point. Here, we provide a numerical study to support this claim. To achieve this we consider the following two models

$$y_i = \begin{cases} \mu_0 + \varepsilon_i, & \text{for } 1 \le i \le \lfloor n/3 \rfloor \\ \mu_1 + \varepsilon_i, & \text{for } \lfloor n/3 \rfloor + 1 \le i \le n \end{cases} \tag{4.18}$$

$$y_i = \begin{cases} \mu_0 + \varepsilon_i, & \text{for } 1 \le i \le \lfloor n/3 \rfloor \\ \mu_1 + \varepsilon_i, & \text{for } \lfloor n/3 \rfloor + 1 \le i \le \lfloor 2n/3 \rfloor \\ \mu_2 + \varepsilon_i, & \text{for } \lfloor 2n/3 \rfloor + 1 \le i \le n \end{cases} \tag{4.19}$$

where $(\mu_0, \mu_1, \mu_2) = (0, 1, 1.5)$, $\varepsilon_i \sim \mathcal{N}(0, 1)$, the sample size n ranges from 200 to 2000 and we repeat the experiment $B = 100$ times for every specific sample size. For both the models (4.18) and (4.19) we examine the ability of FLSA and BS in locating the first change-point η_1, i.e. the jump between μ_0 and μ_1. This is due to our earlier observation that multisegmentation does not allow to pin the locations of the estimated change-points exactly. Hence, we only consider the first knot to be returned from SPA. Finally, we use the following metric to assess the performance of BS and FLSA

$$\text{MSE} = \sum_{\ell'=1}^{B} (\eta_1 - \hat{\eta}_1^{(\ell')})^2 / B$$

where $\hat{\eta_1}$ is the estimated change-point obtained from either of the two methods.

Figure 4.4 indicates that BS does well in both models and particularly in the model (4.19), where two change-points are present, the BS method significantly outperforms FLSA.

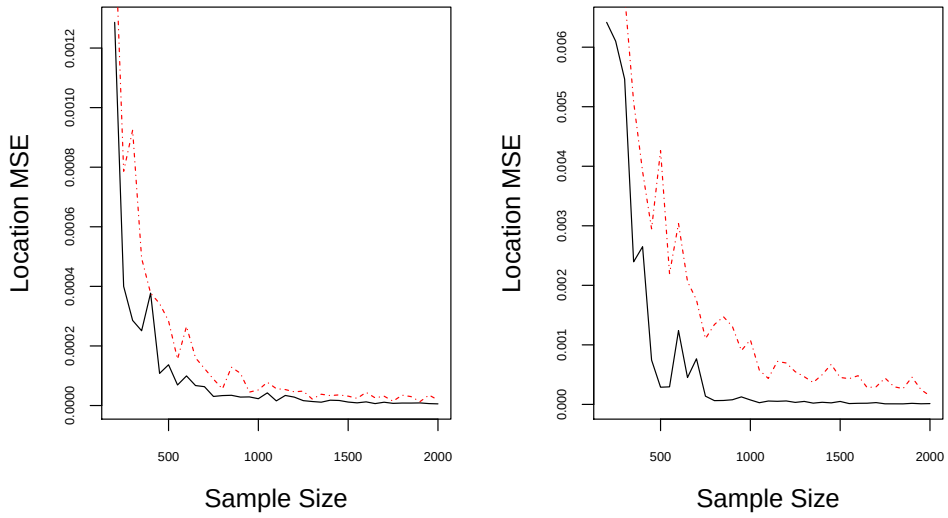

Figure 4.4: MSE calculated for increasing sample size. The left panel shows the performance of BS (black line) and FLSA (red dotted line) on the first change-point of the model (4.18). The right panel is for the model (4.19) and for the first change-point only.

4.5.2 Multiple change-point performance in the ℓ_2 sense

We conduct another simulation study to compare the performance of the FLSA estimator and the BS method on simulated data. We take the underlying functions to be the DJ data (*djdata*) and test the consistency of the two methods in the ℓ_2 sense. The reason we choose this measure is to examine whether the FLSA method can do better on noisy signals where "peaks" (big jumps of small magnitude) are observed (the *bumps* data), on smooth signals (the *heavisine* data) or on wiggly signals (the left part of the *doppler* data). For a fair comparison between the FLSA, mFLSA and BS methods, we need to control that a stopping rule will not impact the performance. This is why we allow the algorithms to run assuming that there are many change-points. Given that in practice either we do not know the real number of change-points in a signal or a signal has smoother transitions (heavisine and doppler data) we are

able to perform a knot-by-knot (change-point) comparative study between the three methods. We repeat the experiment 100 times and report the mean and plus or minus one standard deviation, see Figures 4.5 - 4.7. For robustness, we examine their performance for three different signal-to-noise-ratio scenarios by contaminating the series with iid Gausian noise with mean zero and $\sigma = 1, 2, 3$.

For all the simulated examples the BS method clearly outperforms both FLSA and mFLSA, see Figures 4.5, 4.6 and 4.7. Particularly, in signals with a "blocky" structure (blocks and bumps data) BS achieves a low value of MSE even in the very noisy scenario (top panels in Figure 4.7). It is also interesting that BS outperforms FLSA and mFLSA in the signals with smoother transitions (heavisine and doppler data) and in the high SNR cases it is competitive with the other two methods when increased flexibility is allowed. In addition, as noticed above mFLSA improves the estimation performance of the FLSA estimator in every scenario.

4.6 Extensions to other settings

In this section, we extend our previous arguments about algorithm's computational complexity to other set-ups. In the next lines we present possible extensions of SPA and we see that the locating functions can be found even though they are difficult to obtain.

4.6.1 The two-dimensional FLSA case

We start our discussion with the denoising of an image or the $2d$ FLSA problem. In this setting, the FLSA estimator penalises the differences between adjacent pixels.

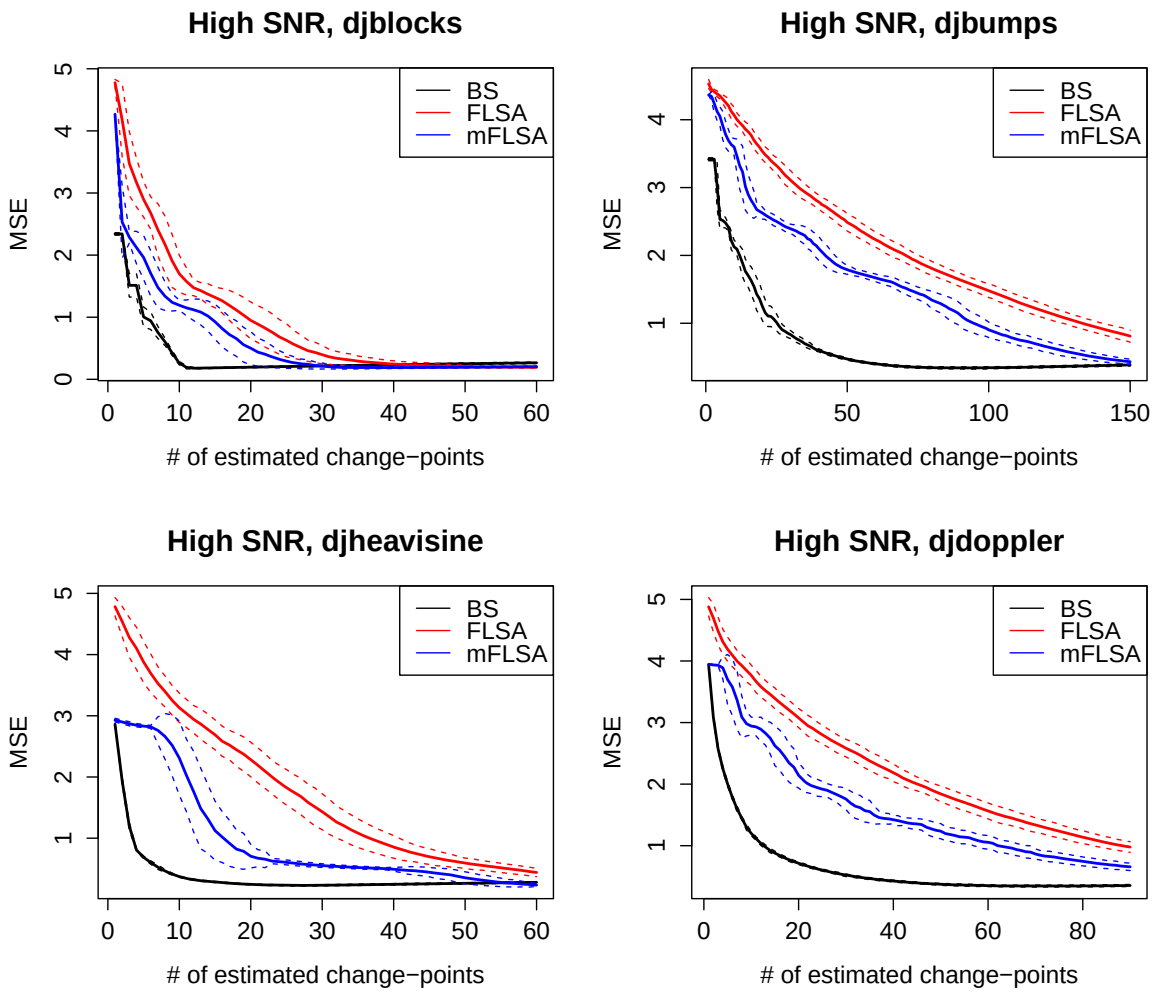

Figure 4.5: Shown is the squared error loss ($\|\mu - \hat{\mu}\|_2$) in predicting the true function μ averaged over 100 *simulated data sets where the signals are the DJ data. The red curves display the loss for FLSA, the blue for mFLSA and the black for BS. The dotted lines are the standard deviations. All simulations are based on a high signal-to-noise ratio.*

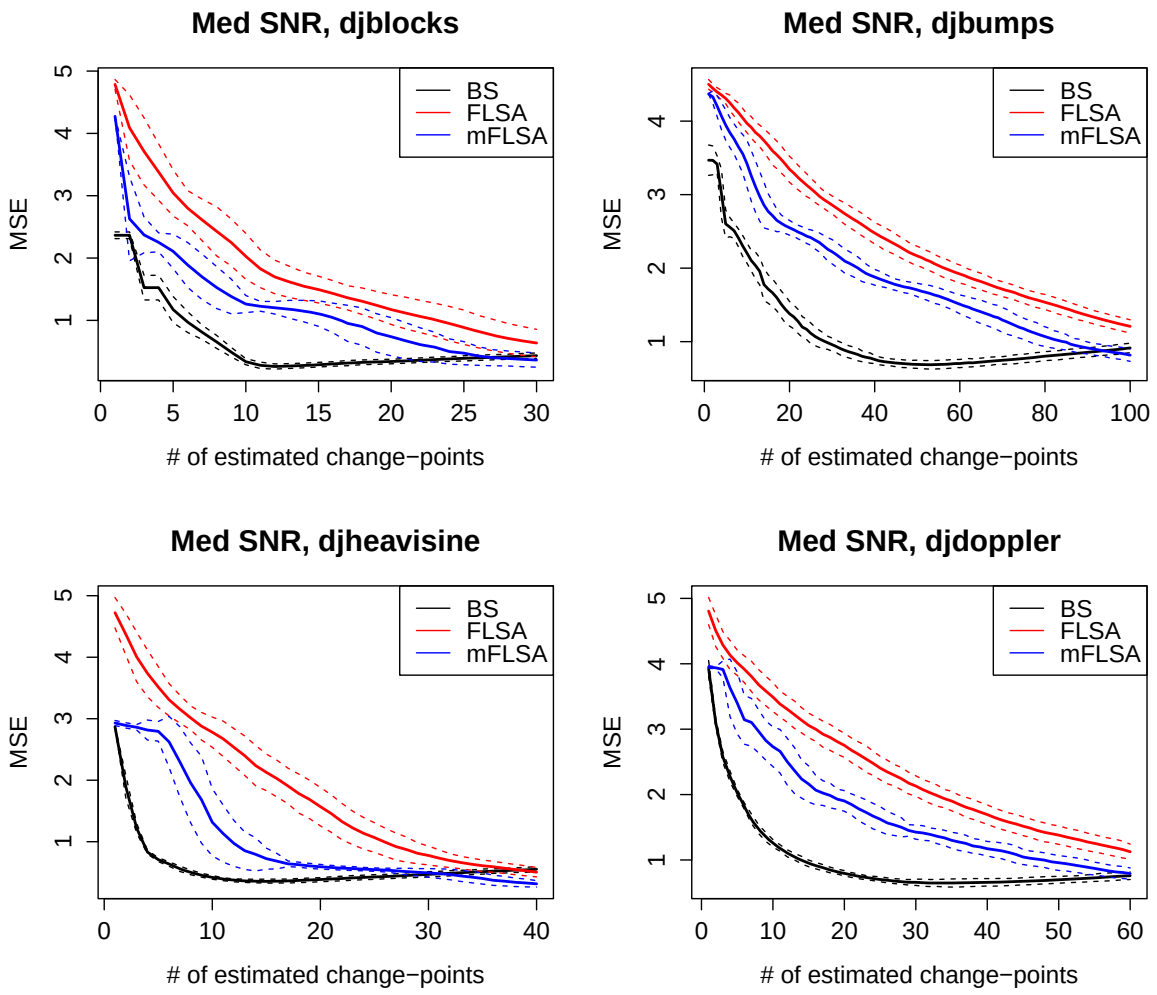

Figure 4.6: Shown is the squared error loss ($\|\mu - \hat{\mu}\|_2$) in predicting the true function μ averaged over 100 *simulated data sets where the signals are the DJ data. The red curves display the loss for FLSA, the blue for mFLSA and the black for BS. The dotted lines are the standard deviations. All simulations are based on a medium signal-to-noise ratio.*

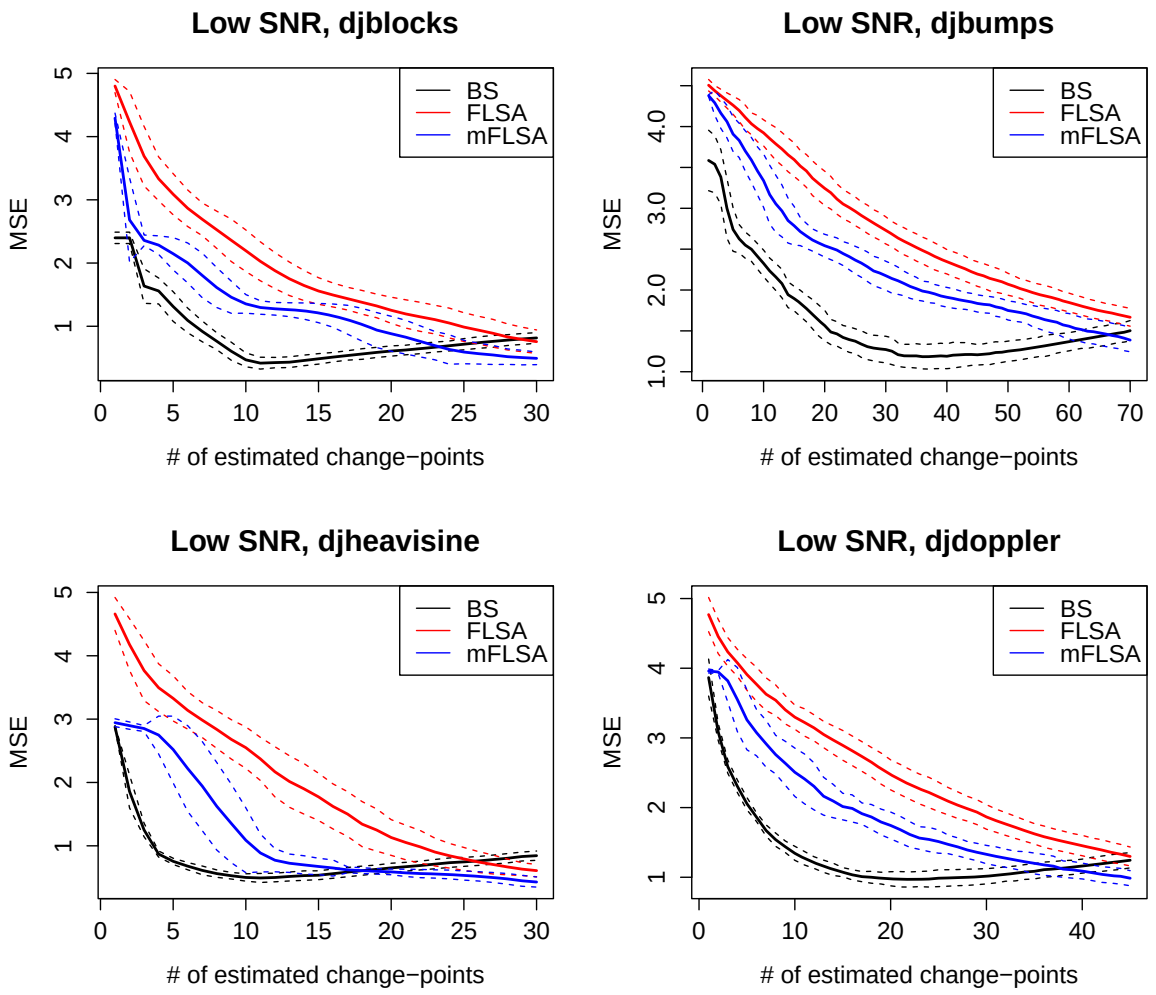

Figure 4.7: Shown is the squared error loss ($\|\mu - \hat{\mu}\|_2$) in predicting the true function μ averaged over 100 *simulated data sets where the signals are the DJ data. The red curves display the loss for FLSA, the blue for mFLSA and the black for BS. The dotted lines are the standard deviations. All simulations are based on a low signal-to-noise ratio.*

Penalty matrix $D^{2\text{d}}$ has a similar structure with (4.4), i.e. every row contains a 1 and -1, but arranged such that the differences are not only horizontal but vertical, i.e.

$$D^{2\text{d}} = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 & -1 & 1 \\ -1 & 0 & 0 & \cdots & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & \cdots & 0 & 0 & 1 \end{pmatrix}.$$

We have not established a closed-end formula in the sense of (4.10) and (4.11) for the 2d FLSA. However, Fryzlewicz (2007) argues that a top-down approach would be less suitable for image denoising due to the fact that the particular form of the basis functions would result in undesirable "blocky" artefacts. Hence, a bottom-up approach, i.e. searching all the pairs of neighbours for the smallest detail coefficient, would be more efficient. A natural choice for image denoising in the context of total variation penalty is the work by Hoefling (2010) and Kovac and Smith (2011).

4.6.2 The piecewise polynomial case

Another interesting extension of function estimation is when μ_i of the model (4.1) is a smooth function of time. Tibshirani (2014) shows that the solutions from total variation penalty estimators resemble the structure of a piecewise kth degree polynomial filtering where the discrete derivative operators can be defined in a recursive manner starting with $D^{(1)}$ and then letting

$$D^{(k+1)} = D^{(1)}D^{(k)} \text{ for } k = 1, 2, 3, ...$$

This means that the method can estimate the underlying piecewise polynomial function of any order such as constant ($k = 0$), linear ($k = 1$), quadratic ($k = 2$) etc.

The changes in kth derivative (knots) are selected adaptively based on the data and this simultaneous selection and estimation phenomenon does not occur in regression splines (Hastie and Tibshirani (1990)) or smoothing splines (de Boor (1978), Wahba (1990), Green and Silverman (1994)). The former operate on a fixed set of knots and the user needs to select the number of knots and their placement. The latter place a knot at each data point. Through a generalised ridge regression on the coefficients in a natural spline basis smoothing splines implement smoothness.

For trend filtering the main algorithm still applies and only the locating function (4.8) which now involves the quantity $(D^{(k)}D^{(k)^T})^{-1}D^{(k)}$ needs to be adjusted, but we have not established closed-end formulae (see Figure 4.8 for different forms of $\xi^{1,i,n}$ for $k = 0, 1, 2$). We notice that all other settings will always involve more operations due to the fact that the knots are now allowed to leave the active set. In other words, if a knot is located at a value $\lambda_2^{(1)}$ it is not necessary that it will remain a knot at a value $\lambda_2^{(0)} < \lambda_2^{(1)}$. Therefore, the calculation of the whole path will be more computationally intensive compared with the piecewise constant estimation method.

4.7 Proofs

Before we prove Theorem 4.1 we prove Lemma 4.1. Lemma 4.2 provides a bound for the regularisation parameter λ_2. In addition, and w.l.o.g, we assume that the estimated change-point $\hat{\eta}$ is such that $\hat{\eta} > \eta$.

Proof of Lemma 4.1. Consider the model (4.13). Since the series y_i is blocky, we assume there exists a partition $\{\mathcal{F}_1, \mathcal{F}_2\}$. Following e.g. Hoefling (2010), that is, differentiating (4.3) with respect to μ_i and setting it equal to 0 we get that $\hat{\mu}_i$ is the

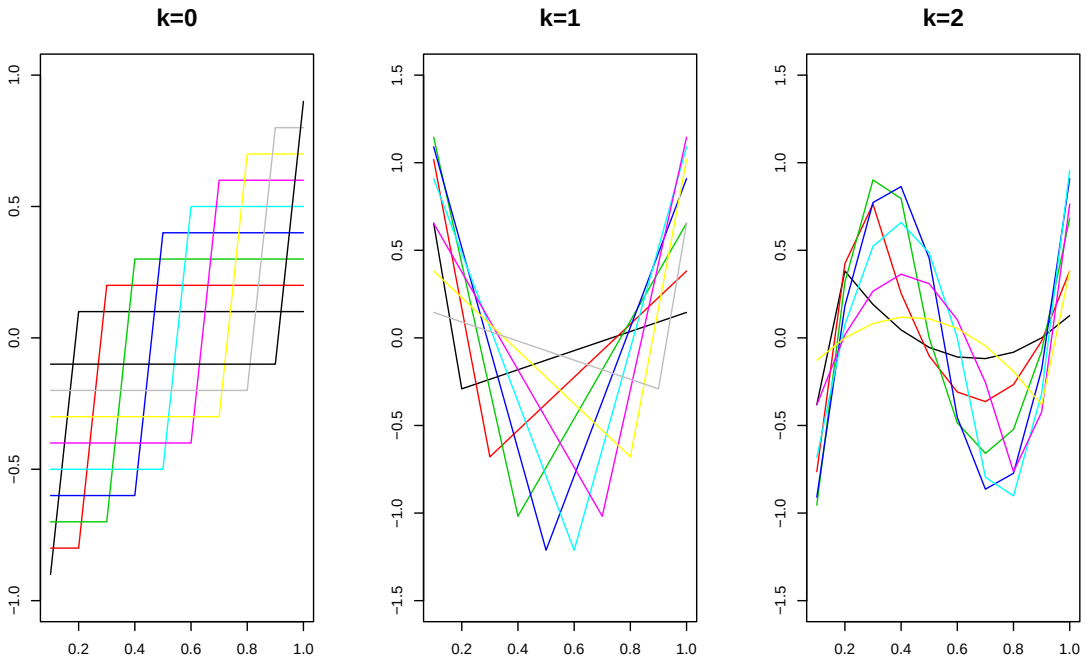

Figure 4.8: From left to right: $\xi^{1,i,n}$ functions of order $k = 0$ (piecewise constant), $k = 1$ (piecewise linear), and $k = 2$ (piecewise quadratic). In all the cases the sample size is $n = 10$. The knots are adaptively chosen based on the data. At the initiation of the SPA method the first knot b is located by $b = \arg\max_{i\in\{k+1,\dots,n\}} |\langle y, \xi^{1+k,i,n}\rangle|$.

unique solution to the subgradient equation

$$\hat{\mu}_i = y_i - \lambda_2 \left(\text{sign}(\hat{\mu}_i - \hat{\mu}_{i-1}) - (\hat{\mu}_{i+1} - \hat{\mu}_i)\right)$$

where the $\text{sign}(x)$ function is defined as

$$\text{sign}(x) = \begin{cases} 1 & \text{if } x > 0 \\ -1 & \text{if } x < 0 \\ [-1, 1] & \text{if } x = 0. \end{cases}$$

Now, summing over every partition $\hat{\mathcal{F}}_1$ and $\hat{\mathcal{F}}_2$ we get (4.14) and (4.15).

□

Lemma 4.2. *Let* $A_n = \{|\sum_{i=1}^n \varepsilon_i| \leq \lambda_2\}$*; then the following holds*

$$\mathbb{P}(A_n) \to 1$$

where $\lambda_2 \geq \sqrt{2n \log n}$ *is the regularisation parameter in (4.2).*

Proof. Since $\sigma = 1$ we have that $\sum_{i=1}^n \varepsilon_i \sim \mathcal{N}(0, n)$ and using a (standard) Gaussian bound

$$\mathbb{P}\left(\left|\sum_{i=1}^n \varepsilon_i\right| > \lambda_2\right) \leq \exp\left(-\frac{\lambda_2^2}{2n}\right)$$

the lemma holds. □

We now proceed with the proof of the main Theorem 4.1 where we use a similar procedure with Lemma A3 of Fryzlewicz (2014). On the set A_n defined in Lemma 4.1 we start with the following

$$\sum_{i=1}^n (y_i - \hat{\mu}_i)^2 + \lambda_2 ||\hat{\mu}||_{TV} \leq \sum_{i=1}^n (y_i - \mu_i)^2 + \lambda_2 ||\mu||_{TV}$$

$$\sum_{i=1}^n (y_i - \hat{\mu}_i)^2 - (y_i - \mu_i)^2 \leq \lambda_2 \sum_{i=2}^n |\mu_i - \mu_{i-1}| - \lambda_2 \sum_{i=2}^n |\hat{\mu}_i - \hat{\mu}_{i-1}|$$

$$\sum_{i=1}^n (\hat{\mu}_i - \mu_i)^2 \leq 2 \underbrace{\sum_{i=1}^n \varepsilon_i (\hat{\mu}_i - \mu_i)}_{\text{I}} + \lambda_2 \left(\underbrace{\sum_{i=2}^n |\mu_i - \mu_{i-1}|}_{\text{II}} - \underbrace{\sum_{i=2}^n |\hat{\mu}_i - \hat{\mu}_{i-1}|}_{\text{III}} \right) \tag{4.20}$$

We decompose the RHS of (4.20) starting with I

$$\sum_{i=1}^n \varepsilon_i (\hat{\mu}_i - \mu_i) = \underbrace{\sum_{i=1}^{\eta} \varepsilon_i (\hat{\mu}_i - \mu_i)}_{\text{I.A}} + \underbrace{\sum_{i=\eta+1}^{\hat{\eta}} \varepsilon_i (\hat{\mu}_i - \mu_i)}_{\text{I.B}} + \underbrace{\sum_{i=\hat{\eta}+1}^{n} \varepsilon_i (\hat{\mu}_i - \mu_i)}_{\text{I.C}}$$

I.A becomes

$$\sum_{i=1}^{\eta} \varepsilon_i \left(\frac{\sum_{u=1}^{\hat{\eta}} y_u}{\hat{\eta}} + \frac{\lambda_2}{\hat{\eta}} - \mu_i \right) = \left(\frac{\sum_{u=1}^{\hat{\eta}} y_u}{\hat{\eta}} \right) \sum_{i=1}^{\eta} \varepsilon_i + \frac{\lambda_2}{\hat{\eta}} \sum_{i=1}^{\eta} \varepsilon_i - \sum_{i=1}^{\eta} \varepsilon_i \mu_i.$$

Using the model (4.13) we have that the first term on the RHS

$$\left(\frac{\eta\mu_0}{\hat{\eta}}+\frac{\sum_{u=1}^{\hat{\eta}}\varepsilon_u}{\hat{\eta}}+\frac{(\hat{\eta}-\eta)\mu_1}{\hat{\eta}}\right)\sum_{i=1}^{\eta}\varepsilon_i+\frac{\lambda_2}{\hat{\eta}}\sum_{i=1}^{\eta}\varepsilon_i-\sum_{i=1}^{\eta}\varepsilon_i\mu_0=\frac{\eta}{\hat{\eta}}\underline{\mu}\sum_{i=1}^{\eta}\varepsilon_i+\frac{\lambda_2}{\hat{\eta}}\sum_{i=1}^{\eta}\varepsilon_i-\underline{\mu}\sum_{i=1}^{\eta}\varepsilon_i.$$

Hence, a bound for I.A is given by

$$\text{I.A}\leq\frac{\lambda_2}{\hat{\eta}}\sum_{i=1}^{\eta}\varepsilon_i+\underline{\mu}\sum_{i=1}^{\eta}\varepsilon_i.$$

Following a similar argument for I.B we have that

$$\begin{aligned}\text{I.B} &= \left(\frac{\eta\mu_0}{\hat{\eta}}+\frac{\sum_{u=1}^{\hat{\eta}}\varepsilon_u}{\hat{\eta}}+\frac{(\hat{\eta}-\eta)\mu_1}{\hat{\eta}}\right)\sum_{i=\eta+1}^{\hat{\eta}}\varepsilon_i+\frac{\lambda_2}{\hat{\eta}}\sum_{i=\eta+1}^{\hat{\eta}}\varepsilon_i-\sum_{i=\eta+1}^{\hat{\eta}}\varepsilon_i\mu_1\\ &= \frac{\eta\underline{\mu}}{\hat{\eta}}\sum_{i=\eta+1}^{\hat{\eta}}\varepsilon_i+\frac{\sum_{u=1}^{\hat{\eta}}\varepsilon_u}{\hat{\eta}}\sum_{i=\eta+1}^{\hat{\eta}}\varepsilon_i+\frac{\lambda_2}{\hat{\eta}}\sum_{i=\eta+1}^{\hat{\eta}}\varepsilon_i\leq\underline{\mu}\sum_{i=\eta+1}^{\hat{\eta}}\varepsilon_i+\frac{\lambda_2}{\hat{\eta}}\sum_{i=\eta+1}^{\hat{\eta}}\varepsilon_i.\end{aligned}$$

Hence, I.A and I.B are of the same order. For I.C we have that

$$\begin{aligned}\text{I.C} &= \left(\mu_1+\frac{\sum_{u=\hat{\eta}+1}^{n}\varepsilon_u}{n-\hat{\eta}}\right)\sum_{i=\hat{\eta}+1}^{n}\varepsilon_i-\left(\frac{\lambda_2}{n-\hat{\eta}}\right)\sum_{i=\hat{\eta}+1}^{n}\varepsilon_i-\sum_{i=\hat{\eta}+1}^{n}\varepsilon_i\mu_1\\ &= \frac{\sum_{u=\hat{\eta}+1}^{n}\varepsilon_u}{n-\hat{\eta}}\sum_{i=\hat{\eta}+1}^{n}\varepsilon_i-\left(\frac{\lambda_2}{n-\hat{\eta}}\right)\sum_{i=\hat{\eta}+1}^{n}\varepsilon_i\leq\frac{\lambda_2^2}{\delta_n}.\end{aligned}$$

For II we have that $\sum_{i=2}^{n}|\mu_i-\mu_{i-1}|=|\underline{\mu}|$ from Assumptions 4.1.

We now examine III. Note that

$$\sum_{i=1}^{n}|\hat{\mu}_i-\hat{\mu}_{i-1}|=|\hat{\mu}_1-\hat{\mu}_0|.$$

Hence, from (4.14) and (4.15), and the model (4.13) we have that

$$\begin{aligned}|\hat{\mu}_1-\hat{\mu}_0| &= \left|\frac{\sum_{i=\hat{\eta}+1}^{n}y_i}{n-\hat{\eta}}-\frac{\lambda_2}{n-\hat{\eta}}-\frac{\sum_{i=1}^{\hat{\eta}}y_i}{\hat{\eta}}-\frac{\lambda_2}{\hat{\eta}}\right|\\ &= \left|\mu_1+\frac{\sum_{i=\hat{\eta}+1}^{n}\varepsilon_i}{n-\hat{\eta}}-\frac{\lambda_2}{n-\hat{\eta}}-\frac{\eta\mu_0}{\hat{\eta}}-\frac{\epsilon_n\mu_1}{\hat{\eta}}-\frac{\sum_{i=1}^{\hat{\eta}}\varepsilon_i}{\hat{\eta}}-\frac{\lambda_2}{\hat{\eta}}\right|\\ &\leq |\underline{\mu}|+\left|\frac{\sum_{i=\hat{\eta}+1}^{n}\varepsilon_i}{n-\hat{\eta}}\right|+\frac{\lambda_2}{n-\hat{\eta}}+\left|\frac{\sum_{i=1}^{\hat{\eta}}\varepsilon_i}{\hat{\eta}}\right|+\frac{\lambda_2}{\hat{\eta}}.\end{aligned}$$

On the set A_n we combine I.A, I.B, I.C, II and III, and we get

$$\sum_{i=1}^{n}(\hat{\mu}_i-\mu_i)^2<C_2\underline{\mu}\lambda_2+\frac{\lambda_2^2}{\delta_n}$$

for a big enough C_2 and by noticing that $\mathcal{O}(\lambda_2)$ dominates $\mathcal{O}(\lambda_2^2/\hat{\eta})$.

We decompose the LHS of (4.20)

$$\sum_{i=1}^{n}(\hat{\mu}_i-\mu_i)^2=\left(\sum_{i=1}^{\eta}+\sum_{i=\eta+1}^{\hat{\eta}}+\sum_{i=\hat{\eta}+1}^{n}\right)(\hat{\mu}_i-\mu_i)^2. \tag{4.21}$$

We start with the first term on the LHS. Using (4.14) and the model (4.13) we have that

$$\begin{aligned}
\sum_{i=1}^{\eta}(\hat{\mu}_i-\mu_i)^2 &= \sum_{i=1}^{\eta}\left(\frac{\sum_{u=1}^{\hat{\eta}}y_u}{\hat{\eta}}+\frac{\lambda_2}{\hat{\eta}}-\mu_i\right)^2\\
&= \sum_{i=1}^{\eta}\left(\frac{\sum_{u=1}^{\hat{\eta}}\mu_u}{\hat{\eta}}+\frac{\sum_{u=1}^{\hat{\eta}}\varepsilon_u}{\hat{\eta}}+\frac{\lambda_2}{\hat{\eta}}-\mu_i\right)^2\\
&= \sum_{i=1}^{\eta}\left(\frac{\eta}{\hat{\eta}}\mu_0+\frac{(\hat{\eta}-\eta)\mu_1}{\hat{\eta}}+\frac{\sum_{u=1}^{\hat{\eta}}\varepsilon_u}{\hat{\eta}}+\frac{\lambda_2}{\hat{\eta}}-\mu_0\right)^2\\
&= \sum_{i=1}^{\eta}\left(\frac{\epsilon_n\underline{\mu}}{\hat{\eta}}+\frac{\sum_{u=1}^{\hat{\eta}}\varepsilon_u}{\hat{\eta}}+\frac{\lambda_2}{\hat{\eta}}\right)^2\geq\frac{\delta_n\epsilon_n^2\underline{\mu}^2}{n^2}+\frac{4\delta_n\lambda_2^2}{n^2}+\frac{4\delta_n\lambda_2\epsilon_n\underline{\mu}}{n^2}.
\end{aligned}$$

We proceed with the second term.

$$\begin{aligned}
\sum_{i=\eta+1}^{\hat{\eta}}(\hat{\mu}_i-\mu_i)^2 &= \sum_{i=\eta+1}^{\hat{\eta}}\left(\frac{\sum_{u=1}^{\hat{\eta}}y_u}{\hat{\eta}}+\frac{\lambda_2}{\hat{\eta}}-\mu_i\right)^2\\
&= \sum_{i=\eta+1}^{\hat{\eta}}\left(\frac{\eta\underline{\mu}}{\hat{\eta}}+\frac{\sum_{u=1}^{\hat{\eta}}\varepsilon_u}{\hat{\eta}}+\frac{\lambda_2}{\hat{\eta}}\right)^2\geq\frac{\delta_n\epsilon_n\underline{\mu}^2}{n^2}+\frac{4\epsilon_n\lambda_2^2}{n^2}+\frac{4\lambda_2\epsilon_n\underline{\mu}}{n^2}.
\end{aligned}$$

Similarly for the third term

$$\sum_{i=\hat{\eta}+1}^{n}\left(\frac{\sum_{u=\hat{\eta}+1}^{n}\varepsilon_u}{n-\hat{\eta}}-\frac{\lambda_2}{n-\hat{\eta}}\right)^2\geq\frac{4\lambda_2^2}{\delta_n}.$$

Combining all the inequalities we get the result.

4.8 Connecting Chapter 3 and Chapter 5

In this chapter, we have provided evidence that the total variation penalty estimator is suboptimal in detecting both the location and the number of the change-points in

the mean of a stochastic process. The simulation studies conducted clearly indicate that the BS method outperforms the FLSA method. In Chapter 3 we showed that the Wild Binary Segmentation (WBS) method - an improved version of BS - does well in detecting the change-points occurred in the autocovariance function of a time series. It is unlikely therefore for the FLSA method to exhibit better performance in other settings, for example in the model (4.1) when the error ε_i is autocorrelated.

Despite that, the fused lasso method and, particularly, the solution path algorithm of Tibshirani and Taylor (2011), which solves many types of lasso problems, provides the user with a flexible tool to estimate a different class of models of the following form

$$y = \beta_1(u) + \beta_2(u)X_2 + ... + \beta_p(u)X_p + \varepsilon \tag{4.22}$$

where $y \in \mathbb{R}^n$ is the response vector, X_j for $j = 1, ..., p$ are the inputs, $\varepsilon \in \mathbb{R}^n$ are iid random errors and the coefficients $\beta_j(u)$ are piecewise constant, linear, quadratic or cubic functions of u. This class of models is considered in Chapter 5.

Given the good performance of the BS or WBS methods, one might choose an appropriate loss function to estimate the piecewise constant varying coefficients $\beta_j(u)$ for $j = 1, ..., p$ and then apply the binary segmentation search. Even though appealing, this kind of estimation will be restricted to piecewise constant models since to the best of our knowledge the BS method has not been applied in models where the coefficients $\beta_j(u)$ are piecewise polynomials.

There are, of course, alternative methods developed for estimating the model (4.22), such as, the smoothing splines (Hastie and Tibshirani (1993)), the kernel estimators (Hoover et al. (1998)), the local polynomial least squares (Fan and Zhang (1999)), or the polynomial splines (Huang et al. (2002)), to name but a few. We do

not presume that the fused lasso method should be regarded as the preferred method over other techniques in the estimation of the model (4.22), but that it represents a useful contribution in making use of an L_1 penalty in the estimation process.

Chapter 5

Adaptive Estimation of Time-Varying Models

Introduction

In a standard linear regression set-up we are interested in modelling the influence of covariates $x^{(1)}, ..., x^{(p)}$ on the response variable y via

$$y_t = \beta_0 + \beta_1 x_t^{(1)} + ... + \beta_p x_t^{(p)} + \varepsilon_t \tag{5.1}$$

where $\varepsilon_t \in \mathbb{R}^n$ are iid random errors. A useful extension of this linear regression is to assume a model where the regression coefficients β_j for $j = 1, ..., p$ vary, for example, over time, different age groups or other covariates (and thus termed *varying coefficient* models; henceforth, VC).

The full potential of VC models was not explored until the seminal works of Cleveland et al. (1991) and Hastie and Tibshirani (1993). VC models are used in a range of applications, including longitudinal studies aiming at investigating how covariates affect responses through time (Hoover et al. (1998), Fan and Zhang (1999),

Fan and Zhang (2000), Eubank et al. (2004) among others). For example, in Eubank et al. (2004) the authors examine a data set which concerns patients with multiple sclerosis who had been admitted to nursing homes. The response variable y in this study is a performance index measuring the activities of daily living which has been measured along with other variables such as race, ethnicity, body mass index or gender. The authors allow the covariates to vary with time (in addition to the coefficients) which in this case is taken to be the age of a patient, and they develop a method to estimate this model. Other applications involve financial data, such as the work by Criton and Scaillet (2011) who examine the time-varying alpha, a measure of financial performance, in order to show that market exposures differ between two crises. Especially in the time series context (note that if $x_t^{(1)} = y_{t-1}, ..., x_t^{(p)} = y_{t-p}$ the model (5.1) is an AR(p) model) there is a substantial interest for models where coefficients evolve with respect to some variable; see, for example, Robinson (1989), Chen and Tsay (1993) or Cai et al. (2000). Bai and Perron (1998) and Qu and Perron (2007) assume that the coefficients in the model (5.1) are piecewise constant which is of interest in many studies where relationships between economic indices are likely to contain structural breaks (e.g. Stock and Watson (1996), McConnell and Perez-Quiros (2000)).

In this chapter, we are interested in fitting a linear model in several predictor variables. We believe that each coefficient is varying with respect to some underlying parameter t. We consider the following *time-varying model*

$$y_t = \sum_{j=1}^{p} \beta_t^{(j)} x_t^{(j)} + \varepsilon_t, \text{ for } j = 1, ..., p \text{ and } t = 1, ..., n \tag{5.2}$$

where $\beta_t^{(j)}$ are p piecewise polynomial functions of time and $\varepsilon_t \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. The

model (5.2) covers the case of a varying intercept by setting $x_t^{(1)} = 1 \; \forall t = 1, ..., n$.

We assume that the coefficients vary with respect to a single index t which, for simplicity, is taken to be time, but other indices can be also used as in the applications we consider in Section 5.8. Further, the underlying parameter t is univariate and the same for all the covariates. Hence, (5.2) is a special case of a VC model where $\beta^{(j)}$ is not necessarily a function of time $g(t)$, but also $g(R_j)$ where R_j can be taken to be x_j or a linear combination of regressors and/or other variables (Fan et al. (2003)). Finally, the time-varying model (5.2) differs from the generalised additive models (GAM) of Hastie and Tibshirani (1990) in that GAMs assume a linear regression model where some or all of the regressors are smooth non-parametric functions. A special case of a GAM is non-parametric regression which we discuss in Section 5.1.

Fan and Zhang (2008) categorise the VC estimation into three approaches. One possible way to conduct variable smoothing is by using the smoothing spline approach of Hastie and Tibshirani (1993). The varying coefficients can be estimated using the backfitting algorithm, an iterative "one at a time" method, typically adopted in the estimation of GAMs. Smoothing splines have also been studied by Hoover et al. (1998) and Chiang et al. (2001) in the context of longitudinal studies. Another approach of estimating the model (5.2) is by adopting polynomial splines first proposed by Huang et al. (2002). Polynomial splines are favoured over other basis systems, such as the Fourier basis, since they perform well in approximating the local features of a function and provide stable numerical solutions (de Boor (1978)). Finally, a branch of estimators make use of a kernel-local polynomial smoothing, see Cleveland et al. (1991), Wu et al. (1998), Hoover et al. (1998), Kauermann and Tutz (1999) and Cai et al. (2000). But a major drawback of ordinary least squares kernels and

local polynomial estimators is that they rely on a single bandwidth and, hence, they assume that all functions posses the same degree of smoothness. To overcome this limitation, Fan and Zhang (1999) propose a two-step-estimation where the obtained initial estimates are input into a second local least-squares regression. By doing so it is expected that the estimation is not sensitive to the bandwidth of the first step.

We propose an estimation method based on the Fused Lasso method of Tibshirani et al. (2005). Due to the adoption of an L_1 penalty in the estimation process, our method falls in the penalised regression category similarly with the smoothing splines which utilise L_2 penalties. We adopt the solution path algorithm of Tibshirani and Taylor (2011) (henceforth, T&T), a method used to solve lasso-type problems such as the non-parametric regression. In a non-parametric regression setting Tibshirani (2014) shows that estimators with total-variation penalties perform better than smoothing splines in terms of minimax convergence rates and empirical evidence. Since we adopt this new class of estimators to estimate $\beta_t^{(j)}$ for $j = 1, ..., p$ in the model (5.2) we expect our method (termed Fused Lasso estimator for Time Varying models - FuLTV) to do better than smoothing splines in time-varying estimation. Indeed, in the simulated examples and real data sets that we consider FuLTV performs well in most cases. Finally, a notable result of our method is that it permits an exact calculation of the degrees of freedom and, hence, a more efficient way of model selection.

This chapter contributes in the time-varying model estimation literature in four ways. First, it proposes a new class of estimators for the time-varying coefficients $\beta_t^{(j)}$ in the model (5.2) making use of L_1 penalties in the estimation process. Second, it suggests a path algorithm for this lasso-type problem. Path algorithms provide an

exact solution in contrast with general purpose convex optimisation techniques. In addition, they offer an interpretation advantage where the user is able to examine the solution path for a decreasing regularisation parameter. Third, it shows how FuLTV estimates models where the underlying coefficient structure is not only piecewise constant, but also piecewise polynomial of degree k and, hence, piecewise linear, quadratic, cubic etc. Finally, the adoption of the fused lasso method enables a comparison with penalised least squares method first proposed by Hastie and Tibshirani (1993) who adopt L_2 penalties for estimating the model (4.22). From that perspective, this chapter also serves as a comparative study between L_1 - and L_2 -type of penalised regression.

This chapter is organised as follows. After introducing the solution path algorithm of T&T (Section 5.1), in Section 5.2 we examine some computational aspects of the univariate time-varying model. Then we extend T&T's path algorithm in a multi-covariate setting in Section 5.3. This is followed by a comparison between FuLTV and the smoothing splines of Hastie and Tibshirani (1993) and a sketch of the theoretical consistency of FuLTV for the piecewise constant model in Sections 5.4 and 5.5, respectively. In Section 5.6 we discuss model selection for the FuLTV by looking at the degrees of freedom, a measure of the complexity of a model. After performing a simulation study to assess the performance of our method in Section 5.7, the penultimate part (Section 5.8) consists of an account of two applications to real data sets. Finally, Section 5.9 contains proofs of the lemmas related to the FuLTV method.

5.1 Preliminaries

The method we propose in this chapter for estimating the model (5.2) is a natural extension of the non-parametric regression, i.e. the observations $y_1, ..., y_n \in \mathbb{R}$ are generated from the following model

$$y_i = f_0(x_i) + \varepsilon_i, \quad i = 1, ..., n \tag{5.3}$$

where x_i are input points, f_0 is the underlying function (signal) to be estimated and $\varepsilon_1, ..., \varepsilon_n$ are independent errors. It is important to notice that the model (5.3) can contain multiple functions, and not just $f_0(x_i)$. Therefore, the model (5.3) is a univariate GAM. Finally, the model (5.2) differs from (5.3) (apart from the multiple predictors involved) in that y_i is related to x_i through a coefficient that varies with respect to another variable.

Many methods for estimating f_0 have been proposed such as local polynomials, splines or wavelets. A special case of the model (5.3) is when f_0 exhibits a piecewise constant behaviour which can be also described as abrupt changes, termed change-points or break-points, in the mean of a series. Algorithms for estimating the piecewise constant signal in the context of the Fused Lasso approach, the main topic of interest in this chapter, can be found in Friedman et al. (2007) (Fused Lasso Signal Approximator - FLSA) and Hoefling (2010). However, an algorithm has already been devised by Davies and Kovac (2001), but with a different name, i.e. the taut string. Kim et al. (2009a) introduce the trend filtering method to estimate the underlying function of (5.3) when f_0 is a piecewise linear function. Tibshirani (2014) proposes a kth order trend filtering which estimates the structure of a piecewise polynomial of any order i.e. piecewise quadratic, cubic etc and not only piecewise constant or linear.

In addition, it is worth mentioning that trend filtering is practically very similar to the locally adaptive regression splines of Mammen and van de Geer (1997), a total variation type of method for estimating the model (5.3).

The algorithm of T&T is designed to solve the following lasso problem

$$\beta \in \arg\min_{\beta \in \mathbb{R}^p} ||y - X\beta||_2^2 + \lambda ||D\beta||_1 \tag{5.4}$$

where $y \in \mathbb{R}^n$ is an outcome vector, $X \in \mathbb{R}^{n \times p}$ is a predictor matrix and $D \in \mathbb{R}^{m \times p}$ a penalty matrix. A special case of the problem is when $D = I$ which reduces to the standard lasso problem (Tibshirani (1996)). For the model (5.3) $X = I$ and $\beta_i = f_0(x_i)$. If the underlying signal β_i is piecewise constant we use the penalty matrix

$$D = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix} \in \mathbb{R}^{(n-1) \times n}. \tag{5.5}$$

In Chapter 4 we presented the algorithm in the case where $X = I$ and D has the form of (5.5). Of utmost importance is to note a difference in algorithmic terms between this case and the case where D is an arbitrary penalty matrix. When the penalty matrix D is as in (5.5), i.e. the piecewise constant estimator, the dual coordinates (knots) always remain in the active set $\mathcal{B}$ until the termination of the algorithm. From the "primal" perspective this means that

$$\hat{\beta}_{\lambda_0,i} = \hat{\beta}_{\lambda_0,i+1} \Rightarrow \hat{\beta}_{\lambda,i} = \hat{\beta}_{\lambda,i+1} \; \forall \; \lambda \geq \lambda_0.$$

However, for an arbitrary penalty matrix D the dual coordinates included in the active set $\mathcal{B}$ (hitting coordinates) can also leave $\mathcal{B}$ for decreasing regularisation parameter and this will happen frequently. Therefore, at the qth iteration, apart from

testing when a dual coordinate will join the active set $\mathcal{B}$, i.e.

$$i_{q+1} = \arg\max_i h_i$$

where

$$h_i = \frac{[D_{-\mathcal{B}}(D_{-\mathcal{B}})^T)^+ D_{-\mathcal{B}} y]_i}{[D_{-\mathcal{B}}(D_{-\mathcal{B}})^T)^+ D_{-\mathcal{B}}(D_{\mathcal{B}})^T \mathcal{S}]_i \pm 1}$$

(either -1 or 1 in (4.8) will yield a value in $[0, \lambda_q]$) it is essential to know when it will leave $\mathcal{B}$ (leaving coordinates), i.e.

$$i_{q+1} = \arg\max_i l_i$$

where

$$l_i = \begin{cases} \gamma_i/\delta_i & \text{if } \gamma_i, \delta_i < 0 \\ 0 & \text{otherwise} \end{cases}$$

and

$$\gamma_i = \mathcal{S}_i \left[D_{\mathcal{B}}[I - (D_{-\mathcal{B}})^T (D_{-\mathcal{B}}(D_{-\mathcal{B}})^T)^+ D_{-\mathcal{B}}] y \right]_i$$

$$\delta_i = \mathcal{S}_i \left[D_{\mathcal{B}}[I - (D_{-\mathcal{B}})^T (D_{-\mathcal{B}}(D_{-\mathcal{B}})^T)^+ D_{-\mathcal{B}}] (D_{-\mathcal{B}})^T \mathcal{S} \right]_i$$

where A^+ denotes the Moore-Penrose pseudoinverse of A.

Below, we summarise the solution path algorithm of T&T.

Algorithm 1 (Dual Path Algorithm for $X = I$ and a general matrix D)

Given $y \in \mathbb{R}^n$ and $D \in \mathbb{R}^{m \times n}$

1. Find $\hat{u}$ by minimising the L_2 norm solution of

$$\min_{u \in \mathbb{R}^m} ||y - D^T u||_2^2.$$

2. Find the first hitting time λ_1 and the hitting coordinate i_1; record the solution $\hat{u}(\lambda) = \hat{u}$ for $\lambda \in [\lambda_1, \infty])$. Initialize $\mathcal{B} = \{i_1\}$, $\mathcal{S} = \text{sign}(\hat{u}_{i_1})$ and $q = 1$.

3. While $\lambda_q > 0$:

   (a) Find $\hat{a}$ and $\hat{b}$ by minimising the L_2 norm solution of

   $$\min_{a \in \mathbb{R}^{m-|\mathcal{B}|}} ||y - D_{-\mathcal{B}}^T a||_2^2 \quad \text{and} \quad \min_{b \in \mathbb{R}^{m-|\mathcal{B}|}} ||D_{\mathcal{B}}^T \mathcal{S} - D_{-\mathcal{B}}^T b||_2^2.$$

   (b) Compute the next hitting time $\Pi_{q+1} = \max_i h_i$ and leaving time $\Gamma_{q+1} = \max_i l_i$.

   Set $\lambda_{q+1} = \max\{\Pi_{q+1}, \Gamma_{q+1}\}$.

   If $\Pi_{q+1} > \Gamma_{q+1}$ add the hitting coordinate to $\mathcal{B}$ and its sign $\mathcal{S}$; otherwise, remove the leaving coordinate from $\mathcal{B}$ and $\mathcal{S}$.

   Record the solution $\hat{u}(\lambda) = \hat{a} - \lambda\hat{b}$ for $\lambda \in [\lambda_{q+1}, \lambda_q])$ and update $q = q + 1$.

Algorithm 1 can be extended to other settings where $X \neq I$ as is the case with the model (5.2) we consider here. This is achieved by applying a transformation to the response vector y and the penalty matrix D. Specifically, let $\tilde{y} = XX^+y$ and $\tilde{D} = DX^+$. Then, to find the solution path of the transformed problem, we can apply Algorithm 1 on $\tilde{y}$ and $\tilde{D}$. However, this transformation does not change the properties of the estimator: the values of the coefficients can be obtained from the dual coordinates by

$$\hat{\beta} = \tilde{y} - \tilde{D}^T\hat{u}.$$

5.2 The univariate time-varying model

5.2.1 Computational aspects

As a preparatory exercise and before we consider the estimation of the model (5.2) in Section 5.3 we examine the simpler time-varying model with a single covariate ($p = 1$)

$$y_t = \beta_t x_t + \varepsilon_t \tag{5.6}$$

where $t = 1, ..., n$ and ε_t are independent errors. Note the similarities of this model with the model (5.3) when taking $x_t = 1$ for $\forall t \in \{1, ..., n\}$.

Our aim is to estimate β_t which, for example, can be sparse and blocky. We form the loss function

$$f(\beta_t) = \sum_{t=1}^{n}(y_t - \beta_t x_t)^2 + \lambda_1 \sum_{t=1}^{n} |\beta_t| + \lambda_2 \sum_{t=2}^{n} |\beta_t - \beta_{t-1}| + \lambda_3 \sum_{t=1}^{n} \beta_t^2. \tag{5.7}$$

where $\lambda_3 > 0$. The reasons for adding the extra (ridge) penalty $\lambda_3 \sum_{t=1}^{n} \beta_t^2$ are explained later in this section. The fused lasso penalty term in the piecewise constant time-varying model set-up takes the following form

$$\|D\beta\|_1 = \sum_{t=1}^{n} |\beta_t - \beta_{t-1}|.$$

This type of penalty encourages sparsity in the differences of the coefficients and hence some of the terms $|\beta_t - \beta_{t-1}|$ will be zero. We define the matrix

$$X = \begin{pmatrix} x_1 & 0 & \cdots & 0 \\ 0 & x_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & x_n \end{pmatrix} \quad \in \mathbb{R}^{n \times n} \tag{5.8}$$

which is the identity matrix (as in the FLSA) multiplied by the vector x_t. In order to proceed with the estimation method of the above model we present the following lemma which is a modification of Lemma A.1 of Friedman et al. (2007) [pg 326]:

Lemma 5.1. *When the solution for* $\lambda_1 = 0$ *and* $\lambda_2 \geq 0$ *denoted by* $\hat{\beta}(0, \lambda_2)$ *is known then for a fixed* λ_3 *the solution for* $\lambda_1 > 0$ *is*

$$\hat{\beta}_t(\lambda_1, \lambda_2) = sign(\hat{\beta}_t(0, \lambda_2))(|\hat{\beta}_t(0, \lambda_2)| - \frac{\lambda_1}{\underline{x}_t^2})^+$$

where $\underline{x}_t^2 = x_t^2 + \lambda_3$.

Proof: See Section 5.9.

Simply, due to the special structure of the design matrix X, the estimated coefficients for the lasso penalty can be obtained by soft-thresholding. Hence, we do not need to solve the problem over a grid of values of the pair (λ_1, λ_2) but only over a grid of values of λ_2 and then use Lemma 5.1 to find the solution for different values of λ_1. The division by $\underline{x}_t^2$ is permitted thanks to the ridge penalty in (5.7). Since we do not consider the lasso penalty in time-varying estimation, henceforth, we will take $\lambda_1 = 0$ and for notational simplicity, $\lambda = \lambda_2$.

We now show how the fused lasso estimator returns a blocky solution by fusing neighbouring coefficients for increasing λ. Let us consider the loss function for the model (5.6)

$$f(\beta) = \sum_{t=1}^{n}(y_t - \beta_t x_t)^2 + \lambda \sum_{t=2}^{n} |\beta_t - \beta_{t-1}| + \lambda_3 \sum_{t=1}^{n} \beta_t^2$$

where we fix all $\beta_k = \hat{\beta}_k$, $k \neq t$ at their global minimising values (those values that minimise the loss function) and we only consider β_t. The loss function is not differentiable with respect to β_t at $\{\hat{\beta}_{t-1}, \hat{\beta}_{t+1}\}$ (left panel of Figure 5.1) and standard rules of subdifferential calculus are adopted (Bertsekas (1999)) i.e. the subdifferential

of $f(\beta_t)$ with respect to β_t is

$$\partial f(\beta) = \begin{cases} -(y_t - \beta_t x_t)x_t + 2\lambda_3\beta_t + \lambda\left(\text{sign}(\hat{\beta}_t - \hat{\beta}_{t-1}) - \text{sign}(\hat{\beta}_{t+1} - \beta_t)\right) & \text{if } \hat{\beta}_t \notin \{\hat{\beta}_{t-1}, \hat{\beta}_{t+1}\} \\ [-(y_t - \hat{\beta}_{t-1}x_t)x_t + 2\lambda_3\hat{\beta}_{t-1} - 2\lambda, -(y_t - \hat{\beta}_{t-1}x_t)x_t + 2\lambda_3\hat{\beta}_{t-1}] & \text{if } \hat{\beta}_t = \hat{\beta}_{t-1} \\ [-(y_t - \hat{\beta}_{t+1}x_t)x_t + 2\lambda_3\hat{\beta}_{t+1}, -(y_t - \hat{\beta}_{t+1}x_t)x_t + 2\lambda + 2\lambda_3\hat{\beta}_{t+1}] & \text{if } \hat{\beta}_t = \hat{\beta}_{t+1} \end{cases}$$

The above expression is a piecewise linear function of $\hat{\beta}_t$ with one solution, if that exists. The breaks occur at points $\hat{\beta}_{t-1}$ and $\hat{\beta}_{t+1}$ and we check whether $0 \in \partial f(\beta)$ by inspecting each of the three intervals that are created (see also the right panel of Figure 5.1). In the illustration of Figure 5.1 this occurs when $\hat{\beta}_t = \hat{\beta}_{t-1}$, i.e. the two coefficients are now fused.

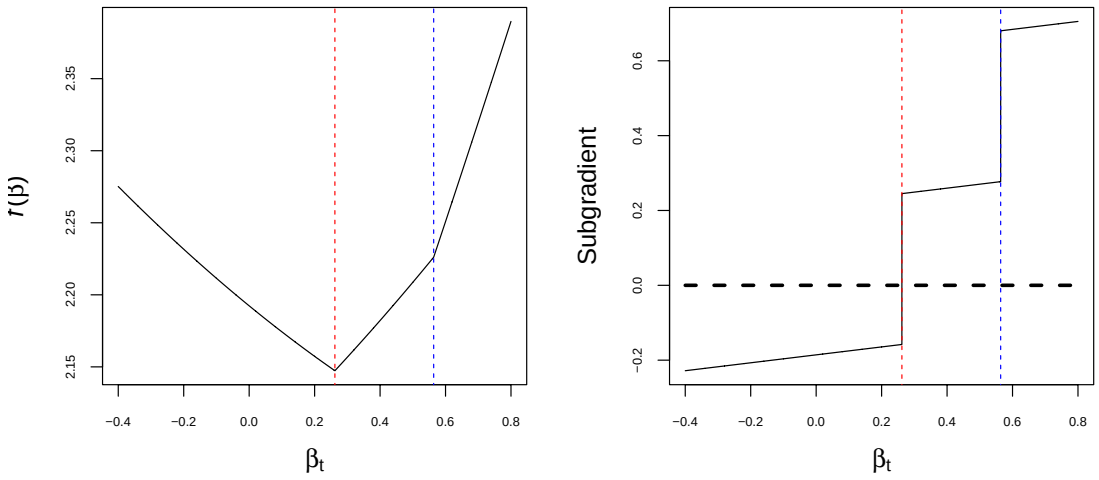

Figure 5.1: Loss function $f(\beta)$ with respect to β_t with the rest of the parameters set at their global minimising values (left). The subgradient $\partial f(\beta)$ of β_t with discontinuities at $\hat{\beta}_{t-1}$ and $\hat{\beta}_{t+1}$ (right). The blue line is the break at $\hat{\beta}_{t+1}$ while the red is at the point where β_t takes its optimal value i.e. equal to $\hat{\beta}_{t-1}$.

In Section 5.2.2 we devise a solution path algorithm to estimate univariate time-varying models. Since the ridge penalty provides numerical stability (the X matrix is

singular if some of the values of x_t are zero or close to zero) we choose to keep it by making a variable transformation. To achieve this we re-write (5.7) in the following form

$$f(\beta_t) = \sum_{t=1}^{n}(y_t - \beta_t x_t)^2 + \lambda \sum_{t=2}^{n} |\beta_t - \beta_{t-1}| + \sum_{t=1}^{n}(0 - \sqrt{\lambda_3}\beta_t)^2. \tag{5.9}$$

We can now define vector $\check{y} = [y, \mathbf{0}]^T$ and $\check{X} = \left[\begin{smallmatrix} X \\ \sqrt{\lambda_3} I \end{smallmatrix}\right]$ where I is a $n \times n$ diagonal matrix and X as in (5.8). This transformation allows us to work with $\check{y}$ and $\check{X}$ instead of y, X. The addition of a ridge penalty into the loss function is required only for the purpose of applying the solution path algorithm and not for the consistency result in Section 5.5.

5.2.2 A solution path algorithm for the univariate time-varying model

Using the transformed variables $\check{y}$ and $\check{X}$ we proceed with the estimation of a univariate model with piecewise constant (time-varying) coefficients by using the penalty matrix as in (5.5).

In matrix notation the optimisation problem (5.7) has the following form

$$\min_{\beta \in \mathbb{R}^n, z \in \mathbb{R}^m} \frac{1}{2}||\check{y} - \check{X}\beta||_2^2 + \lambda||z||_1 \ \text{ s.t. } \ D\beta = z. \tag{5.10}$$

In the new optimisation problem (5.10) the ridge penalty is taken into consideration through the transformed variables $\check{y}$ and $\check{X}$. We now follow the same argument as in T&T (see also Section 5.1), but for the time-varying set-up.

We rewrite the problem (5.10) into its Lagrangian form

$$\mathcal{L}(\beta, z, u) = \frac{1}{2}||\check{y} - \check{X}\beta||_2^2 + \lambda||z||_1 + u^T D\beta - u^T z$$

and minimise it over β and z. Starting with β,

$$\left(\frac{1}{2}(\check{y}-\check{X}\hat{\beta})^T(\check{y}-\check{X}\hat{\beta})+u^TD\hat{\beta}\right)'=0$$

or

$$-\check{X}^T\check{y}+\check{X}^T\check{X}\hat{\beta}+D^Tu=0$$

or

$$\hat{\beta}=(\check{X}^T\check{X})^{-1}(\check{X}^T\check{y}-D^Tu).$$

Define $\mathcal{Q}=(\check{X}^T\check{X})^{-1}$. We now invoke an argument by T&T to show that the dual problem has the following form

$$\min_{u\in\mathbb{R}^{n-1}}\frac{1}{2}(\check{X}^T\check{y}-D^Tu)^T\mathcal{Q}(\check{X}^T\check{y}-D^Tu)\ \text{ s.t. }\ ||u||_\infty\leq\lambda. \tag{5.11}$$

We transform $\check{y}$ and D according to T&T and more specifically

$$\tilde{y}=\check{X}\mathcal{Q}\check{X}^T\check{y} \tag{5.12}$$

$$\tilde{D}=D\mathcal{Q}^T\check{X}^T \tag{5.13}$$

In order to obtain the time-varying coefficients from the dual variables $\hat{u}_\lambda$ we use

$$\hat{\beta}_\lambda=\mathcal{Q}\check{X}^T(\check{y}-\tilde{D}^T\hat{u}_\lambda).$$

The optimisation problem has the following form

$$\min_{u\in R^{n-1}}\frac{1}{2}||\tilde{y}-\tilde{D}^Tu||_2^2\ \text{ s.t. }\ ||u||_\infty\leq\lambda. \tag{5.14}$$

Note that $(\check{X}^T\check{X})^{-1}$ is a $n\times n$ diagonal matrix with entries $\frac{x_t^2}{x_t^2+\lambda_3}$ for $t=1,...,n$ and λ_3 as in (5.7). Then, we can take $\tilde{y}\approx\check{y}$ and, therefore no transformation is required. If, in addition, $\lambda_3\to 0$ and matrix X is invertible then $\tilde{y}:=y$ and the similarity of problems (4.5) and (5.14) becomes apparent.

We now turn to the reasons for adding an L_2 penalty into the loss function. Assume that $\lambda_3=0$, then $\mathcal{Q}=(X^TX)^{-1}$ which is not always invertible (to see that

take some elements of diag(X) to be zero). By adding the extra penalty we permit the calculation of $\mathcal{Q}$ even if X^TX is singular. Recall from the ridge regression that the L_2 penalty permits estimation of the regression coefficients when X^TX is not of full rank which is the case when the sample size n is smaller than the number of predictors. The ridge penalty performs a similar task in the univariate case.

Instead of adding a ridge penalty in (5.7) we can follow a different approach. First, consider the model

$$y_t x_t = \beta_t x_t^2 + \varepsilon_t x_t. \tag{5.15}$$

We define M non-overlapping partitions of the model each of size m such that the response matrix $\tilde{Y} \in \mathbb{R}^{M\times 1}$ has the form

$$\tilde{Y} = \left[\sum_{t=1}^{m} y_t x_t, \sum_{t=m+1}^{2m} y_t x_t, ..., \sum_{t=(M-1)m+1}^{Mm} y_t x_t\right]^T$$

and $\tilde{X} \in \mathbb{R}^{M\times M}$ such that

$$\text{diag}(\tilde{X}) = \left[\sum_{t=2}^{m} x_t^2, \sum_{t=m+1}^{2m} x_t^2, ..., \sum_{t=(M-1)m+1}^{Mm} x_t^2\right].$$

We can repeat the primal-dual transformation to obtain the optimisation problem (5.14) by considering the following model

$$\underline{\mathrm{y}}_{t^\star} = \beta_{t^\star} \underline{\mathrm{x}}_{t^\star} + u_{t^\star} \tag{5.16}$$

where $t^\star = 1, ..., M-1$, $\underline{\mathrm{y}}_{t^\star} = \sum_{t=t^\star m}^{(t^\star+1)m} y_t x_t$ and $\underline{\mathrm{x}}_{t^\star} = \sum_{t=t^\star m}^{(t^\star+1)m} x_t^2$ for a partition of size m.

However, grouping the data as shown has three main disadvantages. Firstly, it reduces the effective sample size which becomes even less desirable as we increase the size of the small segments. In applications with large samples this might not be a major drawback, but in small samples this method can significantly reduce

the performance of the estimator. Secondly, the fact that we define non-overlapping segments, it implies that it is possible to miss change-points that occur inside the segments. Finally, a tuning process is needed to select size m. We do not make further use of this transformation apart from proving Lemma 5.2 below.

In order to derive a solution path algorithm for the model (5.6) we examine whether the boundary lemma holds (Lemma 5.2). The boundary lemma in the FLSA is the equivalent of Proposition 2 of Friedman et al. (2007) which states that two parameters that are fused in the solution for (λ_1, λ_2) will be fused for all $(\lambda_1, \lambda_2' > \lambda_2)$. T&T notice that the lemma holds when DD^T is diagonally dominant, that is

$$(DD^T)_{ii} \geq \sum_{j \neq i} \left|(DD^T)_{ij}\right| \text{ for } i = 1, ..., n-1.$$

In Section 5.9 (Lemma 5.3) we show that $\tilde{D}\tilde{D}^T$ is also diagonally dominant and, hence, the following lemma holds:

Lemma 5.2. *For a univariate time-varying model we have that for any coordinate i, the solution $\hat{u}_\lambda$ of (5.14) satisfies*

$$\hat{u}_{\lambda_0,i} = \lambda_0 \Rightarrow \hat{u}_{\lambda,i} = \lambda \;\; for \;\; all \;\; \lambda \;\; \in [0, \lambda_0]$$

and

$$\hat{u}_{\lambda_0,i} = -\lambda_0 \Rightarrow \hat{u}_{\lambda,i} = -\lambda \;\; for \;\; all \;\; \lambda \;\; \in [0, \lambda_0].$$

Proof: See Section 5.9.

Simply, the lemma states that for decreasing λ the coordinate u_i stays within the boundary i.e. $u_i = \lambda$. Thus at every iteration we only need to find the interior coordinates.

The solution path algorithm for a univariate time varying model with piecewise-constant coefficients is a direct modification of Algorithm 1. In other words, we only

need to solve the following linear system

$$\tilde{D}\tilde{D}^T u = \tilde{D}\check{y}. \tag{5.17}$$

This corresponds to an amendment in steps 1 and 3a of Algorithm 1.

5.2.3 Beyond piecewise-constant structure

In this section we argue that FuLTV can estimate a univariate time-varying where the underlying varying coefficient structure is not necessarily piecewise-constant. In the non-parametric regression set-up Tibshirani (2014) suggests that the solutions from total variation penalty estimators resemble the structure of a piecewise kth degree polynomial filtering where the discrete derivative operators can be defined in a recursive manner starting with $D^{(1)}$ and then letting

$$D^{(k+1)} = D^{(1)}D^{(k)} \text{ for } k = 1, 2, 3, ... \tag{5.18}$$

We also refer the reader to the work by the same author for theoretical support of estimation properties of trend filtering and its comparison (in terms of minimax convergence rates) with smoothing splines (see e.g. de Boor (1978), Wahba (1990), Green and Silverman (1994)) and locally adaptive regression splines by Mammen and van de Geer (1997).

The penalty matrix D in (5.5) is, hence, $D^{(k+1)}$ for $k = 0$. Another type of penalty is the L_1 trend filtering ($k = 1$) of Kim et al. (2009a) which penalises variations in the trend, i.e.

$$\left\|D^{(2)}\beta\right\|_1 = \sum_{t=1}^{n} |\beta_{t-1} - 2\beta_t + \beta_{t+1}|.$$

An example of a simulated time-varying univariate model and its estimated varying coefficients β_t is shown in Figure 5.2. The coefficients are assumed to admit a

piecewise-linear function and, hence, a $D^{(2)}$ penalty is used. This type of penalty also appears in Mammen and van de Geer (1997) and from that perspective trend filtering is the same with locally adaptive regression splines when $k = 0$ or $k = 1$.

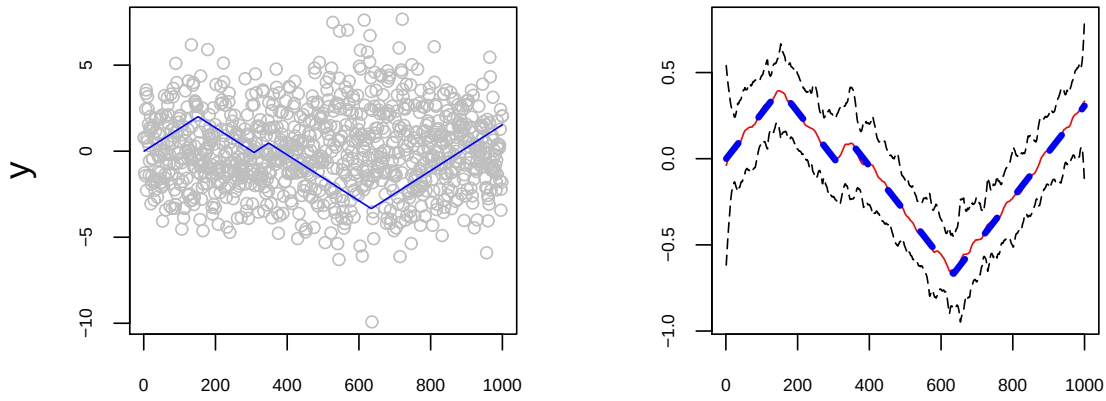

Figure 5.2: Left panel is an instance of 100 *simulated TV models of the form* $y_t = \beta_t x_t + \varepsilon_t$ *where* $x_t \sim \mathcal{N}(1,1)$, $\varepsilon_t \sim \mathcal{N}(0,2)$ *and* β_t *follows a piecewise linear function (in blue and multiplied by* 5 *for scale reasons). Right panel shows the estimated coefficients averaged over* 100 *simulations denoted by the (red) solid line while the standard deviations (multiplied by* 2 *for scale reasons) are denoted by the two (black) symmetric, dashed lines. The underlying, true function* β_t *is denoted by the (blue) dashed line.*

5.3 Multi-covariate time-varying model estimation

We extend FuLTV in the multi-covariate setting and we estimate the time-varying coefficients in the model (5.2). We form the following loss function

$$f(\beta) = \sum_{t=1}^{n} \left(y_t - \sum_{j=1}^{p} \beta_t^{(j)} x_t^{(j)} \right)^2 + \lambda \sum_{j=1}^{p} ||D\beta_t^{(j)}||_1 \tag{5.19}$$

or (in matrix notation)

$$f(\beta) = ||\mathbf{y} - \mathbf{X}\beta||_2^2 + \lambda \|\mathbf{D}\beta\|_1 \tag{5.20}$$

where $\mathbf{y} \in \mathbb{R}^{n\times 1}$ is the response vector; $\mathbf{X} \in \mathbb{R}^{n\times p^\star}$ is the design block matrix where its partition contains p diagonal matrices of size $n \times n$ and

$$p^\star = np. \tag{5.21}$$

In addition, $\beta \in \mathbb{R}^{p^\star \times 1}$ is the coefficient matrix and $\mathbf{D} \in \mathbb{R}^{p(n-1)\times p(n-1)}$ is the penalty matrix the form of which is described shortly after.

In Section 5.2, we presented the univariate piecewise constant model ($p = 1$) and we showed that it shares many features with the FLSA method. In the multi-covariate case we stack the p matrices $X^{(j)} = \text{diag}(x_t^{(j)}) \in \mathbb{R}^{n\times n}$ for $j = 1, ..., p$ into one single design matrix, i.e.

$$\mathbf{X} = \left(X^{(1)}| \quad X^{(2)}| \quad \dots | \quad X^{(p)} \right).$$

The penalty matrix $\mathbf{D}$ has the following form

$$\mathbf{D} = \begin{pmatrix} D_1^{(k_1+1)} & 0 & \cdots & 0 \\ 0 & D_2^{(k_2+1)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & D_p^{(k_p+1)} \end{pmatrix}$$

where $D_j^{(k_j+1)}$ is the penalty matrix (5.18) with discrete difference operator of order $k_j + 1$ where $k_j \geq 0$ for $j = 1, ..., p$. An interesting feature of the penalty matrix $\mathbf{D}$ is that it allows the use of different orders of piecewise polynomial functions across the covariates, even though practically this means that the user has a priori knowledge of the underlying structure of each of the varying coefficients.

We now apply the solution path algorithm for the model (5.2). Firstly, we note that in the multi-covariate time-varying case the predictor matrix X does not have

full rank (recall that $m = np$). One way to deal with that is to add a ridge penalty to the original problem (5.20)

$$f(\beta) = ||y - \mathbf{X}\beta||_2^2 + \lambda||\mathbf{D}\beta||_1 + \lambda_3||\beta||_2^2.$$

This is analogous to the elastic net of Zou and Hastie (2005) which adds a second constraint to the lasso problem. The above can be re-written as follows

$$f(\beta) = ||y^* - \mathbf{X}^*\beta||_2^2 + \lambda||\mathbf{D}\beta||_1 \tag{5.22}$$

where $y^* = [y, 0]^T$, $\mathbf{X}^* = \left[\frac{\mathbf{X}}{\sqrt{\lambda_3} I}\right]$ and I is a $p^\star \times p^\star$ diagonal matrix. The extra ridge penalty provides more computational stability especially when the diagonals of the $X^{(j)}$ contain values close or equal to 0.

For the multi-covariate $\mathbf{X}^*$, we can apply the same argument with the univariate setting as in Section 5.2.2 to derive the dual of (5.22)

$$\min_{u \in \mathbb{R}^{p^\star}} \frac{1}{2}||\tilde{y^*} - \tilde{\mathbf{D}}^T u||_2^2 \text{ s.t. } ||u||_\infty \leq \lambda \tag{5.23}$$

where the transformation of y^* and $\mathbf{D}$ permits the use of the solution path algorithm as with the univariate setting. The only difference now between (5.14) and (5.23) is that in the latter the newly defined penalty matrix $\tilde{\mathbf{D}}$ is not diagonally dominant and the boundary lemma does not hold. This means that in addition to checking when a coordinate will hit the boundary we have to determine when a boundary coordinate will leave the boundary.

5.4 Comparison with smoothing splines

Hastie and Tibshirani (1993) propose the following penalised least squares criterion

in order to solve (5.2)

$$\ell(\beta_1, ..., \beta_p) = \sum_{t=1}^{n} \{y_t - \sum_{j=1}^{p} \beta_t^{(j)} x_t^{(j)}\} + \sum_{j=1}^{p} \lambda_j \int [\beta_t^{(j)''}]^2 \mathrm{d}t. \tag{5.24}$$

They parameterise the problem by adopting the natural cubic spline basis. Denote them $N_j^1(t), ..., N_j^{n_j}(t)$ where n_j are the unique values of t. Further, let the basis matrix $\mathbf{N}_j$ have tqth element $N_j^q(t)$. Then each $\beta_t^{(j)}$ can be expressed in terms of its basis functions

$$\beta_t^{(j)} = \sum_{\nu=1}^{n_j} \gamma_{\nu j} N_j^{\nu}(t)$$

which can be rewritten as $\boldsymbol{\beta}_j = \boldsymbol{N}_j \boldsymbol{\gamma}_j$. The penalised least squares equation (5.24) can now be written as

$$\ell(\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_2, ...\boldsymbol{\gamma}_p) = ||\boldsymbol{y} - \sum_{j=1}^{p} \boldsymbol{W}_j \boldsymbol{N}_j \boldsymbol{\gamma}_j||_2^2 + \sum_{j=1}^{p} \lambda_j ||\boldsymbol{\gamma}_j||_{\boldsymbol{\Omega}_j}^2 \tag{5.25}$$

where $\boldsymbol{\Omega}_j$ has tqth element $\int N_j^t(t)'' N_j^q(t)'' \mathrm{d}t$, the penalty seminorm $||\boldsymbol{\gamma}_j||_{\boldsymbol{\Omega}_j}^2 = \boldsymbol{\gamma}_j^T \boldsymbol{\Omega}_j \boldsymbol{\gamma}_j$ and $\boldsymbol{W}_j \in \mathbb{R}^{n \times n}$ is a diagonal matrix with the values of $X_t^{(j)}$ on its diagonal. The estimated coefficients can then be obtained

$$\hat{\boldsymbol{\beta}}_j = \boldsymbol{N}_j \boldsymbol{\gamma}_j = \boldsymbol{S}_j(\lambda_j) \boldsymbol{W}_j^{-} (\boldsymbol{y} - \sum_{k \neq j} \boldsymbol{W}_k \boldsymbol{N}_k \boldsymbol{\gamma}_k)$$

where $\boldsymbol{S}_j(\lambda_j) = \boldsymbol{N}_j (\boldsymbol{N}_j^T \boldsymbol{W}_j^2 \boldsymbol{N}_j + \lambda_j \boldsymbol{\Omega}_j)^{-1} \boldsymbol{N}_j^T \boldsymbol{W}_j^2$ and $\boldsymbol{W}_j^{-}$ is the generalized inverse of $\boldsymbol{W}_j$, necessary if some elements of $\boldsymbol{W}_j$ are 0.

The matrix operator $\boldsymbol{S}_j(\lambda_j)$ is a weighted cubic smoothing spline with weights $\boldsymbol{W}_j$ and one can see that this reduces to a cubic smoothing spline when $\mathrm{diag}(\boldsymbol{W}_j) = [1, ..., 1]$. The minimisers $\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, ..., \hat{\boldsymbol{\beta}}_p$ can be found in an iterative "one at a time" manner by using backfitting procedures.

To compare the varying coefficient smoothing spline problem (5.24) with the FuLTV method we re-write the fitted values in the following form (we remove sub-

script j for convenience)

$$N\gamma = N(N^T W^2 N + \lambda\Omega)^{-1} N^T W y = (W^2 + \lambda K)^{-1} W y \tag{5.26}$$

where $K = N^{-T}\Omega N^{-1}$ (the expression in (5.26) is termed the *Reinsch* form). After setting $\hat{u} = N\gamma$ into (5.25) we have the following minimisation problem

$$u \in \arg\min_{u\in\mathbb{R}^n} ||\boldsymbol{y} - \boldsymbol{W}u||_2^2 + \lambda u^T K u \tag{5.27}$$

which has a similar form with (5.4) above (note that $W = \mathbf{X}$). Extending Tibshirani (2014) who studies the differences between trend filtering and smoothing splines for the case of $W = I$ we discuss them in a time-varying setting. A first observation is that $K^{1/2}$ is similar to the discrete derivative operators. For instance, when $k = 3$ Tibshirani (2014) shows that $||K^{1/2}u||_2^2 = ||C^{-1/2}D^{(2)}u||_2^2$ where $C^{-1/2} \in \mathbb{R}^{n\times n}$ is a tridiagonal matrix, with diagonal and off-diagonal elements equal to 2/3 and 1/6 respectively. The main distinction between the two methods, however, lies in the two types of penalties applied i.e. L_2 (ridge) for the smoothing splines and L_1 (lasso) for the FuLTV method. It is well known that the former type shrinks coefficients towards zero (but never set them equal to zero) while the latter gives a sparse solution, i.e. it adaptively sets coefficients equal to zero. Making the analogy to the time-varying model we would expect FuLTV to have better adaptivity properties than smoothing splines. The simulation study of Section 5.7 supports this claim.

5.5 Time-varying estimation as a lasso problem

We now transform problem (5.20) into its lasso equivalent

$$f(\alpha) = \|\mathbf{y} - \mathbf{X}\mathcal{H}\alpha\|_2^2 + \lambda\|\alpha\|_1 \tag{5.28}$$

where $\mathcal{H}$ is a block diagonal matrix having in its diagonal p lower triangular matrices. This transformation derives from the invertibility of the penalty matrix D as given in (5.3) and the fact that the inverse operation inverts the matrices in the diagonal. An action of this kind is generally allowed when the penalty matrix $D \in \mathbb{R}^{m \times p}$ has $\text{rank}(D) = m$ in the sense that a fused lasso problem (or generalised lasso if the penalty matrix does not necessarily have the form of (5.5)) can be transformed to a lasso problem. Now, the two problem formulations (5.20) and (5.28) are the *analysis* and *synthesis* approaches in the context of L_1 penalised estimation with varying coefficients (the terms are used by Elad et al. (2007) to categorise two branches of estimation methods commonly embraced in the signal processing literature). It is yet unclear which approach is easier to work with and Elad et al. (2007) establish the existence of an "unbridgeable gap" between them, even though they favour the analysis approach.

We denote by $H^{(j)}$ each submatrix of the $\mathcal{H}$ diagonal matrix. The form of each $H^{(j)}$ depends on k, the degree of polynomial filtering as in (5.18). From Lemma 2 of Tibshirani (2014) the predictor matrix $H^{(j)} \in \mathbb{R}^{n \times n}$ is given by

$$H^{(j)}_{i,i'} = \begin{cases} i^{i'-1}/n^{i'-1} & \text{for } i = 1, ..., n, i' = 1, ..., k+1 \\ 0 & \text{for } i \leq i' - 1, i' \geq k+2 \\ \sigma^{(k)}_{i-i'+1} k!/n^k & \text{for } i > i' - 1, i' \geq k+2 \end{cases} \tag{5.29}$$

where $\sigma^{(0)} = 1$ for all i and

$$\sigma^{(k)} = \sum_{i'=1}^{i} \sigma^{(k-1)}_{i'} \text{ for } k = 1, 2, 3, ...$$

where $\sigma^{(k)}$ is the kth order cumulative sum of $(1, 1, ..., 1) \in \mathbb{R}^i$. For the piecewise constant and piecewise linear estimators the basis matrices are respectively

$$H^{(j)} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 1 & 1 & \cdots & 0 \\ \vdots & & & & \\ 1 & 1 & 1 & \cdots & 1 \end{pmatrix} \quad \text{if } k = 0 \qquad H^{(j)} = \frac{1}{n} \begin{pmatrix} n & 1 & 0 & 0 & \cdots & 0 \\ n & 2 & 1 & 0 & \cdots & 0 \\ n & 3 & 2 & 1 & \cdots & 0 \\ \vdots & & & & & \\ n & n & n-2 & n-3 & \cdots & 1 \end{pmatrix} \quad \text{if } k = 1 \tag{5.30}$$

Let us assume that $\widetilde{X} = X\mathcal{H}$ and $\hat{\alpha} - \alpha = v$ for notational simplicity. In addition, define the active set $\mathcal{B} = \{t \in [1,n] \quad \text{and} \quad j \in [1,p] : \alpha_{t,j} \neq 0\}$ and $\mathcal{B}^c = \{1,...,n\} \cap \{1,...,p\} \backslash \mathcal{B}$. We use the notation $\mathcal{B}^c$ and not $-\mathcal{B}$ (practically they are the same) to distinguish a set (former) from a set of rows (latter). Finally, denote $s_0 = |\mathcal{B}|$ the cardinality of the active set and $C^n := X^T X$ the Gram-type matrix where $C^n \in \mathbb{R}^{p^\star \times p^\star}$. The matrix C^n can be expressed in a block-wise form

$$C^n = \begin{pmatrix} C^n_{11} & C^n_{12} & \cdots & C^n_{1p} \\ C^n_{21} & C^n_{22} & \cdots & C^n_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ C^n_{p1} & C^n_{p2} & \cdots & C^n_{pp} \end{pmatrix}$$

where $C^n_{\kappa\kappa'} \in \mathbb{R}^{n \times n}$ for $\kappa, \kappa' = 1,...,p$ are diagonal matrices containing $x^\kappa_t x^{\kappa'}_t$ for $t = 1,...,n$. For the lasso problem (5.28) the Gram matrix $C^{\mathcal{H}}$ is $\mathcal{H}^T C_n \mathcal{H}$ and its specific form will depend on the choice of k for the different predictor variables.

An important feature of the lasso estimator is its ability to recover the true pattern of a high-dimensional model asymptotically with high probability. Zhao and Yu (2006) give the following definition of sign consistency.

Definition 5.1. *An estimator β_λ is called sign consistent if and only if*

$$P(sign(\beta) = sign(\beta_\lambda)) \to 1 \ \ as \ \ n \to \infty. \tag{5.31}$$

In order for the above condition to hold for the time-varying linear model the irrepresentable condition should be met. Assume, w.l.o.g, that $\text{diag}(X) = [1,...,1]$

and $j = 1$ such that $C^{\mathcal{H}} = \frac{1}{n}H^T H$. Then, Meinshausen and Yu (2009) give the following definition

Definition 5.2. *The sub-matrix C_{FQ} of C_n is obtained by keeping rows with index in the set F and columns with index in the set Q. Then the irrepresentable condition is fulfilled when the following inequality holds element-wise:*

$$|C^{\mathcal{H}}_{\mathcal{B}^c\mathcal{B}} \left(C^{\mathcal{H}}_{\mathcal{B}\mathcal{B}}\right)^{-1} sign(\beta_{\mathcal{B}})| < 1. \tag{5.32}$$

In our case, with the use of counterexamples we can show that for $k \geq 0$ there exists at least one component i_0 such that (5.32) does not hold. Hence, the asymptotic properties of the lasso, normally adopted in a general regression framework, cannot directly be applied here. However, we can still examine the convergence rate of our estimation method in the ℓ_2 sense i.e. $||\hat{\beta}-\beta||_2^2$. This is simply equal to $||\mathcal{H}\hat{\alpha}-\mathcal{H}\alpha||_2^2 = ||\mathcal{H}(\hat{\alpha}-\alpha)||_2^2$ and hence we can find a bound by examining (5.28). We impose the following assumptions:

(A1) - For any $t = 1, ..., n$ and $j = 1, ..., p$ $\underline{\mathcal{M}} \leq \left(x_t^{(j)}\right)^2 \leq \overline{\mathcal{M}}$.

(A2) - For $p \to \infty$ and $n \to \infty$

$$\log p_n^\star / n \to 0$$

where $p_n^\star$ as in (5.21).

(A3) - The regularisation parameter $\lambda = \sigma\sqrt{\frac{2\log p_n^\star \overline{\mathcal{M}}}{n}}$.

We now prove the following result for the time-varying FL estimator when the underlying coefficient structure of the model (5.2) is piecewise constant.

Proposition 5.1. *Under Assumptions (A1)-(A3), for $k = 0$ and $\{\alpha_{t,j}\}_{t,j\in\mathcal{B}} \in (\alpha_{\min}, \alpha_{\max})$ the following event*

$$\frac{1}{n}||\mathcal{H}(\hat{\alpha}-\alpha)||_2^2 \leq 2\sigma\sqrt{\frac{\log p_n^\star \overline{\mathcal{M}}}{n}} s_0 \alpha_{\max} \underline{\mathcal{M}}^{-1}$$

holds with probability tending to 1.

Proof. If $\hat{\alpha}$ are the minimisers of the lasso problem, then the following holds

$$\frac{1}{n}||y-\widetilde{X}\hat{\alpha}||_2^2+\lambda||\hat{\alpha}||_1 \leq \frac{1}{n}||y-\widetilde{X}\alpha||_2^2+\lambda||\alpha||_1$$

$$\frac{1}{n}||\varepsilon-\widetilde{X}v||_2^2+\lambda||\alpha+v||_1 \leq \frac{1}{n}||\varepsilon||_2^2+\lambda||\alpha||_1$$

$$\frac{1}{n}||\widetilde{X}v||_2^2 \leq \underbrace{\frac{2}{n}v^T\widetilde{X}^T\varepsilon}_{\text{I}}+\lambda||\alpha||_1-\lambda||\alpha+v||_1. \tag{5.33}$$

We turn to the process I which we can write as $\frac{1}{n}v^T\mathcal{H}^T\boldsymbol{X}^T\varepsilon$. Recall from (5.30) that $H^{(j)}$ is a lower triangular matrix. Matrix multiplication of the form $\Theta = H^{(j)}\Delta$ where Δ is a $n \times 1$ vector returns a vector Θ the elements of which are cumulative sums of decreasing length. That is,

$$\Theta = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 0 & 1 & 1 & \cdots & 1 \\ 0 & 0 & 1 & \cdots & 1 \\ \vdots & & & & \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \begin{bmatrix} \Delta_{1,1} \\ \Delta_{2,1} \\ \Delta_{3,1} \\ \vdots \\ \Delta_{n,1} \end{bmatrix} = \begin{bmatrix} \sum_{t=1}^n \Delta_{t,1} \\ \sum_{t=2}^n \Delta_{t,1} \\ \sum_{t=3}^n \Delta_{t,1} \\ \vdots \\ \Delta_{n,1} \end{bmatrix}$$

We invoke the above property to show that

$$\frac{1}{n}v^T\mathcal{H}^T\boldsymbol{X}^T\varepsilon = \frac{1}{n}\sum_{j=1}^{p}\sum_{\ell=1}^{n}v_\ell^{(j)}\left(\sum_{t=\ell}^{n}x_t^{(j)}\varepsilon_t\right) \leq \sum_{j=1}^{p}\sum_{\ell=1}^{n}v_\ell^{(j)}\lambda \tag{5.34}$$

where the last inequality derives from Lemma 5.4 (see Section 5.9).

The inequality (5.33) becomes

$$\frac{1}{n}||\widetilde{X}v||_2^2 \leq \sum_{j=1}^{p}\sum_{\ell=1}^{n}v_\ell^{(j)}\lambda+\lambda||\alpha||_1-\lambda||\alpha+v||_1$$

$$\frac{1}{n}||\widetilde{X}(\hat{\alpha}-\alpha)||_2^2 \leq \lambda||\alpha||_1+\lambda||\hat{\alpha}||_1+\lambda||\alpha||_1-\lambda||\hat{\alpha}||_1 \leq 2\lambda\sum_{\{t,j\}\in\mathcal{B}}\alpha_{tj} \leq 2\sigma\sqrt{\frac{\log p_n^\star}{n}}s_0\alpha_{\max}\overline{\mathcal{M}}.$$

With this result and by Assumption (A1) the proof concludes. □

Some remarks are in order. The rate obtained is the same with FLSA up to the terms $\log p_n^\star$, $\underline{\mathcal{M}}$ and $\overline{\mathcal{M}}$. In general, we note that for the nonparametric regression problem (5.3) Tibshirani (2014) shows that the kth order trend filtering attains the minimax rate, i.e. $\mathcal{O}(n^{-(2k+2)/(2k+3)})$ which for $k = 0$ is better than that of Proposition 5.1. Proposition 5.1 can be extended to high order cases ($k > 0$) by finding an appropriate bound for the process (5.34).

5.6 Degrees of freedom and model selection

The degrees of freedom measures the complexity of the model and quantitatively describes the effective number of parameters used in the fit by a given procedure. An estimate of degrees of freedom allows us to use model selection criteria. Before discussing in detail model selection criteria, we provide the definition of the degrees of freedom as in e.g. Efron (1986) or Efron et al. (2004). Let us assume that $y \in \mathbb{R}^n$ is drawn from the following normal model

$$y \sim \mathcal{N}(\mu, \sigma^2 I)$$

where $\mathbf{X}$ is fixed. For a function $h : \mathbb{R}^n \to \mathbb{R}^n$ (with ith coordinate function $h_i : \mathbb{R}^n \to \mathbb{R}$), the degrees of freedom of function h is defined as

$$\mathrm{df}(h) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \mathrm{Cov}(h_i(y), y_i).$$

In our context, $h(y) = \mathbf{X}\hat{\beta}_\lambda(y)$ for fixed λ.

Tibshirani and Taylor (2012) give an expression for the degrees of freedom with minimal assumptions (Theorem 3). For an arbitrary $\mathbf{X}$, $\mathbf{D}$ and $\lambda \geq 0$, an unbiased estimate for the degrees of freedom for the generalised lasso is given by

$$\mathrm{df}(\mathbf{X}\beta_\lambda) = \mathbb{E}[\dim(\mathbf{X}\mathrm{null}(\mathbf{D}_{-\mathcal{B}}))]. \tag{5.35}$$

We now consider null($\mathbf{D}_{-\mathcal{B}}$). Since $\mathbf{D}$ is a block diagonal matrix of $D_j^{k_j+1}$ for $j = 1, ..., p$ (e.g. $k_j = 0$ gives the 1-dimensional fused lasso penalty which leads to piecewise constant solutions), then

$$\text{null}(\boldsymbol{D}_{-\mathcal{B}}) = \text{span}\{\text{null}(D_{-\mathcal{B}}^{k_1+1})\times\{0\}\times...\cap\{0\}\times\text{null}(D_{-\mathcal{B}}^{k_2+1})\times\{0\}\times...\cap...\{0\}\times\text{null}(D_{-\mathcal{B}}^{k_p+1})\}. \tag{5.36}$$

Therefore, at every iteration of the algorithm we can find the null space of $\mathbf{D}$ by looking only at the null space of $D_j^{k_j+1}$ associated with the covariate $X^{(j)}$.

In addition, the multiplication with $\mathbf{X}$ does not change the dimension of the null space of $\mathbf{D}$ and, therefore (5.35) reduces to $\text{df}(\mathbf{X}\beta_\lambda) = \mathbb{E}[\dim(\text{null}(\mathbf{D}_{-\mathcal{B}}))]$. Practically, this means the degrees of freedom are given by

$$\text{df}(\mathbf{X}\beta_\lambda) = \sum_{j=1}^{p} \dim(\text{null}(D_{-\mathcal{B}}^{k_j+1}))$$

or, put it simply, the degrees of freedom are given by examining the null space of each $D_{-\mathcal{B}}^{k_j+1}$ for $j = 1, ..., p$ which in turn depends on the number of knots in $\hat{\beta}^{(j)}$ and the kth degree of polynomial filtering. Hence, for each $\hat{\beta}^{(j)}$ the degrees of freedom derives from known results on the generalised lasso and trend filtering (see Tibshirani and Taylor (2012)), i.e.

$$\text{df}(\hat{\beta}^{(j)}) = \mathbb{E}[\text{number of knots in } \hat{\beta}^{(j)}] + k + 1$$

where the number of number of knots in $\hat{\beta}^{(j)}$ is the number of non-zero entries in $D^{(k_j+1)}\hat{\beta}^{(j)}$.

From the above analysis we can see that the FuLTV method gives an exact representation of the degrees of freedom of any fitted model. In the context of other linear estimation methods (linear in the data y), such as in Hastie and Tibshirani (1993), the calculation of the degrees of freedom is based on approximate methods.

For a simple smooth fit $\hat{y} = \mathbf{S}y$ where $\mathbf{S}$ is the operator that produces the fitted term $\mathbf{X}\hat{\beta}_\lambda$, Hastie and Tibshirani (1990) consider three definitions of the degrees of freedom $\mathrm{df}(\beta_\lambda)$, i.e. $\mathrm{tr}(\mathbf{S}), \mathrm{tr}(\mathbf{S}^T\mathbf{S}), \mathrm{tr}(2\mathbf{S} - \mathbf{S}\mathbf{S}^T)$. These are not easy to calculate (see Zhang (2003) for a discussion on this topic as well as for empirical formulas for degrees of freedom). For instance, for the latter definition Hastie and Tibshirani (1993) make use of the following approximation (in the context of VC models)

$$\mathrm{tr}(2\mathbf{S} - \mathbf{S}\mathbf{S}^T) \approx 1.25\mathrm{tr}(\mathbf{S}) - 0.5$$

from Hastie and Tibshirani (1990).

Once we obtain the degrees of freedoms at certain values of the regularisation parameter λ, the final model needs to be chosen. One approach would be cross-validation, in which the observations are divided into a training set and a test set. Then, a model is estimated on the former set and its accuracy is tested on the latter set using an appropriate error measure; this procedure is repeated and the error measure is averaged over different test sets. This method, however, is computationally intensive and it does not take advantage of the path-following structure of our algorithm. Instead of cross-validation we can use certain information criteria, such as the Cp statistic (Mallows (1973)) or the Bayesian Information Criterion (BIC), also known as the Schwarz criterion (Schwarz (1978)) and considered e.g. by Bai and Perron (2003) in the context of change-point estimation in linear models. The criteria are given below respectively,

$$C_p(\lambda) = \left\| y - \mathbf{X}\hat{\beta}_\lambda \right\|_2^2 - n\sigma^2 + 2\sigma^2 \mathrm{df}(\mathbf{X}\hat{\beta}_\lambda)$$

and

$$\mathrm{BIC}(\lambda) = \log \left\| \left(y - \mathbf{X}\hat{\beta}_\lambda \right) / n \right\|_2^2 + \mathrm{df}(\mathbf{X}\hat{\beta}_\lambda) \log(n)/n$$

where $\hat{\beta}_\lambda$ are the estimated coefficients at a fixed value of λ. Due to the fact that FuLTV is a path algorithm which consists of sub-models then one can choose the final model such that $\lambda^\star = \arg\min_\lambda \text{BIC}(\lambda)$ or $\lambda^\star = \arg\min_\lambda C_p(\lambda)$.

For decreasing λ the ℓ_2 norm of the estimated residuals are monotonically decreasing and, hence, the minimum Cp or BIC(λ) value will be found somewhere between the critical points. Ideally, this implies that at every iteration we can stop the algorithm as soon as we calculate a value of C_p or BIC that is larger from the one obtained in the previous iteration. However, in our simulations, we notice that stopping the algorithm as soon as a minimum value of C_p or BIC is obtained does not work efficiently. The main reason behind this is that it is difficult to know how many times a coordinate will leave the boundary. Both information criteria penalise for extra complexity through the number of the estimated knots. A direct consequence is the early termination of the algorithm as it gives the signal that complexity has increased very fast.

We propose to allow the algorithm to run several steps and then choose the estimated coefficients that return the global minimum for either criterion. It is difficult to bound the number of steps required, but we find that $\mathcal{O}(p_n^\star)$ iterations work well. It is noted that Mammen and van de Geer (1997), who devise an algorithm in the context of nonparametric regression using total variation penalties, conjecture that roughly $\mathcal{O}(n)$ cycles are necessary.

The above discussion is illustrated by an example. We consider a simple piecewise-

stationary AR(1) process

$$y_t = \begin{cases} 0.9y_{t-1} + \varepsilon_t, \text{ for } 1 \leq t \leq 100 \\ \varepsilon_t, \text{ for } 101 \leq t \leq 300 \\ 0.9y_{t-1} + \varepsilon_t, \text{ for } 301 \leq t \leq 500 \end{cases} \tag{5.37}$$

where $\varepsilon_t \sim \mathcal{N}(0, 1)$. A realisation of the process is shown at the top-left panel of Figure 5.3. From the bottom-right panel of the same figure it is evident that the BIC criterion would signal an early termination of the path algorithm if we chose to stop the algorithm as soon as a minimum value for BIC is achieved. This is due to the fact that until the fourth point (iteration) the BIC function monotonically increases. Allowing the algorithm to run for many cycles shows a significant decrease, reaching the minimum BIC value after ten iterations. Similar arguments can be made for the C_p criterion (bottom-right panel).

Finally, in our simulations we find that BIC works better than C_p which tends to over-fit the data, see for example Figure 5.3 where we observe that the minimum BIC is obtained earlier at iteration 10 while for C_p at 20.

5.7 Simulation study

We conduct a set of simulations to assess the performance of our method and compare it against the smoothing splines (henceforth, SStv). For the latter we use the *R* package *mgcv* (Wood (2014)) which is a repository of generalised additive modelling and varying coefficient models functions; we refer the reader to Wood (2006) for a guide to the *mgcv* package. The main function adopted for the simulations is *gam*[1].

[1]We acknowledge help from Simon Wood in regards to the use of this function.

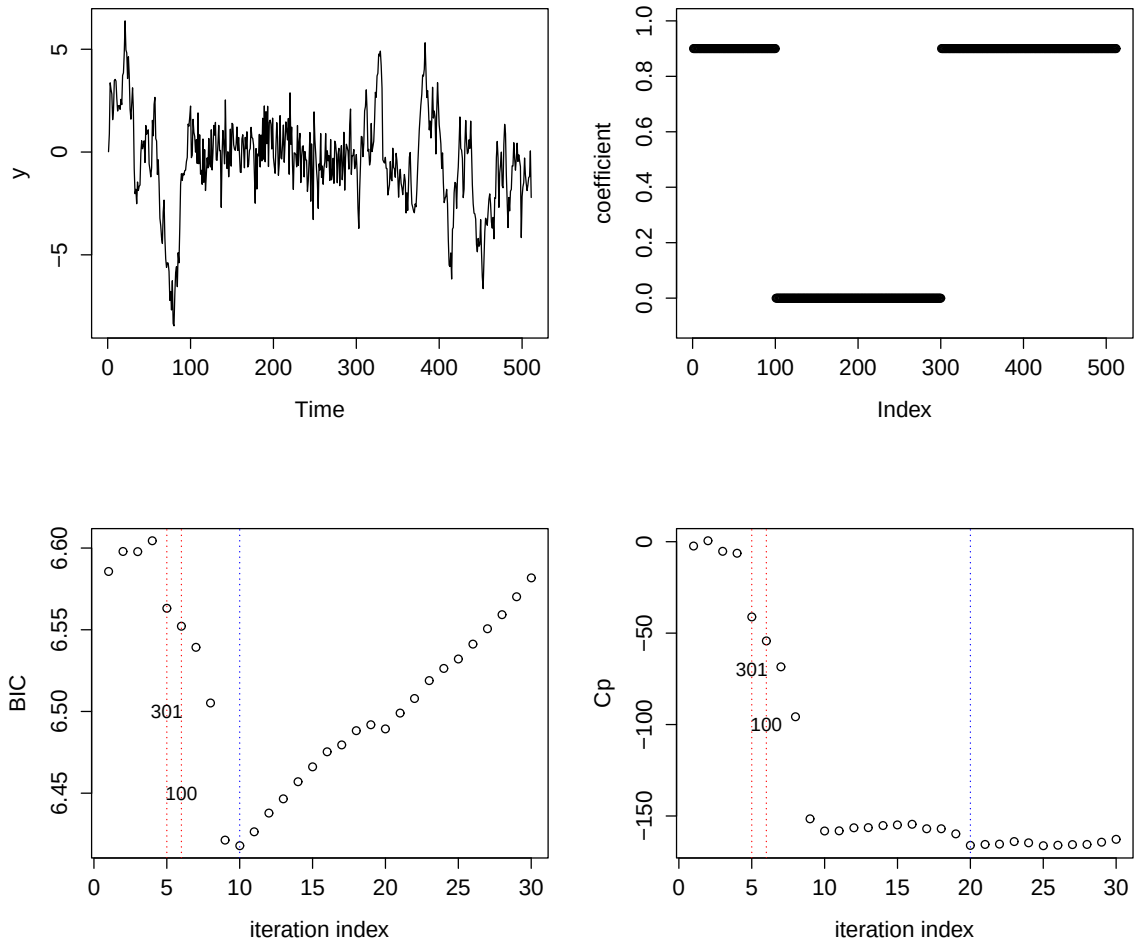

Figure 5.3: Plots of two Information Criteria: BIC (down-left) and C_p(down-right) for a non-stationary model (top-left) with time-varying AR(1) coefficients (top-right)

To compare the performance of the methods we calculate the mean absolute deviation error

$$\mathcal{E}_{\text{MAD}}^{(j)} = \frac{1}{nB} \sum_{\ell'}^{B} \sum_{t=1}^{n} \left| \beta_t^{(j);\ell'} - \hat{\beta}_t^{(j);\ell'} \right|$$

where $j = 0, ..., p$ and $\hat{\beta}_t^{(j);\ell'}$ are the estimated coefficients obtained from either of the two methods and B is the number of experiments.

The Model: We generate y_t according to the model (5.2) for $p = 3$ of sample size $n = 200, 500, 1000$ and $\sigma = 0.5, 1, 2$. The coefficients curves for the first two

covariates are taken from Huang et al. (2002), i.e.

$$\beta_t^{(0)} = 15 + 20\,\sin(t\pi/60)$$

$$\beta_t^{(1)} = 2 - 3\,\cos\{(t - 25\pi)/15\}$$

The third curve is a combination of $\beta_t^{(0)}$ and $\beta_t^{(1)}$

$$\beta_t^{(2)} = \begin{cases} 15 + 20\,\sin(t\pi/60), & \text{for } \lfloor n/2 \rfloor \\ c^0 - 3\,\cos\{(t - 25\pi)/15\}, & \text{for } \lfloor n/2 \rfloor + 1 \end{cases}$$

where c^0 is a constant selected such that the curve is roughly continuous at $t = \lfloor n/2 \rfloor$. The purpose of $\beta_t^{(2)}$ is to assess the performance of the two methods when a segment of a curve exhibits more variability than another. From panel c of Figure 5.4 it appears that the right part of the curve is "wiggly", while the left part is smooth.

Finally, the independent variables $X_t^{(j)} \sim \mathcal{N}(1, 1)$ for $j = 2, 3$ and $X_t^{(1)} = [1, ..., 1]^T$. We select $k = 3$ i.e. the cubic trend filtering and BIC(λ) for model selection. For every pair (n, σ) we repeat the experiment $B = 100$ times.

Table 5.1 summarises the results. For small sample size SStv shows good performance, close to that of FuLTV. Particularly, when $n = 200$ and in low signal-to-noise cases SStv outperforms FuLTV yet by a margin. However, as the sample size increases FuLTV does well compared with SStv and the difference in $\mathcal{E}_{\text{MAD}}^{(j)}$ for all covariates is higher in many instances. We also notice that FuLTV almost always outperforms SStv when "wiggliness" is present in the underlying coefficient curve (like $\beta_t^{(1)}$). However, for $n = 1000$ the results indicate that SStv gives better estimates when a curve is both wiggly and smooth which is the case with $\beta_t^{(2)}$. Perhaps, the reason is that SStv does well in the smooth part of the curve and this is reflected in the total $\mathcal{E}_{\text{MAD}}^{(j)}$.

Table 5.1: Simulation results for the model described in Section 5.7. For every coefficient curve the mean of $\mathcal{E}_{MAD}^{(j)}$ for $j = 0, 1, 2$ is reported over $B = 100$ repetitions.

$n = 200$	$\mathcal{E}_{MAD}$					
	FuLTV			**SStv**		
Coefficients	0.5	1	2	0.5	1	2
$\beta^{(0)}$	0.0679	0.1179	0.2217	0.0723	0.1198	0.2134
$\beta^{(1)}$	0.0832	0.1509	0.2791	0.0932	0.1658	0.2840
$\beta^{(2)}$	0.0879	0.1559	0.2988	0.0954	0.1613	0.2790

$n = 500$	$\mathcal{E}_{MAD}$					
	FuLTV			**SStv**		
Coefficients	0.5	1	2	0.5	1	2
$\beta^{(0)}$	0.1216	0.1980	0.4274	0.1197	0.1992	0.4202
$\beta^{(1)}$	0.1660	0.2763	0.5812	0.1701	0.2920	0.5638
$\beta^{(2)}$	0.1594	0.2857	0.5153	0.1744	0.3091	0.5195

$n = 1000$	$\mathcal{E}_{MAD}$					
	FuLTV			**SStv**		
Coefficients	0.5	1	2	0.5	1	2
$\beta^{(0)}$	0.2720	0.3718	0.5229	0.3640	0.5561	0.6558
$\beta^{(1)}$	0.9027	1.0107	0.9589	1.0317	1.5765	1.7687
$\beta^{(2)}$	0.6087	0.7517	0.7477	0.5040	0.7616	0.8848

Overall, FuLTV shows a better performance even though this should not be seen as a criticism of SStv, as it performs well in other cases (especially in the small samples).

5.8 Applications

5.8.1 Ethanol data

For the purpose of motivation, we consider the same example from Hastie and Tibshirani (1993) who estimate a varying-coefficient model using 88 observations on the exhaust from an engine fueled by ethanol. The data set, first analysed by Cleveland

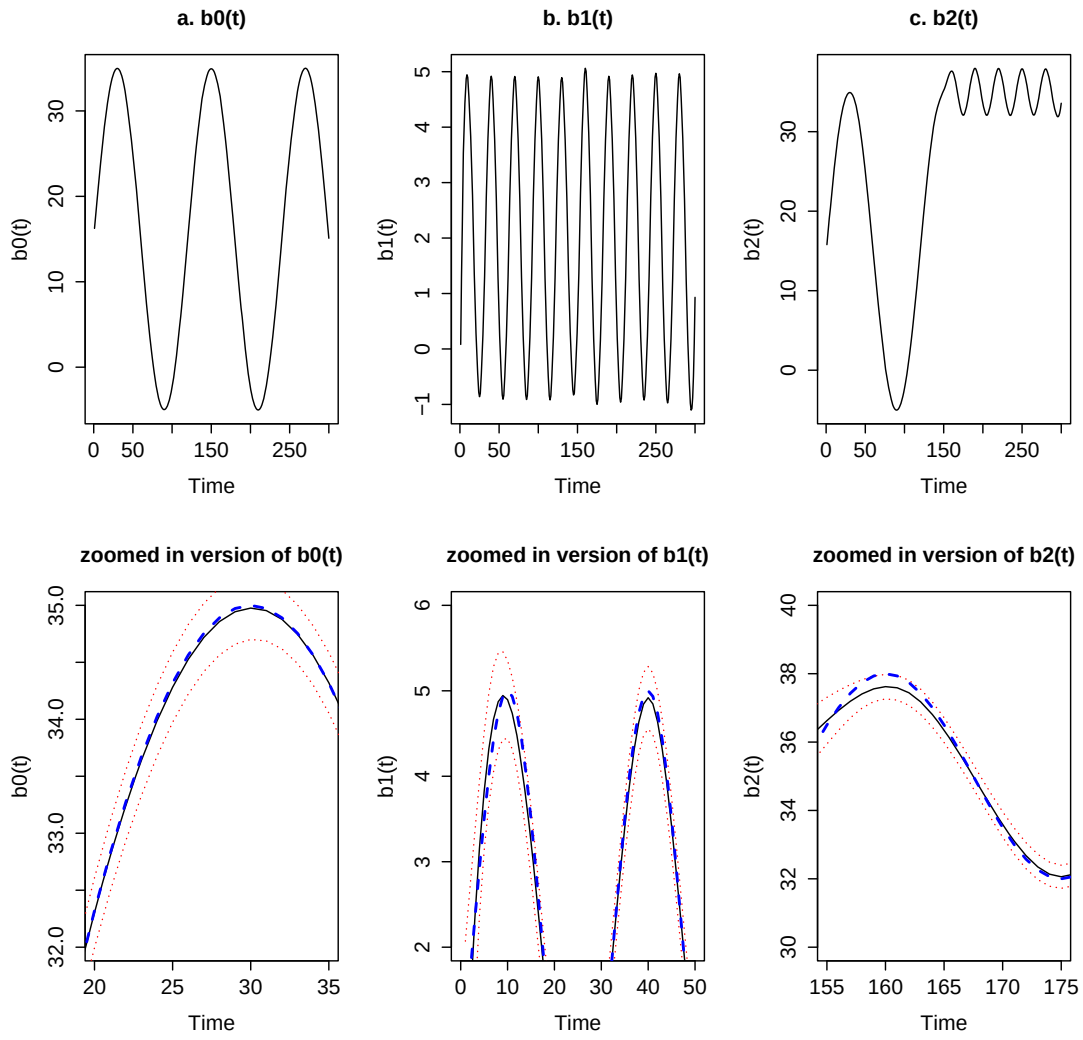

Figure 5.4: Top panels show the estimated coefficients averaged over 100 *repetitions for the model described in Section 5.7. Bottom panels are zoomed in versions of the estimated coefficients. The estimated coefficients are denoted by the black solid line and their point-wise standard deviations (calculated over* 100 *repetitions and multiplied by* 2 *for scale reasons) are denoted by the two red symmetric lines. The true coefficient functions* $\beta_t^{(j)}$ *for* $j = 0, ..., 2$ *are denoted by the blue dashed lines.*

et al. (1991), is available from the *R* package *lattice* (Sarkar (2008)). It consists of the response variable NOx_i (concentration of nitrogen dioxide) and two predictors E_i and C_i which measure the fuel-air ratio and the compression ratio of the engine,

respectively. The authors observe that C_i interacts with E_i and they suggest the following model

$$NOx_i = \beta_0(E_i) + \beta_1(E_i)C_i + \varepsilon_i. \tag{5.38}$$

To estimate this model we choose the following penalty matrix

$$\mathbf{D} = \begin{pmatrix} D_{\beta_0}^{(k_1+1)} & 0 \\ 0 & D_{\beta_1}^{(k_2+1)} \end{pmatrix}.$$

Without restricting the choice of different trend filtering orders we set $k_1 = k_2 = 3$ which is the cubic trend filtering matrix. Higher orders did not significantly improve the estimation. We do not use an information criterion, but instead we extract a solution when the total degrees of freedom are 20. These are roughly the degrees of freedom obtained from the *gam* function in the *mgcv* package which uses cross-validation to select the penalty parameter (Hastie and Tibshirani (1993) choose 8 degrees of freedom for each of the two predictors).

In addition to FuLTV and SStv, we also estimate the following least squares model

$$NOx_i = \beta_0 + \beta_1 E_i^2 + \varepsilon_i.$$

Results are shown in Figure 5.5. The FuLTV method does well in this example indicated by the q-q plots and it achieves a residual sum of squares of 2.53, while the smoothing splines achieve 2.62 (Hastie and Tibshirani (1993) estimate it to be 2.65) and the least squares 107.73.

5.8.2 Boston Housing data

We present an application of the FuLTV method where we use BIC to select the regularisation parameter. The Boston Housing data, found in the *R* package *ml-*

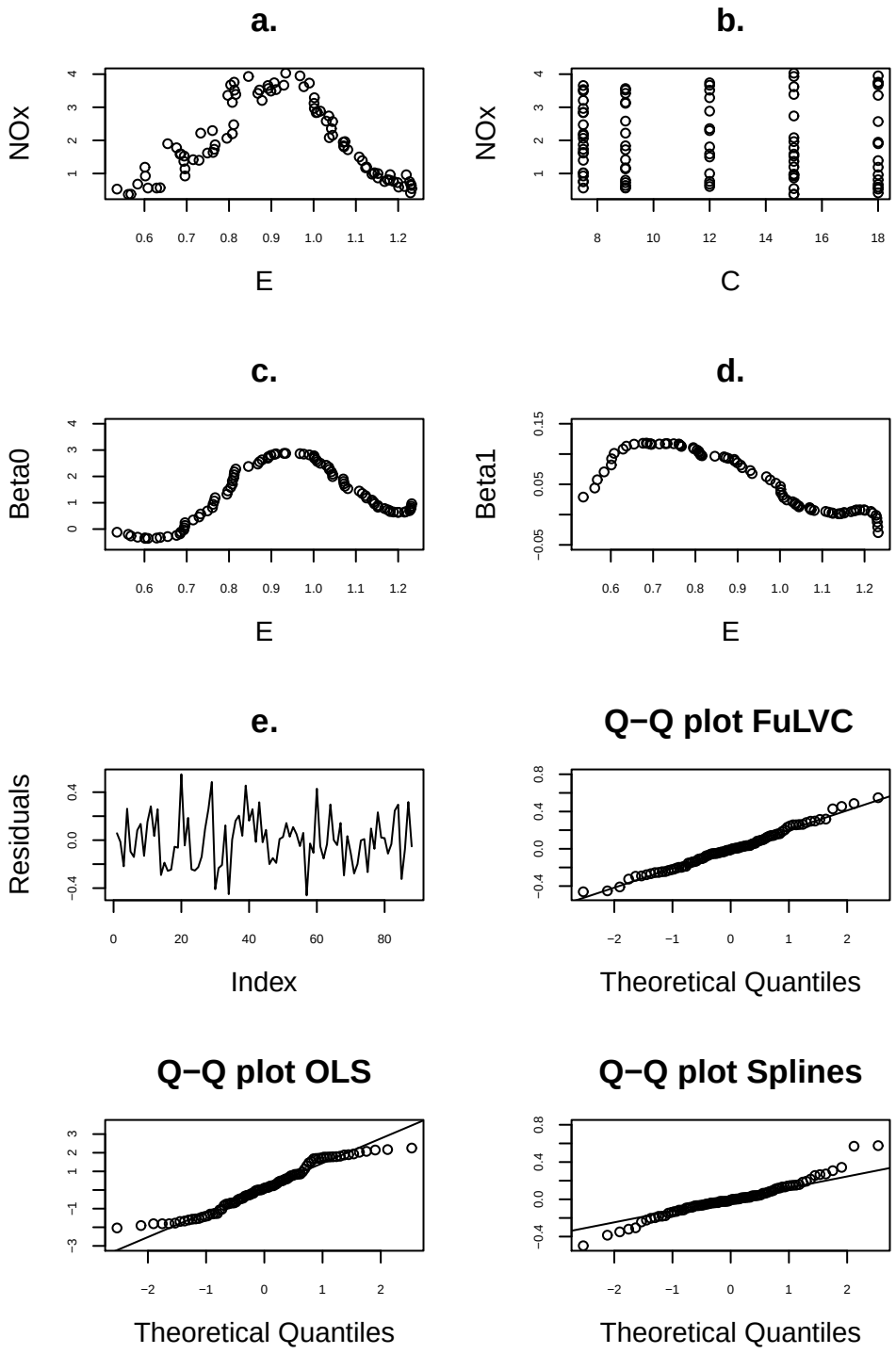

Figure 5.5: The estimated varying coefficients β_0 (c.) and β_1 (d.) for the ethanol example for $\lambda = 30.76$.

bench (Leisch and Dimitriadou (2010)) and first analysed in the context of varying-coefficient model by Fan and Huang (2005), consists of the median value of owner-occupied homes (MEDV) in the Boston area along with other variables. Here, we only consider the predictors that have been shown to be the most relevant in predicting MEDV when the coefficients are allowed to vary (Wang and Xia (2009) and Antoniadis et al. (2013)), i.e. CRIM (a measure of crime), RM (average number of rooms in a dwelling), PRATIO (student-teacher ratio by town) and TAX (full-value property-tax rate per $10,000). Following Fan and Huang (2005) and Wang and Xia (2009) we take the underlying covariate to be LSTAT (we denote it by u), the percentage of the lower status of the population. In addition, the response and predictor variables are transformed so that their marginal distribution to be approximately $\mathcal{N}(0,1)$. To achieve this we use Box-Cox transformations as in Antoniadis et al. (2013). In accordance with all these studies the intercept is also allowed to vary. Similarly with the ethanol example we choose the cubic trend filtering for each of the predictors. By fitting the model

$$MEDV_i = \beta_0(u) + \beta_1(u)CRIM_i + \beta_2(u)RM_i + \beta_3(u)PRATIO_i + \varepsilon_i \quad \text{for} \quad i = 1, ..., 504$$

we get the coefficient curves shown in Figure 5.6. From panel c it is evident that e.g. the house prices are positively related to the number of rooms in a dwelling (RM), but this relation diminishes when moving to areas with lower status. The q-q plot in the same figure confirms that our method performs well. Finally, we note that the RSS for the FuLTV method is 77.28, lower than that of the smoothing splines (79.53), but only marginally.

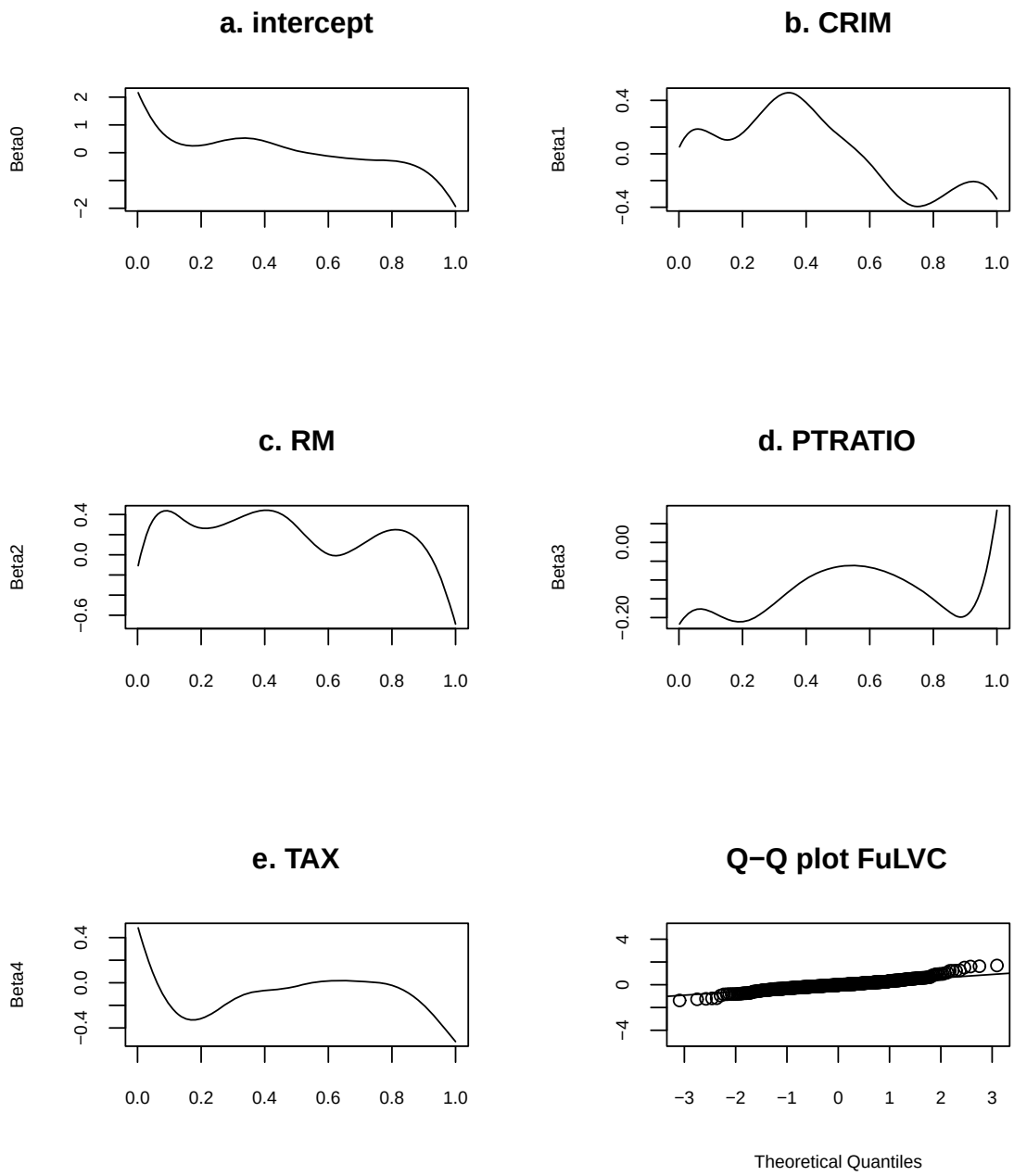

Figure 5.6: The estimated varying coefficients β_0 (a.), β_1 (b.), β_2 (c.), β_3 (d.), and β_4 (e.), and the q-q *plot (bottom right) for the Boston Housing data.*

5.9 Proofs

Proof of Lemma 5.1. First we have the subgradient equation which is

$$g_t = \hat{\beta}_t \underline{x}_t^2 - \underline{y}_t \underline{x}_t + \lambda_1 s_t^{(1)} + \lambda_2 s_t^{(2)}$$

We now insert $\hat{\beta}_t(\lambda_1) = \text{sign}(\hat{\beta}_t(0))(|\hat{\beta}_t(0)| - \frac{\lambda_1}{\underline{x}_t^2})^+$ into the subgradient equation and examine two cases:

Case 1: $|\beta_t(0)| > \frac{\lambda_1}{\underline{x}_t^2}$

$$g_t(\lambda_1) = \hat{\beta}_t(0)\underline{x}_t^2 - \lambda_1 \text{sign}(\hat{\beta}_t(0)) - \underline{y}_t \underline{x}_t + \lambda_1 s_t^{(1)}(\lambda_1) + \lambda_2 s_{t+1}^{(2)}(\lambda_1) - \lambda_2 s_t^{(2)}(\lambda_1)$$

$$\begin{aligned} g_t(\lambda_1) \;=\; & \hat{\beta}_t(0)\underline{x}_t^2 - \lambda_1 \text{sign}(\hat{\beta}_t(0)) - \underline{y}_t \underline{x}_t + \lambda_1 s_t^{(1)}(\lambda_1) + \\ & + \lambda_2 \{\text{sign}(\text{sign}(\hat{\beta}_{t+1}(0))(|\hat{\beta}_{t+1}(0)| - \frac{\lambda_1}{\underline{x}_t^2}) - \text{sign}(\hat{\beta}_t(0))(|\hat{\beta}_t(0)| - \frac{\lambda_1}{\underline{x}_t^2}))\} - \\ & - \lambda_2 \{\text{sign}(\text{sign}(\hat{\beta}_t(0))(|\hat{\beta}_t(0)| - \frac{\lambda_1}{\underline{x}_t^2}) - \text{sign}(\hat{\beta}_{t-1}(0))(|\hat{\beta}_{t-1}(0)| - \frac{\lambda_1}{\underline{x}_t^2}))\}. \end{aligned}$$

Note from the above two equations that the signs of the total variation penalties do not change, since soft-thresholding does not change the ordering of β_t, β_{t-1} and β_{t+1}. So for any $\lambda_1 > 0$ it holds that $s_t^{(2)}(\lambda_1) = s_t^{(2)}(0)$. Hence we have that

$$g_t(\lambda_1) = \hat{\beta}_t(0)\underline{x}_t^2 - \underline{x}_t \underline{y}_t + \lambda_2 s_t^{(2)}(0) - \lambda_1 \text{sign}(\hat{\beta}_t(0)) + \lambda_1 s_t^{(1)}(\lambda_1).$$

By the assumption of $\hat{\beta}_t(0)$ being a solution, the first three terms of the equation are equal to zero and thus $g_t(\lambda_1) = 0$.

Case 2: $|\beta_t(0)| \leq \frac{\lambda_1}{\underline{x}_t^2}$. We have that

$$\begin{aligned} g_t(\lambda_1) \;=\; & -\underline{y}_t \underline{x}_t + \lambda_1 s_t^{(1)}(\lambda_1) + \\ & + \lambda_2 \{\text{sign}(\text{sign}(\hat{\beta}_{t+1}(0))(|\hat{\beta}_{t+1}(0)| - \frac{\lambda_1}{\underline{x}_t^2}) - \text{sign}(\hat{\beta}_t(0))(|\hat{\beta}_t(0)| - \frac{\lambda_1}{\underline{x}_t^2}))\} - \\ & - \lambda_2 \{\text{sign}(\text{sign}(\hat{\beta}_t(0))(|\hat{\beta}_t(0)| - \frac{\lambda_1}{\underline{x}_t^2}) - \text{sign}(\hat{\beta}_{t-1}(0))(|\hat{\beta}_{t-1}(0)| - \frac{\lambda_1}{\underline{x}_t^2}))\} \\ & = 0 \end{aligned}$$

by choosing $s_t^{(1)}(\lambda_1) = \beta_t(0)\underline{x}_t^2/\lambda_1 \in [-1, 1]$.

Note that in the extreme and without interest - in terms of real applications - case, the soft-thresholding quantity would be $\lambda_1/\underline{x}_t^2$. One can see that by applying a lasso regression on the univariate model $\underline{y}_t = \beta_t \underline{x}_t + \varepsilon_t$. □

Lemma 5.3. *The matrix $\tilde{D}\tilde{D}^T \in \mathbb{R}^{(n-1)\times(n-1)}$ is diagonally dominant i.e.*

$$(\tilde{D}\tilde{D}^T)_{i,i} \geq \sum_{j \neq i} (\tilde{D}\tilde{D}^T)_{i,j} \quad \textit{for} \quad i = 1, ..., n-1.$$

Proof. First note that $\tilde{D}\tilde{D}^T = D\mathcal{Q}^T\check{X}^T\check{X}\mathcal{Q}D^T = D\mathcal{Q}D^T$ from (5.13). The matrix $\mathcal{Q}$ is diagonal with entries $\frac{1}{x_t^2+\lambda_3}$ for $t = 1, ..., n$; Every row $i = 2, ..., n-2$ of $\tilde{D}\tilde{D}^T$ is

$$\left[0, ..., 0, \frac{1}{x_i^2+\lambda_3}, \frac{1}{x_i^2+\lambda_3} + \frac{1}{x_{i+1}^2+\lambda_3}, \frac{1}{x_{i+1}^2+\lambda_3}, 0, ..., 0\right]$$

where the middle term is on the diagonal (i, i), hence the off-diagonal terms $\frac{1}{x_i^2+\lambda_3}, \frac{1}{x_{i+1}^2+\lambda_3}$ is equal to the diagonal term. When $i = 1$

$$\left[\frac{1}{x_1^2+\lambda_3} + \frac{1}{x_2^2+\lambda_3}, \frac{1}{x_2^2+\lambda_3}, 0, ..., 0\right].$$

Hence, the first term (on the diagonal) is always larger than the second (off the diagonal). Finally, when $i = n-1$

$$\left[0, ..., 0, \frac{1}{x_{n-1}^2+\lambda_3} + \frac{1}{x_n^2+\lambda_3}\right].$$

This concludes the proof. □

Proof of Lemma 5.2. The boundary proof below applies to the time-varying model (5.16) over M partitions of y_t. Recall that by partitioning the (5.16) model we can safely make divisions since $\underline{x}_t > 0$ for $\forall t = 1, ..., n$.

$$\min_{u_t} \frac{1}{2}\left(\underline{y}_t - (\frac{u_{t-1}}{\underline{x}_t} - \frac{u_t}{\underline{x}_t})\right)^2 + \frac{1}{2}\left(\underline{y}_{t+1} - (\frac{u_t}{\underline{x}_{t+1}} - \frac{u_{t+1}}{\underline{x}_{t+1}})\right)^2$$

s.t. $|u_t| \leq \lambda$. Differentiating with respect to u_t,

$$\left(\underset{\sim}{\mathrm{y}}_t - \frac{u_{t-1}}{\underset{\sim}{\mathrm{x}}_t} + \frac{u_t}{\underset{\sim}{\mathrm{x}}_t}\right)\frac{1}{\underset{\sim}{\mathrm{x}}_t} - \left(\underset{\sim}{\mathrm{y}}_{t+1} - \frac{u_t}{\underset{\sim}{\mathrm{x}}_{t+1}} + \frac{u_{t+1}}{\underset{\sim}{\mathrm{x}}_{t+1}}\right)\frac{1}{\underset{\sim}{\mathrm{x}}_{t+1}} = 0.$$

This is a quadratic function where its solution lies in an interval, hence

$$u_t = T_\lambda\left(\frac{\underset{\sim}{\mathrm{x}}_t\underset{\sim}{\mathrm{y}}_{t+1} - \underset{\sim}{\mathrm{x}}_{t+1}\underset{\sim}{\mathrm{y}}_t + u_{t-1}\frac{\underset{\sim}{\mathrm{x}}_{t+1}}{\underset{\sim}{\mathrm{x}}_t} + u_{t+1}\frac{\underset{\sim}{\mathrm{x}}_t}{\underset{\sim}{\mathrm{x}}_{t+1}}}{\frac{\underset{\sim}{\mathrm{x}}_{t+1}}{\underset{\sim}{\mathrm{x}}_t} + \frac{\underset{\sim}{\mathrm{x}}_t}{\underset{\sim}{\mathrm{x}}_{t+1}}}\right).$$

Then, the proof proceeds as in T&T, i.e.

$$|u_{\lambda_0,i} - u_i^{(1)}| = T_{\lambda_0}\left(\frac{u^{(1)}_{\lambda_0,t-1}\frac{\underset{\sim}{\mathrm{x}}_{t+1}}{\underset{\sim}{\mathrm{x}}_t} + u^{(0)}_{\lambda_0,t+1}\frac{\underset{\sim}{\mathrm{x}}_t}{\underset{\sim}{\mathrm{x}}_{t+1}}}{\frac{\underset{\sim}{\mathrm{x}}_{t+1}}{\underset{\sim}{\mathrm{x}}_t} + \frac{\underset{\sim}{\mathrm{x}}_t}{\underset{\sim}{\mathrm{x}}_{t+1}}}\right) - T_\lambda\left(\frac{u^{(1)}_{t-1}\frac{\underset{\sim}{\mathrm{x}}_{t+1}}{\underset{\sim}{\mathrm{x}}_t} + u^{(0)}_{t+1}\frac{\underset{\sim}{\mathrm{x}}_t}{\underset{\sim}{\mathrm{x}}_{t+1}}}{\frac{\underset{\sim}{\mathrm{x}}_{t+1}}{\underset{\sim}{\mathrm{x}}_t} + \frac{\underset{\sim}{\mathrm{x}}_t}{\underset{\sim}{\mathrm{x}}_{t+1}}}\right)$$

which is $\leq \max\left\{\left|\frac{u^{(1)}_{\lambda_0,t-1}\frac{\underset{\sim}{\mathrm{x}}_{t+1}}{\underset{\sim}{\mathrm{x}}_t} - u^{(1)}_{t-1}\frac{\underset{\sim}{\mathrm{x}}_{t+1}}{\underset{\sim}{\mathrm{x}}_t}}{\frac{\underset{\sim}{\mathrm{x}}_{t+1}}{\underset{\sim}{\mathrm{x}}_t} + \frac{\underset{\sim}{\mathrm{x}}_t}{\underset{\sim}{\mathrm{x}}_{t+1}}}\right|, \lambda_0 - \lambda\right\}$ and thus $||u_{\lambda_0} - u^{(1)}||_\infty \leq \lambda_0 - \lambda$ by noticing that $\left|\frac{\frac{\underset{\sim}{\mathrm{x}}_{t+1}}{\underset{\sim}{\mathrm{x}}_t}}{\frac{\underset{\sim}{\mathrm{x}}_{t+1}}{\underset{\sim}{\mathrm{x}}_t} + \frac{\underset{\sim}{\mathrm{x}}_t}{\underset{\sim}{\mathrm{x}}_{t+1}}}\right| \leq 1.$ □

Lemma 5.4. *Let* $V_i^{(j)} = \sum_{t=i}^n x_t\varepsilon_t$ *for* $i = 1, ..., n$ *and* $j = 1, ..., p$. *Define the event* $\Lambda_n = \left\{\max_{i=1,...,n,j=1,...,p}|V_i^{(j)}| > n\lambda\right\}$. *Then, if* $\lambda = \sigma\sqrt{\frac{2\log p_n^\star}{n}}$, *the following holds*

$$\mathbb{P}(\Lambda_n) \to 1.$$

Proof. From standard results (see Knight and Fu (2000) or Hebiri and van de Geer (2011)) $V_i^{(j)} \sim \mathcal{N}(0, (n-i+1)\sigma^2\overline{\mathcal{M}})$. Now, we have that

$$\begin{aligned}\mathbb{P}\left(\left\{\max_{i=1,...,n,j=1,...,p}|V_i^{(j)}| \leq n\lambda\right\}\right) &\leq p_n^\star \max_{i=1,...,n,j=1,...,p}\mathbb{P}\left(|V_i^{(j)}| \leq n\lambda\right)\\ &\leq p_n^\star\exp\left(-\frac{n^2\lambda^2}{2\sigma^2 n\overline{\mathcal{M}}}\right) \to 1.\end{aligned}$$

This concludes the proof. □

Chapter 6

Conclusions and future directions

In this thesis we have considered randomised and L_1 penalty approaches to the segmentation of time series and regression models. In this chapter, we summarise our main contributions and findings of Chapters 3, 4 and 5 and we discuss possible directions for future research.

In Chapter 3, we adopted the Wild Binary Segmentation method (WBS) proposed by Fryzlewicz (2014) in order to detect the number and locations of the change-points in the second-order structure of a time series. Thanks to a certain randomised mechanism, WBS works in cases where the spacings between change-points are short, unlike the standard Binary Segmentation. In addition, we developed a method to combine the change-points detected at different scales of the wavelet periodogram, our main change-point detection statistic. We tested our algorithm on a series of stationary and non-stationary time series models for both small and large samples. The results indicate the good performance of the WBS method. We also applied our method to two real data sets: the US Gross National Product where we detected peaks and troughs in the growth of the US economy; and the infant electrocardiogram

data where we identified the sleep states.

In Chapter 4, we focused on the estimation of the piecewise constant structure of a signal+noise model using the fused lasso method of Tibshirani et al. (2005), a total variation penalty regression. In particular, we showed a fast way of implementing the solution path algorithm of Tibshirani and Taylor (2011). This was achieved by replacing the matrix multiplications with simple CUSUM-type statistics. Based on this observation we were also able to make a connection between the taut string algorithm of Davies and Kovac (2001) and its "multiscale" version of Cho and Fryzlewicz (2011). In addition, by considering a piecewise constant model with a single change-point we proved a result about the consistency of the fused lasso estimator. The main output of this result is that the detection of the exact location of a change-point is not feasible. We supported this claim through a simulation study for different scenarios.

In Chapter 5, we proposed a path algorithm based on Tibshirani and Taylor (2011) and the fused lasso of Tibshirani et al. (2005), termed FuLTV, to estimate regression models where the coefficients are piecewise constant functions of an index variable such as time. Thanks to the adaptability of the fused lasso penalty, our proposed method is capable of estimating models where the underlying coefficient function is not only piecewise constant, but piecewise linear, quadratic or cubic. In addition, we considered various simulated examples and real data sets and we showed that FuLTV did better than smoothing splines of Hastie and Tibshirani (1993) in most cases. From that perspective, Chapter 5 also serves as a comparative study between L_1 - and L_2 -type of penalised regression in time-varying model estimation.

We conclude with a discussion of a few possible avenues to extend the work presented in this thesis. The WBS method in Chapter 3 can be extended to the

estimation of regression models with change-points. We have considered in Chapter 5 the fused lasso methodology to estimate regression models with change-points. However, given the good performance of the Binary Segmentation method over the fused lasso in the context of non-parametric regression (Chapter 4), it is natural to expect that WBS will do better than the latter in the estimation of piecewise constant regression models. The new method could build upon that of Bai and Perron (2003) with the main change-point detection statistic being the sum of squared residuals. At least one advantage of the WBS method over that of Bai and Perron (2003) will be the lack of a selection process for the minimum segment size (trimming parameter).

Furthermore, at least two directions for further research stand out with regards to Chapter 4 and the fused lasso estimator. The first is to explore trend detection, which has received considerable attention by practitioners and academics in different fields including biological/medical sciences (e.g. Greenland and Longnecker (1992)), geophysics (Baillie and Chung (2002)) and macroeconomics (e.g. Hodrick and Prescott (1997), Singleton (1988)). Even though trend detection with an L_1 penalty has been already documented and tested (Kim et al. (2009a)), a consistency result about the number and locations of the change-points remains, to the best of our knowledge, still unexplored.

The second direction is to recognise the change-point detection as a model selection procedure and examine whether other variable selection methods can do better than lasso. This is due to our findings in Chapter 4 that the (fused) lasso is sub-optimal in detecting the location of the change-points. We can still use the same basis functions and the reasons for doing this are twofold: i. they can be computed quickly, without matrix multiplications, and, other methods can take advantage of

this property and ii. the user is not required to select the number of knots and their placement. An adaptive way of knot selection, but not in the lasso framework, could perform better in the non-parametric regression set-up.

Finally, the FuLTV method in Chapter 5 can be also extended in at least two different directions. First, in Section 5.5 we provided a sketch of the theoretical consistency of FuLTV for the piecewise constant model. For the case where the regression coefficients admit different smoothness levels the simulation studies confirm FuLTV's good performance in comparison with the smoothing splines. Future research should focus on the proof of a consistency result for the FuLTV method for different levels of smoothness. Second, in addition to estimating the regression coefficient functions of a time-varying model (Section 5.3) it is important to consider the problem of selecting the relevant variables among a large set of variables. The variable selection problem in the context of a time-varying model has been shown to improve its forecasting performance, see Wang and Huang (2008), Wang and Xia (2009) and Antoniadis et al. (2013). Therefore, variable selection for the FuLTV method should be considered in future studies.

Bibliography

F. Abramovich, T. Bailey, and T. Sapatinas. Wavelet analysis and its statistical applications. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 49(1):1–29, 2000. 24

J. Antoch and M. Hušková. Detection of structural changes in regression. *Tatra Mt. Math. Publ*, 26:201–215, 2003. 104

A. Antoniadis. Wavelet methods in statistics: Some recent developments and their applications. *Statistics Surveys*, 1:16–55, 2007. 24

A. Antoniadis, I. Gijbels, and S. Lambert-Lacroix. Penalized estimation in additive varying coefficient models using grouped regularization. *Statistical Papers*, pages 1–24, 2013. 176, 184

I. Auger and C. Lawrence. Algorithms for the optimal identification of segment neighborhoods. *Bulletin of mathematical biology*, 51(1):39–54, 1989. 55, 61

R. Averkamp and C Houdré. Wavelet thresholding for non-necessarily gaussian noise: idealism. *The Annals of Statistics*, 31(1):110–151, 2003. 51

J. Bai and P. Perron. Estimating and testing linear models with multiple structural changes. *Econometrica*, pages 47–78, 1998. 140

J. Bai and P. Perron. Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, 18(1):1–22, 2003. 167, 183

R. Baillie and S. Chung. Modeling and forecasting from trend-stationary long memory models with applications to climatology. *International Journal of Forecasting*, 18 (2):215–226, 2002. 183

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009. 46

R. Bellman. On the approximation of curves by line segments using dynamic programming. *Communications of the ACM*, 4(6):284, 1961. 124

R. Bellman and S. Dreyfus. *Applied dynamic programming*, volume 2. Princeton University Press, 1966. 61, 105

I. Berkes, E. Gombay, and L. Horváth. Testing for changes in the covariance structure of linear processes. *Journal of Statistical Planning and Inference*, 139(6):2044–2063, 2009. 62

B. Bernanke. *The great moderation*, volume 20. February, 2004. 92

D. Bertsekas. *Nonlinear programming*. Athena Scientific, 1999. 37, 149

L. Birgé and P. Massart. Gaussian model selection. *Journal of the European Mathematical Society*, 3(3):203–268, 2001. 105

O. Blanchard and J. Simon. The long and large decline in us output volatility. *Brookings papers on economic activity*, 2001(1):135–174, 2001. 92

S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004. 32, 44

J. Braun, R. Braun, and H. Müller. Multiple changepoint fitting via quasilikelihood, with application to dna sequence segmentation. *Biometrika*, 87(2):301–314, 2000. 55

L. Breiman. Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):373–384, 1995. 39

L. Breiman, J. Friedman, RA Olshen, D. Steinberg, and P. Colla. *CART: classification and regression trees*. Wadsworth, Belmont, 1983. 104

P. Brockwell and R. Davis. *Introduction to time series and forecasting*, volume 1. Taylor & Francis, 2002. 21

B. Brodsky and B. Darkhovsky. *Nonparametric methods in change point problems*, volume 243. Springer, 1993. 19, 106, 117

Z. Cai, J. Fan, and R. Li. Efficient estimation and inferences for varying-coefficient models. *Journal of the American Statistical Association*, 95(451):888–902, 2000. 140, 141

J. Chen and A. Gupta. *Parametric Statistical Change Point Analysis: With Applications to Genetics, Medicine, and Finance*. Springer, 2011. 55

R. Chen and R. Tsay. Functional-coeffcient autoregressive models. *Journal of the American Statistical Association*, 88:298–308, 1993. 140

X. Chen, S. Kim, Q. Lin, J. Carbonell, and E. Xing. Graph-structured multi-task regression and an efficient optimization method for general fused lasso. *arXiv preprint arXiv:1005.3579*, 2010. 44, 46

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*, pages 999–1018, 1964. 104

C. Chiang, J. Rice, and C. Wu. Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables. *Journal of the American Statistical Association*, 96(454):605–619, 2001. 141

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. 19, 105, 106, 117, 182

H. Cho and P. Fryzlewicz. Multiscale and multilevel technique for consistent segmentation of nonstationary time series. *Statistica Sinica*, 22(1):207–229, 2012. 57, 63, 69, 76, 78, 79, 80, 82, 83, 96, 97

H. Cho and P. Fryzlewicz. Multiple change-point detection for high-dimensional time series via sparsified binary segmentation. *Preprint*, 2013. 63, 75, 76, 77, 96, 102

T. Clark. Is the great moderation over? an empirical analysis. *Federal Reserve Bank of Kansas City Economic Review*, 94(4):5–42, 2009. 92

W. Cleveland, E. Grosse, and W. Shyu. Local regression models. *Statistical models in S*, pages 309–376, 1991. 139, 141, 172

J. Cooley and J. Tukey. An algorithm for the machine calculation of complex fourier series. *Math. comput*, 19(90):297–301, 1965. 53

G. Criton and O. Scaillet. Time-varying analysis in risk and hedge fund performance:

How forecast ability increases estimated alpha. Technical report, working paper, 2011. 140

R. Dahlhaus. Fitting time series models to nonstationary processes. *The Annals of Statistics*, 25(1):1–37, 1997. 22

I. Daubechies. *Ten lectures on wavelets*, volume 61. SIAM, 1992. 23, 25

P. Davies and A. Kovac. Local extremes, runs, strings and multiresolution. *The Annals of Statistics*, pages 1–48, 2001. 3, 19, 105, 106, 107, 108, 115, 122, 124, 144, 182

S. Davies and D. Bland. Interestingness detection in sports audio broadcasts. In *Machine Learning and Applications (ICMLA), 2010 Ninth International Conference on*, pages 643–648. IEEE, 2010. 60

R. Davis, D. Huang, and Y. Yao. Testing for a change in the parameter values and order of an autoregressive model. *The Annals of Statistics*, pages 282–304, 1995. 61

R. Davis, T. Lee, and G. Rodriguez-Yam. Structural break estimation for nonstationary time series models. *Journal of the American Statistical Association*, 101 (473):223–239, 2006. 61, 62, 83, 105

R. Davis, T. Lee, and G. Rodriguez-Yam. Break detection for a class of nonlinear time series models. *Journal of Time Series Analysis*, 29(5):834–867, 2008. 62

C. de Boor. *A Practical Guide to Splines,*. Springer, 1978. 48, 132, 141, 155

R. DeVore. Nonlinear approximation. *Acta numerica*, 7:51–150, 1998. 53

D. Donoho. Unconditional bases are optimal bases for data compression and for statistical estimation. *Applied and computational harmonic analysis*, 1(1):100–115, 1993. 51

D. Donoho. De-noising by soft-thresholding. *Information Theory, IEEE Transactions on*, 41(3):613–627, 1995. 51

D. Donoho. High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Math Challenges Lecture*, pages 1–32, 2000. 30

D. Donoho and J. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994. 10, 35, 36, 51, 52, 104

D. Donoho, I. Johnstone, G. Kerkyacharian, and D. Picard. Wavelet shrinkage: asymptopia? *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 301–369, 1995. 51

I. A Eckley, G. Nason, and R. Treloar. Locally stationary wavelet fields with application to the modelling and analysis of image texture. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(4):595–616, 2010. 23

B. Efron. How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association: Theory and Methods*, 81(394):461–470, 1986. 165

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004. 32, 38, 124, 165

M. Elad, P. Milanfar, and R. Rubinstein. Analysis versus synthesis in signal priors. *Inverse problems*, 23(3):947, 2007. 161

R. Eubank, C. Huang, Y. Maldonado, N. Wang, S. Wang, and R. Buchanan. Smoothing spline estimation in varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(3):653–667, 2004. 140

J. Fan and T. Huang. Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli*, pages 1031–1057, 2005. 176

J. Fan and W. Zhang. Statistical estimation in varying coefficient models. *The Annals of Statistics*, pages 1491–1518, 1999. 137, 139, 142

J. Fan and W. Zhang. Simultaneous confidence bands and hypothesis testing in varying-coefficient models. *Scandinavian Journal of Statistics*, 27(4):715–731, 2000. 140

J. Fan and W. Zhang. Statistical methods with varying coefficient models. *Statistics and its Interface*, 1(1):179, 2008. 141

J. Fan, Q. Yao, and Z. Cai. Adaptive varying-coefficient linear models. *Journal of the Royal Statistical Society: series B (statistical methodology)*, 65(1):57–80, 2003. 141

W. Fisher. On grouping for maximum homogeneity. *Journal of the American Statistical Association*, 53(284):789–798, 1958. 124

K. Frick, A. Munk, and H. Sieling. Multiscale change point inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(3):495–580, 2014. 105

J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimiza-

tion. *The Annals of Applied Statistics*, 1(2):302–332, 2007. 3, 19, 35, 38, 39, 40, 105, 106, 107, 108, 144, 148, 154

P. Fryzlewicz. Modelling and forecasting financial log-returns as locally stationary wavelet processes. *Journal of Applied Statistics*, 32(5):503, 2005. 23, 29

P. Fryzlewicz. Unbalanced haar technique for nonparametric function estimation. *Journal of the American Statistical Association*, 102(480):1318–1327, 2007. 41, 104, 111, 112, 131

P. Fryzlewicz. Wild binary segmentation for multiple change-point detection. *Preprint*, 2014. 2, 18, 57, 64, 65, 66, 67, 69, 75, 95, 96, 97, 98, 101, 104, 120, 124, 134, 181

P. Fryzlewicz and G. Nason. Haar–fisz estimation of evolutionary wavelet spectra. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(4): 611–634, 2006. 29, 71, 72, 101

P. Fryzlewicz and S. Subba Rao. Multiple-change-point detection for auto-regressive conditional heteroscedastic processes. *Journal of the Royal Statistical Society: series B (statistical methodology)*, 2013. 57, 63

P. Fryzlewicz, T. Sapatinas, and S. Rao. A haar–fisz technique for locally stationary volatility estimation. *Biometrika*, 93(3):687–704, 2006. 29

S. Gey and E. Lebarbier. Using cart to detect multiple change points in the mean for large sample. 2008. 104

E. Gombay. Change detection in autoregressive time series. *Journal of Multivariate Analysis*, 99(3):451–464, 2008. 61

E. Gombay and D. Serban. Monitoring parameter change in time series models. *Journal of Multivariate Analysis*, 100(4):715–725, 2009. 61

P. Green and B. Silverman. *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman & Hall/CRC Press, 1994. 48, 132, 155

S. Greenland and M. Longnecker. Methods for trend estimation from summarized dose-response data, with applications to meta-analysis. *American journal of epidemiology*, 135(11):1301–1309, 1992. 183

J. Groen, G. Kapetanios, and S. Price. Multivariate methods for monitoring structural change. *Journal of Applied Econometrics*, 2011. 77

J. Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica: Journal of the Econometric Society*, pages 357–384, 1989. 92

J. Hamilton. *Time series analysis*, volume 2. Princeton university press, 1994. 21

F. Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393, 1974. 120

Z. Harchaoui and C. Lévy-Leduc. Multiple change-point estimation with a total variation penalty. *Journal of the American Statistical Association*, 105(492), 2010. 55, 106, 124

T. Hastie and R. Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 757–796, 1993. 49, 137, 139, 141, 143, 158, 166, 167, 172, 174, 182

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning; Data Mining, Inference and Prediction*. Springer, New York. Second edition, 2009. 34

T. J Hastie and R. Tibshirani. *Generalized Additive Models*. Number 43. CRC Press, 1990. 132, 141, 167

M. Hebiri and S. van de Geer. The smooth-lasso and other $\ell_1+\ell_2$-penalized methods. *Electronic Journal of Statistics*, 5:1184–1226, 2011. 180

R. Hodrick and E. Prescott. Postwar us business cycles: an empirical investigation. *Journal of Money, credit, and Banking*, pages 1–16, 1997. 183

H. Hoefling. A path algorithm for the fused lasso signal approximator. *Journal of Computational and Graphical Statistics*, 19(4):984–1006, 2010. 10, 35, 40, 41, 106, 108, 131, 132, 144

D. Hoover, J. Rice, C. Wu, and L. Yang. Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, 85(4):809–822, 1998. 137, 139, 141

L. Horváth, Z. Horváth, and M. Husková. Ratio tests for change point detection. *Inst. Math. Stat*, 1:293–304, 2008. 73

J. Huang, C. Wu, and L. Zhou. Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika*, 89(1):111–128, 2002. 137, 141, 171

C. Inclan and G. 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. 57, 62, 63, 78

B. Jackson, J. Scargle, D. Barnes, S. Arabhi, A. Alt, P. Gioumousis, E. Gwin, P. Sangtrakulcharoen, L. Tan, and T. Tsai. An algorithm for optimal partitioning of data on an interval. *Signal Processing Letters, IEEE*, 12(2):105–108, 2005. 62, 105

Z. Kander and S. Zacks. Test procedures for possible changes in parameters of statistical distributions occurring at unknown time points. *The Annals of Mathematical Statistics*, pages 1196–1210, 1966. 104

G. Kauermann and G. Tutz. On model diagnostics using varying coefficient models. *Biometrika*, 86(1):119–128, 1999. 141

S. Kay. *Fundamentals of Statistical signal processing, Volume 2: Detection theory*. Prentice Hall PTR, 1998. 61, 105

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. 55, 57, 62, 105

R. Killick, I. Eckley, and P. Jonathan. A wavelet-based approach for detecting changes in second order structure within nonstationary time series. *Electronic Journal of Statistics*, 7:1167–1183, 2013. 61, 63, 86

S. Kim, S. Cho, and S. Lee. On the cusum test for parameter changes in garch (1, 1) models. *Communications in Statistics-Theory and Methods*, 29(2):445–462, 2000. 63

S. Kim, K. Koh, S. Boyd, and D. Gorinevsky. ℓ_1 trend filtering. *Siam Review*, 51(2): 339–360, 2009a. 42, 50, 106, 144, 155, 183

S. Kim, K. Sohn, and E. Xing. A multivariate regression approach to association analysis of a quantitative trait network. *Bioinformatics*, 25(12):i204–i212, 2009b. 34

K. Knight and W. Fu. Asymptotics for lasso-type estimators. *The Annals of statistics*, pages 1356–1378, 2000. 180

E. Kolaczyk and R. Nowak. Multiscale likelihood analysis and complexity penalized estimation. *The Annals of Statistics*, 32(2):500–527, 2004. 104

A. Kovac and A. Smith. Nonparametric regression on a graph. *Journal of Computational and Graphical Statistics*, 20(2):432–447, 2011. 131

M. Lavielle and E. Moulines. Least-squares estimation of an unknown number of shifts in a time series. *Journal of time series analysis*, 21(1):33–59, 2000. 61, 105

M. Lavielle and G. Teyssiere. Detection of multiple change-points in multivariate time series. *Lithuanian Mathematical Journal*, 46(3):287–306, 2006. 61

M. Lavielle and G. Teyssiere. *Adaptive detection of multiple change-points in asset price volatility*. Springer, 2007. 61

S. Lee and S. Park. The cusum of squares test for scale changes in infinite order moving average processes. *Scandinavian Journal of Statistics*, 28(4):625–644, 2001. 63

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. 63

F. Leisch and E. Dimitriadou. *mlbench: Machine Learning Benchmark Problems*, 2010. R package version 2.1-1. 176

Y. Li and G. Arce. A maximum likelihood approach to least absolute deviation regression. *EURASIP Journal on Applied Signal Processing*, 2004:1762–1769, 2004. 39

X. Lin, M. Pham, and A. Ruszczynski. Alternating linearization for structured regularization problems. *arXiv preprint arXiv:1201.0306*, 2011. 46

J. Liu, L. Yuan, and J. Ye. An efficient algorithm for a class of fused lasso problems. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 323–332. ACM, 2010. 46

S. Mallat. Multiresolution approximations and wavelet orthonormal bases of $l^2(r)$. *Transactions of the American Mathematical Society*, 315(1):69–87, 1989. 25, 53, 106, 111

S. Mallat. *A wavelet tour of signal processing*. Academic press, 1999. 23

C. Mallows. Some comments on c p. *Technometrics*, 15(4):661–675, 1973. 105, 167

E. Mammen and S. van de Geer. Locally adaptive regression splines. *The Annals of Statistics*, 25(1):387–413, 1997. 50, 105, 107, 115, 145, 155, 156, 168

M. McConnell and G. Perez-Quiros. Output fluctuations in the united states: What has changed since the early 1980's? *American Economic Review*, pages 1464–1476, 2000. 140

N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010. 33

N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, pages 246–270, 2009. 118, 163

D. Mercurio and V. Spokoiny. Statistical inference for time-inhomogeneous volatility models. *The Annals of Statistics*, pages 577–602, 2004. 22, 60

G. Nason. *Wavelet methods in statistics with R*. Springer, 2008. 24

G. Nason. A test for second-order stationarity and approximate confidence intervals for localized autocovariances for locally stationary time series. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2013a. 80

G. Nason. *wavethresh: Wavelets statistics and transforms*, 2013b. R package version 4.6.2. 10, 36, 92

G. 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 (Statistical Methodology)*, 62(2):271–292, 2000. 23, 26, 28, 29, 30, 63, 70, 71, 92

M. Neumann and R. Von Sachs. *Wavelet thresholding: beyond the Gaussian iid situation*. Springer, 1995. 29

M. Nunes, M. Knight, and G. Nason. Adaptive lifting for nonparametric regression. *Statistics and Computing*, 16(2):143–159, 2006. 51

A. 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. 58, 66

H. Ombao, J. Raz, R. von Sachs, and B. Malow. Automatic statistical analysis of bivariate nonstationary time series. *Journal of the American Statistical Association*, 96(454):543–560, 2001. 60, 61

D. Percival and A. Walden. Wavelet methods for time series analysis (cambridge series in statistical and probabilistic mathematics). 2000. 24

P. Perron. Dealing with structural breaks. *Palgrave handbook of econometrics*, 1: 278–352, 2006. 61

M. Priestley. Spectral analysis and time series. 1981. 21

Z. Qu and P. Perron. Estimating and testing structural changes in multivariate regressions. *Econometrica*, 75(2):459–502, 2007. 140

G. Rigaill. Pruned dynamic programming for optimal multiple change-point detection. *arXiv preprint arXiv:1004.0887*, 2010. 105

A. Rinaldo. Properties and refinements of the fused lasso. *The Annals of Statistics*, 37(5B):2922–2952, 2009. 118

O. Rioul and M. Vetterli. Wavelets and signal processing. *IEEE signal processing magazine*, 8(LCAV-ARTICLE-1991-005):14–38, 1991. 23

P. Robinson. *Nonparametric estimation of time-varying parameters*. Springer, 1989. 140

C. R Rojas and B. Wahlberg. On change point detection using the fused lasso method. *arXiv preprint arXiv:1401.5408*, 2014. 118

L. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268, 1992. 50

D. Salomon. *Data compression: the complete reference*. Springer, 2004. 23

J. Sanderson, P. Fryzlewicz, and M. Jones. Estimating linear dependence between nonstationary time series using the locally stationary wavelet model. *Biometrika*, 97(2):435–446, 2010. 23

D. Sarkar. *Lattice: Multivariate Data Visualization with R*. Springer, New York, 2008. 173

A. 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. 63

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2): 461–464, 1978. 167

A. Sen and M. Srivastava. On tests for detecting change in mean. *The Annals of Statistics*, 3(1):98–108, 1975. 61

J. Shapiro. Embedded image coding using zerotrees of wavelet coefficients. *Signal Processing, IEEE Transactions on*, 41(12):3445–3462, 1993. 23

R. Shumway and D. Stoffer. *Time series analysis and its applications: with R examples*. Springer, 2011. 89

D. Siegmund. Confidence sets in change-point problems. *International Statistical Review/Revue Internationale de Statistique*, pages 31–48, 1988. 104

D. Siegmund and E. Venkatraman. Using the generalized likelihood ratio statistic for sequential detection of a change-point. *The Annals of Statistics*, pages 255–271, 1995. 104

K. Singleton. Econometric issues in the analysis of equilibrium business cycle models. *Journal of Monetary Economics*, 21(2):361–386, 1988. 183

C. Stărică and C. Granger. Nonstationarities in stock returns. *Review of economics and statistics*, 87(3):503–522, 2005. 60

J. Stock and M. Watson. Evidence on structural instability in macroeconomic time series relations. *Journal of Business & Economic Statistics*, 14(1):11–30, 1996. 140

J. Stock and M. Watson. Has the business cycle changed? evidence and explanations. In *Monetary Policy and Uncertainty: Adapting to a Changing Economy, Federal Reserve Bank of Kansas City Symposium, Jackson Hole, Wyoming*, 2003. 92

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996. 32, 33, 145

R. Tibshirani. Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics*, 42(1):285–323, 2014. 20, 50, 106, 115, 131, 142, 144, 155, 160, 161, 165

R. Tibshirani and J. Taylor. The solution path of the generalized lasso. *The Annals of Statistics*, 39(3):1335–1371, 2011. 3, 19, 35, 41, 42, 51, 53, 105, 106, 107, 108, 110, 122, 137, 142, 182

R. Tibshirani and J. Taylor. Degrees of freedom in lasso problems. *The Annals of Statistics*, 40(2):1198–1232, 2012. 165, 166

R. Tibshirani and P. Wang. Spatial smoothing and hot spot detection for cgh data using the fused lasso. *Biostatistics*, 9(1):18–29, 2008. 34

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005. 3, 19, 34, 50, 107, 142, 182

P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109(3):475–494, 2001. 39

E. Venkatraman. *Consistency results in multiple change-point problems*. PhD thesis, Department of Statistics. Stanford University, 1992. 57, 65, 96, 104

E. Venkatraman and A. Olshen. A faster circular binary segmentation algorithm for the analysis of array cgh data. *Bioinformatics*, 23(6):657–663, 2007. 59

B. Vidakovic. *Statistical modeling by wavelets*, volume 503. John Wiley & Sons, 2009. 24, 26

L. Vostrikova. Detecting disorder in multidimensional random processes. In *Soviet Mathematics Doklady*, volume 24, pages 55–59, 1981. 56, 65, 104

G. Wahba. *Spline models for observational data*, volume 59. Siam, 1990. 48, 132, 155

H. Wang and Y. Xia. Shrinkage estimation of the varying coefficient model. *Journal of the American Statistical Association*, 104(486), 2009. 176, 184

L. Wang and J. Huang. Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *Journal of American Statistical Association*, 103(484):1556–1569, 2008. 184

L. Wang, Y. You, and H. Lian. A simple and efficient algorithm for fused lasso signal approximator with convex loss function. *Computational Statistics*, 28(4): 1699–1714, 2013. 46

S. Wang, B. Nan, S. Rosset, and J. Zhu. Random lasso. *The Annals of Applied Statistics*, 5(1):468, 2011. 33

S. Wood. *Generalized additive models: an introduction with R*. CRC press, 2006. 169

S. Wood. *mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML smoothness estimation*, 2014. R package version 1.7-27. 169

K. Worsley. Confidence regions and tests for a change-point in a sequence of exponential family random variables. *Biometrika*, 73(1):91–104, 1986. 104

C. Wu, C. Chiang, and D. Hoover. Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. *Journal of the American statistical Association*, 93(444):1388–1402, 1998. 141

Y. Yao. Estimating the number of change-points via schwarz'criterion. *Statistics & Probability Letters*, 6(3):181–189, 1988. 55, 61, 105

Y. Yao and S. Au. Least-squares estimation of a step function. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 370–381, 1989. 61, 105

G. Ye and X. Xie. Split bregman method for large scale fused lasso. *Computational Statistics & Data Analysis*, 55(4):1552–1569, 2011. 46

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006. 39

C. Zhang. Calibrating the degrees of freedom for automatic data smoothing and effective curve checking. *Journal of the American Statistical Association*, 98(463): 609–628, 2003. 167

P. Zhao and B. Yu. On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006. 162

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006. 33

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67(12):301–320, 2005. 33, 39, 158