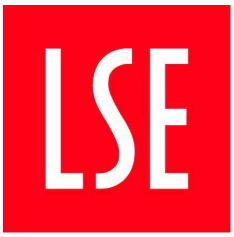
Sparse modelling and estimation for nonstationary time series and high-dimensional data



Haeran Cho

Department of Statistics

London School of Economics

A thesis submitted for the degree of

Doctor of Philosophy

September 2010

I declare that the thesis hereby submitted for the PhD degree of the London School of Economics and Political Science is my own work and that it has not previously been submitted for any other degree. Where other sources of information have been used, they have been acknowledged.

Haeran Cho

To my loving family and friends.

Acknowledgements

First of all, I would like to thank my supervisor, Dr. Piotr Fryzlewicz, for his immense support and extremely valuable guidance throughout my PhD study. I would also like to show my deep gratitude to all the staff and students in the Department of Mathematics at the University of Bristol and in the Department of Statistics at the London school of Economics, for making the last three years one of the most interesting experiences of my life. Besides, my research would not have been possible without the sponsorships from the two universities and the Higher Education Funding Council for England.

Further thanks to my loving friends and relatives here in the UK, back in Korea, and elsewhere in the world to whom I am indebted for their constant encouragement. Finally, I wish to thank my dear parents and brother for their boundless love and support.

Abstract

Sparse modelling has attracted great attention as an efficient way of handling statistical problems in high dimensions. This thesis considers sparse modelling and estimation in a selection of problems such as breakpoint detection in nonstationary time series, nonparametric regression using piecewise constant functions and variable selection in high-dimensional linear regression.

We first propose a method for detecting breakpoints in the second-order structure of piecewise stationary time series, assuming that those structural breakpoints are sufficiently scattered over time. Our choice of time series model is the locally stationary wavelet process (Nason *et al.*, 2000), under which the entire second-order structure of a time series is described by wavelet-based local periodogram sequences. As the initial stage of breakpoint detection, we apply a binary segmentation procedure to wavelet periodogram sequences at each scale separately, which is followed by within-scale and across-scales post-processing steps. We show that the combined methodology achieves consistent estimation of the breakpoints in terms of their total number and locations, and investigate its practical performance using both simulated and real data.

Next, we study the problem of nonparametric regression by means of piecewise constant functions, which are known to be flexible in approximating a wide range of function spaces. Among many approaches developed for this purpose, we focus on comparing two well-performing techniques, the taut string (Davies & Kovac, 2001) and the Unbalanced Haar (Fryzlewicz, 2007) methods. While the multiscale nature of the latter is easily observed, it is not so obvious that the former

can also be interpreted as multiscale. We provide a unified, multiscale representation for both methods, which offers an insight into the relationship between them as well as suggesting some lessons that both methods can learn from each other.

Lastly, one of the most widely-studied applications of sparse modelling and estimation is considered, variable selection in high-dimensional linear regression. High dimensionality of the data brings in many complications including (possibly spurious) non-negligible correlations among the variables, which may result in marginal correlation being unreliable as a measure of association between the variables and the response. We propose a new way of measuring the contribution of each variable to the response, which adaptively takes into account high correlations among the variables. A key ingredient of the proposed *tilting* procedure is hard-thresholding sample correlation of the design matrix, which enables a data-driven switch between the use of marginal correlation and *tilted correlation* for each variable. We study the conditions under which this measure can discriminate between relevant and irrelevant variables, and thus be used as a tool for variable selection. In order to exploit these theoretical properties of tilted correlation, we construct an iterative variable screening algorithm and examine its practical performance in a comparative simulation study.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

One of the most challenging problems in modern statistics is to effectively analyse complex and possibly high-dimensional data. Sparse modelling has often been found attractive when it is believed that there exists a sparse structure which can well-describe the data. For example, sparse modelling is widely adopted in high-dimensional linear regression, where substantial progress has been made over the last few decades under the assumption that only a small number of variables have significant contribution to the response.

This thesis is divided into three parts where different statistical problems are discussed under the common theme of sparse modelling and estimation, which are: breakpoint detection in piecewise stationary time series, nonparametric regression using piecewise constant estimators and variable selection in high-dimensional linear regression. In Chapter 2, we first review the literature in the relevant areas, as well as the basic of wavelet theory which is frequently used throughout this thesis. The rest of the thesis is organised as follows.

Chapter 3. Multiscale and multilevel technique for consistent breakpoint detection in piecewise stationary time series

Being one of the simplest forms of departure from stationarity, piecewise stationary modelling can be useful for analysing a wide class of time series, where a time series is assumed to be (approximately) stationary between two adjacent breakpoints in its dependence structure. A commonly adopted assumption in the relevant literature is that those structural breakpoints

are sufficiently scattered over time and thus sparse in the time domain. Therefore classifying this problem as an application of sparse modelling and estimation, we propose a breakpoint detection method for a class of piecewise stationary, linear processes, which is a combined procedure of a binary segmentation algorithm and post-processing steps. We show that the breakpoints detected by our methodology are consistent estimates of the breakpoints in the second-order structure of the time series, in terms of their total number and locations, and apply the breakpoint detection method to simulated data as well as Dow Jones Industrial Average index to see its practical performance.

Chapter 4. Multiscale interpretation of piecewise constant estimators: taut string and Unbalanced Haar techniques

In nonparametric regression, piecewise constant estimators are favoured for their flexibility in approximating a wide range of function spaces. Chapter 4 compares two piecewise constant estimators, the taut string (see e.g. Davies & Kovac (2001)) and the Unbalanced Haar (Fryzlewicz, 2007) techniques, both of which show good performance in numerical experiments as well as achieving theoretical consistency. We present a unified, multiscale representation for both methods, which offers an insight into the links between them and provides avenues for further improving the two techniques.

Chapter 5. High-dimensional variable selection via tilting

In high-dimensional linear regression problems, variable selection can improve estimation accuracy and model interpretability when it is assumed that only a small number of variables actually contribute to the response. With growing dimensionality of data, the problem of correctly identifying the relevant variables becomes more challenging, one of the complications being the presence of (possibly spurious) non-negligible correlations among the variables. In Chapter 5, a procedure termed *tilting* is proposed in order to measure the association between each variable and the response in a way that adaptively takes into account high correlations among the variables. We study the conditions under which the *tilted correlations* of the relevant variables dominate those of the irrelevant variables, and construct an itera-

tive algorithm based on this new measure, whose performance is compared with other competitors in a simulation study.

In Fan & Lv (2010), the term "high" dimensionality was used to refer to the general case where the dimensionality, or the complexity of the data, grew with the sample size, and "ultra-high" to refer to the case where the dimensionality increased at a non-polynomial rate. Therefore the first two problems can be classified as high-dimensional problems, the dimensionality of the data being equal to the number of observations in both problems, whereas the third problem can include ultra-high dimensional cases in this thesis.

We note that the problems discussed in Chapters 3–5 are distinct from each other in several aspects. For example, in Chapter 4, although a breakpoint in a piecewise constant estimate can indicate where the mean of the data changes significantly, the underlying function may not be piecewise constant itself; on the other hand, the time series model used in Chapter 3 has piecewise constant components in its decomposition, whose breakpoints correspond to the breakpoints in the second-order structure of the time series. Also, the target data for a breakpoint detection method or a piecewise constant estimator have natural (temporal) ordering and thus structured differently from the data used in Chapter 5. This difference is reflected in the model assumptions made in Chapter 3 and Chapter 5. In the former, the structural breakpoints are assumed to be both sparse in the time domain and of sufficient distance from each other, while in the latter, the parameter vector is assumed only to be sparse in terms of the number of non-zero coefficients.

However, we can also draw connections between these different statistical problems under the overall theme of this thesis, sparsity. Being located between the two other chapters, Chapter 4 contains our attempt at establishing some links between breakpoint detection and high-dimensional variable selection problems, using the piecewise constant estimators discussed in that chapter as a "bridge".

Chapter 2

Literature review

In this chapter, we provide a review of the literature on the sparse modelling and estimation problems covered in this thesis, which include breakpoint detection in nonstationary time series, nonparametric regression using piecewise constant estimators and high-dimensional variable selection.

We begin with an overview of the wavelet theory, which has been applied to a broad range of statistical analysis. Wavelets are frequently employed throughout this thesis in different contexts.

2.1 Wavelets

A wavelet function is a wave-like oscillation whose compact support sets it apart from the big waves such as sine and cosine functions. An excellent overview of wavelet theory and its application can be found in Vidakovic (1999). In this section, we provide a brief introduction to wavelets which is vital in expanding the discussion of this thesis, including the multiscale nature of wavelets, discrete wavelet transform and non-decimated wavelets.

We first present some properties of wavelets in connection with continuous wavelet decomposition. A *mother wavelet* ψ is defined as any function in $\mathbb{L}_2(\mathbb{R})$, the space of all square-integrable functions, which satisfies the following *admis-*

sibility condition,

$$C_\psi = \int_{\mathbb{R}} \frac{|\hat{\psi}(\omega)|^2}{|\omega|} d\omega < \infty, \tag{2.1}$$

where $\hat{\psi}(\omega)$ is the Fourier transform of $\psi(x)$. From the admissibility condition, we can derive that

$$\int \psi(x)dx = \hat{\psi}(0) = 0. \tag{2.2}$$

From ψ, a family of functions $\psi_{a,b}$ are generated as translated and dilated versions of the mother wavelet ψ for $a \in \mathbb{R} \setminus \{0\}$ and $b \in \mathbb{R}$, i.e.

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

Example 1.2.2 given in Vidakovic (1999) notes that classical orthonormal bases, such as Fourier basis for $\mathbb{L}_2(\mathbb{R})$, are non-local, since many basis functions have substantial contributions at any value of a decomposition. The properties of ψ as noted in (2.1) and (2.2) indicate that the bases generated from a wavelet function can be localised both in frequency and time by their construction, and such localisation in time can be made arbitrarily fine when an appropriate dilation parameter a is chosen.

For any function $f \in \mathbb{L}_2(\mathbb{R})$, the continuous wavelet transform (CWT) is defined as a function of two variables a and b,

$$\text{CWT}_f(a,b) = \langle f, \psi_{a,b} \rangle = \int f(x)\overline{\psi_{a,b}(x)}dx,$$

and under the admissibility condition, the original function f is recovered via the following inverse transform,

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

The CWT of a function of one variable is a function of two variables, which implies that the CWT is redundant. This redundancy in transform can be reduced

by selecting discrete values of a and b. The following *critical sampling*

$$a = 2^{-i},\ b = k2^{-i};\ \psi_{i,k} = 2^{i/2}\psi(2^i x - k) \text{ for } i, k \in \mathbb{Z},$$

produces the minimal basis in the sense that, it preserves all the information about the decomposed function and any coarser sampling does not give a unique inverse transform. A generalisation of the above sampling can be obtained as

$$a = a_0^{-i},\ b = kb_0a_0^{-i}; i, k \in \mathbb{Z},\ a_0 > 1,\ b_0 > 0.$$

Indices i and k are commonly referred to as "scale" and "location" parameters, respectively. Large values of the scale parameter i denote finer scales where the wavelet functions are more localised and oscillatory. On the other hand, small values of i denote coarser scales with less oscillatory wavelet functions. A theoretical framework for the critically sampled wavelet transform was developed in Mallat *et al.* (1989) and Mallat (1989), which is known as the Mallat's multiresolution analysis, and we describe it in the next section.

2.1.1 Multiresolution analysis

A multiresolution analysis is a sequence of closed subspaces $\{V_i\}_{i\in\mathbb{Z}}$ in $\mathbb{L}_2(\mathbb{R})$ satisfying the follows conditions.

(i) There exists a *scaling function* $\phi \in V_0$ whose integer translations $\{\phi(x - k)\}_{k\in\mathbb{Z}}$ form an orthonormal basis of V_0.

(ii) Spaces $\{V_i\}_{i\in\mathbb{Z}}$ lie in a containment hierarchy as

$$\cdots \subset V_{-2} \subset V_{-1} \subset V_0 \subset V_1 \subset V_2 \subset \cdots. \tag{2.3}$$

(iii) Spaces are self-similar in the sense that $f(2^i x) \in V_i \iff f(x) \in V_0$.

(iv) $\cap_i V_i = \{\mathbf{0}\}$ and $\overline{\cup_i V_i} = \mathbb{L}_2(\mathbb{R})$.

From (i) and (iii), the set $\{\sqrt{2}\phi(2x-k)\}_{k\in\mathbb{Z}}$ is an orthonormal basis for V_1. Then, since $V_0 \subset V_1$, the function ϕ also belongs to V_1 with the following representation

$$\phi(x) = \sqrt{2}\sum_k h_k \phi(2x-k) \tag{2.4}$$

for some coefficients $h_k,\ k \in \mathbb{Z}$. We refer to this (possibly infinite) vector $\underline{\mathrm{h}} = \{h_k\}_{k\in\mathbb{Z}}$ as a *wavelet filter*.

When there is a sequences of subspaces of $\mathbb{L}_2(\mathbb{R})$ satisfying (i)–(iv) with the scaling function ϕ, there exists an orthonormal basis for $\mathbb{L}_2(\mathbb{R})$ in the following form

$$\{\psi_{i,k}(x) = 2^{i/2}\psi(2^i x - k) :\ i,k \in \mathbb{Z}\},$$

such that each $\{\psi_{i,k}(x) :\ k \in \mathbb{Z}\}$ for a fixed i is an orthonormal basis of W_i, which is defined as the orthogonal complement space of V_i in V_{i+1}. We denote this relationship between the function spaces by $V_{i+1} = V_i \oplus W_i$. Then we have

$$V_{i+1} = V_i \oplus W_i = V_{i-1} \oplus W_{i-1} \oplus W_i = \cdots = V_0 \oplus \bigoplus_{j=0}^{i} W_j,$$

and taking $i \to \infty$,

$$\mathbb{L}_2(\mathbb{R}) = V_{i_0} \oplus \bigoplus_{j=i_0}^{\infty} W_j$$

for any $i_0 \in \mathbb{Z}$.

The function $\psi = \psi_{0,0}$ is called a wavelet function or the mother wavelet, and since $\psi(x) \in V_1$, the following representation is satisfied for some coefficients $\{g_k\}_{k\in\mathbb{Z}}$,

$$\psi(x) = \sqrt{2}\sum_k g_k \phi(2x-k). \tag{2.5}$$

Derivation of the mother wavelet ψ from the scaling function ϕ was discussed in Section 3.3.1 of Vidakovic (1999), where $\{g_k\}_{k\in\mathbb{Z}}$ and $\{h_k\}_{k\in\mathbb{Z}}$ were shown to be related as $g_k = (-1)^k h_{1-k}$.

Section 3.4 of the same monograph contains some examples of important families of wavelets. By way of example, we introduce Haar wavelets, whose

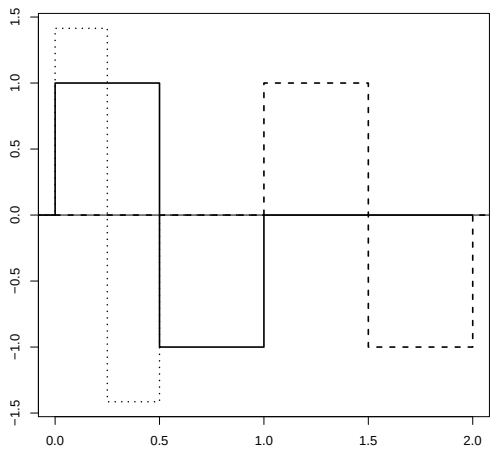


Figure 2.1: Examples of Haar wavelets: ψ (solid), $\psi_{0,1}$ (dashed) and $\psi_{1,0}$ (dotted).

scaling function is of the following form

$$\phi(x) = \mathbb{I}(0 \le x < 1) = \begin{cases} 1 & \text{if } 0 \le x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Examining this scaling function, we have

$$\phi(x) = \phi(2x) + \phi(2x-1) = \frac{1}{\sqrt{2}} \cdot \sqrt{2}\phi(2x) + \frac{1}{\sqrt{2}} \cdot \sqrt{2}\phi(2x-1),$$

and thus the filter coefficients in (2.4) are derived as $h_0 = h_1 = 1/\sqrt{2}$. Then the corresponding $g_0 = -g_1 = 1/\sqrt{2}$ and thus the wavelet function of Haar wavelets satisfies

$$\psi(x) = \phi(2x) - \phi(2x-1) = \begin{cases} 1 & \text{if } 0 \le x < \frac{1}{2}, \\ -1 & \text{if } \frac{1}{2} < x \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Figure 2.1 shows the Haar wavelet function ψ, its shifted version $\psi_{0,1}$ and its rescaled version $\psi_{1,0}$.

2.1.2 Discrete wavelet transform

A *discrete wavelet transform* (Mallat, 1989; Mallat *et al.*, 1989, DWT) is a wavelet algorithm for fast decomposition and reconstruction of discrete datasets, which is analogous to the fast Fourier transform (FFT). Wavelet transforms are linear and can be defined using $n \times n$-orthonormal matrices for the input data of size n. The DWT avoids the matrix representation by exploiting the nested structure of the multiresolution analysis, and thus saves time and memory.

Recalling the definition of function spaces V_i and W_i in Section 2.1.1, any function $f \in V_i$ has a unique representation as $f(x) = v(x)+w(x)$, where $v \in V_{i-1}$ and $w \in W_{i-1}$. Thus f can be decomposed as

$$\begin{aligned} f(x) &= \sum_k c_{i,k}\phi_{i,k}(x) \\ &= \sum_l c_{i-1,l}\phi_{i-1,l}(x) + \sum_l d_{i-1,l}\psi_{i-1,l}(x) \\ &= v(x) + w(x). \end{aligned} \tag{2.6}$$

Note that from (2.4) and (2.5), we have

$$\phi_{i-1,l}(x) = 2^{i/2}\sum_k h_k\phi(2^i x - 2l - k) = \sum_k h_{k-2l}\phi_{i,k}(x), \tag{2.7}$$

$$\psi_{i-1,l}(x) = 2^{i/2}\sum_k g_k\phi(2^i x - 2l - k) = \sum_k g_{k-2l}\phi_{i,k}(x). \tag{2.8}$$

Since V_i and W_i are orthogonal, applying the above results to (2.6), we obtain

$$\begin{aligned} c_{i-1,l} &= \langle f, \phi_{i-1,l}\rangle = \langle f, \sum_k h_{k-2l}\phi_{i,k}(x)\rangle \\ &= \sum_k h_{k-2l}\langle f, \phi_{i,k}(x)\rangle = \sum_k h_{k-2l}c_{i,k} \end{aligned} \tag{2.9}$$

and similarly

$$d_{i-1,l} = \sum_k g_{k-2l}c_{i,k}. \tag{2.10}$$

Therefore coefficients $\{c_{i-1,k}\}$, $\{d_{i-1,k}\}$ can be computed using the coefficients

from the next finer scale, $\{c_{i,k}\}$. In the reverse direction, a single step in the reconstruction algorithm can be written as

$$\begin{aligned} c_{i,k} &= \langle f, \phi_{i,k}\rangle = \langle v, \phi_{i,k}\rangle + \langle w, \phi_{i,k}\rangle \\ &= \sum_l c_{i-1,l}\langle \phi_{i-1,l}, \phi_{i,k}\rangle + \sum_l d_{i-1,l}\langle \phi_{i-1,l}, \psi_{i,k}\rangle \\ &= \sum_l c_{i-1,l} h_{k-2l} + \sum_l d_{i-1,l} g_{k-2l}. \end{aligned}$$

In summary, we only need $O(n)$ operations to perform the DWT for a finite sequence of length n. Denoting the space of square-summable sequences by $l_2(\mathbb{Z})$, let $\underline{\mathrm{f}} = \{f_k\}_{k=0}^{2^I-1}$ be an input sequence of length 2^I in $l_2(\mathbb{Z})$. Then viewing $\underline{\mathrm{f}}$ as the vector of scaling coefficients of a function f, i.e. $f_k = c_{I,k} = \langle f, \phi_{I,k}\rangle$, the DWT of $\underline{\mathrm{f}}$ is obtained by using (2.9) and (2.10),

$$\mathrm{DWT}(\underline{\mathrm{f}}) = (c_{0,0}, d_{0,0}, d_{1,0}, d_{1,1}, d_{2,0}, \ldots, d_{2,3}, \ldots, d_{I-1,0}, \ldots, d_{I-1,2^{I-1}-1}). \quad (2.11)$$

Roughly speaking, the wavelet coefficients $d_{i,k}$ capture the local behaviour of $\underline{\mathrm{f}}$ at scale i and location $2^{I-i}k$, while $c_{0,0}$ captures its overall average behaviour.

The DWT of $\underline{\mathrm{f}}$ in (2.11) can also be represented using the *decimation* and *convolution* operators, which are defined as below.

- The decimation operator $[\downarrow 2]$ is a mapping from $l_2(\mathbb{Z})$ to $l_2(2\mathbb{Z})$ as

  $$([\downarrow 2]\underline{\mathrm{f}})_k = \sum_l f_l \mathbb{I}(l - 2k) = f_{2k},$$

  where $\mathbb{I}(x)$ is an indicator function satisfying $\mathbb{I}(x) = 0$ except for $\mathbb{I}(0) = 1$.

- The convolution operator $\mathbf{H}$ with respect to the filter $\underline{\mathrm{h}} = \{h_k\}_{k\in\mathbb{Z}}$ is defined as

  $$\mathbf{H} : l_2(\mathbb{Z}) \to l_2(\mathbb{Z}),\ (\mathbf{H}\underline{\mathrm{f}})_k = \sum_l h_{l-k} f_l,$$

  and $\mathbf{G}$ is similarly defined with respect to $\underline{\mathrm{g}} = \{g_k\}_{k\in\mathbb{Z}}$.

We further define the operators $\mathcal{H} = [\downarrow 2]\mathbf{H}$ and $\mathcal{G} = [\downarrow 2]\mathbf{G}$. Then by applying $\mathcal{H}$

to $\underline{c}_I = \{c_{I,k}\}_{k=0}^{2^I-1}$, we move to the next coarser scale "approximation", $\underline{c}_{I-1} = \mathcal{H}\underline{c}_I$ where $\underline{c}_{I-1}$ is of length 2^{I-1}. The "detail" information lost by this approximation is captured by $\underline{d}_{I-1} = \mathcal{G}\underline{c}_I$, which is again of length 2^{I-1}. By repeatedly applying these two operators, we obtain another representation of the DWT as follows.

$$\begin{aligned} \mathrm{DWT}(\underline{f}) &= (\underline{c}_0, \underline{d}_0, \underline{d}_1, \ldots, \underline{d}_{I-2}, \underline{d}_{I-1}) \\ &= (\mathcal{H}^I\underline{f}, \mathcal{G}\mathcal{H}^{I-1}\underline{f}, \mathcal{G}\mathcal{H}^{I-2}\underline{f}, \ldots, \mathcal{G}\mathcal{H}\underline{f}, \mathcal{G}\underline{f}). \end{aligned}$$

2.1.3 Non-decimated wavelet transform

In the *non-decimated wavelet transform* (NDWT), or the *stationary wavelet transform*, wavelet coefficients are not decimated as in the DWT. Nason & Silverman (1995) provided a detailed description of the NDWT and its potential applications in nonparametric regression.

One limitation of the DWT is that it is not translation-invariant in the following sense: the wavelet coefficients of $\underline{f}_\tau = \{f_{k-\tau}\}_{k\in\mathbb{Z}}$ are generally not the delayed versions of $\mathrm{DWT}(\underline{f})$. Due to the decimation operator $[\downarrow 2]$ which takes the elements of even indices only ($([\downarrow 2]\underline{f})_k = f_{2k}$), information about the input data used by the DWT is restricted at dyadic locations.

Note that, by defining the *shifting* operator as

$$\mathcal{S} : l_2(\mathbb{Z}) \to l_2(\mathbb{Z}) \text{ for which } (\mathcal{S}\underline{f})_k = f_{k+1},$$

a simple modification of $[\downarrow 2]$ is defined as

$$[\downarrow 2]_1 = [\downarrow 2]\mathcal{S} \text{ such that } ([\downarrow 2]_1\underline{f})_k = f_{2k+1}.$$

The NDWT tackles the limitation of the DWT with a redundant decomposition of $\underline{f}$, which contains the wavelet coefficients obtained from all possible alterations between $[\downarrow 2]$ and $[\downarrow 2]_1$ at every scale.

To have a close look at the NDWT, we need to define the *dilation* operator $[\uparrow 2]$ which alternates an input sequence with zeros, such that

$$([\uparrow 2]\underline{f})_{2k} = f_k \text{ and } ([\uparrow 2]\underline{f})_{2k+1} = 0.$$

Then, an operator defined as $\mathbf{H}^{(r)} = [\uparrow 2]^r \mathbf{H}$ is a convolution operator with respect to the filter $\underline{\mathrm{h}}^{(r)} = \{h_k^{(r)}\}_{k\in\mathbb{Z}}$ satisfying

$$h_{2^r k}^{(r)} = h_k, \text{ and } h_k^{(r)} = 0 \text{ if } k \text{ is not a multiple of } 2^r.$$

By its construction, $\underline{\mathrm{h}}^{(r)}$ is obtained by inserting a zero between every adjacent pair of elements of $\underline{\mathrm{h}}^{(r-1)}$. We similarly define $\mathbf{G}^{(r)} = [\uparrow 2]^r \mathbf{G}$.

Given an input sequence $\underline{\mathrm{f}} = \{f_k\}_{k=0}^{2^I-1} \in l_2(\mathbb{Z})$, let $\underline{\mathrm{a}}_I = \underline{\mathrm{f}}$ and recursively define

$$\underline{\mathrm{a}}_{i-1} = \mathbf{H}^{(I-i)} \underline{\mathrm{a}}_i \text{ and } \underline{\mathrm{b}}_{i-1} = \mathbf{G}^{(I-i)} \underline{\mathrm{a}}_i,$$

for $i = I, I-1, \ldots, 1$. Then the NDWT of $\underline{\mathrm{f}}$ is $\underline{\mathrm{b}}_{I-1}, \underline{\mathrm{b}}_{I-2}, \ldots, \underline{\mathrm{b}}_{I-i_0}, \underline{\mathrm{a}}_{I-i_0}$ for a fixed $i_0 \in \{1, \ldots, I\}$ indicating the depth of transform. Since there is no decimation step in the NDWT, all the subsequent $\underline{\mathrm{a}}_i$ and $\underline{\mathrm{b}}_i$ are of the same length $(= 2^I)$ as the input sequence. Therefore performing the NDWT takes $O(n \log n)$ operations rather than $O(n)$ of the DWT.

2.1.4 Wavelets in this thesis

Vidakovic (1999) discussed a broad range of wavelet applications in statistical problems, such as nonparametric regression, density estimation, time series analysis and deconvolution. Antoniadis (1997) provided a survey of wavelet techniques for nonparametric curve estimation, including both "linear" and "non-linear" methods (see Section 2.4 for the definitions of these two different approaches).

In this thesis, a wavelet-based time series model is adopted as a framework for developing a time series segmentation method in Chapter 3. The chosen model is the locally stationary wavelet model, which was first introduced in Nason *et al.* (2000) and further studied in Van Bellegem & von Sachs (2004) and Fryzlewicz & Nason (2006). We provide a detailed description of the locally stationary wavelet model in Section 2.2.2, and justify this choice as a suitable framework for developing our segmentation procedure in Section 3.1.

Another wavelet application of interest in this thesis is in the context of nonparametric regression. In Section 2.4.1, a non-linear shrinkage method named wavelet thresholding (Donoho & Johnstone, 1994) is described, which automati-

cally adapts to the unknown smoothness of the signal to be estimated. In Chapter 4, the Unbalanced Haar wavelet estimator (Fryzlewicz, 2007) is discussed in details, which combines the wavelet thresholding technique with an adaptive selection of Haar-like wavelet basis.

2.2 Nonstationary time series analysis

For the theoretical treatment of time series procedures, the (weak) stationarity assumption has often been adopted, under which the autocovariance functions are constant over time depending only on the time lag. Although stationarity is a well-studied assumption in time series, it is not necessarily a realistic one when the time series under observation evolves in naturally nonstationary environments. One such example can be found in finance, where return series are considered to have time-varying variance in response to the events taking place in the market. Mikosch & Stărică (1999), Kokoszka & Leipus (2000) and Stărică & Granger (2005), among many others, argued in favour of nonstationary modelling of financial returns. For instance, given the explosion of market volatility during the recent financial crisis, it is unlikely that the same stationary time series model can accurately describe the evolution of market prices before and during the crisis.

Great efforts have been made to relax the assumption of second-order stationarity, and a selective review of *linear* nonstationary time series models is provided in Section 2.2.1 below. As for non-linear processes, Dahlhaus & Subba Rao (2006) generalised the class of autoregressive conditional heteroscedastic (ARCH) processes to include the processes whose parameters were allowed to slowly change over time. Similarly, Polzehl & Spokoiny (2006) introduced a more general class of GARCH models with time varying coefficients, which admitted both abrupt change and smooth transition in the parameters. In this thesis, however, we restrict our attention to linear nonstationary processes only.

Among many nonstationary time series models, Section 2.2.2 is devoted to describing the class of locally stationary wavelet time series (Nason *et al.*, 2000), which is adopted for the development of a breakpoint detection procedure in Chapter 3 of this thesis.

2.2.1 Nonstationary time series models

Spectral analysis has been a fundamental tool in time series analysis, and under the weak stationarity assumption, the frequency domain characteristics of a zero-mean process X_t can be explained by the following Cramér representation (Cramér, 1942),

$$X_t = \int_{-\pi}^{\pi} A(\omega) \exp(i\omega t) dZ(\omega), \ t \in \mathbb{Z}, \tag{2.12}$$

where $A(\omega)$ denotes the amplitude of the process X_t at frequency ω, and $dZ(\omega)$ is an orthonormal increment process satisfying

$$\text{cov}(dZ(\omega), dZ(\omega')) = \begin{cases} d\omega & \text{if } \omega = \omega' \\ 0 & \text{otherwise.} \end{cases}$$

Between the spectrum density of X_t, defined as $f_X(\omega) = |A(\omega)|^2$, and the autocovariance function c_X, there exists the following relationship

$$c_X(\tau) = \int_{-\pi}^{\pi} f_X(\omega) \exp(i\omega\tau) d\omega. \tag{2.13}$$

To relax the stationarity assumption, Priestley (1965) proposed a class of *oscillatory processes* as a modified version of (2.12), where the amplitude function $A(\omega)$ was replaced with a slowly-varying, time-dependent function $A_t(\omega)$. Then, the spectra functions of this oscillatory process had a physical interpretation of being local energy distributions over frequency. However, it is not an easy task to establish rigorous asymptotic theory for oscillatory processes; for the observations $\{X_t\}_{t=1}^T$ from an arbitrary nonstationary process, taking the sample size T to infinity would simply imply the extension of the process to the future, which does not throw any light on the behaviour of the process at the beginning of the time interval.

To tackle this drawback, Dahlhaus (1997) proposed a framework analogous to that of nonparametric regression by regarding the observations as being obtained on a finer grid with increasing T. Then, adopting the notation of a triangular stochastic array $\{X_{t,T}\}_{t=0}^{T-1}$, we can construct asymptotic theory of nonstationary

time series. In Dahlhaus (1997), the class of *locally stationary processes* was defined with its transfer function A^0 and a continuous trend function μ as below;

$$X_{t,T} = \mu\left(\frac{t}{T}\right) + \int_{-\pi}^{\pi} A^0_{t,T}(\omega)\exp(i\omega t)dZ(\omega),\ t = 0,\ldots,T-1;\ T > 0, \quad (2.14)$$

and there exists a 2π-periodic function $A : [0,1] \times \mathbb{R} \to \mathbb{C}$ satisfying

- $A(u,-\omega) = \overline{A(u,\omega)}$,
- $A(u,\omega)$ is continuous in u, and
- for some $C > 0$,

$$\sup_{t,\omega}\left|A^0_{t,T}(\omega) - A\left(\frac{t}{T},\omega\right)\right| \leq \frac{C}{T}.$$

Instead of replacing the transfer function $A(\omega)$ in (2.12) by a smooth function $A(t/T,\omega)$ directly, the above definition requires only that the time-dependent transfer function $A^0_{t,T}(\omega)$ is "close" to $A(t/T,\omega)$. In this manner, the class of locally stationary processes was shown to include autoregressive processes with time-varying AR parameters (Dahlhaus, 1996). Another example of locally stationary processes is a time-modulated process of the following form

$$X_{t,T} = \mu\left(\frac{t}{T}\right) + \alpha\left(\frac{t}{T}\right)Y_t,$$

provided Y_t is stationary and the functions $\mu,\ \alpha : [0,1] \to \mathbb{R}$ are continuous.

Adak (1998) extended the locally stationary process in (2.14) to the class of *piecewise locally stationary processes*. $X_{t,T}$ is piecewise locally stationary if it is locally stationary at all time points $z = t/T \in [0,1]$, except possibly at finitely many breakpoints. Piecewise stationary process (as a concatenation of finite number of stationary processes) belong to the class of piecewise locally stationary processes.

Ombao *et al.* (2002) introduced the Smooth Localised complex EXponential (SLEX) basis vectors, which were simultaneously orthogonal and localised both in time and frequency. Since the SLEX basis vectors overlap, the SLEX transform

can be adopted to smoothly partition the time axis in a dyadic manner and hence to represent discrete random processes whose spectral properties change over time. The partitioning outcome can be used in modelling the data as a *blended stationary process*, which assumes smooth transitions between adjoining stationary blocks rather than abrupt changes as in piecewise stationary processes. A time series segmentation method based on SLEX transform was introduced in their paper, and its brief description can be found in Section 2.3.

2.2.2 Locally stationary wavelet model

Nason *et al.* (2000) defined the class of *locally stationary wavelet (LSW) processes*, which could roughly be described as replacing the harmonic system $\exp(i\omega t)$ in the locally stationary processes (2.14) by a wavelet system.

Before introducing the formal definition of the LSW model, we note that in this section (and also in Chapter 3), $\psi_{i,k}$ is used to denote discrete, non-decimated wavelets rather than wavelet functions as in Section 2.1 of this thesis. That is, a discrete, non-decimated wavelet vector is denoted by

$$\psi_i = (\psi_{i,0}, \psi_{i,1}, \ldots, \psi_{i,\mathcal{L}_i})^T$$

such that e.g. for Haar wavelets, $\psi_{i,k}$ satisfies

$$\psi_{i,k} = 2^{i/2}\mathbb{I}_{\{0,\ldots,2^{-i-1}-1\}}(k) - 2^{i/2}\mathbb{I}_{\{2^{-i-1},\ldots,2^{-i}-1\}}(k) \tag{2.15}$$

($\mathbb{I}_{\mathcal{A}}(k)$ is an indicator function which takes 1 if $k \in \mathcal{A}$ and 0 otherwise), for all $i = -1, -2, \ldots$ and $k \in \mathbb{Z}$.

The data inhabit in scale zero, and small negative values of the scale parameter i denote "finer" scales where the wavelet vectors are more localised and oscillatory, whereas large negative values of i denote "coarser" scales with longer, less oscillatory wavelet vectors. As for $\mathcal{L}_i$, the length of a wavelet vector at scale i, it can be shown that $\mathcal{L}_i = (2^{-i} - 1)(\mathcal{L}_{-1} - 1) + 1$ for all $i < 0$. As seen in (2.15), discrete, non-decimated wavelets can be shifted to any location defined by the finest-scale wavelets, unlike in the DWT where its shifts are restricted to "dyadic" locations (i.e. multiples of 2^{-i} at scale i). Therefore discrete, non-

decimated wavelets are no longer an orthonormal, but an overcomplete collection of shifted vectors.

Now, we present the LSW model as defined in Nason *et al.* (2000).

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

$$X_{t,T} = \sum_{i=-I(T)}^{-1} \sum_{k=-\infty}^{\infty} \omega_{i,k;T} \psi_{i,t-k} \xi_{i,k} \tag{2.16}$$

where $I(T) = -\min\{i : \mathcal{L}_i \leq T\}$. *The parameters* $i \in \{-1, -2, \ldots, -I(T)\}$ *and* $k \in \mathbb{Z}$ *are used to denote the scale and the location respectively,* $\psi_i = (\psi_{i,0}, \ldots, \psi_{i,\mathcal{L}_i})$ *are discrete, real-valued, compactly supported, non-decimated wavelet vectors, and* $\xi_{i,k}$ *are zero-mean, orthonormal, identically distributed random variables. For each* i*, there exists a Lipschitz-continuous function* $W_i : [0, 1] \to \mathbb{R}$ *such that*

- $\sum_{i=-\infty}^{-1} |W_i(z)|^2 < \infty$ *uniformly in* $z \in (0, 1)$,
- *the Lipschitz constants* L_i *are uniformly bounded in* i *as well as satisfying* $\sum_{i=-\infty}^{-1} 2^{-i} L_i < \infty$, *and*
- *there exists a sequence of constants* C_i *satisfying* $\sum_{i=-\infty}^{-1} C_i < \infty$ *and*

$$\sup_{0 \leq k \leq T-1} \left| \omega_{i,k;T} - W_i\left(\frac{k}{T}\right) \right| \leq \frac{C_i}{T}, \tag{2.17}$$

  for each T *and* $i = -1, \ldots, -I(T)$.

As wavelets are parameterised by scale i and location k, the representation in (2.16) is naturally scale- and location-dependent, and the local power in the autocovariance of $X_{t,T}$ is decomposed with respect to scales (instead of frequencies) along time. That is, like the transfer function $A_{t,T}^0$ in (2.14), each $\omega_{i,k;T}^2$ measures the local power (i.e. contribution to the autocovariance) of the time series at scale i and location k. To obtain meaningful estimation results, $\{\omega_{i,k;T}^2\}_{k \in \mathbb{Z}}$ are allowed to evolve slowly by being sufficiently close to regular Lipschitz functions $W_i(k/T)$ as in (2.17).

In the LSW framework, the asymptotic *evolutionary wavelet spectrum* (EWS) is defined on the rescaled unit interval $(0, 1)$ as

$$S_i(z) = W_i^2(z) = \lim_{T\to\infty} \omega^2_{i,\lfloor zT\rfloor;T}, \quad z \in (0, 1),$$

which has an interpretation of being the analogue of the spectrum of stationary processes. By way of example, Nason *et al.* (2000) showed in Proposition 2.17 that any stationary process with absolutely summable covariance is an LSW process, since each EWS $S_i(z)$ does not change over the rescaled time z in such a case.

For stationary time series, the spectral density and the autocovariance function are related as one being the Fourier transform of the other, see (2.13). Such relationship can also be derived in the LSW model framework using the same wavelet system in (2.16). First, define the autocorrelation wavelets $\Psi_i(\tau) = \sum_{k=-\infty}^{\infty} \psi_{i,k}\psi_{i,k+\tau}$, and the autocorrelation wavelet inner product matrix $\mathbf{A} = (A_{i,j})_{i,j<0}$ with its elements

$$A_{i,j} = \sum_{\tau} \Psi_i(\tau)\Psi_j(\tau).$$

Further, let $c_T(z, \tau)$ denote the finite-sample autocovariance function of $X_{t,T}$ at lag τ and rescaled location z, i.e.

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

and let $c(z, \tau)$ denote the asymptotic local autocovariance function, which is defined as a transform of $S_i(z)$ with respect to the set of autocorrelation wavelets, i.e.

$$c(z, \tau) = \sum_{i=-\infty}^{-1} S_i(z)\Psi_i(\tau). \tag{2.18}$$

Then, Proposition 2.11 of Nason *et al.* (2000) showed that, under the assumptions in Definition 2.1, $c_T(z, \tau)$ and $c(z, \tau)$ are close in the following sense,

$$|c_T(z, \tau) - c(z, \tau)| = O(T^{-1}) \tag{2.19}$$

as $T \to \infty$, uniformly in $\tau \in \mathbb{Z}$ and $z \in (0,1)$. Also the representation in (2.18) is invertible as

$$S_i(z) = \sum_{\tau} \left(\sum_{j} \Psi_j(\tau) A^{-1}_{i,j} \right) c(z, \tau),$$

see Proposition 2.14 in Nason *et al.* (2000). In summary, the above results show that there is a one-to-one correspondence between the EWS and the asymptotic local autocovariance function for the LSW time series.

An estimate for the EWS of an LSW process $X_{t,T}$ can be obtained by using the set of its squared wavelet coefficients, which is referred to as the *wavelet periodogram*.

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

$$I^{(i)}_{t,T} = \left| \sum_{s} X_{s,T} \psi_{i,s-t} \right|^2 \tag{2.20}$$

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

We quote the following result from Nason *et al.* (2000).

Proposition 2.1.

$$\mathbb{E} I^{(i)}_{t,T} = \sum_{j=-\infty}^{-1} S_j \left(\frac{t}{T} \right) A_{i,j} + O\left(\frac{2^{-i}}{T} \right). \tag{2.21}$$

If $X_{t,T}$ is Gaussian, then

$$var\left(I^{(i)}_{t,T} \right) = 2 \left\{ \sum_{j=-\infty}^{-1} S_j \left(\frac{t}{T} \right) A_{i,j} \right\}^2 + O\left(\frac{2^{-i}}{T} \right).$$

Defining $\beta_i(z)$ as a linear transform of evolutionary wavelet spectra with respect

to the autocorrelation wavelet inner product matrix

$$\beta_i(z) = \sum_{j=-\infty}^{-1} S_j(z) A_{i,j},$$

Proposition 2.1 implies that the wavelet periodogram $I_{t,T}^{(i)}$ is an inconsistent but asymptotically unbiased estimator of $\beta_i(t/T)$. Therefore, from (2.21), we can derive an estimate of $S_i(z)$ as

$$\hat{S}_i(z) = \sum_{j=-I(T)}^{-1} I_{\lfloor zT \rfloor, T}^{(j)} A_{i,j}^{-1}.$$

For the discussion on the smoothing of wavelet periodograms and the estimates of EWS obtained from the smoothed $I_{t,T}^{(i)}$, see Nason *et al.* (2000). Using these estimates of wavelet spectra, the local autocovariance function $c(z, \tau)$ can also be estimated from (2.18).

In Van Bellegem & von Sachs (2004) and Van Bellegem & von Sachs (2008), a new definition for LSW model was introduced, enlarging the class of LSW processes to contain the processes whose spectral density function may change abruptly over time. It was achieved by replacing the Lipschitz condition in (2.17) by a condition on the total variation of amplitudes. Fryzlewicz & Nason (2006) presented a modified version of the LSW model in Definition 2.1, which assumed the scale-dependent transfer function $W_i(z) : [0, 1] \to \mathbb{R}$ to be piecewise constant with a finite (but unknown) number of jumps, imposing a similar condition on the total variation of $W_i(z)$ as that in Van Bellegem & von Sachs (2004). In Chapter 3, we describe this modified LSW model in Section 3.1, and adopt it as a framework for developing a procedure which detects breakpoints in the second-order structure of nonstationary time series.

As for multivariate time series analysis, Sanderson *et al.* (2010) proposed a new bivariate LSW time series model, based on which they developed a method of wavelet coherence for estimating the dependence between neuroscience data recorded from different brain regions. In Eckley *et al.* (2010), an extended version of the LSW model into two-dimensions was used to model and analyse image texture data.

2.3 Breakpoint detection in nonstationary time series

As noted in Section 2.2, for the processes which evolve in naturally nonstationary environments, nonstationary modelling appears more realistic than its stationary counterpart. Piecewise stationarity is arguably the simplest form of departure from stationarity, and one task when faced with data of this form is to detect breakpoints in the dependence structure. The breakpoint detection problem in nonstationary time series can be divided into two categories, as retrospective (*a posteriori*) breakpoint detection and *on-line* breakpoint detection.

The on-line approach is adopted when the aim of analysis lies in detecting any change while the monitoring of the data is still in progress. A survey of the literature in this area can be found in Lai (2001), and more recent efforts include Hawkins *et al.* (2003), Tartakovsky *et al.* (2006) and Mei (2006). On the other hand, the retrospective approach takes into account the entire set of observations at once and detects breakpoints which occurred in the past. Using the term "segmentation" interchangeably with multiple breakpoint detection, the outcome of *a posteriori* segmentation can be of interest for several purposes; for example, the information from the last (approximately) stationary segment can be useful in forecasting the future. We classify the time series segmentation problem as an application of sparse modelling and estimation, since the breakpoints in the dependence structure of the time series are often assumed to be sufficiently scattered over time and thus sparse in the time domain.

In Section 2.3.1, we review a selection of breakpoint detection methods which were proposed for detecting single or multiple breakpoints, in the dependence structure of either independent or correlated observations. Among many procedures developed for time series segmentation, Section 2.3.2 focuses on the binary segmentation procedure, which is a key ingredient of our breakpoint detection methodology proposed in Chapter 3.

2.3.1 Retrospective breakpoint detection methods

Early breakpoint detection literature was mostly devoted to testing the existence of a single breakpoint in the mean or variance of independent observations (Chernoff & Zacks, 1964; Hawkins, 1977; Hsu, 1977; Sen & Srivastava, 1975; Worsley, 1986).

When the presence of more than one breakpoint is suspected, an algorithm for detecting multiple breakpoints is needed to extend the testing procedures for a single breakpoint. Being a method of solving complex problems by breaking them down into simpler steps, dynamic programming was adopted in the literature when the proposed segmentation procedure looked for the "optimal" segmentation, according to a criterion tailored e.g. in the framework of maximum likelihood estimation (Hawkins, 2001) or reproducing kernel Hilbert space (Harchaoui & Cappe, 2007). One drawback of dynamic programming is that its application involves the difficult choice of the total number of breakpoints. Another method for tackling the multiple breakpoint detection problem is the binary segmentation procedure, whose detailed description can be found in Section 2.3.2.

Various multiple breakpoint detection methods have been proposed for time series of correlated observations. Adak (1998) proposed a segmentation procedure which divided the time series into dyadic blocks using binary trees, and then chose the best segmentation which minimised the discrepancy between estimated spectra within each segment. Ombao *et al.* (2001) adopted a similar approach of performing the data partitioning followed by the best segmentation selection. Their Auto-SLEX procedure used the SLEX transform (see Section 2.2.1) to produce a collection of overlapping dyadic partitions, from which it selected the best segmentation by applying the best basis algorithm (for the details of the best basis algorithm, see e.g. Wickerhauser (1994)).

In Lavielle & Moulines (2000), a method was developed for obtaining the least squares estimates of multiple breakpoints in linear processes with changing mean, extending the work of Bai & Perron (1998) who considered the single breakpoint case. Based on this method, Andreou & Ghysels (2002) studied a heuristic segmentation procedure for the GARCH model with changing parameters. In Lavielle & Teyssière (2005), a breakpoint detection method was de-

veloped for weakly or strongly dependent processes with time-varying volatility, which minimised a penalised contrast function based on a Gaussian likelihood. For solving the optimisation problem, it also used dynamic programming along with an automatic procedure for choosing the final number of breakpoints.

Davis *et al.* (2006) developed the Auto-PARM procedure for segmenting a piecewise stationary AR process, which was defined as a concatenation of stationary AR processes. Based on the idea that the best fitting model for a given time series was the one that enabled the maximum compression of the data, the Auto-PARM procedure was designed to look for a combination of the total number and locations of breakpoints as well as the values of AR parameters, which would minimise a certain criterion developed under the minimum description length (MDL) principle. They adopted a search heuristic termed the genetic algorithm, which mimicked the process of natural evolution for traversing the vast parameter space. This procedure was later extended to the segmentation of non-linear processes in Davis *et al.* (2008).

2.3.2 Binary segmentation

Vostrikova (1981) introduced a binary segmentation procedure, which recursively performed locating and testing for multiple breakpoints to achieve computationally efficient and multilevel breakpoint detection. It was shown that the breakpoint estimates from the binary segmentation were consistent for a class of random processes with piecewise constant means. However, one limitation of the proposed procedure was that the critical value of the test at each iteration was difficult to compute in practice, due to the stochasticity in previously selected breakpoints.

Venkatraman (1993) employed a similar idea to find multiple breakpoints in the mean of independent and normally distributed variables with a test criterion depending only on the length of data sequence. A brief sketch of the proposed binary segmentation procedure is as below.

Let $\{Y_t\}_{t=1}^T$ denote the sequence to be segmented and

$$Y_t = \mu_t + \epsilon_t, \quad t = 1, \ldots, T, \tag{2.22}$$

where μ_t may change over time in a piecewise constant manner, and ϵ_t are independent random noise following $\mathcal{N}(0, \sigma^2)$. The binary segmentation procedure performs a hypothesis testing of $t = b$ being a breakpoint by checking whether $\{Y_t\}_{t=1}^{b}$ and $\{Y_t\}_{t=b+1}^{T}$ have the same distribution or not, which is equivalent to checking whether two segments have the same mean under the model (2.22). To test the null hypothesis

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_n$$

against the alternative hypothesis

$$H_1 : \mu_1 = \cdots = \mu_b \neq \mu_{b+1} = \cdots = \mu_n,$$

the likelihood ratio statistic is of the form

$$\mathbb{Y}_{1,T}^{b} = \sqrt{\frac{T-b}{T \cdot b}} \sum_{t=1}^{b} Y_t - \sqrt{\frac{b}{T \cdot (T-b)}} \sum_{t=b+1}^{T} Y_t. \tag{2.23}$$

Thus the test statistic at the first level of binary segmentation is obtained as

$$\mathbb{Y}_{1,T} = \max_{b=1,\ldots,T} \left| \mathbb{Y}_{1,T}^{b} \right|,$$

and if $\mathbb{Y}_{1,T}$ is greater than a critical value, say C_T, the null hypothesis of no breakpoint is rejected and a breakpoint is estimated as $\hat{b} = \arg\max_b \left| \mathbb{Y}_{1,T}^{b} \right|$. The next step is to divide the sequence into two, to the left and right of the estimated breakpoint $\hat{b}$ (i.e., $\{Y_t\}_{t=1}^{\hat{b}}$ and $\{Y_t\}_{t=\hat{b}+1}^{T}$), and the same searching and testing procedure is performed at the next level, separately within each segment.

Venkatraman (1993) showed that the breakpoints detected by the procedure described above were consistent in terms of their total number and locations with the test criterion $C_T = T^{3/8}$.

The binary segmentation procedure was also used in detecting multiple shifts in the variance of independent observations (Chen & Gupta, 1997; Inclán & Tiao, 1994). Whitcher *et al.* (2000, 2002) and Gabbanini *et al.* (2004) suggested to segment long memory processes by applying the iterative cumulative sum of squares (ICSS) algorithm (originally proposed in Inclán & Tiao (1994)) to discrete wavelet coefficients of time series, which were approximately Gaussian and decorrelated.

However, their approach does not take into account the autocorrelation still remaining in the wavelet coefficient sequence of time series.

Chapter 3 of this thesis addresses the problem of retrospective breakpoint detection in the framework of a piecewise stationary time series model, which is based on the LSW model described in Section 2.2.2. We propose a binary segmentation procedure which permits autocorrelation in the sequence to be segmented, with its test criterion depending on the sample size only and thus being easy to compute. We show that this binary segmentation achieves consistency in identifying the total number and locations of multiple breakpoints in correlated sequences of a multiplicative form, instead of the additive form in (2.22).

2.4 Nonparametric regression

A canonical problem in nonparametric regression is the estimation of a one-dimensional function f from noisy observations y in the following additive model

$$y_t = f\left(\frac{t}{n}\right) + \epsilon_t, \ t = 1, \cdots, n. \tag{2.24}$$

In the simplest version of (2.24), $\{\epsilon_t\}_{t=1}^n$ are assumed to be i.i.d. Gaussian variables satisfying $\mathbb{E}(\epsilon_t) = 0$ and $\text{var}(\epsilon_t) = \sigma^2$. Although it is not necessarily a realistic assumption in some applied problems, it serves as a good benchmark for comparing estimation techniques and judging their potential performance in more complex models. In other words, if a method performs poorly for the model (2.24) with i.i.d. Gaussian noise, there is often little chance of it performing well in more complex settings.

The problem of estimating f in the model in (2.24) is of interest for at least two purposes: an estimate of f provides insights into the relationship between the design variable x_t (in the model (2.24), $x_t = t/n$) and the response variable y_t, and it can also be used for predicting observations which have not been made yet. Many approaches have been proposed to tackle this nonparametric regression problem, and we list a few of them which have been widely studied since their introduction.

Kernel estimation (Nadaraya, 1964; Watson, 1964).

The Nadaraya-Watson kernel estimator is defined as

$$\hat{f}^K(x) = \left\{ \sum_{s=1}^{n} K\left(\frac{x - x_s}{h}\right) \right\}^{-1} \cdot \sum_{t=1}^{n} K\left(\frac{x - x_t}{h}\right) y_t, \tag{2.25}$$

where the kernel K is any smooth function satisfying

$$K(x) \geq 0, \int K(x)dx = 1, \int xK(x)dx = 0 \text{ and } \int x^2 K(x)dx > 0.$$

The bandwidth $h > 0$ determines the amount of smoothing, i.e. the estimates $\hat{f}^K$ using small values of h are "rough", while the estimates gets "smoother" for large values of h. By re-writing (2.25), we can show that the kernel estimator is linear in the observations $\{y_t\}_{t=1}^n$ in the following sense: $\hat{f}^K$ satisfies

$$\hat{f}^K(x) = \sum_{t=1}^{n} l_t(x) y_t \text{ where } l_t(x) = \frac{K\left(\frac{x-x_t}{h}\right)}{\sum_{s=1}^{n} K\left(\frac{x-x_s}{h}\right)}.$$

Spline smoothing (Silverman, 1985; Wegman & Wright, 1983).

The smoothing spline estimator is defined as the minimiser of

$$\sum_{t=1}^{n} \left(y_t - \tilde{f}(x_t)\right) + \lambda \int \tilde{f}''(x)^2 dx, \tag{2.26}$$

over the class of twice differentiable functions $\tilde{f}$. $\lambda \geq 0$ is a smoothing parameter which controls the trade-off between fidelity to the data and smoothness of the estimator $\hat{f}^S$. Due to the quadratic nature of (2.26), $\hat{f}^S$ is also a linear smoother, i.e. there exists a weight function $G(z, x)$ (which depends on the design points x_t, $t = 1, \ldots, n$ and smoothing parameter λ) satisfying

$$\hat{f}^S(z) = \frac{1}{n} \sum_{t=1}^{n} G(z, x_t) y_t.$$

Local polynomials (Cleveland & Devlin, 1988; Fan & Gijbels, 1996).

Let z be some fixed value at which we wish to estimate f and let $p \geq 0$ be a fixed integer. Then local polynomial regression finds $\hat{\mathbf{a}} = (\hat{a}_0, \hat{a}_1, \ldots, \hat{a}_p)^T$ which minimises the following locally weighted sum of squares

$$\sum_{t=1}^{n} w_t(z) \left(y_t - P_z(x_t; a)\right), \text{ where}$$
$$P_z(x; a) = a_0 + a_1(x - z) + \frac{a_2}{2!}(x - z)^2 + \cdots + \frac{a_p}{p!}(x - z)^p,$$

and returns the local estimate $P_z(x; \hat{\mathbf{a}})$. To be precise, the estimator $\hat{\mathbf{a}}$ depends on the value of z as $\hat{\mathbf{a}}(z)$, and therefore at the target value $x = z$, we have the localy polynomial regression estimate $\hat{f}^P(z) = \hat{a}_0(z)$. The special case of $p = 1$ is called local linear regression, and when $p = 0$, local polynomial regression coincides with kernel estimation. As with the kernel estimator, $\hat{f}^P$ also has a relationship similar to that in (2.26) with the observations $\{y_t\}_{t=1}^n$. For the specific expression of l_t for local polynomials, see e.g. Fan & Gijbels (1995).

We note that the above description is intended as a brief taster of the well-known nonparametric regression techniques, and that much work has been done to improve and extend these methods. In their simplest form, however, all three estimators are linear smoothers. When the underlying function f is smooth, linear approximation methods can achieve optimal performance in terms of the mean-square error (MSE) of the estimator $\hat{f}$,

$$\text{MSE}(\hat{f}, f) = \frac{1}{n}\mathbb{E}\|\hat{f} - f\|_2^2.$$

For example, Fan (1993) showed that the minimax risk of local linear regression smoothers was optimal for a class of smooth functions f, attaining both optimal rates of convergence and (nearly) optimal constant factors.

On the other hand, when the underlying function f is irregular (e.g. discontinuous), non-linear approximation can achieve better performance (DeVore, 1998), and in what follows, our focus is on presenting some well-performing non-linear methods. Especially, Section 2.4.1 is devoted to a wavelet smoothing technique

named wavelet thresholding (Donoho & Johnstone, 1995, 1994), and in Section 2.4.2, we review a selection of non-linear approximation methods which produce piecewise constant estimators. We continue the discussion of non-linear, piecewise constant estimators in Chapter 4, where a new, multiscale estimation framework is introduced for the better understanding of some piecewise constant estimators.

2.4.1 Wavelet thresholding estimator

Donoho & Johnstone (1994) introduced a non-linear smoothing method called *wavelet thresholding* technique. The first step of wavelet thresholding is to obtain the DWT of the observations $\{y_t\}_{t=1}^n$ as

$$d_{i,k} = \theta_{i,k} + z_{i,k},$$

where $d_{i,k}$ ($\theta_{i,k}$, $z_{i,k}$) denotes the DWT of y_t ($f(t/n)$, ϵ_t). Then a threshold λ is applied to shrink some wavelet coefficients $d_{i,k}$ towards 0, and the wavelet thresholding estimator of f, say $\hat{f}$, is returned as the inverse DWT of the thresholded wavelet coefficients.

Since the DWT is orthonormal, $z_{i,k}$, the DWT of i.i.d. Gaussian noise variables ϵ_t, are still i.i.d. Gaussian in the wavelet domain. Meanwhile, wavelet transforms tend to concentrate the "energy" in data in the sense that, wavelet coefficients $\theta_{i,k}$ corresponding to where f is smooth are likely to be close to 0, while those corresponding to where f has irregularities are likely to be significantly different from 0 (see Section 6.3.3 of Vidakovic (1999)). From the above observations, it is expected that important features of f are represented sparsely in the wavelet domain, and thus an unknown function f can accurately be recovered by "killing" small $d_{i,k}$ with an appropriately chosen threshold.

Donoho & Johnstone (1994) proposed *hard* and *soft* thresholding rules

$$\begin{aligned} T_{\text{hard}}(d_{i,k}, \lambda) &= d_{i,k} \cdot \mathbb{I}(|d_{i,k}| > \lambda), \\ T_{\text{soft}}(d_{i,k}, \lambda) &= \text{sign}(d_{i,k}) \cdot \max\left(|d_{i,k}| - \lambda, 0\right), \end{aligned}$$

with the use of the *universal* threshold $\lambda_{\text{univ}} = \sigma\sqrt{2\log n}$. The term *VisuShrink* was used to describe the application of thresholding with this universal threshold,

and the resulting estimator $\hat{f}^{VS}$ was shown to have the "oracle" property, i.e. the MSE of VisuShrink estimator $\hat{f}^{VS}$ is close (within a logarithmic factor $\log n$) to the ideal risk that can be achieved when equipped with an oracle telling which wavelet coefficients $d_{i,k}$ should be "killed" and "kept".

The *SureShrink* procedure proposed in Donoho & Johnstone (1995) adaptively selects a threshold for each scale i, by minimising the Stein's unbiased estimator for risk (Stein, 1981, SURE). SureShrink estimator was also shown to be near minimax within the whole range of Besov space (a set of functions in Lebesgue space which have certain smoothness, see DeVore & Popov (1988)), by automatically adapting to the unknown smoothness of f.

More recent approaches to the thresholding technique include the study of threshold selection in the framework of multiple-hypothesis testing (Abramovich & Benjamini, 1995, 1996; Ogden & Parzen, 1996a,b), or cross-validation (Nason, 1995, 1996). Wang (1996) and Johnstone & Silverman (1997) discussed the application of wavelet thresholding technique under the presence of correlated noise. Abramovich *et al.* (1998) considered wavelet thresholding within a Bayesian framework, where a prior distribution was designed to capture the sparseness of wavelet decomposition $\theta_{i,k}$. In Johnstone & Silverman (2004), an adaptive threshold selection procedure termed *empirical Bayesian thresholding* was proposed.

2.4.2 Piecewise constant estimators

DeVore (1998) noted that the class of piecewise constant functions was flexible in approximating a wide range of function spaces. It was further shown in the paper that, when the underlying function f was spatially inhomogeneous, the performance of non-linear, piecewise constant estimators was superior to that of linear methods, as they chose the partition of $[0, 1]$ (on which piecewise constant estimates were taken) in a data-driven way, unlike the linear piecewise constant estimators on fixed partitions. In this section, we present a list of non-linear, piecewise constant approximation methods which have shown good performance.

The wavelet thresholding estimation discussed in Section 2.4.1 returns a piecewise constant estimate when Haar wavelets are used, whose specific form can be

found in Section 2.1.1 of this thesis. The CART methodology (Breiman *et al.*, 1983, Classification and Regression Trees) performs greedy binary splitting to grow a partition, whose terminal nodes yield a piecewise constant estimator. In Engel (1997), a method for locally adaptive histogram construction was introduced, which was based on a tree of dyadic partitions and hence obtained a multiscale, piecewise constant estimator. The adaptive weight smoothing (Polzehl & Spokoiny, 2000) produces a piecewise constant estimator using an iterative local averaging procedure with an adaptive choice of weights. Comte & Rozenholc (2004) and Kolaczyk & Nowak (2005) proposed to estimate an unknown function using piecewise polynomials by optimising a complexity-penalised likelihood, and their approaches can be adopted to obtain piecewise constant estimators.

In Chapter 4, our interest lies in comparing two non-linear methods for producing piecewise constant estimates, the taut string (Barlow *et al.*, 1972; Davies & Kovac, 2001) and the Unbalanced Haar (Fryzlewicz, 2007) techniques, both of which are computationally fast, achieve theoretical consistency, and exhibit excellent performance in numerical experiments. The former is a penalised least squares estimator with its penalty imposed on the total variation of unknown function, whereas the latter involves wavelet thresholding (Section 2.4.1) with respect to an orthonormal, Haar-like basis vectors, whose breakpoints are no longer constrained to be in the middle of their support as in Haar wavelets. We propose a new multiscale framework, which both methods are instances of, and it is this new framework that provides better insight into the two techniques as well as some directions for future research.

2.5 High-dimensional linear regression

One of the most important and widely-studied statistical problems is to infer the relationship between the response and the explanatory variables in the following linear regression model

$$\mathbf{y} = \mathbf{X}\beta + \epsilon, \tag{2.27}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$ is an n-vector of the response, $\mathbf{X} = (X_1, \ldots, X_p)$ is an $n \times p$ design matrix, and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T \in \mathbb{R}^n$ is an n-vector of i.i.d. random errors satisfying $\mathbb{E}(\epsilon_i) = 0$ and $\text{var}(\epsilon_i) = \sigma^2 < \infty$.

Technological advances have led to the explosion of data across many scientific disciplines, e.g. genomics, functional MRI, tomography and finance, to name a few, such that the dimensionality of the data p can be very large, sometimes much larger than the number of observations n. Donoho (2000) listed specific scientific problems which demanded the development of tools for high-dimensional data analysis, due to the apparent inability of classical methods in coping with the explosive growth of dimensionality.

For the last few decades, substantial progress has been made to tackle the problem of high dimensionality in linear regression, under the assumption that only a small number of variables actually contribute to the response, i.e.,

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

has its cardinality much smaller than p. Fan & Lv (2010) noted that sparsity arose naturally in many scientific problems. By way of example, in disease classification, it is often believed that only dozens of genes out of tens of thousands have significant contributions to the development of a disease. Assuming sparsity on the data structure, identifying the subset $\mathcal{S}$ can improve both model interpretability and estimation accuracy.

There exists a long list of works devoted to the high-dimensional variable selection problem and an excellent survey of literature can be found in Fan & Lv (2010). In this section, we review a selection of variable selection methods which approach the problem from different angles. First, we provide an overview of the penalised least squares (PLS) estimation, of which classical model selection methods as well as more recent works, such as the ridge regression, the Lasso (Tibshirani, 1996) and the SCAD (Fan & Li, 2001), are instances. Implementation of the PLS estimation methods is discussed in Section 2.5.2.

The Dantzig selector (Candès & Tao, 2007) is an l_1-regularisation method which is closely related to the Lasso. Section 2.5.3 provides a detailed description of the Dantzig selector and its connection with the Lasso. Then follows the

discussion of Sure Independence Screening (Fan & Lv, 2008), a dimensionality reduction procedure for ultra high-dimensional problems, in Section 2.5.4.

In Section 2.5.5, we note that the presence of (possibly spurious) non-negligible correlations among the large number of variables renders the high-dimensional variable selection problem very difficult, and present a list of methods which take into account the correlation structure of $\mathbf{X}$. In Chapter 5, we also recognise the importance of this issue, and propose a new way of measuring the association between each variable and the response, by accounting for the sample correlation structure of $\mathbf{X}$ in a data-driven manner.

Finally, we introduce some notations which are used throughout this section, and revisited later in Chapter 5. The l_q-norm for an n-vector $\mathbf{u} \in \mathbb{R}^n$ is defined as

$$\|\mathbf{u}\|_q = \left(\sum_{i=1}^{n} |u_i|^q \right)^{1/q},$$

and therefore

$$\|\mathbf{u}\|_0 = \sum_i \mathbb{I}(u_i \neq 0), \ \|\mathbf{u}\|_1 = \sum_i |u_i| \text{ and } \|\mathbf{u}\|_2 = \sqrt{\sum_i u_i^2}.$$

This definition is extended to $q = \infty$ as $\|\mathbf{u}\|_\infty = \max_i |u_i|$, and we often refer to the l_2-norm as the norm. The ith row of $\mathbf{X}$ is denoted by $\mathbf{x}_i$, i.e. $\mathbf{x}_i = (X_{i,1}, \ldots, X_{i,p})$. Let $\mathcal{D}$ be a subset of the index set $\mathcal{J} = \{1, \ldots, p\}$. Then, for any $n \times p$ matrix $\mathbf{X}$, we use $\mathbf{X}_\mathcal{D}$ to denote an $n \times |\mathcal{D}|$-submatrix of $\mathbf{X}$ with $X_j,\ j \in \mathcal{D}$ as its columns. In a similar manner, $\beta_\mathcal{D}$ denotes a $|\mathcal{D}|$-subvector of a p-vector β with $\beta_j,\ j \in \mathcal{D}$ as its elements. For a given $\mathcal{D}$ satisfying $|\mathcal{D}| < n$, we denote the projection matrix onto the column space of $\mathbf{X}_\mathcal{D}$ by $\Pi_\mathcal{D}$. Finally, C and C' are used to denote generic positive constants.

2.5.1 Penalised least squares estimators

Classical model selection methods using the tools such as Akaike's information criterion (Akaike, 1973, AIC), Mallows' C_p (Mallows, 1973) and Bayesian information criterion (Schwarz, 1978, BIC), belong to the l_0-norm PLS estimation of

the following form,

$$\min_{\tilde{\beta}\in\mathbb{R}^p} \|\mathbf{y} - \mathbf{X}\tilde{\beta}\|_2^2 + \lambda\|\tilde{\beta}\|_0, \tag{2.28}$$

with different choice of the penalty parameter $\lambda > 0$. These approaches have a unified interpretation that they all search for the best trade-off between the goodness of fit and the complexity of the model.

One critical limitation of this l_0-norm minimisation framework was noted in Candès & Tao (2007). That is, searching for a solution to the problem (2.28) requires an exhaustive search over all the subsets of columns of $\mathbf{X}$, which clearly has exponential complexity except for in a few circumstances, e.g. when $\mathbf{X}$ is an orthonormal matrix. In general, finding a solution for (2.28) may be feasible only when p ranges in a few dozens.

The l_0-norm PLS estimation is a special case of l_q-norm penalised regression which places the penalty on the l_q-norm of the parameter vector. In what follows, we provide a list of PLS estimation methods with various penalty terms.

2.5.1.1 Ridge regression

The ridge regression proposed in Hoerl & Kennard (1970) replaces the l_0-norm in (2.28) with $\|\tilde{\beta}\|_2^2$,

$$\min_{\tilde{\beta}\in\mathbb{R}^p} \|\mathbf{y} - \mathbf{X}\tilde{\beta}\|_2^2 + \lambda\|\tilde{\beta}\|_2^2. \tag{2.29}$$

As a result, the ridge regression achieves continuous shrinkage and its solution $\hat{\beta}^R$ shows good prediction performance through a bias-variance trade-off. However, a parsimonious model representation cannot be obtained by the ridge regression, and below we explain this point with a simple example.

When $\mathbf{X}^T\mathbf{X} = \mathbf{I}_p$, the ordinary least squares (OLS) estimate $\hat{\beta}^{\text{OLS}}$ is equal to $\mathbf{X}^T\mathbf{y}$. Then the problem in (2.29) is reduced to

$$\min_{\tilde{\beta}\in\mathbb{R}^p} \sum_{j=1}^{p} \left\{ \left(\hat{\beta}_j^{\text{OLS}} - \tilde{\beta}_j\right)^2 + \lambda|\tilde{\beta}_j|^2 \right\},$$

and therefore

$$\hat{\beta}_j^R = \frac{\hat{\beta}_j^{\text{OLS}}}{1+\lambda} \text{ for } j = 1, \ldots, p.$$

Thus, it is clear that as the result of ridge regression, every variable is kept in the model with each $\hat{\beta}_j^R$ shrunken towards zero.

2.5.1.2 Lasso

Least absolute shrinkage and selection operator (Tibshirani, 1996, Lasso) belongs to the class of PLS estimators with its penalty on the l_1-norm of β, i.e., it solves the following problem

$$\min_{\tilde{\beta} \in \mathbb{R}^p} \|\mathbf{y} - \mathbf{X}\tilde{\beta}\|_2^2 + \lambda \|\tilde{\beta}\|_1, \tag{2.30}$$

where the penalty on $\|\tilde{\beta}\|_1$ leads to a sparse solution with certain coefficients set to be exactly zero. This property of the Lasso can be understood in connection with the soft-thresholding rule (Section 2.4.1).

As in Section 2.5.1.1, suppose $\mathbf{X}^T\mathbf{X} = \mathbf{I}_p$. Then, the problem in (2.30) is equal to

$$\min_{\tilde{\beta} \in \mathbb{R}^p} \sum_{j=1}^{p} \left\{ \left(\hat{\beta}_j^{\text{OLS}} - \tilde{\beta}_j \right)^2 + \lambda |\tilde{\beta}_j| \right\},$$

whose solution satisfies

$$\begin{aligned} \hat{\beta}_j^L &= \text{sign}(\hat{\beta}_j^{\text{OLS}}) \cdot \left(\left| \hat{\beta}_j^{\text{OLS}} \right| - \frac{\lambda}{2} \right)_+ \\ &= \begin{cases} \hat{\beta}_j^{\text{OLS}} - \frac{\lambda}{2} & \text{if } \hat{\beta}_j^{\text{OLS}} > 0 \text{ and } |\hat{\beta}_j^{\text{OLS}}| > \frac{\lambda}{2}, \\ \hat{\beta}_j^{\text{OLS}} + \frac{\lambda}{2} & \text{if } \hat{\beta}_j^{\text{OLS}} < 0 \text{ and } |\hat{\beta}_j^{\text{OLS}}| > \frac{\lambda}{2}, \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{2.31}$$

(2.31) can be interpreted in a way that the l_1-norm penalty acts like a soft-thresholding rule and automatically sets β_j with small values of OLS estimates $\hat{\beta}_j^{\text{OLS}}$ to zero.

We note that the minimisation problem in (2.30) is convex and thus considered tractable, i.e. it can be solved in polynomial time, unlike the l_0-norm minimisation

problem. Donoho (2006) showed that if there was a sparse solution to the l_0-norm minimisation problem in (2.28), it could be well-approximated by solving the l_1-norm PLS estimation problem in (2.30).

Zhao & Yu (2006) provided a condition under which the Lasso estimator was consistent in the sense that, the Lasso solution $\hat{\beta}^L$ satisfied

$$\mathbb{P}\left(\text{sign}(\hat{\beta}^L) = \text{sign}(\beta)\right) \to 1 \text{ as } n \to \infty.$$

This condition, termed *irrepresentable* condition, requires that there exists $C < 1$ satisfying

$$\max_{j \notin \mathcal{S}} \left|\text{sign}(\beta_{\mathcal{S}})^T (\mathbf{X}_{\mathcal{S}}^T \mathbf{X}_{\mathcal{S}})^{-1} \mathbf{X}_{\mathcal{S}}^T X_j\right| \leq C < 1. \tag{2.32}$$

Focusing on the term $(\mathbf{X}_{\mathcal{S}}^T \mathbf{X}_{\mathcal{S}})^{-1} \mathbf{X}_{\mathcal{S}}^T X_j$, (2.32) can roughly be interpreted as imposing a constraint on the regression coefficients of the irrelevant variables on the relevant variables. If we re-write (2.32) to hold for every possible combination of $\text{sign}(\beta_{\mathcal{S}})$ (which is unknown), the condition amounts to requiring the l_1-norms of such regression coefficient vectors to be uniformly smaller than 1, i.e. the total amount of an irrelevant variable represented by the relevant variables is uniformly bounded from above (thus the term "irrepresentable").

The model selection consistency of the Lasso was studied in the context of graphical models by Meinshausen & Bühlmann (2008), where a condition similar to the irrepresentable condition was termed *neighbourhood stability*. The condition (2.32) can easily be violated in the presence of non-negligible correlations among the variables, and thus it is rather an unrealistic assumption for high-dimensional datasets.

Zhang & Huang (2008) showed the variable selection consistency of the Lasso under the sparse Riesz condition. Before going into the details of this condition, we need to define the sparse eigenvalues.

Definition 2.3. *The minimal sparse eigenvalue* $\phi_{\min}(d)$ *is defined for* $d \leq p$ *as*

$$\phi_{\min}(d) = \inf_{\mathbf{u} \in \mathbb{R}^d; \mathcal{D} \subset \mathcal{J}; |\mathcal{D}| \leq d} \frac{\|\mathbf{X}_{\mathcal{D}} \mathbf{u}\|_2}{\|\mathbf{u}\|_2},$$

and analogously for the maximal sparse eigenvalue $\phi_{\max}(d)$.

The sparse Riesz condition requires the existence of $C,\ C' > 0$ for which

$$\frac{\phi_{\max}((2+4C)|\mathcal{S}|+1)}{\phi_{\min}((2+4C)|\mathcal{S}|+1)} \leq C'. \tag{2.33}$$

Provided that the sparse Riesz condition holds, the Lasso estimator $\hat{\beta}^L$ has the same support as the true regression coefficients vector β with asymptotic probability 1.

There have been substantial efforts to extend the Lasso, of which we mention a few. The adaptive Lasso proposed in Zou (2006) selected the amount of penalisation adaptively for each $\tilde{\beta}_j$ as

$$\lambda_j = \lambda \cdot \left|\hat{\beta}_j^{\text{OLS}}\right|^{-\gamma}$$

for some $\gamma > 0$. It was shown that the adaptive Lasso estimator possessed the so-called oracle property for a carefully chosen λ. That is, as if equipped with an oracle furnishing complete information about which β_j's are nonzero, the adaptive Lasso estimator $\hat{\beta}^{AL}$ achieves the following:

consistency in variable selection

$$\lim_{n\to\infty} \mathbb{P}(\hat{\mathcal{S}}^{AL} = \mathcal{S}) = 1 \text{ for } \hat{\mathcal{S}}^{AL} = \{1 \leq j \leq p :\ \hat{\beta}_j^{AL} \neq 0\}.$$

asymptotic normality

$$\sqrt{n}\left(\hat{\beta}_{\mathcal{S}}^{AL} - \beta_{\mathcal{S}}\right) \to_d \mathcal{N}_{|\mathcal{S}|}\left(\mathbf{0}, \frac{\sigma^2}{n} \cdot \mathbf{X}_{\mathcal{S}}^T \mathbf{X}_{\mathcal{S}}\right).$$

The randomised Lasso proposed in Meinshausen & Bühlmann (2010) repeatedly produced the penalty parameters as $\lambda_j = \lambda \cdot W_j^{-1}$ with W_j following a uniform distribution. Then it took the frequency of each variable being estimated to be non-zero as the variable selection criterion, rather than the estimated coefficient values themselves. Due to the randomness brought in from the selection of penalty parameters, the set of variables identified by the randomised Lasso was shown to be consistent even when the irrepresentable condition was violated.

2.5.1.3 Elastic net

Zou & Hastie (2005) proposed the *elastic net*, where the penalisation was imposed on a linear combination of $\|\tilde{\beta}\|_1$ and $\|\tilde{\beta}\|_2^2$,

$$\min_{\tilde{\beta}\in\mathbb{R}^p} \|\mathbf{y} - \mathbf{X}\tilde{\beta}\|_2^2 + \lambda_1\|\tilde{\beta}\|_1 + \lambda_2\|\tilde{\beta}\|_2^2. \tag{2.34}$$

Due to its penalty on both l_1- and l_2-norms, the elastic net attains the grouping effect, i.e. regression coefficients of a group of highly correlated variables tend to be equal (up to a sign change if negatively correlated), a property that the Lasso does not possess.

2.5.1.4 SCAD

Fan & Li (2001) proposed conditions for a penalty function in the PLS estimation to return an estimator which was unbiased, sparse (i.e. small estimated coefficients were automatically set to zero) and continuous in the data. It was noted in the paper that, while l_q-penalty with $q \in (1, 2]$ did not meet the sparsity condition, the l_1-penalty did not satisfy the unbiasedness, and l_q-penalty with $q \in [0, 1)$ did not satisfy the continuity.

Their *smoothly clipped absolute deviation* (SCAD) penalty function $\sum_j p_\lambda(\tilde{\beta}_j)$ was designed to meet all three requirements, and its derivative satisfied

$$p_\lambda(t)' = \lambda\left\{\mathbf{I}(t \le \lambda) + \frac{(a\lambda - t)\cdot\mathbf{I}(a\lambda > t)}{(a-1)\lambda}\mathbf{I}(t > \lambda)\right\} \tag{2.35}$$

for some $a > 2$. The SCAD estimator was shown to achieve the aforementioned oracle property of the adaptive Lasso under some regularity conditions on the distribution of $(\mathbf{x}_i, y_i)$.

2.5.2 Implementation of PLS estimation

Efron *et al.* (2004) proposed the *least angle regression* (LARS) algorithm, whose simple modification could compute the Lasso solution path for a range of penalty parameter λ. Here we provide a rough description of the LARS algorithm.

Let $\mathcal{A}$ denote the "active set" which represents the variables included in the current model. Starting with an empty $\mathcal{A}$ (i.e., all the coefficients are set to be zero), the LARS searches for the variable which attains the maximum marginal correlation (in the absolute value) with the response $\mathbf{y}$. Denoting such variable by X_1, the LARS adds the index 1 to $\mathcal{A}$. Then the current residual $\mathbf{z}$ is updated by taking the largest step possible in the direction of X_1 away from $\mathbf{y}$, until some other variable, say X_2, achieves as much marginal correlation with the current residual as X_1. After adding the index 2 to active set $\mathcal{A}$, the current residual vector $\mathbf{z}$ is updated again, by proceeding in the "equiangular" direction between the two variables X_1 and X_2, i.e. $\mathbf{z}$ keeps the same distance from both X_1 and X_2 simultaneously, until the third variable X_3 has as much marginal correlation with $\mathbf{z}$ as X_1 and X_2. The LARS continues its equiangular progression between X_1, X_2 and X_3, until the fourth index enters $\mathcal{A}$ and so forth.

To obtain the Lasso solution path, an additional constraint is needed throughout the LARS algorithm, on the sign of each non-zero coefficient estimate, say $\tilde{\beta}(\mathcal{A})_j$, $j \in \mathcal{A}$, to agree with that of the corresponding marginal correlations between X_j, $j \in \mathcal{A}$ and the current residual $\mathbf{z}$. This Lasso-LARS modification can also be adopted to compute the solution paths for the Lasso extensions such as the adaptive Lasso, the randomised Lasso and the elastic net (provided λ_2 in (2.34) is fixed).

Another way of implementing the Lasso is the coordinate-wise descent approach. When there is only one variable X_1 in the linear model (2.27) and $\|X_1\|_2 = 1$, the Lasso solution is a soft-thresholded version of the OLS estimate $\hat{\beta}^{\text{OLS}} = X_1^T\mathbf{y}$, i.e.

$$\hat{\beta}^L \;=\; T_{\text{soft}}\left(X_1^T\mathbf{y}, \frac{\lambda}{2}\right) = \text{sign}(X_1^T\mathbf{y}) \cdot \left(|X_1^T\mathbf{y}| - \frac{\lambda}{2}\right)_+, \tag{2.36}$$

see also (2.31). Based on this observation, an iterative algorithm was proposed in Friedman *et al.* (2007) for the general case of multiple variables. It applies the soft-thresholding step in (2.36) to update each coefficient separately, with "partial residuals" in place of the response. In other words, for the column-wise

normalised $\mathbf{X}$ (i.e. $\|X_j\|_2 = 1$), each $\tilde{\beta}_j$ is updated as

$$\tilde{\beta}_j \leftarrow T_{\text{soft}}\left(X_j^T(\mathbf{y} - \tilde{\mathbf{y}}^{(j)}), \frac{\lambda}{2}\right),$$

where the partial residuals satisfy $\tilde{y}_i^{(j)} = \sum_{k \neq j} \tilde{\beta}_k x_{i,k}$, $i = 1, \ldots, n$, until all $\tilde{\beta}_j$, $j = 1, \ldots, p$ converge. This coordinate-wise descent algorithm is an attractive tool whenever a single-parameter problem is easy to solve. For the discussion of extensions of this approach as well as its computational efficiency, see Friedman *et al.* (2007) and the references therein.

For the PLS estimation problems with non-concave penalty functions such as the SCAD penalty (2.35), Fan & Li (2001) proposed the local quadratic approximation (LQA) algorithm. By locally approximating the penalty function using a quadratic function as

$$p_\lambda(|t|) \approx p_\lambda(|t^*|) + \frac{1}{2}\frac{p'_\lambda(|t^*|)}{|t^*|}(|t|^2 - |t^*|^2)$$

for a given initial value t^*, the PLS estimation problem itself becomes quadratic and thus admits a closed-form solution. In Zou & Li (2008), it was shown that a better approximation could be achieved using the local linear approximation (LLA) algorithm

$$p_\lambda(|t|) \approx p_\lambda(|t^*|) + p'_\lambda(t^*)(|t| - |t^*|).$$

It is due to the fact that, although both LLA and LQA are convex majorants of the SCAD function, the LLA is the minimum (tightest) convex majorant of the concave function on $[0, \infty)$ (Fan & Lv, 2010).

2.5.3 Dantzig selector

The Dantzig selector presented in Candès & Tao (2007) solves the following l_1-regularisation problem

$$\text{minimise } \|\tilde{\beta}\|_1 \text{ subject to } \|\mathbf{X}^T(\mathbf{y} - \mathbf{X}\tilde{\beta})\|_\infty \leq \lambda, \tag{2.37}$$

assuming that $\mathbf{X}$ is column-wise normalised, i.e. $\|X_j\|_2 = 1$. Candès & Tao (2007) introduced a condition on the correlation structure of $\mathbf{X}$ called the "uniform uncertainty principle" (UUP), which was defined using the following two quantities.

- The *s-restricted isometry constant* δ_s of $\mathbf{X}$ is the smallest quantity satisfying

$$(1-\delta_s)\|\mathbf{c}\|_2^2 \leq \|\mathbf{X}_{\mathcal{D}}\mathbf{c}\|_2^2 \leq (1+\delta_s)\|\mathbf{c}\|_2^2,$$

  for all the sets $\mathcal{D} \subset \mathcal{J}$ with $|\mathcal{D}| \leq s$ and all coefficient vectors $\mathbf{c} \in \mathbb{R}^{|\mathcal{D}|}$.

- The *s, s′-restricted orthogonality constant* $\theta_{s,s'}$ for $s + s' \leq p$ is defined as the smallest quantity which satisfies

$$|\langle \mathbf{X}_{\mathcal{D}}\mathbf{c}, \mathbf{X}_{\mathcal{D}'}\mathbf{c}' \rangle| \leq \theta_{s,s'}\|\mathbf{c}\|_2 \cdot \|\mathbf{c}'\|_2$$

  for all the disjoint sets $\mathcal{D}, \mathcal{D}' \subset \mathcal{J}$ of cardinalities $|\mathcal{D}| \leq s$ and $|\mathcal{D}'| \leq s'$ and all coefficient vectors $\mathbf{c} \in \mathbb{R}^{|\mathcal{D}|}$, $\mathbf{c}' \in \mathbb{R}^{|\mathcal{D}'|}$.

Then the UUP requires $\delta_{2|\mathcal{S}|} + \theta_{|\mathcal{S}|,2|\mathcal{S}|} < 1$, under which the Dantzig selector $\hat{\beta}^D$ satisfies

$$\|\hat{\beta}^D - \beta\|_2^2 \leq C^2 \cdot 2\log p \cdot \left(\sigma^2 + \sum_{j=1}^{p} \min(\beta_j^2, \sigma^2)\right) \tag{2.38}$$

with $\lambda = \sigma\sqrt{2\log p}$. We note that $\hat{\beta}^D$ achieves a non-asymptotic oracle inequality under l_2-loss, in the sense that the right-hand side of (2.38) is within a logarithmic factor $(\log p)$ to the ideal MSE attained with an oracle telling the indices of non-zero coefficients. We note that essentially, the UUP requires that every submatrix $\mathbf{X}_{\mathcal{D}}$, with its number of columns $|\mathcal{D}|$ comparable to $|\mathcal{S}|$, should behave as if they were orthonormal, a condition that can be stringent in high-dimensional problems.

The similarities between the Dantzig selector and the Lasso can be drawn by re-writing (2.37) and (2.30) as

$$\text{minimise } \|\tilde{\beta}\|_1 \text{ subject to } \|\mathbf{X}^T\mathbf{X}(\tilde{\beta} - \hat{\beta}^{\text{OLS}})\|_\infty \leq \lambda_D,$$

$$\text{minimise } \|\tilde{\beta}\|_1 \text{ subject to } \|\mathbf{X}(\tilde{\beta} - \hat{\beta}^{\text{OLS}})\|_2^2 \leq \lambda_L,$$

respectively. Based on this relationship between the Dantzig selector and the Lasso, James *et al.* (2009) proposed a new algorithm named DASSO for fitting the entire solution path of the Dantzig selector at a computational cost similar to that of the LARS algorithm.

2.5.4 Sure independence screening

Fan & Lv (2008) noted that, when the dimensionality p grew exponentially with the sample size n, there could exist non-negligible correlations even among those $X_1, \ldots, X_p$ generated as i.i.d. Gaussian. If so, conditions such as the irrepresentable condition or the UUP are not likely to be met, and even when the UUP holds, we may not be able to ignore the multiplicative factor $\log p$ in (2.38).

As a way of tackling these difficulties with ultra-high dimensionality, Fan & Lv (2008) proposed the Sure Independence Screening (SIS) which reduced the dimensionality of the data from ultra-high level to that below the sample size in a computationally efficient way. The SIS achieves this by applying the component-wise regression, i.e. marginal correlation screening. It screens $\mathbf{X}^T\mathbf{y} = (X_1^T\mathbf{y}, \ldots, X_p^T\mathbf{y})^T$, ranks the importance of each variable according to the magnitude of corresponding marginal correlation, and selects a submodel of cardinality $d = d_n$ as

$$\hat{\mathcal{A}}_d = \left\{1 \leq j \leq p : |X_j^T\mathbf{y}| \text{ is among the first } d_n \text{ largest of all.}\right\}.$$

Fan & Lv (2008) showed that under certain conditions, the SIS achieved the *sure screening property*

$$\mathbb{P}\left(\mathcal{S} \subset \hat{\mathcal{A}}_d\right) \to 1.$$

When this sure screening property is satisfied, we can expect better estimation accuracy by applying the PLS estimation based methods or the Dantzig selector to a submodel chosen by the SIS, which is of the dimensionality comparable to the data size.

There are some limitations of the SIS, however, which arise from the failure of marginal correlation screening in the presence of high correlations among the variables:

(i) some irrelevant variables that are highly correlated with the relevant ones can have higher priority to be selected by the SIS than other relevant variables that are relatively weakly related to the response;

(ii) a relevant variable that is marginally uncorrelated but jointly correlated with the response cannot be picked by the SIS;

(iii) there may exist collinearity between the variables.

To overcome these difficulties, an iterative version of the SIS (ISIS) was proposed, which applied the SIS based variable selection methods, such as the SIS-Dantzig selector or the SIS-SCAD, in an iterative manner. In the following section, we present some other methods which also address this issue of non-negligible correlations among the variables, by using the measures other than marginal correlation to infer the strength of association between each variable and the response.

2.5.5 High correlations among the variables

One of the major complications encountered in high-dimensional variable selection is the presence of possibly spurious, non-negligible correlations among the variables, an example of which is shown in Figure 2.2. We generated an $n \times p$-matrix $\mathbf{X}$ with i.i.d. Gaussian entries, where $n = 100$ and $p = 2000$, and plotted the magnitude of correlations among the columns of $\mathbf{X}$ in increasing order. The figure shows that when p is large, the absolute values of sample correlations even among i.i.d. variables can be greater than 0.5, which is clearly non-negligible.

As noted in (i)–(iii) of Section 2.5.4, when the variables are highly correlated, marginal correlation can be misleading as a measure of association between the variables and the response. We summarise below some iterative algorithms, where measures other than marginal correlation are adopted to determine which variable(s) should be included in (or removed from) the model at each iteration.

Bühlmann *et al.* (2009) proposed the PC-simple algorithm, which used partial correlation instead of marginal correlation in order to iteratively remove irrelevant

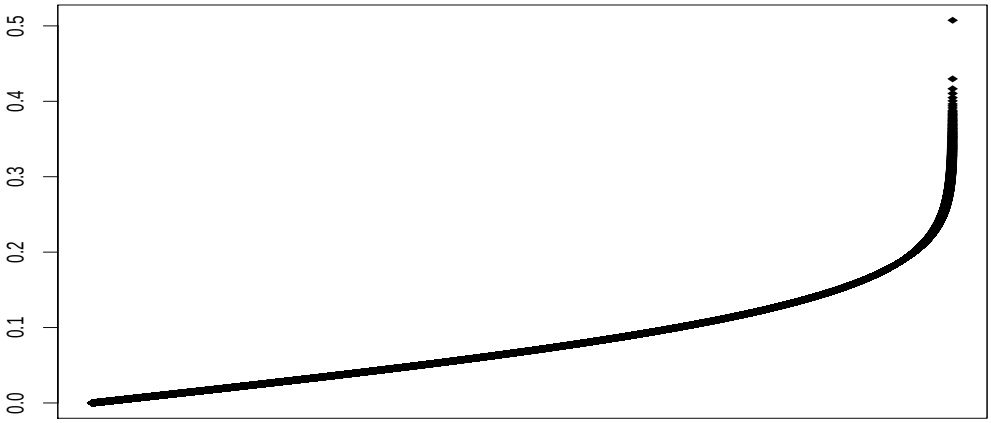


Figure 2.2: Correlations among i.i.d. Gaussian variables in increasing order when $n = 100,\ p = 2000$.

variables from the model. The partial correlation between X_j and $\mathbf{y}$ conditional on $\mathbf{X}_{\mathcal{D}}$ for some $\mathcal{D} \subset \mathcal{J} \setminus \{j\}$ is defined as the correlation between the residuals, which result from the linear regression of X_j with $\mathbf{X}_{\mathcal{D}}$ and that of $\mathbf{y}$ with $\mathbf{X}_{\mathcal{D}}$.

Also, we note that "greedy" algorithms such as the traditional forward selection (see e.g. Chapter 8.5 of Weisberg (1980)) or the forward regression (Wang, 2009) have an interpretation in this context due to their greediness (see below for an explanation of the term), unlike less greedy algorithms for generating a solution path such as the LARS.

At each iteration, both the forward selection and the forward regression algorithms update the current residual $\mathbf{z}$ after taking the greediest step towards the variables included in the current model $\mathcal{A}$, i.e., $\mathbf{z}$ is obtained as the projection of $\mathbf{y}$ onto the orthogonal complement of the current model space spanned by $\mathbf{X}_{\mathcal{A}}$. This greedy progression can be seen as taking into account the correlations between the variables which are in the current model and those which are not, as measuring the marginal correlation between $X_j,\ j \notin \mathcal{A}$ and the current residual $\mathbf{z} = (\mathbf{I}_n - \Pi_{\mathcal{A}})\mathbf{y}$ is equivalent to measuring the association between such X_j and $\mathbf{y}$ conditional on $X_k,\ k \in \mathcal{A}$.

As a non-iterative method, the regression framework proposed in Witten & Tibshirani (2009) accounts for the correlation structure of $\mathbf{X}$ using the so-called "scout" procedure. It first identifies non-negligible partial correlations between

X_j and X_k conditional on all the other variables $X_l,\ l \neq j, k$ for all $j \neq k$, which is achieved by obtaining a shrunken estimate of the inverse covariance matrix of $\mathbf{X}$, and then applies this estimate for computing a PLS estimate of β.

In Chapter 5, we propose to measure the contribution of each variable to the response by adaptively taking into account the sample correlation structure for each X_j, and present an iterative algorithm based on this new measure. A more detailed description of the methods discussed in this section, in comparison with our proposed methodology, is provided later in Section 5.3.3.

Chapter 3

Multiscale and multilevel technique for consistent breakpoint detection in piecewise stationary time series

The (weak) stationarity assumption implies that

- the mean and the variance of the underlying process are constant,
- and its autocovariance function depends only on the time lag.

Although great efforts in the theoretical treatment of time series analysis have been made under the stationarity assumption, it may not be a realistic assumption for modelling the time series data which are observed in naturally nonstationary environments; examples of such datasets include speech signals, Electroencephalography (EEG) data, seismic signals and financial time series.

Piecewise stationarity is arguably the simplest departure from stationarity, and one task when faced with a time series of this form is to detect breakpoints in its dependence structure. In this chapter, we propose a procedure for detecting breakpoints in the second-order structure of a piecewise stationary process, which is linear but otherwise does not follow any particular parametric model. The nonparametric time series model chosen for this purpose is the locally sta-

tionary wavelet (LSW) model, which was first proposed by Nason *et al.* (2000). Section 2.2.2 of this thesis provides a detailed introduction to the LSW model as well as its extensions and applications. In this chapter, we use the LSW model as presented in Fryzlewicz & Nason (2006), which is a modified version of the model described in Section 2.2.2. Under this LSW model, the piecewise-constant, second-order structure of a time series is completely described by its wavelet-based, local periodogram sequences at multiple scales, and these wavelet periodograms are the basic statistics of our segmentation procedure.

To achieve multiple breakpoint detection, we propose a new binary segmentation method which is applied to wavelet periodogram sequences at each scale separately. For an overview of binary segmentation algorithm, see Section 2.3.2 and references therein. We then introduce our within-scale and across-scales post-processing steps which succeed the binary segmentation procedure, and show that the combined methodology achieves consistent estimation of the breakpoints in the second-order structure of the original time series, in terms of their total number and locations.

We note that our method can simultaneously be termed *multiscale* and *multilevel*, as the basic time series model used for our purpose is a wavelet-based and thus a multiscale model, and the core methodology to segment the wavelet periodogram sequence at each scale is based on binary segmentation and is thus a multilevel procedure.

The rest of this chapter is organised as follows. Section 3.1 describes the LSW model and justifies its choice for our purpose. Our breakpoint detection methodology, which consists of a binary segmentation procedure as well as post-processing steps, is introduced in Section 3.2, where we also demonstrate its theoretical consistency in estimating the total number and locations of breakpoints. In Section 3.3, the outcome of a simulation study is presented, where the performance of our method is compared with that of the state of the art. In Section 3.4, we apply our technique to the segmentation of the historical Dow Jones index. All the proofs of our theoretical results are provided in Section 3.5.

3.1 Locally stationary wavelet time series

In this section, we first introduce the time series model first presented in Fryzlewicz & Nason (2006), which is a slightly modified version of the LSW time series model in Section 2.2.2. Then follow its properties and the justification of this choice of LSW model as an attractive framework for developing our time series segmentation methodology.

The modified version of LSW model for piecewise stationary time series is as below.

Definition 3.1. *A triangular stochastic array* $\{X_{t,T}\}_{t=0}^{T-1}$ *for* $T = 1, 2, \ldots,$ *is in a class of LSW processes, if there exists a mean-square representation*

$$X_{t,T} = \sum_{i=-\infty}^{-1} \sum_{k=-\infty}^{\infty} W_i(k/T)\psi_{i,t-k}\xi_{i,k} \tag{3.1}$$

where $i \in \{-1, -2, \ldots\}$ *and* $k \in \mathbb{Z}$ *are scale and location parameters respectively,* $\psi_i = (\psi_{i,0}, \ldots, \psi_{i,\mathcal{L}_i-1})^T$ *are discrete, real-valued, compactly supported, non-decimated wavelet vectors with support lengths* $\mathcal{L}_i = O(2^{-i})$*, and* $\xi_{i,k}$ *are zero-mean, orthonormal, identically distributed random variables.*

For each $i \leq -1$*, there exists a real-valued, piecewise constant function* $W_i(z) : [0, 1] \to \mathbb{R}$ *which has a finite (but unknown) number of jumps. Let* L_i *denote the total magnitude of jumps in* $W_i^2(z)$*. The variability of functions* $W_i(z)$ *is controlled such that* $W_i(z)$ *satisfy*

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

We assume that random variables $\xi_{i,k}$ follow the standard normal distribution. Extensions to non-Gaussian distributions may be possible but technically difficult, and thus not discussed in this thesis. Comparing the above definition with the Cramér's representation of stationary processes (see (2.12) in Section 2.2.1), $W_i(k/T)$ is a scale- and location-dependent transfer function, the wavelet vectors ψ_i are analogous to the Fourier exponentials, and the innovations $\xi_{i,k}$ correspond to the orthonormal increment process. By assuming that each $W_i(z)$ is piecewise

constant, we are able to model time series data with a piecewise constant second-order structure where, between any two breakpoints in $W_i(z)$, the second-order structure remains constant.

Here we briefly recall some properties of the LSW model provided in Section 2.2.2. For an LSW process, its evolutionary wavelet spectrum (EWS) is defined as $S_i(z) = W_i(z)^2$. In Nason *et al.* (2000), it was shown that the EWS had the following one-to-one correspondence with the asymptotic local autocovariance function of the process,

$$c(z, \tau) = \lim_{T \to \infty} \operatorname{cov}(X_{[zT],T}, X_{[zT]+\tau,T}),$$

see (2.18) for further details. We note the validity of $W_i(z)$ as a transfer function, since the variance of the resulting time series $X_{t,T}$ is uniformly bounded over t, and the one-to-one correspondence between $c(z, \tau)$ and $S_i(z)$ leads to the model identifiability.

Our objective is to develop a consistent method for detecting breakpoints in the EWS, and consequently provide a segmentation of the original time series. The following assumption is placed on the breakpoints present in the EWS, which implies that there are finite number of breakpoints in the second-order structure of the process.

Assumption 3.1. *The set of those locations z where (possibly infinitely many) functions $S_i(z)$ contain a jump, is finite. That is, let*

$$\mathcal{B} = \{z : \ \exists\, i \text{ for which } \lim_{u \to z-} S_i(u) \neq \lim_{u \to z+} S_i(u)\},$$

then $B = |\mathcal{B}| < \infty$.

We recall the definition of the wavelet periodogram of $X_{t,T}$ given in Section 2.2.2, which is essentially a sequence of squared wavelet coefficients of $X_{t,T}$.

Definition 3.2. *Let $X_{t,T}$ be an LSW process constructed using the wavelet system ψ. Then, the triangular stochastic array*

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

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

We further recall the definition of autocorrelation wavelets

$$\Psi_i(\tau) = \sum_k \psi_{i,k}\psi_{i,k-\tau},$$

and that of the autocorrelation wavelet inner product matrix

$$\mathbf{A} = \left(\sum_\tau \Psi_i(\tau)\Psi_j(\tau) \right)_{i,j<0}.$$

Then, a function $\beta_i(z)$ is defined as the linear transform of EWS with respect to the autocorrelation wavelet inner product matrix, i.e.,

$$\beta_i(z) = \sum_{j=-\infty}^{-1} S_j(z)A_{i,j}.$$

Fryzlewicz & Nason (2006) showed that $\mathbb{E}I_{t,T}^{(i)}$, the expectation of a wavelet periodogram, is "close" to $\beta_i(z)$ in the following sense.

Proposition 3.1 (Propositions 2.1-2.2 of Fryzlewicz & Nason (2006))**.** *Let $I_{t,T}^{(i)}$ be the wavelet periodogram at a fixed scale i. Under Assumption 3.1, the integrated bias between $\mathbb{E}I_{t,T}^{(i)}$ and $\beta_i(t/T)$ satisfies*

$$T^{-1}\sum_{t=0}^{T-1}\left|\mathbb{E}I_{t,T}^{(i)} - \beta_i(t/T)\right|^2 = O(T^{-1}2^{-i}) + b_{i,T}, \tag{3.2}$$

where $b_{i,T}$ depends on the sequence $\{L_i\}_i$. For example, if $L_i = O(a^i)$ for $a > 2$ then $b_{i,T} = O(T^{\frac{1}{2\log_2 a-1}-1})$, which implies, in particular, that the rate of convergence in (3.2) is

$$O\left\{T^{-\min\left(1-\frac{1}{2\log_2 a-1},1-\varphi\right)}\right\}$$

uniformly over $i = -1, \ldots, -\varphi\log_2 T$.

Further, each $\beta_i(z)$ is a piecewise constant function with at most B jumps, all of which occur in the set $\mathcal{B}$. Additionally, if there exists $C > 0$ for which $S_i(z) \leq C2^i$ for all i, we have $\beta_i(z) \leq C$ uniformly over all i.

In summary, we conclude that there exists a one-to-one correspondence between the EWS, asymptotic local autocovariance function $c(z, \tau)$ and the functions $\beta_i(z)$ (being the asymptotic expectation of the wavelet periodograms). Therefore, every breakpoint in the second-order structure results in a breakpoint in at least one of the $\beta_i(z)$'s and is thus detectable, at least with $T \to \infty$, by analysing the wavelet periodogram sequences.

We note that $\mathbb{E}I_{t,T}^{(i)}$ itself is piecewise constant by definition, except on the intervals of length $C2^{-i}$ around the discontinuities occurring in $\mathcal{B}$, where C is a positive constant depending on the wavelet system ψ used to construct $X_{t,T}$. Given a breakpoint $\nu \in \mathcal{B}$, the computation of $I_{t,T}^{(i)}$ for $t \in [\nu - C2^{-i}, \nu + C2^{-i}]$ involves the observations from two different stationary segments, which results in $\mathbb{E}I_{t,T}^{(i)}$ being "almost" piecewise constant yet not completely so.

The finiteness of $\mathcal{B}$ implies that there exists a fixed index $I^* < \lfloor \log_2 T \rfloor$, such that each breakpoint in $\mathcal{B}$ can be found in at least one of the functions $S_i(z)$ for $i = -1, \ldots, -I^*$. Thus, from the invertibility of the matrix $(A_{i,k})_{i,k=-1}^{-I^*}$ and the closeness between $\beta_i(z)$ and $\mathbb{E}I_{t,T}^{(i)}$ (as noted in Proposition 3.1), we conclude that every breakpoint is detectable from the wavelet periodogram sequences at scales $i = -1, \ldots, -I^*$, and we only consider $I_{t,T}^{(i)}$ at these scales for our breakpoint detection procedure.

Since I^* is fixed but unknown, in our theoretical considerations we permit it to increase slowly to infinity with T, see Section 3.5.1 for more discussion on the rate at which I^* is allowed to increase. A further reason for disregarding the coarse scales $i < -I^*$ is that the autocorrelation within each wavelet periodogram sequence becomes stronger at coarser scales. Similarly, the intervals on which $\mathbb{E}I_{t,T}^{(i)}$ is not piecewise constant become longer (being of the length $C2^{-i}$). In summary, for coarse scales, wavelet periodograms provide little useful information about the breakpoints and thus can safely be omitted.

We close this section by listing the rationale behind the choice of the LSW model as a suitable framework for developing our methodology.

(i) The entire piecewise constant second-order structure of the process is "encoded" in the (asymptotically) piecewise constant sequences $\mathbb{E}I_{t,T}^{(i)}$. That is, any breakpoints in the second-order structure must be detectable by analysing the wavelet periodograms, which are relatively easy to handle as

they follow a multiplicative model (see Section 3.2.1) and are localised due to the compact support of the wavelets used in their computation.

(ii) Due to the "whitening" property of wavelets, the wavelet periodogram sequences are often much less autocorrelated than the original process. In Section 9.2.2 of Vidakovic (1999), the whitening property of wavelets is formalised for a second-order stationary time series X_t with a sufficiently smooth spectral density. Defining its wavelet coefficient as $r_{i,k} = \sum_s X_s \psi_{i,s-k}$, the across-scales and within-scale covariance of the wavelet coefficients satisfies the following, provided the wavelet used is also sufficiently smooth.

- $\mathbb{E}(r_{i,k} r_{i',k'})$ vanishes for $|i - i'| > 1$.
- $\mathbb{E}(r_{i,k} r_{i',k'})$ is arbitrarily small for $|i - i'| = 1$.
- $\mathbb{E}(r_{i,k} r_{i',k'})$ decays as $o(|k - k'|^{-1})$ within each scale, i.e. when $i = i'$.

Although this whitening property of wavelets motivated our choice of the LSW model, we emphasise that our segmentation method does permit autocorrelation in the wavelet periodogram sequences, as specified later in Section 3.2.1. Therefore, our procedure does not formally rely on the whitening effect of wavelets on the periodogram sequences.

(iii) The entire array of the wavelet periodograms at all scales is easily and rapidly computable via the non-decimated wavelet transform (see Section 2.1.3).

(iv) The use of the *rescaled time* $z = k/T$ in (3.1) and the associated regularity assumptions on the transfer functions $W_i(z)$ allow us to establish rigorous asymptotic properties of our procedure.

3.2 Binary segmentation algorithm

In this section, we first note that each wavelet periodogram sequence follows a multiplicative model, and introduce a binary segmentation algorithm for a generic

class of multiplicative sequences. Binary segmentation is a computationally efficient tool which searches for multiple breakpoints in a recursive manner, and thus can be categorised as a greedy and multilevel algorithm. As noted in Section 2.3.2, Venkatraman (1993) applied the procedure to a sequence of independent, normal variables with multiple breakpoints in its mean, and showed that the detected breakpoints were consistent in terms of their total number and locations. In the following, we aim at extending these consistency results to the multiplicative model where dependence between observations is permitted.

3.2.1 Generic multiplicative model

Recall that each wavelet periodogram ordinate is simply a squared wavelet coefficient of a zero-mean Gaussian time series, which is distributed as a scaled χ_1^2 variable with the following decomposition:

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

where $\{Z_{t,T}\}_{t=0}^{T-1}$ are autocorrelated standard normal variables. Hence we first develop a generic breakpoint detection tool for multiplicative sequences

$$Y_{t,T}^2 = \sigma_{t,T}^2 \cdot Z_{t,T}^2, \quad t = 0, \ldots, T-1. \tag{3.3}$$

$I_{t,T}^{(i)}$ and $\mathbb{E}I_{t,T}^{(i)}$ can be viewed as special cases of $Y_{t,T}^2$ and $\sigma_{t,T}^2$, respectively. We assume the following additional conditions, which are satisfied by $I_{t,T}^{(i)}$ and $\mathbb{E}I_{t,T}^{(i)}$ under the assumptions in Theorem 3.2 later on.

(i) $\sigma_{t,T}^2$ is deterministic and "close" to a piecewise constant function $\sigma^2(t/T)$ in the sense that

- $\sigma_{t,T}^2$ is piecewise constant apart from intervals of length at most $C2^{I^*}$ around the discontinuities in $\sigma^2(z)$ for some constant $C > 0$;
- the integrated squared bias between $\sigma_{t,T}^2$ and $\sigma^2(t/T)$ satisfies

$$T^{-1} \sum_{t=0}^{T-1} |\sigma_{t,T}^2 - \sigma^2(t/T)|^2 = o(\log^{-1} T),$$

where the latter rate comes from the rate of convergence of the integrated squared bias between $\beta_i(t/T)$ and $\mathbb{E}I_{t,T}^{(i)}$ (see Proposition 3.1), and from the fact that our attention is limited to wavelet periodograms at I^* finest scales only.

Further, $\sigma^2(z)$ is bounded from above and away from zero, with a finite but unknown number of jumps.

(ii) $\{Z_{t,T}\}_{t=0}^{T-1}$ is a sequence of standard normal variables and its autocorrelation sequence is absolutely summable asymptotically. That is, the function

$$\rho(\tau) = \sup_{t,T} |\mathrm{cor}(Z_{t,T}, Z_{t-\tau,T})|$$

satisfies $\rho_\infty^1 < \infty$, where $\rho_\infty^p = \sum_\tau |\rho(\tau)|^p$.

A simple example of $\{Z_{t,T}\}_{t=0}^{T-1}$ satisfying this requirement is a short-memory stationary process, for which $\rho_\infty^1 = \sum_\tau |\mathrm{cor}(Z_{t,T}, Z_{t-\tau,T})|$. Then the process $Y_{t,T}$ is a time-modulated stationary process.

Once the breakpoint detection algorithm for the generic model (3.3) has been established, it can be applied to segment each wavelet periodogram sequence.

3.2.2 Algorithm

The first step of the binary segmentation procedure is to find the most likely location of a breakpoint. We locate such a point in the interval $(0, T-1)$ as the one which maximises the absolute value of

$$\mathbb{Y}_{0,T-1}^b = \sqrt{\frac{T-b}{T\cdot b}}\sum_{t=0}^{b-1} Y_{t,T}^2 - \sqrt{\frac{b}{T\cdot(T-b)}}\sum_{t=b}^{T-1} Y_{t,T}^2 \tag{3.4}$$

$$= \sqrt{\frac{(T-b)\cdot b}{T}}\left(\frac{1}{b}\sum_{t=0}^{b-1} Y_{t,T}^2 - \frac{1}{T-b}\sum_{t=b}^{T-1} Y_{t,T}^2\right). \tag{3.5}$$

From (3.5), $\mathbb{Y}_{0,T-1}^b$ can be interpreted as a scaled difference between the partial means of two segments $\{Y_{t,T}^2\}_{t=0}^{b-1}$ and $\{Y_{t,T}^2\}_{t=b}^{T-1}$, where the scaling factor is chosen

such that the variance of $\mathbb{Y}^b_{0,T-1}$ remains constant over b in the idealised case of $Y^2_{t,T}$ being i.i.d. Once we find

$$b_{1,1} = \arg \max_{b \in (0,T-1)} \left| \mathbb{Y}^b_{0,T-1} \right|,$$

we use $\left| \mathbb{Y}^{b_{1,1}}_{0,T-1} \right|$ to test the null hypothesis of $\sigma^2(t/T)$ being constant over $[0, T-1]$. The test statistic and its critical value are established such that when a breakpoint is present in a given interval, the null hypothesis is rejected with probability converging to 1. If the null hypothesis is rejected, we continue the simultaneous locating and testing of breakpoints, separately on the two segments to the left and right of $b_{1,1}$, in a recursive manner until no further breakpoints are detected.

The algorithm is summarised below, where j is the level index and l is the location index of the node at each level. We note that the term "level" is used to indicate the progression of the segmentation procedure, in contrast to "scale" which is used to describe the multiscale nature of our wavelet model.

Binary segmentation algorithm

Step 0 Begin with $(j, l) = (1, 1)$. Let $s_{j,l} = 0$ and $e_{j,l} = T - 1$.

Step 1 Let $n_{j,l} = e_{j,l} - s_{j,l} + 1$. Iteratively compute $\mathbb{Y}^b_{s_{j,l},e_{j,l}}$ as in (3.4) for $b \in \mathcal{D}_{j,l}$ where

$$\mathcal{D}_{j,l} = \left\{ b \in (s_{j,l}, e_{j,l}) : \max \left(\sqrt{\frac{e_{j,l} - b}{b - s_{j,l} + 1}}, \sqrt{\frac{b - s_{j,l} + 1}{e_{j,l} - b}} \right) \leq c \right\} \quad (3.6)$$

with $c \geq 1$ being a fixed constant. Next, find $b_{j,l}$ which maximises the absolute value of $\mathbb{Y}^b_{s_{j,l},e_{j,l}}$, i.e.,

$$b_{j,l} = \arg \max_{b \in \mathcal{D}_{j,l}} \left| \mathbb{Y}^b_{s_{j,l},e_{j,l}} \right|,$$

and compute $m_{j,l} = \sum_{t=s_{j,l}}^{e_{j,l}} Y^2_{t,T} / \sqrt{n_{j,l}}$.

Step 2 Perform hard thresholding on $|d_{j,l}|/m_{j,l}$ with the threshold chosen as

$t_{j,l} = \tau T^{\theta}\sqrt{\log T/n_{j,l}}$, such that

$$\hat{d}_{j,l} = \begin{cases} d_{j,l} & \text{if } |d_{j,l}| > t_{j,l} \cdot m_{j,l}, \\ 0 & \text{otherwise.} \end{cases}$$

The choice of θ and τ is discussed in Section 3.2.4.

Step 3 If either $\hat{d}_{j,l} = 0$ or $\max\{b_{j,l} - s_{j,l} + 1, e_{j,l} - b_{j,l}\} < \Delta_T$ for l, stop the algorithm on the segment $(s_{j,l}, e_{j,l})$. If not, divide the segment $(s_{j,l}, e_{j,l})$ into two to the left and to the right of the detected breakpoint $b_{j,l}$ as

$$(s_{j+1,2l-1}, e_{j+1,2l-1}) = (s_{j,l}, b_{j,l}) \text{ and } (s_{j+1,2l}, e_{j+1,2l}) = (b_{j,l} + 1, e_{j,l}),$$

and update the level j as $j \leftarrow j + 1$. Again, the choice of Δ_T is discussed in Section 3.2.4.

Step 4 Repeat Steps 1–3.

The set of detected breakpoints from the above algorithm is $\{b_{j,l} : \hat{d}_{j,l} \neq 0\}$. The condition (3.6) imposed on b in Step 1 comes from the fact that the breakpoints should be sufficiently scattered over time without being too close to each other. Note that a similar condition is required of the true breakpoints in $\sigma^2(t/T)$ in Assumption 3.2 of Section 3.2.3.

The test statistic $|d_{j,l}|/m_{j,l}$ is a scaled version of the test statistics used in the iterative cumulative sum of squares (ICSS) algorithm, which is another binary segmentation procedure introduced in Inclán & Tiao (1994) for detecting multiple shifts in the variance of observations. However, their test criterion is derived empirically under the assumption of observations being independent, and thus there is no guarantee that the ICSS algorithm produces consistent breakpoint estimates. On the other hand, our algorithm permits the presence of autocorrelation in target sequences, and its test criterion enables the consistent identification of the total number and locations of breakpoints, which is further discussed in Section 3.2.3. Kouamo *et al.* (2010) proposed a CUSUM-type test, which was applied to wavelet variance at one or several scales to detect the presence of nonstationarity for a class of processes. We note that, although it also permits

correlation in the target statistic, the test procedure is designed for detecting a single change in the data.

Finally, we note the relationship between our binary segmentation procedure and the *Haar-Fisz* technique which was proposed by Fryzlewicz & Nason (2006) and Fryzlewicz *et al.* (2006) in different contexts. In the former, the Haar-Fisz technique was adopted for estimating the time-varying local variance of an LSW time series, and in the latter for estimating time-varying volatility in a locally stationary model for financial log-returns. Each Haar-Fisz method has a device (termed the Fisz transform) for stabilising the variance of the Haar wavelet coefficients of the data, such that the distribution of resulting statistics is brought close to Gaussianity with constant variance. This is similar to Step 2 in our algorithm, where the differential statistic $d_{j,l}$ is divided by the local mean $m_{j,l}$ (up to a multiplicative factor $\sqrt{n_{j,l}}$), with the convention $0/0 = 0$. However, the Fisz transform was defined only for the case $b = (e_{j,l} + s_{j,l} - 1)/2$, i.e. when the segments were split in half, and it was not used for the purpose of breakpoint detection.

3.2.2.1 Post-processing within a sequence

We further equip the segmentation procedure with an extra step which is aimed at reducing the risk of overestimating the number of breakpoints. The ICSS algorithm has a "fine-tune" step: if more than one breakpoint is found, each breakpoint is checked against the adjacent ones to reduce the risk of overestimation. We propose a post-processing procedure performing a similar task within the single-sequence multiplicative model (3.3). That is, at each breakpoint, the test statistic is re-calculated over the interval between its two neighbouring breakpoints and compared with the threshold again.

Denote the breakpoint estimates as $\hat{\eta}_p,\ p = 1, \ldots, \hat{N}$ and let $\hat{\eta}_0 = 0$, $\hat{\eta}_{\hat{N}+1} = T$. For each $\hat{\eta}_p$, we examine whether

$$\left| \mathbb{Y}^{\hat{\eta}_p}_{\hat{\eta}_{p-1}+1,\hat{\eta}_{p+1}} \right| > \tau T^{\theta} \sqrt{\log T} \cdot \frac{1}{\hat{\eta}_{p+1} - \hat{\eta}_{p-1}} \sum_{t=\hat{\eta}_{p-1}+1}^{\hat{\eta}_{p+1}} Y^2_{t,T}, \tag{3.7}$$

for $\mathbb{Y}^b_{s,e}$ defined as in (3.4). If the above inequality does not hold, $\hat{\eta}_p$ is removed

and the same procedure is repeated with the reduced set of breakpoints until the set does not change.

Note that the fine-tune step of the ICSS algorithm re-calculates both the location and test statistic at each iteration, and therefore the locations of breakpoints are subject to change after tuning. However in our post-processing procedure, only the test statistic is re-calculated at existing breakpoints and thus their locations are preserved. We thus emphasise that our within-scale post-processing step is in line with the theoretical derivation of breakpoint detection consistency in the sense that

(a) the extra check in (3.7) is of the same form as Step 2 in the original algorithm, and

(b) the locations of the breakpoints that survive the post-processing are unchanged.

For more discussion on this point, see our Lemmas 3.5–3.6 and the subsequent discussion in Section 3.5.1.

3.2.3 Consistency of detected breakpoints

In a breakpoint detection problem, it is desirable that the proposed procedure should correctly identify the total number and locations of breakpoints. In this section, Theorem 3.1 first shows the consistency of our algorithm for a multiplicative sequence as in (3.3), which corresponds to the wavelet periodogram sequence at a single scale. Later, Theorem 3.2 demonstrates how this consistency result for a single scale carries over to the consistency of our methodology in detecting breakpoints in the entire second-order structure of the original LSW process $X_{t,T}$.

Denote the number of breakpoints in $\sigma^2(t/T)$ by N and the breakpoints themselves by

$$0 < \eta_1 < \ldots < \eta_N < T - 1,$$

with $\eta_0 = 0$ and $\eta_{N+1} = T - 1$. The following assumption states that the breakpoints η_p should sufficiently be scattered over time without being too close to each other.

Assumption 3.2. *For $\theta \in (1/4, 1/2)$ and $\Theta \in (\theta + 1/2, 1)$, the length of each segment in $\sigma^2(t/T)$ is bounded from below by $\delta_T = CT^{\Theta}$ for some $C > 0$. Further, the breakpoints cannot be too close to each other in the sense that there exists a fixed constant $c \geq 1$ satisfying*

$$\max_{1 \leq p \leq N} \left\{ \sqrt{\frac{\eta_p - \eta_{p-1}}{\eta_{p+1} - \eta_p}}, \sqrt{\frac{\eta_{p+1} - \eta_p}{\eta_p - \eta_{p-1}}} \right\} \leq c.$$

The relationship between Assumption 3.2 and the condition (3.6) imposed in the binary segmentation algorithm can readily be noted. Then, the following theorem shows the consistency of binary segmentation algorithm for a sequence following the multiplicative model in (3.3).

Theorem 3.1. *Suppose that $\{Y_{t,T}\}_{t=0}^{T-1}$ follows model (3.3). Assume there exist $M, m > 0$ such that*

- $\sup_t |\sigma^2(t/T)| \leq M$ *and*
- $\inf_{1 \leq i \leq N} \left| \sigma^2\left(\frac{\eta_i + 1}{T}\right) - \sigma^2\left(\frac{\eta_i}{T}\right) \right| \geq m.$

Then, under Assumption 3.2, the breakpoints detected by our binary segmentation procedure are consistent in terms of their total number and locations. That is,

$$\mathbb{P}\left\{ \hat{N} = N : \ |\hat{\eta}_p - \eta_p| \leq C\epsilon_T, \ 1 \leq p \leq N \right\} \to 1 \ \text{as} \ T \to \infty,$$

where $\hat{\eta}_p, \ p = 1, \ldots, \hat{N}$ denote the detected breakpoints, $\epsilon_T = T^{1/2} \log T$ and C denotes an arbitrary positive constant.

Interpreting this result in the rescaled time interval $[0, 1]$, we have $\epsilon_T / T = T^{-1/2} \log T \to 0$ as $T \to \infty$.

3.2.3.1 Post-processing across the scales

As mentioned in Section 3.1, we only consider wavelet periodograms $I_{t,T}^{(i)}$ at scales $i = -1, \ldots, -I^*$, where I^* is chosen to satisfy

$$2^{I^*} \ll \epsilon_T = T^{1/2} \log T$$

such that the bias between $\sigma^2_{t,T}$ and $\sigma^2(t/T)$ does not influence the derivation of consistency result in Theorem 3.1.

Recall that any breakpoint in the second-order structure of the original process $X_{t,T}$ must be reflected as a breakpoint in at least one of the asymptotic wavelet periodogram expectations $\beta_i(z)$, $i = -1, \ldots, -I^*$, and vice versa: a breakpoint in one of the $\beta_i(z)$'s implies a breakpoint in the second-order structure of $X_{t,T}$. Thus, it is sensible to combine the estimated breakpoints across the multiple scales of wavelet periodograms by, roughly speaking, selecting a breakpoint as significant if it appears in any of the wavelet periodogram sequences. In what follows, we first provide an algorithm which performs the selection of the final set of breakpoints as above, and show that it extends the within-scale consistency results in Theorem 3.1 to the original time series.

Let $\hat{\mathcal{B}}_i$ be the set of detected breakpoints from the sequence $I^{(i)}_{t,T}$, i.e.

$$\hat{\mathcal{B}}_i = \left\{ \hat{\eta}^{(i)}_p : \ p = 1, \ldots, \hat{N}_i \right\}, \quad i = -1, \ldots, -I^*.$$

Then the post-processing finds a subset of $\cup_{i=-1}^{-I^*} \hat{\mathcal{B}}_i$, say $\hat{\mathcal{B}}$, as formulated below.

Across-scales post-processing algorithm

Step 1 Arrange all the breakpoints into groups so that those from different sequences and within the distance of Λ_T from each other are categorised as belonging to the same group, and denote the groups by $\mathcal{G}_1, \ldots, \mathcal{G}_{\hat{B}}$.

Step 2 Find the finest scale with the most breakpoints as

$$i_0 = \max \left\{ i : \ \arg \max_{-I^* \le i \le -1} \hat{N}_i \right\}.$$

Step 3 Check whether there exists $\hat{\eta}^{(i_0)}_{p_0}$ which satisfies

$$\left| \hat{\eta}^{(i)}_p - \hat{\eta}^{(i_0)}_{p_0} \right| < \Lambda_T,$$

for every $\hat{\eta}^{(i)}_p$ with $i \neq i_0$; $1 \le p \le \hat{N}_i$. In other worlds, check whether $\hat{N}_{i_0} = \hat{B}$ and $\hat{\mathcal{B}}_{i_0}$ contains a member of each group $\mathcal{G}_1, \ldots, \mathcal{G}_{\hat{B}}$. If so, let

$\hat{\mathcal{B}} = \hat{\mathcal{B}}_{i_0}$ and quit the across-scales post-processing.

Step 4 Otherwise, choose the final set of detected breakpoints as

$$\hat{\mathcal{B}} = \left\{ \hat{\nu}_p : \ p = 1, \ldots, \hat{B} \right\},$$

where each $\hat{\nu}_p \in \mathcal{G}_p$ is chosen as the member of the group with the maximum i (finest scale).

We set $\Lambda_T = O(\epsilon_T)$ in order to take into account the bias between $\hat{\eta}_p$ and η_p which arises in deriving the results of Theorem 3.1. As argued previously, breakpoints detected at coarser scales are likely to be less accurate than those detected at finer scales, and thus Step 4 of the above post-processing algorithm prefers the latter. The across-scales post-processing procedure preserves the number of distinct breakpoints as well as their locations determined by the binary segmentation algorithm. Hence the breakpoints in the final set $\hat{\mathcal{B}}$ are still consistent estimates of the true breakpoints in the second-order structure of the original process $X_{t,T}$. Although it is not the only possible way of combining the breakpoints across scales consistently with our theory, the above post-processing algorithm shows good practical performance in our simulation study.

We denote the set of the true breakpoints in the second-order structure of $X_{t,T}$ by $\mathcal{B} = \{\nu_p : \ p = 1, \ldots, B\}$ (with a slight abuse of notation; recall Assumption 3.1), and the finally selected breakpoints from the across-scales post-processing by $\hat{\mathcal{B}} = \left\{ \hat{\nu}_p : \ p = 1, \ldots, \hat{B} \right\}$. Then the following theorem states the consistency of our breakpoint detection methodology, which consists of the binary segmentation algorithm and two post-processing steps.

Theorem 3.2. *Suppose that $X_{t,T}$ satisfies Assumption 3.1, and that the breakpoints ν_p's in $\mathcal{B}$ satisfy the same condition as that required of η_p's in Assumption 3.2. Further assume that the conditions on $\sigma^2(z)$ in Theorem 3.1 hold for each $\beta_i(z)$. Then, the breakpoints detected as in $\hat{\mathcal{B}}$ are consistent in terms of their total number and locations. That is, for an arbitrary positive constant C,*

$$\mathbb{P}\left\{ \hat{B} = B : \ |\hat{\nu}_p - \nu_p| \le C\epsilon_T, \ 1 \le p \le B \right\} \to 1$$

as $T \to \infty$.

3.2.4 Choice of Δ_T, θ, τ and I^*

To ensure that each estimated segment is of sufficiently large length so as not to distort our theoretical results, Δ_T is chosen to satisfy $\Delta_T \geq C\epsilon_T$ for some $C > 0$. However, in practice our method works well for smaller values of Δ_T too, e.g. in the forthcoming simulation experiments, $\Delta_T = C\sqrt{T}$ is used.

As for the choice of θ which is constrained to be within $(1/4, 1/2)$, we use $\theta = 0.251$, since we have found that the method works best when θ is close to the lower end of its permitted range. Instead, we elaborate on the choice of τ as below, although our asymptotic theoretical results hold for any fixed positive τ.

The selection of τ is not a straightforward task, and to get some insight into its choice, a set of numerical experiments was conducted. A vector of random variables was generated as $\mathbf{x} \sim \mathcal{N}_T(0, \Sigma)$, where $\mathbf{x} = (X_1, \ldots, X_T)^T$, and transformed into sequences of wavelet periodograms $I_{t,T}^{(i)}$. The covariance matrix was chosen as $\Sigma = \left(\rho^{|i-j|}\right)_{i,j=1}^{T}$ such that with varying ρ, the level of correlations among the variables $X_t,\ t = 1, \ldots, T$ was controlled. Then we found $\nu \in (1, T)$ which maximised

$$\mathbb{I}_i^b = \left| \sqrt{\frac{T-b}{T \cdot b}} \sum_{t=1}^{b} I_{t,T}^{(i)} - \sqrt{\frac{b}{T(T-b)}} \sum_{t=b+1}^{T} I_{t,T}^{(i)} \right|, \quad b \in (1, T),$$

and computed

$$\mathbb{U}_{i,\rho,T} = \mathbb{I}_i^{\nu} \cdot \{T^{-1} \sum_{t=1}^{T} I_{t,T}^{(i)} \cdot T^{\theta} \sqrt{\log T}\}^{-1}.$$

This was repeatedly conducted with varying choice of the covariance matrix ($\rho = 0, 0.3, 0.6, 0.9$) and sample size ($T = 512, 1024, 2048$), 100 times for each combination.

The quantity $\mathbb{U}_{i,\rho,T}$ is the ratio between our test statistic and the data size-dependent factor $T^{\theta}\sqrt{\log T}$, which appears in our threshold defined in the algorithm of Section 3.2.2. $\mathbb{U}_{i,\rho,T}$ is computed under the null hypothesis of constant second-order structure of X_t, and thus its magnitude serves as a guideline so as to the choice of τ for each scale i, preventing spurious breakpoint detection in the null hypothesis case.

The results showed that the values of $\mathbb{U}_{i,\rho,T}$ and their ranges tended to increase

Table 3.1: Values of τ for each scale $i = -1, \ldots, -4$.

scale i	-1	-2	-3	-4
$\tau_{i,1}$	0.39	0.46	0.67	0.83
$\tau_{i,2}$	0.48	0.52	0.75	0.96

for coarser scales, due to the increasing dependence in the wavelet periodogram sequences. In comparison to the scale factor i, the parameters ρ or T had relatively little impact on $\mathbb{U}_{i,\rho,T}$.

Based on the above numerical experiments, we propose to use different values of scale-dependent τ_i, in Step 2 of the binary segmentation algorithm (Section 3.2.2), and in the within-scale post-processing procedure (Section 3.2.2.1). Denoting the former by $\tau_{i,1}$ and the latter by $\tau_{i,2}$, we choose $\tau_{i,1}$ as the 95% quantile, and $\tau_{i,2}$ as the 97.5% quantile of $\mathbb{U}_{i,T}$ for given i and T. The disappearance of ρ in the subscript indicates that $\mathbb{U}_{i,\rho,T}$ for different values of ρ are all combined. By way of example, the values of τ when $T = 1024$ are summarised in Table 3.1.

Finally, we discuss the choice of I^*, the coarsest wavelet periodogram scale at which we still apply our breakpoint detection procedure. The default choice of I^* is given as $\lfloor \log_2 T/3 \rfloor$. Therefore, we first detect breakpoints in wavelet periodograms at scales $i = -1, \ldots, -\lfloor \log_2 T/3 \rfloor$, and perform the across-scale post-processing as described in Section 3.2.3.1, to obtain the set of breakpoints

$$\hat{\mathcal{B}} = \left\{ \hat{\nu}_p : \ p = 1, \ldots, \hat{B} \right\}.$$

Then, for the wavelet periodogram at the next finest scale $i = -(\lfloor \log_2 T/3 \rfloor + 1)$, we compute the quantities $\mathbb{V}_p, \ p = 1, \ldots, \hat{B} + 1$ as

$$\mathbb{V}_p = \max_{b \in (\hat{\nu}_{p-1}, \hat{\nu}_p)} \left| \frac{\sqrt{\frac{\hat{\nu}_p - b}{(\hat{\nu}_p - \hat{\nu}_{p-1}) \cdot (b - \hat{\nu}_{p-1})}} \sum_{t=\hat{\nu}_{p-1}+1}^{b} I_{t,T}^{(i)} - \sqrt{\frac{b - \hat{\nu}_{p-1}}{(\hat{\nu}_p - \hat{\nu}_{p-1}) \cdot (\hat{\nu}_p - b)}} \sum_{t=b+1}^{\hat{\nu}_p} I_{t,T}^{(i)}}{\sum_{\hat{\nu}_{p-1}+1}^{\hat{\nu}_p} I_{t,T}^{(i)} / (\hat{\nu}_p - \hat{\nu}_{p-1})} \right|$$

where $\hat{\nu}_0 = -1$ and $\hat{\nu}_{\hat{B}+1} = T - 1$. Note that $\mathbb{V}_p$ is again of the same form as our basic test statistic in Step 2 of the binary segmentation algorithm.

Then each $\mathbb{V}_p$ is compared to $\tau_{i,1} \cdot T^{\theta} \sqrt{\log T}$ to see whether there are any further breakpoints yet to be detected in $I_{t,T}^{(i)}$ which have not been included in $\hat{\mathcal{B}}$.

If any $\mathbb{V}_p$ exceeds the threshold, I^* is updated as $I^* \leftarrow I^* + 1$, and we apply our breakpoint detection methodology to the wavelet periodogram at the updated scale $-I^*$, eventually updating $\hat{\mathcal{B}}$. This procedure is repeatedly conducted until either no further changes are made, or I^* reaches $I^* = \lfloor \log_2 T/2 \rfloor$.

We note the similarity between this approach and the within-scale post-processing. Both make use of the test statistics which are of the same form as the basic test statistic of the binary segmentation procedure (in the former for determining the progression to the next finest scale, while in the latter for checking the validity of detected breakpoints within each scale). Thus this procedure for updating I^* is also in line with the theoretical consistency of our breakpoint detection procedure. That is, $\mathbb{V}_p$ being of the same form as the test statistic of our binary segmentation algorithm, Lemma 3.6 in Section 3.5.1 implies that, if there are no more breakpoints to be detected from $I_{t,T}^{(i)}$ for $i < -I^*$ other than those already included in $\hat{\mathcal{B}}$, then $\mathbb{V}_p$ does not exceed the threshold, and vice versa by Lemma 3.5.

3.3 Simulation study

In Davis *et al.* (2006), the performance of the Auto-PARM procedure was assessed and compared with the Auto-SLEX procedure (Ombao *et al.*, 2001) through simulation in various settings (for a brief description of both methods, see Section 2.3). They reported the superior performance of the Auto-PARM in identifying both dyadic and non-dyadic breakpoints in piecewise stationary time series. Some examples in the following are adopted from their paper for the comparative study between the methodology developed in this chapter and the Auto-PARM, alongside some other new examples.

We also applied the breakpoint detection method based on minimising a penalised Gaussian likelihood (Lavielle & Teyssière, 2005, referred to as L&T henceforth) to the same simulated processes, using the Matlab code available on `http://www.math.u-psud.fr/~lavielle/programmes_lavielle.html`. Overall, the performance of L&T was found to be inferior to that of both Auto-PARM and our method for these particular examples. Therefore, the results from L&T are reported in Tables 3.2–3.3, yet comments on the behaviour of breakpoint

detection procedures for specific simulation models are restricted to our method and the Auto-PARM.

In the simulations below, Haar wavelets were used to compute wavelet periodograms. The number of observations in all examples was $T = 1024$. Therefore the default value for I^* was set as $3(= \lfloor \log_2 T/3 \rfloor)$ at the start of each application of our methodology, and then I^* was updated automatically if necessary, as described in Section 3.2.4. Simulation outcome from (A) is given in Table 3.2 and that from the rest of models in Table 3.3, which report the total number of detected breakpoints over 100 simulations.

(A) Stationary AR(1) process with no breakpoints

We consider a stationary AR(1) process,

$$X_t = aX_{t-1} + \epsilon_t \quad \text{for } 1 \leq t \leq 1024,$$

where $\epsilon_t \sim_{\text{i.i.d.}} \mathcal{N}(0, 1)$ (as in all the following examples unless specified otherwise). This model is chosen to evaluate the performance of breakpoint detection methods in not returning any "false alarm" under the null hypothesis of constant second-order structure. For a range of values of a, we summarise the breakpoint detection outcome in Table 3.2.

(B) Piecewise stationary AR process with clearly observable changes

This example is taken from Davis *et al.* (2006). The target nonstationary process is generated from the model below,

$$X_t = \begin{cases} 0.9X_{t-1} + \epsilon_t & \text{for } 1 \leq t \leq 512, \\ 1.68X_{t-1} - 0.81X_{t-2} + \epsilon_t & \text{for } 513 \leq t \leq 768, \\ 1.32X_{t-1} - 0.81X_{t-2} + \epsilon_t & \text{for } 769 \leq t \leq 1024, \end{cases}$$

where the AR parameters change over time in a piecewise constant manner. As seen in Figure 3.1 (a), there is a clear visual difference between the three stationary segments in this model. Figure 3.1 (b) shows the wavelet periodogram at scale -4 and the estimation results, where the dotted lines indicate the true breakpoints ($\eta_1 = 512$, $\eta_2 = 768$) while the dashed lines indicate the detected ones ($\hat{\eta}_1 = 519$, $\hat{\eta}_2 = 764$). Note that although

initially the binary segmentation algorithm returned three breakpoints, the within-sequence post-processing successfully removed the false one.

(C) Piecewise stationary AR process with less clearly observable changes

In this example, the piecewise stationary AR model in (B) is revisited, but its breakpoints are less clear-cut, as seen in Figure 3.2.

$$X_t = \begin{cases} 0.4X_{t-1} + \epsilon_t & \text{for } 1 \leq t \leq 400, \\ -0.6X_{t-1} + \epsilon_t & \text{for } 401 \leq t \leq 612, \\ 0.5X_{t-1} + \epsilon_t & \text{for } 613 \leq t \leq 1024. \end{cases}$$

Figure 3.2 (b) shows the wavelet periodogram at scale -1 for the realisation in the left panel with its breakpoint estimates ($\hat{\eta}_1 = 403$, $\hat{\eta}_2 = 622$). Table 3.3 shows that both our method and the Auto-PARM achieved good performance, identifying exactly two breakpoints in over 95% of the cases.

(D) Piecewise stationary AR process with a short segment

This example is again from Davis *et al.* (2006), which is designed such that there is a single breakpoint and one resulting segment is much shorter than the other.

$$X_t = \begin{cases} 0.75X_{t-1} + \epsilon_t & \text{for } 1 \leq t \leq 50, \\ -0.5X_{t-1} + \epsilon_t & \text{for } 51 \leq t \leq 1024. \end{cases}$$

A typical realisation of the above model, its wavelet periodogram at scale -3 and the estimation outcome are shown in Figure 3.3, where the jump at $\eta_1 = 50$ was identified as $\hat{\eta}_1 = 49$. Even though one segment is substantially shorter than the other, our procedure was able to detect exactly one breakpoint in 97% of the cases and underestimation did not occur even when it failed to detect exactly one.

(E) Piecewise stationary near-unit-root process with changing variance

Financial time series, such as stock indices, individual share or commodity prices, or currency exchange rates, are for certain purposes (such as e.g. pricing of derivative instruments) often modelled as a random walk with a time-varying variance. Motivated by this, we generated a piecewise sta-

tionary, near-unit-root example following the model below, where its AR parameter, being very close to 1, remains constant, while the variance has two breakpoints over time. Note that within each stationary segment, the process can be seen as a special case of the local stationary alternative to a unit-root process (Phillips & Perron, 1988),

$$X_t = (1 + c/T)X_{t-1} + \epsilon_t \text{ with } c < 0. \tag{3.8}$$

A typical realisation generated from this model is given in Figure 3.4 (a).

$$X_t = \begin{cases} 0.999X_{t-1} + \epsilon_t, & \epsilon_t \sim \text{i.i.d.}\mathcal{N}(0,1) & \text{for } 1 \leq t \leq 400, \\ 0.999X_{t-1} + \epsilon_t, & \epsilon_t \sim \text{i.i.d.}\mathcal{N}(0,1.5^2) & \text{for } 401 \leq t \leq 750, \\ 0.999X_{t-1} + \epsilon_t, & \epsilon_t \sim \text{i.i.d.}\mathcal{N}(0,1) & \text{for } 751 \leq t \leq 1024. \end{cases}$$

Recall that the Auto-PARM was designed to find the best fitting AR model for a given time series, by adopting an algorithm which mimicked the process of natural evolution. However, due to the stochastic nature of this algorithm, the Auto-PARM occasionally fails to return consistent estimates. This instability was emphasised in this example, as each run of the Auto-PARM often returned different breakpoints. For one typical realisation, it detected $t = 21,\ 797$ as breakpoints and then only $t = 741$ in the next run on the same sample path.

Overall, our method performed better than the Auto-PARM for this particular example, and here we briefly discuss the reasons behind its good performance. Note that it was at scale -1 of the wavelet periodogram that both breakpoints were consistently identified the most frequently by our procedure. The computation of the wavelet periodogram at scale -1 with Haar wavelets is a differencing operation, and naturally "whitens" the near-unit-root process in this example to clearly reveal any changes of variance in the sequence.

(F) Piecewise stationary AR process with high autocorrelation

The features of the following AR model are: high degree of autocorrelation and less obvious breakpoints compared to previous examples. Its typical

realisation is shown in Figure 3.5 (a).

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

Again, the instability of Auto-PARM was notable for this example, with the second breakpoint at $t = 750$ often left undetected. Our procedure correctly identified both breakpoints in 84% of the cases.

(G) Piecewise stationary ARMA$(1, 1)$ process

In this simulation study, we generated piecewise stationary ARMA processes from the following model,

$$X_t = \begin{cases} 0.7X_{t-1} + \epsilon_t + 0.6\epsilon_{t-1} & \text{for } 1 \leq t \leq 125, \\ 0.3X_{t-1} + \epsilon_t + 0.3\epsilon_{t-1} & \text{for } 126 \leq t \leq 532, \\ 0.9X_{t-1} + \epsilon_t & \text{for } 533 \leq t \leq 704, \\ 0.1X_{t-1} + \epsilon_t - 0.5\epsilon_{t-1} & \text{for } 705 \leq t \leq 1024. \end{cases}$$

As illustrated in Figure 3.6 (a), the first breakpoint $t = 125$ is less apparent than the other two. The Auto-PARM procedure often left this breakpoint undetected, while our procedure found all three in 76% of cases.

We note that it was scale $i = -4$ at which $t = 125$ was detected most frequently by our procedure. With a time series of length $T = 1024$, default scales provided by our algorithm are $i = -1, -2, -3$, and therefore this example demonstrates the effectiveness of the updating procedure for I^* described in Section 3.2.4. That is, after completing the examination of $I^{(i)}_{t,T}$ for $i = -1, -2, -3$, our procedure checked if there were more breakpoints to be detected from $I^{(i)}_{t,T}$ for the next finest scale $i = -4$, and since it was the case, updated I^* to 4. Figure 3.6 (b) shows the wavelet periodogram at scale -4 for the time series example in the left panel.

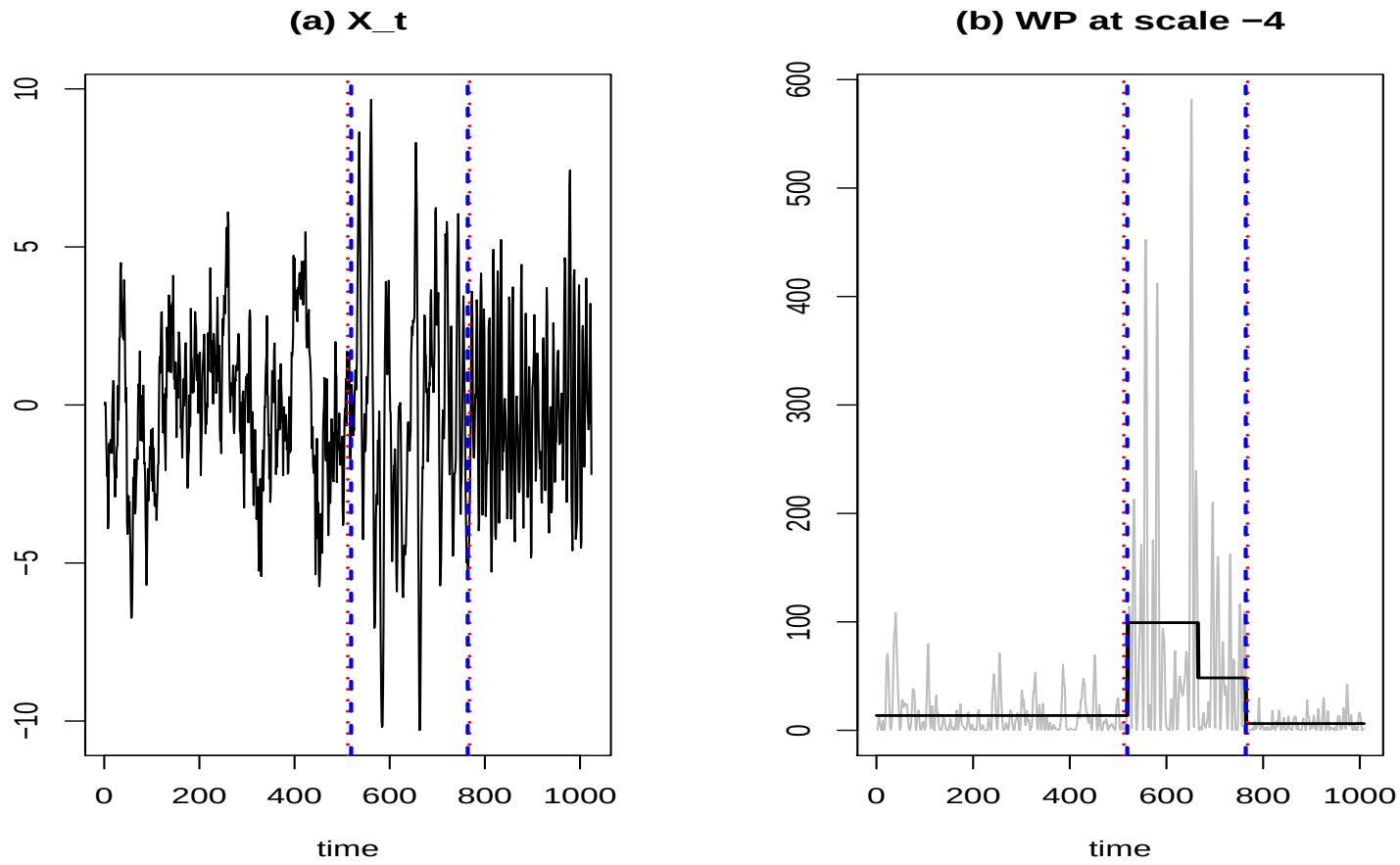


Figure 3.1: (a) A realisation of model (B), and true (red dotted) and detected (blue dashed) breakpoints; (b) $I_{t,T}^{(i)}$ at $i = -4$, its estimate (solid) and true (red dotted) and detected (blue dashed) breakpoints.

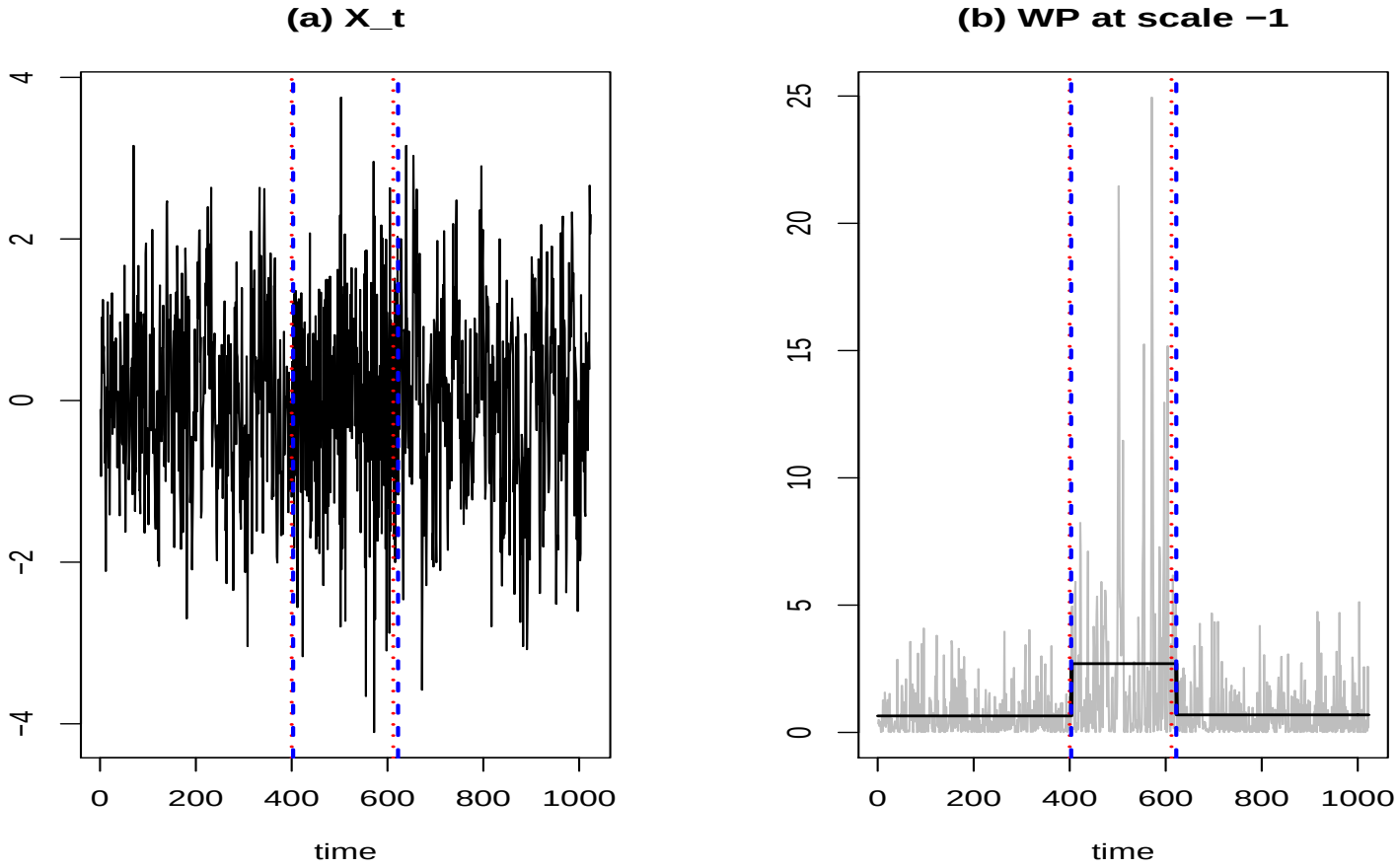


Figure 3.2: (a) A realisation of model (C), and true (red dotted) and detected (blue dashed) breakpoints; (b) $I^{(i)}_{t,T}$ at $i = -1$, its estimate (solid) and true (red dotted) and detected (blue dashed) breakpoints.

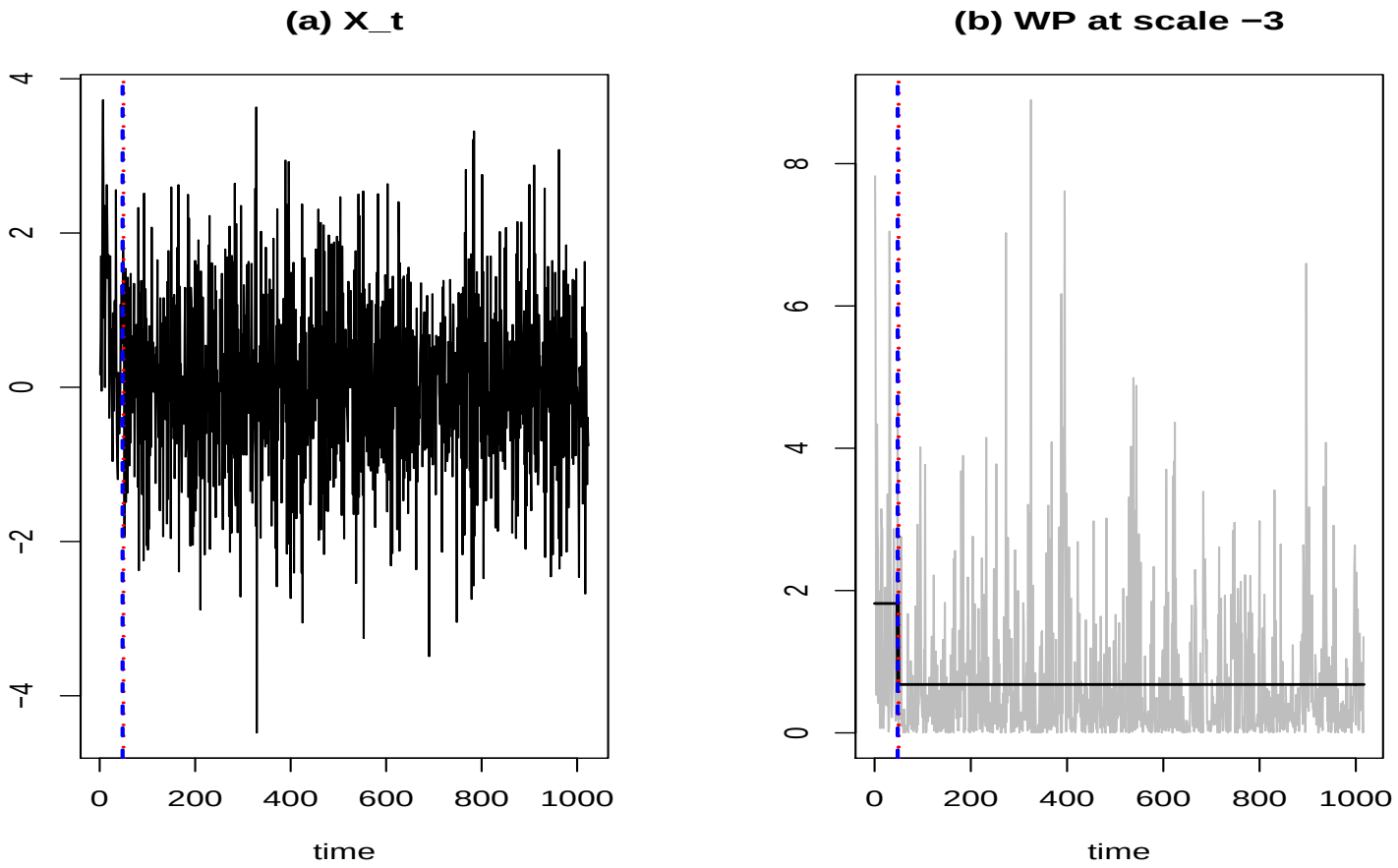


Figure 3.3: (a) A realisation of model (D), and true (red dotted) and detected (blue dashed) breakpoints; (b) $I^{(i)}_{t,T}$ at $i = -3$, its estimate (solid) and true (red dotted) and detected (blue dashed) breakpoints.

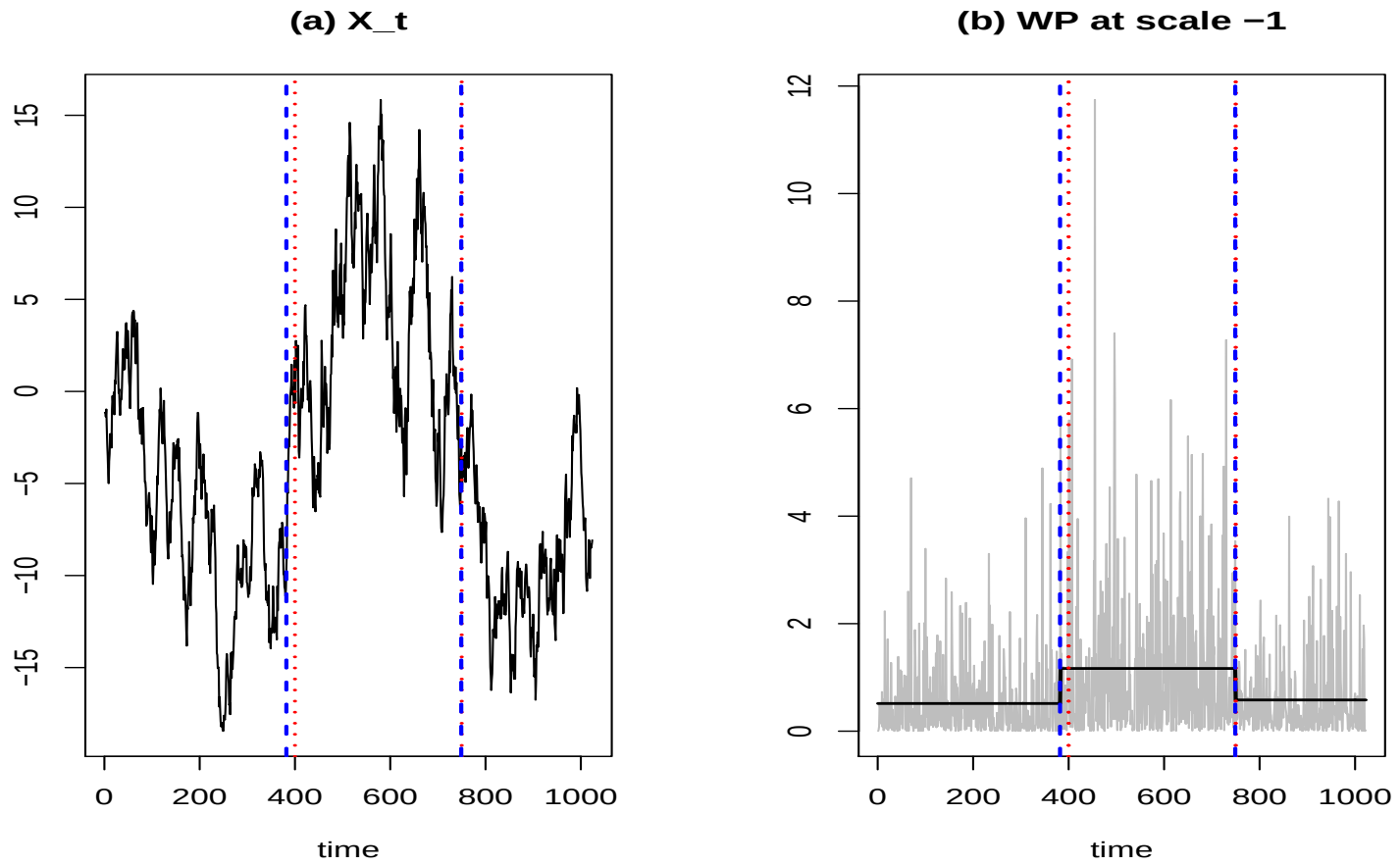


Figure 3.4: (a) A realisation of model (E), and true (red dotted) and detected (blue dashed) breakpoints; (b) $I_{t,T}^{(i)}$ at $i = -1$, its estimate (solid) and true (red dotted) and detected (blue dashed) breakpoints.

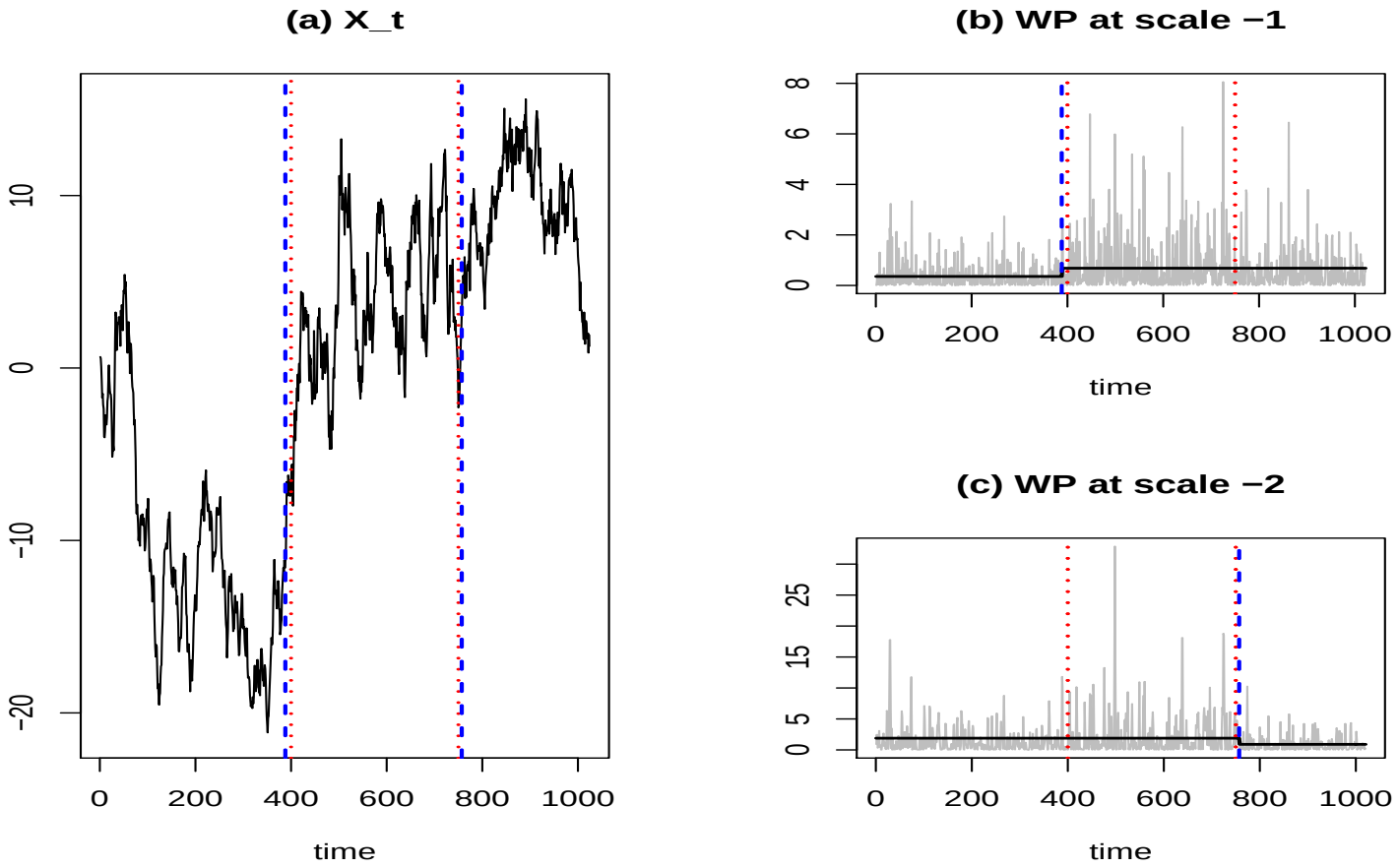


Figure 3.5: (a) A realisation of model (F), and true (red dotted) and detected (blue dashed) breakpoints; (b) $I_{t,T}^{(i)}$ at $i = -1$, its estimate (solid) and true (red dotted) and detected (blue dashed) breakpoints; (c) $I_{t,T}^{(i)}$ at $i = -2$, its estimate (solid) and true (red dotted) and detected (blue dashed) breakpoints.

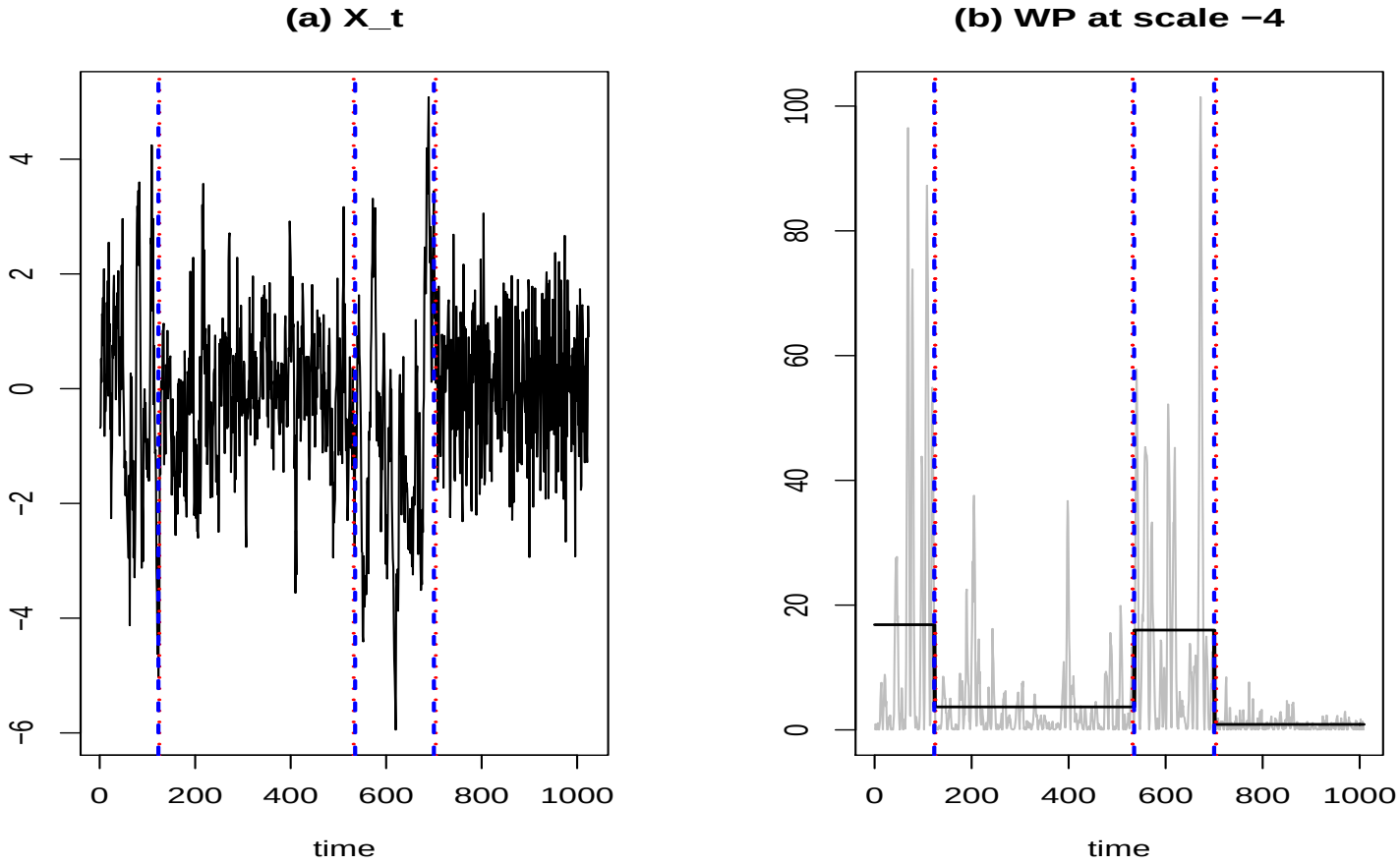


Figure 3.6: (a) A realisation of model (G), and true (red dotted) and detected (blue dashed) breakpoints; (b) $I_{t,T}^{(i)}$ at $i = -4$, its estimate (solid) and true (red dotted) and detected (blue dashed) breakpoints.

Table 3.2: Summary of breakpoint detection from Simulation (A); Our segmentation method (Seg), Auto-PARM (AP) and Lavielle & Teyssière (2005). Results over 100 simulations.

	number of breakpoints																	
a		0.7			0.4			0.1			-0.1			-0.4			-0.7	
	Seg	AP	L&T	Seg	AP	L&T	Seg	AP	L&T	Seg	AP	L&T	Seg	AP	L&T	Seg	AP	L&T
0	**100**	**100**	**70**	**100**	**100**	**80**	**100**	**100**	**80**	**99**	**100**	**73**	**99**	**100**	**71**	**94**	**100**	**64**
1	0	0	0	0	0	0	0	0	1	1	0	3	1	0	2	5	0	1
≥ 2	0	0	30	0	0	20	0	0	19	0	0	24	0	0	27	1	0	35
total	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100

Table 3.3: Summary of breakpoint detection from Simulations (B)–(G); Our segmentation method (Seg), AutoPARM (AP) and Lavielle & Teyssière (2005). Results over 100 simulations.

	number of breakpoints																	
	model (B)			model (C)			model (D)			model (E)			model (F)			model (G)		
	Seg	AP	L&T	Seg	AP	L&T	Seg	AP	L&T	Seg	AP	L&T	Seg	AP	L&T	Seg	AP	L&T
0	0	0	1	0	0	59	0	0	55	1	42	87	1	20	5	0	0	0
1	0	0	1	0	0	2	**97**	**100**	**24**	0	31	0	14	68	9	1	16	1
2	**93**	**99**	**51**	**96**	**100**	**20**	3	0	3	**97**	**16**	**1**	**84**	**7**	**8**	6	55	5
3	4	1	12	3	0	6	0	0	5	2	9	0	1	3	11	**76**	**29**	**43**
4	3	0	10	1	0	4	0	0	1	0	0	0	0	1	16	17	0	10
≥ 5	0	0	25	0	0	9	0	0	12	0	2	12	0	1	51	0	0	41
total	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100

3.4 U.S stock market data analysis

Many authors, including Stărică & Granger (2005), argued in favour of nonstationary modelling of financial returns. In this section, we analyse the Dow Jones Industrial Average index by regarding it as a process with an extremely high degree of autocorrelation (such as in the near-unit-root model of Phillips & Perron (1988), see (3.8)) and a time-varying variance, similar to the simulation model in Section 3.3 (E).

(A) Dow Jones weekly closing values 1970–1975

The time series of weekly closing values of the Dow Jones Industrial Average index between July 1971 and August 1974 was studied in Hsu (1979) and revisited by Chen & Gupta (1997). Historical data are available on `www.google.com/finance/historical?q=INDEXDJX:.DJI`, where daily and weekly prices can be extracted for any time period. Both papers concluded that there was a change in the variance of the index around the third week of March 1973.

For the ease of computation of the wavelet periodograms, we chose the same weekly index between 1 July 1970 and 19 May 1975 so that the data size was $T = 256$ and the aforementioned time period was contained in this interval. In this dataset, the third week of March 1973 corresponds to $t = 141$ and our procedure detected $\hat{\eta} = 142$ as a breakpoint, as illustrated in Figure 3.8.

As for the other breakpoint detection method used in our simulation study, the Auto-PARM did not return any breakpoint. Since Lavielle & Teyssière (2005), when analysing financial time series, applied their segmentation procedure (L&T) to the *log-returns* ($\log(X_t/X_{t-1})$) of the data rather than the original data X_t themselves, we followed this practice and applied L&T to the log-returns of the Dow Jones data. It returned $t = 141$ as a breakpoint, which is very close to $\hat{\eta}$ detected by our procedure.

(B) Dow Jones daily closing values 2007–2009

We further investigated more recent, daily data from the same source, be-

tween 8 January 2007 and 16 January 2009. Over this period, the global financial market experienced one of the worst crises in history.

Our breakpoint detection algorithm estimated two breakpoints (see Figure 3.9), one in the last week of July 2007 ($\hat{\eta}_1 = 135$), and the other in mid-September 2008 ($\hat{\eta}_2 = 424$). The Auto-PARM returned three breakpoints on average, although the estimated breakpoints were unstable as noted in Section 3.3 (E). $t = 35,\ 426,\ 488$ were detected most frequently as breakpoints, and $t = 100$ or $t = 140$ were detected in place of $t = 35$ on other occasions. L&T, when applied to the log-returns ($\log(X_t/X_{t-1})$) of the data as in the above (A), detected $t = 127,\ 424$ as breakpoints, which are very close to $\hat{\eta}_1$ and $\hat{\eta}_2$ by our method.

The first breakpoint $\hat{\eta}_1$ coincided with the outbreak of the worldwide "credit crunch", as subprime mortgage backed securities were discovered in portfolios of banks and hedge funds around the world. The second breakpoint $\hat{\eta}_2$ coincided with the bankruptcy of Lehman Brothers, a major financial services firm, an event which brought even more volatility to the market. One evidence supporting our breakpoint detection outcome is the TED spread (available on http://www.bloomberg.com/apps/quote?ticker=.tedsp:ind), which is an indicator of perceived credit risk in the general economy. As shown in Figure 3.7, it spiked up in late July 2007, remained volatile for a year, then spiked even higher in September 2008, and these movements coincide almost exactly with our detected breakpoints.

3.5 Proofs

3.5.1 The proof of Theorem 3.1

The consistency of our algorithm is first proved for the sequence below,

$$\tilde{Y}^2_{t,T} = \sigma^2(t/T) \cdot Z^2_{t,T},\ t = 0, \ldots, T-1, \tag{3.9}$$

where the true piecewise constant function $\sigma^2(t/T)$ replaces $\sigma^2_{t,T}$ in (3.3).

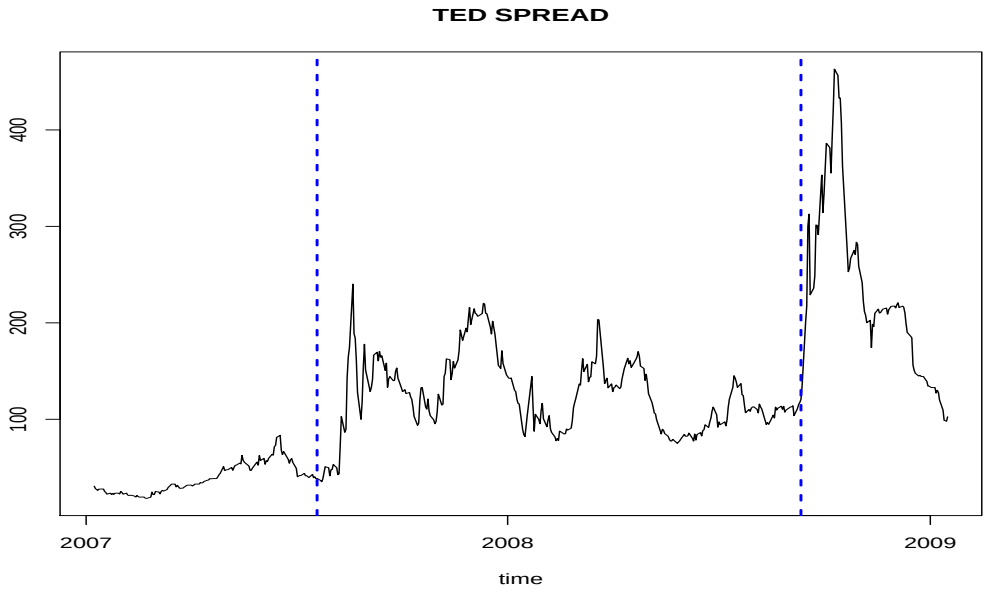


Figure 3.7: TED spread between January 2007 and January 2009 and the breakpoints detected by our procedure (blue dashed).

We denote $n = e - s + 1$ and define

$$\begin{aligned}
\tilde{\mathbb{Y}}^b_{s,e} &= \sqrt{\frac{e-b}{n(b-s+1)}} \sum_{t=s}^{b} \tilde{Y}^2_{t,T} - \sqrt{\frac{b-s+1}{n(e-b)}} \sum_{t=b+1}^{e} \tilde{Y}^2_{t,T}, \\
\tilde{\mathbb{S}}^b_{s,e} &= \sqrt{\frac{e-b}{n(b-s+1)}} \sum_{t=s}^{b} \sigma^2(t/T) - \sqrt{\frac{b-s+1}{n(e-b)}} \sum_{t=b+1}^{e} \sigma^2(t/T), \text{ and} \\
\mathbb{S}^b_{s,e} &= \sqrt{\frac{e-b}{n(b-s+1)}} \sum_{t=s}^{b} \sigma^2_{t,T} - \sqrt{\frac{b-s+1}{n(e-b)}} \sum_{t=b+1}^{e} \sigma^2_{t,T}.
\end{aligned}$$

Note that these quantities are simply inner products of the respective sequences and a vector

$$(\underbrace{0,\cdots,0}_{s-1}, \underbrace{\sqrt{\tfrac{e-b}{n(b-s+1)}},\cdots,\sqrt{\tfrac{e-b}{n(b-s+1)}}}_{b-s+1}, \underbrace{-\sqrt{\tfrac{b-s+1}{n(e-b)}},\cdots,-\sqrt{\tfrac{b-s+1}{n(e-b)}}}_{e-b}, \underbrace{0,\cdots,0}_{T-e})^T,$$

whose support starts at s, is constant and positive until b, then constant negative until e and normalised such that it sums to zero and sums to one when squared.

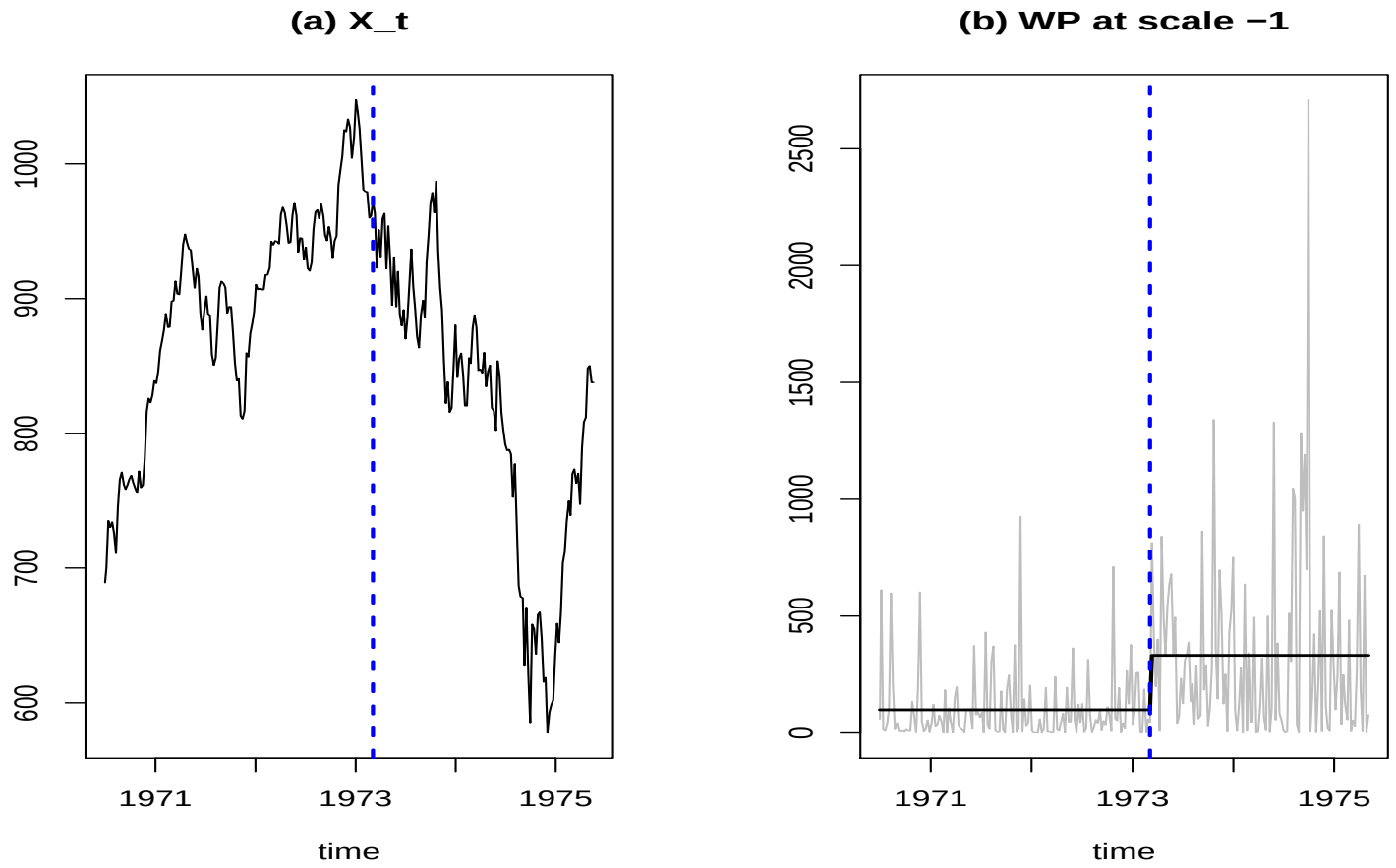


Figure 3.8: (a) Weekly average values of the Dow Jones IA index (July 1970–May 1975); (b) Wavelet periodogram at scale -1, its estimate (solid) and a detected breakpoint (blue dashed).

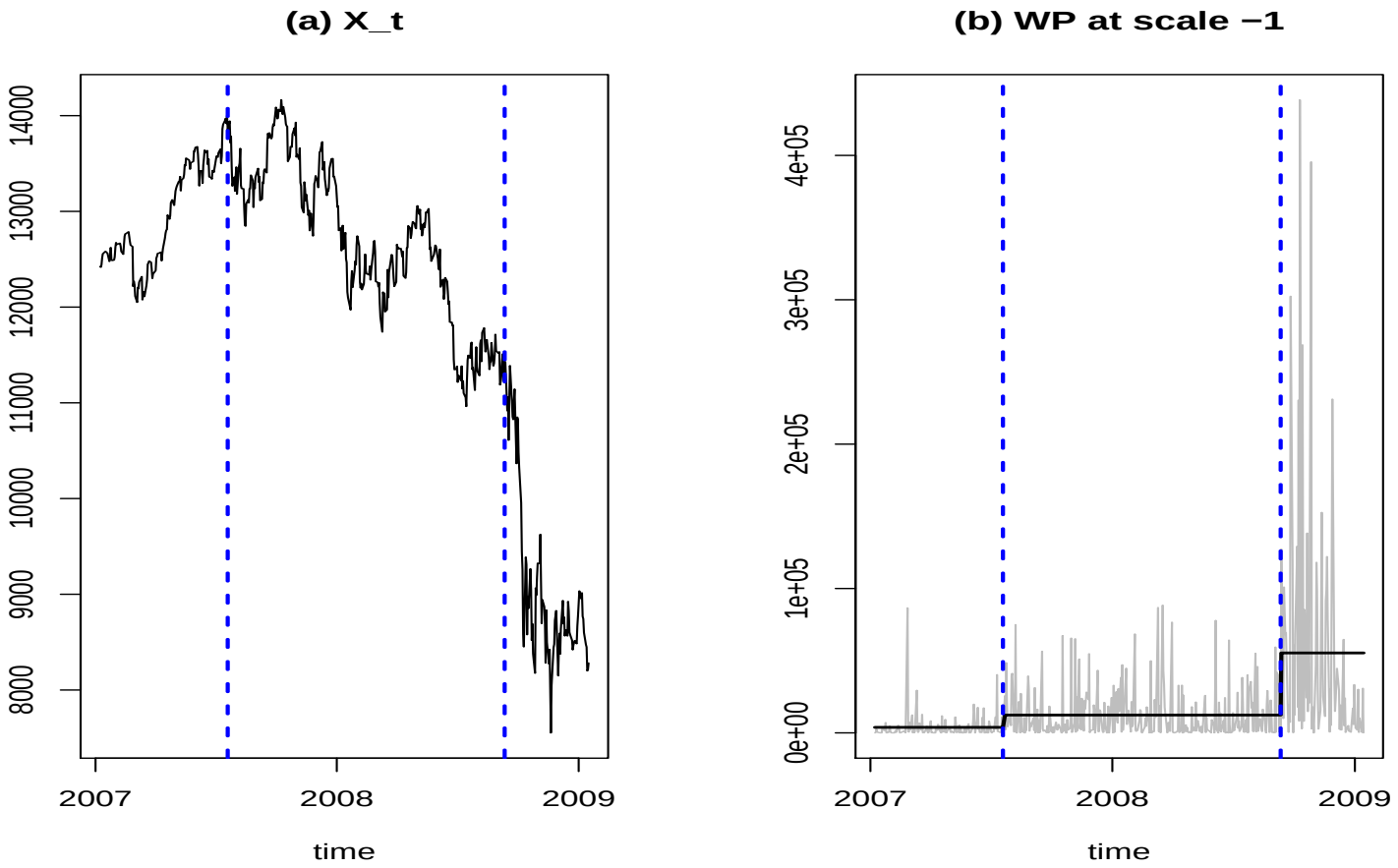


Figure 3.9: (a) Daily average values of the Dow Jones IA index (Jan 2007–Jan 2009); (b) Wavelet periodogram at scale −1, its estimate (solid) and detected breakpoints (blue dashed).

Let s, e satisfy

$$\eta_{p_0} \leq s < \eta_{p_0+1} < \ldots < \eta_{p_0+q} < e \leq \eta_{p_0+q+1}$$

for $0 \leq p_0 \leq B - q$. It is shown throughout the proof that this will always be the case at all stages of the algorithm. In Lemmas 3.1–3.5 below, we impose at least one of following conditions:

$$s < \eta_{p_0+r} - C\delta_T < \eta_{p_0+r} + C\delta_T < e \text{ for some } 1 \leq r \leq q, \tag{3.10}$$

$$\{(\eta_{p_0+1} - s) \wedge (s - \eta_{p_0})\} \vee \{(\eta_{p_0+q+1} - e) \wedge (e - \eta_{p_0+q})\} \leq C\epsilon_T, \tag{3.11}$$

where $\wedge$ and $\vee$ are the minimum and maximum operators, respectively, and C denotes an arbitrary positive constant (as in what follows unless specified otherwise).

Recall that throughout the algorithm, a segment is defined by previously detected breakpoints s and e. Then (3.10) implies that when there is at least one true breakpoint within the segment (s, e) which has not been detected yet, it is of sufficient distance from both s and e. On the other hand, (3.11) implies that for each s and e, there exists a true breakpoint within a sufficiently short distance. In the proofs of following lemmas, it is shown that both conditions (3.10) and (3.11) hold throughout the algorithm for all those segments starting at s and ending at e. As Lemma 3.6 concerns the case when all breakpoint have already been detected, it does not use either of these conditions.

The proof of the Theorem is constructed as follows. Once Lemma 3.1 is shown, the result is used in the proof of Lemma 3.2, which in turn is used alongside Lemma 3.3 in the proof of Lemma 3.4. From the result of Lemma 3.4, we derive Lemma 3.5 and finally, Lemmas 3.5 and 3.6 are used to prove Theorem 3.1.

Lemma 3.1. *Let s and e satisfy (3.10), then there exists $1 \leq r^* \leq q$ such that*

$$\left|\tilde{\mathbb{S}}_{s,e}^{\eta_{p_0+r^*}}\right| = \max_{s<t<e} |\tilde{\mathbb{S}}_{s,e}^{t}|, \text{ and} \tag{3.12}$$

$$\left|\tilde{\mathbb{S}}_{s,e}^{\eta_{p_0+r^*}}\right| \geq C\delta_T/\sqrt{T}. \tag{3.13}$$

Proof. (3.12) is proved by Lemmas 2.2 and 2.3 of Venkatraman (1993). For

the inequality part in (3.13), we note that in the case of a single breakpoint in $\sigma^2(z)$, r in (3.10) coincides with r^* and we can use the constancy of $\sigma^2(z)$ to the left and to the right of the breakpoint to show that

$$\left|\tilde{\mathbb{S}}_{s,e}^{\eta_{p_0+r}}\right| = \left|\frac{\sqrt{\eta_{p_0+r}-s+1}\sqrt{e-\eta_{p_0+r}}}{\sqrt{n}}\left(\sigma^2\left(\frac{\eta_{p_0+r}}{T}\right)-\sigma^2\left(\frac{\eta_{p_0+r}+1}{T}\right)\right)\right|,$$

which is bounded from below by $C\delta_T/\sqrt{T}$. In the case of multiple breakpoints, we remark that for any r satisfying (3.10), the above order remains the same and thus (3.13) follows. □

Lemma 3.2. *Suppose s and e satisfy (3.10) and further assume that $\tilde{\mathbb{S}}_{s,e}^{\eta_{p_0+r}} > 0$ for some $1 \le r \le q$. Then for any b satisfying $|\eta_{p_0+r} - b| = C\epsilon_T$, we have*

$$\tilde{\mathbb{S}}_{s,e}^{\eta_{p_0+r}} \ge \tilde{\mathbb{S}}_{s,e}^{b} + 2\log T$$

for large T.

Proof. Without loss of generality, assume $\eta_{p_0+r} < b$. As done in Lemma 3.1, we first derive the result in the case of a single breakpoint in $\sigma^2(z)$. The following holds;

$$\tilde{\mathbb{S}}_{s,e}^{b} = \frac{\sqrt{\eta_{p_0+r}-s+1}\sqrt{e-b}}{\sqrt{e-\eta_{p_0+r}}\sqrt{b-s+1}}\tilde{\mathbb{S}}_{s,e}^{\eta_{p_0+r}}, \text{ and} \tag{3.14}$$

$$\begin{aligned}
\tilde{\mathbb{S}}_{s,e}^{\eta_{p_0+r}} - \tilde{\mathbb{S}}_{s,e}^{b} &= \left(1 - \frac{\sqrt{\eta_{p_0+r}-s+1}\sqrt{e-b}}{\sqrt{e-\eta_{p_0+r}}\sqrt{b-s+1}}\right)\tilde{\mathbb{S}}_{s,e}^{\eta_{p_0+r}} \\
&= \frac{\sqrt{1+\frac{b-\eta_{p_0+r}}{\eta_{p_0+r}-s+1}} - \sqrt{1-\frac{b-\eta_{p_0+r}}{e-\eta_{p_0+r}}}}{\sqrt{1+\frac{b-\eta_{p_0+r}}{\eta_{p_0+r}-s+1}}} \cdot \tilde{\mathbb{S}}_{s,e}^{\eta_{p_0+r}} \\
&\ge \frac{\left(1+\frac{c_1\epsilon_T}{2\delta_T}\right) - \left(1+\frac{c_2\epsilon_T}{2\delta_T}\right) + o\left(\frac{\epsilon_T}{\delta_T}\right)}{\sqrt{2}} \cdot \tilde{\mathbb{S}}_{s,e}^{\eta_{p_0+r}} \\
&\ge C\frac{\epsilon_T}{\delta_T} \cdot \frac{\delta_T}{\sqrt{T}} \ge 2\log T
\end{aligned}$$

for large T, where c_1 and c_2 are positive constants. The Taylor expansion

$$\sqrt{1+x} = 1 + \frac{1}{2}x - \frac{1}{8}x^2 + \frac{1}{16}x^3 - \cdots, \text{ for } |x| < 1,$$

is applied to derive the first inequality and Lemma 3.1 to the second inequality. Similar arguments are applicable when $b < \eta_{p_0+r}$. Since the order of (3.14) remains the same in the case of multiple breakpoints, the lemma is proved. □

Lemma 3.3. *There exists $C > 1$ such that, with probability converging to 1 with T,*

$$\left|\tilde{\mathbb{Y}}^b_{s,e} - \tilde{\mathbb{S}}^b_{s,e}\right| \leq C \log T$$

uniformly over $(s, b, e) \in \mathcal{D}$, where $\mathcal{D}$ is defined as

$$\mathcal{D} = \{(s,b,e) : \ 1 \leq s < b < e \leq T, \ n = e - s + 1 \geq C\delta_T \text{ and } \max\left\{\sqrt{\frac{b-s+1}{e-b}}, \sqrt{\frac{e-b}{s-b+1}}\right\} \leq c\}$$

for the same $c \geq 1$ used in Assumption 3.2.

Proof. We need to show that

$$\mathbb{P}\left(\max_{(s,b,e)\in\mathcal{D}} \frac{1}{\sqrt{n}} \left|\sum_{t=s}^{e} \sigma^2(t/T)(Z^2_{t,T} - 1) \cdot c_t\right| > C \log T\right) \to 0, \tag{3.15}$$

where we define

$$c_t = \frac{\sqrt{e-b}}{\sqrt{b-s+1}} \text{ for } t \in [s, b] \text{ and } c_t = \frac{\sqrt{b-s+1}}{\sqrt{e-b}} \text{ otherwise.}$$

Let $\{U_t\}_{t=s}^e$ be i.i.d. standard normal variables, and define an $n \times n$-matrix $\mathbf{V} = (v_{i,j})_{i,j=1}^n$ and an $n \times n$-diagonal matrix $\mathbf{W} = (w_{i,j})_{i,j=1}^n$ with their elements satisfying

$$v_{i,j} = \text{cor}\left(Z_{i+s-1,T}, Z_{j+s-1,T}\right) \text{ and } w_{i,i} = \sigma^2\left(\frac{i+s-1}{T}\right) \cdot c_{i+s-1},$$

respectively. By standard results (see e.g. Johnson & Kotz (1970), page 151),

showing (3.15) is equivalent to showing that the following holds

$$\left|\sum_{t=s}^{e} \lambda_{t-s+1}(U_t^2 - 1)\right| < C\sqrt{n}\log T$$

uniformly over $(s, b, e) \in \mathcal{D}$ with probability converging to 1, where λ_i are eigenvalues of the matrix $\mathbf{VW}$. Due to the Gaussianity of U_t, it can be shown that $\lambda_{t-s+1}(U_t^2 - 1)$ satisfies the Cramér's condition, i.e., there exists a constant $C > 0$ such that

$$\mathbb{E}\left|\lambda_{t-s+1}(U_t^2 - 1)\right|^p \leq C^{p-2}p!\mathbb{E}\left|\lambda_{t-s+1}(U_t^2 - 1)\right|^2, \quad p = 3, 4, \ldots.$$

Therefore we can apply Bernstein's inequality (Bosq, 1998) and obtain

$$\begin{aligned}
&\mathbb{P}\left(\left|\sum_{t=s}^{e} \frac{1}{\sqrt{n}}\sigma^2(t/T)(Z_{t,T}^2 - 1)\cdot c_t\right| > C\log T\right) \\
&\leq 2\exp\left(-\frac{n\log^2 T}{4\sum_{i=1}^{n}\lambda_i^2 + 2\max_i|\lambda_i|C\sqrt{n}\log T}\right). \qquad (3.16)
\end{aligned}$$

Note that

$$\sum_{i=1}^{n}\lambda_i^2 = \operatorname{tr}(\mathbf{VW})^2 \leq c^2\max_z \sigma^4(z)n\rho_\infty^2.$$

Also it follows that

$$\max_i|\lambda_i| \leq c\max_z \sigma^2(z)\|\mathbf{V}\|,$$

where $\|\cdot\|$ denotes the spectral norm of a matrix. Since $\mathbf{V}$ is non-negative definite, we have $\|\mathbf{V}\| \leq \rho_\infty^1$ and then (3.15) is bounded by

$$\begin{aligned}
&\sum_{(s,b,e)\in\mathcal{D}} 2\exp\left(-\frac{n\log^2 T}{4c^2\max_z \sigma^4(z)n\rho_\infty^2 + 2c\max_z \sigma^2(z)\sqrt{n}\log T\rho_\infty^1}\right) \\
&\leq CT^3\exp\left(-\log^2 T\right) \to 0. \qquad (3.17)
\end{aligned}$$

The convergence in (3.17) follows from the fact that $\rho_\infty^p \leq C2^{I^*}$ and this can be made to be of order $\log T$, since the only requirement on I^* is that it converges to infinity but no particular speed is specified. Thus the lemma follows. □

Lemma 3.4. *Assume that (3.10) and (3.11) hold. Then for* $b = \arg\max_{s<t<e} |\tilde{\mathbb{Y}}^t_{s,e}|$, *there exists* $1 \le r \le q$ *such that, for large* T,

$$|b - \eta_{p_0+r}| \le \epsilon_T. \tag{3.18}$$

Proof. Let $\tilde{\mathbb{S}}_{s,e} = \max_{s<t<e} |\tilde{\mathbb{S}}^t_{s,e}|$. From Lemma 3.3, we have

$$\tilde{\mathbb{Y}}^b_{s,e} \ge \tilde{\mathbb{S}}_{s,e} - \log T \text{ and } \tilde{\mathbb{S}}^b_{s,e} \ge \tilde{\mathbb{Y}}^b_{s,e} - \log T$$

for large T. Hence it can be derived that $\tilde{\mathbb{S}}^b_{s,e} \ge \tilde{\mathbb{S}}_{s,e} - 2\log T$.

Assume that (3.18) does not hold and thus $b \in (\eta_{p_0+r} + \epsilon_T, \eta_{p_0+r+1} - \epsilon_T)$ for any r. From Lemma 2.2 in Venkatraman (1993), $\tilde{\mathbb{S}}^t_{s,e}$ is either monotonic or decreasing and then increasing on $[\eta_{p_0+r}, \eta_{p_0+r+1}]$, which implies that $\tilde{\mathbb{S}}^{\eta_{p_0+r}}_{s,e} \vee \tilde{\mathbb{S}}^{\eta_{p_0+r+1}}_{s,e} > \tilde{\mathbb{S}}^b_{s,e}$.

Suppose $\tilde{\mathbb{S}}^{\eta_{p_0+r}}_{s,e} > \tilde{\mathbb{S}}^b_{s,e}$. Then there exists $b' \in (\eta_{p_0+r}, \eta_{p_0+r} + \epsilon_T]$ satisfying $\tilde{\mathbb{S}}^{\eta_{p_0+r}}_{s,e} - 2\log T \ge \tilde{\mathbb{S}}^{b'}_{s,e}$ from Lemma 3.2. Since $b > b'$, we also get $\tilde{\mathbb{S}}^{\eta_{p_0+r+1}}_{s,e} > \tilde{\mathbb{S}}^b_{s,e}$ (as $\tilde{\mathbb{S}}^t_{s,e}$ is locally increasing at $t = b$), and there will again be a $b'' \in [\eta_{p_0+r+1} - \epsilon_T, \eta_{p_0+r+1})$ satisfying $\tilde{\mathbb{S}}^{\eta_{p_0+r}}_{s,e} - 2\log T \ge \tilde{\mathbb{S}}^{b''}_{s,e}$. Since $b'' > b$, it contradicts that $\tilde{\mathbb{S}}^b_{s,e} \ge \tilde{\mathbb{S}}_{s,e} - 2\log T$. Similar arguments are applicable when $b < \eta_{p_0+r}$ and therefore the lemma follows. □

Lemma 3.5. *Under (3.10) and (3.11),*

$$\mathbb{P}\left(\left| \tilde{\mathbb{Y}}^b_{s,e} \right| < \tau T^\theta \sqrt{\log T} \cdot n^{-1} \sum_{t=s}^{e} \tilde{Y}^2_{t,T} \right) \to 0 \tag{3.19}$$

for $b = \arg\max_{s<t<e} |\tilde{\mathbb{Y}}^t_{s,e}|$.

Proof. From Lemma 3.4, there exists some r such that $|b - \eta_{p_0+r}| < \epsilon_T$. Denote

$$\tilde{d} = \tilde{\mathbb{Y}}^b_{s,e} = \tilde{d}_1 - \tilde{d}_2 \text{ and } \tilde{m} = \frac{1}{\sqrt{n}} \sum_{t=s}^{e} \tilde{Y}^2_{t,T} = c_1 \tilde{d}_1 + c_2 \tilde{d}_2,$$

where

$$\tilde{d}_1 = \sqrt{\frac{e-b}{n(b-s+1)}} \sum_{t=1}^{e} \tilde{Y}^2_{t,T}, \quad \tilde{d}_2 = \sqrt{\frac{b-s+1}{n(e-b)}} \sum_{t=1}^{e} \tilde{Y}^2_{t,T},$$

$$c_1 = \sqrt{\frac{b-s+1}{e-b}} \text{ and } c_2 = \sqrt{\frac{e-b}{b-s+1}}.$$

For simplicity, we assume $c_2 > c_1$. Further, let $\mu_i = \mathbb{E}\tilde{d}_i$ and $w_i = \mathrm{var}(\tilde{d}_i)$ for $i = 1, 2$ and define $\mu = \mathbb{E}\tilde{d}$ and $w = \mathrm{var}(\tilde{d})$. Finally, t_n denotes the threshold dependent on n as $t_n = \tau T^\theta \sqrt{\log T/n}$. Then showing (3.19) is equivalent to showing that $\mathbb{P}(|\tilde{d}| \le t_n \cdot \tilde{m}) \to 0$.

We first note that $w_i \le 2c^2 \sup_z \sigma^4(z)\rho_\infty^2$ and $\mu_i \le c\sqrt{n} \sup_z \sigma^2(z)$. Using Markov's and the Cauchy-Schwarz inequalities,

$$\begin{aligned}
&\mathbb{P}(\tilde{d} \le t_n \cdot \tilde{m}) \le \\
&\mathbb{P}\left\{(\tilde{d}_1 - \mu_1)(c_1 t_n - 1) + (\tilde{d}_2 - \mu_2)(c_2 t_n + 1) + 2c_1 t_n \mu_1 + (c_2 - c_1) t_n \mu_2 \ge (1 + c_1 t_n)\mu\right\} \\
&\le 4\mu^{-2}(1 + c_1 t_n)^{-2} \left\{(c_1 t_n - 1)^2 w_1 + (c_2 t_n + 1)^2 w_2 + 4c_1^2 t_n^2 \mu_1^2 + (c_2 - c_1)^2 t_n^2 \mu_2^2\right\} \\
&\le O\left\{\mu^{-2} c^2 \sup_z \sigma^4(z) \left(\rho_\infty^2 + \tau^2 T^{2\theta} \log T\right)\right\},
\end{aligned}$$

and since for large T,

$$\mu = \tilde{\mathbb{S}}_{s,e}^b \ge \delta_T/\sqrt{T} \ge T^{\Theta - 1/2} \gg T^\theta \sqrt{\log T},$$

the conclusion follows. □

Lemma 3.6. *For some positive constants c_1, c_2, let s, e satisfy either*

(i) there exists $p \in \{1, \ldots, B\}$ such that $s \le \eta_p \le e$ and $(\eta_p - s + 1) \wedge (e - \eta_p) \le c_1 \epsilon_T$, or

(ii) there exists $p \in \{1, \ldots, B\}$ such that $s \le \eta_p < \eta_{p+1} \le e$ and $(\eta_p - s + 1) \vee (e - \eta_{p+1}) \le c_2 \epsilon_T$.

Then

$$\mathbb{P}\left(\left|\tilde{\mathbb{Y}}_{s,e}^b\right| > \tau T^\theta \sqrt{\log T} \cdot n^{-1} \sum_{t=s}^{e} \tilde{Y}_{t,T}^2\right) \to 0$$

for large T, where $b = \arg\max_{s<t<e} |\tilde{\mathbb{Y}}_{s,e}^t|$.

Proof. First we assume (i) holds. Define two events $\mathcal{A}_1$ and $\mathcal{A}_2$ as

$$\begin{aligned}\mathcal{A}_1 &= \left\{ \left| \tilde{\mathbb{Y}}^b_{s,e} \right| > \tau T^\theta \sqrt{\log T} \cdot n^{-1} \sum_{t=s}^{e} \tilde{Y}^2_{t,T} \right\}, \\ \mathcal{A}_2 &= \left\{ \frac{1}{n} \left| \sum_{t=s}^{e} \left(\tilde{Y}^2_{t,T} - \mathbb{E}\tilde{Y}^2_{t,T} \right) \right| < h \equiv \frac{(\eta_p - s + 1)\sigma_1^2 + (e - \eta_p)\sigma_2^2}{2n} \right\},\end{aligned}$$

where $\sigma_1^2 = \sigma^2(\eta_p/T)$ and $\sigma_2^2 = \sigma^2((\eta_p + 1)/T)$. Then, it follows

$$\mathbb{P}(\mathcal{A}_1) = \mathbb{P}(\mathcal{A}_1 \cap \mathcal{A}_2) + \mathbb{P}(\mathcal{A}_1 | \mathcal{A}_2^c)\,\mathbb{P}(\mathcal{A}_2^c) \leq \mathbb{P}(\mathcal{A}_1 \cap \mathcal{A}_2) + \mathbb{P}(\mathcal{A}_2^c).$$

The first probability is bounded as

$$\mathbb{P}(\mathcal{A}_1 \cap \mathcal{A}_2) \leq \mathbb{P}\left(\left| \tilde{\mathbb{Y}}^b_{s,e} \right| > \tau T^\theta \sqrt{\log T} \cdot n^{-1} \sum_{t=s}^{e} \left(\mathbb{E}\tilde{Y}^2_{t,T} - h \right) \right). \tag{3.20}$$

From Lemma 3.3, we have $|\tilde{\mathbb{Y}}^b_{s,e} - \tilde{\mathbb{S}}^b_{s,e}| \leq \log T$ for large T. Also Lemmas 2.2 and 2.3 of Venkatraman (1993) indicate that

$$\max_{s<t<e} |\tilde{\mathbb{S}}^t_{s,e}| = |\tilde{\mathbb{S}}^{\eta_p}| \leq \sqrt{c_1 \epsilon_T (n - c_1 \epsilon_T)/n} \leq C\sqrt{\epsilon_T}.$$

Therefore $|\tilde{\mathbb{Y}}^b_{s,e}| \leq |\tilde{\mathbb{S}}^{\eta_p}| + \log T \leq C\sqrt{\epsilon_T}$ for some $C > 0$ and by applying Markov's inequality, (3.20) is bounded by

$$\frac{\mathbb{E}\left(\tilde{\mathbb{Y}}^b_{s,e} \right)^2}{\tau^2 T^{2\theta} \log T \cdot h^2} \leq C T^{1/2 - 2\theta} \to 0.$$

Turning our attention to $\mathbb{P}(\mathcal{A}_2^c)$, we need to show that

$$\mathbb{P}\left(\frac{1}{n} \left| \sum_{t=s}^{e} \sigma^2(t/T)(Z^2_{t,T} - 1) \right| > h \right) \to 0,$$

which is achieved by applying Bernstein's inequality as in (3.16) (in the proof of Lemma 3.3). Similar arguments are applied when (ii) holds, and thus the lemma is proved. □

We now prove the consistency of the binary segmentation procedure for the model given in (3.9). At the start of the algorithm, as $s = 0$ and $e = T - 1$, all conditions for Lemma 3.5 are met. Therefore, by Lemma 3.4, the binary segmentation procedure finds a breakpoint within the distance of $C\epsilon_T$ from a true breakpoint. Under the conditions on the distance between the two adjacent breakpoints (Assumption 3.2), both (3.10) and (3.11) are satisfied within each segment defined by previously detected breakpoints, until every breakpoint in $\sigma^2(t/T)$ is identified. Then, either of two conditions (i) or (ii) in Lemma 3.6 is met within each segment, and therefore no further breakpoint is detected with probability converging to 1.

Next, we study how the bias present in $\mathbb{E}I^{(i)}_{t,T}(= \sigma^2_{t,T})$ affects the consistency. Proposition 3.1 states that $\mathbb{E}I^{(i)}_{t,T}$ is close to $\beta_i(t/T)(= \sigma^2(t/T))$ in the sense that the integrated bias between $\mathbb{E}I^{(i)}_{t,T}$ and $\beta_i(t/T)$ converges to zero (Proposition 3.1). Suppose the interval $[s, e]$ includes a true breakpoint η_p as in (3.10), and let

$$b = \arg\max_{t\in(s,e)} |\tilde{\mathbb{S}}^t_{s,e}| \text{ and } \hat{b} = \arg\max_{t\in(s,e)} \left|\mathbb{S}^t_{s,e}\right|.$$

Recall that $\mathbb{E}I^{(i)}_{t,T}$ remains constant within each stationary segment, apart from short (of length $C2^{-i}$) intervals around the discontinuities in $\beta_i(t/T)$. Suppose a jump occurs at $t = \eta_p$ in $\beta_i(t/T)$ yet there is no change in $\mathbb{E}I^{(i)}_{t,T}$ for $t \in [\eta_p - C2^{-i}, \eta_p + C2^{-i}]$. Then the integrated bias is bounded from below by $C'\delta_T/T$ from Assumption 3.2, and it contradicts Proposition 3.1.

Therefore there will be a change in $\mathbb{E}I^{(i)}_{t,T}$ as well on such intervals around $t = \eta_p$ such that

$$\mathbb{E}I^{(i)}_{t_1,T} \neq \mathbb{E}I^{(i)}_{t_2,T} \text{ for } t_1 \leq \eta_p - C2^{-i} \text{ and } t_2 \geq \eta_p + C2^{-i}.$$

Although the bias of $\mathbb{E}I^{(i)}_{t,T}$ in relation to $\beta_i(t/T)$ may cause a discrepancy between $\hat{b}$ and b, it is expected that

$$|\hat{b} - b| \leq C2^{I^*} \ll \epsilon_T$$

for $I^* = O(\log\log T)$, which is an admissible rate for I^*. Besides, once one

breakpoint is detected in such intervals, the algorithm by its construction does not allow any more breakpoints to be detected within the distance of Δ_T from the detected breakpoint. Hence the bias in $\mathbb{E}I_{t,T}^{(i)}$ does not affect the results of Lemmas 3.1–3.6 for wavelet periodograms at finer scales, and the consistency still holds for $Y_{t,T}^2$ in (3.3), in place of $\tilde{Y}_{t,T}^2$.

Finally, we note that the within-scale post-processing step in Section 3.2.2.1 is in line with the theoretical consistency of our procedure.

- Lemma 3.5 implies that our test statistic exceeds the threshold when there is a breakpoint η within a segment $[s, e]$, which is of sufficient distance from both s and e and thus remained to be detected.

- Lemma 3.6 shows that it does not exceed the threshold when (s, η, e) does not satisfy the conditions required in Lemma 3.5.

3.5.2 The proof of Theorem 3.2

From Assumption 3.1 and the invertibility of the autocorrelation wavelet inner product matrix $\mathbf{A}$, there exists at least one sequence of wavelet periodograms among $I_{t,T}^{(i)},\ i = -1, \ldots, -I^*$ from which any breakpoint in $\mathcal{B}$ is detected.

Suppose there is only one such scale, i_0, for $\nu_p \in \mathcal{B}$ and denote the detected breakpoint as $\hat{\eta}_{p_0}^{(i_0)}$. After the across-scales post-processing, $\hat{\eta}_{p_0}^{(i_0)}$ is selected as $\hat{\nu}_p$, since no other $\hat{\eta}_q^{(i)}$, $i \neq i_0$, is within the distance of $\Lambda_T = C\epsilon_T$ from either $\hat{\nu}_p$ or $\hat{\eta}_{p_0}^{(i_0)}$, and thus

$$\left|\nu_p - \hat{\eta}_{p_0}^{(i_0)}\right| = |\nu_p - \hat{\nu}_p| \leq \epsilon_T$$

with probability converging to 1 from Theorem 3.1.

On the other hand, suppose that there are $D(\leq I^*)$ breakpoints detected for ν_p detected from D different wavelet periodogram sequences $I_{t,T}^{(i)},\ i = -i_1, \ldots, -i_D$, and denote them as $\hat{\eta}_{p_1}^{(i_1)}, \ldots, \hat{\eta}_{p_D}^{(i_D)}$. For any $1 \leq a < b \leq D$, it holds that

$$\left|\hat{\eta}_{p_a}^{(i_a)} - \hat{\eta}_{p_b}^{(i_b)}\right| \leq \left|\hat{\eta}_{p_a}^{(i_a)} - \nu_p\right| + \left|\hat{\eta}_{p_b}^{(i_b)} - \nu_p\right| \leq C\epsilon_T$$

by Theorem 3.1, and therefore all $\hat{\eta}_{p_1}^{(i_1)}, \ldots, \hat{\eta}_{p_D}^{(i_D)}$ are classified as belonging to the same group, say $\mathcal{G}$. Then, our across-scales post-processing procedure is con-

structed to select only the one from the finest scale as $\hat{\nu}_p$ among those breakpoints in $\mathcal{G}$. Hence the post-processing preserves the consistency for the breakpoints selected as its outcome in terms of their total number and locations.

Chapter 4

Multiscale interpretation of piecewise constant estimators: taut string and Unbalanced Haar techniques

Both the Unbalanced Haar (UH) technique (Fryzlewicz, 2007) and the taut string (TS) based method (see e.g. Barlow *et al.* (1972) and Davies & Kovac (2001)) estimate a one-dimensional function f from noisy observations $\{y_t\}_{t=1}^n$ by means of piecewise constant functions under the following additive model:

$$y_t = f\left(\frac{t}{n}\right) + \epsilon_t, \quad t = 1, \cdots, n. \tag{4.1}$$

Both techniques are computationally fast, achieve theoretical consistency, and exhibit good performance in numerical simulation studies. The UH method involves the decomposition of the data with respect to an adaptively chosen, Haar-like wavelet basis and therefore it is easy to comprehend its multiscale nature. On the other hand, being a penalised least squares estimator, the multiscale character of the TS method is not so obvious and has not previously been noted in the literature to our best knowledge.

In this chapter, our interest lies in studying the two methods and establish-

ing a link between the two and as the first step, we present a unified estimation methodology which both the UH and the TS techniques are instances of. It is this unified framework that provides ground for a multiscale interpretation of the TS technique, as well as better understanding of the similarities and differences between the two methods. Then taking advantage of this common framework, we derive lessons which either method can learn from the other. Further, being located between the chapters addressing two different problems, time series segmentation (Chapter 3) and high-dimensional variable selection (Chapter 5), this chapter concludes by connecting these problems using the UH and the TS techniques, with emphasis on the unifying theme of this thesis, sparsity.

The rest of the chapter is organised as follows. In Section 4.1, we first provide an overview of the UH and TS techniques as well as their algorithms in the form of flowcharts within the unified framework, which offer an insight into the relationship between their physical interpretations. Then follows a comparison study, including the understanding of the two techniques in the context of breakpoint detection (Section 4.2). In Section 4.3, we list some ways of improving and extending both techniques, which suggest avenues for possible future research, and finally in Section 4.4, we link this chapter to other applications of sparse modelling and estimation discussed in Chapter 3 and Chapter 5.

4.1 Unbalanced Haar and taut string techniques

4.1.1 Unbalanced Haar technique

The UH technique consists of three steps: the transformation of observations $\{y_t\}_{t=1}^n$ with respect to an adaptively chosen UH wavelet basis, hard-thresholding of the wavelet coefficients, and the inverse UH transformation of the thresholded coefficients. The principles of traditional wavelet thresholding estimation, which does not have the adaptive basis selection step, can be found in Section 2.4.1.

Before introducing the UH wavelets, we first recall the wavelet function ψ of

the Haar wavelet

$$\psi(t) = \begin{cases} 1 & \text{for } t \in [0, 1/2), \\ -1 & \text{for } t \in [1/2, 1), \\ 0 & \text{otherwise,} \end{cases} \tag{4.2}$$

which clearly shows that Haar wavelets have a breakpoint in the middle of their supports.

The UH wavelets were first studied in Girardi & Sweldens (1997) as an extended version of classical Haar wavelet vectors, the extension being that the breakpoint was permitted to occur anywhere within their support. Let s and e denote the start and the end of a generic interval, respectively, and let $b \in (s, e)$ denote the location of the breakpoint. Then, we denote a UH vector which is supported on the interval $[s, e]$ and has a breakpoint b by $\psi_{s,b,e}$, with its elements $\psi_{s,b,e}(l)$ satisfying

$$\psi_{s,b,e}(l) = \sqrt{\frac{e-b}{(e-s+1)\cdot(b-s+1)}} \cdot \mathbb{I}_{[s,b]}(l) - \sqrt{\frac{b-s+1}{(e-s+1)\cdot(e-b)}} \cdot \mathbb{I}_{[b+1,e]}(l).$$

Classical Haar wavelet vectors are a special case with $b = (s + e - 1)/2$.

Note that on a given support $\{s, \ldots, e\}$, the choice of breakpoints b defines the choice of a UH vector. Fryzlewicz (2007) presented an adaptive way of UH basis selection, which can be seen as iteratively detecting a breakpoint within a segment defined by the breakpoints detected at the previous iterations. Denote the vector of observations as $\tilde{y} = (y_1, \ldots, y_n)^T$ and its subvector on a generic support $\{s, \ldots, e\}$ as $\tilde{y}_{s,e} = (y_s, \ldots, y_e)^T$. Then the first breakpoint $b_{1,1}$ is chosen from $\{1, \ldots, n\}$ such that the inner product between $\tilde{y}$ and $\psi_{1,b_{1,1},n}$ is maximised in absolute value, i.e.

$$b_{1,1} = \arg \max_{b \in \{1,\ldots,n\}} \left| \langle \tilde{y}, \psi_{1,b,n} \rangle \right|. \tag{4.3}$$

The explicit expression of the UH wavelet coefficient is given later in (4.8). The next breakpoints are chosen similarly on the supports defined by the previously chosen breakpoint, i.e. $\{1, \ldots, b_{1,1}\}$ and $\{b_{1,1}+1, \ldots, n\}$, and the same procedure

is repeated until it is no longer possible to divide any support into two.

Then $\tilde{y}$ is transformed with respect to the orthonormal basis defined by these selected breakpoints. Once the UH transform is finished, the next step of UH technique is hard-thresholding the wavelet coefficients by setting to zero those which fall below the universal threshold $\sigma\sqrt{2\log n}$. In practice, σ, the standard deviation of the noise ϵ_t, is likely to be unknown yet can be estimated as the median of the sequence $\left\{|y_t - y_{t-1}|/\sqrt{2},\ t = 2, \ldots, n\right\}$ divided by the 0.75-quantile of the standard normal distribution (which is approximately equal to 0.6745).

Finally, the inverse transform of the thresholded wavelet coefficients is taken to obtain the final estimate $\hat{f}^{UH}$. It is shown that $\hat{f}^{UH}$ is mean-square consistent for a wide range of functions, uniformly over those UH bases (however they have been selected) which are not "too unbalanced" in the following sense: there exists a fixed $c \in [1/2, 1)$ such that each basis vector should satisfy

$$\max\left\{\frac{b-s+1}{e-s+1}, \frac{e-b}{e-s+1}\right\} \leq c. \tag{4.4}$$

Thus in practice, the maximisation of the inner products as described above (see e.g. (4.3)) is performed in such a way that each time, the maximum is taken over only those wavelets which satisfy condition (4.4), to ensure mean-square consistency of the resulting estimator.

We note that at the beginning of the UH basis selection procedure, the entire observation vector is scanned in the search for $b_{1,1}$. However, the scope of the search is iteratively narrowed down, as each "parent" vector of observations $\tilde{y}_{s,e}$ gets iteratively divided into two "children", i.e. the subvectors to the left and right of the previously detected breakpoint $\tilde{y}_{s,b}$ and $\tilde{y}_{b+1,e}$. Because of this natural parent-child structure of the search, the UH estimation technique can be viewed as multiscale.

The recursive, binary nature of the UH technique shows its connection to the CART methodology, which is also a greedy binary splitting procedure (see Section 2.4.2). However, the UH technique is more than a binary decision tree; the key ingredient of the UH technique is that it furnishes a decomposition of the data into wavelet coefficients, which can then be further processed depending on the aim of the analysis, and thus fully enjoys the benefits of its being a wavelet

technique.

We also note that the binary decision tree is only one, "top-down", way of adaptively choosing a UH basis. Another way, which can be referred to as a "bottom-up" approach, was introduced in Fryzlewicz (2007). Besides, even the top-down UH estimator and the CART differ significantly in that the former employs the usual universal wavelet thresholding, whereas the latter employs a "hereditary" structure by which further splitting is stopped on a subinterval which is judged to be a node. An interesting connection between the dyadic ("balanced") Haar approach and the dyadic CART is given in Donoho (1997), where again, it is noted that the dyadic CART estimator differs from the Haar thresholding estimator due to its heredity rule imposed on the tree structure.

In Section 4.1.3, a more physical interpretation of the UH technique is provided along with its flowchart representation.

4.1.2 Taut strings

The TS technique was introduced in Barlow *et al.* (1972) in the context of isotonic (monotonic) function estimation. For the more general model (4.1), it is shown that a taut string solves a penalised least squares functional, where the penalty is based on the total variation norm (Davies & Kovac, 2001; Mammen & van de Geer, 1997). That is, the TS technique searches for a $\hat{f}^{TS}$ which satisfies

$$\hat{f}^{TS} = \arg\min_{\tilde{f}} \left\{ \sum_{t=1}^{n} (\tilde{f}_t - y_t)^2 + \gamma \sum_{t} |\tilde{f}_{t+1} - \tilde{f}_t| \right\}, \tag{4.5}$$

where γ is a tuning parameter. Then, $\hat{f}^{TS}$ is a piecewise constant function whose number of breakpoints is a non-increasing function of γ.

One way of describing the computation of $\hat{f}^{TS}$ is using the following "string" and "tube" arguments, which is referred to as the *uniscale* TS algorithm throughout this chapter. Denote the integrated process of observations $\{y_t\}_{t=1}^n$ as $\mathbf{Y} = \{Y_t\}_{t=1}^n$, i.e.,

$$Y_t = \sum_{u=1}^{t} y_u \text{ with } Y_0 = 0.$$

Then the graph of $\mathbf{Y}$ on the interval $[0,1]$ connects $\{(t/n, Y_t),\ 0 \le t \le n\}$. Now, imagine a tube of radius, say $\lambda > 0$, which surrounds the graph of $\mathbf{Y}$. This tube consists of the lower bound $l_t = Y_t - \lambda$ and the upper bound $u_t = Y_t + \lambda$, and its radius λ is related to the penalty parameter γ from (4.5).

Then, suppose there is a string connecting $(0, Y_0)$ and $(1, Y_n)$, while being constrained to lie within the tube. The string is now pulled until it is *taut*, thus the name taut string, touching the tube on either lower or upper side at possibly multiple "knots". In other words, the taut string has the smallest length among the functions $\tilde{z}$ satisfying

$$\tilde{z} : [0,1] \to \mathbb{R}; \quad \tilde{z}_0 = Y_0,\ \tilde{z}_n = Y_n \text{ and } l_t \le \tilde{z}_t \le u_t,$$

and its derivative coincides with the above $\hat{f}^{TS}$ for an appropriately chosen λ (Davies & Kovac, 2001).

Note that between the two knots at which the string only touches the upper bound $\mathbf{u}$, it coincides with the greatest convex minorant (GCM) of $\mathbf{u}$. Similarly, between the two knots where the string only touches the lower bound $\mathbf{l}$, it is the least concave majorant (LCM) of $\mathbf{l}$. Finally where the string switches from touching $\mathbf{u}$ to touching $\mathbf{l}$, a local maximum occurs in its derivative and a local minimum occurs in the opposite manner.

Davies & Kovac (2001) proposed the *taut string multiresolution method* for nonparametric regression with emphasis on consistent estimation of the number and locations of local extremes. Its final estimate $\hat{f}$ is obtained from a TS estimate $\hat{f}^{TS}$, by further squeezing $\hat{f}^{TS}$ locally such that the empirical residuals $\{y_t - \hat{f}_t\}_{t=1}^n$ would satisfy

$$\max_{I \in \mathcal{I}} \frac{1}{\sqrt{|I|}} \left| \sum_{t \in I} (y_t - \hat{f}_t) \right| \le \lambda, \tag{4.6}$$

where $\mathcal{I}$ denotes a collection of every support set $\{s, s+1, \ldots, e\}$ for $1 \le s \le e \le n$. The authors also introduced an algorithm to obtain $\hat{f}^{TS}$, which simultaneously computed the GCM of $\mathbf{u}$ and the LCM of $\mathbf{l}$ to find the knots from left ($t = 0$) to right ($t = n$).

In Section 4.1.3, we provide an alternative algorithm accompanied by a flowchart,

which reveals the multiscale nature of the TS method. It is this multiscale interpretation of the TS algorithm through which we derive the similarities and differences between the UH and TS techniques in Section 4.2.

4.1.3 Unified multiscale description of UH and TS algorithms

In introducing the unified framework for both UH and TS techniques, we revisit the concept of a string and its knots. Using the same notation as in Section 4.1.2, consider a string, denoted by $\mathbf{z}$, which connects $(0, Y_0)$ and $(1, Y_n)$ with a straight line.

We note that the algorithm for the UH technique is established in an *adjusted* y-axis. We define a multiplying factor ρ^{UH} on $t \in [s, e)$ as

$$\rho^{UH}(t; s, e) = \sqrt{\frac{e - s + 1}{(t - s + 1)(e - t)}}, \tag{4.7}$$

which is applied to the string $\mathbf{z}$ and the integrated process $\mathbf{Y}$ in order to yield their adjusted versions $\mathbf{z}^*$ and $\mathbf{Y}^*$ as

$$z_t^* = \rho^{UH}(t; s, n) \cdot z_t \text{ and } Y_t^* = \rho^{UH}(t; s, e) \cdot Y_t.$$

The adjusting factor ρ^{UH} comes from the UH wavelet basis which is used to compute the wavelet coefficient. It is designed such that the wavelet coefficient defined on the support $\{s, \ldots, e\}$ with a breakpoint at t is equal to the product of $\rho^{UH}(t; s, e)$ and the differential term between the local sum ($\sum_{u=s}^{t} y_u = Y_t - Y_{s-1}$) and the scaled global sum ($\frac{t-s+1}{e-s+1} \cdot (Y_e - Y_{s-1})$) of the observations, see (4.8) for further details.

Next, consider a tube of radius r surrounding the integrated process $\mathbf{Y}$ (or its adjusted version $\mathbf{Y}^*$ in the UH technique). This time, the radius is chosen to be large enough that the string $\mathbf{z}$ (or $\mathbf{z}^*$) does not touch the tube surrounding the integrated process $\mathbf{Y} \pm r$ (or $\mathbf{Y}^* \pm r$). With this starting set-up, our algorithmic interpretation of the UH and TS techniques is summarised in the flowcharts provided in Figures 4.1–4.2. Based on these flowcharts, Section 4.2 provides

a detailed comparison study between the UH and TS techniques.

The two algorithms proceed similarly by "squeezing" the tube and "re-arranging" the string simultaneously. By squeezing the tube, the first knot is detected at, say $t = b$, as the point where the tube first touches $\mathbf{z}$ ($\mathbf{z}^*$). If the radius of the squeezed tube is greater than a pre-specified value $\lambda > 0$, the string is re-arranged in a manner that is described in the point (ii) below, and two segments are defined by the previously detected knot at $t = b$, i.e. $[0, b/n]$ and $[(b+1)/n, 1]$. The same tube squeezing (in other words, knot detection) and string re-arrangement steps are repeated on each segment separately, as long as

Case 1. the length of the segment is large enough for further division of the segment to be possible in the next iteration, and

Case 2. the squeezed tube radius is greater than λ on the given segment.

If, on any segment, the tube is squeezed to have its radius less than λ, we set its radius back to λ. The estimation procedure is finished once the progression of the algorithm is terminated on every segment by the violation of either Case 1 or Case 2 above, and the final estimate is obtained as the derivative of the string $\mathbf{z}$. In both algorithms, the current parent segment is always split into two children subsegments. Therefore the same procedure is applied to the data at multiple scales and thus we can conclude that not only the UH technique but also the TS technique is multiscale.

While the above description shows the similarities between the basic steps of the two algorithms, they differ in the following details.

(i) The tube squeezing in the UH algorithm is performed in the adjusted y-axis with its adjusting factor defined in (4.7), while that in the TS algorithm is performed in the original y-axis.

(ii) When a knot is detected with the squeezed tube having its radius larger than λ, the string re-arrangement is done differently. On a generic segment $[s/n, e/n]$, the UH algorithm arranges $\mathbf{z}$ (in the original y-axis) to connect $(s/n, Y_s)$ and $(b/n, Y_b)$ with a straight line, as well as $(b/n, Y_b)$ and $(e/n, Y_e)$ with a straight line.

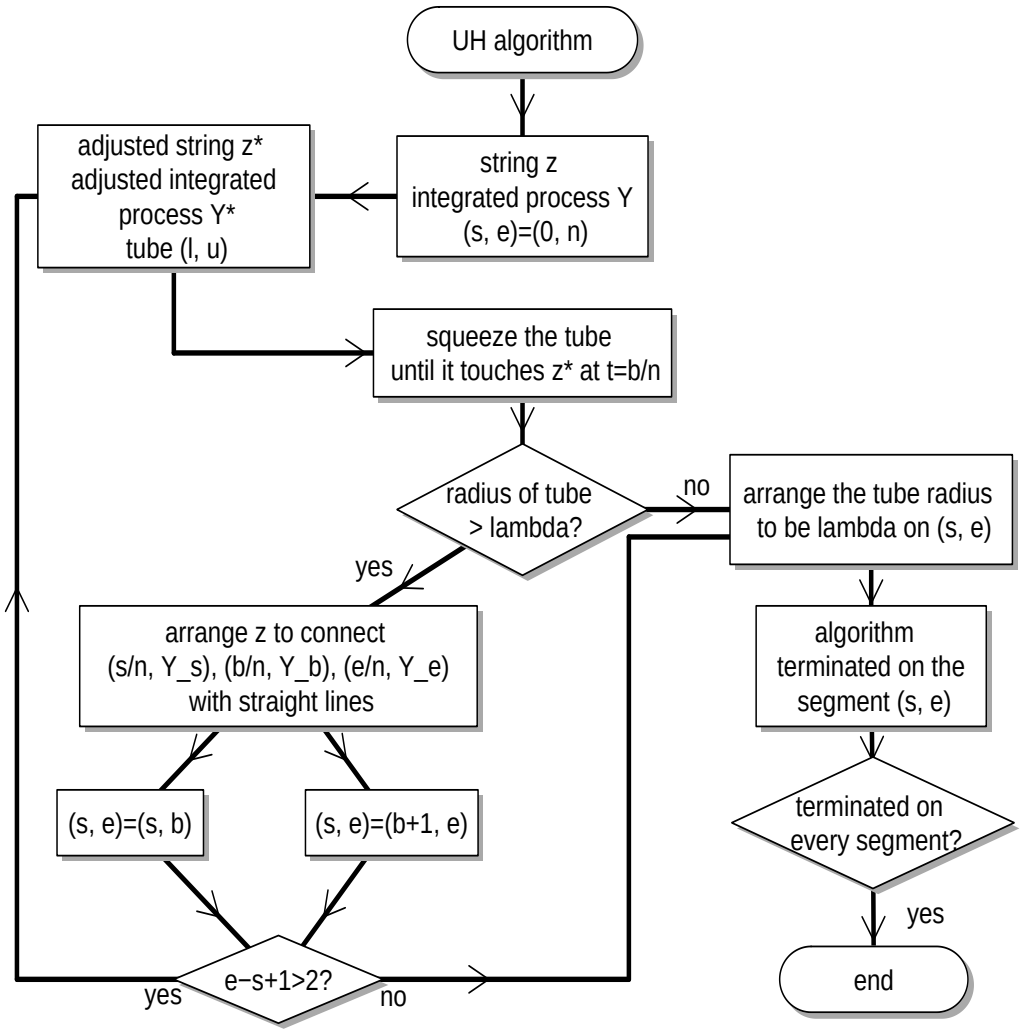


Figure 4.1: Flowcharts of UH algorithm.

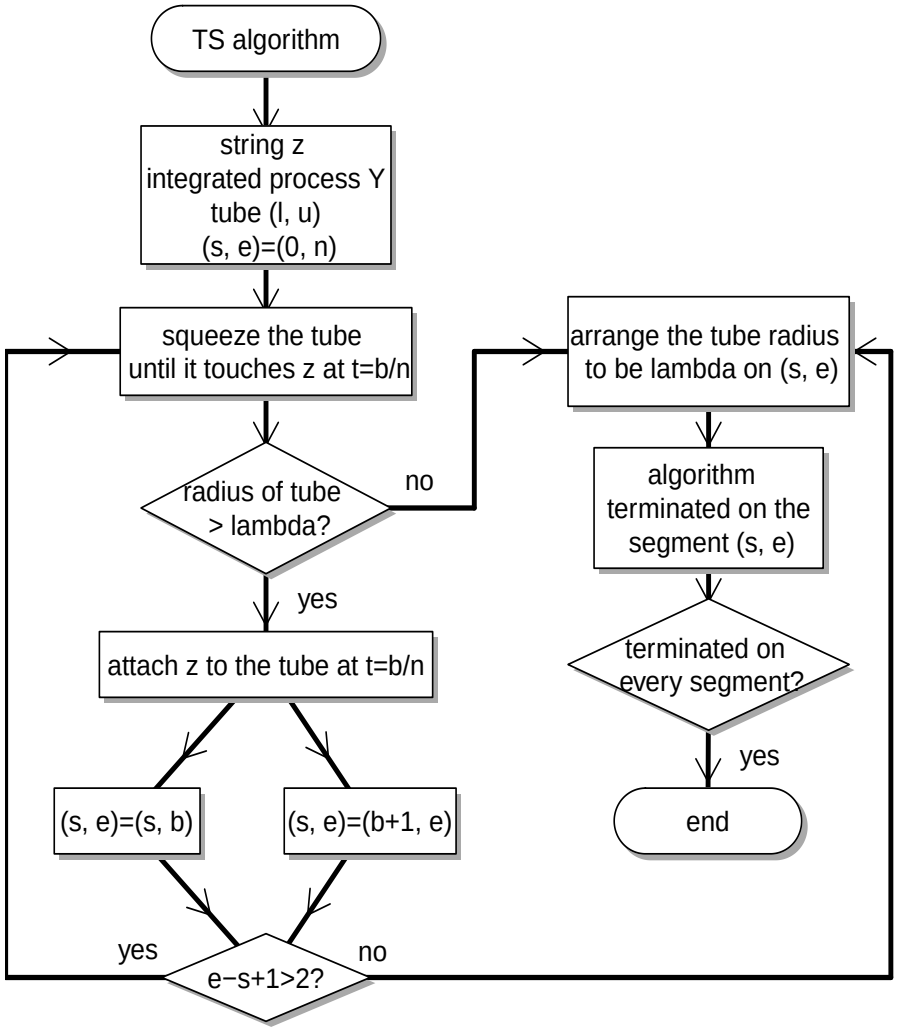


Figure 4.2: Flowcharts of TS algorithm.

On the other hand, the TS algorithm attaches $\mathbf{z}$ to the tube at the detected knot, and further squeezing of the tube is applied with $\mathbf{z}$ still being attached to it. Note that the tube itself remains as a symmetric band around the integrated process $\mathbf{Y}$ throughout the algorithm. However, since $\mathbf{z}$ consists of straight lines connecting two neighbouring knots (including $(0, Y_0)$ and $(1, Y_n)$), the slope of each line changes constantly as the radius of the tube decreases, and as a result, it is a constantly changing function on $[0, 1]$. The attachment of $\mathbf{z}$ to the tube can be observed in Figure 4.5, where the upper right and lower middle figures show the state in between the detection of two knots.

In summary, our TS algorithm returns its final estimate as the derivative of the string $\mathbf{z}$, which is attached to the tube of radius λ at zero, one, or multiple knots and connects neighbouring knots with straight lines.

As opposed to the uniscale TS algorithm presented in Section 4.1.2, the TS algorithm from our unified approach is referred to as the *multiscale* TS algorithm. We emphasise that the multiscale TS algorithm returns exactly the same estimator as that obtained from the uniscale TS algorithm, and thus it also solves the penalised least squares problem in (4.5). In the application of the multiscale TS algorithm, suppose that the first knot is detected with the tube squeezed just enough to touch the string. If the radius of the tube at such a state is λ_1, the string in that state is equal to the string from the uniscale TS algorithm with the tube radius equal to λ_1. Then recursively applying the same argument, it can be seen that the multiscale TS algorithm produces exactly the same state of the tube and the string as the uniscale TS algorithm.

We note that the UH algorithm as presented in the flowchart (Figure 4.1) is a slight modification of the UH technique described in Section 4.1.1. The modification simplifies the graphical representation as well as the comparison between the UH and TS techniques. In the flowchart, the algorithm terminates on a segment if the squeezed tube radius is smaller than λ (Case 2), which can be seen as imposing the heredity rule discussed in Section 4.1.1. On the other hand, the original algorithm terminates only when the segment is too short to be further divided into two (Case 1), and then applies hard-thresholding with λ as

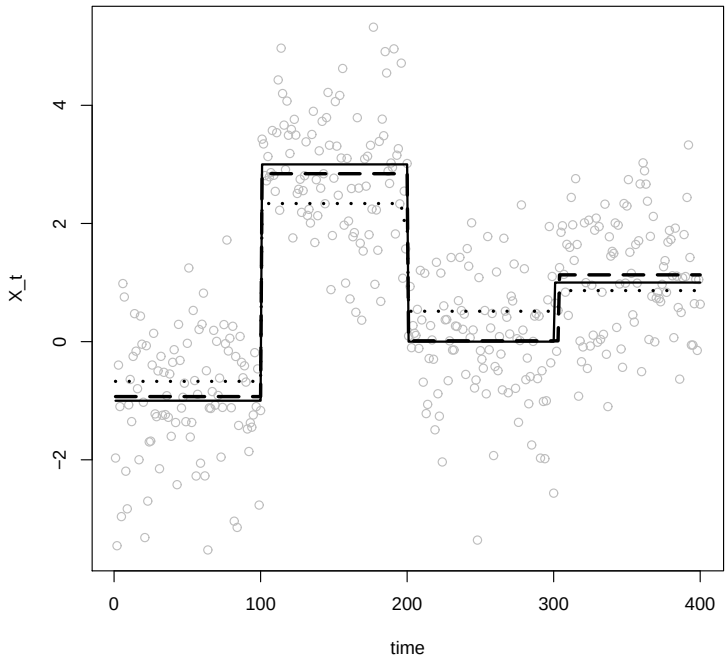


Figure 4.3: A toy example: observations y_t (grey dots), unknown function f (solid line), UH estimate $\hat{f}^{UH}$ (dashed line), TS estimate $\hat{f}^{TS}$ (dotted line).

the threshold. This difference can affect the adaptivity of the final estimate $\hat{f}^{UH}$ depending on the shape of underlying function f, and it is further discussed in Section 4.3 under the heading **Local squeezing**.

We also note that the algorithm in Figure 4.1 does not take into account the condition imposed in (4.4) when selecting $b \in (s, e)$, unlike the original UH algorithm as proposed in Fryzlewicz (2007). However, this condition can easily be incorporated in both UH and TS algorithms and is only omitted for the simplicity of presentation.

We conclude this section by presenting, in Figures 4.4–4.5, iteration-by-iteration progression of both algorithms from our unified approach, as applied to the toy example from Figure 4.3. Iteration (j, k) indicates that the knot is detected in the jth iteration on the kth segment from the left.

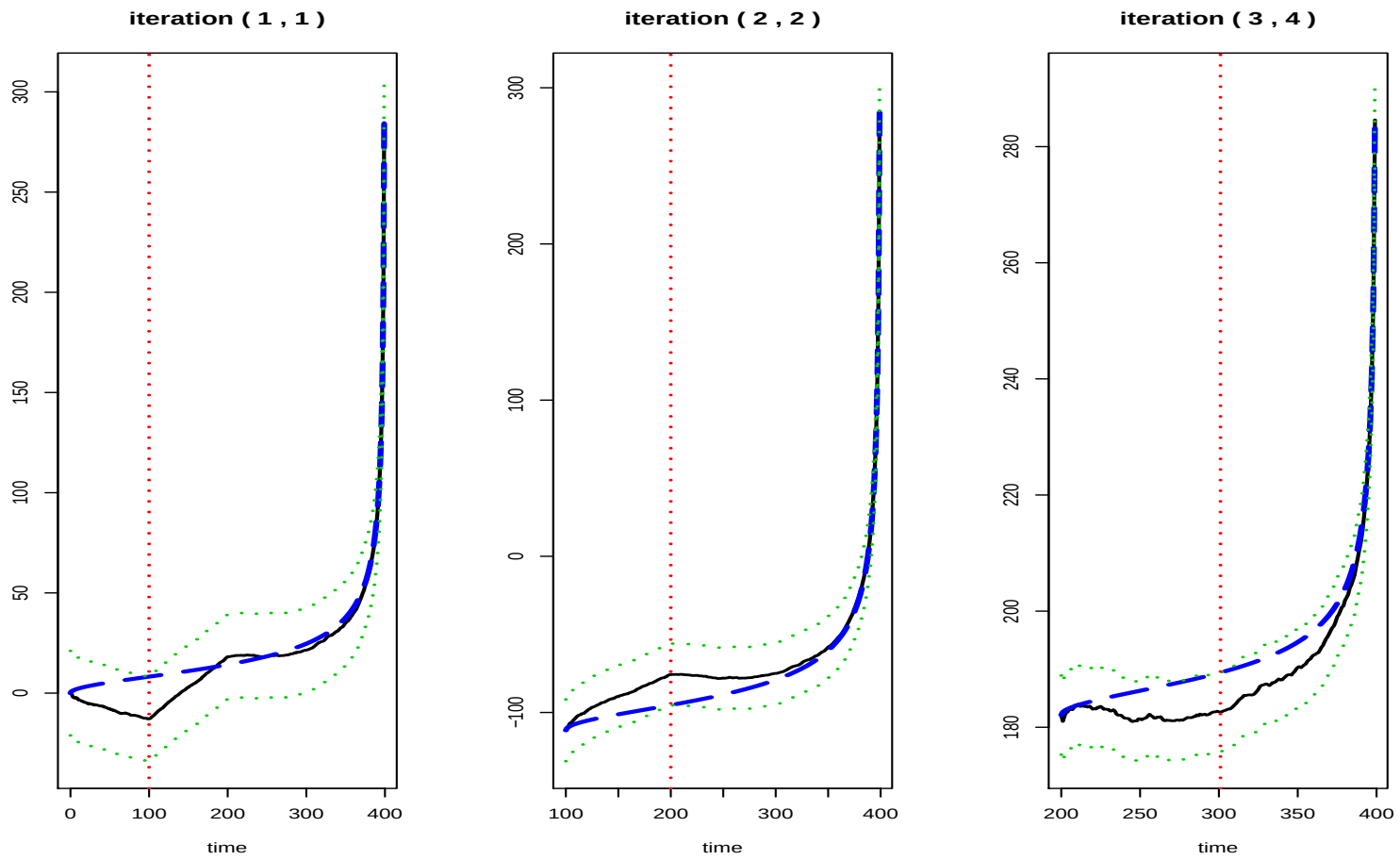


Figure 4.4: An application of UH algorithm to the model in Figure 4.3: adjusted integrated process $\mathbf{Y}^*$ (black solid), string $\mathbf{z}^*$ (blue dashed), tube $\mathbf{Y}^* \pm r$ (green dotted) and the locations of the knots (vertical, red dotted)

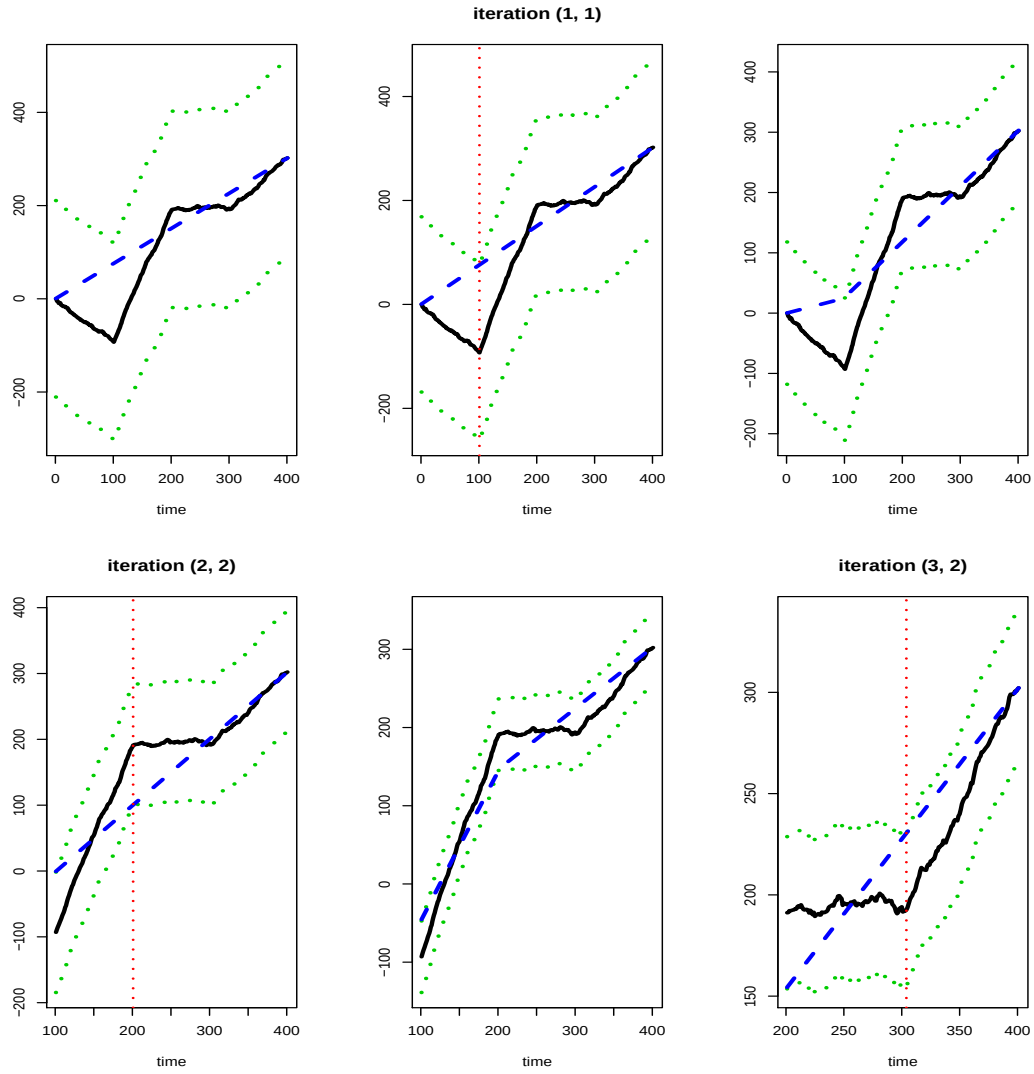


Figure 4.5: An application of TS algorithm to the model in Figure 4.3: integrated process $\mathbf{Y}$ (black solid), string $\mathbf{z}$ (blue dashed), tube $\mathbf{Y} \pm r$ (green dotted) and the locations of the knots (vertical, red dotted); the upper left figure shows the state of the tube and string at the beginning of algorithm; the upper right and lower middle figures show the state in between the detection of knots.

4.2 Comparison of UH and TS techniques

Based on the multiscale algorithms established in the previous section, we now provide a detailed comparison study between the UH and TS techniques. Firstly, in Section 4.2.1, we define the "locating" functions for both techniques, which are used to find the locations of a knot within a given segment. Then the comparison study continues in Section 4.2.2 in the framework of breakpoint detection, which provides a theoretical insight into reasons why the UH and TS techniques often perform differently.

4.2.1 Locating functions of UH and TS techniques

In the UH technique, the selection of a UH basis on a generic interval (s, e) involves the computation of the inner product between $\tilde{y}_{s,e}$ and a set of UH wavelet vectors $\psi_{s,t,e}$ for $t \in (s, e)$. The break $b \in (s, e)$ in a wavelet vector $\psi_{s,b,e}$ corresponds to the knot on the segment $(s/n, e/n)$ in the UH algorithm, and it is located as

$$\begin{aligned} & b = \arg\max_{t\in(s,e)} |\langle \tilde{y}_{s,e}, \psi_{s,t,e}\rangle| \\ = \ & \arg\max_{t\in(s,e)} \left| \sqrt{\frac{e-t}{(e-s+1)(t-s+1)}} (Y_t - Y_{s-1}) - \sqrt{\frac{t-s+1}{(e-s+1)(e-t)}} (Y_e - Y_t) \right| \\ = \ & \arg\max_{t\in(s,e)} \left| \sqrt{\frac{e-s+1}{(t-s+1)(e-t)}} \left\{ \frac{t-s+1}{e-s+1} (Y_e - Y_{s-1}) - (Y_t - Y_{s-1}) \right\} \right| \qquad (4.8) \\ = \ & \arg\max_{t\in(s,e)} \phi^{UH}(t; s, e). \end{aligned}$$

As note in Section 4.1.3, (4.8) can be seen as the product between the adjusting factor ρ^{UH} and the differential term between the local sum $\left(\sum_{u=s}^{t} y_u\right)$ and the scaled global sum $\left(\frac{t-s+1}{e-s+1} \sum_{u=s}^{e} y_u\right)$. In what follows, we denote this differential term by $\mathcal{D}_{s,e}^{t}$, i.e.,

$$\mathcal{D}_{s,e}^{t} = \frac{t-s+1}{e-s+1} (Y_e - Y_{s-1}) - (Y_t - Y_{s-1})$$

In summary, $\phi^{UH}(b; s, e)$ can be seen as the radius of the tube in its adjusted y-axis when it touches the string at b/n, as well as having the interpretation of being the UH wavelet coefficient of $\tilde{y}_{s,e}$ with respect to $\psi_{s,b,e}$ in absolute value. Therefore the step comparing the squeezed tube radius to λ is equivalent to the hard-thresholding of wavelet coefficients, and it justifies setting λ equal to the universal threshold.

We now derive the locating function for the TS algorithm. Conditional on the string touching the tube at t/n, let g_t indicate whether it touches its upper bound ($g_t = 1$) or lower bound ($g_t = -1$). Initially, as the bounds of the tube approach the string, we note that the first knot is chosen as

$$b = \arg_{t\in(0,n)} \max_{g_t=\pm 1} g_t \cdot \left(\frac{t}{n} Y_n - Y_t \right). \tag{4.9}$$

With the convention that $g_0 = g_n = 0$, further knots on a generic interval (s, e) are located as

$$\begin{aligned} b &= \arg_{t\in(s,e)} \max_{g_t=\pm 1} \phi^{TS}(t; s, e), \text{ where} \\ \phi^{TS}(t; s, e) &= \begin{cases} g_t \cdot \mathcal{D}^t_{s,e} & \text{if } g_{s-1} = g_e, \\ \frac{e-s+1}{(e-s+1)(g_t - g_{s-1}) - (t-s+1)(g_e - g_{s-1})} \cdot \mathcal{D}^t_{s,e} & \text{if } g_{s-1} \neq g_e. \end{cases} \end{aligned}$$

To compare the factors multiplied to the differential term $\mathcal{D}^t_{s,e}$ in ϕ^{UH} and ϕ^{TS}, we quote the following lemma from Venkatraman (1993). Supposing the signal f is piecewise constant and there is no noise present in the observations, Lemma 4.1 implies that the maximum of ϕ^{UH} is then attained only at the true breakpoints of f at every iteration of the UH algorithm.

Lemma 4.1 (Lemma 2.2 of Venkatraman (1993))**.** *Let $m > 0$ be an integer and $0 = a_0 < a_1 < \ldots < a_m < a_{m+1} = 1$. We choose μ_i, $i = 0, \ldots, m$ such that $\mu_i \neq \mu_{i+1}$ and*

$$\sum_{i=0}^{m} (a_{i+1} - a_i)\mu_i = 0.$$

Then we can define a piecewise constant function f whose breakpoints are denoted

by $a_i,\ i = 1, \ldots, m$ *as*

$$f(x) = \mu_i \text{ for } x \in (a_i, a_{i+1}],\ i = 0, \ldots, m.$$

Define the function Φ^{UH} *as*

$$\Phi^{UH}(x) = \frac{\sum_{j=1}^{i}(a_j - a_{j-1})\mu_{j-1} + (x - a_i)\mu_i}{\sqrt{x(1-x)}}, \tag{4.10}$$

for $x \in [a_i, a_{i+1}]; 0 \le i \le m$. *Denote*

$$\Phi^* = \max_{x\in(0,1)} \left|\Phi^{UH}(x)\right| \text{ and } x^* = \arg\max_{x\in(0,1)} \left|\Phi^{UH}(x)\right|$$

such that $\Phi^{UH}(x^*) = \Phi^*$. *Then there exists* $1 \le i \le m$ *such that* $a_i = x^*$, *i.e., the maximum of* $|\Phi^{UH}|$ *can only be attained at one of* a_i*'s.*

Simple algebra shows that Φ^{UH} is equivalent to ϕ^{UH} for $x = t/n \in (0, 1)$. The equivalent of Φ^{UH} for the TS technique is defined using the notation of Lemma 4.1 as

$$\Phi^{TS}(x) = \frac{\sum_{j=1}^{i}(a_j - a_{j-1})\mu_{j-1} + (x - a_i)\mu_i}{\alpha_1 x + \alpha_2(1-x)} \tag{4.11}$$

for $x \in [a_i, a_{i+1}];\ 0 \le i \le m$, where $\alpha_1, \alpha_2 \in \{0, \pm 1, \pm 2\}$ subject to the condition $|\alpha_1 + \alpha_2| = 2$. The particular values taken by α_1, α_2 depend on whether the string touches the lower or upper bound at the start and end of the segment defined by $[a_i, a_{i+1}]$.

Figure 4.6 shows interesting characteristics of the two locating functions, where the UH and TS algorithms are applied to both noiseless and noisy observations of $n = 300$, which are generated from the following model,

$$f(u) = \begin{cases} -4 & \text{for } u \in (0, 1/3], \\ 0 & \text{for } u \in (1/3, 2/3], \\ 5 & \text{for } u \in (2/3, 1]. \end{cases} \tag{4.12}$$

First, consider the example with noiseless observations (dashed lines). The upper

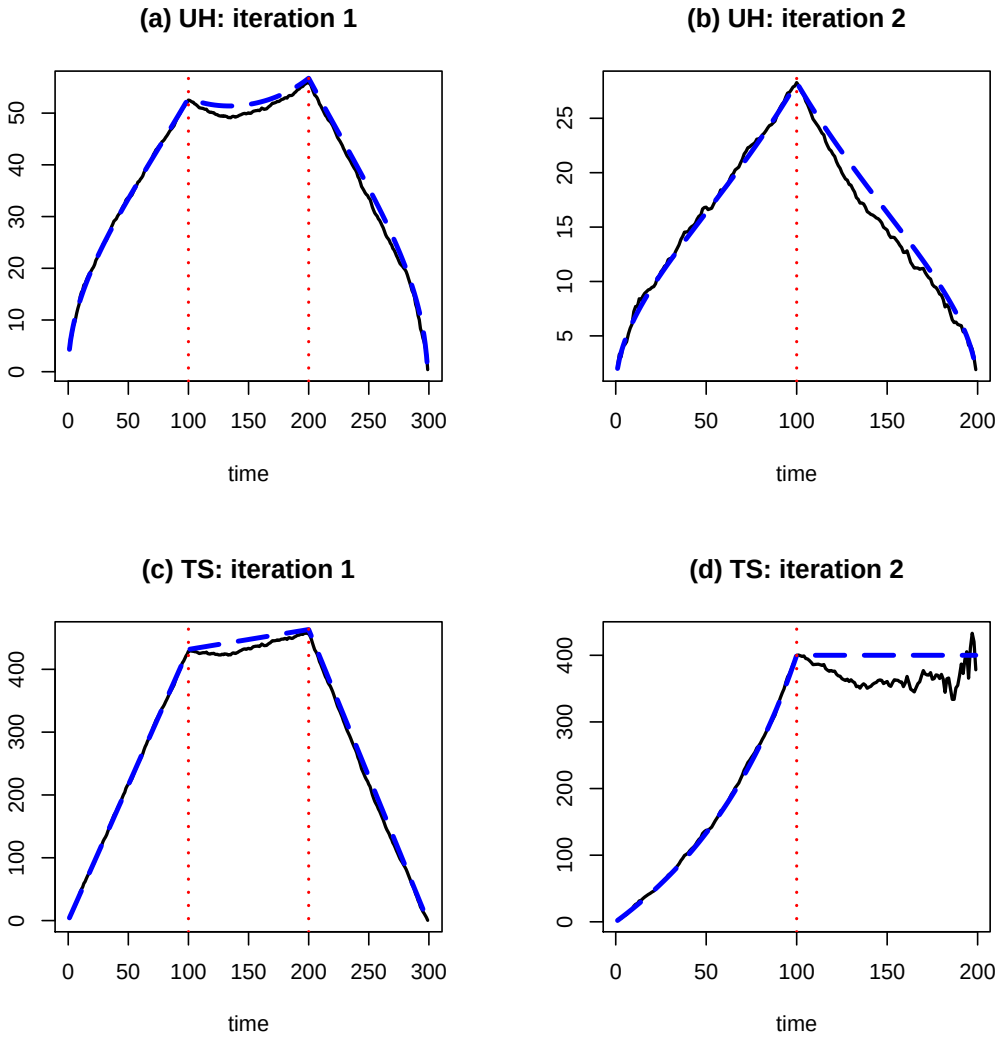


Figure 4.6: (a), (b) $\left|\phi^{UH}(t;s,e)\right|$ at iteration 1, 2; (c), (d) $\left|\phi^{TS}(t;s,e)\right|$ at iteration 1, 2; red dotted (vertical): true breakpoints, blue dashed: noiseless observations, black solid: noisy observations.

panel shows ϕ^{UH} at first two iterations ($(s, e) : (1, 300) \to (1, 200)$), where it is clear that the (local) maxima are attained exactly at the true breakpoints ($t = 100,\ 200$). The lower panel shows ϕ^{TS} at first two iterations, where two different shapes of the locating function are observed. ϕ^{TS} is piecewise linear at the first iteration, while at the second iteration, it reaches a plateau at $t = 100$ and remains constant on $[100, 200)$.

For a piecewise constant signal function f, either shape can occur at each iteration of the TS algorithm, depending on which side of the tube the string has been attached to in previous iterations, i.e. on the values of g_s, g_e and g_b. In either case, it is clear that ϕ^{TS} does not "point out" the locations of true breakpoints as distinctively as ϕ^{UH} does, since the change in the derivative of ϕ^{TS} is not as dramatic as in that of ϕ^{UH} around each breakpoint. Thus we conclude that there is no theoretical equivalent of Lemma 4.1 for Φ^{TS}. This difference may lead to the TS estimate reflecting the true breakpoint structure less accurately than the UH estimate when the underlying function f is piecewise constant, and we expand more on this point in the context of breakpoint detection in the next section.

4.2.2 Link to breakpoint detection

A theoretical study of a family of test statistics for breakpoint detection was made in Brodsky & Darkhovsky (1993). Their study, in light of the relationship of these test statistics to ϕ^{UH} and ϕ^{TS}, supports our observations of the previous section on the "alertness" of the locating function of the UH technique in comparison with that of the TS technique.

In Chapter 3.5 of the book, the problem of *a posteriori* (retrospective) breakpoint detection was considered, where the task was to find an abrupt change in the mean value of a random sequence. Section 2.3 of this thesis provides a description of *a posteriori* breakpoint detection in contrast to the *on-line* approach as well as a survey of the retrospective breakpoint detection methods.

Let $\{x_t\}_{t=1}^n$ be a realisation of a Gaussian process with at most one breakpoint in its mean and otherwise i.i.d., and let X_t be the integrated process of x_t, i.e.

$X_t = \sum_{u=1}^{t} x_u$. Then, a family of test statistics indexed by δ was proposed as

$$d_\delta(t) = \left\{ \frac{t}{n} \left(1 - \frac{t}{n} \right) \right\}^\delta \cdot \left\{ \frac{1}{t} X_t - \frac{1}{n-t} \left(X_n - X_t \right) \right\}, \tag{4.13}$$

where $t \in \{1, \ldots, n-1\}$ and $\delta \in [0, 1]$. A breakpoint candidate is chosen as

$$\hat{b}_\delta = \arg\max_t |d_\delta(t)|$$

and if $|d_\delta(\hat{b}_\delta)|$ exceeds a test criterion, $\hat{b}_\delta$ becomes the estimated breakpoint. It can be shown with simple algebra that $d_{1/2}$ corresponds to ϕ^{UH}. In the case of d_1, it corresponds to ϕ^{TS} only when the string is attached to the same side of tube at $t = s/n$ and $t = e/n$, i.e. at the first iteration of the TS algorithm and each time when $g_s = g_e$ later on.

Below we summarise the asymptotic results from Brodsky & Darkhovsky (1993) on the probabilities of type I error (false alarm, i.e. the test statistic exceeding the test criterion although there is no breakpoint), type II error (false tranquillity, i.e. the test statistic being smaller than the test criterion although there is a breakpoint) and the estimation error in the distance between the detected and true breakpoints. Note that the single breakpoint in the following (b), (c) is constrained to exist within $[a_1, a_2]$ where $0 < a_1 < a_2 < 1$, which is in accordance with the assumption (4.4) made in Fryzlewicz (2007) for the consistency of the UH technique.

(a) When there is no breakpoint present in the observations, the asymptotic rate of convergence for the probability of a type I error increases in δ, i.e. d_1 is asymptotically the best in not causing any false alarm.

(b) When there is a single breakpoint, the asymptotic rate of convergence for the probability of a type II error decreases in δ, i.e. d_0 is asymptotically the best at detecting that there is a breakpoint.

(c) When there is a single breakpoint, say b, the asymptotic rate of convergence for the estimation error probability $\mathbb{P}\left(\left| \hat{b}_\delta - b \right| > \xi \right) \to 0$ is maximised when $\delta = 1/2$, i.e., $d_{1/2}$ is asymptotically the best at estimating the location of the breakpoint.

Note that the above (a) and (b) are obtained under the assumption that the same critical value is used for all $d_\delta(t)$, $\delta \in [0,1]$. Then, for a fixed critical value, the rate of convergence for the probabilities of type I and type II errors are optimised when $\delta = 0$ and $\delta = 1$, respectively.

Suppose now that we choose the critical value C_δ (depending on δ) such that the probability of a type I error is fixed at α. Provided the i.i.d. noise satisfies $\epsilon_t \sim \mathcal{N}(0,1)$, Theorem 3.5.1 of Brodsky & Darkhovsky (1993) implies that

$$C_0 = \sqrt{\frac{2A}{\Delta n}},\ C_{1/2} = \sqrt{\frac{2A}{n}} \text{ and } C_1 = \sqrt{\frac{A}{2n}}, \tag{4.14}$$

where $A = -\log\alpha$ and $\Delta = \min(a_1(1-a_1), a_2(1-a_2))$.

With the above critical values, we can compare the rate of convergence at which the probability of a type II error tends to 0 for different choices of δ. Let $\beta_\delta(n)$ denote the probability of a type II error for each δ. Further, denote the magnitude of the jump at the breakpoint b by h, and define $p = b(1-b) \leq 1/4$. It was noted in Brodsky & Darkhovsky (1993) that when the critical value did not satisfy $C_\delta < hp^\delta$, the probability of a type II error was positive for all n and tended to 1 as $n \to \infty$. Therefore assuming $C_\delta < hp^\delta$, we obtain the following from their Theorem 3.5.2.

$$\beta_\delta(n) \sim \exp\left(-\frac{n(hp^\delta - C_\delta)^2}{2p^{2\delta-1}}\right) = \exp\left(-\frac{n\Theta_\delta}{2}\right). \tag{4.15}$$

By plugging in C_δ from (4.14), each Θ_δ is obtained as

$$\Theta_0 = \left(h\sqrt{p} - \sqrt{\frac{2pA}{\Delta n}}\right)^2,\ \Theta_{1/2} = \left(h\sqrt{p} - \sqrt{\frac{2A}{n}}\right)^2,\ \Theta_1 = \left(h\sqrt{p} - \sqrt{\frac{A}{2pn}}\right)^2.$$

Recalling that the true breakpoint (if it exists) satisfies $b \in [a_1, a_2]$, we have $p \geq \Delta$ and thus $2p/\Delta \geq 2$ and $1/(2p) \geq 2$. Therefore we derive that $\Theta_{1/2} \geq \Theta_\delta$, $\delta = 0, 1$, i.e. when the type 1 error probability is fixed, the rate of convergence for probability of a type II error is better for $\delta = 1/2$ than for $\delta = 0, 1$.

In the above sense, ϕ^{UH} is more alert at breakpoint detection, in detecting both its presence and location, compared with ϕ^{TS}. Combined with the ob-

servation made in Section 4.2.1, we conclude that when estimating a piecewise constant signal with the emphasis on breakpoint detection, it is likely that the UH technique would perform better than the TS technique.

4.3 Possible lessons and directions for future research

While the comparison study between the UH and TS techniques is interesting in itself, it also provides, by establishing links between them, common ground on which the two methods can learn lessons from each other. Below we list some of such lessons that can potentially lead to new developments in the area of nonparametric function estimation.

Choice of threshold

The UH algorithm uses the universal threshold $\sigma\sqrt{2\log n}$ as the critical radius λ. By comparing the multiplying factors of ϕ^{UH} and ϕ^{TS}, we can derive the corresponding critical radius for the multiscale TS algorithm. The equivalent of ρ^{UH} for the multiscale TS algorithm, say ρ^{TS}, satisfies

$$\frac{\rho^{TS}(b;s,e)}{\rho^{UH}(b;s,e)} = C_{s,e}\sqrt{e-s+1},$$

where $C_{s,e}$ is a constant depending on $(b-s+1)/(e-s+1)$, g_s and g_e. Therefore $C_{s,e}\sigma\sqrt{2n\log n}$ can be used as a stopping radius in the multiscale TS algorithm.

In Davies & Kovac (2001), the use of $C_0\sigma\sqrt{n}$ as the global radius was proposed for the uniscale TS algorithm, where C_0 was chosen as a certain quantile of the sup-norm of standard Brownian motion. In order to achieve consistency (in the sense that e.g. constant signals are estimated as constant with probability tending to 1), C_0 may need to converge slowly to infinity, which leads to the two radii (or thresholds) being comparable in terms of their order of magnitude.

UH basis selection

The mean-square consistency result given in Fryzlewicz (2007) holds for any UH basis as long as the breakpoint in each wavelet vector is not too "unbalanced" in the sense of (4.4). The TS algorithm provides yet another way of constructing a UH basis besides the top-down selection method proposed in Fryzlewicz (2007).

Local squeezing

To improve the convergence rate at local extremes, Davies & Kovac (2001) combined the TS technique with a multiresolution method (see (4.6)), applying an additional local squeezing step to the TS estimate. It may be possible to derive a similar theoretical result on the estimated UH residuals $\{y_t - \hat{f}_t^{UH}\}_{t=1}^n$ and apply an analogous local squeezing to obtain a sharper estimate.

On the other hand, although it does not contain explicit local squeezing, the original UH algorithm as described in Section 4.1.1 (and by Fryzlewicz (2007)) obtains the UH wavelet decomposition down to the finest scale, and then applies the hard-thresholding of wavelet coefficients. This can be seen as a replacement for / equivalent of the local squeezing used in Davies & Kovac (2001), as it enhances the adaptivity of the UH estimator. Similar modification can readily be made to our version of the TS algorithm.

Controlling the total variation

The total variation penalty in (4.5) restricts the string to be attached to one of the bounds of the tube at a knot, rather than connecting thus found knot and its adjacent knots with straight lines. Therefore, by modifying the re-arrangement step in the UH algorithm, similar control over the total variation of the estimated function could be achieved.

Extensions to non-Gaussian error distributions

In practice, the assumption of Gaussian noise imposed on $\epsilon_t,\ t = 1, \ldots, n$ is violated in many nonparametric estimation problems, such as Poisson intensity or volatility estimation. In Dümbgen & Kovac (2009), extensions of taut strings were discussed under the assumption that the noise followed a distribution from the exponential family. Their final estimate was obtained

as the transformation of $\hat{f}^{TS}$, the estimate from the least squares setting in (4.5), via a known function. The same arguments may be applied to $\hat{f}^{UH}$ when the prior knowledge on the noise distribution is available.

On the other hand, for the cases where the exact form of the relationship between the mean and variance of noise distribution is unknown, a data-driven, wavelet-based estimation technique was proposed in Fryzlewicz *et al.* (2008), where the use of UH wavelets is readily applicable. By treating the variance stabilisation step of the proposed technique as the adjustment of the y-axis, its extension to the TS technique is also attainable via applying an appropriate adjusting factor to the string and the integrated process.

4.4 Link to Chapter 3 and Chapter 5

The connection between the piecewise constant estimation and the time series segmentation problems can easily be drawn; in Chapter 3, it is assumed that the autocovariance functions of the time series change over time in a piecewise constant manner. A less apparent link between the problem discussed in this chapter and the high-dimensional variable selection problem, which is addressed in Chapter 5, is shown later in this section by treating the additive model in (4.1) as a linear regression model.

We first show that the UH technique has a close relationship with the time series segmentation methodology developed in Chapter 3, especially with its binary segmentation step. Recalling how a breakpoint candidate ν is found in the binary segmentation algorithm in Section 3.2.2,

$$\nu = \arg\max_{b\in(s,e)} \left| \sqrt{\frac{e-b}{(e-s+1)\cdot(b-s+1)}} \sum_{t=s}^{b} Y_{t,T}^2 - \sqrt{\frac{b-s+1}{(e-s+1)\cdot(e-b)}} \sum_{t=b+1}^{e} Y_{t,T}^2 \right|,$$

it is clear that ν is chosen among $b \in (s,e)$ at which the inner product between an Unbalanced Haar vector $\psi_{s,b,e}$ and $\left\{Y_{t,T}^2\right\}_{t=s}^{e}$ is maximised. Thus from (4.8), we can derive the connection between the breakpoint detection methodology and the basis selection step of the UH technique.

Chapter 3 shows the consistency of the detected breakpoints in terms of their

total number and locations, which is in line with the arguments on the alertness of the locating function ϕ^{UH} in Section 4.2.2. However, the test criterion of our breakpoint detection method is greater than that of the UH method. This can be understood from the fact that, when the aim of analysis is to obtain consistent breakpoint estimates from correlated observations, we need a test criterion greater than that for producing a consistent piecewise constant estimate of an unknown function, which may or may not be piecewise constant itself.

We now discuss the TS technique in the context of both breakpoint detection and high-dimensional variable selection problems. One way to re-write the model in (4.1) is in the following linear regression form,

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\beta + \epsilon, \\ \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} &= \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ & \vdots & & \ddots & \vdots \\ 1 & 1 & 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} f_1 \\ f_2 - f_1 \\ \vdots \\ f_n - f_{n-1} \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}. \end{aligned} \tag{4.16}$$

Under the assumption that the mean of $\{y_t\}_{t=1}^n$ does not change too frequently (i.e. the breakpoints in $\{f_t\}_{t=1}^n$ are sparse and only a small number of $\beta_t = f_t - f_{t-1}$ are nonzero), we can view this function estimation problem as both segmentation problem and high-dimensional linear regression problem with a sparse coefficient vector (the dimensionality of (4.16) being equal to the number of observations and thus growing with n).

We note that controlling the total variation of f as in (4.5) is equivalent to controlling the l_1-norm of β in (4.16), which is an approach commonly taken in the variable selection literature (see Section 2.5 for more details). Then, solving the penalised least squares problem in (4.5) corresponds to finding a Lasso solution for the linear regression model in (4.16), see Section 2.5.1.2 for more details of Lasso.

Another variable selection method with l_1-penalty is the Dantzig selector (Section 2.5.3), whose application in this framework has a natural interpretation of imposing a bound over the empirical residuals $\{y_t - \hat{f}_t\}_{t=1}^n$ of the final estimate $\hat{f}$. Originally adopted by Davies & Kovac (2001) in order to control the num-

ber of local extremes (see (4.6) and Section 4.3, **Local squeezing**), this bound results from the particular structure of the design matrix $\mathbf{X}$ in (4.16). Since its column-wise normalised version is

$$\mathbf{X}^* = \begin{pmatrix} 1/\sqrt{n} & 0 & 0 & \cdots & 0 \\ 1/\sqrt{n} & 1/\sqrt{n-1} & 0 & \cdots & 0 \\ & \vdots & & \ddots & \vdots \\ 1/\sqrt{n} & 1/\sqrt{n-1} & 1/\sqrt{n-2} & \cdots & 1 \end{pmatrix},$$

the condition imposed on $\mathbf{y} - \mathbf{X}\hat{\beta} = \mathbf{y} - \hat{\mathbf{f}}$ in (2.37) can be re-written as

$$\|\mathbf{X}^{*T}(\mathbf{y} - \mathbf{X}\hat{\beta})\|_\infty = \max_{1 \le k \le n} \frac{1}{\sqrt{n-k+1}} \left| \sum_{t=k}^{n} (y_t - \hat{f}_t) \right| \le \lambda. \tag{4.17}$$

There is more than one way to re-write the model (4.16) as a linear regression model, and therefore we can impose the multiresolution bound over numerous sums of empirical residuals of the form (4.17).

In fact, our proposed variable selection methodology in Chapter 5 is not related to these l_1-norm regularisation methods. Also, due to high correlations among the columns of $\mathbf{X}$, the conditions imposed for the consistency of the Lasso or the Dantzig selector (e.g. the irrepresentable condition or the uniform uncertainty principle) are not likely to be met by the design matrix in (4.16), and thus this approach to the function estimation problem may not be successful in terms of identifying the breakpoints in piecewise constant functions f. However, we emphasise that the main objective of this section is to connect the seemingly different problems discussed throughout this thesis by means of the piecewise constant estimators studied in this chapter, and to unify them eventually under the common theme of sparse modelling and estimation.

Chapter 5

High-dimensional variable selection via tilting

5.1 Introduction

Inferring the relationship between the response and the explanatory variables in linear models has been widely studied from the point of view of both practical applications and theory. We recall the linear regression model described in Section 2.5

$$\mathbf{y} = \mathbf{X}\beta + \epsilon, \tag{5.1}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$ is an n-vector of the response, $\mathbf{X} = (X_1, \ldots, X_p)$ is an $n \times p$ design matrix and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T \in \mathbb{R}^n$ is an n-vector of i.i.d. random errors.

As noted in Section 2.5, the necessity for an efficient way of handling high-dimensional data has increased dramatically in many fields of sciences, engineering and humanities. For example, a DNA microarray consists of microscopic spots of DNA "features" and it is often the case that the number of features ranges from thousands to tens of thousands, all of which can be viewed as potential explanatory variables (Fan & Lv, 2010). To tackle the challenging problem of estimating the coefficient vector β in high-dimensional situations, substantial progress has been made over the last two decades under the assumption that only

a small number of variables actually contribute to the response, i.e.,

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

is of cardinality much smaller than p. Under such an assumption, identifying $\mathcal{S}$ leads to the improvement of both model interpretability and estimation accuracy, and Section 2.5 of this thesis provides a survey of the literature devoted to the high-dimensional variable selection problem under the sparsity assumption.

One of the difficulties in high-dimensional variable selection is the presence of (possible spurious) non-negligible correlations among the variables. Below we list typical complications encountered in high-dimensional problems due to high correlations among the variables, which were originally pointed out by Fan & Lv (2008).

(a) Irrelevant variables which are highly correlated with the relevant ones can have high marginal correlations with the response.

(b) A relevant variable can be marginally uncorrelated but jointly correlated with the response.

(c) Collinearity can exist among the variables, i.e., $|X_j^T X_k|$ for $j \neq k$ can be close to 1.

In summary, (a)–(c) imply that marginal correlation screening can be misleading as a measure of association between the variables and the response, especially in analysing high-dimensional data. In Section 2.5.5, we review some methods which approach the variable selection problem by taking into account non-negligible correlations among the variables. They examine the strength of association between each variable and the response using measures that are a step further from simple marginal correlation.

We propose another way of measuring the contribution of each variable to the response, which also accounts for the correlation structure among variables. It is accomplished by *tilting* each column X_j such that the impact of other variables X_k, $k \neq j$ on the *tilted correlation* between X_j and $\mathbf{y}$ is reduced, and thus the relationship between the jth covariate and the response can be identified more

accurately. One main ingredient of this methodology is the adaptive choice of those variables X_k whose impact on X_j is to be removed, which is achieved by hard-thresholding the sample correlation matrix of $\mathbf{X}$.

Other key steps in our methodology are: projection of each variable onto a subspace chosen in the hard-thresholding step, and rescaling of such projected variables to obtain a measure of association between the variables and the response which we refer to as the tilted correlation. We show that under certain conditions, the tilted correlation can discriminate between relevant and irrelevant variables, and thus can be used as a tool for variable selection. We also propose an iterative algorithm based on tilting and present its unique features in relation to other existing methods.

The remainder of this chapter is organised as follows. In Section 5.2, we introduce the tilting procedure and study the theoretical properties of tilted correlation in various scenarios. Then in Section 5.3, we propose the TCS algorithm, which iteratively screens the tilted correlations to identify the relevant variables, and compare it to other existing methods. Section 5.4 reports the outcome of a comparative simulation study and Section 5.5 concludes the chapter. Proofs of theoretical results are in Section 5.6.

5.2 Tilting: motivation, definition and properties

5.2.1 Notation and model description

We recap the notation introduced in Section 2.5 as well as providing a description of our model in (5.1). For an n-vector $\mathbf{u} \in \mathbb{R}^n$, we define the l_1- and l_2-norms as $\|\mathbf{u}\|_1 = \sum_i |u_i|$ and $\|\mathbf{u}\|_2 = \sqrt{\sum_i u_i^2}$, and the latter is often referred to as the norm. We denote the ith row of $\mathbf{X}$ by $\mathbf{x}_i = (X_{i,1}, \ldots, X_{i,p})$. Let $\mathcal{D}$ denote a subset of the index set $\mathcal{J} = \{1, \ldots, p\}$. Then $\mathbf{X}_\mathcal{D}$ denotes an $n \times |\mathcal{D}|$-submatrix of $\mathbf{X}$ with $X_j,\ j \in \mathcal{D}$ as its columns for any $n \times p$ matrix $\mathbf{X}$. In a similar manner, $\beta_\mathcal{D}$ denotes a $|\mathcal{D}|$-subvector of a p-vector β with $\beta_j,\ j \in \mathcal{D}$ as its elements. For a given submatrix $\mathbf{X}_\mathcal{D}$, we denote the projection matrix onto the column space

of $\mathbf{X}_{\mathcal{D}}$ by $\Pi_{\mathcal{D}}$. We use the expression $|a_n| \gg |b_n|$ to describe that $|a_n b_n^{-1}| \to \infty$. Finally, C and C' are frequently used to denote generic positive constants.

In what follows, we assume that each column of $\mathbf{X}$ is normalised to have unit norm, and thus the sample correlation matrix of $\mathbf{X}$ is defined as $\mathbf{C} = \mathbf{X}^T\mathbf{X} = (c_{j,k})_{j,k=1}^p$. Further, $\epsilon_i,\ i = 1, \ldots, n$ are assumed to be i.i.d. random noise following a normal distribution $\mathcal{N}(0, \sigma^2/n)$ with $\sigma^2 < \infty$. We note that in the relevant literature, without the unit norm imposed on the columns of $\mathbf{X}$, the sample correlation matrix of $\mathbf{X}$ is defined as $\mathbf{C} = n^{-1}\mathbf{X}^T\mathbf{X}$. It implies that this normalisation step can be seen as dividing every element of $\mathbf{X}$ by $\sqrt{n}$, and therefore the term n^{-1} in the noise variance is justified.

5.2.2 Motivation and definition of tilting

In this section, we introduce the procedure of tilting a variable and define the tilted correlation between each variable and the response.

First, we note that the marginal correlation between each variable X_j and $\mathbf{y}$ has the following decomposition.

$$X_j^T\mathbf{y} = X_j^T\left(\sum_{k=1}^{p}\beta_k X_k + \epsilon\right) = \beta_j + \underline{\sum_{k\in\mathcal{S}\setminus\{j\}}\beta_k X_j^T X_k} + X_j^T\epsilon, \tag{5.2}$$

It shows that the issues (a) and (b) noted in Section 5.1 arise when the underlined summand in (5.2) is non-negligible, due to the large values of $|X_j^T X_k|$ for $k \in \mathcal{S} \setminus \{j\}$. The main idea behind tilting is to transform each X_j in such a way that the corresponding underlined summand for the transformed X_j is zero or negligible, while not distorting the contribution of the jth covariate to the response. Treating the underlined summand as a "bias" term, it is apparent that by projecting X_j onto the space orthogonal to those X_k's which attain large $|X_j^T X_k|$, a corresponding bias term for a thus-transformed X_j would be significantly reduced.

For each X_j, denote the set of such X_k's by $\mathcal{C}_j$. Without prior knowledge of $\mathcal{S}$, one way of selecting $\mathcal{C}_j$ for each X_j is to identify those variables $X_k,\ k \neq j$ which have non-negligible correlations with X_j. A careful choice of $\mathcal{C}_j$ is especially

important when the dimensionality p is high. Informally speaking, when $\mathbf{X}$ has more columns than rows ($p > n$), an n-vector (whether it is X_j or $\mathbf{y}$) may well be approximated by a large number of columns $X_k,\ k \neq j$, which leads to the conclusion that including too many variables in $\mathcal{C}_j$ would distort the association between the jth covariate and the response. However, we also observe that intuitively, those X_k's having small sample correlations with X_j do not significantly contribute to the underlined bias term, and thus can safely be omitted from the set $\mathcal{C}_j$. Therefore, it appears natural to include in $\mathcal{C}_j$ only those variables X_k whose correlations with X_j exceed a certain threshold in magnitude, and this hard-thresholding step is an important element of our methodology.

Based on the above observation, we propose a procedure for selecting $\mathcal{C}_j$ adaptively for each j depending on the sample correlation structure of $\mathbf{X}$. We first find $\pi_n \in (0, 1)$ which acts as a threshold on each off-diagonal entry $c_{j,k},\ j \neq k$ of the sample correlation matrix $\mathbf{C}$, identifying whether the sample correlation between X_j and X_k is non-negligible. Then, the subset $\mathcal{C}_j$ for each variable X_j is obtained as

$$\mathcal{C}_j = \{k \neq j : \ |X_j^T X_k| = |c_{j,k}| > \pi_n\}.$$

Tilting a variable X_j is defined as the procedure of projecting X_j onto the orthogonal complement of the space spanned by $X_k,\ k \in \mathcal{C}_j$, which reduces to zero the impact of those X_k's on the association between the projected version of X_j and $\mathbf{y}$.

Hard-thresholding was previously adopted for the estimation of a high-dimensional covariance matrix, although this has not been done in the context of variable selection to the best of our knowledge. In Bickel & Levina (2008), an estimator obtained by hard-thresholding the sample covariance matrix was shown to be consistent with the choice of $C\sqrt{\log p/n}$ as the threshold, provided the covariance matrix was appropriately sparse and the dimensionality p satisfied $\log p/n \to 0$. A similar result was reported in El Karoui (2008) with the threshold of magnitude $Cn^{-\gamma}$ for some $\gamma \in (0, 1/2)$. Our theoretical choice of threshold π_n is described in Section 5.2.3, where we also briefly compare it to the aforementioned thresholds. In practice, π_n is chosen from the off-diagonal elements of the sample correlation matrix $\mathbf{C}$ by controlling the false discovery rate, as presented in Section 5.3.4.

Now we describe the effect of tilting. Let $\tilde{\mathbf{X}}_j$ denote a submatrix of $\mathbf{X}$ with $X_k,\ k \in \mathcal{C}_j$ as its columns, and Π_j the projection matrix onto the space spanned by $X_k,\ k \in \mathcal{C}_j$, i.e.,

$$\Pi_j = \tilde{\mathbf{X}}_j(\tilde{\mathbf{X}}_j^T\tilde{\mathbf{X}}_j)^{-1}\tilde{\mathbf{X}}_j^T.$$

The tilted variable X_j^* for each X_j is defined as

$$X_j^* = (\mathbf{I}_n - \Pi_j)X_j.$$

Then, the correlation between the tilted variable X_j^* and $X_k,\ k \in \mathcal{C}_j$ is reduced to zero, and therefore such X_k's no longer have any impact on $(X_j^*)^T\mathbf{y}$. However, $(X_j^*)^T\mathbf{y}$ cannot directly be used as a measure of association between X_j and $\mathbf{y}$, since the norm of the tilted variable X_j^*, provided $\mathcal{C}_j$ is non-empty, satisfies

$$\|X_j^*\|_2 = X_j^T(\mathbf{I}_n - \Pi_j)X_j < X_j^TX_j = 1.$$

Therefore, we need to rescale $(X_j^*)^T\mathbf{y}$ so as to make it a reliable criterion for gauging the contribution of each X_j to $\mathbf{y}$.

Let a_j and a_{jy} denote the squared proportion of X_j and $\mathbf{y}$ (respectively) represented by $X_k,\ k \in \mathcal{C}_j$, i.e.,

$$a_j = \frac{\|\Pi_j X_j\|_2^2}{\|X_j\|_2^2} \text{ and } a_{jy} = \frac{\|\Pi_j \mathbf{y}\|_2^2}{\|\mathbf{y}\|_2^2}.$$

We denote the tilted correlation between X_j and $\mathbf{y}$ with respect to a rescaling factor s_j by

$$c_j^*(s_j) = s_j^{-1} \cdot (X_j^*)^T\mathbf{y},$$

and propose two rescaling rules below.

Rescaling 1. Decompose $(X_j^*)^T\mathbf{y}$ as

$$\begin{aligned}(X_j^*)^T\mathbf{y} = X_j^T(\mathbf{I}_n - \Pi_j)\mathbf{y} = X_j^T\left\{\sum_{k=1}^{p}\beta_k(\mathbf{I}_n - \Pi_j)X_k + (\mathbf{I}_n - \Pi_j)\epsilon\right\}\\ = \beta_j X_j^T(\mathbf{I}_n - \Pi_j)X_j + \sum_{k\in\mathcal{S}\setminus\mathcal{C}_j, k\neq j}\beta_k X_j^T(\mathbf{I}_n - \Pi_j)X_k + X_j^T(\mathbf{I}_n - \Pi_j)\epsilon. \quad (5.3)\end{aligned}$$

Provided the second and third summands in (5.3) are negligible in comparison with the first, rescaling the inner product $(X_j^*)^T\mathbf{y}$ by $1 - a_j = X_j^T(\mathbf{I}_n - \Pi_j)X_j$ can isolate β_j, which amounts to the contribution of X_j to $\mathbf{y}$, in the sense that $(X_j^*)^T\mathbf{y}/(1-a_j)$ can be represented as β_j plus a "small" term (our theoretical results later make this statement more precise). Motivated by this, we use the rescaling factor $\lambda_j = (1-a_j)$ to define a rescaled version of X_j^* as

$$X_j^\bullet = (1 - a_j)^{-1} \cdot X_j^*$$

and the corresponding tilted correlation as

$$c_j^*(\lambda_j) = (1 - a_j)^{-1} \cdot (X_j^*)^T\mathbf{y} = (X_j^\bullet)^T\mathbf{y}.$$

Rescaling 2. Since $\mathbf{I}_n - \Pi_j$ is also a projection matrix, we note that $(X_j^*)^T\mathbf{y}$ is equal to the inner product between $X_j^* = (\mathbf{I}_n - \Pi_j)X_j$ and $\mathbf{y}_j^* = (\mathbf{I}_n - \Pi_j)\mathbf{y}$, with their norms satisfying $\|X_j^*\|_2 = \sqrt{1 - a_j}$ and $\|\mathbf{y}_j^*\|_2 = \sqrt{1 - a_{jy}} \cdot \|\mathbf{y}\|_2$. By rescaling X_j^* and $\mathbf{y}_j^*$ by $\sqrt{1 - a_j}$ and $\sqrt{1 - a_{jy}}$ respectively, we obtain the vectors

$$X_j^\circ = (1 - a_j)^{-1/2} \cdot X_j^* \text{ and } \mathbf{y}_j^\circ = (1 - a_{jy})^{-1/2} \cdot \mathbf{y}_j^*,$$

whose norms satisfy $\|X_j^\circ\|_2 = \|X_j\|_2$ and $\|\mathbf{y}_j^\circ\|_2 = \|\mathbf{y}\|_2$. Therefore, with the rescaling factor set equal to $\Lambda_j = \{(1 - a_j)(1 - a_{jy})\}^{1/2}$, we define the tilted correlation as

$$c_j^*(\Lambda_j) = \{(1 - a_j)(1 - a_{jy})\}^{-1/2} \cdot (X_j^*)^T\mathbf{y} = (X_j^\circ)^T\mathbf{y}_j^\circ.$$

Figure 5.1 illustrates the above rescaling steps visualised in a three-dimensional space, where a variable X_j is assumed to attain a non-negligible correlation with X_k, $k \neq j$ (i.e. $|X_j^T X_k| > \pi_n$). In the left panel, $c_j^*(\lambda_j)$ (rescaling 1) is equal to the inner product between $X_j^\bullet$ and $\mathbf{y}$, while in the right panel, $c_j^*(\Lambda_j)$ (rescaling 2) is equivalent to the inner product between X_j° and $\mathbf{y}_j^\circ$.

We note that, with the rescaling factor λ_j (rescaling 1), the tilted correlation $c_j^*(\lambda_j)$ coincides with the ordinary least squares estimate of β_j from regressing $\mathbf{y}$

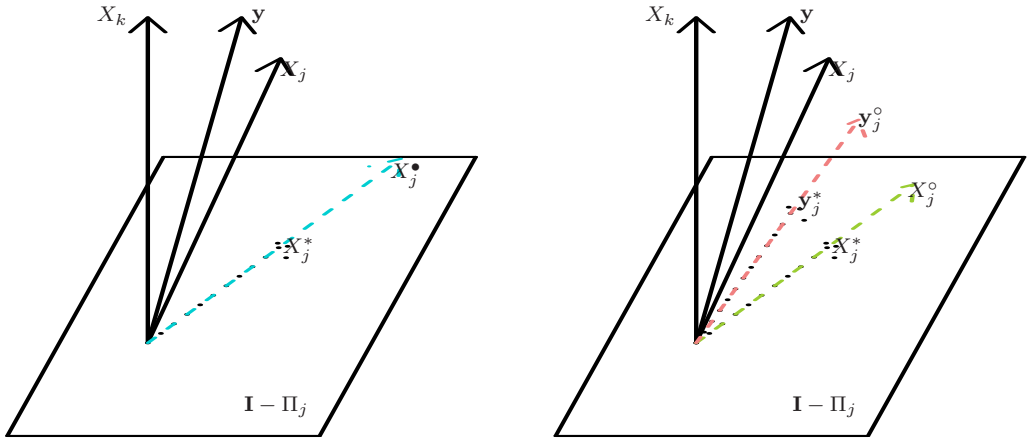


Figure 5.1: 3-dimensional visualisation of the rescaling methods. Rescaling 1: $c_j^*(\lambda_j) = \langle X_j^\bullet, \mathbf{y} \rangle$ (left); rescaling 2: $c_j^*(\Lambda_j) = \langle X_j^\circ, \mathbf{y}_j^\circ \rangle$ (right); X_j^* and $\mathbf{y}_j^*$ are dotted vectors, while their rescaled versions $X_j^\bullet$, X_j° and $\mathbf{y}_j^\circ$ are dashed vectors.

onto X_k, $k \in \mathcal{C}_j \cup \{j\}$. When rescaled by Λ_j (rescaling 2), the tilted correlation coincides with the sample partial correlation between X_j and $\mathbf{y}$ given X_k, $k \in \mathcal{C}_j$ (denoted by $\hat{\rho}_n(j, \mathbf{y}|\mathcal{C}_j)$), up to a constant multiplicative factor $\|\mathbf{y}\|_2$, i.e.,

$$c_j^*(\Lambda_j) = \|\mathbf{y}\|_2 \cdot \hat{\rho}_n(j, \mathbf{y}|\mathcal{C}_j).$$

Although partial correlation is also used in the PC-simple algorithm (see Section 2.5.5), we emphasise that there exists a crucial difference between tilting and PC-simple algorithm, since tilting has an adaptive way of selecting the conditioning subset $\mathcal{C}_j$ for each X_j as described earlier in this section. A detailed discussion on the difference between the two methods is provided in Section 5.3.3. In what follows, whenever the tilted correlation is denoted by c_j^* without specifying the rescaling factor s_j, the relevant statement is valid for either of the rescaling factors λ_j and Λ_j.

Finally, we note that if the set $\mathcal{C}_j$ turns out to be empty for a certain index j, then for such X_j, its tilted correlation with either of the rescaling factors would reduce to standard marginal correlation, which in this case is expected to work well (in measuring the association between the jth covariate and the response) since no other variables are significantly correlated with X_j. In summary, our pro-

posed tilting procedure enables an adaptive "switch" between the use of marginal correlation and tilted correlation for each variable X_j, depending on the sample correlation structure of $\mathbf{X}$.

In the following section, we study some properties of tilted correlation and show that the corresponding properties do not always hold for marginal correlation. This prepares ground for the algorithm proposed in Section 5.3.1 which adopts tilted correlation for variable screening.

5.2.3 Properties of the tilted correlation

In the high-dimensional linear regression literature, various assumptions on the correlation structure of the variables have been made for the theoretical treatment of proposed methods. When establishing such assumptions, two different approaches have been adopted frequently: imposing the conditions either at the "population" level or at the "sample" level. For example, Fan & Li (2001) took the former approach and assumed that the observations $(\mathbf{x}_i, y_i)$ were independent and identically distributed with probability density obeying some regularity conditions. On the other hand, the irrepresentable condition (Zhao & Yu, 2006) and the sparse Riesz condition (Zhang & Huang, 2008) for the lasso, the uniform uncertainty principle (UUP) for the Dantzig selector (Candès & Tao, 2007) and the asymptotic identifiability condition for the extended BIC (Chen & Chen, 2008) impose restrictions on the behaviour of design matrix $\mathbf{X}$ itself, regardless of its being deterministic or a realisation from a random distribution (detailed descriptions of these conditions can be found in Section 2.5.1.2 and Section 2.5.3 of this thesis). To investigate the implications of their conditions, Candès & Tao (2007) showed that a random matrix with i.i.d. Gaussian entries would satisfy the UUP with high probability, and similar arguments were made by Zhang & Huang (2008) as well in support of the sparse Riesz condition.

In studying the theoretical properties of tilted correlation, we make the following assumptions (A1)–(A6) on the linear model in (5.1). Where $\mathbf{X}$ is concerned, we follow the latter approach and impose the restrictions directly on the design matrix itself. Then follow some comments on our conditions in comparison with other assumptions from the relevant literature, either at the population level or

not, to study their implications.

(A1) The number of non-zero coefficients $|\mathcal{S}|$ satisfies $|\mathcal{S}| = O(n^{\delta})$ for $\delta \in [0, 1/2)$.

(A2) The number of variables satisfies $\log p = O(n^{\theta})$ with $\theta \in [0, 1-2\gamma)$ for $\gamma \in (\delta, 1/2)$.

(A3) With the same γ as in (A2), the threshold is chosen as $\pi_n = C_1 n^{-\gamma}$ for some positive constant C_1. Then, we assume that there exists $C > 0$ such that

$$\mathcal{C}_j = \{k \neq j : \ |c_{j,k}| > \pi_n\}$$

is of cardinality $|\mathcal{C}_j| \leq C n^{\xi}$ uniformly over all j, where $\xi \in [0, 2(\gamma - \delta))$.

(A4) Non-zero coefficients satisfy

$$\max_{j \in \mathcal{S}} |\beta_j| < M \text{ and } n^{\mu} \cdot \min_{j \in \mathcal{S}} |\beta_j| \to \infty$$

for $M \in (0, \infty)$ and $\mu \in [0, \gamma - \delta - \xi/2)$.

(A5) There exists $\alpha \in (0, 1)$ satisfying, for all j,

$$1 - X_j^T \Pi_j X_j = 1 - a_j > \alpha.$$

(A6) For those j whose corresponding $\mathcal{C}_j$ satisfies $\mathcal{S} \nsubseteq \mathcal{C}_j$, we have

$$n^{\kappa} \cdot \frac{\|(\mathbf{I}_n - \Pi_j)\mathbf{X}_{\mathcal{S}}\beta_{\mathcal{S}}\|_2^2}{\|\mathbf{X}_{\mathcal{S}}\beta_{\mathcal{S}}\|_2^2} \to \infty,$$

for κ satisfying $\kappa/2 + \mu \in [0, \gamma - \delta - \xi/2)$.

We note that the assumptions (A3), (A5) and (A6), which are imposed on the sample correlation structure of $\mathbf{X}$, are not directly comparable with the sparse Riesz condition or the UUP in the sense that, our assumptions are subject to the specific choice of $\mathcal{C}_j$ for each X_j, while the others are imposed on the submatrices of $\mathbf{X}$, denoted by $\mathbf{X}_{\mathcal{D}}$, uniformly over every $\mathcal{D} \subset \mathcal{J}$ whose cardinality is bounded by $C|\mathcal{S}|$ for some $C > 0$. Below we further discuss the implications of (A1)–(A6).

In (A1) and (A2), we let the sparsity $|\mathcal{S}|$ and the dimensionality p of the linear model grow with the sample size n. The choice of $\pi_n = C_1 n^{-\gamma}$ in (A3) is in agreement with Bickel & Levina (2008) and El Karoui (2008) as their proposed thresholds are also greater than $n^{-1/2}$. We note that this theoretical threshold is not easily applicable, as the rate parameter γ is bounded but unknown. In practice, π_n is chosen by controlling the false discovery rate (Section 5.3.4). The cardinality of $\mathcal{C}_j$ needs to be bounded to guarantee the existence of the projection matrix Π_j as well as to prevent tilted correlations from being distorted (see Section 5.2.2). We now give an example of when (A3) is satisfied.

Suppose for instance that each observation $\mathbf{x}_i$, $i = 1, \ldots, n$ is independently generated from a multivariate normal distribution $\mathcal{N}_p(\mathbf{0}, \Sigma)$ with $\Sigma_{j,k} = \varphi^{|j-k|}$ for some $|\varphi| \in [0, 1)$. Then summarising the Lemma 1 and the subsequent arguments of Kalisch & Bühlmann (2007), we have

$$\max_{j \neq k} \mathbb{P}\left(|c_{j,k} - \Sigma_{j,k}| > C_2 n^{-\gamma}\right) \leq Cn \exp\left(-\frac{C_2(n-4)n^{-2\gamma}}{2}\right)$$

for some $C_2 \in (0, C_1)$ and $C > 0$, which implies that

$$\mathbb{P}\left(\max_{j \neq k} |c_{j,k} - \Sigma_{j,k}| \leq C_2 n^{-\gamma}\right) \geq 1 - \frac{Cnp(p-1)}{2} \cdot \exp\left(-\frac{C_2(n-4)n^{-2\gamma}}{2}\right) \quad (5.4)$$

The right-hand side of (5.4) tends to 1, provided $\log p = O(n^\theta)$ with $\theta \in [0, 1/2 - \gamma)$. Then (A3) holds with probability converging to 1, since for $|j - k| \gg \log n$,

$$|c_{j,k}| \leq |\varphi|^{|j-k|} + C_2 n^{-\gamma} < \pi_n.$$

Intuitively, if some non-zero coefficients converge to zero too rapidly, identifying the corresponding variables as relevant is very difficult. (A4) imposes a lower bound on the non-zero coefficients, which still allows the minimum of the magnitude of non-zero coefficients to decay to 0 as n grows. It also imposes an upper bound, which is needed to ensure that the ratio between the maximum and minimum non-zero coefficients in absolute value does not grow too quickly with n.

(A5) is required to rule out strong collinearity among the variables. Since

$$0 < \alpha < 1 - a_j = \frac{\det\left(\mathbf{X}_{\mathcal{C}_j \cup \{j\}}^T \mathbf{X}_{\mathcal{C}_j \cup \{j\}}\right)}{\det\left(\tilde{\mathbf{X}}_j^T \tilde{\mathbf{X}}_j\right)},$$

we can find a connection between (A5) and the condition requiring strict positive definiteness of the population covariance matrix of $\mathbf{X}$, which is often found in the variable selection literature including Bühlmann *et al.* (2009) and Zou (2006).

Chen & Chen (2008) introduced a new asymptotic identifiability condition for high-dimensional problems, which can be re-written as below (after taking into account the column-wise normalisation of $\mathbf{X}$),

$$\lim_{n\to\infty} \min_{\mathcal{D}\subset\mathcal{J}, |\mathcal{D}|\leq|\mathcal{S}|, \mathcal{D}\neq\mathcal{S}} n(\log n)^{-1} \cdot \frac{\|(\mathbf{I}_n - \Pi_{\mathcal{D}})\mathbf{X}_{\mathcal{S}}\beta_{\mathcal{S}}\|_2^2}{\|\mathbf{X}_{\mathcal{S}}\beta_{\mathcal{S}}\|_2^2} \to \infty. \tag{5.5}$$

Chen & Chen (2008) showed that this identifiability condition was weaker than (i.e. implied by) the sparse Riesz condition. The similarity between (5.5) and (A6) can readily be seen. The difference is that (5.5) is a uniform condition over the entire collection of sets $\mathcal{D}$, whereas (A6) is only required to hold for $\mathcal{C}_j$; however, the rate $n^{-\kappa}$ is less favourable than $\log n/n$. Even with this slower rate replacing $\log n/n$ in (5.5), our condition (A6) is still weaker than the sparse Riesz condition for a certain configuration of δ and ξ.

As far as variable selection is concerned, if the absolute values of tilted correlations for $j \in \mathcal{S}$ are markedly larger than those for $j \notin \mathcal{S}$, we can use the tilted correlations for the purpose of variable screening. In the following Sections 5.2.3.1–5.2.3.3, we study the conditions under which the tilted correlations (with either rescaling factor) satisfy such properties.

5.2.3.1 Scenario 1

In the first scenario, we assume the following condition on $\mathbf{X}$.

Condition 5.1. *There exists $C > 0$ such that*

$$\left|(\Pi_j X_j)^T X_k\right| \leq C n^{-\gamma}$$

for all $j \in \mathcal{J}$ *and* $k \in \mathcal{S} \setminus \mathcal{C}_j,\ k \neq j$.

This condition implies that when X_j is projected onto the space spanned by $X_l,\ l \in \mathcal{C}_j$, any $X_k \in \mathcal{S}$ which is not close to X_j (in the sense that $k \in \mathcal{S} \setminus \mathcal{C}_j$) remains not "too close" to the projected X_j $(= \Pi_j X_j)$. In Section 5.6.1.1, it is shown that Condition 5.1 holds asymptotically when each column X_j is generated independently as a random vector on a sphere of radius 1, which is the surface of the Euclidean ball

$$B_2^n = \left\{ \mathbf{x} \in \mathbb{R}^n : \sum_{i=1}^{n} x_i^2 \leq 1 \right\}.$$

The following theorem states that, under Condition 5.1, the tilted correlations of the relevant variables dominate those of the irrelevant variables.

Theorem 5.1. *Under assumptions (A1)–(A6), if Condition 5.1 holds, then* $\mathbb{P}(\mathcal{E}_1) \to 1$ *where*

$$\mathcal{E}_1 = \left\{ \frac{|c_k^*(s_k)|}{\min_{j \in \mathcal{S}} |c_j^*(s_j)|} \to 0 \text{ for all } k \notin \mathcal{S} \right\}, \tag{5.6}$$

regardless of the choice of the rescaling factor (that is, with $s_j = \lambda_j$ *or* $s_j = \Lambda_j$*). On the event* $\mathcal{E}_1$*, the following holds.*

- $n^{\mu} \cdot c_j^* \to 0$ *for* $j \notin \mathcal{S}$.
- $n^{\mu} \cdot |c_j^*| \to \infty$ *for* $j \in \mathcal{S}$.
- *With the rescaling 1,* $c_j^*(\lambda_j)/\beta_j \to 1$ *when* $\beta_j \neq 0$.

5.2.3.2 Scenario 2

Let $\mathcal{K}$ denote a subset of $\mathcal{J}$ such that $X_k,\ k \in \mathcal{K}$ are either relevant $(k \in \mathcal{S})$ or highly correlated with at least one of the relevant variables $(k \in \cup_{j \in \mathcal{S}} \mathcal{C}_j)$. That is,

$$\mathcal{K} = \mathcal{S} \cup \{ \cup_{j \in \mathcal{S}} \mathcal{C}_j \},$$

and we impose the following condition on the sample correlation structure of $\mathbf{X}_{\mathcal{K}}$.

Condition 5.2. *For each* $j \in \mathcal{S}$*, if* $k \in \mathcal{K} \setminus \{ \mathcal{C}_j \cup \{j\} \}$*, then* $\mathcal{C}_k \cap \mathcal{C}_j = \emptyset$.

In other words, this condition implies that for each relevant variable X_j, if $X_k,\ k \in \mathcal{K}$ is not highly correlated with X_j, there does not exist an $X_l,\ l \neq j, k$, which achieves sample correlations greater than the threshold π_n with both X_j and X_k simultaneously.

Suppose that the sample correlation matrix of $\mathbf{X}_{\mathcal{K}}$ is "approximately bandable", i.e., $|c_{j,k}| > \pi_n$ for any $j, k \in \mathcal{K}$ satisfying $|j-k| \leq B$ and $|c_{j,k}| < \pi_n$ otherwise, with the band width B satisfying $B|\mathcal{S}|^2/p \to 0$. Then, if $\mathcal{S}$ is selected randomly from $\mathcal{J}$ with each $j \in \mathcal{J}$ having equal probability to be selected in $\mathcal{S}$, Condition 5.2 holds with probability bounded from below by

$$\left(1-\frac{4B}{p-1}\right)\cdot\left(1-\frac{8B}{p-2}\right)\cdots\left(1-\frac{4(|\mathcal{S}|-1)B}{p-|\mathcal{S}|+1}\right) \geq \left(1-\frac{4|\mathcal{S}|B}{p-|\mathcal{S}|+1}\right)^{|\mathcal{S}|-1} \to 1.$$

Another example satisfying Condition 5.2 is when each column of $\mathbf{X}_{\mathcal{K}}$ is generated as a linear combination of common factors in such a way that every off-diagonal element of the sample correlation matrix of $\mathbf{X}_{\mathcal{K}}$ exceeds the threshold π_n.

Under this condition, we can derive a similar result as in Scenario 1, with the dominance of the tilted correlations for the relevant variables restricted within $\mathcal{K}$.

Theorem 5.2. *Under (A1)–(A6), if Condition 5.2 holds, then* $\mathbb{P}(\mathcal{E}_2) \to 1$ *where*

$$\mathcal{E}_2 = \left\{ \frac{|c_k^*(s_k)|}{\min_{j\in\mathcal{S}} |c_j^*(s_j)|} \to 0 \text{ for all } k \in \mathcal{K}\setminus\mathcal{S} \right\},$$

regardless of the choice of the rescaling factor (that is, with $s_j = \lambda_j$ *or* $s_j = \Lambda_j$*). On the event* $\mathcal{E}_2$*, the following holds.*

- $n^\mu \cdot c_j^* \to 0$ *for* $j \notin \mathcal{S}$.
- $n^\mu \cdot |c_j^*| \to \infty$ *for* $j \in \mathcal{S}$.
- *With the rescaling 1,* $c_j^*(\lambda_j)/\beta_j \to 1$ *when* $\beta_j \neq 0$.

5.2.3.3 Scenario 3

Finally, we consider a case when $\mathbf{X}$ satisfies a condition weaker than Condition 5.2.

Condition 5.3. *(C1) For each $j \in \mathcal{S}$, if $k \in \mathcal{K} \setminus \{\mathcal{C}_j \cup \mathcal{S}\}$, then $\mathcal{C}_k \cap \mathcal{C}_j = \emptyset$.*

(C2) The marginal correlation between $X_j^* = (\mathbf{I}_n - \Pi_j)X_j$ for $j \in \mathcal{S}$ and $\mathbb{E}\mathbf{y} = \mathbf{X}_{\mathcal{S}}\beta_{\mathcal{S}}$ satisfies

$$n^{\mu} \cdot \inf_{j \in \mathcal{S}} \left|(X_j^*)^T \mathbf{X}_{\mathcal{S}} \beta_{\mathcal{S}}\right| \to \infty.$$

It is clear that Condition 5.2 is stronger than (C1), as the latter does not impose any restriction between $\mathcal{C}_j$ and $\mathcal{C}_k$ if both $j, k \in \mathcal{S}$. Bühlmann *et al.* (2009) placed a similar lower bound as that in (C2) on the population partial correlation $\rho_n(j, \mathbf{y}|\mathcal{D})$ of the relevant variables X_j, $j \in \mathcal{S}$ for any subset $\mathcal{D} \subset \mathcal{J} \setminus \{j\}$ satisfying $|\mathcal{D}| \leq |\mathcal{S}|$. Combined with the assumptions (A4)–(A5), (C2) rules out the ill-posed case where the configuration of non-zero parameters β_j, $j \in \mathcal{S}$ cancels out the "tilted covariance" among the relevant variables. We clarify this statement more precisely in the proof of Theorem 5.3. It is shown in Section 5.6.3 that Condition 5.3 is satisfied if Condition 5.2 holds, and thus Condition 5.3 itself is weaker than Condition 5.2.

With Condition 5.3, we can show similar results to those in Theorem 5.2.

Theorem 5.3. *Under (A1)–(A6), if Condition 5.3 holds, then $\mathbb{P}(\mathcal{E}_3) \to 1$ where*

$$\mathcal{E}_3 = \left\{ \frac{|c_k^*(s_k)|}{\min_{j \in \mathcal{S}} |c_j^*(s_j)|} \to 0 \text{ for all } k \in \mathcal{K} \setminus \mathcal{S} \right\},$$

regardless of the choice of the rescaling factor (that is, with $s_j = \lambda_j$ or $s_j = \Lambda_j$). On the event $\mathcal{E}_3$, the following holds.

- $n^{\mu} \cdot c_j^* \to 0$ *for* $j \notin \mathcal{S}$.
- $n^{\mu} \cdot |c_j^*| \to \infty$ *for* $j \in \mathcal{S}$.

In contrast to Scenario 2, tilted correlations $c_j^*(\lambda_j)$ no longer necessarily converge to β_j as $n \to \infty$ in this scenario.

Marginal correlations $X_j^T \mathbf{y}$ for $j \in \mathcal{S}$ cannot be expected to have the same dominance over those for $j \notin \mathcal{S}$ as in Theorems 5.1–5.3, unless every off-diagonal element of the sample correlation matrix $\mathbf{C}$ is uniformly small, which is an unrealistic assumption especially in high-dimensional problems. On the other hand,

Conditions 5.1–5.3 specify when the tilted correlation can satisfy the desired properties, while allowing the presence of high correlations among the variables. Below we further expand on this point with a simple example.

The following set-up is consistent with Condition 5.3: $p = 3$, $\mathcal{S} = \{1, 2\}$, noise is not present, $|c_{1,3}|$ and $|c_{2,3}|$ exceed the threshold. In this case, even when $c_{1,2}$, $c_{1,3}$, $c_{2,3}$ and the non-zero coefficients β_1, β_2 are chosen so that the marginal correlation screening fails in the sense that

$$|X_3^T \mathbf{y}| > \max(|X_1^T \mathbf{y}|, \ |X_2^T \mathbf{y}|),$$

we have $|(X_3^*)^T \mathbf{y}| = 0$ and thus tilted correlation screening is successful.

Scenarios 1–3 do not imply tilting fails when the conditions therein are not met. Rather, they are imposed in order to study when tilting can succeed and what can be expected in such cases. In the next section, we use the theoretical properties of tilted correlations derived in this section to construct a variable screening algorithm.

5.3 Application of tilting

Recalling the issues (a)–(c) listed at the beginning of Section 5.1, which are typically encountered in high-dimensional problems, it is clear that tilting is specifically designed to tackle the occurrence of (a) and (b).

First turning to (a), for an irrelevant variable X_j which attains high marginal correlation with $\mathbf{y}$ due to its high correlations with the relevant variables X_k, $k \in \mathcal{C}_j \cap \mathcal{S}$, the impact of those high correlations is reduced to 0 in the tilted correlation of X_j and $\mathbf{y}$, and thus tilted correlation provides a more accurate measure of its association with $\mathbf{y}$. Similar arguments apply to (b), where tilting is capable of fixing small marginal correlations between relevant variables and $\mathbf{y}$. As for (c), it is common practice to impose assumptions which rule out strong collinearity among variables, and we have also followed this route.

In what follows, we present one way of exploiting our theoretical study in Section 5.2.3, in the form of an algorithm which iteratively applies the tilting procedure.

5.3.1 Tilted correlation screening algorithm

In Scenario 3, under a relatively weaker condition than those in Scenarios 1–2, it is shown that the tilted correlations of relevant variables dominate those of irrelevant variables within $\mathcal{K}$. Even though $\mathcal{K}$ is unknown in practice, as its knowledge involves that of $\mathcal{S}$, we can exploit the theoretical results by iteratively screening both marginal correlations and tilted correlations within a carefully chosen subset of variables.

When every off-diagonal entry of the sample correlation matrix is small, marginal correlation screening can be used as a reliable way of measuring the strength of association between each X_j and $\mathbf{y}$, and indeed, c_j^* for the variable X_j with an empty $\mathcal{C}_j$ is equal to the marginal correlation $X_j^T\mathbf{y}$ regardless of the choice of the rescaling factor s_j. Therefore if a variable X_k with $\mathcal{C}_k = \emptyset$ achieves the maximum marginal correlation in absolute value, such X_k is likely to be relevant. On the other hand, if $\mathcal{C}_k \neq \emptyset$, then the high marginal correlation between X_k and $\mathbf{y}$ may have resulted from the high correlations of X_k with $X_j,\ j \in \mathcal{C}_k \cap \mathcal{S}$, even when X_k itself is not relevant. In this case, by screening the tilted correlations of $X_j,\ j \in \mathcal{C}_k \cup \{k\}$, we can choose the variable attaining the maximum $|c_j^*|$ as a relevant variable. In either way, one variable is selected and we add it to the "active set" $\mathcal{A}$ which represents the currently chosen model.

As the next step, we update the linear model by projecting it onto the orthogonal complement of the current model space $\mathbf{X}_\mathcal{A}$, i.e.,

$$(\mathbf{I}_n - \Pi_\mathcal{A})\mathbf{y} = (\mathbf{I}_n - \Pi_\mathcal{A})\mathbf{X}\beta + (\mathbf{I}_n - \Pi_\mathcal{A})\epsilon. \tag{5.7}$$

With the updated response and design matrix, we continue the above screening procedure iteratively. Below we summarise the above arguments in the form of an algorithm, which is referred to as the tilted correlation screening algorithm (TCS algorithm) throughout the chapter.

TCS algorithm

Step 0 Start with an empty active set $\mathcal{A} = \emptyset$, current residual $\mathbf{z} = \mathbf{y}$, and current design matrix $\mathbf{Z} = \mathbf{X}$.

Step 1 Find the variable which achieves the maximum marginal correlation with $\mathbf{z}$ in absolute value, and let

$$k = \arg\max_{j \notin \mathcal{A}} |Z_j^T \mathbf{z}|.$$

Identify $\mathcal{C}_k = \{j \notin \mathcal{A},\ j \neq k :\ |Z_k^T Z_j| > \pi_n\}$ and if $\mathcal{C}_k = \emptyset$, let $k^* = k$ and go to Step 3.

Step 2 If $\mathcal{C}_k \neq \emptyset$, screen the tilted correlations c_j^* between Z_j and $\mathbf{z}$ for $j \in \mathcal{C}_k \cup \{k\}$ and find

$$k^* = \arg\max_{j \in \mathcal{C}_k \cup \{k\}} |c_j^*|.$$

Step 3 Add k^* to $\mathcal{A}$, and update the current residual $\mathbf{z}$ and the current design matrix $\mathbf{Z}$ as

$$\mathbf{z} \leftarrow (\mathbf{I}_n - \Pi_{\mathcal{A}})\mathbf{y} \text{ and } \mathbf{Z} \leftarrow (\mathbf{I}_n - \Pi_{\mathcal{A}})\mathbf{X},$$

respectively. Further, rescale each column $Z_j,\ j \notin \mathcal{A}$ of $\mathbf{Z}$ to have unit norm.

Step 4 Repeat Steps 1–3 until the cardinality of active set $|\mathcal{A}|$ reaches a pre-specified $m < n$.

As noted at the beginning of this section, the results in Theorems 5.2–5.3 are restricted within $\mathcal{K} \subset \mathcal{J}$, which is unknown without the knowledge of $\mathcal{S}$. However, Steps 1–2 can be interpreted as an attempt to remain within the set $\mathcal{K}$, since we either

- directly choose an index k which is believed to lie in the set $\mathcal{S}$ (its corresponding Z_k attains the maximum marginal correlation with the current residual $\mathbf{z}$), or
- screen the tilted correlations within $\mathcal{C}_k \cup \{k\}$ which is likely to contain at least one relevant variable, recalling that $\mathcal{K} = \mathcal{S} \cup \{\cup_{j \in \mathcal{S}} \mathcal{C}_j\}$.

In Step 4, we need to specify m which acts as a stopping index in the TCS algorithm. The TCS algorithm iteratively builds a solution path of the active set $\mathcal{A}_{(1)} \subset \cdots \subset \mathcal{A}_{(m)} = \mathcal{A}$, and therefore the final model $\hat{\mathcal{S}}$ can be chosen as

either one of the submodels $\mathcal{A}_{(i)}$ or a subset of $\mathcal{A}$. We discuss the selection of $\hat{\mathcal{S}}$ in Section 5.3.2.

5.3.1.1 Updating step in the TCS algorithm

During the application of the TCS algorithm, the linear regression model (5.1) is updated in Step 3 by projecting both $\mathbf{y}$ and $\mathbf{X}$ onto the orthogonal complement of the current model space spanned by $\mathbf{X}_{\mathcal{A}}$. Therefore it is interesting to observe that in Step 1, with a non-empty active set $\mathcal{A}$, the subset of indices $j \notin \mathcal{A},\ j \neq k$ whose corresponding $Z_j (= (\mathbf{I}_n - \Pi_{\mathcal{A}})X_j)$ attain non-negligible sample correlations with $Z_k (= (\mathbf{I}_n - \Pi_{\mathcal{A}})X_k)$ is equal to the following set

$$\mathcal{C}_{k|\mathcal{A}} = \{j \notin \mathcal{A}, j \neq k :\ \hat{\rho}_n(j, k|\mathcal{A}) > \pi_n\}, \tag{5.8}$$

where $\hat{\rho}_n(j, k|\mathcal{A})$ denotes the sample partial correlation between X_j and X_k conditional on $\mathbf{X}_{\mathcal{A}}$. Then, with a non-empty $\mathcal{A}$, the tilted correlation c_j^* in Step 2 measures the association between X_j and $\mathbf{y}$ conditional on both the current model $\mathbf{X}_{\mathcal{A}}$ and the subset of variables $X_l,\ l \in \mathcal{C}_{j|\mathcal{A}}$ adaptively chosen for each $j \in \mathcal{C}_{k|\mathcal{A}} \cup \{k\}$,

While (A1)–(A2) and (A4) remain unchanged after the updating step, the assumptions (A3), (A5)–(A6) can be re-written for the updated current residual and current design matrix as below.

(A3') We assume that there exists $C > 0$ such that $\mathcal{C}_{j|\mathcal{A}}$ defined as in (5.8) is of cardinality $|\mathcal{C}_{j|\mathcal{A}}| \leq Cn^{\xi}$ uniformly over all $j \notin \mathcal{A}$.

(A5') There exists $\alpha \in (0, 1)$ satisfying, for all $j \notin \mathcal{A}$,

$$\frac{X_j^T(\mathbf{I}_n - \Pi_{\mathcal{A}\cup\mathcal{C}_{j|\mathcal{A}}})X_j}{X_j^T(\mathbf{I}_n - \Pi_{\mathcal{A}})X_j} > \alpha.$$

(A6') For those $j \notin \mathcal{A}$ whose corresponding $\mathcal{C}_{j|\mathcal{A}}$ satisfies $\mathcal{S} \setminus \mathcal{A} \nsubseteq \mathcal{C}_{j|\mathcal{A}}$, we have

$$n^{\kappa} \cdot \frac{\|(\mathbf{I}_n - \Pi_{\mathcal{A}\cup\mathcal{C}_{j|\mathcal{A}}})\mathbf{X}_{\mathcal{S}}\beta_{\mathcal{S}}\|_2^2}{\|(\mathbf{I}_n - \Pi_{\mathcal{A}})\mathbf{X}_{\mathcal{S}}\beta_{\mathcal{S}}\|_2^2} \to \infty.$$

As noted in Section 5.2.3, (A5') is related to the condition requiring strict positive definiteness of the population covariance matrix of $\mathbf{X}$. Also we can draw the connection between (A6') and the asymptotic identifiability condition (5.5) introduced in Chen & Chen (2008). By assuming an extended version of (5.5) as

$$\lim_{n\to\infty} \min_{\mathcal{D}\subset\mathcal{J},|\mathcal{D}|\leq C|\mathcal{S}|,\mathcal{D}\neq\mathcal{S}} n^{\kappa} \cdot \frac{\|(\mathbf{I}_n - \Pi_{\mathcal{D}})\mathbf{X}_{\mathcal{S}}\beta_{\mathcal{S}}\|_2^2}{\|\mathbf{X}_{\mathcal{S}}\beta_{\mathcal{S}}\|_2^2} \to \infty,$$

we can expect (A6') to hold as the algorithm progresses, provided $|\mathcal{A}| \leq C|\mathcal{S}|$ and $\xi \leq \delta$ (in conjunction with the fact that the "bias" term of $X_j^T\mathbf{y}$ in (5.2) consists of $\beta_k X_j^T X_k$ for $k \in \mathcal{S} \setminus \{j\}$, requiring $|\mathcal{C}_j| \leq |\mathcal{S}|$ is reasonable).

As for Conditions 5.1–5.3, they may also be extended to account for the updating of $\mathbf{X}$ during the application of the TCS algorithm. This would lead to deriving conditions under which the TCS algorithm is screening consistent, i.e.

$$\mathbb{P}(\mathcal{S} \subset \mathcal{A}) \to 1$$

after a certain number of iterations. In this case, m acts as a stopping rule, which cannot be too large for the updating step to be meaningful, while at the same time, it cannot be too small as a sufficient number of iterations need to be taken for every relevant variable to be included in $\mathcal{A}$.

While it is an interesting research topic to extend the theoretical results in Scenarios 1–3 to the screening consistency of the TCS algorithm, we do not pursue this direction of research here, since the main objective of this chapter is to develop a new measure of association between the variables and the response in a linear model. Instead, the following section presents two methods for the final model selection which are readily applicable to our framework.

5.3.2 Final model selection

5.3.2.1 Extended BIC

In Bogdan *et al.* (2004) and Chen & Chen (2008), an extended version of Bayesian information criterion (BIC) was proposed as

$$\mathrm{BIC}(\mathcal{A}) = \log\left\{\frac{1}{n}\|(\mathbf{I}_n - \Pi_{\mathcal{A}})\mathbf{y}\|_2^2\right\} + \frac{|\mathcal{A}|}{n}(\log n + 2\log p). \tag{5.9}$$

This new BIC takes into account high dimensionality of the data by adding a penalty term dependent on p. Chen & Chen (2008) noted that if $p \approx n^{1/2}$, the maximum (spurious) inflation in the log-likelihood was of order $0.5\log n$, and therefore in the case of $\mathcal{S} = \emptyset$, the probability of selecting a wrong, one-variable model would be positive with the original BIC.

They also showed the consistency of this new BIC under stronger conditions than those imposed in (A1), (A2) and (A4): the level of sparsity was $|\mathcal{S}| = O(1)$, the dimensionality was $p = O(n^C)$ for $C > 0$, and non-zero coefficients satisfied $\min_{j\in\mathcal{S}} |\beta_j| > C'$ for $C' > 0$. Then, under the asymptotical identifiability condition in (5.5), the modified BIC was shown to be consistent in the sense that

$$\mathbb{P}\left(\min_{|\mathcal{D}|\le m,\ \mathcal{D}\neq\mathcal{S}} \mathrm{BIC}(\mathcal{D}) > \mathrm{BIC}(\mathcal{S})\right) \to 1 \text{ for } m \ge |\mathcal{S}|,$$

i.e., the probability of selecting any model other than $\mathcal{S}$ converges to zero.

Since the TCS algorithm generates a solution path which consists of m sub-models $\mathcal{A}_{(1)} \subset \cdots \subset \mathcal{A}_{(m)} = \mathcal{A}$, a natural way of combining our algorithm with the BIC in (5.9) is to choose the final model as $\hat{\mathcal{S}} = \mathcal{A}_{(m^*)}$, where

$$m^* = \arg\min_{1\le i\le m} \mathrm{BIC}(\mathcal{A}_{(i)}).$$

At the price of replacing $\log n/n$ with $n^{-\kappa}$ in (5.5), the consistency of the new BIC can be shown with the level of sparsity growing with n as in (A1) and the dimensionality increasing exponentially with n as in (A2). The proof of this statement follows the exact line of proof in Chen & Chen (2008) and so we omit the details.

5.3.2.2 Multi-stage variable selection

Wasserman & Roeder (2009) proposed a multi-stage procedure for high-dimensional variable selection, which was shown to be able to control the type I error (false positive) at a desired level ν, when combined with the Lasso, the forward selection or the SIS-type marginal correlation screening, i.e.

$$\limsup_{n\to\infty} \mathbb{P}(\mathcal{S}^c \cap \hat{\mathcal{S}} \neq \emptyset) \leq \nu.$$

In this multi-stage procedure, the data is divided into two or three parts such that each part is used either at the model screening stage or at the "cleaning" stage as described below.

Once m variables have been identified in the active set $\mathcal{A}$, we can obtain an estimate of β by regressing $\mathbf{y}$ on $\mathbf{X}_{\mathcal{A}}$ as

$$\hat{\beta}_{\mathcal{A}} = (\mathbf{X}_{\mathcal{A}}^T \mathbf{X}_{\mathcal{A}})^{-1} \mathbf{X}_{\mathcal{A}} \mathbf{y},$$

and set $\hat{\beta}_j = 0$ if $j \notin \mathcal{A}$. Then the final $\hat{\mathcal{S}} \subset \mathcal{A}$ is selected at the cleaning stage by examining the t-statistic of each $\hat{\beta}_j,\ j \in \mathcal{A}$, i.e.

$$\hat{\mathcal{S}} = \{j \in \mathcal{A} :\ |T_j| > z_{\nu/2m}\}, \tag{5.10}$$

where T_j denotes the usual t-statistic of $\hat{\beta}_j$ and z_u is chosen such that $\mathbb{P}(Z > z_u) = u$ for $Z \sim \mathcal{N}(0,1)$. Although our framework does not have the data splitting step, this can easily be incorporated in applying the TCS algorithm.

5.3.3 Relation to existing literature

In Section 2.5.5, we briefly discuss a list of methods which account for correlations among the variables in measuring the association between each variable and the response. Now having been equipped with the complete picture of the TCS algorithm, we provide a detailed comparison between our methodology and the aforementioned methods.

Bühlmann *et al.* (2009) proposed the PC-simple algorithm, which iteratively removed the variables identified as having small association with the response

by partial correlation screening. Behind the adoption of partial correlation lies the concept of *partial faithfulness*, which implies that, at the population level, if the partial correlation between X_j and $\mathbf{y}$ conditional on $\mathbf{X}_{\mathcal{D}}$ was zero for some $\mathcal{D} \subset \mathcal{J} \setminus \{j\}$ (i.e. $\rho_n(j, \mathbf{y}|\mathcal{D}) = 0$), then $\rho_n(j, \mathbf{y}|\mathcal{J} \setminus \{j\}) = 0$. In their PC-simple algorithm, sample partial correlations $\hat{\rho}_n(j, \mathbf{y}|\mathcal{D})$ were used as the measure of association between X_j and $\mathbf{y}$, where $\mathcal{D}$ was any subset of the active set $\mathcal{A}$ (those variables still remaining in the current model excluding X_j) with its cardinality $|\mathcal{D}|$ equal to the number of iterations taken so far.

In details, the PC-simple algorithm starts with $\mathcal{A} = \mathcal{J}$ and iteratively repeats the following:

- calculate sample partial correlations $\hat{\rho}_n(j, \mathbf{y}|\mathcal{D})$ for all $j \in \mathcal{A}$ and for all $\mathcal{D}$ satisfying the above cardinality condition,
- apply the Fisher's Z-transform for testing the null hypotheses $H_0 : \rho_n(j, \mathbf{y}|\mathcal{D}) = 0$, i.e. see if

$$\frac{\sqrt{n - |\mathcal{D}| - 3}}{2} \cdot \left| \log \left(\frac{1 + \hat{\rho}_n(j, \mathbf{y}|\mathcal{D})}{1 - \hat{\rho}_n(j, \mathbf{y}|\mathcal{D})} \right) \right| > \Phi^{-1} \left(1 - \frac{\nu}{2} \right) \tag{5.11}$$

  where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function,
- remove those variables which do not satisfy (5.11) from $\mathcal{A}$, until $|\mathcal{A}|$ falls below the number of iterations taken so far.

Recalling the definition of the rescaling factor Λ_j, we can see the connection between $c_j^*(\Lambda_j)$ and $\hat{\rho}_n(j, \mathbf{y}|\mathcal{D})$, as both are (up to a multiplicative factor $\|\mathbf{y}\|_2$) the partial correlations between X_j and $\mathbf{y}$ conditional on a subset of variables. However, a significant difference comes from the fact that, the PC-simple algorithm takes every $\mathcal{D} \subset \mathcal{A} \setminus \{j\}$ with fixed $|\mathcal{D}|$ at each iteration, whereas our TCS algorithm adaptively selects $\mathcal{C}_j$ (or $\mathcal{C}_{j|\mathcal{A}}$ when $\mathcal{A} \neq \emptyset$) for each j. Also, while λ_j is also a valid rescaling factor in our tilted correlation methodology, partial correlations are by definition computed using Λ_j only.

As for the forward regression (Wang, 2009, FR) and the forward selection (FS), although the initial stage of the two techniques is simple marginal correlation screening, their progression has a new interpretation given a non-empty active

Table 5.1: Comparison of variable selection methods.

	TCS algorithm	PC-simple	FR	FS
Step 0	$\mathcal{A} = \emptyset$	$\mathcal{A} = \mathcal{J}$	$\mathcal{A} = \emptyset$	$\mathcal{A} = \emptyset$
action	one selected	multiple removed	one selected	one selected
conditioning set $\mathcal{D}$	$\mathcal{A} \cup \mathcal{C}_{j\|\mathcal{A}}$ $= \mathcal{A} \cup \{k \notin \mathcal{A}, k \neq j : \|\hat{\rho}_n(j,k\|\mathcal{A})\| > \pi_n\}$	remaining variables, $\|\mathcal{D}\|$ fixed	current model $\mathcal{A}$	current model $\mathcal{A}$
rescaling	λ_j or Λ_j	Λ_j	λ_j	none

set ($\mathcal{A} \neq \emptyset$). Both algorithms obtain the current residual $\mathbf{z}$ by projecting the response $\mathbf{y}$ onto the orthogonal complement of the current model space, i.e., $\mathbf{z} = (\mathbf{I}_n - \Pi_{\mathcal{A}})\mathbf{y}$. That is, they also measure the association between each $X_j,\ j \notin \mathcal{A}$ and $\mathbf{y}$ conditional on the current model $\mathbf{X}_{\mathcal{A}}$ and thus take into account the correlations between $X_j,\ j \notin \mathcal{A}$ and $X_k,\ k \in \mathcal{A}$.

The difference between the FR and the FS comes from the fact that the FR updates not only the current residual $\mathbf{z}$ but also the current design matrix as $\mathbf{Z} = (\mathbf{I}_n - \Pi_{\mathcal{A}})\mathbf{X}$ (as in Step 3 of the TCS algorithm). Therefore the FR eventually screens the rescaled version of $X_j^T(\mathbf{I}_n - \Pi_{\mathcal{A}})\mathbf{y}$ with the rescaling factor defined similarly to λ_j but replacing $\mathcal{C}_j$ with $\mathcal{A}$, i.e., $X_j^T(\mathbf{I}_n - \Pi_{\mathcal{A}})X_j = 1 - X_j^T\Pi_{\mathcal{A}}X_j$. On the other hand, there is no rescaling step in the FS and it screens the terms $X_j^T(\mathbf{I}_n - \Pi_{\mathcal{A}})\mathbf{y},\ j \notin \mathcal{A}$, themselves.

We note that unlike the FR and the FS, which always screen the marginal correlations $|Z_j^T\mathbf{z}|,\ j \notin \mathcal{A}$ (or $|X_j^T\mathbf{z}|$ in the FS) at each stage of their progression (after updating both or either of $\mathbf{z}$ and $\mathbf{Z}$), our method is able to adaptively "switch" between the use of marginal correlation and tilted correlation, depending on the sample correlation structure of the current design matrix $\mathbf{Z}$. Other crucial differences are as already mentioned above in the context of the PC-simple algorithm: the data-driven choice of the conditioning set $\mathcal{C}_j$ and the validity of the two rescaling methods in tilting.

In conclusion, the TCS algorithm, the PC-simple algorithm, the FR and the FS share the common ingredient of measuring the contribution of each variable X_j to $\mathbf{y}$ conditional on certain other variables; however, there are also important differences between them as reported in Table 5.1.

Finally, we note the relationship between the TCS algorithm and the covariance-regularised regression method proposed in Witten & Tibshirani (2009). When the active set $\mathcal{A}$ is not empty, the selection of $\mathcal{C}_j$ in Step 1 of the TCS algorithm is essentially the identification of non-negligible partial correlations among the variables conditional on the current model $\mathbf{X}_{\mathcal{A}}$, see Section 5.3.1.1. The *scout* procedure introduced in Witten & Tibshirani (2009) also has a step identifying the variables which have non-negligible partial correlation with each other (i.e., $\rho_n(j,k|\mathcal{J}\setminus\{j,k\}) \neq 0$). However, in the scout procedure, such identification is achieved by obtaining a regularised estimate of the inverse covariance matrix of $\mathbf{X}$ via penalised likelihood estimation, rather than hard-thresholding as in tilting. Also, thus-obtained estimate is applied to estimate β, again by maximising a penalised least squares problem. By contrast, we note that our tilted correlation method is an iterative technique which does not involve any optimisation problems.

5.3.4 Choice of threshold

In this section, we discuss the practical choice of the unknown threshold π_n from the sample correlation matrix $\mathbf{C}$. Bickel & Levina (2008) proposed a cross-validation method for this purpose, while El Karoui (2008) conjectured the usefulness of a procedure based on controlling the false discovery rate (FDR). Since our aim is not at the accurate estimation of the correlation matrix itself, we propose a threshold selection procedure which is a modified version of the approach taken in the latter paper. In the following, we assume that $\mathbf{X}$ is a realisation of a random matrix with each row generated as $\mathbf{x}_i \sim_{\text{i.i.d.}} (\mathbf{0}, \Sigma)$, where each diagonal element of Σ satisfies $\Sigma_{j,j} = 1$.

Our threshold selection method is a multiple hypothesis testing procedure and thus requires p-values of $d = p(p-1)/2$ hypotheses $H_0 : \ |\Sigma_{j,k}| = 0$ defined for all $j < k$. We propose to compute the p-values as follows. First, an $n \times p$-matrix with i.i.d. Gaussian entries is generated, and sample correlations $\{r_{l,m} : \ 1 \leq l < m \leq p\}$ among its columns are obtained as a "reference". Then, the p-value for

each null hypothesis $H_0 : \ |\Sigma_{j,k}| = 0$ is defined as

$$P_{j,k} = d^{-1} \cdot |\{r_{l,m}, \ 1 \leq l < m \leq p : \ |r_{l,m}| \geq |c_{j,k}|\}| \, .$$

The next step is to apply the testing method proposed in Benjamini & Hochberg (1995) to control the false discovery rate, i.e. the expected proportion of incorrectly rejected null hypotheses. Denoting $P_{(1)} \leq \ldots \leq P_{(d)}$ as the p-values in increasing order, we find the largest i for which

$$P_{(i)} \leq i/d \cdot \nu^*$$

and reject all $H_{(j)}, \ j = 1, \ldots, i$. Then $\hat{\pi}_{thr}$ is chosen as the absolute value of the correlation corresponding to $P_{(i)}$. If the hypotheses tests were independent, Benjamini & Hochberg (1995) proved that the FDR was controlled at level ν^*. Although it was not the case in our framework, our simulation study confirmed good practical performance of the above threshold selection procedure with the choice of $\nu^* = p^{-1/2}$ as suggested in El Karoui (2008).

We conclude this section by remarking on the choice of ν^*. Using (5.4), we can bound the probability that the maximum spurious sample correlation among p independent Gaussian variables $U_1, \ldots, U_p$ exceeds the threshold $\pi_n = C_1 n^{-\gamma}$ as

$$\mathbb{P}\left(\max_{j \neq k} \left| U_j^T U_k \right| > \pi_n \right) \leq \frac{Cnp(p-1)}{2} \cdot \exp\left(-\frac{C_1 (n-4) n^{-2\gamma}}{2} \right) . \qquad (5.12)$$

Interpreting ν^* as the permissible ratio of spuriously large sample correlations (among independent variables) which would not be thresholded by π_n, we can derive the particular choice of $\nu^* = p^{-1/2}$ from (5.12): under our assumption (A2), there exists $C_3 > 0$ for which $\log p < C_3 n^{1-\gamma/2}$, and therefore the right-hand side of (5.12) is bounded from above by $p^{2-C_1/(2C_3)}$ for large p, which can be made to be comparable to $p^{-1/2}$ when $C_1/C_3 = 5$. Whether spurious or not, the presence of large sample correlations can distort the association between the variables and the response in marginal correlation, as noted in Section 5.2.2. By allowing those spurious off-diagonal elements of $\mathbf{C} = \mathbf{X}^T\mathbf{X}$ to pass the thresholding step, tilting can successfully remove their influence in tilted correlation.

5.4 Simulation study

In this section, we study the performance of the TCS algorithm applied to simulated data and compare it to other related methods discussed in Section 5.3.3, which are the PC-simple algorithm, the FR and the FS, as well as some specific cases of penalised least squares estimation: the Lasso and the elastic net (Zou & Hastie, 2005) (see Section 2.5.1 for the details of the two methods).

The TCS algorithm was applied using both rescaling methods, with the maximum cardinality of the active set $\mathcal{A}$ (Step 4) set at $m = \lfloor n/2 \rfloor$, a value also used in the FR method. We used the R package `pcalg` to apply the PC-simple algorithm; the FS and the Lasso solution paths were generated by the R package `lars`, and those of the elastic net by the R package `elasticnet`, with a varying l_1 penalty parameter and a fixed l_2 penalty parameter.

5.4.1 Simulation models

In this section, we describe our simulation models. With the exception of (D)–(E), the procedure for generating the sparse coefficient vectors β is outlined below the following list.

(A) Factor model with 2 factors: Let ϕ_1 and ϕ_2 be two independent standard normal variables. Each variable X_j, $j = 1, \ldots, p$, is generated as $X_j = f_{j,1}\phi_1 + f_{j,2}\phi_2 + \eta_j$, where $f_{j,1}, f_{j,2}, \eta_j$ are also generated independently from a standard normal distribution. The model is taken from Meinshausen & Bühlmann (2010).

(B) Factor model with 10 factors: Identical to (A) but with 10 instead of 2 factors.

(C) Factor model with 20 factors: Identical to (A) but with 20 instead of 2 factors.

(D) Taken from Fan & Lv (2008) Section 4.2.2:

$$\mathbf{y} = \beta X_1 + \beta X_2 + \beta X_3 - 3\beta\sqrt{\varphi} X_4 + \epsilon,$$

where $\epsilon \sim \mathcal{N}_n(0, \mathbf{I}_n)$ and $(X_{i,1}, \ldots, X_{i,p})^T$ are generated from a multivariate normal distribution $\mathcal{N}_n(\mathbf{0}, \Sigma)$ independently for $i = 1, \ldots, n$. The population covariance matrix $\Sigma = (\Sigma_{j,k})_{j,k=1}^p$ satisfies $\Sigma_{j,j} = 1$ and $\Sigma_{j,k} = \varphi, j \neq k$, except $\Sigma_{4,k} = \Sigma_{j,4} = \sqrt{\varphi}$, such that X_4 is marginally uncorrelated with $\mathbf{y}$ at the population level. In the original model of Fan & Lv (2008), $\beta = 5$ and $\varphi = 0.5$ were used, but we chose $\beta = 2.5$ and $\varphi = 0.5, 0.95$ to investigate the performance of the variable selection methods in more challenging situations.

(E) Taken from Fan & Lv (2008) Section 4.2.3:

$$\mathbf{y} = \beta X_1 + \beta X_2 + \beta X_3 - 3\beta\sqrt{\varphi} X_4 + 0.25\beta X_5 + \epsilon,$$

with the population covariance matrix of $\mathbf{X}$ for this model is identical to (D) except $\Sigma_{5,k} = \Sigma_{j,5} = 0$, such that X_5 is uncorrelated with any X_j, $j \neq 5$, and relevant. However, it has only a very small contribution to $\mathbf{y}$.

(F) Leukemia data analysis: Golub *et al.* (1999) analysed the Leukaemia dataset from high-density Affymetrix oligonucloeotide arrays, which has 72 observations and 7129 genes (i.e. variables). The dataset is available on `http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi`. In Fan & Lv (2008), this dataset was used to investigate the performance of Sure Independence Screening in a feature selection problem. Here, instead of using the actual response from the dataset, we used the design matrix to create simulated models as follows.

Each column X_j of the design matrix was normalised to $\|X_j\|_2^2 = n$, and out of 7129 such columns, p were randomly selected to generate an $n \times p$-matrix $\mathbf{X}$. Then we generated a sparse p-vector β and the response $\mathbf{y}$ as in (5.1). In this manner, the knowledge of $\mathcal{S}$ could be used to assess the performance of the competing variable selection techniques. A similar approach was taken in Meinshausen & Bühlmann (2010) to generate simulation models from real datasets.

With the exception of (D)–(E), we generated the sparse coefficient vectors β by randomly sampling the indices of $\mathcal{S}$ from $1, \ldots, p$, with $|\mathcal{S}| = 10$. Then the non-

zero coefficient vector $\beta_{\mathcal{S}}$ was drawn from a zero-mean normal distribution such that

$$\mathbf{C}_{\mathcal{S},\mathcal{S}}\beta_{\mathcal{S}} \sim \mathcal{N}_{|\mathcal{S}|}(\mathbf{0}, n^{-1}\mathbf{I}_{|\mathcal{S}|}),$$

where $\mathbf{C}_{\mathcal{S},\mathcal{S}}$ denotes the sample correlation matrix of $\mathbf{X}_{\mathcal{S}}$. In this manner,

$$\arg\max_{j\in\mathcal{J}} |X_j^T(\mathbf{X}_{\mathcal{S}}\beta_{\mathbf{X}})|$$

may not always be attained within $\mathcal{S}$, which makes the correct identification of relevant variables more challenging. The noise level σ was chosen (except (D)–(E) where it was fixed at 1) to set $R^2 = \text{var}(\mathbf{x}_i^T\beta)/\text{var}(y_i)$ at 0.3, 0.5, or 0.9, adopting a similar approach to that made in Wang (2009). In the models (A)–(E), the number of observations was fixed at $n = 100$ while the dimensionality p varied from 500 to 2000 (except (D)–(E) where it was fixed at 1000), and finally, 100 replicates were generated for each set-up.

5.4.2 Simulation results

We evaluate the performance of the variable selection techniques using the receiver operating characteristic (ROC) curves, which plot the true positive rate (TPR) against the false positive rate (FPR). Bühlmann *et al.* (2009) also adopted the ROC curves, noting that they could assess the capacity for variable selection of different techniques independently from the issue of choosing good tuning parameters. A steep slope of the ROC curve indicates that relevant variables were selected before including many irrelevant variables. In Figures 5.2–5.15, ROC curves of different methods are compared, with vertical dotted lines indicating when the FPR reaches $2.5|\mathcal{S}|/p$.

Not surprisingly, variable selection methods turn out to work better for data with relatively lower dimensionality and higher R^2, in terms of the steepness of the ROC curves. Compared with other methods, the TCS algorithm and the FR achieve high TPR more quickly without including too many irrelevant variables for all models. While the PC-simple algorithm attains low FPR, its TPR is also low even when the significant level for the testing procedure is set to be high. The Lasso and the elastic net tend to result in high TPR at the cost of high FPR,

and their ROC curves are below those of the TCS algorithm or the FR for small FPR.

As can be seen from Figures 5.2-5.4, for the two factor example (A), the TCS algorithm and the FR work equally well with their ROC curves showing steep slopes, although the former achieves higher TPR for the case $p = 2000$ and $R^2 = 0.9$. The FS works almost as well as the above two methods for lower dimensional examples ($p = 500$), but with increasing dimensionality, it fails to achieve as high a TPR as that of the TCS algorithm or the FR, which is also the case for models (B) and (C).

As the number of factors used to generate $\mathbf{X}$ increases (see Figures 5.5-5.7 for the model with 10 factors and Figures 5.8-5.10 for the model with 20 factors), the TCS algorithm performs better than the FR, attaining higher TPR for a similar level of FPR. From substantial numerical experiments, we observed that the increase in the number of factors resulted in an increased chance of marginal correlation screening being misleading at the very first iteration, in the sense that

$$\arg\max_j |X_j^T \mathbf{y}| \notin \mathcal{S}.$$

In such set-ups, the adaptive choice of $\mathcal{C}_j$ used by the TCS algorithm turns out to be helpful in correctly identifying a relevant variable more often than marginal correlation screening. For model (C), although the TPR of the Lasso often reaches the highest level (especially when R^2 is low), the ROC curves of the Lasso remain below those of the TCS algorithm, the FR or the FS for small FPR. Between the two rescaling methods, rescaling 2 works better than rescaling 1 for models (A)–(C). Recalling that rescaling 2 is adopted by the FR (see Table 5.1), it is interesting to see that overall the TCS algorithm with rescaling 2 outperforms the FR for these models.

As for the models (D) and (E), the TCS algorithm and the FR outperform the rest when $\varphi = 0.5$, rapidly identifying all the relevant variables before the FPR reaches $2.5|\mathcal{S}|/p$ (see the left columns of Figures 5.11–5.12). However when the correlations among the variables increase with $\varphi = 0.95$ (see the right columns of Figures 5.11–5.12), the TCS algorithm with rescaling 1 is the only method that can identify all the relevant variables. Other methods, including the TCS

algorithm with rescaling 2 and the FR, often neglect to include X_4 due to its high correlations with the other variables, $\sqrt{\varphi}$ being almost 0.975.

For the examples generated from the Leukemia dataset ((F), Figures 5.13–5.15), the TCS algorithm with either of the rescaling methods always performs the best, with its ROC curves always dominating those of others. The FR performs the second best and the FS, the Lasso and the elastic are not able to identify as many relevant variables as the TCS algorithm or the FR even for high FPR.

5.5 Concluding remarks

In this chapter, a new way of measuring the association between the variables and the response is proposed for high-dimensional linear regression, which adaptively takes into account correlations among the variables. We conclude the discussion by listing some new contributions made in this chapter.

- Although tilting is not the only procedure which measures the association between a variable and the response conditional on other variables, its selection of the conditioning variables is a step further from simply using the current model itself or its submodels, as is done in existing iterative algorithms. The hard-thresholding step in the tilting procedure enables an adaptive choice of the conditioning subset $\mathcal{C}_j$ for each variable X_j, depending on the sample correlation structure of $\mathbf{X}$. Recalling the decomposition of the marginal correlation in (5.2), this adaptive choice can be seen as a vital step in isolating the contribution of each variable to the response. Also, in case of $\mathcal{C}_j = \emptyset$, tilted correlation is identical to marginal correlation, which is an example showing the adaptivity of our procedure.

- We have proposed two rescaling factors to obtain the tilted correlation c_j^*, which are also adopted by other methods (rescaling 1 by the forward regression and rescaling 2 by the PC-simple algorithm). However, tilting is the only method to meaningfully use both rescaling factors in the sense that, our theoretical results in Section 5.2.3 are valid for either of the two factors. It would be of interest to identify a way of combining the two rescaling

methods, possibly depending on the correlation structure of $\mathbf{X}$, which we leave as a topic for future research.

- The separation of relevant and irrelevant variables achieved by tilted correlation (as in our Theorems 5.1–5.3), cannot always be achieved by marginal correlation in the same scenarios, and similar results to these theorems have not been reported previously to the best of our knowledge. Not unexpectedly, conditions which are imposed on the linear model (5.1) for these separation properties to hold, take a different form from those imposed for consistency of other variable selection methods, such the sparse Riesz condition for the Lasso or the UUP for the Dantzig selector.

- The proposed TCS algorithm is designed to fully exploit the theoretical properties of the tilted correlation. Numerical experiments confirm its good performance, showing that it can achieve high true positive rate without including many irrelevant variables. The algorithm is easy to implement and does not require the use of advanced computational tools.

5.6 Proofs

5.6.1 Proof of Theorem 5.1

The proof of Theorem 5.1 is divided into Steps 1–3. Recalling the decomposition of $(X_j^*)^T\mathbf{y}$ in (5.3), we first control the inner product between X_j^* and ϵ uniformly over all j in Step 1. In Steps 2–3, we control the second summand $I = \sum_{k\in\mathcal{S}\setminus\mathcal{C}_j, k\neq j} \beta_k X_j^T(\mathbf{I}_n - \Pi_j)X_k$ for j falling into two different categories, and thus derive the results in Theorem 5.1.

Step 1 For $\epsilon \sim \mathcal{N}_n(\mathbf{0}, n^{-1}\sigma^2 \cdot \mathbf{I}_n)$, with probability converging to 1,

$$\max_{1\le j\le p} |\langle \epsilon, Z_j\rangle| \le \sigma\sqrt{2\log p/n}$$

for $Z_1, \ldots, Z_p \in \mathbb{R}^n$ having unit norm as $\|Z_j\|_2 = 1$. From (A2), we have $\sigma\sqrt{2\log p/n} \le Cn^{-\gamma}$ for some $C > 0$, and from (A5), $\|X_j^*\|_2 > \sqrt{\alpha} > 0$.

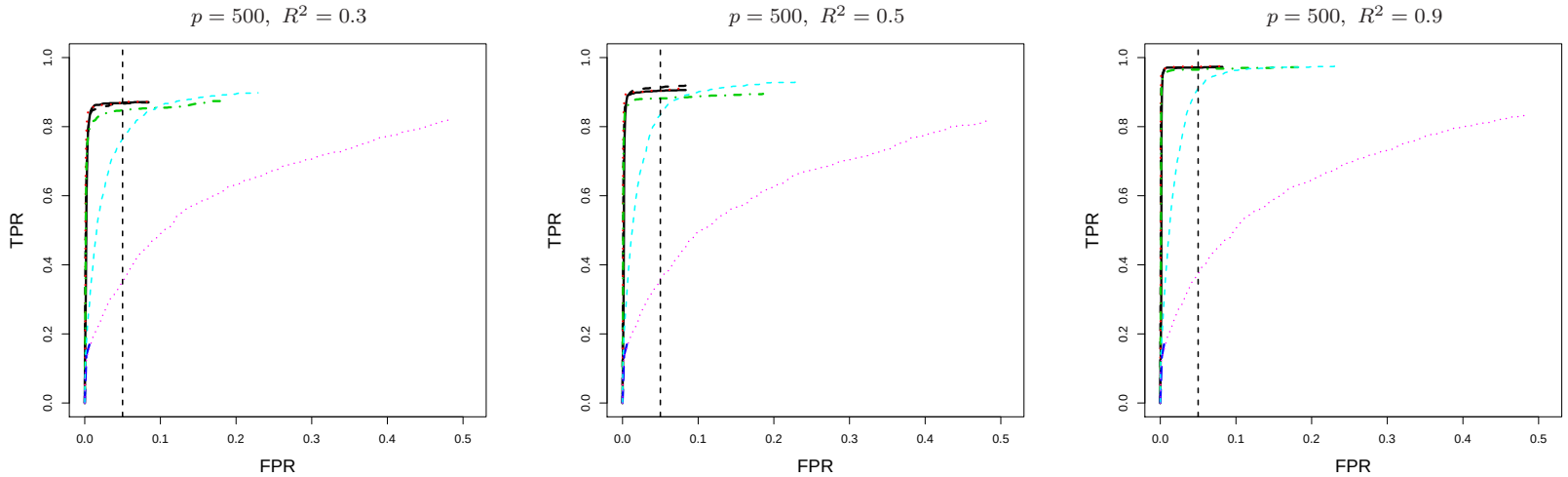


Figure 5.2: ROC curves for the simulation model (A) with $n = 100$ and $p = 500$: TCS algorithm with rescaling 1 (black solid), TCS algorithm with rescaling 2 (black dashed), FR (red dotted), FS (green dotdash), PC-simple algorithm (blue longdash), Lasso (light blue thin dashed), elastic net (magenta thin dotted); FPR= $2.5|S|/p$ (dotted vertical); left: $R^2 = 0.3$, middle: $R^2 = 0.5$, right: $R^2 = 0.9$.

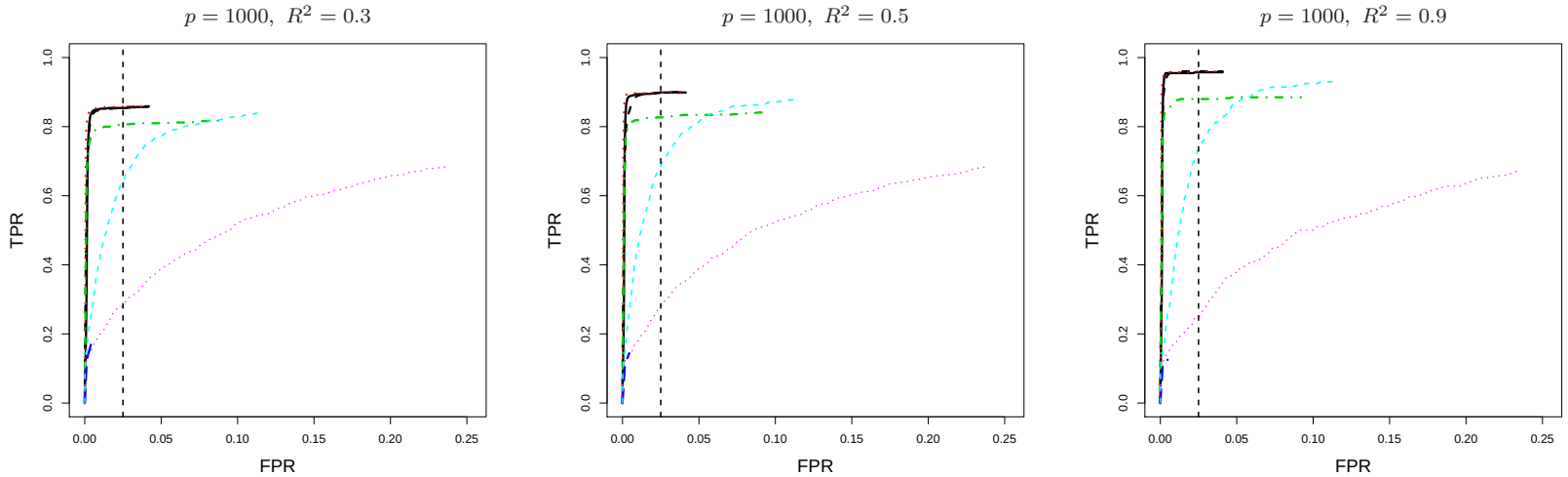


Figure 5.3: ROC curves for the simulation model (A) with $n = 500$ and $p = 1000$: TCS algorithm with rescaling 1 (black solid), TCS algorithm with rescaling 2 (black dashed), FR (red dotted), FS (green dotdash), PC-simple algorithm (blue longdash), Lasso (light blue thin dashed), elastic net (magenta thin dotted); FPR= $2.5|\mathcal{S}|/p$ (dotted vertical); left: $R^2 = 0.3$, middle: $R^2 = 0.5$, right: $R^2 = 0.9$.

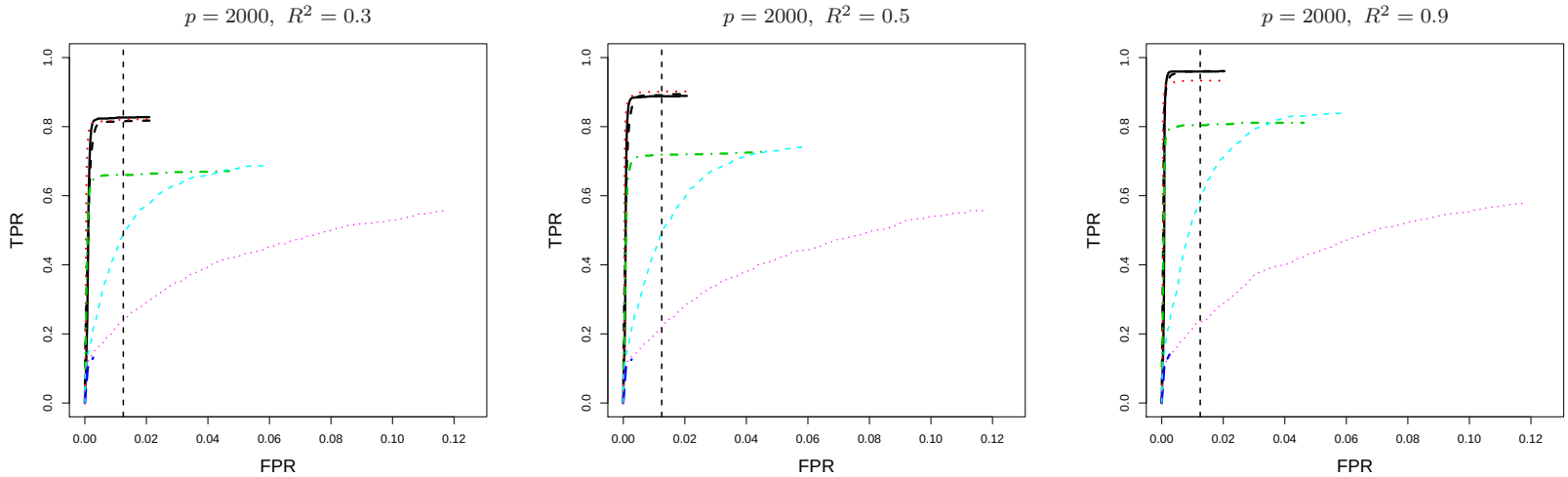


Figure 5.4: ROC curves for the simulation model (A) with $n = 500$ and $p = 2000$: TCS algorithm with rescaling 1 (black solid), TCS algorithm with rescaling 2 (black dashed), FR (red dotted), FS (green dotdash), PC-simple algorithm (blue longdash), Lasso (light blue thin dashed), elastic net (magenta thin dotted); FPR= $2.5|\mathcal{S}|/p$ (dotted vertical); left: $R^2 = 0.3$, middle: $R^2 = 0.5$, right: $R^2 = 0.9$.

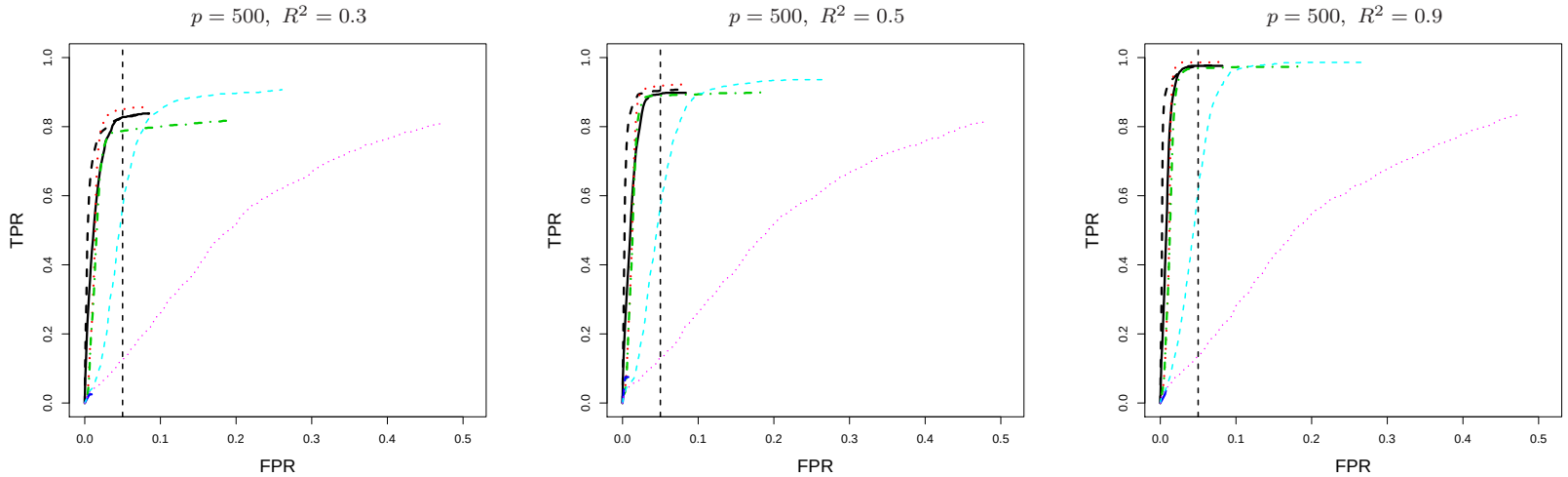


Figure 5.5: ROC curves for the simulation model (B) with $n = 100$ and $p = 500$: TCS algorithm with rescaling 1 (black solid), TCS algorithm with rescaling 2 (black dashed), FR (red dotted), FS (green dotdash), PC-simple algorithm (blue longdash), Lasso (light blue thin dashed), elastic net (magenta thin dotted); FPR$= 2.5|\mathcal{S}|/p$ (dotted vertical); left: $R^2 = 0.3$, middle: $R^2 = 0.5$, right: $R^2 = 0.9$.

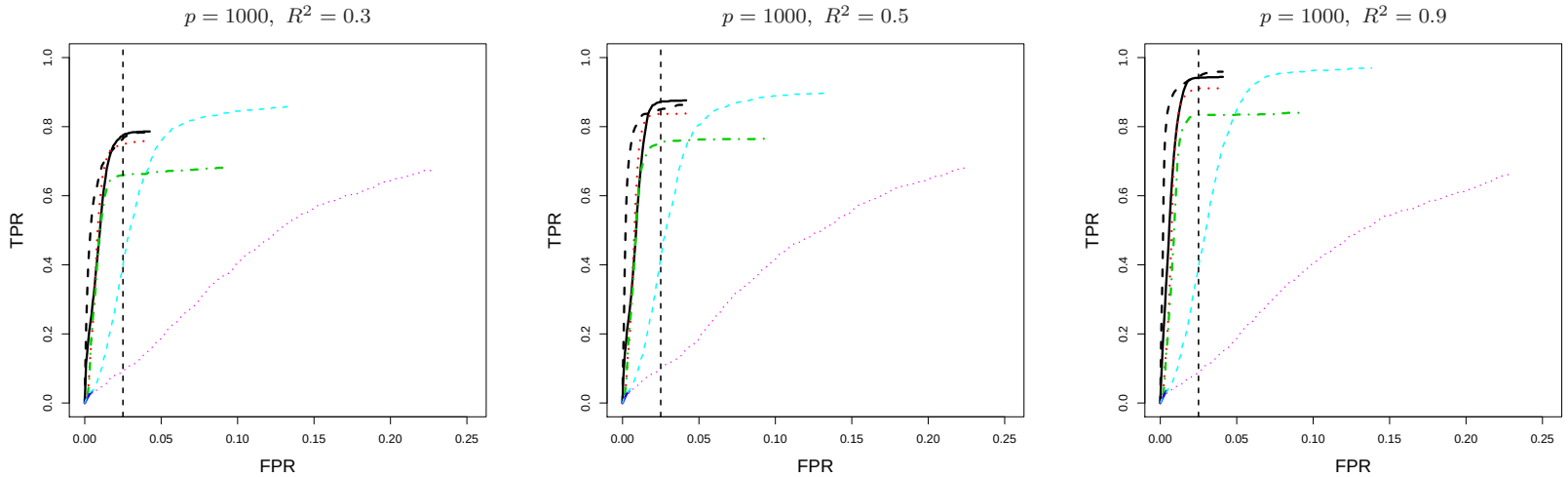


Figure 5.6: ROC curves for the simulation model (B) with $n = 500$ and $p = 1000$: TCS algorithm with rescaling 1 (black solid), TCS algorithm with rescaling 2 (black dashed), FR (red dotted), FS (green dotdash), PC-simple algorithm (blue longdash), Lasso (light blue thin dashed), elastic net (magenta thin dotted); FPR= $2.5|S|/p$ (dotted vertical); left: $R^2 = 0.3$, middle: $R^2 = 0.5$, right: $R^2 = 0.9$.

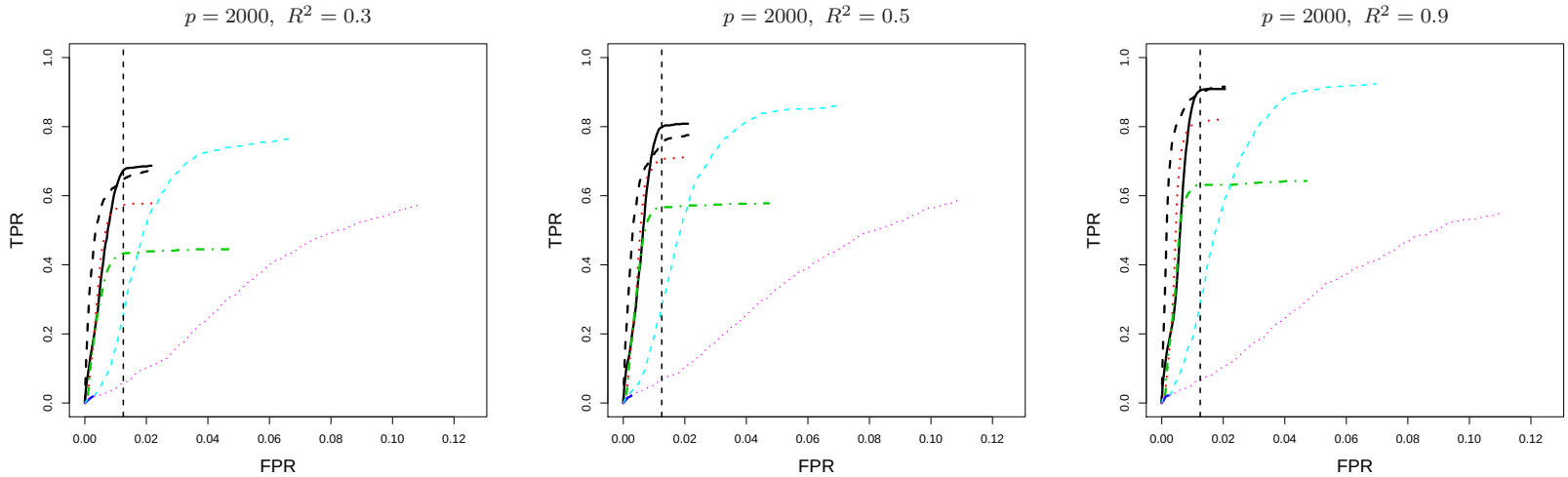


Figure 5.7: ROC curves for the simulation model (B) with $n = 500$ and $p = 2000$: TCS algorithm with rescaling 1 (black solid), TCS algorithm with rescaling 2 (black dashed), FR (red dotted), FS (green dotdash), PC-simple algorithm (blue longdash), Lasso (light blue thin dashed), elastic net (magenta thin dotted); FPR= $2.5|\mathcal{S}|/p$ (dotted vertical); left: $R^2 = 0.3$, middle: $R^2 = 0.5$, right: $R^2 = 0.9$.

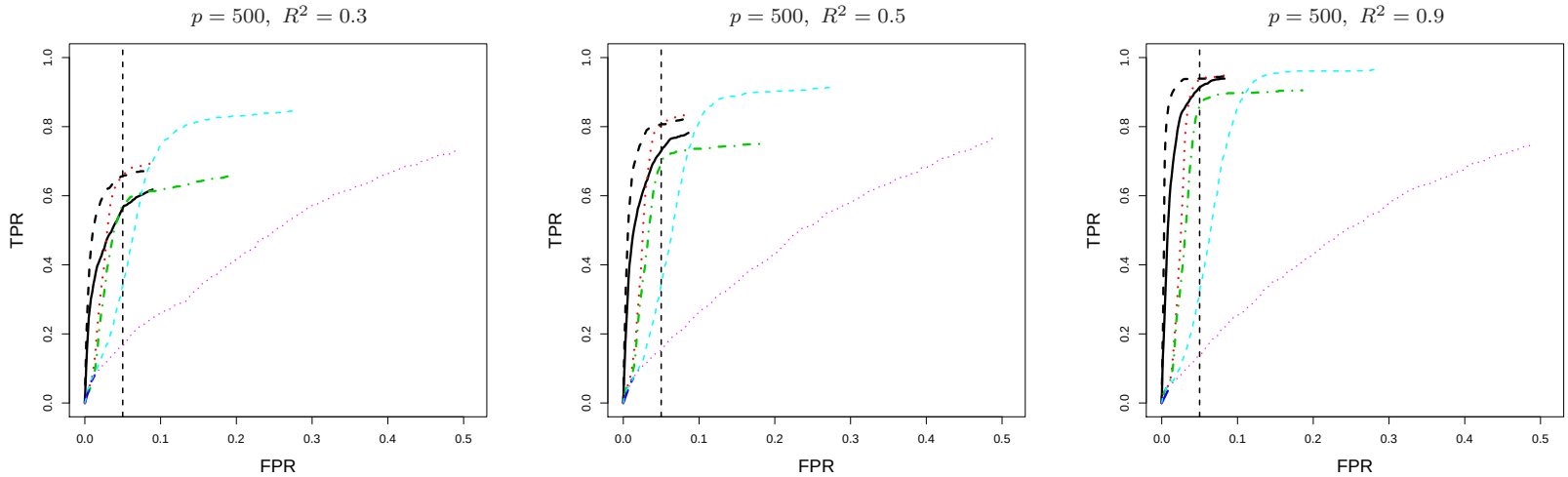


Figure 5.8: ROC curves for the simulation model (C) with $n = 100$ and $p = 500$: TCS algorithm with rescaling 1 (black solid), TCS algorithm with rescaling 2 (black dashed), FR (red dotted), FS (green dotdash), PC-simple algorithm (blue longdash), Lasso (light blue thin dashed), elastic net (magenta thin dotted); FPR= $2.5|\mathcal{S}|/p$ (dotted vertical); left: $R^2 = 0.3$, middle: $R^2 = 0.5$, right: $R^2 = 0.9$.

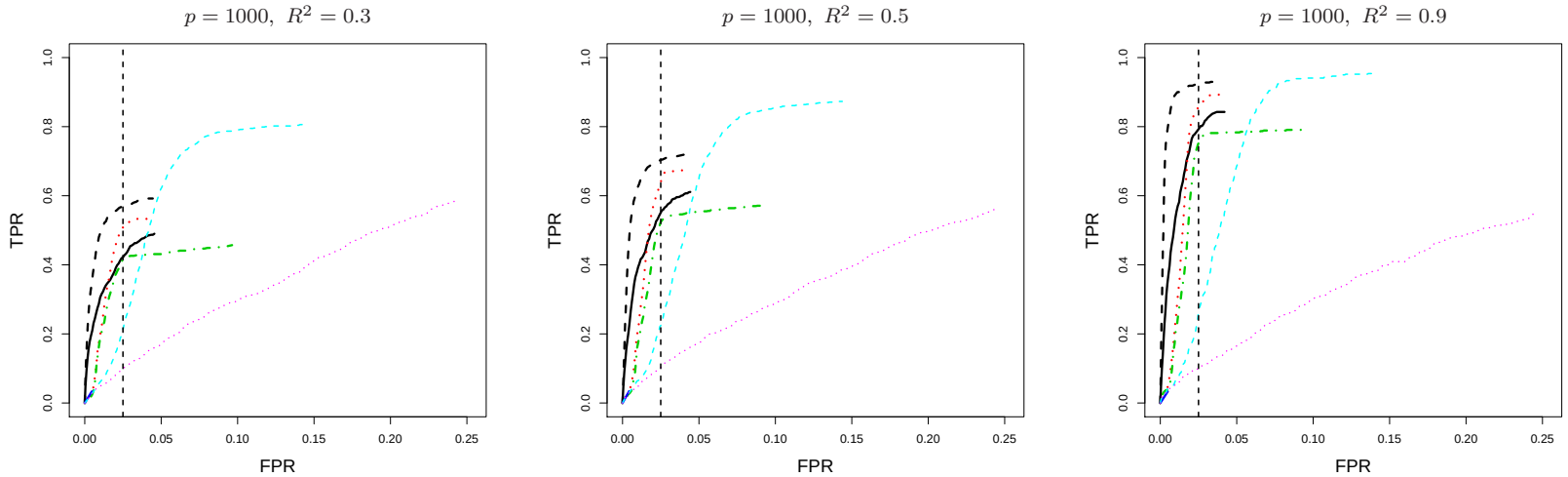


Figure 5.9: ROC curves for the simulation model (C) with $n = 500$ and $p = 1000$: TCS algorithm with rescaling 1 (black solid), TCS algorithm with rescaling 2 (black dashed), FR (red dotted), FS (green dotdash), PC-simple algorithm (blue longdash), Lasso (light blue thin dashed), elastic net (magenta thin dotted); FPR= $2.5|\mathcal{S}|/p$ (dotted vertical); left: $R^2 = 0.3$, middle: $R^2 = 0.5$, right: $R^2 = 0.9$.

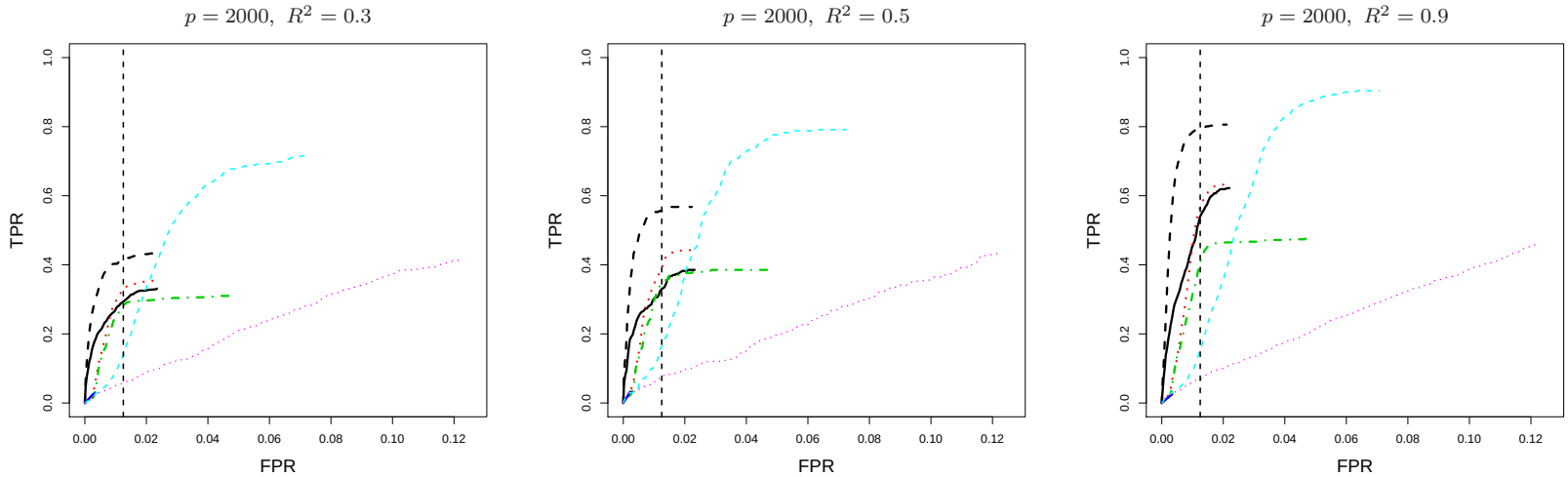


Figure 5.10: ROC curves for the simulation model (C) with $n = 500$ and $p = 2000$: TCS algorithm with rescaling 1 (black solid), TCS algorithm with rescaling 2 (black dashed), FR (red dotted), FS (green dotdash), PC-simple algorithm (blue longdash), Lasso (light blue thin dashed), elastic net (magenta thin dotted); FPR= $2.5|\mathcal{S}|/p$ (dotted vertical); left: $R^2 = 0.3$, middle: $R^2 = 0.5$, right: $R^2 = 0.9$.

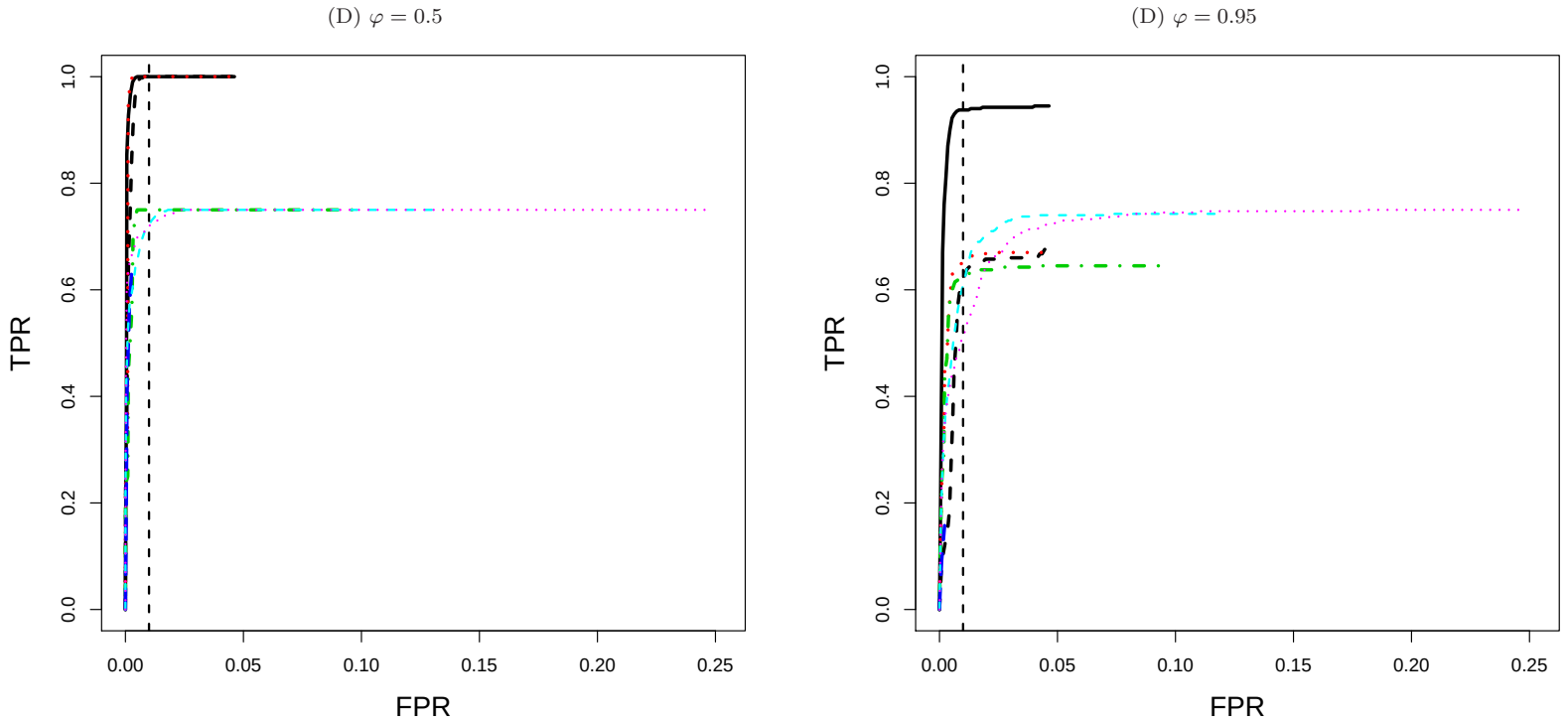


Figure 5.11: ROC curves for the simulation model (D): TCS algorithm with rescaling 1 (black solid), TCS algorithm with rescaling 2 (black dashed), FR (red dotted), FS (green dotdash), PC-simple algorithm (blue longdash), Lasso (light blue thin dashed), elastic net (magenta thin dotted); FPR= $2.5|\mathcal{S}|/p$ (dotted vertical); left: $\varphi = 0.5$, right: $\varphi = 0.95$.

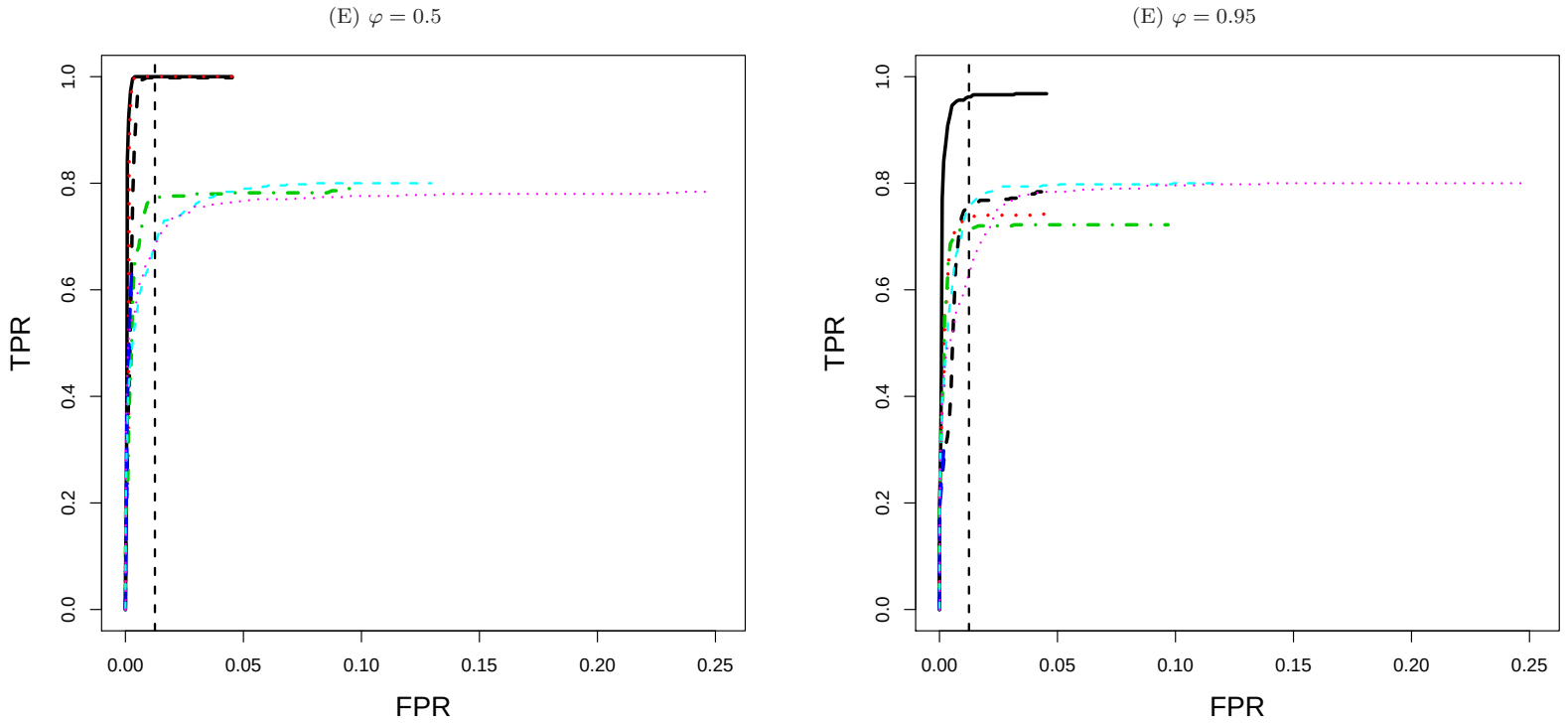


Figure 5.12: ROC curves for the simulation model (E): TCS algorithm with rescaling 1 (black solid), TCS algorithm with rescaling 2 (black dashed), FR (red dotted), FS (green dotdash), PC-simple algorithm (blue longdash), Lasso (light blue thin dashed), elastic net (magenta thin dotted); FPR= $2.5|\mathcal{S}|/p$ (dotted vertical); left: $\varphi = 0.5$, right: $\varphi = 0.95$.

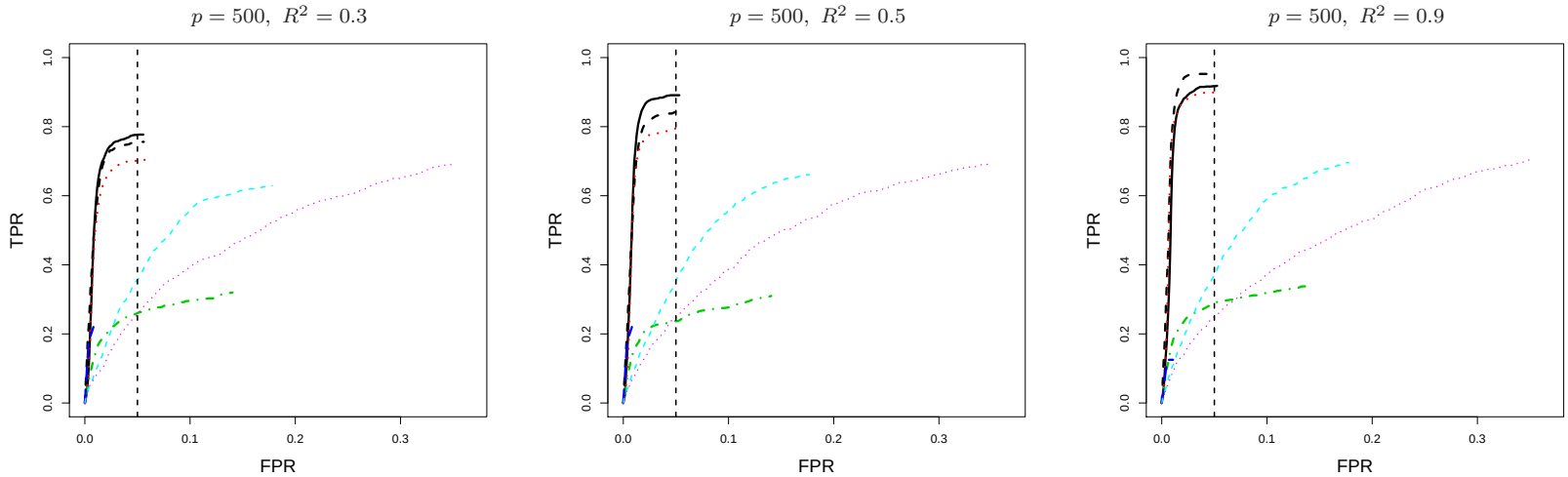


Figure 5.13: ROC curves for the simulation model (F) with $n = 72$ and $p = 500$: TCS algorithm with rescaling 1 (black solid), TCS algorithm with rescaling 2 (black dashed), FR (red dotted), FS (green dotdash), PC-simple algorithm (blue longdash), Lasso (light blue thin dashed), elastic net (magenta thin dotted); FPR= $2.5|\mathcal{S}|/p$ (dotted vertical); left: $R^2 = 0.3$, middle: $R^2 = 0.5$, right: $R^2 = 0.9$.

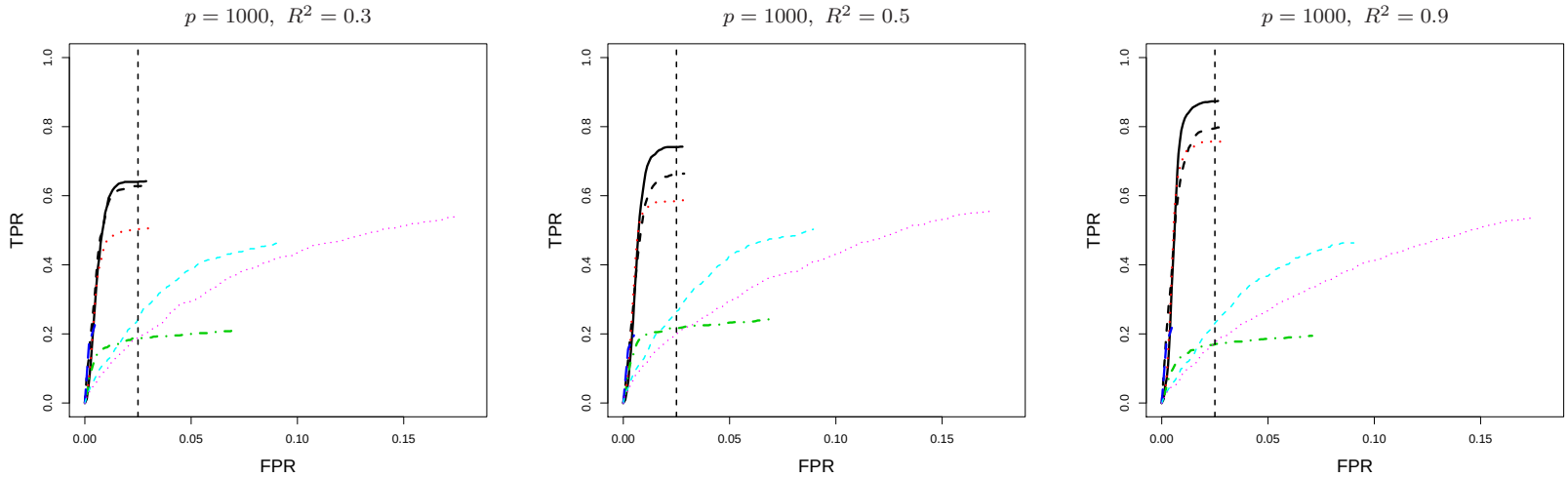


Figure 5.14: ROC curves for the simulation model (F) with $n = 72$ and $p = 1000$: TCS algorithm with rescaling 1 (black solid), TCS algorithm with rescaling 2 (black dashed), FR (red dotted), FS (green dotdash), PC-simple algorithm (blue longdash), Lasso (light blue thin dashed), elastic net (magenta thin dotted); FPR= $2.5|\mathcal{S}|/p$ (dotted vertical); left: $R^2 = 0.3$, middle: $R^2 = 0.5$, right: $R^2 = 0.9$.

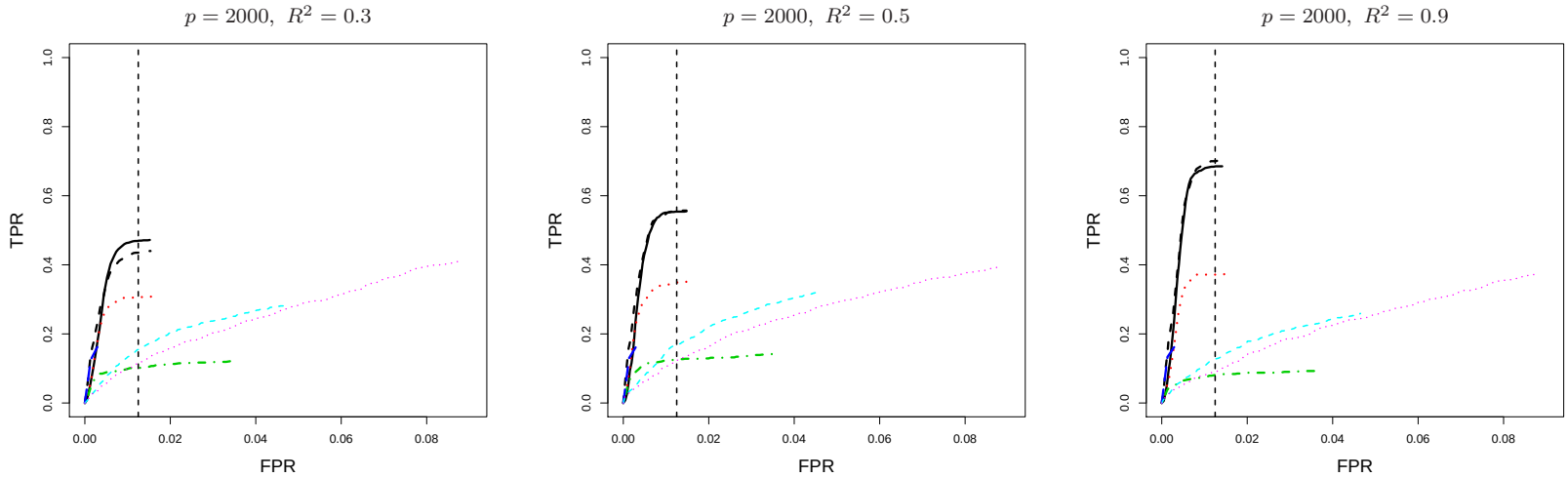


Figure 5.15: ROC curves for the simulation model (F) with $n = 72$ and $p = 2000$: TCS algorithm with rescaling 1 (black solid), TCS algorithm with rescaling 2 (black dashed), FR (red dotted), FS (green dotdash), PC-simple algorithm (blue longdash), Lasso (light blue thin dashed), elastic net (magenta thin dotted); FPR= $2.5|\mathcal{S}|/p$ (dotted vertical); left: $R^2 = 0.3$, middle: $R^2 = 0.5$, right: $R^2 = 0.9$.

Therefore by defining

$$\mathcal{E}_0 = \{\max_j |(X_j^*)^T \epsilon| < Cn^{-\gamma}\},$$

it follows that $\mathbb{P}(\mathcal{E}_0) \to 1$.

Step 2 In this step, we turn our attention to those j whose $\mathcal{C}_j$ satisfies $\mathcal{S} \setminus \{j\} \subseteq \mathcal{C}_j$ and thus the corresponding $I = 0$ and $(X_j^*)^T \mathbf{y} = \beta_j(1 - a_j) + (X_j^*)^T \epsilon$.

Rescaling 1. With the rescaling factor $\lambda_j = (1 - a_j)$ which is bounded away from 0 by (A5), it can be shown that if such j belongs to $\mathcal{S}$, its tilted correlation satisfies $c_j^*(\lambda_j)/\beta_j \to 1$ on $\mathcal{E}_0$, as $|\beta_j| \gg n^{-\mu}$. On the other hand, if $j \notin \mathcal{S}$, we have $\beta_j(1 - a_j) = 0$ which leads to $n^\mu \cdot c_j^*(\lambda_j) \leq n^\mu \cdot Cn^{-\gamma} \to 0$ on $\mathcal{E}_0$.

Rescaling 2. Note that j whose $\mathcal{C}_j$ include all the members of $\mathcal{S}$ cannot be a member of $\mathcal{S}$ itself, and in such case, $(\mathbf{I}_n - \Pi_j)\mathbf{y}$ is reduced to $(\mathbf{I}_n - \Pi_j)\epsilon$. Since (A3) assumes that each $\mathcal{C}_j$ has its cardinality bounded by Cn^ξ, it can be shown that

$$\mathbb{P}\left(\max_j \|\Pi_j \epsilon\|_2 \leq C'\sigma n^{-(\gamma - \xi/2)}\right) \to 1 \tag{5.13}$$

for some $C' > 0$, similarly as in Step 1. Also, Lemma 3 from Fan & Lv (2008) implies that

$$\mathbb{P}\left(\sigma^{-2} \cdot \|\epsilon\|_2^2 < 1 - \omega\right) \to 0 \tag{5.14}$$

for any $\omega \in (0, 1)$. Combining these observations with (A1) and (A4), we derive that

$$1 - a_{jy} = \frac{\|(\mathbf{I}_n - \Pi_j)\epsilon\|_2^2}{\|\mathbf{y}\|_2^2} \geq Cn^{-\delta}$$

with probability tending to 1, and eventually we have $\Lambda_j \geq C'n^{-\delta/2}$ from (A5). Therefore, if $\mathcal{S} \subseteq \mathcal{C}_j$ for some $j \notin \mathcal{S}$, its corresponding tilted correlation satisfies $n^\mu \cdot c_j^*(\Lambda_j) \leq n^\mu \cdot Cn^{-(\gamma - \delta/2)} \to 0$ on $\mathcal{E}_0$.

In the case of $\mathcal{S} \nsubseteq \mathcal{C}_j$, we can derive from (A6), (5.13) and (5.14) that for such j,

$$1 - a_{jy} = \frac{\|(\mathbf{I}_n - \Pi_j)\mathbf{y}\|_2^2}{\|\mathbf{y}\|_2^2} \gg n^{-\kappa},$$

which, combined with (A5), implies that $\Lambda_j \gg n^{-\kappa/2}$. Then the following holds for such j on $\mathcal{E}_0$: $n^\mu \cdot |c_j^*(\Lambda_j)| \geq n^\mu \cdot C|\beta_j| \to \infty$ if $j \in \mathcal{S}$, while $n^\mu \cdot c_j^*(\Lambda_j) \leq n^\mu \cdot Cn^{-(\gamma-\kappa/2)} \to 0$ if $j \notin \mathcal{S}$.

Step 3 We now consider those $j \in \mathcal{J}$ for which $\mathcal{S} \setminus \{j\} \nsubseteq \mathcal{C}_j$ and consequently the corresponding term $I \neq 0$ in general. From (A3) and Condition 5.1, we derive that for each j, there exists some $C > 0$ satisfying the following for all $k \in \mathcal{S} \setminus \mathcal{C}_j,\ k \neq j$,

$$|X_j^T(\mathbf{I}_n - \Pi_j)X_k| \leq |X_j^T X_k| + |(\Pi_j X_j)^T X_k| \leq Cn^{-\gamma}. \tag{5.15}$$

Then from (A1) and (A4), we can bound I as $|I| \leq C'n^{-(\gamma-\delta)}$. Also when $\mathcal{S} \setminus \{j\} \nsubseteq \mathcal{C}_j$, (A5)–(A6) imply that $\Lambda_j \gg n^{-\kappa/2}$.

In summary, we can show that the following claims hold on $\mathcal{E}_0$, similarly as in Step 2: if $j \notin \mathcal{S}$, with either of the rescaling factors, $n^\mu \cdot c_j^*(\lambda_j) \leq n^\mu \cdot Cn^{-(\gamma-\delta-\kappa/2)} \to 0$, whereas if $j \in \mathcal{S}$, its coefficient satisfies $|\beta_j| \gg n^{-\mu}$ and therefore $n^\mu \cdot |c_j^*| \geq n^\mu \cdot C|\beta_j| \to \infty$ with $c_j^*(\lambda_j)/\beta_j \to 1$ for $j \in \mathcal{S}$. □

5.6.1.1 An example satisfying Condition 5.1

In this section, we verify the claim made in Section 5.2.3.1, which states that Condition 5.1 holds with probability tending to 1 when each column X_j is generated independently as a random vector on a n-dimensional unit sphere. We first introduce a result from modern convex geometry reported in Lecture 2 of Ball (1997), which essentially implies that, as the dimension n grows, it is not likely for any two vectors on a n-dimensional unit sphere to be within a close distance to each other.

Lemma 5.1. *Let S^{n-1} denote the surface of the Euclidean ball $B_2^n = \{\mathbf{x} \in \mathbb{R}^n : \sum_{i=1}^n x_i^2 \leq 1\}$ and $\mathbf{u} \in \mathbb{R}^n$ be a vector on S^{n-1} such that $\|\mathbf{u}\|_2 = 1$. Then the*

proportion of spherical cone defined as $\{\mathbf{v} \in S^{n-1} : \ |\mathbf{u}^T\mathbf{v}| \geq \omega\}$ *for any* $\mathbf{u}$ *is bounded from above by* $\exp(-n\omega^2/2)$.

We first note that any $X_k,\ k \neq j$ can be decomposed as the summation of its projection onto X_j and the remainder, i.e., $X_k = c_{j,k}X_j + (\mathbf{I}_n - X_jX_j^T)X_k$. Then

$$(\Pi_j X_j)^T X_k = c_{j,k}(\Pi_j X_j)^T X_j + \left\{(\mathbf{I}_n - X_jX_j^T)\Pi_j X_j\right\}^T X_k,$$

and for $k \in \mathcal{S} \setminus \mathcal{C}_j,\ k \neq j$, the first summand is bounded from above by $a_j \cdot \pi_n \leq C_1 n^{-\gamma}$. As for the second summand, note that

$$\|(\mathbf{I}_n - X_jX_j^T)\Pi_j X_j\|_2^2 = (\Pi_j X_j)^T(\mathbf{I}_n - X_jX_j^T)\Pi_j X_j = a_j(1-a_j),$$

and thus $\mathbf{w} = \{a_j(1-a_j)\}^{-1/2} \cdot (\mathbf{I}_n - X_jX_j^T)\Pi_j X_j$ satisfies $\mathbf{w} \in S^{n-1}$. Then the probability of $|\mathbf{w}^T X_k| > Cn^{-\gamma}$ for any $k \in \mathcal{S} \setminus \mathcal{C}_j,\ k \neq j$ is bounded from above by the proportion of the following spherical cone

$$\left\{X_k \in S^{n-1} : \ |\mathbf{w}^T X_k| > Cn^{-\gamma}\right\}$$

in the unit sphere S^{n-1}. Applying Lemma 5.1, we can show that such proportion is bounded by $\exp\left(-C^2 n^{1-2\gamma}/2\right)$ for each j and k. Therefore, we can find some $C > 0$ satisfying

$$\mathbb{P}\left(\max_{j\in\mathcal{J};\ k\in\mathcal{S}\setminus\mathcal{C}_j,\ k\neq j} |(\Pi_j X_j)^T X_k| > Cn^{-\gamma}\right) \geq 1 - p|\mathcal{S}|\exp\left(-C'n^{1-2\gamma}/2\right),$$

where the right-hand side converges to 1 from assumptions (A1)–(A2).

5.6.2 Proof of Theorem 5.2

For those $j \in \mathcal{K} = \mathcal{S} \cup \{\cup_{j\in\mathcal{S}}\mathcal{C}_j\}$, Condition 5.3 implies that $\mathcal{C}_k \cap \mathcal{C}_j = \emptyset$ if $k \in \mathcal{S} \setminus \mathcal{C}_j$. Then from (A3), we have $\|\Pi_j X_k\|_2 \leq Cn^{-(\gamma-\xi/2)}$ and therefore

$$\left|X_j^T(\mathbf{I}_n - \Pi_j)X_k\right| = \left|X_j^T X_k - (\Pi_j X_j)^T \Pi_j X_k\right| \leq Cn^{-\gamma} + C'n^{-(\gamma-\xi/2)},$$

which leads to

$$\left| \sum_{k \in \mathcal{S} \setminus \mathcal{C}_j, k \neq j} \beta_k X_j^T (\mathbf{I}_n - \Pi_j) X_k \right| = O\left(n^{-(\gamma - \delta - \xi/2)}\right) \tag{5.16}$$

for all $j \in \mathcal{K}$. Using Step 1 of Section 5.6.1, we derive that

$$\mathcal{E}_{01} = \left\{ \max_{j \in \mathcal{K}} \left| \sum_{k \in \mathcal{S} \setminus \mathcal{C}_j, k \neq j} \beta_k X_j^T (\mathbf{I}_n - \Pi_j) X_k + X_j^T (\mathbf{I}_n - \Pi_j) \epsilon \right| \leq C n^{-(\gamma - \delta - \xi/2)} \right\}$$

satisfies $\mathbb{P}(\mathcal{E}_{01}) = \mathbb{P}(\mathcal{E}_0) \to 1$. Since $\mu + \kappa/2 < \gamma - \delta - \xi/2$, we have $n^\mu \cdot c_j^* \to 0$ for $j \notin \mathcal{S}$ on $\mathcal{E}_{01}$, whereas $n^\mu \cdot |c_j^*| \to \infty$ and $c_j^*(\lambda_j)/\beta_j \to 1$ for those $j \in \mathcal{S}$. Therefore the dominance of tilted correlations for $j \in \mathcal{S}$ over those for $j \in \mathcal{K} \setminus \mathcal{S}$ follows. □

5.6.3 Proof of Theorem 5.3

Compared to Condition 5.2, Condition 5.3 does not require any restriction on $\mathcal{C}_j \cap \mathcal{C}_k$ when both X_j and X_k are relevant, although it has an additional assumption (C2). Since $n^\mu \cdot |\beta_j|(1 - a_j) \to \infty$ for $j \in \mathcal{S}$ from (A4)–(A5), (C2) implies that for any $j \in \mathcal{S}$, non-zero coefficients β_k, $k \in \mathcal{S} \setminus \mathcal{C}_j$ do not cancel out all the summands in the following to 0,

$$X_j^T (\mathbf{I}_n - \Pi_j) \mathbf{X}_{\mathcal{S}} \beta_{\mathcal{S}} = \beta_j (1 - a_j) + \sum_{k \in \mathcal{S} \setminus \mathcal{C}_j, k \neq j} \beta_k X_j^T (\mathbf{I}_n - \Pi_j) X_k.$$

If (5.16) in Section 5.6.2 holds, (C2) follows and therefore it can be seen that Condition 5.2 is stronger than Condition 5.3.

On the event $\mathcal{E}_0$ (Step 1 of Section 5.6.1), $|X_j^T (\mathbf{I}_n - \Pi_j) \mathbf{y}| \gg n^{-\mu}$ for $j \in \mathcal{S}$ under (C2) and therefore the tilted correlations of relevant variables satisfy $|c_j^*| \gg n^{-\mu}$ with either of the rescaling factors. On the other hand, for $j \in \mathcal{K} \setminus \mathcal{S}$, we can use the arguments in Section 5.6.2 to show that $n^\mu \cdot c_j^* \to 0$. □

Chapter 6

Conclusions

In this thesis, we have discussed estimation methods which approach some challenging statistical problems under the assumption that the data can be well-described by a sparse model. Below we summarise the main contributions made in Chapters 3–5 and remark on some potential directions for future research.

Chapter 3 was devoted to developing a segmentation procedure for a class of piecewise stationary time series with breakpoints in the second-order structure. Assuming that the breakpoints were sufficiently scattered over time, we developed our methodology in the framework of the locally stationary wavelet model, under which the entire second-order structure of a time series was encoded in its wavelet-based local periodogram sequences. As the initial step of breakpoint detection, a binary segmentation procedure was proposed to segment these wavelet periodogram sequences at each scale separately, which had the following features:

- it permitted autocorrelation within a target sequence, and thus could be applied to wavelet periodogram sequences, and
- its test criterion depended on the data size only and was easy to compute.

The next step was the application of within-scale and across-scales post-processing procedures, and this combined methodology was shown to be consistent in terms of the total number and locations of detected breakpoints. Our method showed good performance in identifying the breakpoints from simulated data, and when it was employed to segment some historical financial time series, the outcome had an interesting interpretation in the context of recent financial crisis.

One immediate way of extending this work may be the investigation of its possibility as an on-line breakpoint detection procedure. Although the binary segmentation procedure itself was initially developed as a tool for *a posteriori* segmentation, it would be interesting to study whether the test statistic and criterion tailored for the consistency of our retrospective segmentation methodology are applicable to sequential data analysis.

In Chapter 4, we compared the two techniques for estimating an unknown signal using piecewise constant functions, the Unbalanced Haar technique and taut strings. The comparison study was conducted by providing a unified, multiscale framework, of which both methods were instances. We also studied the test statistics of the two techniques in the context of breakpoint detection, where it was shown that the UH technique was more alert than the TS technique at both detecting the presence of a breakpoint and estimating its location. While the comparison study between the UH and the TS methods was in itself interesting, we derived some lessons based on the links between the two, which could benefit either of the methods or both. Those lessons concerned the issues which could be encountered within our proposed unified framework, such as

- the choice of a test criterion (or a stopping rule),
- controlling the adaptivity of the estimated function, and
- extensions of these estimation techniques to non-Gaussian error distributions.

We concluded our discussion by observing some connections between the statistical problems addressed in other chapters of this thesis, and the piecewise constant estimators discussed in this chapter under the overall theme of sparsity.

Here we further note that our proposed multiscale approach consists of a "spectrum" of piecewise constant estimators in the sense that, there are many ways of obtaining piecewise constant estimators within this framework, depending on the choice of adjusting factor (which is in turn related to the test statistic) and the re-arrangement of the string. Therefore, as a future research topic, it would be of interest to study those piecewise constant estimators, which belong to the same multiscale framework yet have not been discussed in this thesis. Studying such

estimators is not only useful on its own, but also may lead to the development of a procedure which selects the "best" piecewise constant estimator within the multiscale framework in a data-driven manner.

Finally in Chapter 5, the variable selection problem in linear regression was considered, where the number of variables, or the dimensionality of the data, was possibly much larger than the number of observations. Under the assumption that only a small number of variables actually contributed to the response, we proposed a new way of measuring the association between each variable and the response, which adaptively took into account high correlations among the variables. The proposed tilting procedure had a hard-thresholding step applied to the sample correlation the design matrix, which enabled an adaptive "switch" between the use of marginal correlation and tilted correlation for each variable. We showed that the tilted correlations of the relevant variables dominated those of the irrelevant variables (which were highly correlated with at least one of the relevant variables) under certain conditions, and thus could be used as a tool for variable selection. We then constructed an iterative variable screening algorithm to exploit these theoretical properties of tilted correlation, and investigated its practical performance in a comparative simulation study.

We note that, to the best of our knowledge, similar results to the separation of relevant and irrelevant variables achieved by tilted correlation (Theorems 5.1–5.3), have not been reported previously in the literature. As for the tilted correlation screening (TCS) algorithm, however, it remains as a challenging task to develop a stopping rule which identifies when the TCS algorithm has included every relevant variable in the active set without including too many irrelevant variables (screening consistency). Furthermore, correlation being arguably the most widely used statistical measure of association, we would expect our tilted correlation (which can be viewed as an "adaptive" extension of standard correlation) to be more widely applicable in various statistical contexts beyond the linear regression model.

References

Abramovich, F. & Benjamini, Y. (1995). Thresholding of wavelet coefficients as multiple hypotheses testing procedure. *Lecture notes in statistics: wavelets and statistics*, **103**, 5–14. 29

Abramovich, F. & Benjamini, Y. (1996). Adaptive thresholding of wavelet coefficients. *Computational Statistics and Data Analysis*, **22**, 351–361. 29

Abramovich, F., Sapatinas, T. & Silverman, B. (1998). Wavelet thresholding via a Bayesian approach. *Journal of the Royal Statistical Society, Series B*, **60**, 725–749. 29

Adak, S. (1998). Time-dependent spectral analysis of nonstationary time series. *Journal of the American Statistical Association*, **93**, 1488–1501. 15, 22

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, 267–281, Akadémiai Kiadó. 32

Andreou, E. & Ghysels, E. (2002). Detecting multiple breaks in financial market volatility dynamics. *Journal of Applied Economics*, **17**, 579–600. 22

Antoniadis, A. (1997). Wavelets in statistics: a review. *Statistical Methods and Applications*, **6**, 97–130. 12

Bai, J. & Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica*, **66**, 47–78. 22

Ball, K. (1997). An elementary introduction to modern convex geometry. *Flavors of Geometry*, **31**, 1–58. 164

Barlow, R., Bartholomew, D., Bremner, J. & Brunk, H. (1972). *Statistical inference under order restrictions*. Wiley. 30, 91, 95

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, **57**, 289–300. 142

Bickel, P.J. & Levina, E. (2008). Covariance regularization by thresholding. *Annals of Statistics*, **36**, 2577–2604. 121, 127, 141

Bogdan, M., Ghosh, J. & Doerge, R. (2004). Modifying the Schwarz Bayesian information criterion to locate multiple interacting quantitative trait loci. *Genetics*, **167**, 989–999. 137

Bosq, D. (1998). *Nonparametric statistics for stochastic process: estimation and prediction*. Springer. 84

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., Steinberg, D. & Colla, P. (1983). *CART: Classification and Regression Trees*. Wadsworth: Belmont, CA. 30

Brodsky, B.E. & Darkhovsky, B.S. (1993). *Nonparametric methods in change-point problems*. Springer. 109, 110, 111

Bühlmann, P., Kalisch, M. & Maathuis, M. (2009). Variable selection for high-dimensional models: partially faithful distributions and the PC-simple algorithm. *Biometrika*, **97**, 1–19. 42, 128, 131, 138, 145

Candès, E. & Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics*, **6**, 2313–2351. 31, 33, 39, 40, 125

Chen, J. & Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, **95**, 759–771. 125, 128, 136, 137

Chen, J. & Gupta, A.K. (1997). Testing and locating variance change-points with application to stock prices. *Journal of the American Statistical Association*, **92**, 739–747. 24, 76

Chernoff, H. & Zacks, S. (1964). Estimating the current mean of a normal distribution which is subject to changes in time. *Annals of Mathematical Statistics*, **35**, 999–1028. 22

Cleveland, W. & Devlin, S. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association*, **83**, 596–610. 27

Comte, F. & Rozenholc, Y. (2004). A new algorithm for fixed design regression and denoising. *Annals of the Institute of Statistical Mathematics*, **56**, 449–473. 30

Cramér, H. (1942). On Harmonic Analysis in Certain Function Spaces. *Arkiv för Matematik, Astronomioch Fysik*, **28B**, 1–7. 14

Dahlhaus, R. (1996). On the Kullback-Leibler information divergence of locally stationary processes. *Stochastic processes and their applications*, **62**, 139–168. 15

Dahlhaus, R. (1997). Fitting time series models to nonstationary processes. *Annals of Statistics*, **25**, 1–37. 14, 15

Dahlhaus, R. & Subba Rao, S. (2006). Statistical inference for time-varying ARCH processes. *Annals of Statistics*, **34**, 1075–1114. 13

Davies, P.L. & Kovac, A. (2001). Local extremes, runs, strings and multiresolution. *Annals of Statistics*, **29**, 1–65. iv, 2, 30, 91, 95, 96, 112, 113, 115

Davis, R.A., Lee, T.C.M. & Rodriguez-Yam, G.A. (2006). Structural break estimation for non-stationary time series. *Journal of the American Statistical Association*, **101**, 223–239. 23, 63, 64, 65

Davis, R.A., Lee, T.C.M. & Rodriguez-Yam, G.A. (2008). Break detection for a class of nonlinear time series models. *Journal of Time Series Analysis*, **29**, 834–867. 23

DeVore, R. (1998). Nonlinear approximation. *Acta Numerica*, **7**, 51–150. 27, 29

DeVore, R. & Popov, V. (1988). Interpolation of Besov spaces. *Transactions of the American Mathematical Society*, **305**, 397–414. 29

Donoho, D. (1997). CART and best-ortho-basis: a connection. *Annals of Statistics*, **25**, 1870–1911. 95

Donoho, D. (2000). High-dimensional data analysis: the curses and blessings of dimensionality. *AMS Math Challenges Lecture*. 31

Donoho, D.L. (2006). For most large underdetermined systems of linear equations, the minimal l1-norm near-solution approximates the sparsest near-solution. *Communications on Pure and Applied Mathematics*, **59**, 907–934. 35

Donoho, D.L. & Johnstone, I.M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, **90**, 1200–1224. 28, 29

Donoho, D.L. & Johnstone, J.M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**, 425–455. 12, 28

Dümbgen, L. & Kovac, A. (2009). Extensions of smoothing via taut strings. *Electronic Journal of Statistics*, **3**, 41–75. 113

Eckley, I., Nason, G. & Treloar, R. (2010). Locally stationary wavelet fields with application to the modelling and analysis of image texture. *Journal of the Royal Statistical Society, Series C*, **59**. 20

Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, **32**, 407–499. 37

El Karoui, N. (2008). Operator norm consistent estimation of large dimensional sparse covariance matrices. *Annals of Statistics*, **36**, 2717–2756. 121, 127, 141, 142

Engel, J. (1997). The multiresolution histogram. *Metrika*, **46**, 41–57. 30

Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *Annals of Statistics*, **21**, 196–216. 27

Fan, J. & Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *Journal of the Royal Statistical Society, Series B*, **57**, 371–394. 27

Fan, J. & Gijbels, I. (1996). *Local polynomial modelling and its applications*. Chapman & Hall/CRC. 27

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**, 1348–1360. 31, 37, 39, 125

Fan, J. & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society, Series B*, **70**, 849–911. 32, 41, 118, 143, 144, 163

Fan, J. & Lv, J. (2010). A selective overview of variable selection in high dimensional feature space (invited review article). *Statistica Sinica*, **20**, 101–148. 3, 31, 39, 117

Friedman, J., Hastie, T., Höfling, H. & Tibshirani, R. (2007). Pathwise coordinate optimization. *Annals of Applied Statistics*, **1**, 302–332. 38, 39

Fryzlewicz, P. (2007). Unbalanced Haar technique for nonparametric function estimation. *Journal of the American Statistical Association*, **102**, 1318–1327. iv, 2, 13, 30, 91, 93, 95, 102, 110, 113

Fryzlewicz, P. & Nason, G. (2006). Haar-Fisz estimation of evolutionary wavelet spectra. *Journal of the Royal Statistical Society, Series B*, **68**, 611–634. 12, 20, 46, 47, 49, 56

Fryzlewicz, P., Sapatinas, T. & Subba Rao, S. (2006). A Haar-Fisz technique for locally stationary volatility estimation. *Biometrika*, **93**, 687–704. 56

Fryzlewicz, P., Sapatinas, T. & Subba Rao, S. (2008). Normalised least-squares estimation in time-varying ARCH models. *Annals of Statistics*, **36**, 742–786. 114

Gabbanini, F., Vannucci, M., Bartoli, G. & Moro, A. (2004). Wavelet packet methods for the analysis of variance of time series with application to crack widths on the Brunelleschi Dome. *Journal of Computational and Graphical Statistics*, **13**, 639–658. 24

Girardi, M. & Sweldens, W. (1997). A new class of unbalanced Haar wavelets that form an unconditional basis for Lp on general measure spaces. *Journal of Fourier Analysis and Applications*, **3**, 457–474. 93

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M. *et al.* (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537. 144

Harchaoui, Z. & Cappe, O. (2007). Retrospective multiple change-point estimation with kernels. In *IEEE/SP 14th Workshop on Statistical Signal Processing, 2007*, 768–772. 22

Hawkins, D., PEIHUA, Q. & CHANG, W. (2003). The changepoint model for statistical process control. *Journal of Quality Technology*, **35**, 355–366. 21

Hawkins, D.M. (1977). Testing a sequence of observations for a shift in location. *Journal of the American Statistical Association*, **72**, 180–186. 22

Hawkins, D.M. (2001). Fitting multiple change-point models to data. *Computational Statistics and Data Analysis*, **37**, 323–341. 22

Hoerl, A. & Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **42**, 80–86. 33

Hsu, D.A. (1977). Tests for variance shifts at an unknown time point. *Journal of Applied Statistics*, **26**, 179–184. 22

Hsu, D.A. (1979). Detecting shifts of parameters in gamma sequences with applications to stock price and air traffic flow analysis. *Journal of the American Statistical Association*, **74**, 31–40. 76

Incl´an, C. & Tiao, G.C. (1994). Use of cumulative sums of squares for retrospective detection of changes of variance. *Journal of the American Statistical Association*, **89**, 913–923. 24, 55

James, G., Radchenko, P. & Lv, J. (2009). Dasso: connections between the Dantzig selector and Lasso. *Journal of the Royal Statistical Society, Series B*, **71**, 127–142. 41

Johnson, N. & Kotz, S. (1970). *Distributions in statistics: continuous univariate distributions, vol. 1*. Houghton Mifflin Company. 83

Johnstone, I. & Silverman, B. (1997). Wavelet threshold estimators for data with correlated noise. *Journal of the Royal Statistical Society, Series B*, **59**, 319–351. 29

Johnstone, I. & Silverman, B. (2004). Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Annals of Statistics*, **32**, 1594–1649. 29

Kalisch, M. & Bühlmann, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research*, **8**, 613–636. 127

Kokoszka, P. & Leipus, R. (2000). Change-point estimation in ARCH models. *Bernoulli*, **6**, 513–539. 13

Kolaczyk, E. & Nowak, R. (2005). Multiscale generalised linear models for nonparametric function estimation. *Biometrika*, **92**, 119–133. 30

Kouamo, O., Moulines, E. & Roueff, F. (2010). Testing for homogeneity of variance in the wavelet domain. *Dependence, with applications in statistics*

and econometrics (eds P. Doukhan, G. Lang, D. Surgailis and G. Teyssière), 175–205. 55

Lai, T. (2001). Sequential analysis: some classical problems and new challenges. *Statistica Sinica*, **11**, 303–350. 21

Lavielle, M. & Moulines, E. (2000). Least-squares estimation of an unknown number of shifts in a time series. *Journal of Time Series Analysis*, **21**, 33–59. 22

Lavielle, M. & Teyssière, G. (2005). Adaptive detection of multiple change-points in asset price volatility. *Long memory in economics (eds G. Teyssière and A. Kirman)*, 129–156. 22, 63, 74, 75, 76

Mallat, S. (1989). Multiresolution approximations and wavelet orthonormal bases of l2(r). *Transactions of the American Mathematical Society*, **315**, 69–87. 6, 9

Mallat, S. *et al.* (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11**, 674–693. 6, 9

Mallows, C. (1973). Some comments on Cp. *Technometrics*, **42**, 87–94. 32

Mammen, E. & van de Geer, S. (1997). Penalized quasi-likelihood estimation in partial linear models. *Annals of Statistics*, **25**, 1014–1035. 95

Mei, Y. (2006). Sequential change-point detection when unknown parameters are present in the pre-change distribution. *Annals of Statistics*, **34**, 92–122. 21

Meinshausen, N. & Bühlmann, P. (2008). High dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, **34**, 1436–1462. 35

Meinshausen, N. & Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society, Series B*, **72**, 417–473. 36, 143, 144

Mikosch, T. & Stărică, C. (1999). Changes of structure in financial time series, long range dependence and the GARCH model. *Technical Report, University of Groningen*, **99**, 5–06. 13

Nadaraya, È. (1964). On estimating regression. *Teoriya Veroyatnostei i ee Primeneniya*, **9**, 157–159. 26

Nason, G. (1995). Choice of the threshold parameter in wavelet function estimation. *Lecture notes in statistics: wavelets and statistics*, **103**, 261–280. 29

Nason, G. (1996). Wavelet shrinkage using cross-validation. *Journal of the Royal Statistical Society, Series B*, **58**, 463–479. 29

Nason, G. & Silverman, B. (1995). The stationary wavelet transform and some statistical applications. *Lecture notes in statistics: wavelets and statistics*, **103**, 281–281. 11

Nason, G.P., von Sachs, R. & Kroisandt, G. (2000). Wavelet processes and adaptive estimation of the evolutionary wavelet spectrum. *Journal of the Royal Statistical Society, Series B*, **62**, 271–292. iv, 12, 13, 16, 17, 18, 19, 20, 46, 48

Ogden, T. & Parzen, E. (1996a). Change-point approach to data analytic wavelet thresholding. *Statistics and Computing*, **6**, 93–99. 29

Ogden, T. & Parzen, E. (1996b). Data dependent wavelet thresholding in nonparametric regression with change-point applications. *Computational Statistics and Data Analysis*, **22**, 53–70. 29

Ombao, H.C., Raz, J.A., von Sachs, R. & Malow, B.A. (2001). Automatic statistical analysis of bivariate nonstationary time series. *Journal of the American Statistical Association*, **96**, 543–560. 22, 63

Ombao, H.C., Raz, J.A., von Sachs, R. & Guo, W. (2002). The SLEX model of a non-stationary random process. *Annals of the Institute of Statistical Mathematics*, **54**, 171–200. 15

Phillips, P. & Perron, P. (1988). Testing for a unit root in time series regression. *Biometrika*, **75**, 335–346. 66, 76

Polzehl, J. & Spokoiny, V. (2000). Adaptive weights smoothing with applications to image restoration. *Journal of the Royal Statistical Society, Series B*, **62**, 335–354. 30

Polzehl, J. & Spokoiny, V. (2006). Varying coefficient GARCH versus local constant volatility modeling. Comparison of the predictive power. *SFB 649 Discussion Papers*, **33**. 13

Priestley, M.B. (1965). Evolutionary spectra and non-stationary processes. *Journal of the Royal Statistical Society, Series B*, **27**, 204–23. 14

Sanderson, J., Fryzlewicz, P. & Jones, M. (2010). Measuring dependence between non-stationary time series using the locally stationary wavelet model. *Biometrika*, **97**, 435–446. 20

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464. 32

Sen, A. & Srivastava, M.S. (1975). On tests for detecting change in mean. *Annals of Statistics*, **3**, 98–108. 22

Silverman, B. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society, Series B*, **47**, 1–52. 26

Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *Annals of Statistics*, **9**, 1135–1151. 29

Stărică, C. & Granger, C. (2005). Nonstationarities in stock returns. *Review of Economics and Statistics*, **87**, 503–522. 13, 76

Tartakovsky, A., Rozovskii, B., Blazek, R. & Kim, H. (2006). Detection of intrusions in information systems by sequential change-point methods. *Statistical Methodology*, **3**, 252–293. 21

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, **58**, 267–288. 31, 34

Van Bellegem, S. & von Sachs, R. (2004). On adaptive estimation for locally stationary wavelet processes and its applications. *International Journal of Wavelets, Multiresolution and Information Processing*, **2**, 545–565. 12, 20

Van Bellegem, S. & von Sachs, R. (2008). Locally adaptive estimation of evolutionary wavelet spectra. *Annals of Statistics*, **36**, 1879–1924. 20

Venkatraman, E.S. (1993). Consistency results in multiple change-point problems. *PhD thesis, Stanford University*. 23, 24, 52, 81, 85, 87, 106

Vidakovic, B. (1999). *Statistical modeling by wavelets*. Wiley. 4, 5, 7, 12, 28, 51

Vostrikova, L.J. (1981). Detecting 'disorder' in multidimensional random processes. *Soviet Doklady Mathematics*, **24**, 55–59. 23

Wang, H. (2009). Forward Regression for Ultra-High Dimensional Variable Screening. *Journal of the American Statistical Association*, **104**, 1512–1524. 43, 139, 145

Wang, Y. (1996). Function estimation via wavelet shrinkage for long-memory data. *Annals of Statistics*, **24**, 466–484. 29

Wasserman, L. & Roeder, K. (2009). High-dimensional variable selection. *Annals of Statistics*, **37**, 2178–2201. 138

Watson, G. (1964). Smooth regression analysis. *Sankhyā: Indian Journal of Statistics, Series A*, **26**, 359–372. 26

Wegman, E. & Wright, I. (1983). Splines in statistics. *Journal of the American Statistical Association*, **78**, 351–365. 26

Weisberg, S. (1980). *Applied linear regression*. Wiley-Blackwell. 43

Whitcher, B., Guttorp, P. & Percival, D.B. (2000). Multiscale detection and location of multiple variance changes in the presence of long memory. *Journal of Statistical Computation and Simulation*, **68**, 65–87. 24

Whitcher, B., Byers, S.D., Guttorp, P. & Percival, D.B. (2002). Testing for homogeneity of variance in time series: long memory, wavelets and the Nile River. *Water Resources Research*, **38**, 10–1029. 24

Wickerhauser, M. (1994). *Adapted wavelet analysis from theory to software*. AK Peters Ltd. 22

Witten, D.M. & Tibshirani, R. (2009). Covariance-regularized regression and classification for high-dimensional problems. *Journal of Royal Statistical Society, Series B*, **71**, 615–636. 43, 141

Worsley, K.J. (1986). Confidence regions and tests for a change-point in a sequence of exponential family random variables. *Biometrika*, **73**, 91–104. 22

Zhang, C.H. & Huang, J. (2008). The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Annals of Statistics*, **36**, 1567–1594. 35, 125

Zhao, P. & Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research*, **7**, 2541–2563. 35, 125

Zou, H. (2006). The adaptive Lasso and its oracle property. *Journal of the American Statistical Association*, **101**, 1418–1429. 36, 128

Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, **67**, 301–320. 37, 143

Zou, H. & Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models (with discussion). *Annals of Statistics*, **36**, 1509–1566. 39