Empirical Bayes Block Shrinkage for Wavelet Regression

by Xue Wang, BSc, MSc

Thesis submitted to the University of Nottingham

for the degree of Doctor of Philosophy, September 2005

Contents

List of Figures

List of Tables

Abstract

There has been great interest in recent years in the development of wavelet methods for estimating an unknown function observed in the presence of noise, following the pioneering work of Donoho and Johnstone (1994, 1995) and Donoho *et al.* (1995). In this thesis, a novel empirical Bayes block (EBB) shrinkage procedure is proposed and the performance of this approach with both independent identically distributed (IID) noise and correlated noise is thoroughly explored.

The first part of this thesis develops a Bayesian methodology involving the non-central χ^2 distribution to simultaneously shrink wavelet coefficients in a block, based on the block sum of squares. A useful (and to the best of our knowledge, new) identity satisfied by the non-central χ^2 density is exploited. This identity leads to tractable posterior calculations for suitable families of prior distributions. Also, the families of prior distribution we work with are sufficiently flexible to represent various forms of prior knowledge. Furthermore, an efficient method for finding the hyperparameters is implemented and simulations show that this method has a high degree of computational advantage.

The second part relaxes the assumption of IID noise considered in the first part of this thesis. A semi-parametric model including a parametric component and a nonparametric component is presented to deal with correlated noise situations. In the parametric component, attention is paid to the covariance structure of the noise. Two distinct parametric methods (maximum likelihood estimation and time series model identification techniques) for estimating the parameters in the covariance matrix are investigated. Both methods have been successfully implemented and are believed to be new additions to smoothing methods.

Acknowledgements

I would like to thank my supervisor, Professor Andy Wood, for his support and enthusiasm over the last three and half years. Many of the ideas in the thesis originally came from my discussions with him. I would also like to thank Dr Cliff Litton and Dr Owen Lynn for being my internal assessors and providing useful suggestions during my annual reviews, and Professor Ian Dryden for providing some planar curve data and useful explanations and discussions of these data. Thanks are also due to Dr Huiling Le for her encouragement and advice, to Nikolaos, Alfred, Ali and Mousa for useful discussions of computational questions and to Eleanor, Orawan and Li Yao for their friendship.

I am grateful to School of Mathematical Sciences, the University of Nottingham, and Universities UK for providing scholarships.

Finally, I would like to thank Tao, my husband, as it would not have been written without his support and patience, and to our parents for their full support and help.

Declaration

This thesis is the result of my own work carried out in accordance with the regulations of the University of Nottingham. References to other researchers have been specifically indicated in the text.

None of the material contained in this work has been submitted in support of an application for any other degree or qualification in this or any other higher education establishment.

Xue Wang

Chapter 1

Introduction

The past two decades have witnessed the development of wavelet analysis, a powerful tool which emerged from mathematics and related fields and was adopted by a great variety of researches. The term "wavelet" originates from the work of Morlet *et al.* (1982), in the context of the analysis of seismic reflection data. Since then wavelets have led to exciting applications in many areas, such as signal processing, for example Mallat (1989), and image processing, for example Shapior (1993). The impetus for the application of wavelets in statistics stems from the early 1990s through the work of Donoho and Johnstone, with contributions also from Kerkyacharian and Picard.

Wavelets provide a framework which possesses some key advantages. Firstly, wavelets can be viewed as orthonormal basis functions that are localised in both time and frequency, with time-widths adapted to their frequency. This means that they are able to model a signal with high frequency components, such as discontinuities, in contrast to more traditional statistical methods for estimating an unknown function. A second advantage comes from the fast orthogonal discrete wavelet transform, which makes the application of wavelets available. A third advantage is that wavelets often provide sparse and, therefore, economical representations of functions. These key properties make wavelets an excellent tool for statistical denoising.

With the introduction of nonlinear wavelet methods in statistics by Donoho and Johnstone (1994, 1995, 1998) and Donoho *et al.* (1995), the theory and application

of wavelet approaches to nonparametric regression has developed rapidly. Many papers have been written on this topic. The key points in the mathematical theory of wavelets and their applications in statistics will be reviewed in Chapter 2.

The application of wavelets in the context of nonparametric regression has been already discussed in the literature. Some authors have drawn the conclusion (Hall *et al.*, 1997, 1998, 1999, Cai, 1999, Cai and Silverman, 2001) that thresholding based on blocking the wavelet coefficients has the potential to be more accurate than thresolding obtained term by term since the former method combines the information in neighbouring coefficients. However, block thresholding in the Bayesian framework has not received much consideration. This is one of the principal topics that will be considered in this thesis (Chapter 3 and Chapter 4). An empirical Bayes block (EBB) shrinkage method is proposed. This itself brings forward several questions of its own. The two main questions are how to choose the families of prior distributions and how to shrink or threshold the noisy coefficients. In the existing literature, the parameters in the prior were either chosen by a combination of prior information and data-based estimation or an empirical Bayes (or marginal maximum likelihood) approach which is a completely data-based method. Generally, these estimation methods take longer computation time and make the Bayesian methods less competitive. An effective way to estimate the parameters is still needed; specific proposals are made in the thesis. A special issue arising in the block shrinkage and thresholding approach is how to choose the block size, which will be thoroughly investigated.

In many applications, the possibility of correlated noise arises. This raises new issues which do not arise with models assuming IID noise. The literature on the correlated data situation has mainly concentrated on assuming covariance structures of wavelet coefficients. A natural question to ask is whether these covariance structures of wavelet coefficients can capture features of correlated noise in the time domain. This is the other main area that will be investigated in this thesis (Chapter 5). Under the assumption that the covariance structure of correlated noise is known,

there is a clear need for an estimation procedure to find parameters in the assumed covariance structure. Furthermore, in order to obtain an effective estimation procedure, it is also necessary to explore appropriate ways to make use of the raw data values or the wavelet coefficients of these values.

The entire study was carried out in the **Matlab** programming environment. The algorithms which use the discrete wavelet transform were performed using the **WaveLab802** software that is freely available from

http://www-stat.stanford.edu/software/software.html.

Chapter 2

Review of Wavelets and Nonparametric Regression

2.1 Introduction

In this chapter, an overview of background material relevant to subsequent chapters is given. In the first two sections, some of the necessary mathematical background for wavelets will be summarised. In § 2.2, we will review the definitions of wavelets and multiresolution analysis (MRA). MRA provides a sequence of simple functions to approximate a general unknown function. In § 2.3, some wavelet transforms, the key mathematical tools in the use of wavelets, will be investigated. Some commonly used methods for univariate function estimation in nonparametric regression will be presented in § 2.4. The final section will consider the use of wavelets in statistics, with a focus on application of wavelets to nonparametric function estimation. A brief review of Bayesian analysis is provided in the appendix following this chapter.

2.2 Wavelets and Multiresolution Analysis

Some terminology and background for wavelets, which is required to understand the wavelet methodology and applications considered later, is provided here. More

detailed mathematical descriptions and wavelets can be found in Meyer (1992) and Daubechies (1992).

2.2.1 Wavelets

Wavelets comprise the family of translations and dilations of a single function, denoted ψ. The function ψ, which is called a mother wavelet, was defined by Meyer (1992, page 66) as follows:

Definition 2.1 *Let m be a non-negative integer. A function $\psi(x)$ of a real variable is called a mother wavelet of class m if the following properties hold:*

1. *if $m = 0$, $\psi(x) \in L^\infty(\mathbf{R})$; if $m \geq 1$, $\psi(x)$ and all its derivatives up to order m belong to $L^\infty(\mathbf{R})$;*

2. *$\psi(x)$ and all its derivatives up to order m decrease rapidly as $x \to \pm\infty$;*

3. *$\int_{-\infty}^{+\infty} x^k \psi(x) dx = 0$ for $0 \leq k \leq m$;*

4. *the collection of functions $2^{j/2}\psi(2^j x - k)$, $j, k \in \mathbf{Z}$, is an orthonormal basis of $L^2(\mathbf{R})$.*

In the above, $\mathbf{R}$ is the set of real numbers, $\mathbf{Z}$ is the set of integers, $L^2(\mathbf{R})$ is the set of square integrable real-valued functions on $\mathbf{R}$ and $L^\infty(\mathbf{R})$ is the set of bounded integrable functions on $\mathbf{R}$

The functions $\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k)$ are wavelets. Condition (1) determines the regularity of the mother wavelet. The localization property mentioned in Condition (2) extends also to the frequency domain. With regard to this property, many wavelets used in practice are compactly supported. Condition (3) specifies the oscillatory character, known as the vanishing moments property.

There are many mother wavelets, e.g. the well-known Haar wavelet, discovered by the mathematician Haar in 1910, Symmlet wavelet, Daubechies wavelet and

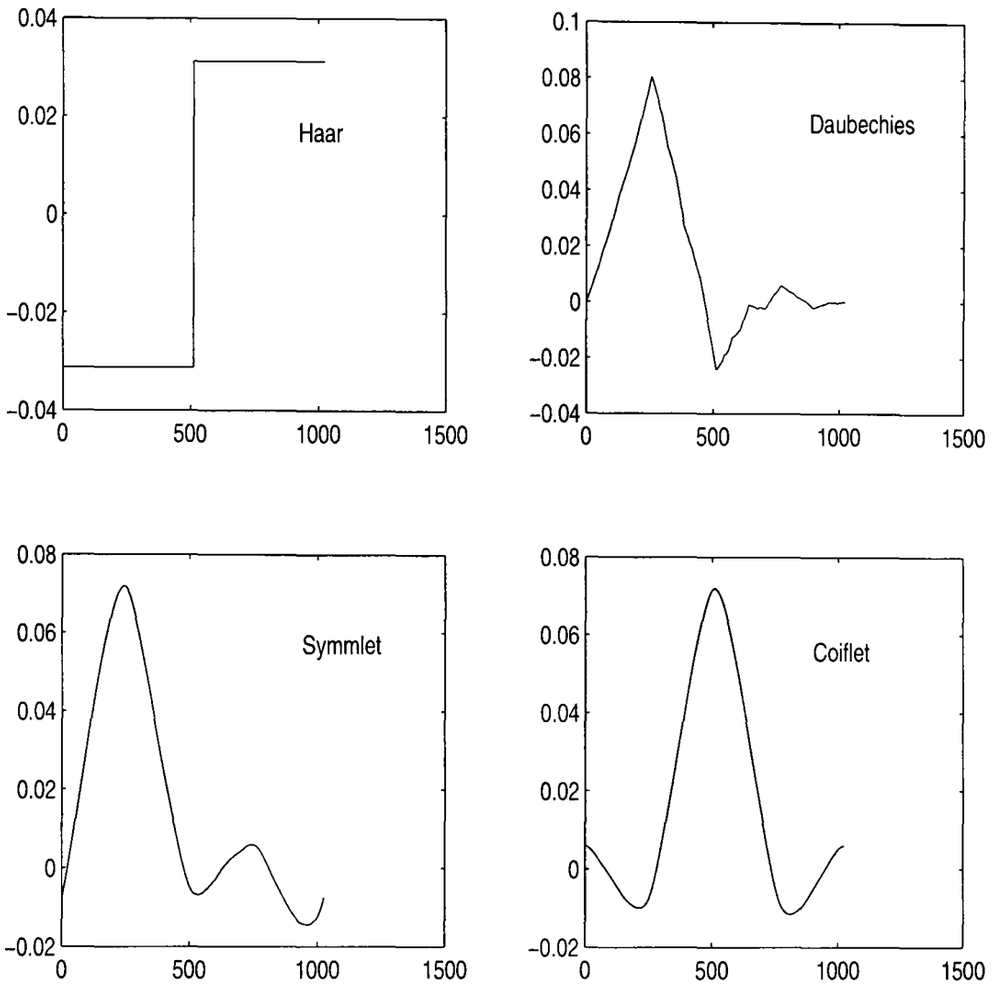

Figure 2.1: Several common wavelets

Coiflet wavelet, all discussed by Daubechies (1992). Although they have different expressions and characteristics, all of them satisfy the above definition. Figure 2.1 shows several different wavelets we will use in the following chapters. These pictures are created using MakeWavelet in **WaveLab802**.

2.2.2 Mallat's Multiresolution Analysis

Multiresolution analysis (MRA) is a tool for the constructive description of different wavelet bases (Mallat, 1989).

Definition 2.2 *A multiresolution analysis of* $L^2(\mathbf{R})$ *consists of a sequence of closed subspaces* $V_j \subset L^2(\mathbf{R})$, $j \in \mathbf{Z}$, *with the following properties:*

1. $V_j \subset V_{j+1}$;

2. $f(\cdot) \in V_j \Leftrightarrow f(2\cdot) \in V_{j+1}$;

3. $f(\cdot) \in V_0 \Leftrightarrow f(\cdot - k) \in V_0, \qquad \forall k \in \mathbf{Z}$;

4. $\bigcap_{j \in Z} V_j = \{0\}$;

5. $\overline{\bigcup_{j \in Z} V_j} = L^2(\mathbf{R}), \qquad \{V_j\},\ j \in \mathbf{Z},$ *is dense in* $L^2(\mathbf{R})$;

6. *a scaling function* $\phi \in V_0$ *with a nonvanishing integral exists such that the collection* $\{\phi(x-k) | k \in \mathbf{Z}\}$ *constitutes an orthonormal basis for* V_0.

From Condition (1), $V_j \subset V_{j+1}$, the orthogonal complement W_j of V_j can be found such that $V_{j+1} = V_j \oplus W_j$, where the symbol $\oplus$ stands for direct sum. Similarly, $V_j = V_{j-1} \oplus W_{j-1}$ and so on, from which it follows that W_{j-1} is also orthogonal to W_j and all the spaces W_j (unlike the spaces V_j) are mutually orthogonal.

Conditions (2) and (3) imply that $\forall j \in \mathbf{Z}$, $\{\phi_{jk} : k \in \mathbf{Z}\}$ constitutes an orthonormal basis for V_j, where

$$\phi_{jk}(x) = 2^{j/2}\phi(2^j x - k).$$

Let P_j be the orthogonal projection operator onto V_j. Condition (4) implies that, when $j \to -\infty$, we lose all the details of f, and $\{P_j f\}$ converging to $\{0\}$ in an L^2 space can be expressed as

$$\lim_{j \to -\infty} P_j f = 0,$$

where convergence of $P_j f$ to 0 in an L^2 space means $\lim_{j \to -\infty} \int_{\mathbf{R}} |P_j f(x)|^2 dx = 0$. On the other hand, Condition (5) ensures that the signal approximation converges to the original signal in the same sense:

$$\lim_{j \to \infty} P_j f = f.$$

Then the approximation $P_j f$ of a function f at resolution level j is given by

$$P_j f(x) = \sum_{k=-\infty}^{\infty} c_{jk}\phi_{jk}(x),$$

where

$$c_{jk} = \int_{-\infty}^{\infty} \phi_{jk}(x) f(x) dx.$$

The MRA and ϕ defined in Condition (6) are called r-regular, if $\phi \in C^r$, where C^r is the set of functions which have derivatives up to order r, and ϕ and every derivative up to order r can be chosen in such a way that for every integer $m \geq 0$ there exists a constant C_m satisfying

$$\left|\phi^{(j)}(x)\right| \leq \frac{C_m}{(1+|x|)^m} \quad \text{for} \quad j = 0, 1, \cdots, r.$$

2.3 Wavelet Transform

2.3.1 Wavelet Expansion

From Definition 2.2, $V_j \subset V_{j+1}$, and so there exists the orthogonal complement W_j such that $V_{j+1} = V_j \oplus W_j$ with $W_j \perp V_j$. Therefore for some $j_0 \in \mathbf{Z}$, there is a series of mutually orthogonal subspace W_j, $j \in \mathbf{Z}$, such that $V_j = V_{j_0} \oplus \bigoplus_{k=j_0}^{j-1} W_k$ for $j > j_0$. In conjunction with Conditions (4) and (5) of Definition 2.2, this implies that

$$\bigoplus_{j \in \mathbf{Z}} W_j = L^2(\mathbf{R}). \tag{2.1}$$

In other words, $L^2(\mathbf{R})$ can be decomposed into mutually orthogonal subspaces. A function ψ can be found (see, for example, Daubechies, 1992 and Mallat, 1989) such that the collection $\{\psi(x-k) | k \in \mathbf{Z}\}$ constitutes an orthonormal basis for W_0 and its integer translations and dilations $\{\psi_{jk} : \psi_{jk}(t) = 2^{j/2}\psi(2^j t - k), \quad j, k \in \mathbf{Z}\}$ form an orthonormal basis for $L^2(\mathbf{R})$.

Now we consider the generation of an orthonormal wavelet basis for functions $f \in L^2(\mathbf{R})$. For some $j_0 \in \mathbf{Z}$, $\{\phi_{j_0 k}, \psi_{jk} : j, k \in \mathbf{Z}, j \geq j_0\}$ forms an orthonormal basis for $L^2(\mathbf{R})$. Using this basis, the wavelet representation of a function f is

$$f(x) = \sum_{k \in Z} c_{j_0 k} \phi_{j_0 k}(x) + \sum_{j=j_0}^{\infty} \sum_{k \in Z} d_{jk} \psi_{jk}(x), \tag{2.2}$$

where the wavelet coefficients are

$$d_{jk} = \int_{-\infty}^{\infty} \psi_{jk}(x) f(x) dx \tag{2.3}$$

and the scaling function coefficients are

$$c_{j_0 k} = \int_{-\infty}^{\infty} \phi_{j_0 k}(x) f(x) dx. \tag{2.4}$$

The first term on the right hand side of (2.2) is the approximation $P_{j_0} f$ of f at resolution level j_0. Using $V_{j+1} = V_j \oplus W_j$, and since $\{\psi_{jk} : k \in \mathbf{Z}\}$ is a basis for W_j, $\sum_{k \in \mathbf{Z}} d_{jk} \psi_{jk}(x)$ is the difference between $P_j f$ and the finer resolution approximation $P_{j+1} f$. So for each value of j, the second term in (2.2) adds another level of detail into the representation.

Because of the vanishing moments property (Condition (3) of Definition 2.1), if f is smooth, the wavelet representation is very economical because there will be few wavelet coefficients d_{jk}, which are noticeably different from 0. Also, because wavelets are localised in time and scale, a discontinuity, or other high frequency feature, in f will only result in large wavelet coefficients for values of k corresponding to the location of the feature. Therefore, many functions can be adequately represented by a small number of wavelet coefficients. This property explains the application of wavelets to data compression and is also important in statistical applications.

2.3.2 The Discrete Wavelet Transform

In statistical settings we are typically concerned with discrete samples, rather than continuous functions, since functions are, in practice, observed at a finite number of discrete time points. Therefore a discrete wavelet transform (DWT) is required.

First consider some properties of ϕ. Since $\phi \in V_0 \subset V_1$, there exist an h_n such that

$$\phi(x) = \sum_n h_n \phi_{1n}(x), \tag{2.5}$$

where

$$h_n = < \phi, \phi_{1n} > = \int \phi(x) \phi_{1n}(x) dx.$$

Therefore, for all $j, k \in \mathbf{Z}$,

$$\phi_{j-1,k}(x) = \sum_n h_{n-2k}\phi_{jn}(x),$$

and since the ϕ_{jk}'s are orthonormal,

$$\sum_{n\in\mathbf{Z}} |h_n|^2 = 1.$$

Similarly, $\psi \in W_0 \subset V_1$, there exist g_n such that

$$\psi(x) = \sum_n g_n\phi_{1n}(x), \tag{2.6}$$

where

$$g_n =< \psi, \phi_{1n} >= \int \psi(x)\phi_{1n}(x)dx.$$

For all $j, k \in \mathbf{Z}$,

$$\psi_{j-1,k}(x) = \sum_n g_{n-2k}\phi_{jn}(x).$$

Mallat (1989) showed that one possible choice is $g_n = (-1)^n h_{1-n}$.

The recursive relationship between the scaling function and wavelet coefficients at successive levels can be obtained from (2.4) and (2.5)

$$\begin{aligned} c_{j-1,k} &= \int f(x)\Big\{\sum_n h_{n-2k}\phi_{jn}(x)\Big\}dx \\ &= \sum_n h_{n-2k}\Big\{\int f(x)\phi_{jn}(x)dx\Big\} \\ &= \sum_n h_{n-2k}c_{jn}, \end{aligned} \tag{2.7}$$

and from (2.3) and (2.6)

$$d_{j-1,k} = \sum_n g_{n-2k}c_{jn}. \tag{2.8}$$

This recursive relationship is another important property of wavelet transform held between scaling function coefficients and wavelet coefficients at successive levels. This property is related to the "pyramid" algorithm, a fast algorithm to calculate the coefficients provided by Mallat (1989).

Consider a vector of function values $\mathbf{f} = (f(t_1), ..., f(t_n))^T$ at equally spaced points t_i, and let n be an integer power of 2, say 2^{J+1}. A function can be constructed at level $J+1$ as follows:

$$\widehat{f}_{J+1}(x) = \sum c_{J+1,k}\phi_{J+1,k}(x),$$

where $c_{J+1,k} = f(t_k)$. The function $\widehat{f}_{J+1}(x)$ is an element of V_{J+1} and can be projected onto spaces V_J and W_J, giving

$$\begin{aligned}\widehat{f}_{J+1}(x) &= (P_{V_J}\widehat{f}_{J+1})(x) + (P_{W_J}\widehat{f}_{J+1})(x) \\ &= \sum_l c_{J,l}\phi_{J,l}(x) + \sum_l d_{J,l}\psi_{J,l}(x).\end{aligned}$$

The corresponding scaling coefficients in level J are

$$\begin{aligned}c_{J,l} &= <\widehat{f}_{J+1}, \phi_{J,l}> \\ &= \sqrt{2} <\widehat{f}_{J+1}, \sum_k h_{k-2l}\phi_{J+1,k}> \\ &= \sqrt{2}\sum_k h_{k-2l}c_{J+1,k} \qquad (2.9)\end{aligned}$$

Similarly, the wavelet coefficients are

$$d_{J,l} = \sqrt{2}\sum_k g_{k-2l}d_{J+1,k}. \qquad (2.10)$$

(2.9) and (2.10) are the same as the formulae (2.7) and (2.8) except the constant $2^{1/2}$. Applying this procedure recursively, we can find the coefficients $c_{j_0,k}$ and $d_{jk}, j_0 \leq j \leq J$.

Mallat (1989) also derived a reconstruction algorithm as follows. Note that at each level of the reconstruction, finer scale coefficients are obtained from coarser ones as illustrated by

$$\begin{aligned}\sum_k c_{j-1,k}\phi_{j-1,k}(x) + \sum_k d_{j-1,k}\psi_{j-1,k}(x) &= (P_{V_{j-1}}\widehat{f}_{J+1})(x) + (P_{W_{j-1}}\widehat{f}_{J+1})(x) \\ &= (P_{V_j}\widehat{f}_{J+1})(x) \\ &= \sum_k c_{j,k}\phi_{j,k}(x),\end{aligned}$$

where

$$\begin{aligned} c_{j,k} &= < \phi_{j,k}, P_{V_j} f_{J+1} > \\ &= \sum_l c_{j-1,l} < \phi_{j,k}, \phi_{j-1,l} > + \sum_l d_{j-1,l} < \phi_{j,k}, \psi_{j-1,l} > . \end{aligned}$$

This gives Mallat's "pyramid" algorithm.

2.3.3 Matrix Expression of DWT

To facilitate the presentation of the DWT later on, we will give the matrix expression of DWT here. For a detailed reference, see Percival and Walden (2000).

Let m represent the number of vanishing moments of the wavelet, j_0 the coarsest resolution level and $n = 2^{J+1}$ is the observations of the function. We can use an orthogonal matrix $\mathcal{W}$ associated with the orthonormal wavelet basis to represent the DWT. This matrix yields a vector $\mathbf{w}$ of the wavelet coefficients of $\mathbf{y}$ via

$$\mathbf{w} = \mathcal{W}\mathbf{y}.$$

We have the inverse formula

$$\mathbf{y} = \mathcal{W}^T\mathbf{w}. \tag{2.11}$$

Here vector $\mathbf{w}$ has $n = 2^{J+1}$ elements, indexed by two integers j and k: the wavelet coefficients $d_{j,k}$, $j = j_0, \ldots, J$, $k = 0, \ldots, 2^j - 1$ and the scaling function coefficients $c_{j_0,k}$, $k = 0, \ldots, 2^{j_0} - 1$.

Let us now decompose the elements of the vector $\mathbf{w}$ into $J - j_0 + 2$ subvectors, where J is called the finest resolution level. The first $J - j_0 + 1$ subvectors are denoted by $\mathbf{d_j}$, $j = J, \ldots, j_0$, and the jth such subvector contains all of the wavelet coefficients at resolution level j. Note that $\mathbf{d_j}$ is a column vector with 2^j elements. The final subvector is denoted as $\mathbf{c_{j_0}}$ and contains just the scaling coefficients at level j_0. Also, we can define the $\mathcal{W}_j$ and $\mathcal{V}_{j_0}$ matrices by partitioning the rows of $\mathcal{W}$ according to the partition of $\mathbf{w}$.

$$\begin{bmatrix} \mathbf{d_J} \\ \mathbf{d_{J-1}} \\ \vdots \\ \mathbf{d_{j_0}} \\ \mathbf{c_{j_0}} \end{bmatrix} = \begin{bmatrix} \mathcal{W}_J \\ \mathcal{W}_{J-1} \\ \vdots \\ \mathcal{W}_{j_0} \\ \mathcal{V}_{j_0} \end{bmatrix} \mathbf{y} \tag{2.12}$$

To interpret the components of $\mathcal{W}$, let $\mathbf{W}_{jk}$ denote the (j,k)th row of $\mathcal{W}_j$ and $\mathbf{V}_{j_0k}$ denote the (j_0,k)th row of $\mathcal{V}_{j_0}$. Write $W_{jk}(i)$ and $V_{j_0k}(i)$ for the components of $\mathbf{W}_{jk}$ and $\mathbf{V}_{j_0k}$, respectively. The inversion (2.11) becomes

$$y_i = \sum_{jk} d_{jk} W_{jk}(i) + \sum_{j_0k} c_{j_0k} V_{j_0k}(i).$$

For j and k bounded away from extreme cases by the conditions $m \ll j \ll J$ and $0 \ll k \ll 2^j$, we have the approximation of the components $\mathbf{W}_{j,k}$ and $\mathbf{V}_{j_0,k}$ of the matrix $\mathcal{W}$ as follows:

$$\sqrt{n} W_{j,k}(i) \approx 2^{j/2} \psi(2^j t), \qquad t = i/n - k2^{-j},$$

$$\sqrt{n} V_{j_0,k}(i) \approx 2^{j_0/2} \phi(2^{j_0} t), \qquad t = i/n - k2^{-j_0}.$$

For more discussion, see Dohono and Johnson (1995).

In matrix notation, each row in $\mathcal{V}_{j_0}$ is orthogonal to every row in $\mathcal{W}_j$ and also satisfies $\mathcal{V}_{j_0}\mathcal{V}_{j_0}^T = I_{2^{j_0}\times 2^{j_0}}$ and $\mathcal{W}_j\mathcal{W}_j^T = I_{2^j\times 2^j}$.

2.3.4 Translation Invariant DWT

In this section we describe a modified version of the discrete wavelet transform called the translation invariant DWT (TIDWT). The TIDWT have been discussed in the wavelet literature under different names (see Percival and Walden, 2000, Chapter 5), for example, "shift invariant DWT" (Beylkin, 1992), "stationary DWT" (Nason and Silverman, 1995) and "maximal overlap DWT" (Percival and Walden, 2000), but the transforms are essentially the same. Here we will use the name "translation invariant DWT" (Coifman and Donoho, 1995).

Firstly, we introduce the circularly shifted operator $\mathcal{T}$. For a signal $(x_t : 0 \leq t < n)$, we let $\mathcal{T}$ denote the circulant shift, $(\mathcal{T}x)_t = x_{(t+1) \bmod n}$. The operator $\mathcal{T}$ is unitary, and hence invertible $\mathcal{T}_{-1} = (\mathcal{T})^{-1}$.

In terms of the operator $\mathcal{T}$, the idea of TIDWT is just this: given a signal $\mathbf{x} = (x_t : 0 \leq t < n)$, apply the usual DWT twice, once to $\mathbf{x}$ and once to $\mathcal{T}\mathbf{x}$. Then we merge the two sets of DWT coefficients together to obtain the whole set of the wavelet coefficients. When reconstructing the signal $\mathbf{x}$, we split the whole set of coefficients into two sets in the same way as we merge them together and apply the usual inverse DWT to these two sets separately. Hence we obtain the two reconstructed signals, denoted as $\widehat{\mathbf{x}}$ and $\mathcal{T}\widehat{\mathbf{x}}$. Once we unshift $\mathcal{T}\widehat{\mathbf{x}}$ to be $\widehat{\mathbf{x}}' = \mathcal{T}_{-1}(\mathcal{T}\widehat{\mathbf{x}})$ and average $\widehat{\mathbf{x}}$ and $\widehat{\mathbf{x}}'$, we get the reconstructed signal.

In contrast to the DWT that restricts the sample size to be an integer power of 2, say 2^{J+1}, the fast algorithm developed for TIDWT is designed for any sample size N (see Percival and Walden, 2000). Also, TIDWT can suppress some visual artifacts produced by DWT, for example, Gibbs phenomena (see Coifman and Donoho, 1995). Of course, a computational price is paid for using the TIDWT.

2.3.5 Wavelet Analysis vs. Fourier Analysis

There are some similarities and some differences between Fourier analysis and wavelet analysis. The fast Fourier transform (FFT) and the discrete wavelet transform (DWT) are both linear operations and the mathematical properties of the matrices involved in the transforms are similar as well. In addition, the basis functions of both transforms are localized in frequency.

The most important difference between these two kinds of transforms is that individual wavelet functions are also localized in space while the Fourier sine and cosine functions are not. This localization feature in both frequency scale (via dilations) and space (via translations) makes wavelets very special in many cases. For example, one major advantage of wavelet methods is their very high adaptability and their ability to capture discontinuities and singularities. A related advantage of

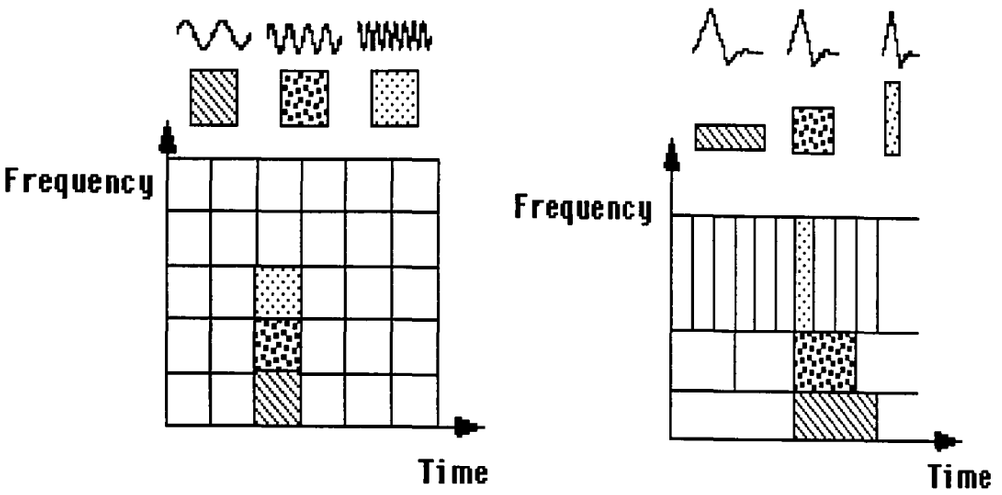

Figure 2.2: (a): Fourier transform (left); (b): wavelet transform (right). (Graps, 1997)

wavelets is sparseness which occurs when functions and operators are transformed into the wavelet domain. This sparseness results in a number of useful applications, such as removing noise from data, and will be discussed later in the thesis.

This time-frequency resolution difference between the Fourier transform and wavelet transform can be demonstrated by looking at the basis function coverage of the time-frequency plane. Figure 2.2(a) shows a windowed Fourier transform, where the windows are fixed for the whole frequency domain. In contrast Figure 2.2(b) shows one wavelet function, the Daubechies wavelet, with a variable time-frequency window which is governed by the dilation parameter and the translation parameter. Figure 2.2 originated in Graps (1997). For high frequencies (possibly representing discontinuities), some basis functions with narrow support are chosen, whereas basis functions with wide support are chosen for detailed analysis.

Another difference is that wavelet transforms do not have a single set of basis functions like the Fourier transform, which has just the sine and cosine functions. Different wavelet bases may be tailored to different applications.

2.4 Nonparametric Regression

Nonparametric regression has been a fundamental tool in data analysis over the past two decades and is still an expanding area of ongoing research. The goal is to recover an unknown function, say f, based on sampled data that are contaminated with random noise.

Suppose we observe responses $y_1, \ldots, y_n$ at nonrandom design points $x_1, \ldots, x_n$, which follow the model

$$y_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots, n, \tag{2.13}$$

where f is the unknown function to be estimated, and the ϵ_i are random errors, often assumed to be independent and identically distributed (IID) as $N(0, \sigma^2)$. In the nonparametric framework, only very general assumptions about f are made such as that it belongs to a certain class of smooth functions.

Nonparametric regression techniques provide a very effective and simple way of finding structure in data sets without the imposition of a parametric regression model. Some of the popular estimators are those based on kernel functions, smoothing splines and orthogonal series. Each of these approaches has its own particular strengths and weaknesses. However, there is a common drawback to these nonparametric regression techniques: they are likely to break down unless strong smoothness assumptions are satisfied everywhere.

2.4.1 Kernel Estimations

Kernel regression is probably the simplest and best understood method of nonparametric regression. Traditional approaches have involved the Nadaraya-Watson (NW) estimator (Nadaraya, 1964 and Watson, 1964) and local polynomial kernel estimators (Stone, 1977, Fan and Gijbels, 1996).

NW estimator

The NW estimator, based on the observation sample pairs $(x_1, y_1), \ldots, (x_n, y_n)$, is

$$\widehat{f}_h(x) = \frac{\sum_{i=1}^{n} K_h(x - x_i) y_i}{\sum_{i=1}^{n} K_h(x - x_i)}, \tag{2.14}$$

where $K_h(x) = h^{-1}K(x/h)$ is the kernel function with scale factor h. The kernel function K is usually chosen to be a continuous, bounded and symmetric function satisfying

$$\int_{-\infty}^{\infty} K(x) = 1,$$

and h is the smoothing parameter (or window width, or bandwidth).

If we define a sequence as

$$W_{hi}(x) = K_h(x - x_i)/\hat{m}_h(x) \quad i = 1, \ldots, n,$$

where

$$\hat{m}_h(x) = \sum_{i=1}^{n} K_h(x - x_i),$$

then

$$\hat{f}_h(x) = \sum_{i=1}^{n} W_{hi} y_i.$$

The NW estimator can be viewed as a weighted kernel estimator, with the kernel weights W_{hi} determined by $K(\cdot)$ and h.

A variety of kernel functions are possible in general, but both practical and theoretical considerations limit the choice. Commonly used kernel functions include the Gaussian kernel $K(t) = (2\pi)^{-1/2}\exp(-t^2/2)$, the "symmetric Beta family"

$$K(t) = \frac{1}{\text{Beta}(1/2, r+1)}(1 - t^2)_+^r, \quad r = 0, 1, \cdots,$$

where the subscript + denotes the positive part and a special case of which is due to Bartlett (1963):

$$K(u) = 0.75(1 - u^2)I(|u| \leq 1).$$

Local Polynomial Fitting

Suppose that the regression function f is smooth enough to be approximated by using Taylor's expansion

$$f(z) \approx \sum_{j=0}^{p} \frac{f^{(j)}(x)}{j!}(z-x)^j \equiv \sum_{j=0}^{p} \beta_j (z-x)^j, \tag{2.15}$$

where z is in a neighborhood of x. From a statistical modeling point of view, (2.15) models $f(z)$ locally by a simple polynomial model. This suggests using a locally weighted polynomial regression

$$\sum_{i=1}^{n} \{y_i - \sum_{j=0}^{p} \beta_j (x_i - x)^j\}^2 K_h(x_i - x), \tag{2.16}$$

where $K(\cdot)$ denotes a kernel function and h is a bandwidth.

The simplest polynomial to fit in such a neighborhood is a constant, which corresponds to $p = 0$. There is a similarity between local polynomial fitting and kernel smoothing. For fixed x, the kernel estimator $\hat{f}_h(x)$ with positive weights $W_{hi}(x)$ is the solution to the following minimization problem

$$\min_t \sum_{i=1}^{n} K_h(x - x_i)(y_i - t)^2 = \sum_{i=1}^{n} K_h(x - x_i)(y_i - \hat{f}_h(x))^2.$$

In this sense, the NW kernel smoother can be understood as a local constant polynomial fit.

2.4.2 Smoothing Spline Estimations

An important and widely used alternative approach is to estimate the regression curve $f(x)$ using a smoothing spline. Suppose that $x_1, \ldots . x_n$ are points in $[a, b]$ satisfying $a < x_1 < \ldots < x_n < b$ and we have observations $y_1, \ldots, y_n$. If we wanted the "smoothest possible" curve, in the sense of minimum curvature, to interpolate the given points, then a natural choice would be to use the curve that had the minimum value of $\int f''^2$ among all smooth curves interpolating the data. Using this measure, find $\widehat{f}_\alpha$ that minimizes the weighted sum

$$\sum_{i=1}^{n} (y_i - f(x_i))^2 + \alpha \int (f''(x))^2 dx. \tag{2.17}$$

The nonnegative real number $\alpha > 0$, called a smoothing parameter, governs the trade-off between smoothness and goodness-of-fit. The estimator $\widehat{f}_\alpha$ is called the smoothing spline estimator.

It turns out that among all curves in the class of twice differentiable functions interpolating the points (x_i, y_i), the one minimizing (2.17) is the so-called cubic spline; see Green and Silverman (1994), Silverman (1985), Reinsch (1967) and Good and Gaskins (1971).

2.4.3 Orthogonal Series Estimations

Let ϕ_j, $j = 1, \ldots, n$, be a sequence of given basis functions which are orthonormal with respect to the counting measure on the design points $x_1, \ldots, x_n$, that is

$$\sum_{i=1}^{n} \phi_j(x_i)\phi_k(x_i) = \delta_{jk} = \begin{cases} 0 & \text{if } j \neq k \\ 1 & \text{if } j = k \end{cases}$$

Then the regression function can be represented as

$$f(x_i) = \sum_{j=1}^{n} a_j \phi_j(x_i), \tag{2.18}$$

where $a_j = \sum_{k=1}^{n} f(x_k)\phi_j(x_k)$.

Generally, one could select an appropriate subset $\{\phi_j\}_{j \in \mathbf{I}}$, with $\mathbf{I} \subseteq \{1, \ldots, n\}$, of the basis functions and f can be estimated with the coefficients estimated as follows

$$\widehat{a}_j = \sum_{k=1}^{n} y_k \phi_j(x_k), \quad j \in \mathbf{I}.$$

Any available information on f would have an impact on choosing an appropriate basis. There are many possible choices of basis functions. Tarter and Lock (1993) consider the use of Fourier Series in Chapter 3 of their book. Other choices of basis functions include the Hermite functions (Schwartz, 1967) and Haar series (Engel, 1990). A desirable property of a basis would be that f may be represented economically by a few basis functions.

2.4.4 Wavelet Estimation

An appropriate choice of basis for the expansion is therefore a key point in relation to the efficiency of orthogonal series estimations we mentioned above. Wavelets provide an orthogonal basis with many attractive properties. It is therefore natural to consider applying the expansion approach using a wavelet series. See the review papers by Antoniadis *et al.* (2001) and Abramovich *et al.* (2000) for a detailed summary.

The wavelet expansion of a function f is given in (2.2) with the coefficients defined in (2.3) and (2.4). Since wavelet estimators are a form of orthogonal series estimator, the obvious estimators of these coefficients are

$$\widehat{d}_{jk} = \frac{1}{n}\sum_{i=1}^{n} y_i \psi_{jk}(x_i)$$

and

$$\widehat{c}_{j_0k} = \frac{1}{n}\sum_{i=1}^{n} y_i \phi_{j_0k}(x_i).$$

2.5 Wavelet Shrinkage and Thresholding

The key advantages of wavelet estimators can be fully exploited only when considering non-linear wavelet estimators. The non-linearity comes from shrinking or thresholding the empirical coefficients $\tilde{d}_{jk}$, while the scaling function coefficients $\tilde{c}_{j_0k}$ are kept untouched. The coefficients $\tilde{d}_{jk}$, $j = j_0, \ldots, J$, $k = 0, \ldots 2^j - 1$ and $\tilde{c}_{j_0k}$, $k = 0, \ldots, 2^{j_0} - 1$, come from DWT of the noisy data. Wavelet shrinkage and thresholding approaches were first introduced by Donoho and Johnstone (1994). The aim in this type of situation is to recover a signal in the presence of the noise, as indicated in (2.13), with nonrandom design point x_i taken to be $x_i = i/n$.

2.5.1 Wavelet Shrinkage and Thresholding Procedure

The method of wavelet shrinkage and thresholding consists of three main steps, which can be summarised as follows:

Step 1 Obtain the empirical wavelet coefficients by applying the DWT to the data;

Step 2 Modify these coefficients according to some procedure (typically by shrinkage or thresholding procedure);

Step 3 Apply the inverse DWT to the modified coefficients to obtain an estimate of f.

Once a wavelet basis has been chosen, Steps 1 and 3 are straightforward to implement, and very fast and efficient algorithms are available for performing the necessary calculations. Step 2, in which the aim is to "de-noise" the empirical wavelet coefficients, has been approached in a number of ways, including the following.

- The classic thresholding scheme, including the "hard" and "soft" thresholding methods, discussed in detail by Donoho and Johnstone (1994, 1995) and Donoho *et al.* (1995), and cross-validation scheme, see Nason (1996, 1999) and also Hall and Penev (2001).

- Frequentist "block" thresholding scheme. See Hall *et al.* (1997, 1998, 1999), Cai (1999, 2002) and Cai and Silverman (2001).

- Shrinkage or thresholding of wavelet coefficients based on Bayes or empirical Bayes (EB) methods. For work on Bayes or EB shrinkage of individual wavelet coefficients, see Chipman *et al.* (1997), Clyde *et al.* (1998), Abramovich *et al.* (1998) and Clyde and George (2000).

- Bayesian block shrinkage scheme. See De Canditiis and Vidakovic (2004) and Abramovich *et al.* (2002).

A useful summary of the above methods is given by Antoniadis *et al.* (2001).

2.5.2 Classical Thresholding Schemes

Since the wavelet representation of many kinds of function is very economical, it is reasonable to assume that there are a few large value wavelet coefficients concen-

trated near the areas of major spatial activity, e.g discontinuities, but the majority of wavelet coefficients are small. Also, owing to the fact that the wavelet transform is orthogonal, if the ϵ_i are assumed to be independent Gaussian noise, then the wavelet coefficients will also be contaminated with independent Gaussian noise. So in this case, the empirical wavelet coefficients can be written as

$$\tilde{d}_{jk} = d_{jk} + \epsilon_{jk} \tag{2.19}$$

and $\tilde{d}_{jk}$ is distributed as

$$\tilde{d}_{jk} \sim N(d_{jk}, \sigma^2). \tag{2.20}$$

Based on these assumptions, Donoho and Johnstone (1994, 1995) suggested two types of thresholding methods: hard and soft thresholding. Hard thresholding sets all the wavelet coefficients to be 0 if their absolute values are below a certain threshold $\lambda \geq 0$:

$$\widehat{d}_{jk} = \eta_\lambda(\tilde{d}_{jk}) = \tilde{d}_{jk} I(|\tilde{d}_{jk}| > \lambda). \qquad (hard \quad thresholding) \tag{2.21}$$

Soft thresholding shrinks the wavelet coefficients that are larger than the threshold by λ:

$$\widehat{d}_{jk} = \eta_\lambda(\tilde{d}_{jk}) = sgn(\tilde{d}_{jk}) \max(0, |\tilde{d}_{jk}| - \lambda). \qquad (soft \quad thresholding) \tag{2.22}$$

Hard and soft thresholdings are illustrated in Figure 2.3.

After studying the performance of these thresholding methods, Donoho and Johnstone (1994, 1995) concluded that the resulting function estimate is asymptotically minimax for a wide variety of loss functions and functions f belonging to a wide range of smoothness classes. More importantly, they show that the wavelet estimator is nearly optimal for a wide variety of objectives.

Choices of Threshold

Clearly, an appropriate choice of a threshold value λ is fundamental to the effectiveness of the procedure described in the previous section. Too large a threshold might

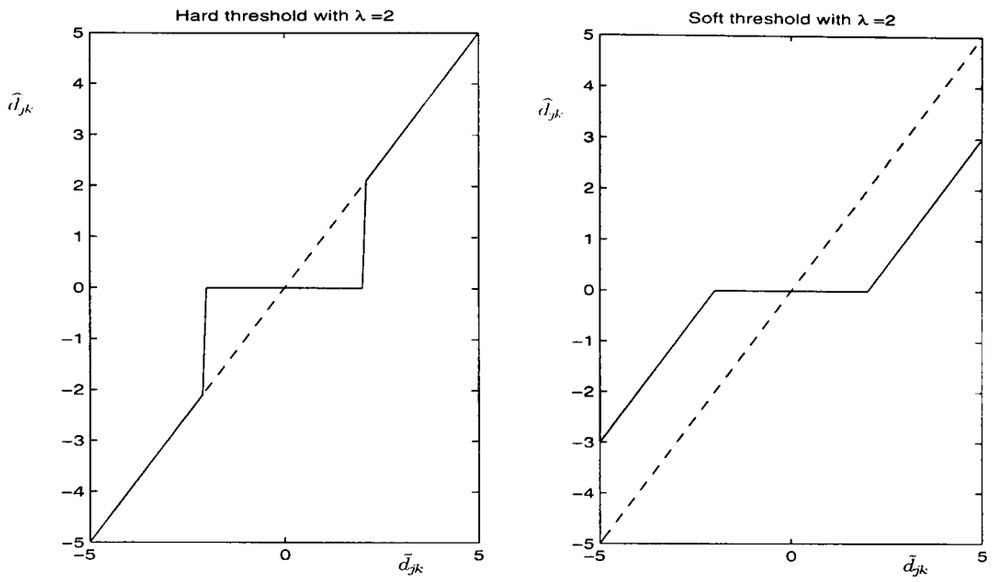

Figure 2.3: Hard thresholding (left) and soft thresholding (right)

cut off important parts of the true function underlying the data, whereas too small a threshold may excessively retain noise in the reconstruction.

Universal Threshold

Donoho and Johnstone (1994) proposed the universal threshold:

$$\lambda_{un} = \sigma\sqrt{2\log(n)}.$$

When σ is unknown, σ may be replaced by a robust estimate $\widehat{\sigma}$, such as the median absolute deviation (MAD) of the wavelet coefficients at the finest level $J = \log(N)-1$ divided by 0.6745 and can be expressed as

$$\widehat{\sigma} = MAD\{d_{Jk}, k = 1, \ldots, 2^J\}/0.6745. \tag{2.23}$$

Despite the simplicity of such a threshold, Donoho and Johnstone (1994) showed that if $\{\epsilon_i\}_{i=1}^n$ is a white noise sequence with variance 1,

$$P(\max|\epsilon_i| > \sqrt{2\log n}) \to 0 \qquad n \to \infty.$$

This means that, with high probability, all the pure noise coefficients will be thresholded to zero.

Although it has good asymptotic properties, the universal threshold depends on the data only through σ (or its estimate). In fact, for large samples, it may be shown that λ_{un} will remove with high probability all the noise in the reconstruction, but part of the real underlying function might also be lost. As a result, the universal threshold tends to oversmooth in practice.

SureShrink Threshold

Donoho and Johnstone (1995) introduced a procedure, SureShrink, which was based on minimizing the Stein unbiased risk estimate (Sure). Let $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \mathbf{1})$ be multivariate normal observations with mean vector μ and covariance matrix $\mathbf{1}$ with 1 along the diagonal and 0 elsewhere. A fixed estimator $\widehat{\boldsymbol{\mu}}(\mathbf{x})$ of $\boldsymbol{\mu}$ can be written as:

$$\widehat{\mu}(\mathbf{x}) = \mathbf{x} + \mathbf{g}(\mathbf{x}),$$

where $\mathbf{g} = (g_i)_{i=1}^p$ is a function from $\mathbf{R}^p$ to $\mathbf{R}^p$. Stein (1981) showed that if $\mathbf{g}(\mathbf{x})$ is weakly differentiable, then

$$E_\mu ||\widehat{\boldsymbol{\mu}}(\mathbf{x} - \boldsymbol{\mu})|| = p + E_\mu \{||\mathbf{g}(\mathbf{x})||^2 + 2\nabla \cdot \mathbf{g}(\mathbf{x})\}$$

where

$$\nabla \cdot \mathbf{g} \equiv \sum_i \frac{\partial}{\partial x_i} g_i.$$

Notice that the soft threshold can be written as

$$\widehat{\mu}(\mathbf{x}) = \mathbf{x} - sgn(\mathbf{x}) \min(|\mathbf{x}|, \lambda).$$

Using Stein's result, the quantity

$$SURE(\lambda, \mathbf{x}) = p - 2 \cdot \#\{i : |x_i| \leq \lambda\} + \sum_{i=1}^{p} (|x_i| \wedge \lambda)^2,$$

$|x_i| \wedge \lambda = \min(|x_i|, \lambda)$, is an unbiased estimate of the risk $E_\mu ||\widehat{\boldsymbol{\mu}}(\mathbf{x} - \boldsymbol{\mu})||$.

This procedure is very simple to implement and attains superior adaptive properties than the Universal threshold method. It has been shown that SureShrink

is smoothness-adaptive: if the unknown function contains jumps, the reconstruction does also; if the unknown function has a smooth piece, the construction is as smooth as the mother wavelet allows. In addition, this shrinkage can be tuned to be asymptotically minimax over a wide range of smoothness classes.

Threshold Selected by Cross-validation

Cross-validation (CV) is widely used as an automatic procedure to choose a smoothing parameter in many areas of statistics, e.g Silverman (1986) and Green and Silverman (1994). Nason (1996) proposed two modified cross-validation (CV) methods for choosing the threshold λ, which minimises the mean integrated square error (MISE) between the wavelet shrinkage estimator $\widehat{f}_\lambda(x)$ and the true function $f(x)$,

$$M(\lambda) = E \int \{\widehat{f}_\lambda(x) - f(x)\}^2 dx. \tag{2.24}$$

Twofolded cross-validation, which works by leaving out half the data points, can be used to select a threshold for a wavelet shrinkage estimator based on $n = 2^{J+1}$ points. Firstly, take all the evenly indexed data points $\{y_{2j}\}$, $j = 1, \ldots, n/2$, to form a wavelet shrinkage estimator $\widehat{f}^E_\lambda$ using a particular threshold while the remaining points are used to estimate the MISE at that threshold. Then, in order to compare the $\widehat{f}^E_\lambda$ with the left out noisy data, an interpolated version of $\overline{f}^E_{\lambda,j}$ can be formed

$$\overline{f}^E_{\lambda,j} = \frac{1}{2}(\widehat{f}^E_{\lambda,j+1} + \widehat{f}^E_{\lambda,j}),$$

where $\widehat{f}^E_{\lambda,n/2+1} = \widehat{f}^E_{\lambda,1}$ is assumed. The same procedure is repeated for the odd indices $\{y_{2j-1}\}$ to give the interpolant $\overline{f}^O_{\lambda,j}$ and the full estimate for $M(\lambda)$ is

$$\widehat{M}(\lambda) = \sum_{j=1}^{n/2} \{(\overline{f}^E_{\lambda,j} - y_{2j-1})^2 + (\overline{f}^O_{\lambda,j} - y_{2j})^2\}.$$

Nason (1996) also discussed a modified leave-one-out CV, which can be used for any number of data points. After moving one point y_i, $1 < i < n$, the remaining points were split into two groups:

$$G_L = \{y_1, \ldots y_{i-1}\}$$

and

$$G_R = \{y_i, ...y_n\}.$$

Two new groups can be obtained by reflecting G_L and G_R to dyadic length from right and left side respectively and wavelet estimators $\widehat{f}_{L,\lambda}$ and $\widehat{f}_{R,\lambda}$ can be derived. The removed point is predicted by averaging the rightmost point of $\widehat{f}_{L,\lambda}$ and the leftmost point of $\widehat{f}_{R,\lambda}$ to give $\widehat{y}_{\lambda,-i}$. The estimated MISE to be minimised over λ is

$$\widehat{M}(\lambda) = \sum_{i=2}^{n-2}(y_i - \widehat{y}_{\lambda,-i})^2$$

However, it is worth noting that cross-validation methods do not work well with serially correlated data.

2.5.3 Frequentist Block Thresholding Schemes

The methods mentioned above involve term by term thresholding, which "kill" or "retain" coefficients on the basis of their individual magnitudes, and information on neighbouring coefficients has no influence on the treatment of particular coefficients. Motivated by the need for spatial adaptivity, Hall *et al.* (1998,1999) first suggested grouping wavelet coefficients into blocks, modelling them blockwise and exploiting the information that coefficients convey about the size of their nearby neighbour.

Cai (1999) and Cai and Silverman (2001) studied local block thresholding rules and provided a new BlockShrink procedure. In this procedure, after wavelet transformation, the empirical wavelet coefficients are grouped at each resolution level j into blocks of length $L = \lfloor \log n \rfloor$, where $\lfloor \cdot \rfloor$ denotes the integer part. All the coefficients of a block (jb) are retained if the energy in this block

$$\sum_{k \in (jb)} d_k^2 > const. \times L\sigma^2, \tag{2.25}$$

in which case it is deemed that this block contains significant information about the function; otherwise the block is deemed insignificant and all the coefficients are set to zero. The aim of BlockShrink procedure is to increase estimation accuracy by

utilizing information about neighbouring wavelet coefficients. It is shown that the estimators produced by this procedure are asymptotically optimal both for global and local estimation and are easy to implement.

The block size L affects the achieving of the global and local adaptivities of the frequentist block thresholding approach. We will summarise discussion about the choice of the block size L in § 4.4.

2.5.4 Bayesian Wavelet Shrinkage and Thresholding

Various Bayesian approaches for thresholding and non-linear shrinkage in general have been proposed recently. See for example Chipman *et al.* (1997), Abromovich and Sapatinas (1999), Abramovich *et al.* (2000), Clyde and George (1999, 2000) and Johnstone and Silverman (1998, 2005). These methods have been shown to be effective. In these approaches, a prior distribution is imposed on the wavelet coefficients, which is designed to capture the sparseness of the wavelet expansions that is common to most applications. The function can then be estimated by applying a suitable Bayesian rule to the resulting posterior distribution of wavelet coefficients.

In general, a Bayesian rule $\eta(x)$ is a shrinkage rule if and only if η is antisymmetric and increasing on $(-\infty, \infty)$ and $0 \leq \eta(x) \leq x$ for all $x \geq 0$. The family of shrinkage rules $\eta(x,t)$ will be a thresholding rule with threshold t if and only if

$$\eta(x,t) = 0 \text{ if and only if } |x| \leq t.$$

A popular prior model for each wavelet coefficient d_{jk} is a mixture of one normal distribution and a point mass at zero. The normal distribution with large variance represents the significant coefficients while a point mass at zero represents the negligible ones. A hierarchical model can be expressed as

$$d_{jk}|r_j \sim r_j N(0, \lambda_j^2) + (1 - r_j)\delta(0), \tag{2.26}$$

where $r_j \sim Bernoulli(p_j)$ for different resolution level j and $\delta(0)$ is a point mass at zero. The binary random variable r_j determines whether the relevant wavelet

coefficient is nonzero ($r_j = 1$), and comes from an $N(0, \lambda_j^2)$ distribution, or zero ($r_j = 0$), and arises from a point mass at zero.

From (2.19), the posterior cumulative distribution of d_{jk} conditional on the empirical wavelet coefficient $\tilde{d}_{jk}$ and σ^2 is given by

$$\begin{aligned} d_{jk}|\tilde{d}_{jk}, \sigma^2 \sim & \Pr(r_{jk} = 1|\tilde{d}_{jk}, \sigma^2) N\Big(\frac{d_{jk}^2}{\sigma^2 + d_{jk}^2}\tilde{d}_{jk}, \frac{\sigma^2 d_{jk}^2}{\sigma^2 + d_{jk}^2}\Big) \\ & + \{1 - \Pr(r_{jk} = 1|\tilde{d}_{jk}, \sigma^2)\}\delta(0). \end{aligned} \tag{2.27}$$

The posterior probabilities can be expressed as

$$\Pr(r_{jk} = 1|\tilde{d}_{jk}) = \frac{1}{1 + O_{jk}(\tilde{d}_{jk}, \sigma^2)}, \tag{2.28}$$

where the posterior odds ratios $O_{jk}(\tilde{d}_{jk}, \sigma^2)$ are given by

$$O_{jk}(\tilde{d}_{jk}, \sigma^2) = \frac{1 - p_j}{p_j} \cdot \frac{(\sigma^2 + \lambda^2)^{1/2}}{\sigma} \exp\Big(-\frac{\lambda^2 \tilde{d}_{jk}^2}{2\sigma^2(\sigma^2 + \lambda_j)}\Big). \tag{2.29}$$

Suitable Bayesian wavelet shrinkage and thresholding estimators are the posterior mean estimator, the posterior median estimator and the "hypothesis testing" estimator, (see below).

Shrinkage Estimates Using Posterior Mean Approaches

Clyde *et al.* (1998) obtained wavelet shrinkage estimates by considering the posterior mean. Assuming that an accurate estimate of the noise variance is available, the closed form expressions for the posterior mean of wavelet coefficient d_{jk} conditionally on $\tilde{d}_{jk}$ and σ^2, can be derived from (2.28), (2.28) and (2.29) as

$$E(d_{jk}|\tilde{d}_{jk}, \sigma^2) = \frac{1}{1 + O_{jk}(\tilde{d}_{jk}, \sigma^2)} \cdot \frac{\lambda_j^2}{\sigma^2 + \lambda_j^2}\tilde{d}_{jk}. \tag{2.30}$$

It shrinks the empirical wavelet coefficients $\tilde{d}_{jk}$ by a nonlinear factor of

$$\frac{1}{1 + O_{jk}(\tilde{d}_{jk}, \sigma^2)} \cdot \frac{\lambda_j^2}{\sigma^2 + \lambda_j^2}.$$

This is an extreme case of Chipman *et al.* (1997), where the prior was chosen to be a mixture of two normal distributions. A normal distribution with small variance

is used to concentrate on the mass near 0 while the distribution with large variance spreads out the rest of the mass across larger values. Antoniadis *et al.* (2001) pointed the important distinction between the uses of a scale mixture of two normal distributions and a scale mixture of a normal distribution and a point mass at zero. In the former case, no wavelet coefficient estimate based on the posterior analysis will be exactly equal to zero. However, in the latter case, with a proper choice of a Bayes rule, it is possible to get wavelet coefficient estimates that are exactly zero.

Thresholding Estimates Using Posterior Median Approaches

Abramovich *et al.* (1998) proposed a Bayesian thresholding rule based on the posterior median. From (2.28), (2.28) and (2.29), solving the equation $F(d_{jk}|\tilde{d}_{jk}) = 0.5$ yields a thresholding procedure

$$Med(d_{jk}|\tilde{d}_{jk}, \sigma^2) = sgn(\tilde{d}_{jk}) \max(0, \varsigma_{jk}) \tag{2.31}$$

where

$$\varsigma_{jk} = \frac{\lambda_j^2}{\sigma^2 + \lambda_j^2}|\tilde{d}_{jk}| - \frac{\lambda_j \sigma}{(\sigma^2 + \sigma_j^2)^{1/2}} \Phi^{-1}\left\{\frac{1 + \min(O_{jk}, 1)}{2}\right\}$$

and Φ is the cumulative distribution function of a standard normal random variable.

We can see that (2.31) corresponds to a thresholding rule with threshold λ_j: for all $\tilde{d}_{jk}$ in some implicitly interval $[-\lambda_j, \lambda_j]$, the quantity ς_{jk} is negative and $Med(d_{jk}|\tilde{d}_{jk}, \sigma^2) = 0$ while for large $\tilde{d}_{jk}$ the thresholding rule is asymptotic to linear shrinkage by a factor of $\lambda_j^2/(\sigma^2 + \lambda_j^2)$.

Thresholding Estimates Using Hypothesis Testing Approaches

Vidakovic (1998) considered a Bayesian method similar to the statistical hypothesis testing method. For each wavelet coefficient d_{jk}, this method involves testing the hypothesis

$$H_0 : d_{jk} = 0 \text{ versus } H_1 : d_{jk} \neq 0.$$

If the hypothesis H_0 is rejected, the d_{jk} is estimated by $\tilde{d}_{jk}$. At each level $j = j_0, \ldots, J$, the prior distribution could therefore be taken as

$$d_{jk} \sim \pi_j \xi(d_{jk}) + (1 - \pi_j)\delta(0),$$

where ξ describes the behaviour of d_{jk} when d_{jk} is nonzero, which occurs with probability π_j.

Abramovich and Sapatinas (1999) obtained the Bayes factor, (a particular case of the "hypothesis testing" approach), by considering the prior mixture (2.26) in the above setting:

$$\widehat{d}_{jk} = \tilde{d}_{jk} I(O_{jk} < 1) \text{ with } O_{jk} = \frac{\Pr(H_0|\tilde{d}_{jk})}{\Pr(H_1|\tilde{d}_{jk})}, \tag{2.32}$$

where I is the indicator function and O_{jk} is the posterior odds ratio that is given by (2.29).

We can see that the thresholding estimator proposed by Abramovich and Sapatinas (1999) mimics the hard thresholding rule (2.21). A wavelet coefficient $\tilde{d}_{jk}$ will be thresholded (i.e. set to zero) if the corresponding posterior odds ratio $O_{jk} > 1$ and will be kept as it is otherwise.

2.5.5 Bayesian Block Wavelet Shrinkage and Thresholding

To increase estimation precision, Abramovich *et al.* (2002) proposed a multivariate normal model to incorporate information about neighbouring empirical wavelet coefficients to form block wavelet shrinkage and block wavelet thresholding estimators. At each resolution level, the wavelet coefficients $\tilde{d}_{jk}$ are grouped into nonoverlapping blocks b_{jK} of certain length $l = l_j$.

Let m_j be the number of blocks ($K = 1, \ldots, m_j$) at level j and consider the following prior model on $\mathbf{d}_{jK}$

$$\boldsymbol{d_{jK}} \mid r_{jK} \sim r_{jK} N(\mathbf{0}, V_j) + (1 - r_{jK})\boldsymbol{\delta(0)}, \tag{2.33}$$

where $\boldsymbol{\delta(0)}$ is a point mass at the zero vector. The matrix V_j is an $l_j \times l_j$ nonsingular covariance matrix given by $V_j = \lambda_j^2 P_j$, where P_j is the $l_j \times l_j$ matrix with elements

$P_j[k,l] = \rho_j^{|k-l|}$ for $k,l = 1,\ldots,l_j$ and $|\rho_j| < 1$. It is also assumed that r_{jK} has the distribution as:

$$\Pr(r_{jK} = 1) = 1 - \Pr(r_{jK} = 0) = p_j, \qquad 0 \leq p_j \leq 1. \tag{2.34}$$

and at each level j the blocks $b_{jK}, K = 1,\ldots,m_j$ are independent. Using Bayes' theorem, results parallel to (2.30) and (2.31) may be derived.

It was reported by Abramovich *et al.* (2002) that the proposed empirical Bayes block wavelet shrinkage and block thresholding estimators outperformed the non-Bayesian block wavelet thresholding estimators in the examples considered.

In practice, the hyperparameters p_j, λ_j^2 and σ^2 need to be estimated before any of above approaches can be used. Several methods have been used, for example estimating the noise level σ with the robust estimate (2.23), and then obtaining maximum likelihood estimation of p_j and λ_j^2 using the EM algorithm (see Clyde and George, 1999). Another possibility is maximum likelihood estimation of σ^2 and p_j and λ_j^2 together using the EM algorithm (see Clyde and George, 2000).

2.6 Summary and Research Plan

Previous work related to wavelet shrinkage and thresholding is summarised in §2.5. From the above survey, it is clear there are some areas that would benefit from further investigation, despite the progress made by authors in recent years. In this thesis, the focus is on the following topics:

1. It would be desirable and of interest to develop a Bayesian approach to block shrinkage/thresholding based on the sum of squares of the wavelet coefficients in a block. Although block thresholding in a frequentist framework has been studied both from a theoretical point of view and in simulation studies, Bayesian block shrinkage approaches have received much less attention.

2. More computationally efficient ways to estimate the hyperparameters in the Bayesian framework are needed. Previous work has mainly used the EM algo-

rithm (or a combination of the EM algorithm and robust estimation of σ^2) to estimate the hyperparameters, which tend to be far slower than the frequentist methods.

3. Correlated data often arise in more realistic settings. There is a need to pay more attention to this issue, and to extend existing methods to the correlated data setting.

This thesis is divided into three main parts: theoretical results, practical issues and extension to correlated data. The theoretical results part (Chapter 3) develops new Bayesian methodology based upon the non-central χ^2 distribution. In the second part (Chapter 4), the Bayesian methodology presented earlier is used to construct Bayesian block shrinkage and thresholding procedures for wavelet coefficients obtained from noisy data. In the third part, a semi-parametric model is discussed in Chapter 5, which is focused on estimating the covariance structure of the correlated noise.

Appendix A: Some Bayesian Analysis Background

We review the basic elements of Bayesian inference which will be used later. It is obvious that we should be able to improve the information or the models we develop and inference if we incorporate whatever a priori qualitative or quantitative knowledge we have available. The Bayesian approach allows us to assign prior distributions to the parameters in the model, and then to update these priors in light of the data, yielding a posterior distribution via Bayes' Theorem:

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}. \tag{2.35}$$

The ability to include prior information in the model is not only an attractive pragmatic feature of the Bayesian approach, but also theoretically vital for guaranteeing coherent inferences.

A.1 Basic Theory

In this section we will review the fundamentals of the Bayesian paradigm. For a rigorous and detailed survey of Bayesian methodology, see Bernardo and Smith (1994), O'Hagan (1994).

Bayes' Theorem

In the Bayesian approach, to specify the model for the observed data $\mathbf{x} = (x_1, \ldots, x_n)$ given the vector of the unknown parameters $\boldsymbol{\theta}$, assumed to lie in a parameter space $\boldsymbol{\Theta}$, we define the likelihood function $L(\boldsymbol{x} \mid \boldsymbol{\theta})$ as a joint probability of observed data $\mathbf{x}$ given the parameter vector $\boldsymbol{\theta}$ and let $\pi(\boldsymbol{\theta})$ denote the prior distribution for the parameters. Inference concerning $\boldsymbol{\theta}$ is then based on its posterior distribution $\pi(\boldsymbol{\theta} \mid \mathbf{x})$, given by

$$\pi(\boldsymbol{\theta} \mid \mathbf{x}) \propto L(\mathbf{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta}) = \frac{L(\mathbf{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta} \in \Theta} L(\mathbf{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta}) d\boldsymbol{\theta}}. \tag{2.36}$$

We refer to this formula as Bayes' Theorem. The integral in the denominator on the right-hand side of (2.36) is a normalising constant and its calculation has tra-

ditionally been a severe obstacle in Bayesian computation. Clearly, the likelihood may be multiplied by any constant (or any function of $\mathbf{x}$ alone) without altering the posterior.

Moreover, the prior distribution $\pi(\boldsymbol{\theta})$ of $\boldsymbol{\theta}$ may also be expressed conditionally on some unknown hyperparameters $\boldsymbol{\phi} \in \boldsymbol{\Phi}$ as $\pi(\boldsymbol{\theta} \mid \boldsymbol{\phi})$. The prior must be completed by a distribution of $\boldsymbol{\phi}$, say $g(\boldsymbol{\phi})$, yielding

$$\pi(\boldsymbol{\theta}) = \int_{\boldsymbol{\phi} \in \boldsymbol{\Phi}} \pi(\boldsymbol{\theta} \mid \boldsymbol{\phi}) g(\boldsymbol{\phi}) d\boldsymbol{\phi}. \tag{2.37}$$

Such a model is called a hierarchical model because of the way in which the distribution of parameters in each level of the hierarchical model depends on the parameters of next level. We could also write the distribution of $\boldsymbol{\phi}$ conditional on some more parameters and this process could continue as far as is needed. For the detailed discussion, see O'Hagan (1994).

A.2 Prior Distributions

The choice of prior distributions represents information available about unknown parameters. Ideally, we would like to work with families of prior distributions which are sufficiently flexible to represent various states of prior knowledge and at the same time result in computationally tractable posterior distribution.

Conjugate Priors. When choosing a prior from a parametric family, some choices may be more computationally convenient than others. In some problems it is possible to select a distribution which is conjugate to the likelihood, that is, one that leads to a posterior belonging to the same family as the prior. It is shown in Morris (1983) that exponential families, a commonly used form of likelihood function, do in fact have conjugate priors, so that this approach will typically be available in practice.

Mixture Priors: Continuous Case. It is common to specify a prior as a mixture of conjugate priors. Such mixtures can offer a very diverse family of distributions

that is capable of representing much more varied prior beliefs than a single conjugate prior. Also, the posterior distribution is then a mixture of posteriors. For example, consider a three-level hierarchical model in which a mixture prior is expressed as (2.37). Then the posterior can be easily calculated as a continuous mixture of component posterior distribution, $\pi(\boldsymbol{\theta} \mid \mathbf{x}, \boldsymbol{\phi})$:

$$\pi(\boldsymbol{\theta} \mid \mathbf{x}) = \int_{\boldsymbol{\phi} \in \boldsymbol{\Phi}} \pi(\boldsymbol{\theta} \mid \mathbf{x}, \boldsymbol{\phi}) g(\boldsymbol{\phi} \mid \mathbf{x}) d\boldsymbol{\phi}, \tag{2.38}$$

where

$$\pi(\boldsymbol{\theta} \mid \mathbf{x}, \boldsymbol{\phi}) = \frac{L(\mathbf{x} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \mid \boldsymbol{\phi})}{\int_{\boldsymbol{\theta} \in \Theta} L(\mathbf{x} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \mid \boldsymbol{\phi}) d\boldsymbol{\theta}}$$

and

$$\begin{aligned} g(\boldsymbol{\phi} \mid \mathbf{x}) &= \frac{g(\boldsymbol{\phi}) \int_{\boldsymbol{\theta} \in \Theta} L(\mathbf{x} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \mid \boldsymbol{\phi}) d\boldsymbol{\theta}}{\int_{\boldsymbol{\theta}, \boldsymbol{\phi}} L(\mathbf{x} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \mid \boldsymbol{\phi}) g(\boldsymbol{\phi}) d\boldsymbol{\phi} d\boldsymbol{\theta}} \\ &= \frac{g(\boldsymbol{\phi}) \int_{\boldsymbol{\theta} \in \Theta} L(\mathbf{x} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \mid \boldsymbol{\phi}) d\boldsymbol{\theta}}{\int_{\boldsymbol{\theta} \in \Theta} L(\mathbf{x} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}}. \end{aligned}$$

Mixture Priors: Discrete Case. Suppose that in a hierarchical model, $\boldsymbol{\phi}$ is a discrete variable taking values $\boldsymbol{\phi_1}, \boldsymbol{\phi_2}, \ldots, \boldsymbol{\phi}_m$. Let $\Pr(\boldsymbol{\phi} = \boldsymbol{\phi}_i) = p_i$ and $\pi(\boldsymbol{\theta} \mid \boldsymbol{\phi} = \boldsymbol{\phi}_i) = \pi_i(\boldsymbol{\theta})$, for $i = 1, 2, \ldots, m$, then the unconditional prior distribution (2.37) is

$$\pi(\boldsymbol{\theta}) = \sum_{i=1}^{m} p_i \pi_i(\boldsymbol{\theta}). \tag{2.39}$$

This is called a mixture of the prior distribution $\pi_i(\boldsymbol{\theta})$ with weights p_i. By simple calculation, we can obtain the posterior as a discrete mixture of component posterior distributions $\pi(\boldsymbol{\theta} \mid \mathbf{x}, \boldsymbol{\phi}_i)$:

$$\pi(\boldsymbol{\theta} \mid \mathbf{x}) = \sum_{i=1}^{m} q_i \pi(\boldsymbol{\theta} \mid \mathbf{x}, \boldsymbol{\phi}_i), \tag{2.40}$$

where

$$\pi(\boldsymbol{\theta} \mid \mathbf{x}, \boldsymbol{\phi}_i) = \frac{L(\mathbf{x} \mid \boldsymbol{\theta}) \pi_i(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta} \in \Theta} L(\mathbf{x} \mid \boldsymbol{\theta}) \pi_i(\boldsymbol{\theta}) d\boldsymbol{\theta}}$$

and

$$q_i = \frac{p_i \int_{\boldsymbol{\theta} \in \Theta} L(\mathbf{x} \mid \boldsymbol{\theta}) \pi_i(\boldsymbol{\theta})}{\sum_{i=1}^{m} p_i \int_{\boldsymbol{\theta} \in \Theta} L(\mathbf{x} \mid \boldsymbol{\theta}) \pi_i(\boldsymbol{\theta}) d\boldsymbol{\theta}}.$$

In fact, the results of the mixture posterior distribution (2.38) and(2.40) are used several times in Chapter 3; see for example Theorem 3.1 and Theorem 3.2.

Non-informative Priors. In many practical situations, no reliable prior information concerning $\boldsymbol{\theta}$ exists, and/or inference based solely on the data is desirable. In this case we typically wish to define a prior distribution $\pi(\boldsymbol{\theta})$ that contains no information about $\boldsymbol{\theta}$ in the sense that it does not favour one $\boldsymbol{\theta}$ value over another. We may refer to a distribution of this kind as a noninformative prior for $\boldsymbol{\theta}$ and argue that the information contained in the posterior about $\boldsymbol{\theta}$ stems from the data only.

Bayesian Computation

The use of Bayesian methods in applied problems has exploded during the 1990s due to the availability of fast computing machines combined with a family of iterative simulation methods known as Markov chain Monte Carlo (MCMC) algorithms. The main idea behind MCMC is to generate a Markov chain which has, as its unique limiting distribution, the posterior distribution of interest. It dates back to the seminal paper of Metropolis *et al.* (1953) although the computational power required was not available at that time. The original generation mechanism was generalised by Hastings (1970) in the so-called Metropolis-Hastings algorithm and has been widely used in Bayesian statistics after about 1990, see Gelfand and Smith (1990). However, Bayesian computation goes beyond the scope of this thesis and we shall not give more detail here.

Chapter 3

Bayesian Results for the Non-Central χ^2 Distribution

The block shrinkage and thresholding methods developed in the frequentist framework (see Cai, 1996, Hall *et al.* 1998, 1999), for estimating regression functions from noisy data by thresholding empirical wavelet coefficients in a block rather than individually, can increase the estimation accuracy of the wavelet coefficients. The aim of the research recorded in this chapter is to develop a parallel methodology in the Bayesian framework. Bayesian results for the non-central χ^2 distribution, which can be used as the theoretical basis for an empirical Bayes block (EBB) shrinkage method, are provided. A useful (and, to the best of our knowledge, newly discovered) identity satisfied by the non-central χ^2 density is exploited. When used with suitable families of prior distributions, it turns out that the EBB shrinkage approach proposed in this thesis (Chapter 3 and Chapter 4) combines a high degree of theoretical and computational tractability with high quality practical performance.

The outline of this chapter is as follows. §3.1 will introduce the definition of a non-central χ^2 distribution used in this chapter, and then discuss some relevant properties of the non-central χ^2 distribution which provide the bases for the EBB approach. §3.2-§3.5 will concentrate on developing Bayesian methodology for the non-centrality parameter of the non-central χ^2 distribution. Several properties of

this approach, which will help to fully understand this methodology, are discussed in § 3.6. Most of the proofs are given in § 3.7.

3.1 Introduction

3.1.1 Definition of the Non-Central χ^2 Distribution

The standard non-central χ^2 distribution can be defined as follows:

Definition 3.1 *Let $x_1, x_2, \ldots, x_m$ be a set of mutually independent normal variables with means $a_1, \ldots, a_m$ respectively and common variance 1. Then $\sum_{i=1}^{m} x_i^2$ has a non-central χ^2 distribution with m degrees of freedom and non-centrality parameter $a = \sum_{i=1}^{m} a_i^2$.*

From the definition of the standard non-central χ^2 distribution given above, we can derive the density of a rescaled non-central χ^2 distribution.

Definition 3.2 *The density of a non-central χ^2 distribution with m degrees of freedom, non-centrality a and scaled by $b > 0$ may be written*

$$\chi_m^2(y|a,b) = \sum_{k=0}^{\infty} e^{-a/(2b)} \frac{\{a/(2b)\}^k}{k!} \chi_{m+2k}^2(y|0.b) \qquad (y > 0) \tag{3.1}$$

where

$$\chi_{m+2k}^2(y|0,b) = \frac{1}{\Gamma\{(m/2)+k\}} \left(\frac{1}{2b}\right)^{(m/2)+k} y^{(m/2)+k-1} e^{-y/(2b)}$$

is the central χ^2 density on $m+2k$ degrees of freedom, also scaled by b.

Note that, since

$$\begin{aligned}
\int_0^\infty \chi_m^2(y|a,b)dy &= \int_0^\infty \sum_{k=0}^{\infty} e^{-a/(2b)} \frac{\{a/(2b)\}^k}{k!} \chi_{m+2k}^2(y|0,b)dy \\
&= \sum_{k=0}^{\infty} e^{-a/(2b)} \frac{\{a/(2b)\}^k}{k!} \int_0^\infty \chi_{m+2k}^2(y|0,b)dy \\
&= e^{-a/(2b)} \sum_{k=0}^{\infty} \frac{\{a/(2b)\}^k}{k!} \\
&= 1,
\end{aligned}$$

we can see that $\chi^2_m(y|a,b)$ is a probability density. Thus the distribution (3.1) may be interpreted as a Poisson mixture of scaled central χ^2 distributions; see e.g. Muirhead (1982, p. 23), Johnson and Kotz (1970, Chapter 28). We shall denote the standard non-central χ^2 distribution with m degrees of freedom and non-centrality parameter a by $\chi^2_m(a)$. Therefore the distribution with density (3.1) may be written as $b\chi^2_m(ab^{-1})$. Note that the distribution $b\chi^2_m(ab^{-1})$ and its density $\chi^2_m(y|a,b)$ are distinguished by the presence of the vertical bar in the latter.

3.1.2 Some Properties of the Non-Central χ^2 Distribution

The cumulative distribution function (CDF) of a non-central χ^2 distribution $b\chi^2_m(ab^{-1})$ (Johnson and Kotz, 1970, p.132) may be written

$$\begin{aligned} F_m(y|a,b) &= \int_0^y \chi^2_m(u|a,b)du \\ &= \sum_{k=0}^{\infty} e^{-a/(2b)} \frac{\{a/(2b)\}^k}{k!} \int_0^y \chi^2_{m+2k}(u|0,b)du \\ &= \sum_{k=0}^{\infty} e^{-a/(2b)} \frac{\{a/(2b)\}^k}{k!} F_{m+2k}(y|0,b) \end{aligned} \tag{3.2}$$

where $F_{m+2k}(y|0,b) = \int_0^y \chi^2_{m+2k}(u|0,b)du$ is the CDF of a central χ^2 distribution, which is given by

$$\begin{aligned} F_{m+2k}(y|0,b) &= \frac{1}{\Gamma\{(m/2)+k\}} \left(\frac{1}{2b}\right)^{(m/2)+k} \int_0^y u^{(m/2)+k-1} e^{-u/(2b)} du \\ &= \frac{1}{\Gamma\{(m/2)+k\}} \left(\frac{1}{2b}\right)^{(m/2)+k} y^{(m/2)+k} \int_0^1 t^{(m/2)+k-1} e^{-yt/(2b)} dt \\ &= \frac{1}{\Gamma\{(m/2)+k\}} \left(\frac{1}{2b}\right)^{(m/2)+k} y^{(m/2)+k} \mathcal{C}_{(m/2)+k-1}\left(\frac{y}{2b}\right), \end{aligned} \tag{3.3}$$

where

$$\mathcal{C}_h(\gamma) = \int_0^1 x^h e^{-\gamma x} dx \tag{3.4}$$

for $h > -1$ and all $\gamma \in \mathbf{R}$. Note that, when $\gamma > 0$, $\mathcal{C}_h(\gamma)$ is proportional to the complete gamma function $\Gamma(h+1)$ multiplied by an incomplete gamma function, and both are available as a standard functions in **Matlab**. Relevant properties of this function will be discussed in §3.3.

The moment generating function of a standard non-central χ^2 distribution $\chi^2_m(a)$ may be written

$$M(t) = (1-2t)^{-m/2}\exp\{at/(1-2t)\}. \tag{3.5}$$

Therefore, the cumulant generating function can be easily obtained as

$$K(t) = log\{M(t)\} = -\frac{m}{2}log(1-2t) + \frac{at}{1-2t}. \tag{3.6}$$

More generally, the moment generating function and the cumulant generating function of the scaled non-central χ^2 distribution with density $\chi^2_m(y|a,b)$ can be written directly as

$$M_{a,b}(t) = (1-2bt)^{-m/2}\exp\{at/(1-2bt)\} \tag{3.7}$$

and

$$K_{a,b}(t) = log\{M_{a,b}(t)\} = -\frac{m}{2}log(1-2bt) + \frac{at}{1-2bt}. \tag{3.8}$$

Since the nth moment μ_n is equal to the nth derivative of $M_{a,b}(t)$ evaluated at $t=0$, we write down the first and second moments

$$\mu_1 = M^{(1)}_{a,b}(0) = bm + a$$

and

$$\mu_2 = M^{(2)}_{a,b}(0) = bm^2 + 2abm + 4ab + 2b^2m + a^2.$$

The mean, variance and third cumulant of this distribution, which will be needed later, can also be obtained in the similar way:

$$E\{b\chi^2_m(ab^{-1})\}(=\kappa_1) = K^{(1)}_{a,b}(0) = bm + a, \tag{3.9}$$

$$Var\{b\chi^2_m(ab^{-1})\}(=\kappa_2) = K^{(2)}_{a,b}(0) = 2mb^2 + 4ab \tag{3.10}$$

and

$$\kappa_3 = K^{(3)}_{a,b}(0) = 8mb^3 + 24ab^2. \tag{3.11}$$

3.2 The Non-Central χ^2 Distribution: Bayesian Results

The approach developed in this chapter and used later is based on Theorem 3.1 and Theorem 3.2 below. Theorem 3.1 tells us that if the non-centrality parameter of a non-central χ^2 distribution has a prior which is a scale mixture of central χ^2 distributions with the appropriate degrees of freedom, then the posterior distribution of the non-centrality parameter is a certain mixture of non-central χ^2 distributions, where the mixture distribution can be readily identified. Theorem 3.2 gives explicit expressions for relevant posterior quantities when the mixture distribution in the prior has the structure given by (3.20) and (3.21).

Theorem 3.1 *Suppose that the likelihood is given by* $f(z|\rho,\sigma^2) = \chi^2_m(z|\rho,\sigma^2)$ *where the prior distribution of* ρ *given* β *has a density* $f(\rho|\beta) = \chi^2_m(\rho|0,\beta^{-1})$, *and* β *given hyperparameters* σ^2 *and* θ *has CDF* $F(\beta|\sigma^2,\theta)$ *with support* $(0,\infty]$. *Write*

$$u \equiv u(\beta) = \frac{\beta\sigma^2}{1+\beta\sigma^2} \tag{3.12}$$

where σ^2 *is treated as a constant. Then*

$$f(\rho|z,\sigma^2,\theta,\beta) = \chi^2_m\{\rho|z(1-u)^2,\sigma^2(1-u)\}, \tag{3.13}$$

$$dF(\beta|z,\sigma^2,\theta) = \frac{\chi^2_m(z|0,\sigma^2/u)dF(\beta|\sigma^2,\theta)}{\int_{\beta\in(0,\infty]}\chi^2_m(z|0,\sigma^2/u)dF(\beta|\sigma^2,\theta)}, \tag{3.14}$$

and therefore the posterior density of ρ *with* β *"integrated out" is given by*

$$f(\rho|z,\sigma^2,\theta) = \int_{\beta\in(0,\infty]}\chi^2_m\{\rho|z(1-u)^2,\sigma^2(1-u)\}dF(\beta|z,\sigma^2,\theta), \tag{3.15}$$

where $u = u(\beta)$ *is defined in (3.12) and* $dF(\beta|z,\sigma^2,\theta)$ *is given in (3.14).*

Thus Theorem 3.1 states that the posterior distribution of ρ is a β-mixture of scaled non-central χ^2 distributions, where the mixture distribution is given in (3.14), for the discussion of mixture distributions, see Appendix A. Before we prove Theorem 3.1, a key identity will be given in the following Lemma.

Lemma 3.1 *The non-central χ^2 density in (3.1) satisfies the following identity:*

$$\chi^2_m(z|\rho,\sigma^2)\chi^2_m(\rho|0,\beta^{-1}) = \chi^2_m\{\rho|z(1-u)^2,\sigma^2(1-u)\}\chi^2_m(z|0,\sigma^2/u). \tag{3.16}$$

where $u = u(\beta)$ is given in (3.12),

A proof is given in §3.7. To the best of our knowledge, identity (3.16) is new.

Proof of Theorem 3.1

We have

$$f(z|\rho,\sigma^2)f(\rho|\beta) = f(\rho,z|\sigma^2,\beta)$$

and

$$f(\rho|z,\sigma^2,\beta)f(z|\sigma^2,\beta) = f(\rho,z|\sigma^2,\beta).$$

By assumption

$$f(z|\rho,\sigma^2) = \chi^2_m(z|\rho,\sigma^2)$$

and

$$f(\rho|\beta) = \chi^2_m(\rho|0,\beta^{-1}),$$

so we can deduce from Lemma 3.1 that

$$f(\rho|z,\sigma^2,\beta) = \chi^2_m\{\rho|z(1-u)^2,\sigma^2(1-u)\},$$

so that (3.13) holds and

$$f(z|\sigma^2,\beta) = \chi^2_m(z|0,\sigma^2/u). \tag{3.17}$$

Moreover

$$dF(\beta|z,\sigma^2,\theta) = \frac{f(z|\sigma^2,\beta)dF(\beta|\sigma^2,\theta)}{\int_{\beta\in(0,\infty]} f(z|\sigma^2,\beta)dF(\beta|\sigma^2,\theta)}, \tag{3.18}$$

so (3.14) follows after substituting (3.17) on the right hand side of (3.18). Finally,

$$f(\rho|z,\sigma^2,\theta) = \int_{\beta\in(0,\infty]} f(\rho|z,\sigma^2,\beta)dF(\beta|z,\sigma^2,\theta), \tag{3.19}$$

so (3.15) follows after substituting (3.13) on the right hand side of (3.19). ■

We now focus on the following mixture structure for $F(\beta|\sigma^2,\theta)$:

$$F(\beta|\sigma^2,\theta=(p,\lambda)) = pF(\beta|\sigma^2,\lambda,J=1) + (1-p)F(\beta|\sigma^2,\lambda,J=0) \tag{3.20}$$

where, in all cases,

$$F(\beta|\sigma^2, \lambda, J = 1) = I_{\{\beta=\infty\}}(\beta), \tag{3.21}$$

and

$$I_{\{\beta=\infty\}}(\beta) = \begin{cases} 1 & \beta = \infty \\ 0 & \beta < \infty \end{cases}$$

is the indicator function; J is a Bernoulli random variable with $J = 1$ corresponding to a unit mass point at $\beta = \infty$ and $J = 0$ corresponding to some other distribution $F(\beta|\sigma^2, \lambda, J = 0)$ to be specified; and $\boldsymbol{\theta} = (p, \lambda)$ where λ is a hyperparameter in the distribution $F(\beta|\sigma^2, \lambda, J = 0)$, and p is the prior probability that $J = 1$. We also define $\delta(\xi)$, the Dirac delta function, by

$$\int_{\mathbf{A}} \delta(\xi)d\xi = \begin{cases} 1 & \text{if } \xi \in \mathbf{A} \\ 0 & \text{otherwise} \end{cases}$$

for $\mathbf{A} \subseteq \mathbf{R}$.

Remark. If we reparameterise β as $\beta' = 1/\beta$, then β' has support $[0, \infty)$ and the use of $\beta = \infty$ is avoided. This alternative approach, which leads to identical results, may be preferred by many readers, as considerable care is needed when working with infinity. However, we stress that no problems arose with the use of infinity in the present context, essentially because all expectations that are considered are finite, due to the fact that when $\beta = \infty$, $u(\beta) = 1$.

The mixture structure (3.20) is designed to capture the sparseness common to most wavelet applications, which is, the majority of the wavelet coefficients are negligible (i.e. close to 0) and the remaining few large coefficients determine most of the function. The term $F(\beta|\sigma^2, \lambda, J = 1)$ produces a point mass at $\rho = 0$ whereas the $F(\beta|\sigma^2, \lambda, J = 0)$ is spread out to accommodate the possibility of larger coefficients. We shall consider some particular choices for $F(\beta|\sigma^2, \lambda, J = 0)$ below.

Theorem 3.2 *If $F(\beta|\sigma^2, \theta)$ has the form indicated in (3.20) and (3.21) then the posterior density for ρ is given by*

$$f(\rho|z, \sigma^2, \theta) = \pi\delta_{\{\rho=0\}}(\rho) + (1 - \pi)f(\rho|z, \sigma^2, \lambda, J = 0) \tag{3.22}$$

where π, the posterior probability of a unit mass point at $\rho = 0$, is given by

$$\pi = \frac{p\chi^2_m(z|0,\sigma^2)}{f(z|\sigma^2,\theta)}; \tag{3.23}$$

$f(z|\sigma^2,\theta)$, the marginal density of z, is given by

$$f(z|\sigma^2,\theta) = p\chi^2_m(z|0,\sigma^2) + (1-p)\int_{\beta\in(0,\infty)} \chi^2_m(z|0,\sigma^2/u)dF(\beta|\sigma^2,\lambda,J=0); \tag{3.24}$$

and

$$f(\rho|z,\sigma^2,\lambda,J=0) = \int_{\beta\in(0,\infty)} \chi^2_m\{\rho|z(1-u)^2,\sigma^2(1-u)\}dF(\beta|z,\sigma^2,\lambda,J=0) \tag{3.25}$$

where

$$dF(\beta|z,\sigma^2,\lambda,J=0) = \frac{\chi^2_m(z|0,\sigma^2/u)dF(\beta|\sigma^2,\lambda,J=0)}{\int_{\beta\in(0,\infty)}\chi^2_m(z|0,\sigma^2/u)dF(\beta|\sigma^2,\lambda,J=0)}. \tag{3.26}$$

The posterior expectation of ρ is given by

$$E[\rho|z,\sigma^2,\theta] = (1-\pi)\int_{\beta\in(0,\infty)}\{m\sigma^2(1-u)+z(1-u)^2\}dF(\beta|z,\sigma^2,\lambda,J=0). \tag{3.27}$$

Proof of Theorem 3.2

Given the specific form of $F(\beta|\sigma^2,\theta)$ in (3.20), we can write down the prior distribution as

$$\begin{aligned} f(\rho|\sigma^2,\theta) &= \int_{\beta\in(0,\infty]} f(\rho|\beta,\sigma^2)dF(\beta|\sigma^2,\theta) \\ &= p\delta_{\{\rho=0\}}(\rho) + (1-p)\int_{\beta\in(0,\infty)} f(\rho|\beta,\sigma^2)dF(\beta|\sigma^2,\lambda,J=0). \end{aligned}$$

The marginal density of z is given by

$$\begin{aligned} f(z|\sigma^2,\theta) &= \int_{\beta\in(0,\infty]} f(z|\sigma^2,\beta)dF(\beta|\sigma^2,\theta) \\ &= \int_0^\infty pf(z|\sigma^2,\beta)dI_{\{\beta=\infty\}}(\beta) \\ &\quad + \int_0^\infty (1-p)f(z|\sigma^2,\beta)dF(\beta|\sigma^2,\lambda,J=0) \\ &= p\chi^2_m(z|0,\sigma^2) + (1-p)\int_{\beta\in(0,\infty)} \chi^2_m(z|0,\sigma^2/u)dF(\beta|\sigma^2,\lambda,J=0). \end{aligned}$$

The quantities (3.25) and (3.26) are obtained similarly using the mixture structure of $F(\beta|\sigma^2,\theta)$ in (3.20). For the posterior expectation of ρ, $E[\rho|z,\sigma^2,\theta]$, we use Fubini's theorem (see Kingman and Taylor, 1966, p. 147) to obtain

$$\begin{aligned} E[\rho|z,\sigma^2,\theta] &= \int_0^\infty \rho(1-\pi)f(\rho|z,\lambda,J=0)d\rho \\ &= (1-\pi)\int_0^\infty \rho\Big\{\int_{\beta\in(0,\infty)} \chi^2_m\{\rho|z(1-u)^2,\sigma^2(1-u)\}dF(\beta|z,\sigma^2,\lambda,J=0)\Big\}d\rho \\ &= (1-\pi)\int_{\beta\in(0,\infty)} \Big\{\int_0^\infty \rho\chi^2_m\{\rho|z(1-u)^2,\sigma^2(1-u)\}d\rho\Big\}dF(\beta|z,\sigma^2,\lambda,J=0) \\ &= (1-\pi)\int_{\beta\in(0,\infty)} \{m\sigma^2(1-u)+z(1-u)^2\}dF(\beta|z,\sigma^2,\lambda,J=0) \end{aligned}$$

as required. ■

3.3 Some Useful Priors

We now consider three cases of the prior (3.20), which we refer to as "mass point", "exponential" and "power" prior, assuming that (3.21) holds in each case. We use the subscripts M, E and P, respectively, to denote these priors. One motivation for considering these priors is that they lead to tractable calculation of posterior quantities. However, as we shall see later, these priors, especially E and P, have other attractive theoretical and practical properties. Each posterior quantity given below may be easily derived using Theorem 3.2. The proofs of all lemmas in this section will be given in §3.7.

Before continuing, a useful lemma is stated.

Lemma 3.2 *Define*

$$C_h(\gamma) = \int_0^1 x^h e^{-\gamma x}dx$$

for $h > -1$ *and all* $\gamma \in \mathbf{R}$, *then*

$$C_h(0) = \frac{1}{h+1}, \tag{3.28}$$

$$C_h(\gamma) \sim \frac{1}{\gamma^{h+1}}\Gamma(h+1) \quad when \ \gamma \to \infty, \tag{3.29}$$

and

$$\frac{C_{h+t}(\gamma)}{C_h(\gamma)} \sim \gamma^{-t}\frac{\Gamma(h+t+1)}{\Gamma(h+1)} \quad when \ \gamma \to \infty. \tag{3.30}$$

3.3.1 Mass Point Prior

In this case we choose $F(\beta|\sigma^2, \lambda, J=0)$ to be

$$F(\beta|\sigma^2, \lambda, J=0) = F_M(\beta|\lambda) = I_{\{\beta \geq \lambda\}}(\beta). \tag{3.31}$$

With straightforward calculations based on Theorem 3.2 we obtain

$$f_M(z|\sigma^2, \theta) = p\chi^2_m(z|0, \sigma^2) + (1-p)\chi^2_m(z|0, \sigma^2/u_\lambda). \tag{3.32}$$

a mixture density of two central χ^2 variables. The posterior probability of the first distribution on the right hand side of (3.32) is

$$\begin{aligned} \pi_M = \pi_M(z) &= \frac{p\chi^2_m(z|0,\sigma^2)}{f_M(z|\sigma^2,\theta)} \\ &= \frac{1}{1+\frac{1-p}{p}R_M}, \end{aligned} \tag{3.33}$$

where

$$R_M = R_M(z) = u_\lambda^{m/2}\exp\left\{\frac{z}{2\sigma^2(1+\lambda\sigma^2)}\right\}.$$

Also,

$$E_M[\rho|z, \sigma^2, \theta] = (1-\pi_M)\{m\sigma^2(1-u_\lambda) + z(1-u_\lambda)^2\} \tag{3.34}$$

where $u_\lambda = u(\lambda) = \lambda\sigma^2/(1+\lambda\sigma^2)$.

As $z \to \infty$, $R_M \to \infty$ and $\pi_M \to 0$, so

$$E_M[\rho|z, \sigma^2, \theta] = z(1-u_\lambda)^2 + m\sigma^2(1-u_\lambda),$$

while as $z \to 0$, $R_M(0) \to u_\lambda^{m/2}$ and

$$\pi_M(0) \to \frac{1}{1+\frac{1-p}{p}R_M(0)},$$

so

$$E_M[\rho|z,\sigma^2,\theta] = \{1-\pi_M(0)\}m\sigma^2(1-u_\lambda). \tag{3.35}$$

3.3.2 Power Prior

The CDF of the "power" prior is defined by

$$F(\beta|\sigma^2,\lambda,J=0) = F_P(\beta|\sigma^2,\lambda) = \left(\frac{\beta\sigma^2}{1+\beta\sigma^2}\right)^{\lambda+1},$$

and the corresponding density is given by

$$f_P(\beta|\sigma^2,\lambda) = (\lambda+1)\frac{\sigma^2}{(1+\beta\sigma^2)^2}\left(\frac{\beta\sigma^2}{1+\beta\sigma^2}\right)^{\lambda}. \tag{3.36}$$

Then from

$$\begin{aligned} f_P(z|\sigma^2,\lambda,J=0) &= \int_{\beta\in(0,\infty)} \chi^2_m(z|0,\sigma^2/u) f_P(\beta|\sigma^2,\lambda)d\beta \\ &= \int_0^1 \chi^2_m(z|0,\sigma^2/u) g_P(u|\sigma^2,\lambda)du, \end{aligned} \tag{3.37}$$

where $g_P(u|\sigma^2,\lambda) = (\lambda+1)u^\lambda$, we have

$$\begin{aligned} f_P(z|\sigma^2,\lambda,J=0) &= \frac{(\lambda+1)z^{(m/2)-1}}{\Gamma(m/2)}\left(\frac{1}{2\sigma^2}\right)^{m/2}\int_0^1 u^{(m/2)+\lambda}\exp\left(-\frac{zu}{2\sigma^2}\right)du \\ &= \frac{(\lambda+1)z^{(m/2)-1}}{\Gamma(m/2)}\left(\frac{1}{2\sigma^2}\right)^{m/2} C_\eta\{z/(2\sigma^2)\}, \end{aligned} \tag{3.38}$$

where $\eta = m/2+\lambda$ and C_η is defined in (3.4). Furthermore, the marginal distribution is

$$f_P(z|\sigma^2,\theta) = p\chi^2_m(z|0,\sigma^2) + (1-p)\frac{(\lambda+1)z^{(m/2)-1}}{\Gamma(m/2)}\left(\frac{1}{2\sigma^2}\right)^{m/2} C_\eta\{z/(2\sigma^2)\}. \tag{3.39}$$

The posterior probability of the unit mass point at zero will be

$$\begin{aligned} \pi_P = \pi_P(z) &= \frac{p\chi^2_m(z|0,\sigma^2)}{f_P(z|\sigma^2,\theta)} \\ &= \frac{1}{1+\frac{1-p}{p}R_P}, \end{aligned} \tag{3.40}$$

where

$$R_P = R_P(z) = (\lambda+1)e^{z/(2\sigma^2)}C_\eta\{z/(2\sigma^2)\}, \tag{3.41}$$

and

$$f(\beta|z,\sigma^2,\lambda,J=0)=\frac{1}{C_\eta\{z/(2\sigma^2)\}}\frac{\sigma^2}{(1+\beta\sigma^2)^2}u^{(m/2)+\lambda}\exp\left\{-uz/(2\sigma^2)\right\}. \quad (3.42)$$

The following results provide more information about the posterior mean in the case of the power prior.

Lemma 3.3 *Following the definition of the CDF of the "power" prior and the mixture marginal distribution given in (3.39), the expectation of posterior distribution will be*

$$E_P[\rho|z,\sigma^2,\theta]=(1-\pi_P)\Big[m\sigma^2+z-(m\sigma^2+2z)\mathcal{A}_{\eta,1}\{z/(2\sigma^2)\}+z\mathcal{A}_{\eta,2}\{z/(2\sigma^2)\}\Big] \quad (3.43)$$

where $\mathcal{A}_{\eta,j}\{z/2\sigma^2)\}=C_{\eta+j}\{z/(2\sigma^2)\}/C_\eta\{z/(2\sigma^2)\}$ $(j=1,2)$ *and* $C_\eta\{z/(2\sigma^2)\}$ *is defined in (3.4).*

Lemma 3.4 *For fixed* σ^2 *and* θ,

$$E_P[\rho|z,\sigma^2,\theta]=z-(m+4\lambda+4)\sigma^2+O(z^{-1}) \quad as \ z\to\infty, \quad (3.44)$$

and

$$E_P[\rho|z,\sigma^2,\theta]=\big(1-\pi_P(0)\big)\frac{m\sigma^2}{\eta+2} \quad as \ z\to 0, \quad (3.45)$$

where

$$\pi_P(0)=\frac{1}{1+\left(\frac{1-p}{p}\right)\left(\frac{\lambda+1}{\eta+1}\right)}.$$

3.3.3 Exponential Prior

The CDF of the "exponential" prior is defined by

$$F(\beta|\sigma^2,\lambda,J=0)=F_E(\beta|\sigma^2,\lambda)=\frac{1-\exp\left(-\lambda\frac{\beta\sigma^2}{1+\beta\sigma^2}\right)}{1-\exp(-\lambda)}$$

and the corresponding density is

$$f_E(\beta|\sigma^2,\lambda)=C_0(\lambda)^{-1}\frac{\sigma^2}{(1+\beta\sigma^2)^2}\exp\left(-\lambda\frac{\beta\sigma^2}{1+\beta\sigma^2}\right). \quad (3.46)$$

The results for the exponential prior can be obtained in similar fashion as those for the power prior. We have

$$f_E(z|\sigma^2,\theta) = p\chi^2_m(z|0,\sigma^2) + (1-p)\frac{z^{(m/2)-1}}{\Gamma(m/2)}\left(\frac{1}{2\sigma^2}\right)^{m/2}\frac{\mathcal{C}_{m/2}(\xi)}{\mathcal{C}_0(\lambda)}, \tag{3.47}$$

$$\begin{aligned}\pi_E = \pi_E(z) &= \frac{p\chi^2_m(z|0,\sigma^2)}{f_E(z|\sigma^2,\theta)} \\ &= \frac{1}{1+\frac{1-p}{p}R_E},\end{aligned} \tag{3.48}$$

where $\xi = \lambda + z/(2\sigma^2)$,

$$R_E = R_E(z) = \frac{e^{z/(2\sigma^2)}\mathcal{C}_{m/2}(\xi)}{\mathcal{C}_0(\lambda)}$$

and

$$f(\beta|z,\sigma^2,\lambda,J=0) = \frac{1}{\mathcal{C}_{m/2}(\xi)}\frac{\sigma^2}{(1+\beta\sigma^2)^2}u^{m/2}\exp\{-\xi u\}. \tag{3.49}$$

Lemma 3.5 *Following the definition of the CDF of the "exponential" prior and the mixture marginal distribution given in (3.47), the expectation of posterior distribution will be*

$$E_E[\rho|z,\sigma^2,\theta] = (1-\pi_E)\{m\sigma^2 + z - (m\sigma^2+2z)\mathcal{A}_{m/2,1}(\xi) + z\mathcal{A}_{m/2,2}(\xi)\}. \tag{3.50}$$

For fixed σ^2 and θ,

$$E_E[\rho|z,\sigma^2,\theta] = z - (m+4)\sigma^2 + O(z^{-1}) \quad as\ z\to\infty \tag{3.51}$$

and

$$E_E[\rho|z,\sigma^2,\theta] = \{1-\pi_E(0)\}\left\{1-\frac{\mathcal{C}_{m/2+1}(\lambda)}{\mathcal{C}_{m/2}(\lambda)}\right\}m\sigma^2 \quad as\ z\to 0, \tag{3.52}$$

where

$$\pi_E(0) = \frac{1}{1+\left(\frac{1-p}{p}\right)\left\{\frac{\mathcal{C}_{m/2}(\lambda)}{\mathcal{C}_0(\lambda)}\right\}}.$$

Proof of Lemma 3.5 is similar to the proofs of Lemma 3.3 and Lemma 3.4 and is omitted.

Following the results (3.35) of § 3.3.1, (3.45) of Lemma 3.4 and (3.52) of Lemma 3.5, we know that the posterior means of the three priors fail to be either a strict shrinkage rule or a thresholding rule (see § 2.5.1). However, according to Johnstone and Silverman (2005), a definition of the bounded shrinkage property is provided as follows:

Definition 3.3 *The family has the bounded shrinkage property if for some constant* b

$$|x - \eta(x,t)| \leq t + b \quad \textit{for all } x \textit{ and } t, \tag{3.53}$$

where $\eta(x,t)$ *is the shrinkage rule.*

The posterior means of the three priors satisfy the bounded shrinkage property. It turns out that these posterior means will closely approximate strict shrinkage and thresholding rules for suitable choices of hyperparameters.

Although the exponential prior is qualitatively similar to the power prior and has performed well in numerical examples, it has generally performed less well than the power prior. For this reason we focus mainly on the power prior to investigate properties of the posterior median in the following sections.

3.3.4 A General Discrete Prior

In some circumstances we may wish to use a prior which is different from the mass point, power or exponential priors discussed above. For a general prior the computations can be performed using numerical integration or MCMC. However, if we are willing to approximate a general prior using a discrete distribution concentrated on a finite set of β values, then the necessary calculations are similar to those given in (3.32).

3.4 Posterior Quantities of Interest

The posterior distribution of ρ is given in Theorem 3.2; see (3.22), (3.23), (3.25) and (3.26). Here, we shall be interested in the following characteristics of the posterior distribution (3.22): ρ_{mean}, the posterior mean; the "hypothesis testing" location estimate $\rho_{hyp} = zI(\pi < 1/2)$ where $I(\pi < 1/2) = 1$ if $\pi < 1/2$ and $I(\pi < 1/2) = 0$ if $\pi \geq 1/2$; and ρ_{med}, the posterior median. In this section we shall assume that the hyperparameters σ^2, p and λ are known. Estimation of the hyperparameters σ^2, p and λ will be discussed in § 4.3.

In the framework of Theorem 3.2, ρ_{mean} and ρ_{hyp} are straightforward to calculate; see (3.27) for ρ_{mean} and (3.23) for ρ_{hyp}. For the three priors M, P and E discussed in the previous subsection, ρ_{mean} is given by (3.34), (3.43) and (3.50), and ρ_{hyp} is obtained using (3.33), (3.40) and (3.48), respectively.

We focus now on the posterior median. From the posterior distribution (3.22), it follows that if $\pi \geq 1/2$ then $\rho_{med} = 0$, while if $\pi < 1/2$ it is necessary to find the α-quantile, with $\alpha = (1/2-\pi)/(1-\pi)$, of the distribution with density $f(\rho|z, \sigma^2, \lambda, J = 0)$ given in (3.25). In the case of the mass point prior this involves finding the α-quantile of the $\sigma^2(1-u_\lambda)\chi^2_m\{z(1-u_\lambda)/\sigma^2\}$ distribution. In the exponential and power prior cases, this involves finding the α-quantile of the CDF

$$F(y) = \int_{u\in(0,1)} H(y|z, \sigma^2, u)g(u)du \tag{3.54}$$

where

$$H(y|z, \sigma^2, u) = Pr[\sigma^2(1-u)\chi^2_m\{z(1-u)/\sigma^2\} \leq y]. \tag{3.55}$$

In the exponential prior case,

$$g(u) = g_E(u) = C_{m/2}\{\lambda + z/(2\sigma^2)\}^{-1}u^{m/2}\exp[-\{\lambda + z/(2\sigma^2)\}u]; \tag{3.56}$$

and in the power prior case,

$$g(u) = g_P(u) = C_{\lambda+m/2}\{z/(2\sigma^2)\}^{-1}u^{\lambda+m/2}\exp\{-uz/(2\sigma^2)\}. \tag{3.57}$$

Note that in (3.54) the variable of integration has been changed from β to $u = \beta\sigma^2/(1+\beta\sigma^2)$, with σ^2 treated as a constant.

$H(y|z, \sigma^2, u)$ is the CDF of a non-central χ^2 distribution. Recall the theoretical expression of the CDF of a non-central χ^2 distribution we gave in (3.2), which can be coded directly using the algorithm given by Narula and Desu (1981). Another method, using the Lugannani-Rice (LR) saddlepoint formula (Lugannani and Rice, 1980) to approximate the real CDF of a non-central χ^2 distribution, is also very appealing and accurate. Since the algorithm by Narula and Desu involves summing a series which in some cases will have a large number of terms while LR formula has explicit form, we choose to use LR formula to approximate the real CDF of a non-central χ^2 distribution in Chapter 4. However, it would be interesting to compare the results provided in Chapter 4 with the results using the algorithm given by Narula and Desu.

3.5 Theoretical Properties of the Posterior Median

In this section, two properties of the median of the posterior distribution (3.22) will be investigated. Here, we focus exclusively on the posterior median, denoted ρ_{med}, given the power prior (3.36).

3.5.1 Asymptotic Behaviour of the Posterior Median

Recall from Lemma 3.4 and Lemma 3.5 in Section 3.3 that the posterior means of the proposed power prior and exponential prior have the bounded shrinkage property. We want to look at the asymptotic behaviour of the posterior median from a similar perspective. In this subsection, $med(\rho_z)$ instead of ρ_{med} will be used to highlight the dependence of ρ and the posterior median on z. The following proposition describes the behaviour of $med(\rho_z)$ as $z \to \infty$.

Proposition 3.1 *In the setting of Lemma 3.4, the median of the posterior distri-*

bution of ρ satisfies

$$med(\rho_z) = z - \sigma^2(5 + 4\lambda + m) + O(z^{-1/2}) \qquad as \;\; z \to \infty. \tag{3.58}$$

A proof will be given in §3.7. Note that, as $z \to \infty$, $med(\rho_z)$ is smaller than the posterior mean given in (3.44) by a fixed quantity, σ^2.

3.5.2 Shrinkage and Thresholding Properties

A natural and important question to ask, particularly in the wavelet context, is whether a given estimation rule, such as the posterior mean or posterior median, has shrinkage and/or thresholding properties. In the case of the posterior median with the power prior, we find a necessary and sufficient condition for it to be a thresholding rule and a necessary and sufficient condition for it to be a shrinkage rule.

Proposition 3.2 *Let $med(\rho_z)$ denote the posterior median with power prior (3.36) and integer $m \geq 1$, $p \in (0,1)$, $\lambda > -1$ and $\sigma^2 > 0$.*

A: *$med(\rho_z)$ is a thresholding rule, in the sense that $med(\rho_z) = 0$ for all sufficiently small positive z, if and only if*

$$p > \frac{2(\lambda+1)}{4(\lambda+1)+m}; \tag{3.59}$$

B: *$med(\rho_z)$ is a shrinkage rule, in the sense that $med(\rho_z) \leq z$ for all $z \geq 0$, if and only if*

$$p \geq \frac{M_m(\sigma^2,\lambda)}{1 + M_m(\sigma^2,\lambda)}, \tag{3.60}$$

where

$$M_m(\sigma^2,\lambda) = \sup_{z \geq 0} \Big[R_P(z)\{1 - 2F(z)\} \Big], \tag{3.61}$$

$R_P(z)$ is defined in (3.41) and $F(z)$ is defined in (3.54).

Proof of Proposition 3.2

Part A: First, note that $R_P = (\lambda+1)e^{z/(2\sigma^2)}\mathcal{C}_\eta\{z/(2\sigma^2)\}$ is a strictly increasing function of z because, using the definition of $\mathcal{C}_\eta$ in (3.4), we have

$$\begin{aligned}\frac{\partial}{\partial z}\Big[(\lambda+1)e^{z/(2\sigma^2)}\mathcal{C}_\eta\{z/(2\sigma^2)\}\Big] &= \frac{\partial}{\partial z}(\lambda+1)\int_0^1 e^{z/(2\sigma^2)}y^\eta e^{-zy/(2\sigma^2)}dy \\ &= (\lambda+1)\frac{\partial}{\partial z}\int_0^1 y^\eta e^{(1-y)z/(2\sigma^2)}dy \\ &= (\lambda+1)\int_0^1 y^\eta\frac{(1-y)}{2\sigma^2}e^{(1-y)z/(2\sigma^2)}dy \\ &> 0,\end{aligned}$$

for fixed λ, η and σ^2. Therefore π_P, defined in (3.40), is a strictly decreasing function of z.

Consequently, the posterior median is a thresholding rule if and only if $\pi_P > 1/2$ when $z = 0$. When $z = 0$,

$$R_P = (\lambda+1)e^0\mathcal{C}_\eta(0) = \frac{\lambda+1}{\eta+1}$$

and

$$\pi_P = \frac{p}{p+(1-p)R_P} = \frac{p}{p+\frac{(1+p)(\lambda+1)}{\eta+1}}.$$

Therefore, the ratio $\pi_P > 1/2$ at $z = 0$ is equivalent to

$$p > \frac{\lambda+1}{\lambda+\eta+2} = \frac{\lambda+1}{2+2\lambda+m/2}$$

as required, since $\eta = \lambda + m/2$.

Part B: $F(z)$, defined in (3.54), is the CDF corresponding to the posterior density (3.25) with the power prior. A necessary and sufficient condition for $med(\rho_z) \leq z$ for all $z \geq 0$ is

$$\pi_P + (1-\pi_P)F(z) \geq \frac{1}{2}, \quad \text{for all } z \geq 0, \tag{3.62}$$

which is equivalent to

$$(1-\pi_P)^{-1} \geq 2\{1-F(z)\}, \quad \text{for all } z \geq 0. \tag{3.63}$$

From the definition of π_P in (3.40), it is seen that

$$(1-\pi_P)^{-1} = 1 + \frac{p}{(1-p)R_P(z)}.$$

Therefore (3.63) is equivalent to

$$\begin{aligned} & 1 + \frac{p}{(1-p)R_P(z)} \geq 2 - 2F(z) \\ \Leftrightarrow \quad & \frac{p}{1-p} \geq R_P(z)\{1-2F(z)\} \\ \Leftrightarrow \quad & p \geq \sup_{z\geq 0}\left[\frac{R_P(z)\{1-2F(z)\}}{1+R_P(z)\{1-2F(z)\}}\right], \end{aligned}$$

for all $z \geq 0$. Since $f(x) = x/(1+x)$ is an increasing function of x for $x \geq 0$, from which it follows that

$$\sup_{z\geq 0}\left[\frac{R_P(z)\{1-2F(z)\}}{1+R_P(z)\{1-2F(z)\}}\right] = \frac{M_m(\sigma^2,\lambda)}{1+M_m(\sigma^2,\lambda)}$$

where $M_m(\sigma^2,\lambda)$ is defined in (3.61). This completes the proof. ■

Conditions (3.59) and (3.60) have a clear and sensible interpretation. Note that small p corresponds to a prior belief in a non-sparse wavelet representation of the unknown function f. Since a key notion underlying wavelet methods is that "most" unknown function can be well approximated by a function with a relatively small proportion of nonzero wavelet coefficients, in case of sufficiently small p, it is far from clear that thresolding and shrinkage are the appropriate things to do, particularly when prior information is presented. Also, the conditions needed for Proposition 3.2 were satisfied in all the numerical examples we have considered in Chapter 4. Similar necessary and sufficient conditions can be derived for the mass prior and exponential prior.

It is worth noting that $R_P\{1-2F(z)\}$, considered as a function of z with m, λ and σ^2 held fixed, is decreasing for all sufficiently small positive z. It is summarised as follows:

Lemma 3.6 *In the setting of Proposition 3.2, for fixed values of m, λ and σ^2, $W_m(z) = R_P\{1-F(z)\}$ is a decreasing function of z for all sufficiently small positive z.*

We will give the proof of this lemma in § 3.7. Also, note that as a consequence of Proposition 3.1, $F(z) > 1/2$ for z sufficiently large, and so $R_P\{1-2F(z)\}$ is negative for z sufficiently large. Moreover, Figure 3.1 shows $R_P\{1-2F(z)\}$, denoted as W_m for short, as a function of z with numerous choices of integer $m \geq 1$ and $\lambda > -1$. In each case, it seems that the maximum occurred at $z = 0$.

Conjecture 3.1 *Based on the above discussion, in particular Lemma 3.7 and Figure 3.1, it seems reasonable to conjecture that the supremum always occurs at $z = 0$, which means that $M_m(\sigma^2, \lambda)$ is equal to $2(\lambda+1)/\{4(\lambda+1)+m\}$, the right-hand side of (3.59).*

3.6 Discussion and Further Work

To facilitate the presentation, we denote the methods corresponding to the three priors (3.31), (3.46) and (3.36) by NCM, NCE and NCP respectively, where the "NC" indicates "based on the non-central chi-squared results in §3.2" and M, E and P indicate "mass point prior", "exponential prior" and "power prior" respectively (see §3.3). In addition, "mean", "hyp" and "med" (the posterior mean, hypothesis testing and posterior median methods, respectively, described in §3.3 and §3.4) will be combined with NCM, NCN and NCP to indicate the denoising method used.

Theoretical motivation for priors. There is an important theoretical reason for preferring the power and exponential priors over the mass point prior. Figure 3.2 shows the Bayes rule for NCMmean, NCPmean and NCEmean, which correspond to using ρ_{mean} with the mass point prior, power prior and exponential prior, respectively. For purposes of comparison, the exact risk of each rule is shown in Figure 3.3. For more discussion of the exact risk, see Marron *et al.*. We can see that the risk

$$R(\rho, \mathcal{S}) = E^{z|\rho}\{\rho - \mathcal{S}(z)\}^2,$$

where $\mathcal{S}(z)$ is a shrinkage operator, stays bounded as $z \to \infty$ when $\mathcal{S}$ is given by

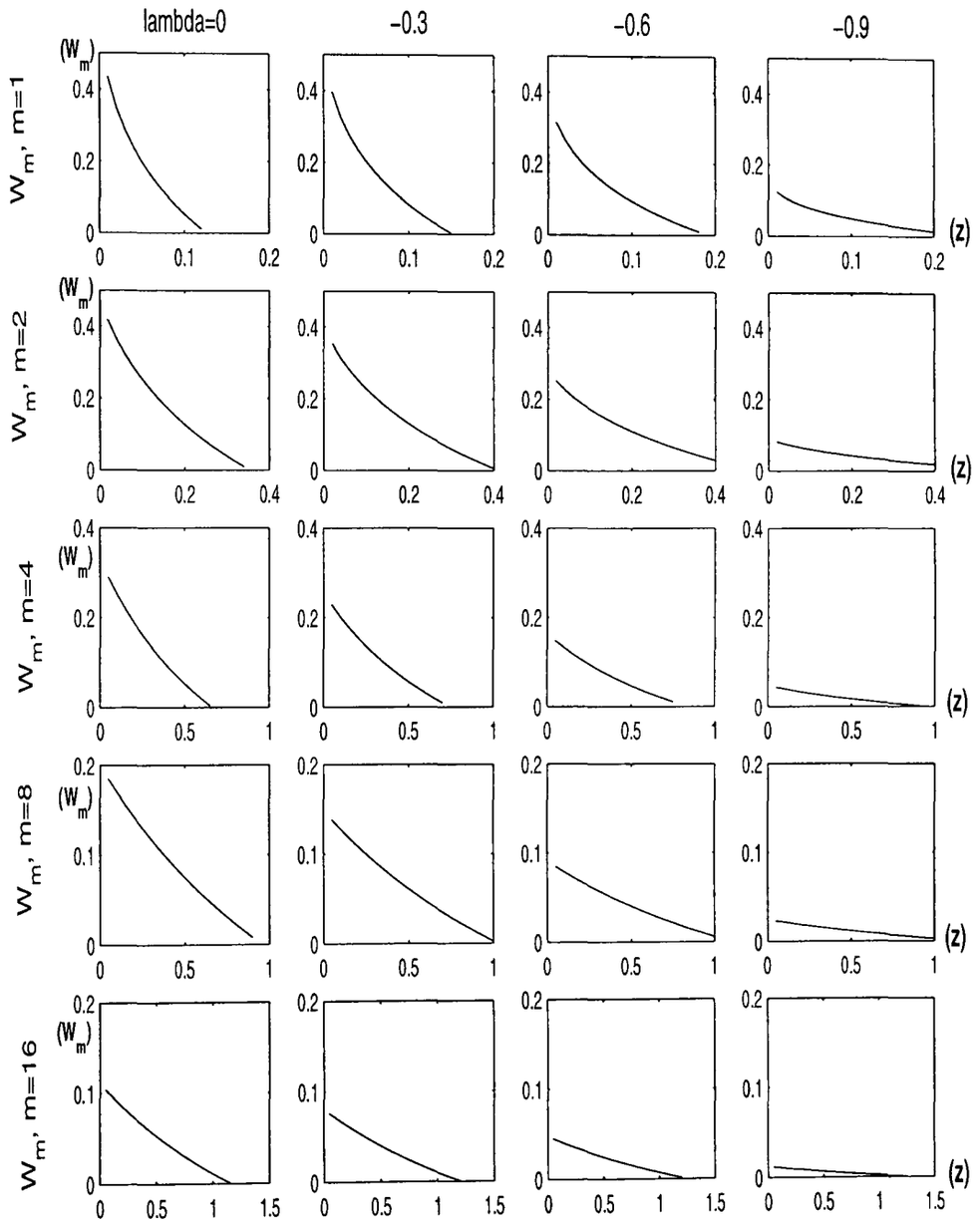

Figure 3.1: $W_m(z)$ with $m = 1, 2, 4, 8, 16$ and $\lambda = 0, -0.3, -0.6, -0.9$.

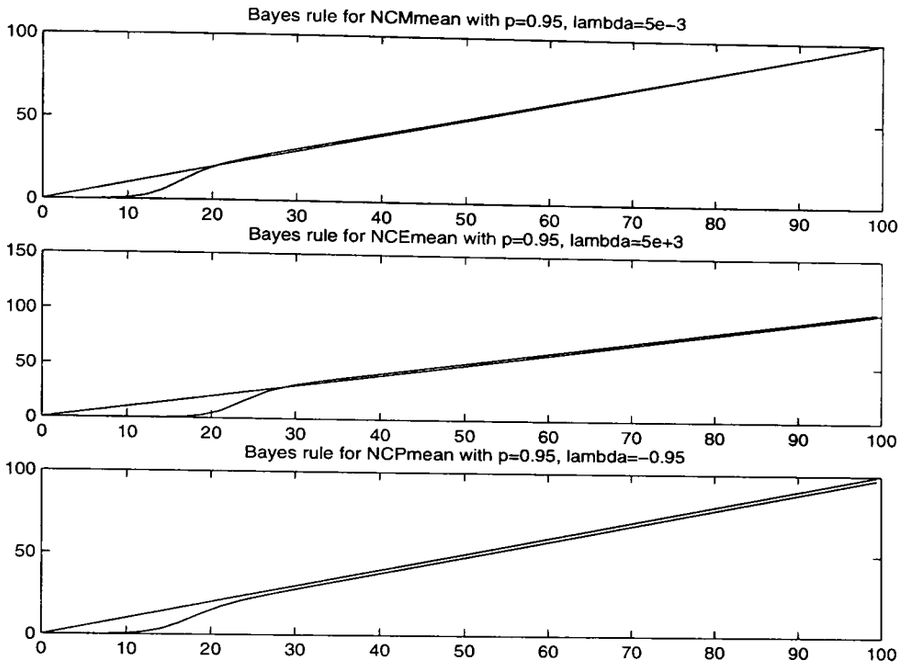

Figure 3.2: Bayesian shrinkage rule for three methods NCMmean, NCEmean and NCPmean.

NCPmean and NCEmean, while the risk of NCMmean is not bounded when ρ goes to ∞.

Heavy tailed priors. The work of Johnstone and Silverman (2005) has shown the theoretical desirability of using heavy-tailed priors in EB approaches to wavelet shrinkage. It is interesting to note that the power and exponential priors both imply heavy-tailed priors for ρ. The following lemma shows this in the power prior case.

Lemma 3.7 *In the case of the power prior, as $\rho \to \infty$, we have*

$$f_p(\rho|\sigma^2, \lambda, J = 0) \sim C_\lambda \frac{1}{\rho^{2+\lambda}},$$

where $\lambda > -1$ and $C_\lambda > 0$ does not depend on ρ.

Therefore, in the case of the power prior, λ determines the tail behavior of the prior distribution of ρ in a simple way.

Posterior mean versus posterior median. From Lemma 3.4 and Lemma 3.5, we can see that the posterior mean, ρ_{mean}, is neither strictly a shrinkage rule nor a

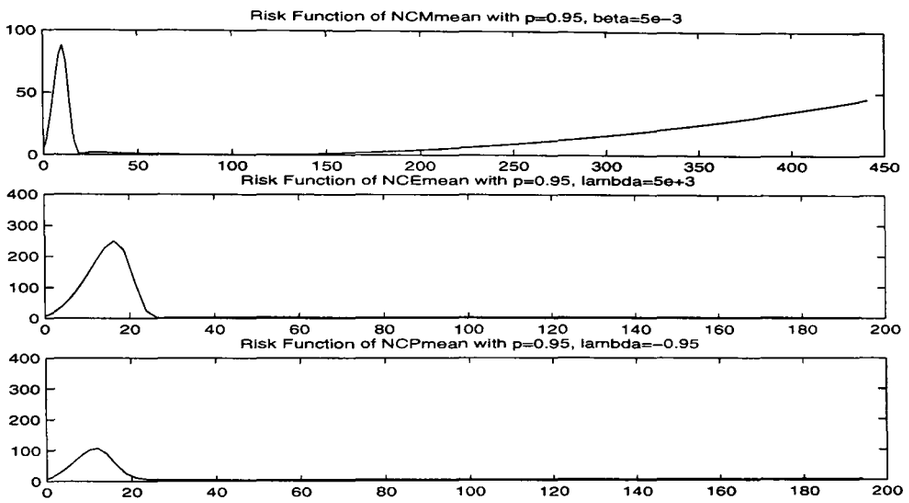

Figure 3.3: Risk functions for three methods NCMmean, NCEmean and NCPmean.

thresholding rule. Consequently, posterior median, ρ_{med} has clear theoretical advantages over ρ_{mean}. However, according to Definition 3.3, the posterior means of the power prior and exponential prior have the bounded shrinkage property. Because ρ_{mean} contains sufficient properties in common with the shrinkage and thresholding rules, it performs as well as ρ_{med}, as shown for example in Figure 3.4, the Bayes rule for NCPmean and NCPmedian. Generally speaking, their performances are quite similar except that, for small z, NCPmean performs like a shrinkage while NCPmedian is a thresholding rule and, for $z \to \infty$, NCPmedian is smaller than NCPmean. Also, we can see their performance in simulations on the standard model functions shown in Chapter 4. Figure 3.5 shows the risk function for NCPmean and NCPmedian.

Empirical block Bayes approach. In this chapter we have assumed that the hyperparameters σ^2, p and λ are known. In practice, we suggest that the following two-stage EBB procedure be adopted: estimate σ^2, p and λ, using the method outlined in §4.1; then substitute these estimates into the relevant formulae, treating them as though they are the known values. This, and other practical matters, are considered in the next chapter.

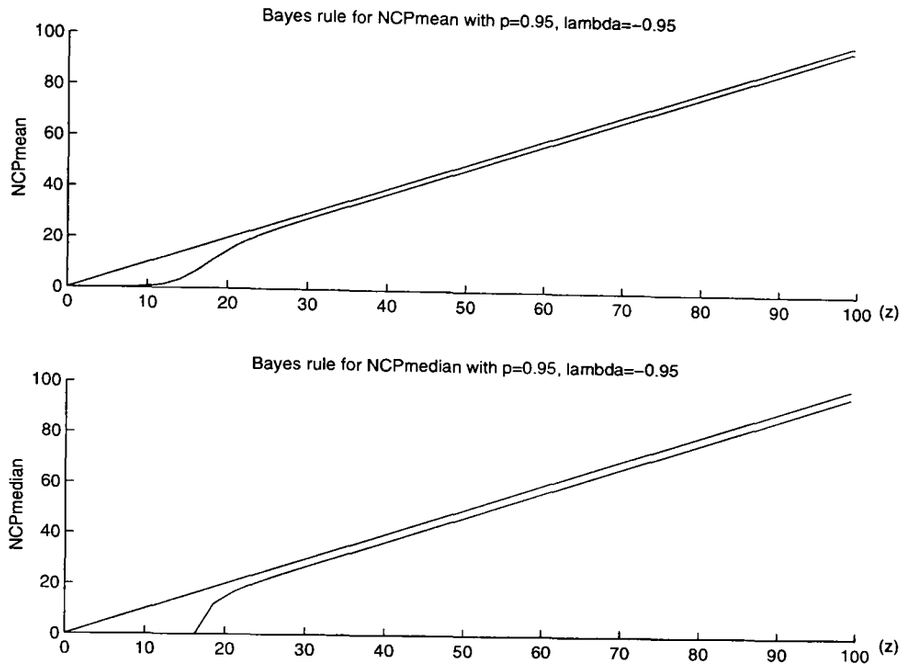

Figure 3.4: Bayesian shrinkage rule for NCPmean and NCPmedian.

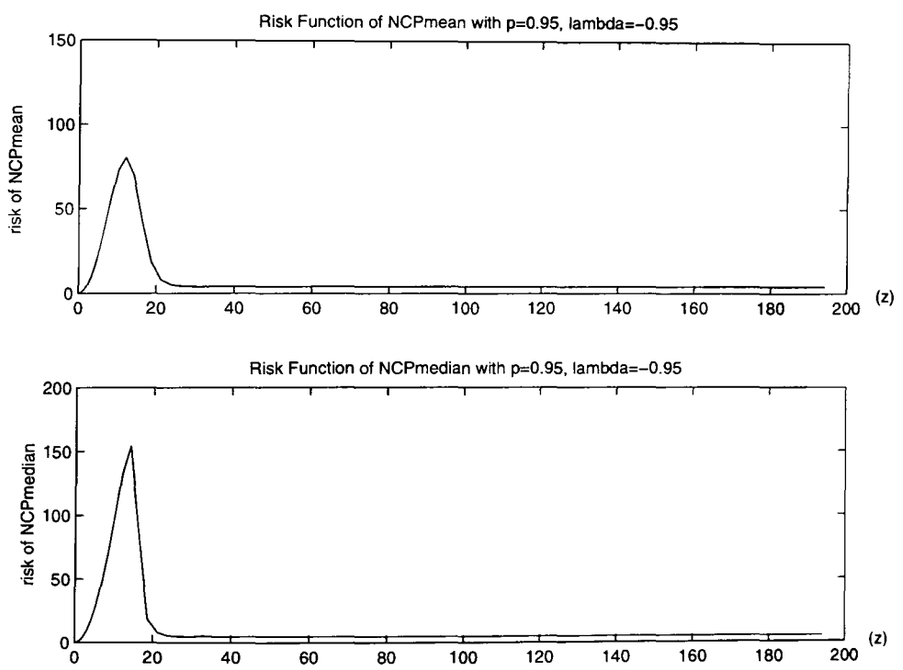

Figure 3.5: Risk function for NCPmean and NCPmedian.

3.7 Proofs

Proof of Lemma 3.1

The marginal density of z is

$$\begin{aligned} f(z|\sigma^2,\beta) &= \int_0^\infty f(z|\rho,\sigma^2)f(\rho|\beta)d\rho \\ &= \int_0^\infty \chi^2_m(z|\rho,\sigma^2)\chi^2_m(\rho|0,\beta^{-1})d\rho. \end{aligned} \tag{3.64}$$

Substitute the definition of the non-central χ^2 distribution in (3.1) to the right side of equation (3.64) and simplify it as

$$\sum_{k=0}^{\infty} \frac{(\beta/2)^{m/2} z^{m/2+k-1} e^{-z/(2\sigma^2)}}{\Gamma(m/2)k!(2\sigma^2)^{m/2+2k}\Gamma(m/2+k)} \int_0^\infty e^{-\rho/(2\sigma^2)} \rho^{m/2+k-1} e^{-\beta\rho/2} d\rho. \tag{3.65}$$

Noticing that

$$\int_0^\infty e^{-\rho/(2\sigma^2)} \rho^{m/2+k-1} e^{-\beta\rho/2} d\rho = \frac{\Gamma(m/2+k)}{\{1/(2\sigma^2)+\beta/2\}^{k+m/2}},$$

we can substitute it into (3.65) and further re-arrange (3.65) to be

$$\frac{(\beta/2)^{m/2} z^{m/2-1}}{\Gamma(m/2)(2\sigma^2)^{m/2}\{1/(2\sigma^2)+\beta/2\}^{m/2}} \sum_{k=0}^{\infty} \frac{e^{-z/(2\sigma^2)}(z/\sigma^2)^k}{k!2^k(2\sigma)^k\{1/(2\sigma^2)+\beta/2\}^k}.$$

Using the fact that

$$\sum_{k=0}^{\infty} \frac{e^{-z/(2\sigma^2)}(z/\sigma^2)^k}{k!2^k(2\sigma^2)^k\{1/(2\sigma^2)+\beta/2\}^k} = e^{-z\beta/(2(1+\sigma^2\beta))},$$

we have

$$\begin{aligned} f(z|\sigma^2,\beta) &= \frac{1}{\Gamma(m/2)}\left\{\frac{\beta}{2(1+\sigma^2\beta)}\right\}^{m/2} z^{m/2-1} e^{-z\frac{\beta}{2(1+\sigma^2\beta)}} \\ &= \chi^2_m\left\{z|0,\left(\frac{\beta}{1+\sigma^2\beta}\right)^{-1}\right\}. \end{aligned}$$

From the definition of u (3.12), we have

$$f(z|\sigma^2,\beta) = \chi^2_m\left(z|0,\frac{\sigma^2}{u}\right). \tag{3.66}$$

The posterior distribution of ρ can then be calculated from Bayes' theorem

$$\begin{aligned} f(\rho|z,\sigma^2,\beta) &= \frac{f(z|\rho,\sigma^2)f(\rho|\sigma^2,\beta)}{f(z|\sigma^2,\beta)} \\ &= \sum_{k=0}^{\infty} \frac{e^{-\frac{z}{2\sigma^2(1+\sigma^2\beta)}}\{\frac{z}{2\sigma^2(1+\sigma^2\beta)}\}^k}{k!} \cdot \frac{e^{-\frac{1+\sigma^2\beta}{2\sigma^2}\rho}\rho^{m/2+k-1}\cdot(\frac{1+\sigma^2\beta}{2\sigma^2})^{m/2+k}}{\Gamma(m/2+k)} \\ &= \chi^2_m\Big\{\rho\Big|\frac{z}{(1+\sigma^2\beta)^2}, \frac{\sigma^2}{1+\sigma^2\beta}\Big\} \\ &= \chi^2_m\{\rho|z(1-u)^2, \sigma^2(1-u)\}. \end{aligned} \tag{3.67}$$

Hence we obtain identity (3.16). ■

Proof of Lemma 3.2

(3.28) is straightforward. For (3.29), substituting $y = \gamma x$ into (3.4), we have

$$C_h(\gamma) = \frac{1}{\gamma^{h+1}} \int_0^{\gamma} y^h e^{-y} dy \tag{3.68}$$

When $\gamma \to \infty$, using the definition $\Gamma(h+1) = \int_0^{\infty} y^h e^{-y} dy$, we obtain (3.29). Using the result of (3.29), as $z \to \infty$ we have

$$\begin{aligned} \lim_{\gamma\to\infty} \frac{C_{h+t}(\gamma)}{C_h(\gamma)} &= \frac{\gamma^{-(h+t+1)}\Gamma(h+t+1)}{\gamma^{-(h+1)}\Gamma(h+1)} \\ &= \gamma^{-t}\frac{\Gamma(h+t+1)}{\Gamma(h+1)}. \end{aligned}$$

■

Proof of Lemma 3.3

In the case of the power prior, we have

$$\begin{aligned} E_P[\rho|z,\sigma^2,\theta] &= (1-\pi_P)\int_0^{\infty} \rho f_P(\rho|z,\sigma^2,\lambda,J=0)d\rho \\ &= (1-\pi_P)\int_0^{\infty} \rho \int_0^{\infty} \chi^2_m\{\rho|z(1-u)^2,\sigma^2(1-u)\} f_P(\beta|z,\sigma,\lambda)d\beta d\rho \\ &= (1-\pi_P)\int_0^{\infty} \rho \int_0^{1} \chi^2_m\{\rho|z(1-u)^2,\sigma^2(1-u)\} g_P(u|z,\sigma,\lambda)du d\rho. \end{aligned}$$

Exchanging the order of integration, we have

$$E_P[\rho|z,\sigma^2,\theta] = (1-\pi_P)\int_0^{1} g_P(u|z,\sigma,\lambda)\int_0^{\infty} \rho\chi^2_m\{\rho|z(1-u)^2,\sigma^2(1-u)\}d\rho du. \tag{3.69}$$

From (3.9), we have

$$\begin{aligned}\int_0^\infty \rho\chi^2_m\{\rho|z(1-u)^2,\sigma^2(1-u)\}d\rho &= (1-u)\sigma^2\Big\{m+\frac{z(1-u)^2}{\sigma^2(1-u)}\Big\}\\ &= m\sigma^2(1-u)+z(1-u)^2\end{aligned}$$

and so (3.69) simplifies to

$$\begin{aligned}E_P[\rho|z,\sigma^2,\theta] &= (1-\pi_P)\int_0^1 \{m\sigma^2(1-u)+z(1-u)^2\}g_P(u|z,\sigma,\lambda)du\\ &= (1-\pi_P)\{m\sigma^2E(1-u)+zE(1-u)^2\},\end{aligned} \tag{3.70}$$

where expectation

$$\begin{aligned}E(u^r) &= \int_0^1 u^{r+m/2+\lambda}\mathcal{C}_\eta\{z/(2\sigma^2)\}^{-1}\exp\{-uz/(2\sigma^2)\}du\\ &= \frac{\mathcal{C}_{\eta+r}\{z/(2\sigma^2)\}}{\mathcal{C}_\eta\{z/(2\sigma^2)\}}\end{aligned} \tag{3.71}$$

for $r=1,2$. Finally, we obtain (3.43). ■

Proof of Lemma 3.4

Using the result (3.30) of Lemma 3.2, as $z\to\infty$, we have

$$\begin{aligned}\mathcal{A}_{\eta,1}(z/2\sigma^2) &= \frac{\mathcal{C}_{\eta+1}\{z/(2\sigma^2)\}}{\mathcal{C}_\eta\{z/(2\sigma^2)\}}\\ &\sim \frac{\eta+1}{z/(2\sigma^2)}\\ &= \frac{m/2+\lambda+1}{z/(2\sigma^2)},\end{aligned}$$

where $\eta=m/2+\lambda$. So

$$\begin{aligned}E_P[\rho|z,\sigma^2,\theta] &= m\sigma^2+z-(m\sigma^2+2z)\frac{2\sigma^2(\lambda+m/2+1)}{z}+O(z^{-1})\\ &\sim z-4\sigma^2(\lambda+m/4+1)\end{aligned}$$

when $z\to\infty$, as required.

As $z\to 0$, $\mathcal{C}_h(0)=1/(h+1)$,

$$\begin{aligned}\mathcal{A}_{\eta,1}(0) &= \mathcal{C}_{\eta+1}(0)/\mathcal{C}_\eta(0)\\ &= \frac{\eta+1}{\eta+2}\end{aligned}$$

and

$$\begin{aligned}\mathcal{A}_{\eta,2}(0) &= \mathcal{C}_{\eta+2}(0)/\mathcal{C}_{\eta}(0) \\ &= \frac{\eta+1}{\eta+3}.\end{aligned}$$

Since,

$$R_P(0) = (\lambda+1)e^0\mathcal{C}_{\eta}(0) = \frac{\lambda+1}{\eta+1},$$

$$\begin{aligned}\pi_P(0) &= \frac{1}{1+\frac{1-p}{p}R_P(0)} \\ &= \frac{1}{1+(\frac{1-p}{p})(\frac{\lambda+1}{\eta+1})}.\end{aligned}$$

Hence

$$\begin{aligned}E_P[\rho|z,\sigma^2,\theta] &\sim \{1-\pi_P(0)\}\Big(m\sigma^2 - m\sigma^2\frac{\eta+1}{\eta+2}\Big) \\ &= \{1-\pi_P(0)\}\frac{m\sigma^2}{\eta+2}\end{aligned}$$

when $z \longrightarrow 0$ as required. ■

Proof of Lemma 3.7

In the case of the power prior, we have

$$\begin{aligned}f_P(\rho|\sigma^2,\lambda,J=0) &= \int_{\beta\in(0,\infty)} f(\rho|\beta)f_P(\beta|\sigma^2,\lambda,J=0)d\beta \\ &= \frac{\lambda+1}{\Gamma(m/2)2^{m/2}}\rho^{(m/2)-1}(\sigma^2)^{\lambda+1}\int_{\beta\in(0,\infty)}\frac{\beta^{(m/2)+\lambda}}{(1+\beta\sigma^2)^{\lambda+2}}e^{-\rho\beta/2}d\beta.\end{aligned}$$

Let $v = \rho\beta/2$, so

$$\begin{aligned}\int_{\beta\in(0,\infty)}\frac{\beta^{m/2+\lambda}}{(1+\beta\sigma^2)^{\lambda+2}}e^{-\rho\beta/2}d\beta &= \int_{v\in(0,\infty)}\frac{(\frac{2v}{\rho})^{m/2+\lambda}}{(1+\frac{2v}{\rho}\sigma^2)^{\lambda+2}}e^{-v}\frac{2}{\rho}dv \\ &= \Big(\frac{2}{\rho}\Big)^{m/2+\lambda+1}\int_{v\in(0,\infty)}\frac{v^{m/2+\lambda}e^{-v}}{\big(1+\frac{2v}{\rho}\sigma^2\big)^{\lambda+2}}dv\end{aligned}$$

As $\rho \longrightarrow \infty$, the dominated convergence theorem (see for example Billingsley, 1968, Theorem 5.5) gives

$$\begin{aligned}\int_{v\in(0,\infty)}\frac{v^{m/2+\lambda}e^{-v}}{(1+\frac{2v}{\rho}\sigma^2)^{\lambda+2}}dv &\longrightarrow \int_0^{\infty} v^{m/2+\lambda}e^{-v}dv \\ &= \Gamma\big(\frac{m}{2}+\lambda+1\big).\end{aligned}$$

Therefore, as $\rho \to \infty$,

$$\begin{aligned} f_P(\rho|\sigma^2, \lambda, J=0) &\sim \rho^{m/2-1}\Big(\frac{2}{\rho}\Big)^{m/2+\lambda+1}\frac{(\lambda+1)(\sigma^2)^{\lambda+1}}{\Gamma(m/2)2^{m/2}}\Gamma\Big(\frac{m}{2}+\lambda+1\Big) \\ &= \mathcal{C}_\lambda \frac{1}{\rho^{\lambda+2}} \end{aligned}$$

where

$$\mathcal{C}_\lambda = \frac{2^{\lambda+1}(\lambda+1)(\sigma^2)^{\lambda+1}}{\Gamma(m/2+\lambda+1)}.$$

■

Proof of Proposition 3.1

Define

$$Y_z = (\rho_z - z)/(2\sigma z^{1/2}), \qquad F_z(y) = Pr[Y_z \leq y]. \tag{3.72}$$

The proof firstly involves calculating asymptotic expressions for the first three posterior cumulants of ρ. Then Esseen's smoothing lemma and related results are used to justify an Edgeworth expansion for the posterior CDF of ρ. Finally, a Cornish-Fisher expansion for the posterior median is obtained, which reduces to (3.58). This proof is broken into four steps:

Step 1: It is shown that as $z \to \infty$,

$$E(Y_z) = z^{-1/2}k_1 + O(z^{-3/2}), \qquad Var(Y_z) = 1 + O(z^{-1})$$

and

$$\kappa_3(Y_z) = z^{-1/2}k_3 + O(z^{-3/2}),$$

where $k_1 = -\sigma(2 + 2\lambda + m/2)$ and $k_3 = 3\sigma$.

Step 2: Define

$$G_z(y) = \Phi(y) - z^{-1/2}\{\kappa_1 + \kappa_3(y^2-1)\}\phi(y), \tag{3.73}$$

where ϕ and Φ are the standard normal density and CDF, respectively. Note that (3.73) is a two-term Edgeworth approximation to a distribution with mean $z^{-1/2}\kappa_1$, variance 1 and third cumulant equal to $z^{-1/2}\kappa_3$; see for example Gnedenko and

Kolmogorov (1968, Chapter 8) and Hall (1992, Chapter 2). In Step 2 it is shown that, for any fixed $\epsilon \in (0,1)$,

$$\sup_{y\in\mathbf{R}} |F_z(y) - G_z(y)| \leq \sup_{0\leq t\leq z^\epsilon} \sup_{y\in\mathbf{R}} |F_z(y|t) - G_z(y|t)| + O(z^{-1}).$$

Step 3: A proof that

$$\sup_{0\leq t\leq z^\epsilon} \sup_{y\in\mathbf{R}} |F_z(y|t) - G_z(y|t)| = O(z^{-1}) \tag{3.74}$$

is given. Therefore

$$\sup_{y\in\mathbf{R}} |F_z(y) - G_z(y)| = O(z^{-1}) \text{ as } z \to \infty. \tag{3.75}$$

Step 4: Using (3.75), it is shown that

$$med(Y_z) = z^{-1/2}(k_1 - \frac{1}{6}k_3) + O(z^{-1}). \tag{3.76}$$

Then (3.58) follows directly from Step 4, because from (3.72) and (3.76),

$$\begin{aligned} med(\rho_z) &= 2\sigma z^{1/2} med(Y_z) + z + O(z^{-1/2}) \\ &= z - \sigma^2(5 + 4\lambda + m) + O(z^{-1/2}) \end{aligned}$$

as required.

Proof of Step 1 The density of the non-zero component of ρ_z is given by

$$f(\rho|z,\sigma^2,\lambda,J=0) = \int_0^{z/(2\sigma^2)} \chi^2_m\Big\{y|z\Big(1-\frac{2\sigma^2 t}{z}\Big)^2, \sigma^2\Big(1-\frac{2\sigma^2 t}{z}\Big)\Big\} \frac{t^\eta e^{-t}}{\{z/(2\sigma^2)\}^{\eta+1} C_\eta\{z/(2\sigma^2)\}} dt.$$

We first calculate the conditional moments of the non-zero component of ρ_z. For $h = 1, 2$ and 3,

$$\begin{aligned} \mu_h(t) &= E\big[(\rho_z - z)^h | t\big] \\ &= \int_0^\infty (y-z)^h \chi^2_m\Big\{y|z\Big(1-\frac{2\sigma^2 t}{z}\Big)^2, \sigma^2\Big(1-\frac{2\sigma^2 t}{z}\Big)\Big\} dy. \end{aligned}$$

Using the results about moment and cumulant of the rescaled non-central χ^2 distribution (3.7) (3.8) and the general relations between moments and cumulants, which are

$$\mu_1 = \kappa_1, \quad \mu_2 = \kappa_2 + \kappa_1^2, \quad \mu_3 = \kappa_3 + 2\kappa_2\kappa_1 + \kappa_1^3; \tag{3.77}$$

see e.g. McCullagh (1987), and putting $a = z(1 - 2\sigma^2 t/z)^2$ and $b = \sigma^2(1 - 2\sigma^2 t/z)$, we obtain the following:

$$\begin{aligned} \mu_1(t) &= E\big[(\rho_z - z)|t\big] \\ &= z(1 - 2\sigma^2 t/z)^2 + m\sigma^2(1 - 2\sigma^2 t/z) - z \\ &= m\sigma^2 - 4\sigma^2 t + O(z^{-1}) \end{aligned} \tag{3.78}$$

$$\begin{aligned} \mu_2(t) &= Var(\rho_z|t) + \{\mu_1(t)\}^2 \\ &= 4ab + 2mb^2 + \{\mu_1(t)\}^2 \\ &= 4\sigma^2 z + O(1) \end{aligned} \tag{3.79}$$

and

$$\begin{aligned} \mu_3(t) &= \kappa_3(\rho_z|t) + 3Var(\rho_z|t)\mu_1(t) + \{\mu_1(t)\}^3 \\ &= 24ab^2 + 8mb^3 + 3(4ab + 2mb^2)\mu_1(t) + \{\mu_1(t)\}^3 \\ &= 24z\sigma^4 + 12z\sigma^2\mu_1(t) + O(1). \end{aligned} \tag{3.80}$$

Next, we will evaluate the unconditional moments $E[(\rho_z - z)^h]$ for $h = 1, 2, 3$. Taking into account the zero component in the posterior distribution of ρ_z, which occurs with posterior probability π, we have

$$E[(\rho_z - z)^h] = \pi(-z)^h + (1 - \pi)E[\mu_h(t)] \tag{3.81}$$

where the expectation on the right hand side is with respect to a gamma variable t truncated above at $z/(2\sigma^2)$. When evaluating (3.81) as $z \to \infty$, the following points are relevant

1. π is exponentially small, so in particular

$$\pi|z|^h = O(z^{-1}) \quad \text{for} \quad h = 0, 1, 2, 3;$$

2. the $O(z^{-1})$ term in (3.78), $O(1)$ term in (3.79) and the $O(1)$ term in (3.80) preserve these orders after expectations are taken over t because each of these remainder terms only depends on t through a low power of t :

3. the error in ignoring the truncation of t at $z/(2\sigma^2)$ when evaluating expectations of a low order power of t is exponentially small, i.e for fixed h

$$\begin{aligned}\int_0^{z/(2\sigma^2)} t^h \frac{t^\eta e^{-t}}{\{z/(2\sigma^2)\}^{\eta+1} C_\eta\{z/(2\sigma^2)\}} dt &= \int_0^\infty t^h \frac{t^\eta e^{-t}}{\gamma(\eta+1)} dt + O(e^{-\epsilon z}) \\ &= \frac{\gamma(\eta+h+1)}{\gamma(\eta+1)} + O(e^{-\epsilon z})\end{aligned}$$

for some $\epsilon > 0$ as $z \to \infty$.

Taking 1, 2 and 3 into account, we evaluate (3.81) as follows: from (3.78),

$$E(\rho_z - z) = m\sigma^2 - 4\sigma^2(\eta+1) + O(z^{-1}); \tag{3.82}$$

from (3.79),

$$E(\rho_z - z)^2 = 4\sigma^2 z + O(1); \tag{3.83}$$

and from (3.80),

$$E(\rho_z - z)^3 = 24z\sigma^4 + 12z\sigma^2\{m\sigma^2 - 4\sigma^2(\eta+1)\} + O(1). \tag{3.84}$$

Therefore, since $\eta = \lambda + m/2$ and using (3.82), we obtain

$$\begin{aligned}E(Y_z) &= E\Big\{(\rho_z - z)/(2\sigma z^{1/2})\Big\} \\ &= -\frac{\sigma^2}{2\sigma z^{1/2}}(4 + 4\lambda + m) + O(z^{-3/2}) \\ &= -z^{-1/2}\sigma(2 + 2\lambda + m/2) + O(z^{-3/2}),\end{aligned}$$

$$\begin{aligned}E(Y_z^2) &= 4\sigma^2 z/(2\sigma z^{1/2})^2 \\ &= 1 + O(z^{-1})\end{aligned}$$

and

$$\begin{aligned}E(Y_z^3) &= E\Big\{(\rho_z - z)^3/(2\sigma z^{1/2})^3\Big\} \\ &= 3z^{-1/2}\sigma + \left(\frac{3}{2}\right) z^{-1/2}\sigma^{-1}\{m\sigma^2 - 4\sigma^2(\eta+1)\} + O(z^{-3/2}).\end{aligned}$$

Using the general relations (see e.g McCullagh, 1987)

$$\kappa_2 = \mu_2 - (\mu_1)^2, \kappa_3 = \mu_3 - 3\mu_2\mu_1 + 2\mu_1^3,$$

we have

$$\begin{aligned} Var(Y_z) &= E(Y_z^2) - \{E(Y_z)\}^2 \\ &= 1 - z^{-1}\sigma^2(2 + 2\lambda + m/2) + O(z^{-1}) \\ &= 1 + O(z^{-1}) \end{aligned}$$

and

$$\begin{aligned} \kappa_3(Y_z) &= E(Y_z^3) - 3E(Y_z^2)E(Y_z) + 2\{E(Y_z)\}^3 \\ &= 3\sigma z^{-1/2} + \Big(\frac{3}{2}\Big)\sigma^{-1}z^{-1/2}(m\sigma^2 - 4\sigma^2\eta) - \Big(\frac{3}{2}\Big)\sigma^{-1}z^{-1/2}(m\sigma^2 - 4\sigma^2\eta) \\ &\quad + O(z^{-3/2}) \\ &= 3\sigma z^{-1/2} + O(z^{-3/2}) \end{aligned}$$

as required. That concludes the proof of step 1.

Proof of Step 2: Define

$$F_z(y|t) = \int_0^{z+(2\sigma z^{1/2})y} \chi^2_m\Big\{\rho|z\Big(1 - \frac{2\sigma^2 t}{z}\Big)^2, \sigma^2\Big(1 - \frac{2\sigma^2 t}{z}\Big)\Big\}d\rho \tag{3.85}$$

and

$$g\{t|z/(2\sigma^2), \eta\} = \frac{t^\eta e^{-t}}{\{z/(2\sigma^2)\}^{\eta+1}\mathcal{C}_\eta\{z/(2\sigma^2)\}}.$$

Note that for each t, $F_z(y|t)$ is the CDF of a noncentral χ^2 distribution which has been translated and scaled. From the definition of $F_z(y)$, it follows that

$$F_z(y) = \pi + (1 - \pi)\tilde{F}_z(y)$$

where

$$\tilde{F}_z(y) = \int_0^{z/(2\sigma^2)} F_z(y|t)g\{t|z/(2\sigma^2), \eta\}dt.$$

Define also

$$G_z(y) = \Phi(y) - z^{-1/2}\{\kappa_1 + \kappa_3(y^2 - 1)\}\phi(y),$$

where $z^{-1/2}\kappa_1$ and $z^{-1/2}\kappa_3$ are, respectively, the first and third cumulants of the distribution with CDF $F_z(y)$, and write

$$G_z(y|t) = \Phi(y) - z^{-1/2}\{\kappa_1(t) + \kappa_3(t)(y^2 - 1)\}\phi(y),$$

where $z^{-1/2}\kappa_1(t)$ and $z^{-1/2}\kappa_3(t)$ are, respectively, the first and third cumulants of the distribution with CDF $F_z(y|t)$. Then

$$\begin{aligned} |F_z(y) - G_z(y)| &= \Big|F_z(y) - \tilde{F}_z(y) + \int_0^{z/(2\sigma^2)} \{F_z(y|t) - G_z(y|t)\} g\{t|z/(2\sigma^2), \eta\} dt \\ &\qquad + \int_0^{z/(2\sigma^2)} G_z(y|t) g\{t|z/(2\sigma^2), \eta\} dt - G_z(y)\Big| \\ &\leq I_1(y) + I_2(y) + I_3(y) \end{aligned} \tag{3.86}$$

where

$$I_1(y) = |F_z(y) - \tilde{F}_z(y)|,$$

$$I_2(y) = \Big| \int_0^{z/(2\sigma^2)} \{F_z(y|t) - G_z(y|t)\} g\{t|z/(2\sigma^2), \eta\} dt \Big|$$

and

$$I_3(y) = \Big| \int_0^{z/(2\sigma^2)} G_z(y|t) g\{t|z/(2\sigma^2), \eta\} dt - G_z(y) \Big|.$$

Now

$$I_1(y) = |\pi + (1-\pi)\tilde{F}_z(y) - \tilde{F}_z(y)| \leq \pi = O(z^{-1}) \tag{3.87}$$

uniformly in y as $z \to \infty$. Next, for any $\epsilon \in (0,1)$ and z large,

$$\begin{aligned} I_2(y) &= \Big| \int_0^{z^\epsilon} + \int_{z^\epsilon}^{z/(2\sigma^2)} \Big| \\ &\leq \sup_{0 \leq t \leq z^\epsilon} |F_z(y|t) - G_z(y|t)| \int_0^{z^\epsilon} g\{t|z/(2\sigma^2), \eta\} dt \\ &\qquad + \big(1 + C(z)\big) \int_{z^\epsilon}^{z/(2\sigma^2)} g\{t|z/(2\sigma^2), \eta\} dt \\ &\leq \sup_{0 \leq t \leq z^\epsilon} |F_z(y|t) - G_z(y|t)| + O(z^{-1}) \end{aligned} \tag{3.88}$$

uniformly in y, where we take

$$C(z) = \sup_{y \in \mathbf{R}} \sup_{0 \leq t \leq z^\epsilon} G_z(y|t)$$

and have used the fact that

$$\int_{z^\epsilon}^{z/(2\sigma^2)} g\{t|z/(2\sigma^2), \eta\} dt$$

decays faster than any power of z as $z \to \infty$. To show this, fix $N < \infty$. Then

$$\begin{aligned} z^N \int_{z^\epsilon}^{z/(2\sigma^2)} g\{t|z/(2\sigma^2),\eta\}dt &\leq z^N \int_{z^\epsilon}^{\infty} g\{t|z/(2\sigma^2),\eta\}dt \\ &= \frac{z^N}{\{z/(2\sigma^2\}^{\eta+1}\mathcal{C}_\eta\{t|z/(2\sigma^2)\}} \int_{z^\epsilon}^{\infty} t^\eta e^{-t}dt \\ &\sim \frac{z^N}{\{z/(2\sigma^2\}^{\eta+1}\mathcal{C}_\eta\{t|z/(2\sigma^2)\}} (z^\epsilon)^\eta e^{-z^\epsilon}. \end{aligned}$$

The above comes from the fact that

$$\int_a^\infty t^\eta e^{-t} \sim C a^\eta e^{-a},$$

see Dudley and Haughton (2002, lemma 4(a)). Also, for $\epsilon \in (0,1)$, $z^{N+\epsilon\eta}e^{-z^\epsilon} \to 0$ as $z \to \infty$ and

$$\{z/(2\sigma^2\}^{\eta+1}\mathcal{C}_\eta\{t|z/(2\sigma^2)\} = \int_0^{z/(2\sigma^2)} y^{\eta+1}e^{-y}dy \to \Gamma(\eta+1)$$

as $z \to \infty$. Therefore

$$z^N \int_{z^\epsilon}^{z/(2\sigma^2)} g\{t|z/(2\sigma^2),\eta\}dt \to 0$$

as $z \to \infty$ for any fixed $N \in (0,\infty)$. Finally,

$$\begin{aligned} I_3(y) = \Big|&\Phi(y) - z^{-1/2}\phi(y)\int_0^{z/(2\sigma^2)} \{\kappa_1(t) + \kappa_3(t)(y^2-1)\}g\{t|z/(2\sigma^2),\eta\}dt \\ &-\phi(y) + z^{-1/2}\phi(y)\{\kappa_1 + \kappa_3(y^2-1)\}\Big|. \end{aligned}$$

Therefore, since

$$\kappa_1 = \int_0^{z/(2\sigma^2)} \kappa_1(t)g\{t|z/(2\sigma^2),\eta\}dt,$$

it follows that

$$\begin{aligned} I_3(y) &= z^{-1/2}\phi(y)|y^2-1|\Big|\int_0^{z/(2\sigma^2)} \kappa_3(t)g\{t|z/(2\sigma^2),\eta\}dt - \kappa_3\Big| \\ &= O(z^{-3/2}) \end{aligned} \tag{3.89}$$

uniformly in y, since by direct calculation,

$$\Big|\int_0^{z/(2\sigma^2)} \kappa_3(t)g\{t|z/(2\sigma^2),\eta\}dt - \kappa_3\Big| = O(z^{-1}),$$

and $\phi(y)|y^2 - 1|$ is a bounded function of y.

Therefore, using (3.86), (3.87), (3.88) and (3.89), it is seen that

$$\sup_{y\in\mathbf{R}} |F_z(y) - G_z(y)| \leq \sup_{0\leq t\leq z^\epsilon} \sup_{y\in\mathbf{R}} \Big|F_z(y|t) - G_z(y|t)\Big| + O(z^{-1}).$$

Proof of Step 3: Firstly, We show that $F_z(y|t)$ can be viewed as the CDF of the sum of $\lfloor z \rfloor$ independent identically distributed random variables, where $\lfloor z \rfloor$ denotes the integer part of z, plus two other independent random variables.

Suppose $N \sim Poisson(\lambda)$, a Poisson random variable with mean λ, and let $X_1, X_2, \ldots$ denote a sequence of IID random variables, independent of N, with common MGF $M_X(\theta)$. Then the MGF of the random sum $S = \sum_{i=1}^{N} X_i$ is given by

$$M_S(\theta) = \sum_{r=0}^{\infty} e^{-\lambda} \frac{\lambda^r}{r!} M_X(\theta)^r = \exp\{\lambda[M_X(\theta) - 1]\}. \tag{3.90}$$

A non-central χ^2 variable W with density $\chi^2_m(y|z,a,b)$ has MGF given by (3.7)

$$M_W(\theta) = (1 - 2b\theta)^{-m/2} \exp\left\{\frac{za\theta}{1 - 2b\theta}\right\}, \tag{3.91}$$

where, in the present setting,

$$a = a(z,t) = \Big(1 - \frac{2\sigma^2 t}{z}\Big)^2 \qquad \textit{and} \qquad b = b(z,t) = \sigma^2\Big(1 - \frac{2\sigma^2 t}{z}\Big) \tag{3.92}$$

and $0 \leq t \leq z^\epsilon$ for some $\epsilon \in (0,1)$. Note that

$$\lim_{z\to\infty} \inf_{0\leq t\leq z^\epsilon} a(z,t) = \lim_{z\to\infty} \sup_{0\leq t\leq z^\epsilon} a(z,t) = 1$$

and

$$\lim_{z\to\infty} \inf_{0\leq t\leq z^\epsilon} b(z,t) = \lim_{z\to\infty} \sup_{0\leq t\leq z^\epsilon} b(z,t) = \sigma^2.$$

It follows from the above that a random variable Y with MGF given in (3.91) may be represented as follows:

$$W = b(U + V + \sum_{i=1}^{\lfloor z \rfloor} X_i) \tag{3.93}$$

where U, V and the X_i are all independent and have the following distributions: $U \sim \chi^2_m$; $V \sim \sum_{j=1}^{N_0} X_{0j}$ $(= 0$ if $N_0 = 0)$, where $N_0 \sim Poisson\{a(z - \lfloor z \rfloor)/(2b)\}$.

N_0 and the X_{0j} are independent; $X_i \sim \sum_{i=1}^{N_i} X_{ij}$ $(= 0$ if $N_i = 0))$, where $N_i \sim$ $Poisson\{a/(2b)\}$, $i = 1, \ldots, \lfloor z \rfloor$, the N_i and X_{ij} are independent with $X_{ij} \sim \chi^2_2$.

To see that (3.93) follows from (3.90) and (3.91), note that

$$E[e^{\theta bU}] = (1 - 2b\theta)^{-m/2},$$

$$E[e^{\theta bV}] = \exp\left\{\frac{a\theta(z - \lfloor z \rfloor)}{1 - 2b\theta}\right\}$$

and

$$E[\exp\{\theta b \sum_{i=1}^{X_i}] = \exp\left\{\frac{\lfloor z \rfloor a\theta}{1 - 2b\theta}\right\}.$$

The product of these three quantities gives (3.91).

A further point to note is that, if a and b are given by (3.92) and W is the random variable in (3.93) with MGF (3.91), then $(W - z)/(2\sigma^2 z^{1/2})$ has CDF $F_z(y|t)$ defined in (3.85).

Arguing heuristically for the moment, when $\lfloor z \rfloor$ is large, $W \approx b \sum_{i=1}^{\lfloor z \rfloor} X_i$, which suggests that we may apply the theorem on page 220 of Gnedenko and Kolmogorov (1968) to obtain the result (3.74). Condition (C) of the theorem is satisfied because the distribution of X_i has an absolutely continuous component.

To justify (3.74) vigorously, the key requirement is to extend Theorem 1(b) on page 204 of Gnedenko and Kolmogorov (1968) to include $U + V$. This can be done without difficulty because U and V are light tailed random variables with an absolutely continuous component.

Proof of Step 4: First, note that $G_z(y)$ converges uniformly in y to $\Phi(y)$ as $z \to \infty$. It follows that the equation $G_z(y) = 1/2$ has a unique solution $y = y_{0,z}$ provided z is sufficiently large. Moreover, the derivative $G'_z(y) = dG_z(y)/dy$ is bounded above zero in a neighbourhood of $y_{0,z}$ in the sense that there exists an $\epsilon > 0$ and a z_0 such that

$$\inf_{z \geq z_0} \inf_{y = |y - y_{0,z}| \leq \epsilon} G'_z(y) > C > 0 \tag{3.94}$$

By definition, $med(Y_z)$ is such that $F_z\{med(Y_z)\} = 1/2$. Therefore, from (3.75)

$$\left|\frac{1}{2} - G_z\{med(Y_z)\}\right| = O(z^{-1}) \tag{3.95}$$

By Taylor's theorem,

$$G_z\{med(Y_z)\} = G_z(y_{0,z}) + \{med(Y_z) - y_{0,z}\}G_z'(y^*) \tag{3.96}$$

where y^* lies between $med(Y_z)$ and $y_{0,z}$. It follows from (3.94)-(3.96) that

$$|med(Y_z) - y_{0,z}| = O(z^{-1}) \tag{3.97}$$

Finally, it follows from the Cornish-Fisher expansion (see Hall, 1992, p68) that

$$y_{0,z} = z^{-1/2}(k_1 - k_3/6) + O(z^{-1})$$

Therefore, using (3.97),

$$med(Y_z) = z^{-1/2}(k_1 - k_3/6) + O(z^{-1})$$

and Step 4 is proved. ■

Proof of Lemma 3.6

By definition,

$$R_P(z) = (\lambda + 1)e^{z/(2\sigma^2)}\mathcal{C}_\eta\{z/(2\sigma^2)\} \tag{3.98}$$

where $\eta = \lambda + m/2$ and $\mathcal{C}_\eta(r)$ is defined in (3.4). The CDF $F(z)$ is given by

$$F(z) = \int_0^z \int_0^1 \chi^2_m\{\rho|z(1-u)^2, \sigma^2(1-u)\}\mathcal{C}_\eta\{z/(2\sigma^2)\}^{-1}u^\eta e^{\{-uz/(2\sigma^2)\}}dud\rho \tag{3.99}$$

A. Behavior of $R_P(z)$ as $z \to 0$. Since

$$R_P(z) = (\lambda + 1)\int_0^1 u^\eta e^{(1-u)z/(2\sigma^2)}du,$$

as $z \to 0$, the Taylor expansion for $e^{(1-u)z/(2\sigma^2)}$ gives

$$\begin{aligned} R_P(z) &= (\lambda+1)\int_0^1 u^\eta\Big\{1 + \Big(\frac{1-u}{2\sigma^2}\Big)z + O(z^2)\Big\}du \\ &= (\lambda+1)\Big\{\frac{1}{\eta+1} + \Big(\frac{1}{\eta+1} - \frac{1}{\eta+2}\Big)\Big(\frac{z}{2\sigma^2}\Big) + O(z^2)\Big\} \\ &= \Big(\frac{\lambda+1}{\eta+1}\Big)\Big\{1 + \Big(\frac{1}{\eta+2}\Big)\Big(\frac{z}{2\sigma^2}\Big) + O(z^2)\Big\}. \end{aligned} \tag{3.100}$$

B. Behavior of $F(z)$ as $z \to 0$. Since

$$\begin{aligned}
&\chi^2_m\{\rho|z(1-u)^2, \sigma^2(1-u)\} \\
&= \sum_{k=0}^{\infty} e^{-z(1-u)/(2\sigma^2)} \frac{\left\{\frac{z(1-u)}{2\sigma^2}\right\}^k}{k!} \chi^2_{m+2k}\{\rho|0, \sigma^2(1-u)\} \\
&= \sum_{k=0}^{\infty} e^{-z(1-u)/(2\sigma^2)} \frac{\left\{\frac{z(1-u)}{2\sigma^2}\right\}^k}{k!} \cdot \frac{1}{\Gamma(m/2+k)} \left\{\frac{1}{2\sigma^2(1-u)}\right\}^{\frac{m}{2}+k} \rho^{\frac{m}{2}+k-1} e^{\frac{-\rho}{2\sigma^2(1-u)}}.
\end{aligned}$$

and $0 \leq \rho \leq z$, as $z \to 0$, $\rho \to 0$, we have

$$\chi^2_m\{\rho|z(1-u)^2, \sigma^2(1-u)\} = \frac{1}{\Gamma(m/2)} \left(\frac{1}{2\sigma^2(1-u)}\right)^{m/2} \rho^{m/2-1} e^{-\rho/\{2\sigma^2(1-u)\}} + O(z). \tag{3.101}$$

Therefore, substituting (??) into (3.99), we obtain

$$\begin{aligned}
F(z) &= C_\eta\{z/2\sigma^2\}^{-1} \frac{1}{\Gamma(m/2)} \\
&\quad \times \int_0^z \left\{ \int_0^1 (1-u)^{-m/2} e^{-\frac{\rho}{2\sigma^2(1-u)}} u^\eta du \right\} \rho^{m/2-1} d\rho + O(z^2).
\end{aligned} \tag{3.102}$$

Also using the fact that $e^{-uz/(2\sigma^2)} = 1 + O(z)$, we have

$$C_\eta\{z/2\sigma^2\}^{-1} = \eta + 1 + O(z).$$

Putting $u = y/(1+y)$, the inner integral of right hand side of (3.102) transforms as follows:

$$\int_0^1 (1-u)^{-m/2} e^{-\frac{\rho}{2\sigma^2(1-u)}} u^\eta du = \int_0^\infty (1+y)^{m/2-2} e^{-\frac{\rho(1+y)}{2\sigma^2}} \left(\frac{y}{1+y}\right)^\eta dy. \tag{3.103}$$

Write

$$H_{m,\eta}\{\rho/(2\sigma^2)\} = \int_0^\infty (1+y)^{m/2-2} e^{-\frac{\rho y}{2\sigma^2}} \left(\frac{y}{1+y}\right)^\eta dy,$$

then the right hand side of (3.103) equals $e^{-\rho/(2\sigma^2)} H_{m,\eta}\{\rho/(2\sigma^2)\}$.

Since $0 \leq \rho \leq z$ and $z \to 0$, we consider what happens to (3.103) when $\rho \to 0$. Three cases arise: $m = 1$, $m = 2$ and $m \geq 3$.

Case 1: $m = 1$. Here,

$$H_{1,\eta}\{\rho/(2\sigma^2)\} \to H_{1,\eta}(0) < 0,$$

where

$$H_{1,\eta}(0) = \int_0^\infty (1+y)^{-3/2}\Big(\frac{y}{1+y}\Big)^\eta dy.$$

Putting $t = y/(1+y)$ and $dt = \frac{1}{(1+y)^2}dy$, we obtain

$$\begin{aligned} H_{1,\eta}(0) &= \int_0^1 t^\eta (1-t)^{1/2-1} dt \\ &= \frac{\Gamma(\eta+1)\Gamma(1/2)}{\Gamma(\eta+3/2)}. \end{aligned}$$

Then, using

$$\begin{aligned} F(z) &= \frac{\eta+1}{\Gamma(1/2)}\Big(\frac{1}{2\sigma^2}\Big)^{\frac{1}{2}} \int_0^z e^{-\rho/(2\sigma^2)} \frac{\Gamma(\eta+1)\Gamma(1/2)}{\Gamma(\eta+3/2)} \rho^{-1/2} d\rho \\ &= \frac{2(\eta+1)\Gamma(\eta+1)}{\Gamma(\eta+3/2)}\Big(\frac{z}{2\sigma^2}\Big)^{1/2} + O(z^{1/2}). \end{aligned}$$

therefore,

$$1 - 2F(z) = 1 - \frac{4(\eta+1)\Gamma(\eta+1)}{\Gamma(\eta+3/2)}\Big(\frac{z}{2\sigma^2}\Big)^{1/2} + O(z^{1/2}). \tag{3.104}$$

Consequently, $W_m(z)$ is decreasing for z sufficiently small since $z^{1/2} > z$ when $0 < z < 1$.

Case 2: $m = 2$.

As $\rho \to 0$, $H_{2,\eta}\{\rho/(2\sigma^2)\} \to \infty$, so this case is different. Put $v = \rho y/(2\sigma^2)$ in (3.103). Then the right hand side of (3.103) is given by

$$e^{-\frac{\rho}{2\sigma^2}} \int_0^\infty \Big(\frac{2\sigma^2}{\rho}\Big)\Big(1 + \frac{2\sigma^2 v}{\rho}\Big)^{-1}\Big(\frac{2\sigma^2 v}{\rho + 2\sigma^2 v}\Big)^\eta e^{-v} dv. \tag{3.105}$$

Then as $\rho \to 0$, (3.105) asymptotic to

$$\begin{aligned} \int_0^\infty \Big(\frac{\rho}{2\sigma^2} + v\Big)^{-1} e^{-v} dv &\sim \int_0^\epsilon \Big(\frac{\rho}{2\sigma^2} + v\Big)^{-1} dv \\ &= \log\Big\{\frac{\epsilon + \rho/(2\sigma^2)}{\rho/(2\sigma^2)}\Big\} \\ &= \log\Big(\frac{2\sigma^1\epsilon}{\rho} + 1\Big) \\ &\sim \log\Big(\frac{1}{\rho}\Big) + \log(2\sigma^2\epsilon) \\ &\sim \log\Big(\frac{1}{\rho}\Big). \end{aligned}$$

From (3.102), as $z \to 0$,

$$\begin{aligned} F(z) &\sim (\eta+1)\Big(\frac{1}{2\sigma^2}\Big)\int_0^z \log\Big(\frac{1}{\rho}\Big)d\rho \\ &\sim (\eta+1)\Big(\frac{z}{2\sigma^2}\Big)\log\Big(\frac{1}{z}\Big)+O(z). \end{aligned}$$

Therefore,

$$1-2F(z) = 1-2(\eta+1)\Big(\frac{z}{2\sigma^2}\Big)\log\Big(\frac{1}{z}\Big)+O(z), \tag{3.106}$$

and the result holds in this case too.

Case 3: $m \geq 3$.

Consider $H_{m,\eta}\{\rho/(2\sigma^2)\}$ when $m \geq 3$. Put $v = \rho y/(2\sigma^2)$ in (3.103) again, we obtain

$$H_{m,\eta}\Big(\frac{\rho}{2\sigma^2}\Big) = e^{-\rho/(2\sigma^2)}\int_0^\infty \Big(\frac{2\sigma^2}{\rho}\Big)\Big(1+\frac{2\sigma^2 v}{\rho}\Big)^{m/2-2}\Big(\frac{2\sigma^2 v}{\rho+2\sigma^2 v}\Big)^{\eta}e^{-v}dv. \tag{3.107}$$

As $\rho \to 0$,

$$\begin{aligned} H_{m,\eta}\Big(\frac{\rho}{2\sigma^2}\Big) &\sim \Big(\frac{2\sigma^2}{\rho}\Big)^{m/2-1}\int_0^\infty v^{m/2-2}e^{-v}dv \\ &= \Big(\frac{2\sigma^2}{\rho}\Big)^{m/2-1}\Gamma\Big(\frac{m}{2}-1\Big), \end{aligned}$$

and

$$\begin{aligned} F(z) &= \frac{\eta+1}{\Gamma(m/2)}\Big(\frac{1}{2\sigma^2}\Big)^{m/2}\int_0^z \Big(\frac{2\sigma^2}{\rho}\Big)^{m/2-1}\Gamma\Big(\frac{m}{2}-1\Big)\rho^{m/2-1}d\rho \\ &= \frac{(\eta+1)\Gamma(m/2-1)}{\Gamma(m/2)}\Big(\frac{1}{2\sigma^2}\Big)\int_0^z d\rho \\ &= \frac{\eta+1}{(m/2-1)}\frac{z}{2\sigma^2}+O(z^2). \end{aligned}$$

Therefore

$$1-2F(z) = 1-\frac{4(\eta+1)}{(m-2)}\frac{z}{2\sigma^2}+O(z^2). \tag{3.108}$$

Comparison of (3.108) with (3.101) tells us that for $m \geq 3$, $H_{m,\sigma^2,\lambda}(z)$ is decreasing for all $z \geq 0$ sufficiently small when

$$\frac{4(\eta+1)}{m-2} > \frac{1}{\eta}. \tag{3.109}$$

Since $\eta = \lambda + m/2$, $\lambda > -1$ and $m \geq 3$, (3.109) is always true. So the conclusion holds. ■

Chapter 4

Empirical Bayes Block Shrinkage: Practical Issues

In this chapter we propose a novel empirical Bayes block (EBB) shrinkage procedure using the Bayesian methodology developed in Chapter 3. The key feature of the approach is that shrinkage is based on the posterior distribution of the sum of squares of the wavelet coefficients in a block. In §4.1, we provide the motivations for basing shrinkage on the block sum of squares. In §4.2, we describe how to apply the Bayesian methodology mentioned in Chapter 3 to the task of performing posterior shrinkage based on the sum of squares. For simplicity, it is assumed in §4.2 that the hyperparameters of the prior are known and that block sizes have already been chosen. Estimation of the hyperparameters and choice of block size are considered in §4.3 and §4.4. In § 4.5, we discuss an equivariance property of the EBB method. Then a simulation study and application of this approach to the denoising of planar curves will be described in §4.7 and §4.8.

4.1 Motivation

As mentioned above, the Bayesian methodology developed for the non-central χ^2 distribution in the previous chapter can be used to perform posterior shrinkage based

on a sum of squares of wavelet coefficients. However, before considering the details of how this can be done, motivation for basing shrinkage on the sum of squares is provided.

1. It seems appealing and natural to base block shrinkage directly on some measure of the "energy" of the block, such as the sum of squares of wavelet coefficients in the block.

2. Frequentist approaches to block thresholding developed by Hall *et al.* (1997, 1998, 1999), Cai (1999, 2002) and Cai and Silverman (2001) employ thresholding based on block sums of squares, and these authors have shown that there are theoretical and practical advantages in this approach. Therefore, it is of interest to develop empirical Bayes block (EBB) methods which directly parallel the frequentist block thresholding techniques.

3. Suppose we wish to denoise a noisy curve $\mathbf{Y(t)} = (\mathbf{Y_1(t)}, \mathbf{Y_2(t)})^{\mathbf{T}} \in \mathbf{R^2}$ observed at $t = t_1, \ldots, t_n$ using wavelet shrinkage. In some situations we may be interested in the shape of the region enclosed by $Y(t)$, in which case we would want to use a method for estimating $Y(t)$ which is invariant with respect to rotations of the ambient space $\mathbf{R}^2$. Let $\{\widehat{d}_{1i}\}_{i=1}^n$ and $\{\widehat{d}_{2i}\}_{i=1}^n$ denote the empirical wavelet coefficients obtained from components $Y_1(t)$ and $Y_2(t)$ respectively. Then if the shrinkage procedure is to be equivariant with respect to rotations of the ambient space $\mathbf{R}^2$, shrinkage should be based on the sums of squares $\{(\widehat{d}_{1i})^2 + (\widehat{d}_{2i})^2\}_{i=1}^n$.

4. If complex wavelets are used (see Lawton, 1993, and Lina and Mayrand, 1995) then it will often be natural to base shrinkage on the amplitude of the complex wavelet coefficient, and to leave the phase unchanged. Once again this leads to a shrinkage procedure based on a sum of squares.

5. If multiwavelets are used then we may wish to develop an EB version of the frequentist thresholding procedure given by Downie and Silverman (1998).

Once again, this is conveniently achieved by basing the shrinkage procedure on a suitable sum of squares.

4.2 Bayesian Block Shrinkage

Consider the standard model (2.13) given in §2.4. After performing the DWT on the noisy observations, we obtain the empirical wavelet coefficients $\tilde{d}_{jk}$, $j = j_0, \ldots, J$, $k = 0, \ldots, 2^j - 1$, which are candidates for shrinkage and the scaling coefficients $\tilde{c}_{j_0 k}$, $k = 0, \ldots, 2^{j_0} - 1$ which will be kept unchanged, as mentioned in §2.5.2. Generally, it is natural to think of wavelet coefficients as having two indices, one for location and one for level. However, for notational convenience we suppress this structure in this chapter and just work with a single index.

Let $\mathcal{K}$ denote the labels of the full set of empirical wavelet coefficients $\{\tilde{d}_i : i \in \mathcal{K}\}$ under consideration and suppose $B \subset \mathcal{K}$. There are various cases we may wish to consider.

4.2.1 Bayesian Block Shrinkage in the Standard Model

In the case of block shrinkage in the standard model (2.13), B may represent a single block where, typically, a block would consist of neighbouring coefficients at the same level. In the case of complex wavelets, B may consist of labels for the real and imaginary part of a single complex wavelet coefficient.

Define $\tilde{\mathbf{d}}_B = \{\tilde{d}_i : i \in B\}$ and let $n(B)$ denote the number of elements (i.e. labels) in B. Under the standard model (2.13),

$$\tilde{\mathbf{d}}_B \sim N_{n(B)}(\mathbf{d}_B, \sigma^2 I_{n(B)}) \tag{4.1}$$

where $\mathbf{d}_B$ is the noiseless version of $\tilde{\mathbf{d}}_B$ and $I_{n(B)}$ is the $n(B) \times n(B)$ identity matrix. Define

$$z = ||\tilde{\mathbf{d}}_B||^2 = \sum_{i \in B} \tilde{d}_i^2 \quad \text{and} \quad \rho = ||\mathbf{d}_B||^2 = \sum_{i \in B} d_i^2.$$

For given values of the hyperparameters σ^2 and $\boldsymbol{\theta} = (p, \lambda)$, let $\mathcal{B}_{\sigma^2,\boldsymbol{\theta}}(z)$ denote ρ_{mean}, ρ_{med} or ρ_{hyp} defined in §3.4. Then, corresponding to Step 2 (the shrinkage step mentioned in §2.5.1), we propose

$$\widehat{\mathbf{d}}_B = \tilde{\mathbf{d}}_B \sqrt{\frac{\mathcal{B}_{\sigma^2,\boldsymbol{\theta}}(z)}{z}}. \tag{4.2}$$

Although (4.2) seems an intuitively reasonable approach, at least when $\mathbf{d}_B$ has an isotropic covariance structure, it is not a fully Bayesian procedure. Here we will look more closely at the motivation for choosing (4.2).

Motivation for (4.2)

In (4.1), the distribution of $\tilde{\mathbf{d}}_B$ under the model is stated. What we would like to know is $\mathbf{d}_B$, the wavelet coefficients in the block B determined by the unknown function f. Given the isotropic covariance structure assumed in the underlying model, and assuming that prior information is non-informative with respect to the directional component, $\mathbf{d}_B/||\mathbf{d}_B||$, of $\mathbf{d}_B$, it seems very natural to estimate the directional components of $\mathbf{d}_B$ by the directional component of $\tilde{\mathbf{d}}_B$, namely $\tilde{\mathbf{d}}_B/||\tilde{\mathbf{d}}_B||$. Indeed, to estimate the directional component of $\mathbf{d}_B$ any other way under the isotropy assumptions of the model and prior would seem inappropriate. If we knew $\rho = ||\mathbf{d}_B||^2$, then it would be natural to estimate $\mathbf{d}_B$ by $z^{-1/2}\rho^{1/2}\tilde{\mathbf{d}}_B$, where $z = ||\tilde{\mathbf{d}}_B||^2$. This is the same as (4.2), but with ρ replacing $\mathcal{B}_{\sigma^2,\boldsymbol{\theta}}(z)$. In practice we do not know $\rho = ||\mathbf{d}_B||^2$. A key goal of the thesis is to set up convenient Bayesian machinery for estimating ρ using a suitable posterior quantity.

However, there is another important point to consider. When considering the posterior mean, it could reasonably be asked why we take $\mathcal{B}_{\sigma^2,\boldsymbol{\theta}}(z)$ to be $\rho_{mean} = E[\rho]$ rather than, say, $\rho_{hmean} = h^{-1}E[h(\rho)]$, where h is any strictly increasing function and $h^{-1}(\cdot)$ is the functional inverse of h. Admittedly, there is a degree of arbitrariness here in the use of the mean (despite the fact that in our numerical work the posterior mean has done rather well).

In contrast, the posterior median has a strong and attractive invariance property:

for any strictly increasing function h, $median(\rho) = h^{-1}\{median[h(\rho)]\}$.

In summary, we believe (4.2) is theoretically well-motivated under the assumed model in the case of the posterior median, but this is not the case to the same extent in the case of the posterior mean (because of the degree of arbitrariness in the choice of h).

4.2.2 Bayesian Block Shrinkage in a General Model

More generally, suppose that instead of (4.1) we have

$$\tilde{\mathbf{d}}_B \sim N_{n(B)}(\mathbf{d}_B, \sigma^2 V) \tag{4.3}$$

where the matrix V is assumed known. Then we may still use shrinkage procedure (4.2), but with z now defined by $z = \tilde{\mathbf{d}}_B^T V^{-1} \tilde{\mathbf{d}}_B$. Thus shrinkage will generally be heavier in those directions in $\tilde{\mathbf{d}}_B$-space with larger variance, which makes good sense.

In practice V will often be unknown. In such cases we suggest estimating V and, subsequently, treating the estimate as though it were known. Estimation of V will be considered in Chapter 5.

If $\mathcal{K}$ is partitioned into N non-overlapping blocks $B(1), \ldots, B(N)$, not necessarily of the same size, then by applying the above procedure to each block we obtain

$$\widehat{\mathbf{d}} = \{\widehat{d}_{B(\ell)} : \ell \in N\} = \{\widehat{d}_i : i \in \mathcal{K}\};$$

see Step 2 mentioned in § 2.5.1.

4.3 Estimation of Hyperparameters

It is necessary to specify values of the hyperparameters σ^2, p and λ before the shrinkage step (4.2) can be used in practice.

To obtain an estimate, $\widehat{\sigma}^2$ say, of σ^2 we suggest using the median of the $z_i = ||\tilde{\mathbf{d}}_{B(i)}||^2$ at the finest level J, divided by the median of the standard central χ^2_m,

assuming that the block sizes at the finest level are all equal to m. Note that this is analogous to the proposal of Donoho and Johnstone (1994, p. 446), who suggest estimating σ by the median absolute deviation of the wavelet coefficients at the finest level J divided by 0.6745.

We have estimated λ by $\widehat{\lambda}$ obtained using the following "quick-and-dirty" method. Suppose that $\mathcal{K}$ is partitioned into blocks $B(1), \ldots, B(N)$ of size $m_1, \ldots, m_N$. We shall use (3.32) to obtain a "central" value of β. Since for each i,

$$\begin{aligned} E_M[z_i|\sigma^2, \theta] &= \int_0^\infty z_i \left[p\chi^2_{m_i}(z_i|0,\sigma^2) + (1-p)\chi^2_{m_i}(z_i|0,\sigma^2/u_\beta)\right] dz_i \\ &= pm_i\sigma^2 + (1-p)\frac{m_i\sigma^2}{u_\beta} \\ &= pm_i\sigma^2 + (1-p)m_i\sigma^2\frac{(1+\beta\sigma^2)}{\beta\sigma^2} \\ &= pm_i\sigma^2 + (1-p)m_i\sigma^2 + (1-p)\frac{m_i}{\beta} \\ &= m_i\sigma^2 + (1-p)\frac{m_i}{\beta} > m_i\sigma^2, \end{aligned}$$

for $0 < p < 1$, it is reasonable to expect that $\sum_{i=1}^N z_i$ is larger than $\sigma^2 \sum_{i=1}^N m_i$. In practice it is usually substantially larger, due to the presence of a signal. Given $Thresh_i = 2m_i\sigma^2 \log N$, a preliminary estimator of p is given by

$$\widehat{p}_0 = \frac{\#\{z_i < Thresh_i : \ i = 1, \ldots, N\}}{N}. \tag{4.4}$$

Then the "central" β is obtained by matching moments

$$\sum_{i=1}^N z_i = \widehat{\sigma}^2 \left(\sum_{i=1}^N m_i\right) + (1-\widehat{p}_0)\left(\sum_{i=1}^N m_i\right)/\beta,$$

which yields

$$\widehat{\beta} = \frac{1-\widehat{p}_0}{\left(m_0^{-1}\sum_{i=1}^N z_i\right) - \widehat{\sigma}^2}.$$

where $m_0 = \sum_{i=1}^N m_i$.

In the case of the mass point prior, we choose $\widehat{\lambda} = \widehat{\beta}$; for the "power" prior, we choose $\widehat{\lambda}$ so that $\widehat{\beta}$ is the median of (3.36). Note that the calculations are facilitated if one transforms from $\widehat{\beta}$ to $\widehat{u}$ using (3.12), and then finds the $\widehat{\lambda}$ so

that $\int_0^{\widehat{u}}(\widehat{\lambda}+1)u^{\widehat{\lambda}}du = 1/2$. This yields $\widehat{\lambda} = -\log 2/\log \widehat{u} - 1$. Similarly, for the "exponential" prior, we may choose $\widehat{\lambda}$ so that $\widehat{\beta}$ is the median of (3.46).

Once $\hat{\sigma}^2$ and $\widehat{\lambda}$ have been obtained, we can then estimate p at each different level by marginal maximum likelihood using (3.24) in the general case, and (3.32), (3.47) and (3.39) in the mass point, exponential and power prior cases, respectively.

Initially, we attempted to use the EM algorithm to estimate λ and p jointly. Results for the mass point prior were satisfactory, but we found that, for the exponential and power priors, the marginal profile likelihood for λ (with p "maximised out") was generally rather flat and for this reason did not lead to a stable or reliable procedure for estimating λ. However, once σ^2 and λ are specified, the marginal maximum likelihood estimator of p is unique. We have found that the "quick and dirty" method described above is very fast and our simulation study indicates that it produces consistently good results.

An alternative to estimating λ is to choose λ to achieve suitable shrinkage or thresholding properties on the basis of Proposition 3.2 in Chapter 3.

4.4 Choice of Block Size

In addition to estimating the hyperparameters σ^2, λ and p, we need to consider the choice of the block size, which corresponds to the degrees of freedom, m, of the non-central χ^2 distribution. The size of the blocks plays an important role in estimation. Theoretical results concerning the choice of block size have been discussed by several authors. In particular, Cai (2002) considered the effect of block size on local and global adaptivity. Global adaptivity can be measured by the mean integrated squared error:

$$R(\hat{f}-f) = E||\hat{f}-f||_2^2 = E\Big\{\int_{-\infty}^{\infty}(\hat{f}(x)-f(x))^2dx\Big\}, \tag{4.5}$$

while local adaptivity can be measured by the expected loss at a point x_0:

$$R\{\hat{f}(x_0)-f(x_0)\} = E\{\hat{f}(x_0)-f(x_0)\}^2. \tag{4.6}$$

Cai (2002) noted that there are conflicting requirements in block size for achieving the global and local adaptivity. For the block size $L = (\log n)^s$, in order to achieve global adaptivity, s must be greater than 1, while for the local adaptivity, s has to be less than 1. The optimal choice, $L = \lfloor \log n \rfloor$, where $\lfloor \cdot \rfloor$ denotes the integer part, is achieved by considering the choice of thresholding constant λ for a given block size. In the regression case, Hall *et al.* (1999) suggested choosing block size $L = (\log n)^2$ and threshold constant $\lambda \geq 48$ to attain the minimax rate of convergence under the global risk measure (4.5). In the density estimation case, Hall *et al.* (1998) indicated that choosing blocks to be of size $c_i(\log n)^2$, where i denotes the resolution level, will achieve the minimax rate of convergence under the global risk measure (4.5).

However, although such results give valuable theoretical insights, they do not provide explicit rules for choosing the block size. The numerical results presented in the next section suggest that in a given problem the best choice of block size will in practice depend not only on sample size, but also fairly strongly on the unknown signal to be estimated. Cai (2002) also pointed out some noticeable discrepancies between the asymptotic and finite sample results.

4.5 Computation of the Posterior Median

As mentioned before, we will use the Lugannani-Rice (LR) saddlepoint formula (see Lugannani and Rice, 1980, Daniels, 1987 and Jensen, 1995) to approximate the CDF of a non-central χ^2 distribution.

4.5.1 Saddlepoint Method

In many contexts, the LR formula has been proved to be a remarkably accurate approximation to the CDF of a sum of independent random variables. In its standard form, the LR approximation for a continuous variable X is given by

$$Pr(X \leq y) \simeq LR(y) = \Phi(\widehat{w}) + \phi(\widehat{w}) \left\{ \frac{1}{\widehat{w}} - \frac{1}{\widehat{u}} \right\}. \tag{4.7}$$

where Φ and ϕ are the standard normal CDF and density, respectively:

$$\widehat{w} = [2\{K(\widehat{t} - \widehat{t}y)\}]^{1/2} sgn(\widehat{t}) \tag{4.8}$$

and

$$\widehat{u} = \widehat{t}K''(\widehat{t})^{1/2}, \tag{4.9}$$

where $sgn(\widehat{t}) = -1, 0$ or 1 depending on whether $\widehat{t}$ is negative, zero or positive. $K(t)$ is the CGF (cumulant generating function) of X and $\widehat{t}$ is the (unique) solution to the saddlepoint equation $K'(t) = y$. We may also investigate the second order LR approximation (see Wood *et al.*, 1993) as

$$\widetilde{Pr}(X \leq y) \simeq LR(y) - \phi(\widehat{w}) \left\{ \frac{1}{\widehat{u}} \left(\frac{1}{8}\widehat{\gamma}_4 - \frac{5}{24}\widehat{\gamma}_3^2 \right) - \frac{1}{\widehat{u}^3} - \frac{\widehat{\gamma}_3}{2\widehat{u}^2} + \frac{1}{\widehat{w}^3} \right\}, \tag{4.10}$$

where $LR(y)$ is given in (4.7), $\widehat{w}$ and $\widehat{u}$ are as before, and $\hat{\gamma}_i = K^{(i)}(\hat{t})/\{K^{(2)}(\hat{t})\}^{i/2}$ $(i = 3, 4)$ are the standardised third and fourth cumulants. We can see that in the first and second LR formula the normal distribution has played a prominent role. Since the CDF and density of the standard normal distribution are available as standard functions in **Matlab**, the LR approximation is numerical realisable.

4.5.2 Finding the Posterior Median

Recall that finding the posterior median is equivalent to finding the α-quantile, with $\alpha = (1/2 - \pi)/(1 - \pi)$, $\pi < 1/2$, of the distribution with density $f(\rho|z, \sigma^2, \lambda, J = 0)$ given in (3.25). This involves finding the α-quantile of the CDF

$$F(y) = \int_{u \in (0,1)} H(y|z, \sigma^2, u) g(u) du$$

where $H(y|z, \sigma^2, u)$ and possible choices for $g(u)$ are defined in (3.55), (3.56) and (3.57).

Since $H(y|z, \sigma^2, u)$ is the CDF of a non-central χ^2 distribution we may approximate it using the LR formula mentioned above. Following (4.7), we have

$$\widehat{H}(y|z, \sigma^2, u) = \Phi(\widehat{w}) + \phi(\widehat{w}) \left\{ \frac{1}{\widehat{w}} - \frac{1}{\widehat{u}} \right\}. \tag{4.11}$$

In order to calculate $\widehat{w} = [2\{K(\widehat{t} - \widehat{t}y\}]^{1/2} sgn(\widehat{t})$ and $\widehat{u} = \widehat{t}K''(\widehat{t})^{1/2}$, we have K, the CGF of $\sigma^2(1-u)\chi^2_m\{z(1-u)/\sigma^2\}$,

$$K(t) = -\frac{m}{2}\log\{1 - 2\sigma^2(1-u)t\} + \frac{z(1-u)^2 t}{1 - 2\sigma^2(1-u)t},$$

and its derivatives

$$K'(t) = \frac{m\sigma^2(1-u)}{1 - 2\sigma^2(1-u)t} + \frac{z(1-u)^2}{\{1 - 2\sigma^2(1-u)t\}^2}$$

and

$$K''(t) = \frac{2m\sigma^4(1-u)^2}{\{1 - 2\sigma^2(1-u)t\}^2} + \frac{4z(1-u)\sigma^2}{\{1 - 2\sigma^2(1-u)t\}^3}.$$

The solution, $\widehat{t}$, of the equation $K'(t) = y$, is given by

$$\widehat{t} = \frac{2y(1-u)^{-1} - m\sigma^2 - \sqrt{m^2\sigma^4 + 4yz}}{4y\sigma^2}.$$

Similarly, we have the second order LR approximation as

$$\widetilde{H}(y|z,\sigma^2,u) = \widehat{H}(y|z,\sigma^2,u) - \phi(\widehat{w})\left\{\frac{1}{\widehat{u}}\left(\frac{1}{8}\widehat{\gamma}_4 - \frac{5}{24}\widehat{\gamma}_3^2\right) - \frac{1}{\widehat{u}^3} - \frac{\widehat{\gamma}_3}{2\widehat{u}^2} + \frac{1}{\widehat{w}^3}\right\} \quad (4.12)$$

with the jth derivative, $K^{(j)}(t)$, $j = 3, 4$, of $K(t)$ given by

$$K^{(j)}(t) = \frac{(j-1)!2^{j-1}m\sigma^{2j}(1-u)^j}{\{1 - 2\sigma^2(1-u)t\}^j} + \frac{j!2^{j-1}z\sigma^{2(j-1)}(1-u)^{j+1}}{\{1 - 2\sigma^2(1-u)t\}^{j+1}}.$$

Substituting $\widehat{H}$ or $\widetilde{H}$ for H in (3.54), we may evaluate the integral over u numerically to obtain an approximation $\widehat{F}$ or $\widetilde{F}$ for F. Then we may solve $\widehat{F}(y) = \alpha$ or $\widetilde{F}(y) = \alpha$ using a root finder to obtain an estimate $\widehat{y}_\alpha$ or $\widetilde{y}_\alpha$ of the α-quantile y_α.

Although, generally, the second order LR approximation is more accurate than the first order LR approximation, it displays numerical problems when $\widehat{t}$ is very close to 0 due to the removable singularity at $t = 0$. We have used the linear interpolation method to avoid this problem in numerical work.

Generally, the posterior median, ρ_{med}, is more difficult to compute than the posterior mean. However, in case of the power prior, it can be accurately approximated using the LR approach described above.

4.6 Equivariance

We now consider what happens if the coordinate system is changed. Starting with the standard model

$$y_i = f_i + \epsilon_i, \quad i = 1, \cdots, n. \tag{4.13}$$

we transform coordinates as follows:

$$\begin{aligned} y_i^* &= a + by_i, \\ f_i^* &= a + bf_i, \quad i = 1, \cdots, n, \\ \epsilon_i^* &= b\epsilon_i. \end{aligned} \tag{4.14}$$

where a and b are constants.

Let $\hat{\mathbf{f}} = (\hat{f}_1, \cdots, \hat{f}_n)^T$ denote an estimator of $\mathbf{f} = (f_1, \cdots, f_n)^T$ based on the data vector $\mathbf{y} = (y_1, \cdots, y_n)^T$ and let $\hat{\mathbf{f}}^* = (\hat{f}_1^*, \cdots, \hat{f}_n^*)^T$ denote the estimator of $\mathbf{f}^* = (f_1^*, \cdots, f_n^*)^T$ based on the transformed data $\mathbf{y}^* = (y_1^*, \cdots, y_n^*)^T$.

Definition 4.1 *The estimation procedure is said to be equivariant if*

$$\hat{\mathbf{f}}^* = a\mathbf{1}_n + b\hat{\mathbf{f}}, \tag{4.15}$$

where $\mathbf{1}_n = (1, \cdots, 1)^T$ *is the n-vector of ones.*

In other words, the estimation procedure is equivariant if applying the procedure to the unknown function in the transformed coordinates gives the same result as applying the procedure to the unknown function in the original coordinates and then transforming to the new coordinates. This definition of equivariance is tailored to the present setting. A more general definition of equivariance in terms of a transformation group is given in Barndorff-Nielson and Cox (1994, page 53 and page 77).

Following (2.12), we can then write the orthogonal matrix $\mathcal{W}$ as

$$\mathcal{W} = \begin{pmatrix} \mathcal{W}_0 \\ \mathcal{V}_{j_0} \end{pmatrix},$$

where $\mathcal{W}_0$ is an $n \times (n - 2^{j_0})$ matrix, $\mathcal{V}_{j_0}$ is an $n \times 2^{j_0}$ matrix and j_0 is the coarest level. Also, according to the vanishing moment property (Condition (3) of Definition (2.1)), when $m = 0$, we have

$$\mathcal{W}_0 \mathbf{1}_n = \mathbf{0}_{n-2^{j_0}} \tag{4.16}$$

where $\mathbf{0}_{n-2^{j_0}}$ is the $(n - 2^{j_0})$-vector of zeros. The three steps of wavelet estimation mentioned in §2.5.1 can be represent as follows.

In Original Coordinates

Step 1 Obtain the empirical wavelet coefficients

$$\mathcal{W}\mathbf{y} = \begin{pmatrix} \mathcal{W}_0\mathbf{y} \\ \mathcal{V}_{j_0}\mathbf{y} \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{d}} \\ \tilde{\mathbf{c}} \end{pmatrix}.$$

Step 2 Adjust $\tilde{\mathbf{d}}$ to $\hat{\mathbf{d}}$ by applying a suitable automatic shrinkage procedure to $\tilde{\mathbf{d}}$.

Step 3 Estimate $\mathbf{f}$ by

$$\hat{\mathbf{f}} = \mathcal{W}^T \begin{pmatrix} \hat{\mathbf{d}} \\ \tilde{\mathbf{c}} \end{pmatrix}.$$

In the transformed coordinates, the same procedure is applied to the transformed observation vector $\mathbf{y}^*$.

Step 1* Obtain

$$\mathcal{W}\mathbf{y}^* = \begin{pmatrix} \mathcal{W}_0\mathbf{y}^* \\ \mathcal{V}_{j_0}\mathbf{y}^* \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{d}}^* \\ \tilde{\mathbf{c}}^* \end{pmatrix}.$$

Step 2* Adjust $\tilde{\mathbf{d}}^*$ to $\hat{\mathbf{d}}^*$ by applying a suitable automatic shrinkage procedure to $\tilde{\mathbf{d}}^*$.

Step 3* Estimate $\mathbf{f}^*$ by

$$\hat{\mathbf{f}}^* = \mathcal{W}^T \begin{pmatrix} \hat{\mathbf{d}}^* \\ \tilde{\mathbf{c}}^* \end{pmatrix}.$$

Proposition 4.1 *Consider the wavelet-based denoising procedure outlined above. Assume that (4.16) holds. Suppose that in step 2 (or step 2*) above, the non-central χ^2 based Bayesian block denoising procedure is used, with power prior, exponential prior or mass point prior, as defined in section 3.3, and with posterior quantity given by the posterior mean, posterior median or hypothesis testing procedure, as defined*

in section 3.4. If the hyperparameters are estimated as indicated in section 3.4, then the estimation procedure is equivariant, i.e. (4.15) holds.

Proof of Proposition 4.1: The relevant z_i consist of sums of squares of components of $\tilde{\mathbf{d}}$, while the z_i^* consist of sums of squares of components of $\tilde{\mathbf{d}}^*$. It follows from (4.16) that $\mathbf{d}^* = b\mathbf{d}$ and therefore $z_i^* = b^2 z_i$, $i = 1, \cdots, n$. Consequently, $(\hat{\sigma}^*)^2 = b^2\hat{\sigma}^2$, where $\hat{\sigma}^2$ is the rescaled median of the z_i. $\hat{\rho}_0$ defined in (4.4) is the same in either coordinate system, while $\hat{\beta}^* = (1/b^2)\hat{\beta}$. It follows from Theorem 3.1 and Theorem 3.2 in Chapter 3 that $\rho^*_{med} = b^2\rho_{med}$ and $\rho^*_{mean} = b^2\rho_{mean}$. Moreover, π, the posterior probability of the unit mass point, does not depend on the scale factor b. Therefore $\rho^*_{hyp} = b^2\rho_{hyp}$. The remaining part of the proof is to note that orthogonality of $\mathcal{W}$ combined with (4.16) implies that $\mathcal{V}_{j_0}^T\mathcal{V}_{j_0} = I_{2^{j_0}\times 2^{j_0}}$. Thus (4.15) holds.

4.7 Simulation Results and an Example

In this section, we present the results of some simulations to illustrate the methods proposed above.

4.7.1 Simulation Study

Four signals, "HeaviSine", "Blocks", "Bumps" and "Doppler", first proposed in Donoho and Johnstone (1994, 1995) as test functions for wavelet estimators, are considered here. Each test function was rescaled so as to achieve a different signal-to-noise ratio (SNR), which here is the ratio of standard deviations of the signal and noise (see Donoho and Johnstone, 1994, 1995). Independent standard normal noise was added at each location x_i according to the standard model (2.13). The value of the rescaled function was obtained as follows:

$$f_{scaled} = f * \left(\frac{SNR}{std(f)}\right), \tag{4.17}$$

where *std* is the standard deviation of the test function, which is defined as

$$std(f) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left[f(x_i) - \{\frac{1}{n}\sum_{i=1}^{n} f(x_i)\}\right]}.$$

The following signal-to-noise ratios were considered: SNR=3, 5, 7 and 10. Each function was sampled at n=256, 512, 1024 and 2048. In order to compare the results, we follow most authors by choosing the following wavelets for different signals: Symmlet 8 for "HeaviSine" and "Doppler", Haar for "Blocks" and Daubechies 3 for "Bumps", where the numbers 8 and 3 indicate the number of vanishing moments for Symmlet and Daubechies wavelets, respectively.

The comparisons discussed below are based on the average mean squared error (MSE), defined as the sum of the mean Squared Bias (MSB) and variance (Var), which are computed as follows.

MSE: The mean squared error is computed for each run and averaged over all simulation runs.

MSB: Let $\overline{f}(x_i)$ be the average of $\widehat{f}(x_i)$ over the number of the simulation runs. The MSB is $(1/n)\sum_{i=1}^{n}(f(x_i) - \overline{f}(x_i))^2$.

Variance: is $(1/n)\sum_{i=1}^{n}(\widehat{f}(x_i) - \overline{f}(x_i))^2$.

Figures 4.1 and 4.2 contain the noisy "Bumps" with SNR=3, and "Doppler" with SNR=7, and the reconstructions were obtained from NCEmean, NCEhyp, NCPmean, NCPhyp, NCPmed based on approximation (4.7) and NCPmed based on approximation (4.10) (see § 3.6 for their definitions). Table 4.1, which is split into parts (a) and (b), shows average MSE results with 100 simulation runs for four test functions, obtained using NCPmean, NCEmean and NCMmean at different SNRs (3, 5, 7 and 10) and in different sample sizes (256, 512, 1024 and 2048) at a fixed block size $m = 2$. These figures and table show that each of the priors did reasonably well. For small SNRs (for example SNR=3), most of the average MSEs of the NCPmean and NCEmean are smaller than those of the NCMmean, while

for large SNRs (for example SNR=5, 7 and 10), most of the average MSEs of the NCPmean and NCMmean are smaller than those of the NCEmean. But overall the procedure based on the power prior (NCPmean) had a clear superiority over the other two.

In what follows we will focus on the signal reconstruction methods based on the posterior mean and posterior median respectively, using the non-central χ^2 model (NC) with power prior (P) to show the effect of choosing different block sizes. We use the notation NCPmean-m and NCPmed-m to denote these two reconstruction methods with block size m.

In Table 4.2, a preliminary study is undertaken to compare the performance of NCPmean and NCPmed methods at the block size $m = 1, 2, 4, 8, 16$ with the sample size $n = 1024$. It is interesting to note that the best choice of block size, as measured by the average MSE, is fairly constant across a range of signal-to-noise ratios, but depends quite heavily on the signal. These simulation results indicate that, in these examples, a small block size ($m = 1, 2$ or 4) is appropriate for the sample size $n = 1024$. Table 4.3 reports the average MSE of NCPmean and NCPmed at the different combinations of the block size $m = 1, 2, 4, 8, 16$ with the sample size $n = 512, 1024, 2048, 4096, 8192$. In Table 4.2 and Table 4.3, the number in each cell is the MSE and an asterisk is used to identify the optimum block size. We can see that the best choice of block size will depend not only on sample size, but also on the unknown signal to be estimated. For example, functions with significant spatial variability, such as Doppler, generally work better with a larger block size ($m = 4$,) while smooth functions, such as HeaviSine and Blocks, can achieve better values for average MSE by reducing m. To a certain extent, the results agree with the simulation results presented by Cai (2002).

Typical reconstructions using the posterior mean and posterior median with block sizes $m = 2, 4$ and 8 are shown in Figures 4.3- 4.5.

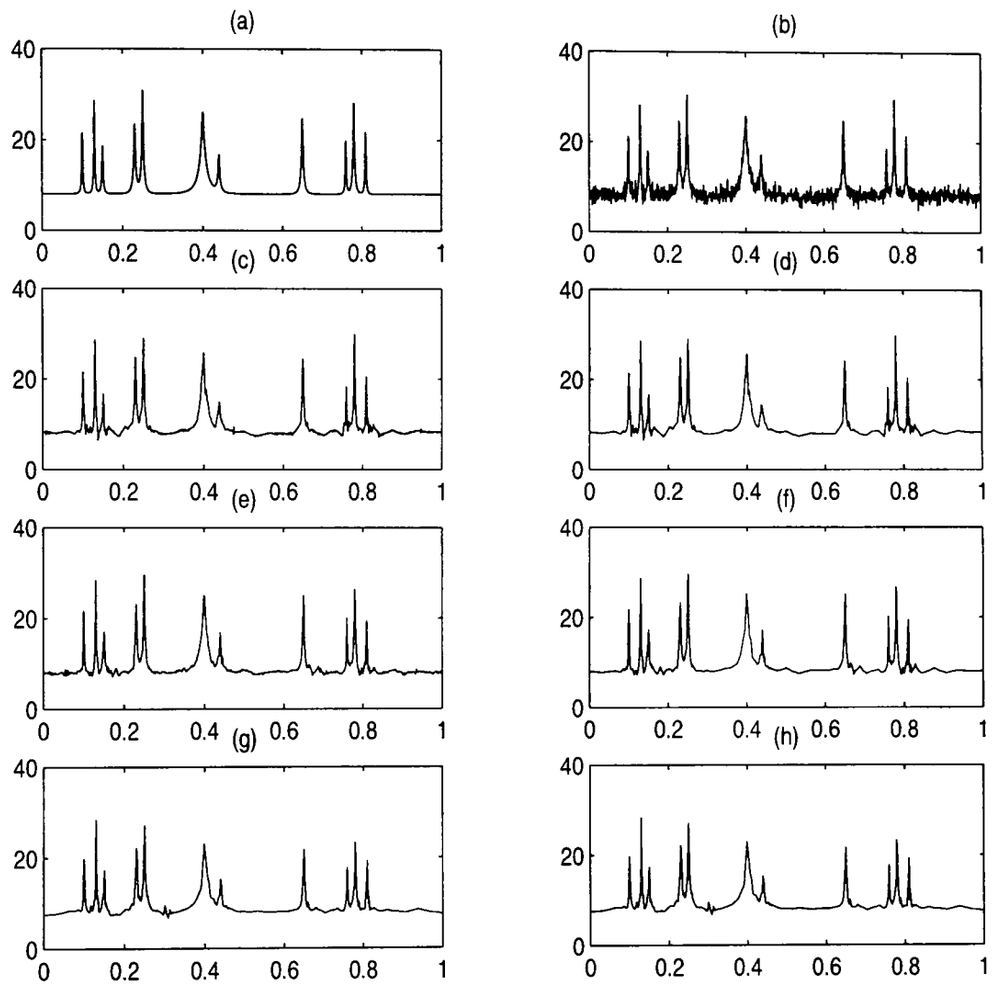

Figure 4.1: The original Bumps function and various reconstructions based on sample size n=1024 and SNR=3: (a) original signal; (b) noisy signal; (c) reconstructed by NCEmean; (d) reconstructed by NCEhyp; (e) reconstructed by NCPmean; (f) reconstructed by NCPhyp; (g) reconstructed by NCPmed using (4.7); (h) reconstructed by NCPmed using (4.10).

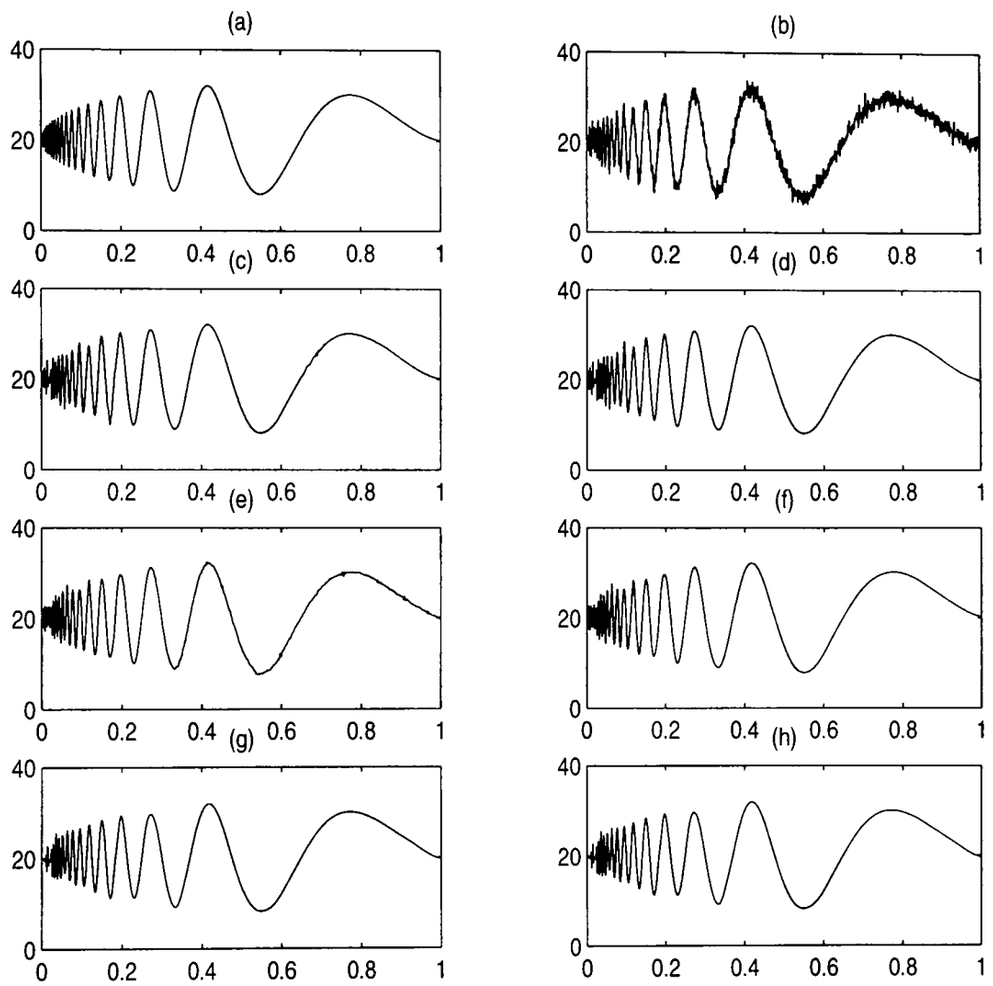

Figure 4.2: The original Doppler function and various reconstructions based on sample size n=1024 and SNR=7: (a) original signal; (b) noisy signal; (c) reconstructed by NCEmean; (d) reconstructed by NCEhyp; (e) reconstructed by NCPmean; (f) reconstructed by NCPhyp; (g) reconstructed by NCPmed using (4.7); (h) reconstructed by NCPmed using (4.10).

Part (a):		HeaviSine				Blocks			
Number	methods	SNR=3	SNR=5	SNR=7	SNR=10	SNR=3	SNR=5	SNR=7	SNR=10
256	NCPmean	0.1409	0.2181	0.2687	0.3148	0.4704	0.3490	0.3116	0.2986
	NCEmean	0.1326	0.2260	0.3136	0.3963	0.4865	0.3756	0.3080	0.2712
	NCMmean	0.1678	0.2317	0.2771	0.3099	0.4085	0.3355	0.2986	0.2775
512	NCPmean	0.0881	0.1340	0.1714	0.1780	0.2744	0.2515	0.2321	0.2123
	NCEmean	0.0870	0.1495	0.1886	0.2615	0.2921	0.2887	0.2630	0.2197
	NCMmean	0.1071	0.1334	0.1717	0.1941	0.2840	0.2514	0.2273	0.2010
1024	NCPmean	0.0528	0.0638	0.0851	0.1088	0.2055	0.1689	0.1533	0.1088
	NCEmean	0.0608	0.0787	0.0928	0.1390	0.2451	0.1941	0.1598	0.1357
	NCMmean	0.0611	0.0680	0.0877	0.1071	0.2185	0.1768	0.1521	0.1346
2048	NCPmean	0.0359	0.0428	0.0496	0.0577	0.1130	0.1038	0.0950	0.0870
	NCEmean	0.0395	0.0528	0.0632	0.0751	0.1317	0.1188	0.0995	0.0801
	NCMmean	0.0372	0.0435	0.0499	0.0585	0.1389	0.1132	0.0965	0.0818

Table 4.1: The comparison of three methods under 100 simulation runs. MSE obtained for different SNRs (3,5,7,10) and sample sizes n (256,512,1024,2048), choosing the block length $m = 2$.

Part (b):		Bumps				Doppler			
Number	methods	SNR=3	SNR=5	SNR=7	SNR=10	SNR=3	SNR=5	SNR=7	SNR=10
256	NCPmean	0.6137	0.6327	0.6765	0.7228	0.3677	0.3642	0.3505	0.3694
	NCEmean	0.6183	0.7034	0.8190	0.9337	0.3821	0.4059	0.4026	0.4598
	NCMmean	0.5696	0.5762	0.6170	0.6688	0.3876	0.3546	0.3479	0.3656
512	NCPmean	0.4580	0.4254	0.4256	0.4262	0.1891	0.2045	0.1954	0.2117
	NCEmean	0.4934	0.4913	0.5043	0.5215	0.1999	0.2236	0.2264	0.2453
	NCMmean	0.4595	0.4198	0.4105	0.4091	0.2477	0.2206	0.2061	0.2169
1024	NCPmean	0.2493	0.2668	0.2753	0.2853	0.1038	0.1228	0.1250	0.1298
	NCEmean	0.2742	0.2992	0.3188	0.3342	0.1126	0.1452	0.1539	0.1497
	NCMmean	0.3132	0.2885	0.2815	0.2856	0.1622	0.1435	0.1350	0.1312
2048	NCPmean	0.1550	0.1728	0.1783	0.1850	0.0554	0.0658	0.0712	0.0732
	NCEmean	0.1692	0.1950	0.2101	0.2237	0.0589	0.0700	0.0850	0.0911
	NCMmean	0.2207	0.1963	0.1882	0.1883	0.1118	0.0902	0.0823	0.0783

	HeaviSine						Blocks					
	NCPmean			NCPmed			NCPmean			NCPmed		
m\SNR	3	7	10	3	7	10	3	7	10	3	7	10
m=1	0.0508	0.0866	0.1077	0.0546	0.0951	0.1260	0.1667*	0.1248*	0.1163*	0.2069*	0.1199*	0.1014*
m=2	0.0501*	0.0840*	0.1027*	0.0539*	0.0880*	0.1149*	0.1928	0.1574	0.1501	0.2359	0.1535	0.1410
m=4	0.0555	0.0922	0.1169	0.0592	0.0973	0.1272	0.2402	0.2135	0.2090	0.2850	0.2054	0.1954
m=8	0.0596	0.1037	0.1348	0.0604	0.1081	0.1426	0.2909	0.2894	0.2903	0.3354	0.2876	0.2862
m=16	0.0608	0.1143	0.1553	0.0614	0.1163	0.1651	0.3333	0.3816	0.3923	0.3780	0.3889	0.4011
	Bumps						Doppler					
	NCPmean			NCPmed			NCPmean			NCPmed		
m\SNR	3	7	10	3	7	10	3	7	10	3	7	10
m=1	0.2847	0.3120	0.3199	0.3616	0.3831	0.3809	0.1184	0.1498	0.1645	0.1384	0.1769	0.1913
m=2	0.2483*	0.2746*	0.2862*	0.2845*	0.3095	0.3225	0.1025	0.1235	0.1281	0.1108	0.1346	0.1332
m=4	0.2710	0.2943	0.3060	0.2947	0.3077*	0.3219*	0.0957*	0.1160*	0.1260*	0.1008*	0.1204	0.1304*
m=8	0.3233	0.3670	0.3792	0.3461	0.3858	0.3977	0.1051	0.1217	0.1371	0.1070	0.1197*	0.1371
m=16	0.3795	0.4648	0.4814	0.4053	0.4924	0.4935	0.1071	0.1254	0.1374	0.1088	0.1247	0.1389

Table 4.2: Simulation results for NCPmean and NCPmed comparing different block sizes $m = 1, 2, 4, 8, 16$ using 100 simulation runs with sample size $n = 1024$.

	HeaviSine					Blocks				
n	m=1	m=2	m=4	m=8	m=16	m=1	m=2	m=4	m=8	m=16
512	0.1261	0.1223*	0.1287	0.1417	0.1402	0.2084*	0.2462	0.3029	0.3602	0.4606
1024	0.0633	0.0628*	0.0703	0.0786	0.0869	0.1376*	0.1694	0.2238	0.2915	0.3681
2048	0.0436	0.0415*	0.0460	0.0525	0.0596	0.0839*	0.1051	0.1389	0.1870	0.2465
4096	0.0257	0.0246*	0.0270	0.0302	0.0359	0.0412*	0.0537	0.0746	0.1061	0.1489
8192	0.0170	0.0155*	0.0161	0.0186	0.0226	0.0245*	0.0315	0.0437	0.0625	0.0896
	Bumps					Doppler				
n	m=1	m=2	m=4	m=8	m=16	m=1	m=2	m=4	m=8	m=16
512	0.4549	0.3978*	0.4389	0.5354	0.6994	0.2400	0.1912	0.1659*	0.1774	0.2049
1024	0.3064	0.2653*	0.2905	0.3598	0.4394	0.1366	0.1191	0.1059*	0.1112	0.1134
2048	0.1931	0.1708*	0.1820	0.2254	0.2844	0.0798	0.0662	0.0612*	0.0635	0.0613
4096	0.1050	0.0956*	0.0998	0.1232	0.1580	0.0431	0.0338	0.0311	0.0299*	0.0305
8192	0.0588	0.0539*	0.0552	0.0679	0.0881	0.1891	0.0209	0.0187	0.0170	0.0157*

Table 4.3: Simulation results for NCPmean comparing different block sizes $m = 1, 2, 4, 8, 16$ with different sample sizes $n = 512, 1024, 2048, 4096, 8192$ using 100 simulation runs.

In Table 4.4, the new methods are compared with a number of recently proposed methods in the literature. This table shows the new methods to be very competitive with the best of the existing methods, with the best of the former surpassing the best of the latter for two of the four test functions: for the the Bumps signal, NCPMean-2 is best; and in the case of the Doppler signal, NCPMean-4 is best. The table also shows that if the translation-invariant DWT is used with the new methods, then major reductions in MSE can be expected; compare methods 9-12 with 13-16, respectively. See, for example, Percival and Walden (2000) for a discussion of the translation-invariant DWT, referred to there as the MODWT. Note: reductions in MSE of similar order may also be expected to occur if the translation-invariant DWT is used with methods 1-8.

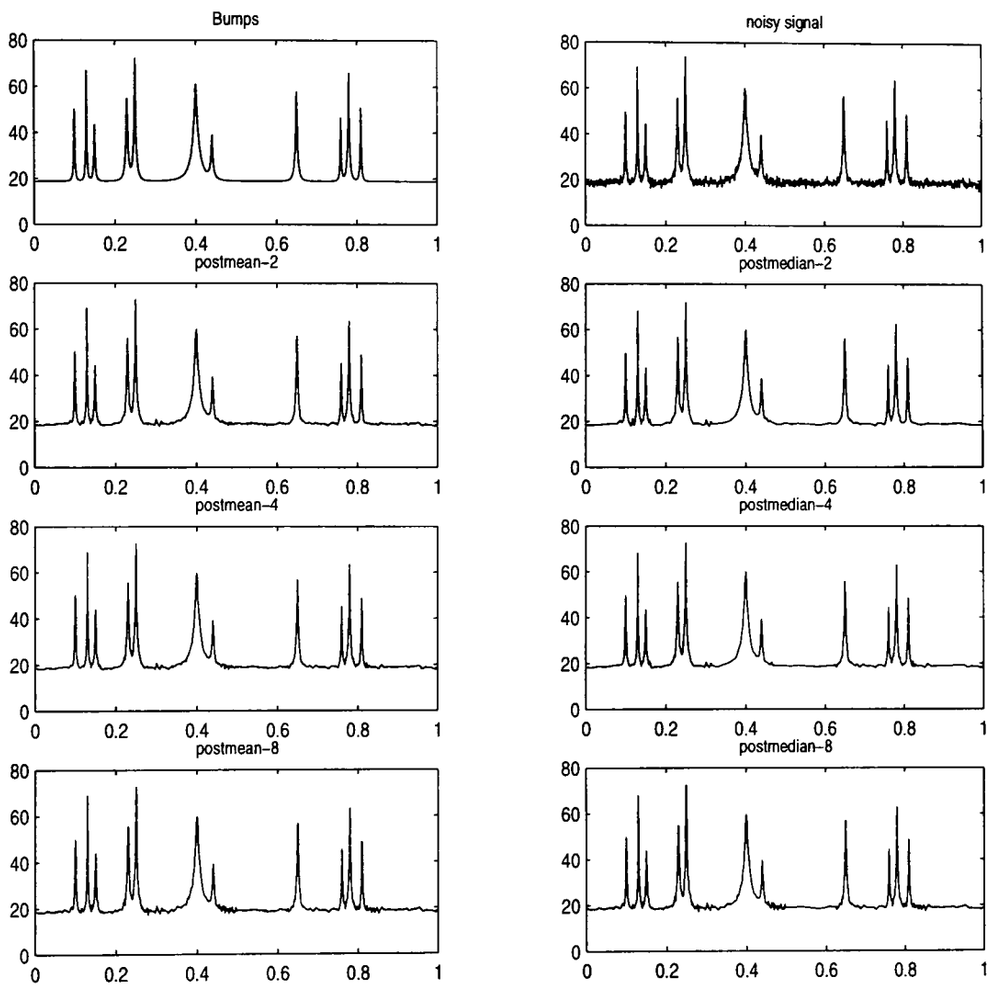

Figure 4.3: The original Bumps function, the Bumps function with noise added and various reconstructions, based on sample size n=1024 and SNR=7. The six reconstructions are obtain using the NCPMean and NCPMed procedures with block sizes 2, 4 and 8.

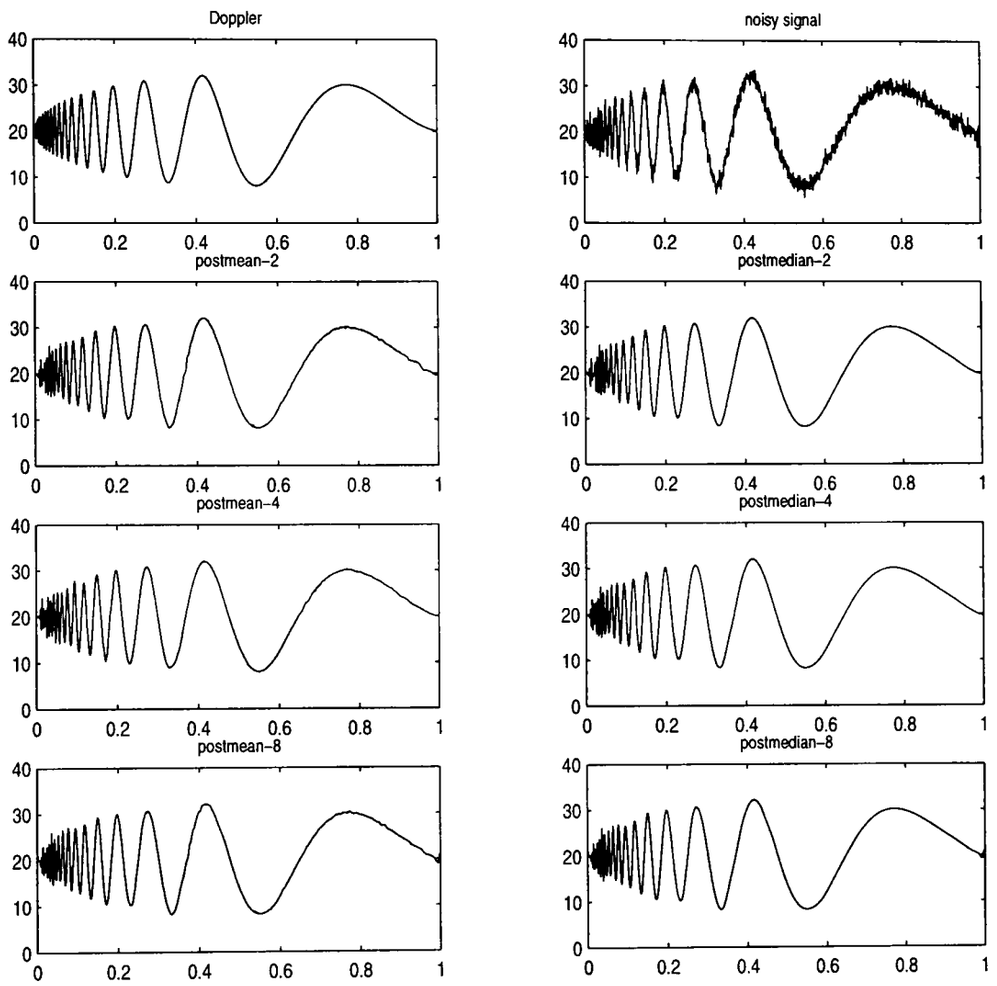

Figure 4.4: The original Doppler function, the Doppler function with noise added and various reconstructions, based on sample size n=1024 and SNR=7. The six reconstructions are obtain using the NCPMean and NCPMed procedures with block sizes 2, 4 and 8.

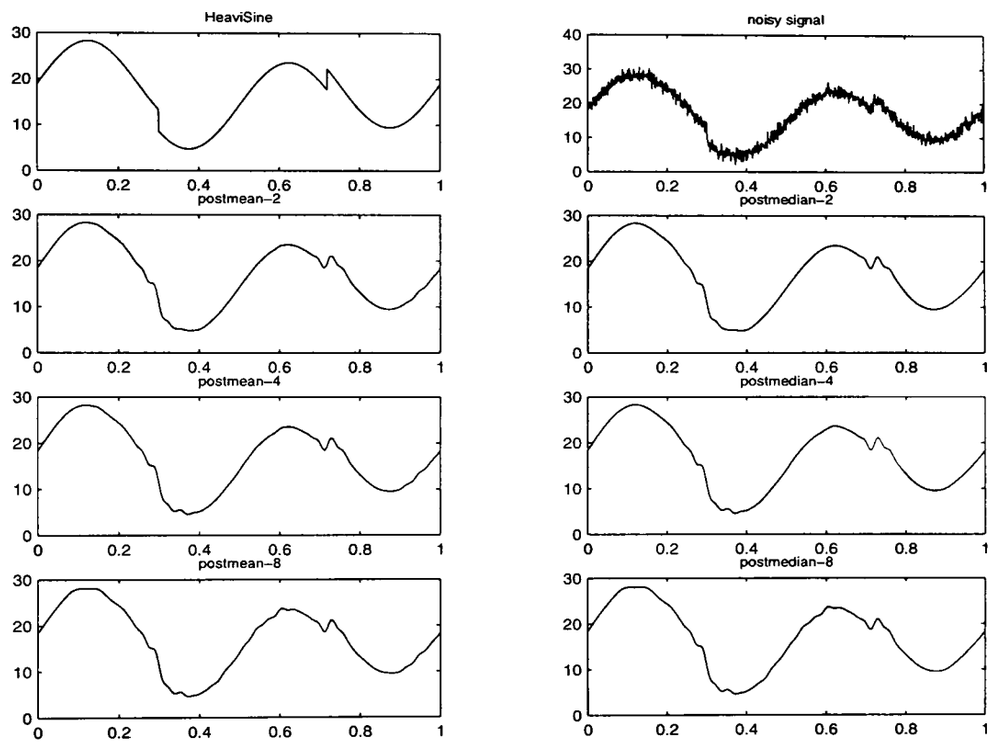

Figure 4.5: The original HeaviSine function, the HeaviSine function with noise added and various reconstructions, based on sample size n=1024 and SNR=7. The six reconstructions are obtain using the NCPMean and NCPMed procedures with block sizes 2, 4 and 8.

	HeaviSine	Blocks	Bumps	Doppler
1. BlockJS	0.1453 (0.1176+0.0276)	0.4593 (0.2797+0.1796)	0.4813 (0.2677+0.2136)	0.1764 (0.1066+0.0698)
2. ABWS	0.0874 (0.0433+0.0442)	0.0995 (0.0121+0.0874)	0.3495 (0.1267+0.2228)	0.1646 (0.0640+0.1006)
3. BAMS	0.0815 (0.0304+0.0511)	0.1107 (0.0142+0.0965)	0.3404 (0.1428+0.1976)	0.1482 (0.0584+0.0899)
4. BBS	0.0860 (0.0394+0.0466)	0.2034 (0.0061+0.1973)	0.2961 (0.0373+0.2588)	0.1185 (0.0288+0.0897)
5. EBTCMean	0.0810 (0.0317+0.0494)	0.3225 (0.1094+0.2131)	0.3589 (0.1052+0.2537)	0.1534 (0.0557+0.0977)
6. EBTCMed	0.0860 (0.0400+0.0460)	0.3537 (0.1379+0.2159)	0.3840 (0.1374+0.2467)	0.1637 (0.0675+0.0958)
7. EBTLMean	0.0816 (0.0304+0.0513)	0.3263 (0.1045+0.2217)	0.3744 (0.1100+0.2644)	0.1606 (0.0549+0.1057)
8. EBTLMed	0.0857 (0.0363+0.0495)	0.3522 (0.1315+0.2207)	0.3908 (0.1347+0.2562)	0.1665 (0.0666+0.0999)
9. NCPmean-2	0.0836 (0.0414+0.0422)	0.1559 (0.0048+0.1511)	0.2701 (0.0422+0.2280)	0.1238 (0.0305+0.0933)
10. NCPmed-2	0.0885 (0.0504+0.0381)	0.1528 (0.0286+0.1242)	0.3087 (0.1073+0.2015)	0.1390(0.0509+0.0881)
11. NCPmean-4	0.0926 (0.0432+0.0494)	0.2103 (0.0041+0.2062)	0.2917 (0.0320+0.2597)	0.1143 (0.0256+0.0887)
12. NCPmed-4	0.0973 (0.0493+0.0480)	0.2023 (0.0153+0.1870)	0.3074 (0.0652+0.2422)	0.1196 (0.0352+0.0844)
13. TINCPmean-2	0.0666 (0.0294+0.0372)	0.1127 (0.0006+0.1121)	0.2228 (0.0150+0.2077)	0.0968 (0.0176+0.0793)
14. TINCPmed-2	0.0677 (0.0372+0.0305)	0.0796 (0.0043+0.0753)	0.2107 (0.0521+0.1586)	0.0857 (0.0245+0.0612)
15. TINCPmean-4	0.0752 (0.0321+0.0431)	0.1556 (0.0015+0.1541)	0.2530 (0.0141+0.2389)	0.0885 (0.0103+0.0782)
16. TINCPmed-4	0.0748 (0.0368+0.0380)	0.1237 (0.0024+0.1213)	0.2298 (0.0281+0.2017)	0.0779 (0.0139+0.0660)

Table 4.4: The Comparison of 16 methods using 1000 simulation runs with n=1024, SNR=7, and signals HeaviSine, Blocks, Bumps and Doppler. For further details of the methods BlocksJS, ABWS, BAMS and BBS see Cai (1999), Chipman et al (1997), Vidakovic and Ruggeri (2001) and De Canditiis and Vidakovic (2004), respectively; the next four methods are variants of EBayesThresh due to Johnstone and Silverman (2005) based on the posterior mean or median, using the Cauchy or Laplace prior, in obvious notation; methods 9–12 are variants of the new methods using the DWT; and methods 13–16 are variants of 9–12 respectively in which the translation-invariant (TI) DWT is used.

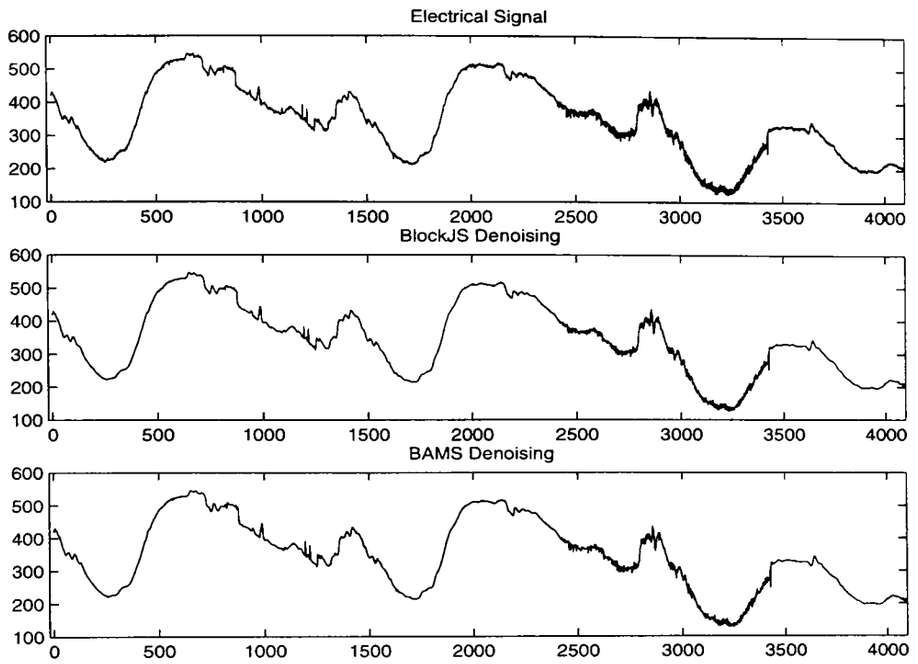

Figure 4.6: Electrical consumption signal and denoising results by BlockJS and BAMS

4.7.2 An Electrical Consumption Example

We further illustrate the performance of the above procedures using an electrical consumption signal measured over a 5-week period (Antoniadis *et al.*, 2001). Figure 4.6(a) shows a selection of the electrical signal, sampled minute by minute over a 3-day period. The noise is introduced whenever a defect occurs in the monitoring equipment. The assumption that the noise is IID is rather doubtful in this example because of its time dependent structure. Nevertheless, it is interesting to see that the proposed method provides sensible and useful results in this example. The denoising results by BlockJS and BAMS procedures are also presented in Figure 4.6.

Figure 4.7 gives the denoising results using NCMmean, NCEmean and NCPmean. It appears that the methods proposed produce a smoother fit, compared with denoising results by BlockJS and BAMS. In this example, the new methods appear to be able to remove the noise more effectively.

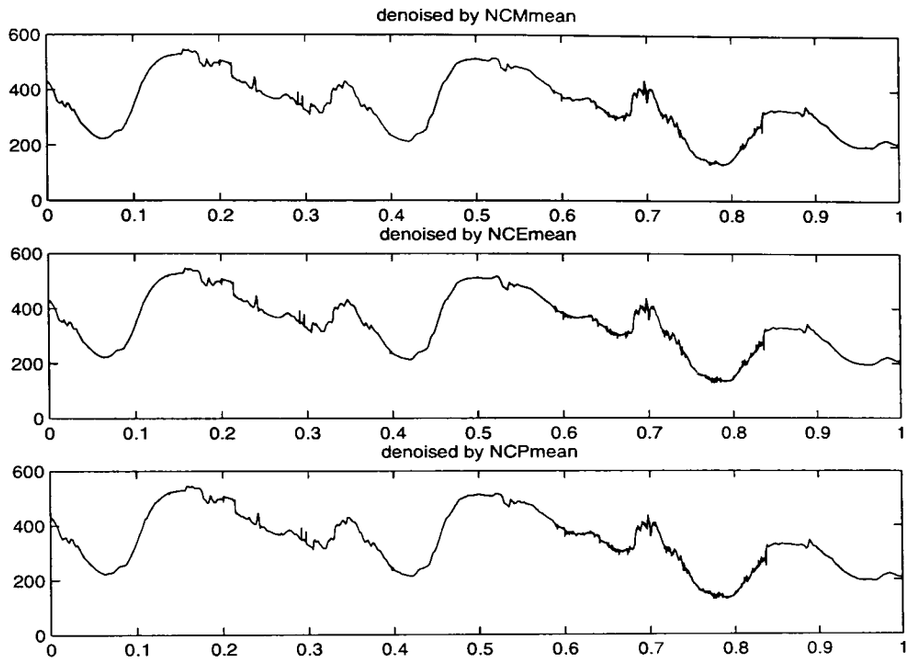

Figure 4.7: Denoising results by NCMmean, NCEmean and NCPmean for the electrical consumption signal.

4.8 Denoising Planar Curves

As an extension of the application of the EBB method mentioned above, the planar curve problem will be investigated. How to recover a contaminated planar curve is essential in shape description and recognition. Furthermore, for most curve matching or shape recognition tasks, it is important that the methods are invariant with respect to rotation, translation.

The problem we consider is that of denoising a noisy closed curve in the plane. A new wavelet estimator based on a sum of squares of empirical wavelet coefficients of two parametric coordinate functions is introduced and the EBB method is then applied to this estimator. The closed planar curve model we consider is given in (4.18). In §4.8.2 the implementation of the approach for calculating the new estimator is explained. The equivariance of this method with respect to translation, rotation and isotropic scaling is discussed in §4.8.3. A simulation study is provided in §4.8.4, giving comparisons of the proposed methods with some existing methods.

4.8.1 The Model

Planar curves can be expressed in terms of two parametric coordinate functions where the two coordinate functions depend on the same parameter as follows:

$$C = \left\{ \begin{pmatrix} f_1(t) \\ f_2(t) \end{pmatrix} : t \in [0, l] \right\} \subset \mathbf{R}^2, \tag{4.18}$$

where $f_i : [0, l] \longrightarrow \mathbf{R}, \quad i = 1, 2$, are continuous, real-valued coordinate functions. If $f_i(0) = f_i(l), \quad i = 1, 2$, then C represents a closed curve in $\mathbf{R}^2$.

In this section it is assumed that a noisy version of C is observed at discrete points $t_i = (l/n) \cdot i$, $i = 1, \ldots, n$, and that the following model is appropriate:

$$\begin{pmatrix} y_{1i} \\ y_{2i} \end{pmatrix} = \begin{pmatrix} f_1(t_i) \\ f_2(t_i) \end{pmatrix} + \sigma^2 \begin{pmatrix} \epsilon_{1i} \\ \epsilon_{2i} \end{pmatrix}, \quad i = 1, \ldots, n, \tag{4.19}$$

where the vector $(\epsilon_{1i}, \epsilon_{2i})^T$ is IID with $N(0, I_{2\times 2})$. It is more convenient if $n = 2^{J+1}$. But if n is not an integer power of 2, the TIDWT can be used.

The method of this section can also apply to non-closed planar curves. However, in this case, boundary problems at the end-points of the curves need to be addressed.

4.8.2 The Denoising Procedure

As parameterized closed curves can be represented by periodic sequences, a corresponding periodic discrete wavelet transformation (Daubechies 1992) that takes observations $(y_{1i}, y_{2i})^T$, $i = 1, \ldots, n$, to the wavelet space will be used. This process can be represented by an orthogonal matrix $\mathcal{W} = (\mathcal{W}_0^T, \mathcal{V}_{j_0}^T)^T$, which yields the relations

$$\begin{aligned} \mathcal{W}\mathbf{y}_j &= \begin{pmatrix} \mathcal{W}_0 \mathbf{y}_j \\ \mathcal{V}_{j_0} \mathbf{y}_j \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{d}}_j \\ \tilde{\mathbf{c}}_j \end{pmatrix} \\ &= \begin{pmatrix} \mathcal{W}_0 \mathbf{f}_j \\ \mathcal{V}_{j_0} \mathbf{f}_j \end{pmatrix} + \sigma^2 \begin{pmatrix} \mathcal{W}_0 \boldsymbol{\epsilon}_j \\ \mathcal{V}_{j_0} \boldsymbol{\epsilon}_j \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{d}_j \\ \mathbf{c}_j \end{pmatrix} + \sigma^2 \begin{pmatrix} \boldsymbol{\epsilon}_j \\ \boldsymbol{\epsilon}_j \end{pmatrix} \quad j = 1, 2, \end{aligned} \tag{4.20}$$

where $\mathbf{y}_j = (y_{j1}, \ldots, y_{jn})^T$, $j = 1, 2$, and, in similar notation, $\mathbf{f}_j$ is the column vector of signal, $\tilde{\mathbf{d}}_j$, $\mathbf{d}_j$ and $\boldsymbol{\epsilon}_j$ are wavelet coefficient vectors of noisy data, signal and noise variable respectively.

Under (4.21), we have

$$\tilde{d}_{ji} \sim N(d_{ji}, \sigma^2), \quad j = 1, 2, \tag{4.21}$$

where $i = 1, \ldots, n-2^{j_0}$ and j_0 is the coarsest level. In order to keep the equivariance with respect to translation and rotation, the following transform of the wavelet coefficients is performed: $\rho_i = d_{1i}^2 + d_{2i}^2$ and $z_i = \tilde{d}_{1i}^2 + \tilde{d}_{2i}^2$. Let $\mathcal{B}_{\sigma^2,\theta}(z)$ denote ρ_{mean}, ρ_{med} or ρ_{hyp} defined in §3.4, the shrinkage step mentioned in §2.5.1 can be expressed as

$$\begin{pmatrix} \widehat{d}_{1i} \\ \widehat{d}_{2i} \end{pmatrix} = \left\{ \frac{\mathcal{B}_{\sigma^2,\theta}(z_i)}{z_i} \right\}^{1/2} \begin{pmatrix} \tilde{d}_{1i} \\ \tilde{d}_{2i} \end{pmatrix}. \tag{4.22}$$

We can see that (4.22) is a straightforward extension of (4.2) with the block size $m = 2$.

An alternative way to obtain the denoised data $(\widehat{y}_{1i}, \widehat{y}_{2i})^T$ is by applying the methods outlined in the previous section to $f_1(\cdot)$ and $f_2(\cdot)$ one by one. However, there is a major drawback of this approach: the method is not equivariant with respect to rotations of the planar curves.

4.8.3 Equivariance

In this section we will show that the shrinkage procedure of the planar curve provided in § 4.8.2 achieves equivariance.

For the standard planar curve model defined in (4.19), we transform coordinates as follows:

$$\begin{pmatrix} \mathbf{y}_1^{*T} \\ \mathbf{y}_2^{*T} \end{pmatrix} = \begin{pmatrix} a_1 \mathbf{1}_n{}^T \\ a_2 \mathbf{1}_n{}^T \end{pmatrix} + bR \begin{pmatrix} \mathbf{y}_1{}^T \\ \mathbf{y}_2{}^T \end{pmatrix}, \tag{4.23}$$

where a_1, a_2 and $b > 0$ are constants, $\mathbf{1}_n = (1, \cdots, 1)^T$ is the n-vector of ones and

$$R = \begin{pmatrix} \cos(\vartheta) & -\sin(\vartheta) \\ \sin(\vartheta) & \cos(\vartheta) \end{pmatrix}.$$

Let $(\hat{\mathbf{f}}_1, \hat{\mathbf{f}}_2)$ denote an estimator of $(\mathbf{f}_1, \mathbf{f}_2)$ based on the noisy data $(\mathbf{y}_1, \mathbf{y}_2)$ and $(\hat{\mathbf{f}}_1^*, \hat{\mathbf{f}}_2^*)$ denote the estimator of $(\mathbf{f}_1^*, \mathbf{f}_2^*)$ based on the transformed data $(\mathbf{y}_1^*, \mathbf{y}_2^*)$.

Definition 4.2 *The estimation procedure is said to be equivariant if*

$$\begin{pmatrix} \hat{\mathbf{f}}_1^{*T} \\ \hat{\mathbf{f}}_2^{*T} \end{pmatrix} = \begin{pmatrix} a_1 \mathbf{1}_n{}^T \\ a_2 \mathbf{1}_n{}^T \end{pmatrix} + bR \begin{pmatrix} \hat{\mathbf{f}}_1{}^T \\ \hat{\mathbf{f}}_2{}^T \end{pmatrix}. \tag{4.24}$$

Following this definition, the three step procedure in the original coordinates and transformed coordinates can be expressed as follows.

Original Coordinates

Step 1 Obtain the empirical wavelet coefficients

$$\mathcal{W}\mathbf{y}_1 = \begin{pmatrix} \mathcal{W}_0\mathbf{y}_1 \\ \mathcal{V}_{j_0}\mathbf{y}_1 \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{d}}_1 \\ \tilde{\mathbf{c}}_1 \end{pmatrix};$$

$$\mathcal{W}\mathbf{y}_2 = \begin{pmatrix} \mathcal{W}_0\mathbf{y}_2 \\ \mathcal{V}_{j_0}\mathbf{y}_2 \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{d}}_2 \\ \tilde{\mathbf{c}}_2 \end{pmatrix}.$$

Step 2 Define $z_i = \tilde{d}_{1i}^2 + \tilde{d}_{2i}^2$, $i = 1, \ldots, n - 2^{j_0}$, and then obtain the adjusted $\hat{d}_{1i}$ and $\hat{d}_{2i}$ by using (4.22).

Step 3 Estimate $\mathbf{f}_1$ and $\mathbf{f}_2$ by

$$\hat{\mathbf{f}}_1 = \mathcal{W}^T \begin{pmatrix} \hat{\mathbf{d}}_1 \\ \tilde{\mathbf{c}}_1 \end{pmatrix};$$

$$\hat{\mathbf{f}}_2 = \mathcal{W}^T \begin{pmatrix} \hat{\mathbf{d}}_2 \\ \tilde{\mathbf{c}}_2 \end{pmatrix}.$$

Transformed Coordinates

Step 1* Obtain

$$\mathcal{W}\mathbf{y}_1^* = \begin{pmatrix} \mathcal{W}_0\mathbf{y}_1^* \\ \mathcal{V}_{j_0}\mathbf{y}_1^* \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{d}}_1^* \\ \tilde{\mathbf{c}}_1^* \end{pmatrix};$$

$$\mathcal{W}\mathbf{y}_2^* = \begin{pmatrix} \mathcal{W}_0\mathbf{y}_2^* \\ \mathcal{V}_{j_0}\mathbf{y}_2^* \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{d}}_2^* \\ \tilde{\mathbf{c}}_2^* \end{pmatrix}.$$

Step 2* Define $z_i^* = (\tilde{d}_{1i}^*)^2 + (\tilde{d}_{2i}^*)^2$, $i = 1, \ldots, n - 2^{j_0}$, and then obtain the adjusted $\hat{d}_{1i}^*$ and $\hat{d}_{2i}^*$ by using (4.22).

Step 3* Estimate $\mathbf{f}_1^*$ and $\mathbf{f}^*{}_1$ by

$$\hat{\mathbf{f}}_1^* = \mathcal{W}^T \begin{pmatrix} \hat{\mathbf{d}}_1^* \\ \tilde{\mathbf{c}}_1^* \end{pmatrix};$$

$$\hat{\mathbf{f}}_2^* = \mathcal{W}^T \begin{pmatrix} \hat{\mathbf{d}}_2^* \\ \tilde{\mathbf{c}}_2^* \end{pmatrix}.$$

Proposition 4.2 *In the setting of Proposition 4.1, the estimation procedure is equivariant, i.e. (4.24) holds.*

Proof of Proposition 4.2: The wavelet transform steps (Step 1 and 3 or Step 1* and 3*) achieve equivariance with respect to translation and rotation (see Chuang and Kuo, 1996). From the proof of Proposition 4.1, we know that $z_i^* = (\tilde{d}_{1i}^*)^2 + (\tilde{d}_{2i}^*)^2 = b^2 z_i$ and $\mathcal{B}_{\sigma^2,\theta}(z_i^*) = \mathcal{B}_{\sigma^2,\theta}(b^2 z_i) = b^2 \mathcal{B}_{\sigma^2,\theta}(z_i)$. Therefore,

$$\frac{\mathcal{B}_{\sigma^2,\theta}(z_i^*)}{z_i^*} = \frac{\mathcal{B}_{\sigma^2,\theta}(z_i)}{z_i}. \tag{4.25}$$

Since from (4.23), we have

$$\begin{pmatrix} \tilde{d}_{1i}^* \\ \tilde{d}_{2i}^* \end{pmatrix} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} + bR \begin{pmatrix} \tilde{d}_{1i} \\ \tilde{d}_{2i} \end{pmatrix}, \tag{4.26}$$

$$\begin{pmatrix} \widehat{d}_{1i} \\ \widehat{d}_{2i} \end{pmatrix} = \left\{ \frac{\mathcal{B}_{\sigma^2,\theta}(z_i)}{z_i} \right\}^{1/2} \begin{pmatrix} \tilde{d}_{1i} \\ \tilde{d}_{2i} \end{pmatrix}$$

and

$$\begin{pmatrix} \widehat{d}_{1i}^* \\ \widehat{d}_{2i}^* \end{pmatrix} = \left\{ \frac{\mathcal{B}_{\sigma^2,\theta}(z_i^*)}{z_i^*} \right\}^{1/2} \begin{pmatrix} \tilde{d}_{1i}^* \\ \tilde{d}_{2i}^* \end{pmatrix},$$

it follows from (4.25) that $(\widehat{d}_{1i}^*, \widehat{d}_{2i}^*)^T$ and $(\widehat{d}_{1i}, \widehat{d}_{2i})^T$ also satisfy the translation rule (4.26). The remainer of the proof is similar to that of Proposition 4.1.

4.8.4 Simulation Results

The purpose of this subsection is to illustrate the practical performance of the proposed Bayesian approaches. The following planar curve, expressed in parametric form, was used in the simulation study:

$$y_{1i} = f_1(t_i) = ((15 + 10\cos(7t_i))\cos(t_i) - sgn(t_i + 1) - sgn(1 - t_i)$$

$$y_{2i} = f_2(t_i) = ((15 + 10\sin(7t_i))\sin(t_i) - sgn(t_i + 1) - sgn(1 - t_i)$$

where $t_i = -\pi + (\pi/512)i$, $i = 0, 1, ...1023$ and $sgn(t) = -1, 0$ or 1 depending on whether t is negative, zero or positive. For this pair of functions, 1024 pairs of points

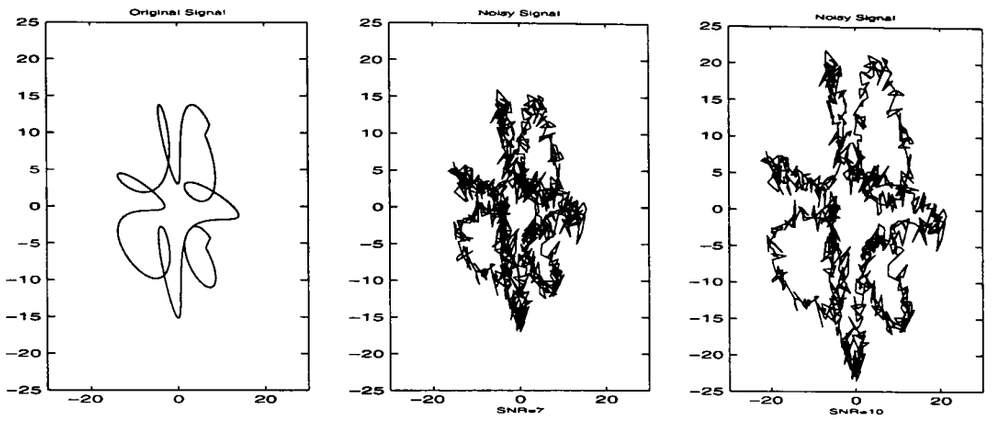

Figure 4.8: The simulated function based on 1024 signal points and the noisy function with SNR=7 and SNR=10

are generated by adding independent random noise $(\epsilon_{1i}\ \epsilon_{2i})^T \sim N(0,\ \sigma^2 I_{2\times 2})$. The value of the rescaled functions are obtained as follows

$$f_{scaled_i} = f_i * (\frac{SNR}{std(f)}).$$

where

$$std(f) = \sqrt{\frac{1}{n-1}\Big\{(f_1(t_i) - \overline{f}_1)^2 + (f_2(t_i) - \overline{f}_2)^2\Big\}}$$

and $\overline{f}_j = (1/n)\sum_{j=1}^{n} f_j(t_i)$, $j = 1, 2$. In simulation study, SNR=7 and SNR=10 are chosen. Daubechies compactly supported periodized wavelets are used. The average MSE for the estimator $(\widehat{f}_1,\ \widehat{f}_2)$ of $(f_1,\ f_2)$ is defined as

$$MSE_f = \frac{1}{n}\sum_{i=1}^{n}\Big[\big\{\widehat{f}_1(t_i) - f_1(t_i)\big\}^2 + \big\{\widehat{f}_2(t_i) - f_2(t_i)\big\}^2\Big] \qquad (4.27)$$

As shown in Figure 4.8, the original signals and noisy signals with SNR=7 and SNR=10 are given. The posterior mean and median methods with the power prior (3.36), NCPmean and NCPmed respectively, specified in Chapter 3, and the "quick-and-dirty" method of § 4.3 to estimate the hyperparameters are used. Simulation results are given in Figure 4.9.

SingleMean (Clyde and George, 1999, 2000) and SingleMed (Abromovich *et al.*, 1998) are methods which are used here to compare with the NCPmean and NCPmed methods. The original programs of SingleMean and SingleMed are from **Matlab**

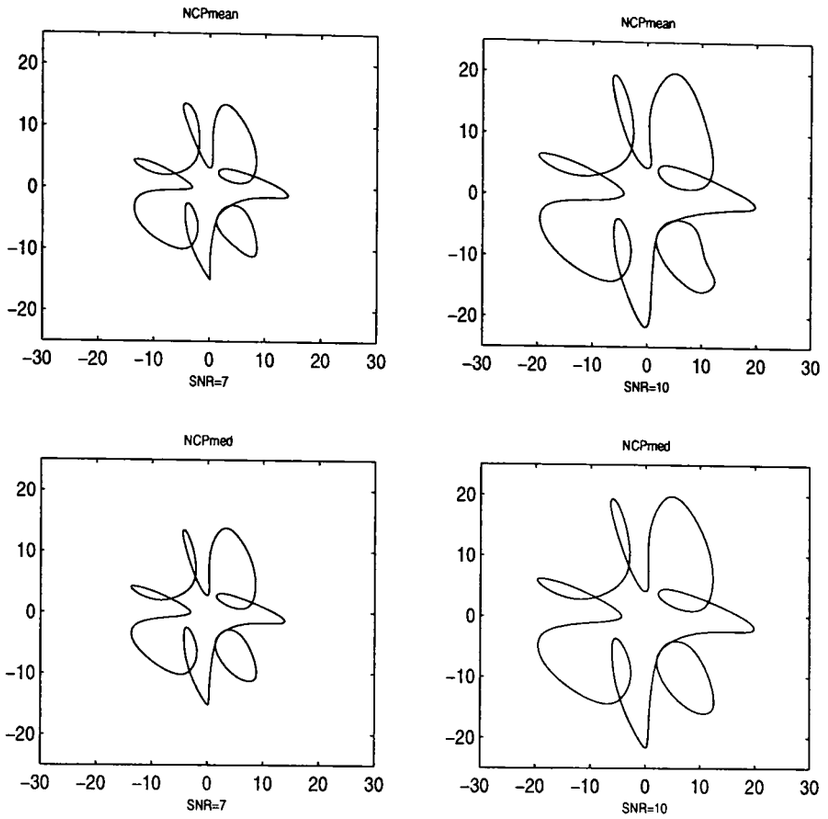

Figure 4.9: NCPmean (top) and NCPmed (bottom) with SNR=7 and SNR=10

package **Waveden** provided by Antoniadis *et al.* (2001). These two methods are applied to the simulated noisy data $\mathbf{y}_1$ and $\mathbf{y}_2$ one by one. Then the average MSEs are calculated by (4.27).

These four methods of reconstruction (NCPmean, NCPmed, SingleMean and SingleMed) with sample size n=1024 and SNR=10 are provided in Figure 4.10 and the average MSEs with mean squared biases (MSB) and variances (Var) over 1000 simulations with SNR=7 and SNR=10 are given in Table 4.5. NCPmean gives the best reconstruction in terms of the average MSEs for both SNR=7 and SNR=10 although the improvement is not dramatic. However, NCPmean and NCPmed methods can be easily applied to higher dimension planar curves, and still remain the equivariance property, which is desirable.

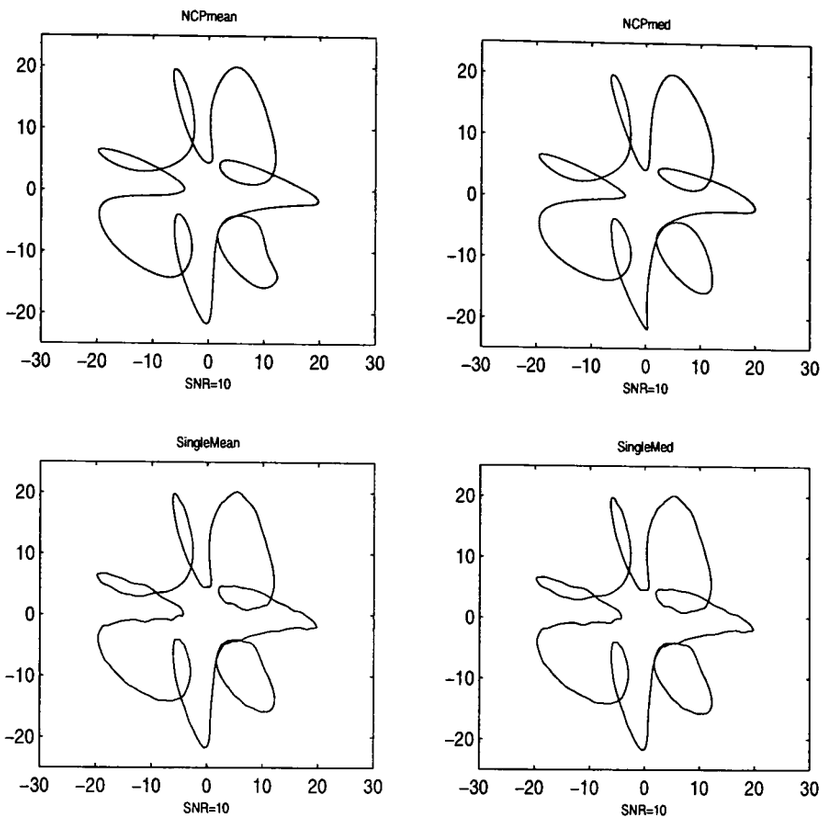

Figure 4.10: Reconstructions of the noisy function based on 1024 signal points with SNR=10, from left to right, NCPmean and NCPmed (top), SingleMean and SingleMed (bottom)

Methods	MSE (MSB+Var)	MSE (MSB+Var)
	SNR=7	SNR=10
NCPmean	0.1065 (0.0334+0.0731)	0.0889 (0.0227+0.0662)
NCPmed	0.1099 (0.0383+0.0716)	0.0893 (0.0239+0.0654)
SingleMean	0.1201 (0.0271+0.0930)	0.1025 (0.0174+0.0851)
SingleMed	0.1192 (0.0293+0.0899)	0.1005 (0.0186+0.0819)

Table 4.5: The comparison of the four methods based on 1024 sample points with SNR=7 and SNR=10 over 1000 simulation runs.

4.9 Conclusions and Further Work

An empirical Bayes approach for block shrinking of wavelet coefficients using the block sum of squares has been proposed. A simulation study was undertaken to investigate the performance of this new approach, and the new methods were compared with both non-Bayesian block wavelet thresholding estimators and with various Bayesian estimators. An application of the proposed methods to a practical data set has also been presented.

Our results indicate that the new methods perform well and are competitive with the existing methods. Our results so far suggest that NCPmean is the best of the new procedures, though not by a big margin.

In conclusion, we present a number of comments.

1. It has been demonstrated that, with an appropriate choice of block length (block lengths 2 and 4 have worked well in our examples), the proposed methods are competitive with several existing methods, for example ABWS and BlockJS. However, there is scope for further study of the effects of varying block length in the framework considered here. We note, however, that implementation of the proposals does not become more difficult when block length is increased.

2. The "quick and dirty" method for estimating the level-independent hyperparameter λ described in section 4.3 is computationally faster than most of the published Bayesian and EB methods and produces good results.

3. It is noted in Johnstone and Silverman (2005) that shrinkage based on the mean does better in terms of MSE than shrinkage based on the median when the standard DWT is used, but the reverse holds when the translation-invariant, or stationary, DWT is used. We note that our numerical results conform to their finding: when we used the standard DWT, the mean outperformed the median in the case of the "power" prior (the only prior for which we have so

far implemented median-based shrinkage), but the median tended to provide the best results when the translation-invariant DWT was used.

4. A simulation study investigating the proposed methods in the planar curve application was described. The equivariance of these methods with respect to translation, rotation and isotropic scaling is a desirable property, though in the simulation study the improvement was not dramatic.

Chapter 5

Estimation of Covariance Parameters in Wavelet Regression with Correlated Noise

5.1 Introduction

Various Bayesian approaches for thresholding and non-linear shrinkage of wavelet coefficients have been proposed and shown to perform well under the IID Gaussian noise assumption. However, in realistic situations noise is often correlated, sometimes highly correlated. To extend the existing methods to handle this situation, a semi-parametric approach for estimating the unknown function f in the presence of correlated noise ϵ has been developed. This approach has been investigated using a simulation study, and numerical results indicate that the proposed method does a good job of reconstructing the signal even with highly correlated data.

The outline of this chapter is as follows. In § 5.2, the model considered in this chapter is given and numerical results show the necessity of accounting for the correlation in the noise. A semi-parametric model is explored in § 5.3. This model includes a parametric part and a nonparametric part. In the parametric part, given the covariance structure of correlated noise, two distinct parametric

approaches to estimate the parameters in the covariance structure are provided . In the nonparametric part, after the parameters are estimated, the EBB method proposed in Chapters 3 and 4 can be used to reconstruct the function. A simulation study is undertaken and graphics are presented in § 5.4.

5.2 A Model with Correlated Noise

5.2.1 Model and Notation

The model to be considered in this chapter is

$$y_i = f(x_i) + \epsilon_i \qquad i = 1, \cdots, n \tag{5.1}$$

where f is the unknown function to be estimated, $\{y_i\}$ is a set of observations, $x_i = i/n$ and $\{\epsilon_i\}$ is a stationary correlated sequence.

When introducing the semi-parametric approach, we allow the error sequence to be a general stationary Gaussian sequence with known covariance structure. Later, in the simulation study, we only consider the $\{\epsilon_i\}$ modelled by a zero-mean stationary Gaussian autoregressive process or a moving average process of orders p and q, (especially, we consider $p, q = 1, 2$ here), which are given by

$$AR(p) \qquad \epsilon_t = \alpha_1 \epsilon_{t-1} + \cdots + \alpha_p \epsilon_{t-p} + \eta_t, \tag{5.2}$$

or

$$MA(q) \qquad \epsilon_t = \beta_1 \eta_{t-1} + \cdots + \beta_q \eta_{t-q} + \eta_t, \tag{5.3}$$

where η_t is independent $N(0, \sigma^2)$. Given that we are assuming the $\{\epsilon_i\}$ is stationary and Gaussian, there is little loss of generality in restricting attention to autoregressive process as most Gaussian stationary processes can be approximated by an autoregressive process of sufficiently high order (see e.g. Brockwell and Davis, 1991).

5.2.2 An Example

The only difference between the standard model (2.13) and the model (5.1) above is the different assumption about the noise ϵ: independent $N(0, \sigma^2)$ noise in the former case and autocorrelated noise in the latter case. Since the properties of the wavelet transform show that wavelets are "almost eigenfunctions" of many operators (see Frazier *et al.*, 1991, Meyer, 1992), which means that the autocorrelation of the wavelet coefficients of a noisy signal within each level dies away rapidly and little or no correlation between the wavelet coefficients at different levels exits (see Johnstone and Silverman, 1997), one may ask if it is possible to get away with using standard methods (i.e. methods designed for the standard model (2.13)) in the correlated data situation (5.1).

We will use a numerical example to answer this question. For the standard test function "HeaviSine" (Donoho and Johnstone 1994, 1995), three types of noise are added: IID normal, AR(1) with $\alpha = 0.3$ and AR(1) with $\alpha = 0.7$. For correlated noise (AR(1) noise with $\alpha = 0.3, 0.7$) cases, we normalise the noise by using

$$\epsilon_{normalised} = \frac{\epsilon}{std(\epsilon)} \tag{5.4}$$

and rescale the signal for all three situations as usual by using

$$f_{scaled} = f * (\frac{SNR}{std(f)}), \tag{5.5}$$

where *std* is the standard deviation of the test function or noise. The signal-to-noise ratio (SNR) equals 7 in each level. Four methods, BlockJS, BAMS, EBTCMean and NCPmean-2 (see Cai, 1999, Vidakovic and Ruggeri, 2001, Johnstone and Silverman, 2005 and posterior mean method of non-central χ^2 distribution with block size $m = 2$ mentioned in Chapters 3 and 4 of this thesis), which we used for comparison in Chapter 4, will be considered here. Reconstructions of the signal from three types of noise are given in Figure 5.1. Clearly, all methods can give the reasonably well denoised results in the IID noise case. However, in the correlated noise cases, we can still see lots of wiggles in the reconstructed functions, especially for AR(1) with $\alpha = 0.7$.

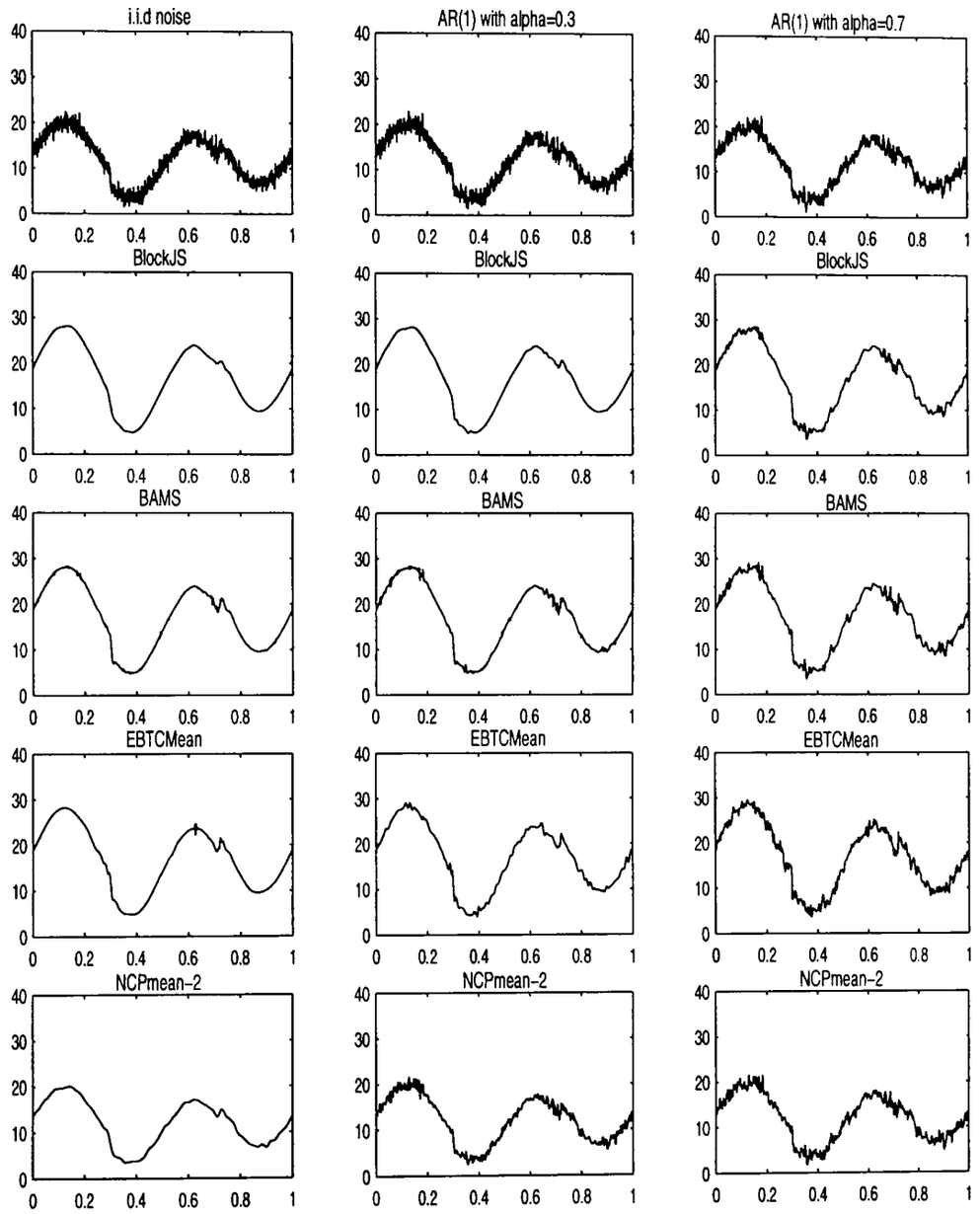

Figure 5.1: Denoising three noisy signals with SNR=7 using four denoising methods: BlockJS, BAMS, EBTCMean and NCPmean-2

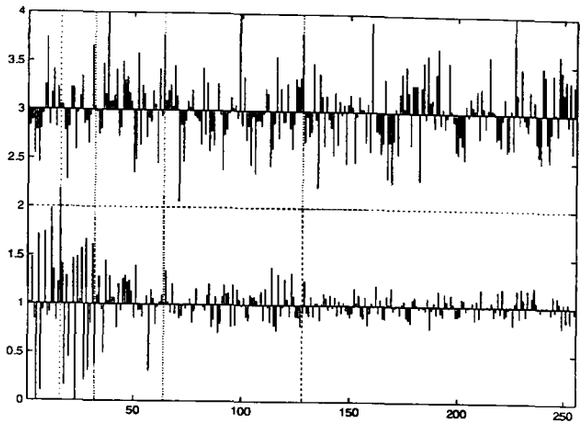

Figure 5.2: Wavelet coefficients of IID noise (above) and AR(1) noise (bottom)

More extensive study of numerical examples shows that if standard methods are used on correlated data, it can seriously affect the quality of the reconstruction of f, particularly when the noise are highly correlated. Thus there is a clear need to account for correlation in the noise when it is present.

Furthermore, Figure 5.2 plots the wavelet coefficients of IID noise (above) and AR(1) noise with $\alpha = 0.7$ (bottom). The difference between two groups of wavelet coefficients at each level shows clearly: wavelet coefficients of IID noise are still IID at each level while those of correlated noise are definitely not IID. Thus there is a need to have a close look at the dependence structure of wavelet coefficients of correlated data.

5.2.3 Analysis of Existing Work

In published work, some consideration has been given to modifications of shrinkage and thresholding procedures to deal with the correlated noise situation. Johnstone and Silverman (1997) pointed out that if the noise in the data is stationary and correlated, then the variance of the wavelet coefficients $\tilde{d}_{jk}$ will depend on the level in the wavelet decomposition but will be constant at each level. A natural extension of the standard wavelet thresholding method is to apply level-dependent thresholding to each level of wavelet coefficients after the wavelet transform. A specific example

was proposed by Johnstone and Silverman (1997). Let λ_j be a sequence of thresholds to be applied to the coefficients $\tilde{d}_{jk}$ at level j and $\widehat{d}_{jk}$ be the estimator,

$$\widehat{d}_{jk} = \eta(\tilde{d}_{jk}, \sigma_j \lambda_j).$$

Hence η denotes soft or hard thresholding, or some compromise between the two; the noise variance σ_j^2 at each level can be estimated from the data and one possibility is to use a robust estimator such as

$$\widehat{\sigma}_j = MAD(d_{jk}, k = 0, \cdots, 2^j - 1)/0.6745,$$

where MAD is the median absolute deviation, and a conservative choice (in the sense of tending to oversmooth) of threshold from certain theoretical perspectives is

$$\widehat{\lambda}_j = \sqrt{2 \log n}.$$

This is a quick and convenient way to cope with the problem of adaptive estimation in correlated noise, especially under a wide range of possible forms of correlation, but it is not a final answer to this question. In the results we give later, we will see that this method does not cope well with rough signals.

Another possibility is to specify the covariance structure of the wavelet coefficients at each level, which allows various forms of dependency between the wavelet coefficients. For example, Abramovich *et al.* (2002) specified the covariance matrix V_j of each block of l_j wavelet coefficients to be an $l_j \times l_j$ matrix with elements

$$V_j[k, l] = \tau_j^2 \rho_j^{|k-l|} \quad \text{where} \quad |\rho| < 1, \quad k, l = 1, \cdots, l_j. \tag{5.6}$$

Although correlated data were discussed in some of these papers, more attention was paid to the resulting covariance structure of wavelet coefficients. As far as we know, the specific forms of covariance structure were given according to prior knowledge about the characteristic of wavelet coefficients. For example, (5.6) was chosen because it was believed that as the distance between two wavelet coefficients increases the correlation between them generally weakens. However, in this chapter, we will consider the covariance structure of the underlying noise vector ϵ.

5.2.4 Finding the Variances of the Wavelet Coefficients

It is important to find the variances of the wavelet coefficients if the covariance matrix of the original data is known. Consider the general case in which $\boldsymbol{\epsilon} \sim N_n(0, V)$, where V is the covariance matrix. Vannucci and Corradi (1999) and Kovac and Silverman (2000) provided insights into correlation structure of wavelet coefficients for large classes of common processes that the correlated noise may belong to.

The DWT of $\boldsymbol{\epsilon}$, $\mathcal{W}\boldsymbol{\epsilon}$ say, has the distribution

$$N_n(0, \Sigma), \tag{5.7}$$

where $\Sigma = \mathcal{W}V\mathcal{W}^T$ is the covariance matrix of wavelet coefficients.

Vidakovic and Müller (1995) suggested incorporating correlation within each level. Vannucci and Corradi (1999) gave a model allowing full correlation between and within levels and, furthermore, developed a recursive algorithm to calculate the covariance of wavelet coefficients within and across levels. Here, we use this recursive algorithm to obtain the covariance structure.

Figure 5.3 shows the covariance matrix, Σ in (5.7), of AR(1) noise with $\alpha = 0.7$. The small squares along the diagonal of the matrix mark the existing correlation after applying the DWT to the (correlated) data. From the finest level to coarsest level, the colour of the squares tends to be darker when the correlation of the wavelet coefficients is higher.

As a comparison, Figure 5.4 shows the covariance matrix, Σ, of the IID noise. Once again we can see that the covariance matrix is an identity matrix with the same variance.

5.3 Semi-Parametric Approaches

Having provided strong motivation for considering the correlation structure of the original data, we now propose a semi-parametric approach to identify this structure.

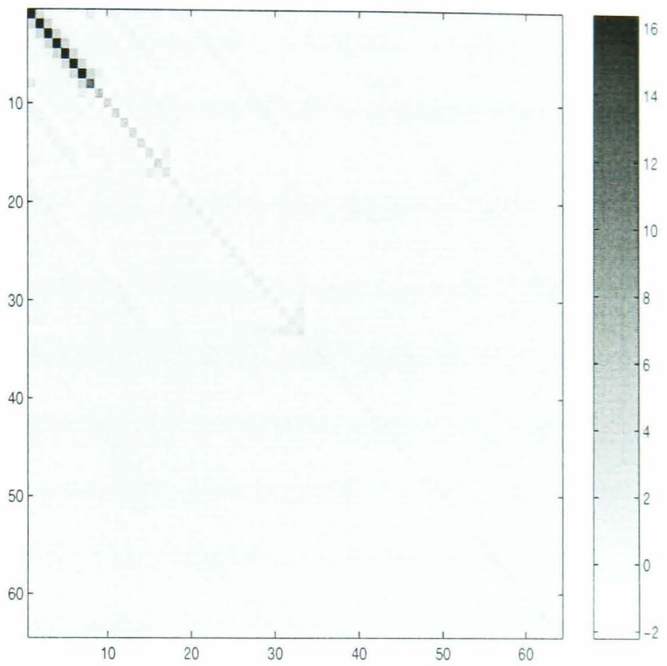

Figure 5.3: Covariance structure of DWT of AR(1) with α=0.7

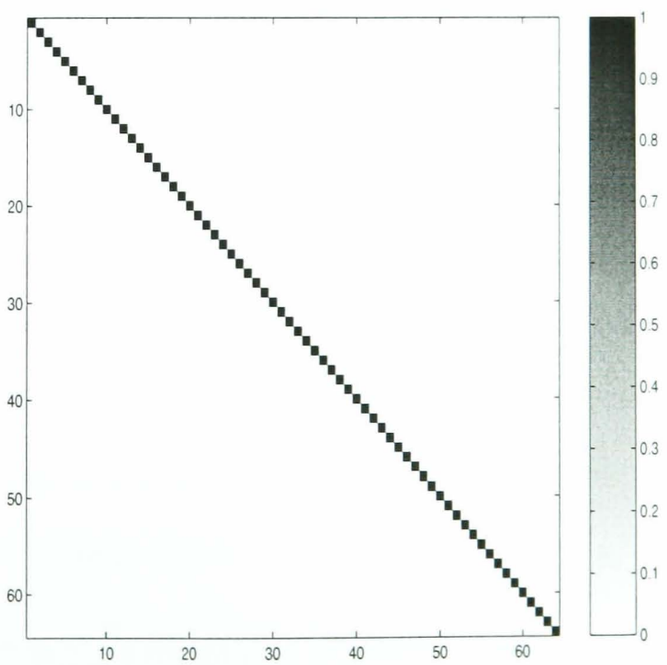

Figure 5.4: Covariance structure of DWT of IID noise

We set

Parametric part: ϵ is multivariate Gaussian with a general (parametric) covariance structure, $V = V(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a parameter vector to be estimated;

Nonparametric part: f is treated nonparametrically as usual.

It is worth noting that in Vannucci and Corradi (1999), they specified a covariance structure of wavelet coefficients according to the properties of wavelet transform, while here we specify the covariance structure $V(\boldsymbol{\theta})$ of the observed data.

To estimate the covariance parameters we can either use estimators in wavelet domain or estimators in time domain. Although these two parametric procedures use the data in different ways, the aims are the same: to estimate parameters in the covariance structure $V = V(\boldsymbol{\theta})$. We shall see from the simulation results that both methods do a good job in the examples considered.

5.3.1 The Parametric Procedure in Time Domain

Firstly, we will have a look at the parametric procedure in time domain. If we know that the ϵ comes from the time series model and a preliminary estimator $\widehat{\epsilon}$ of ϵ can be obtained, then we can use the time series model identification techniques to identify a suitable time series model for $\widehat{\epsilon}$. As we know that the level-dependent "Universal" threshold method mentioned in Johnstone and Silverman (1997) is a quick thresholding method, we will use it here as a preliminary estimator.

The approach proposed in this subsection can be summarised as follows:

Step 1 Start with the model $y_i = f_i + \epsilon_i$, obtain a preliminary estimation $\widehat{f_i}$ of f_i by using the level dependent "Universal" threshold;

Step 2 Obtain $\widehat{\epsilon_i} = y_i - \widehat{f_i}$, which can be regarded as an estimate of the noise, ϵ_i;

Step 3 Use standard time series model identification techniques to determine a suitable parametric covariance structure for vector $\widehat{\epsilon}$, and then estimate the parameters in this covariance structure.

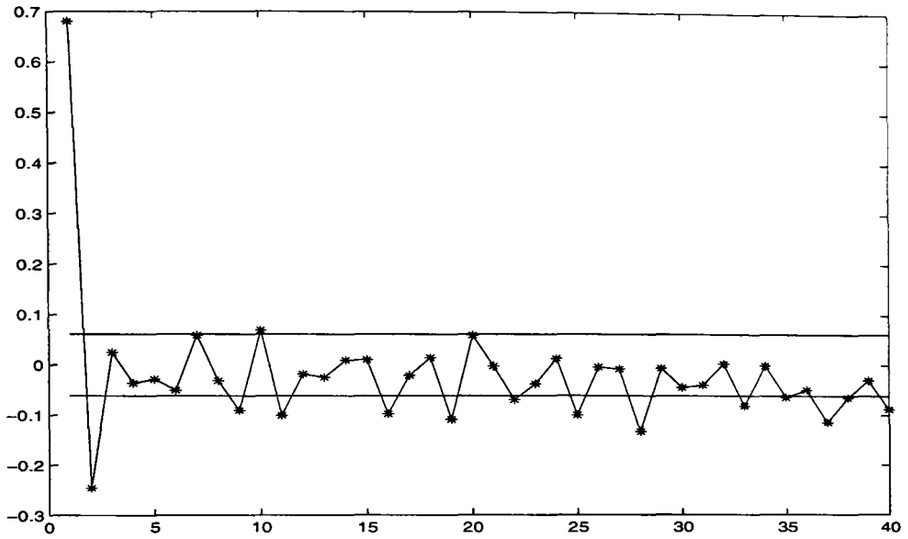

Figure 5.5: The first 40 numbers of the sample pac.f for the estimated data $\widehat{\epsilon}_i$ with the bounds $\pm 1.96n^{-1/2}$.

Steps 1 and 2 are quite straightforward. Here we will look at Step 3 by two examples. Appendix B will give the basic background of time series model identification techniques based on the Durbin-Levinson algorithm and Innovation algorithm. For more detail, see Brockwell and Davis (1991).

Example 5.1 *: 1024 data from a simulated AR(2) process with coefficients* $\alpha_1 = 0.7$ *and* $\alpha_2 = -0.2$ *are added to the HeaviSine signal* f*. Using the level dependent "Universal" threshold method, we obtain the smoothed signal* $\widehat{f}$*. Hence we estimate the noise* ϵ_i *as* $\widehat{\epsilon}_i$ *according to the above steps. By applying the Durbin-Levinson algorithm (see Brockwell and Davis, 1991) to fit successively higher order autoregressive processes to* $\widehat{\epsilon}_i$*, we obtain the sample partial autocorrelation function (the sample pac.f)* $\widehat{\alpha}_{jj}$*. The first 40 numbers of the sample pac.f with the bounds* $\pm 1.96n^{-1/2}$ *are shown in Figure 5.5. Inspection of the graph supports the view that the appropriate model for the noise is an AR(2) process.*

Example 5.2 *: 512 data from a simulated MA(1) process with coefficients* $\beta = 0.5$ *are added to the Doppler signal* f*. Using the same steps as Example 5.1, we obtain* $\widehat{\epsilon}_i$*. By applying the Innovation algorithm (see Brockwell and Davis, 1991) to fit*

	$\hat{\beta}_{mj}$								$\hat{v}_m$
$m \setminus j$	1	2	3	4	5	6	7	8	
1	0.437								1.043
2	0.513	0.051							0.844
3	0.527	0.059	0.013						0.818
4	0.532	0.053	0.025	-0.017					0.813
5	0.532	0.055	0.021	-0.005	-0.029				0.810
6	0.532	0.054	0.023	-0.009	-0.020	-0.029			0.810
7	0.533	0.054	0.023	-0.008	-0.021	-0.028	-0.015		0.810
8	0.534	0.051	0.023	-0.015	-0.012	-0.048	0.032	-0.095	0.809
9	0.544	0.051	0.024	-0.015	-0.011	-0.050	0.037	-0.101	0.798
10	0.550	0.050	0.024	-0.016	-0.011	-0.051	0.037	-0.102	0.791
50	0.526	-0.002	-0.024	-0.053	-0.039	-0.067	0.008	-0.115	0.709
100	0.534	0.007	-0.034	-0.053	-0.031	-0.067	0.007	-0.129	0.646

Table 5.1: The estimated coefficient values $\hat{\beta}_{mj}$, $j = 1, \cdots, 8$ and noise variances $\hat{v}_m$, $m = 1, \cdots, 10, 50, 100$ for the estimated error vector $\widehat{\boldsymbol{\epsilon}}$.

successively higher order moving average processes to the $\widehat{\epsilon}_i$, we obtain the estimated coefficient values $\hat{\beta}_{mj}$ and noise variances $\hat{v}_m$. Table 5.1 shows $\hat{\beta}_{mj}$, $j = 1, \ldots, 8$ and $\hat{v}_m$, $m = 1, \ldots, 10, 50, 100$. This table suggests that MA(1) is the appropriate model for the noise.

5.3.2 The Parametric Procedure in Wavelet Domain

Now we consider the parametric procedure which uses the finest-level wavelet coefficients to estimate the covariance matrix $V(\boldsymbol{\theta})$. As we mentioned in § 2.3.1, the wavelet transform has the property that a wide range of functions have economical wavelet expansions. This means that a function f can be well approximated by

a function whose wavelet coefficients are mostly zero. These do not just include functions that are smooth in a conventional sense, but also those that have discontinuities of values or of gradient. It was proved using mathematical results discussed by Donoho *et al.* (1995). However, wavelet expansions of noise do not have such a property, which can be seen from Figure 5.2. Based on these facts, we suggest the following steps to obtain the wavelet coefficients of the noise to estimate the parameters of the covariance matrix $V(\boldsymbol{\theta})$.

Step 1 Threshold the finest level wavelet coefficients, $\tilde{\mathbf{d}}_J$ say, to obtain $\widehat{\mathbf{d}}_{signal}$;

Step 2 Estimate the portion of the finest level wavelet coefficients attributable to the noise by $\widehat{\mathbf{d}}_{noise} = \tilde{\mathbf{d}}_J - \widehat{\mathbf{d}}_{signal}$;

Step 3 Use maximum likelihood, or if more convenient, a pseudo-likelihood procedure, with estimated data $\widehat{\mathbf{d}}_{noise}$, to estimate the unknown covariance parameters of $V(\boldsymbol{\theta})$.

The reason for basing estimation of the covariance parameters on the finest-level wavelet coefficients only is that these coefficients should be least affected by the smooth parts of the signal. However, if the signal has a few discontinuities, then the finest-level coefficients may have a few very large values due to discontinuities in the signal rather than due to the noise. The purpose of Steps 1 and 2 is to remove these very large coefficients and obtain the wavelet coefficients which come from the noise.

Since we are interested in the finest-level of wavelet coefficients, we look closely at the first part of the wavelet transform, see § 2.3.3. Assume for the moment that we know the noise $\boldsymbol{\epsilon}$ added to the signal. Let $\mathcal{W}_J$ be the wavelet transform to transform $\boldsymbol{\epsilon}$ to the finest level wavelet coefficients $\mathcal{W}_J\boldsymbol{\epsilon}$, which has distribution $N_{n/2}(0, \Sigma_J)$, where $\Sigma_J = \mathcal{W}_J^T V \mathcal{W}_J$ and $\mathcal{W}_J$ is a rectangular matrix of dimension $n \times n/2$, which is a submatrix of an $n \times n$ orthogonal matrix. Then we have an $n/2$-vector of the finest level wavelet coefficients $\mathbf{d}_{noise}$ given by

$$\mathbf{d}_{noise} = \mathcal{W}_J\boldsymbol{\epsilon}. \tag{5.8}$$

In the following we show how to estimate the parameter vector $\boldsymbol{\theta}$ of $V(\boldsymbol{\theta})$ using $\widehat{\mathbf{d}}_{noise}$, an estimate of $\mathbf{d}_{noise}$.

5.3.3 Maximum Likelihood Estimation

In order to estimate the parameter vector $\boldsymbol{\theta}$ in a Gaussian model with covariance structure $V = V(\boldsymbol{\theta})$, maximum likelihood estimation may be used. Assume that we have $\hat{\mathbf{d}}_{noise}$ obtained from Step 2 in § 5.3.2. To simplify the presentation, the notation $\mathbf{d}$ rather than $\hat{\mathbf{d}}_{noise}$ is used in this subsection though in practice we use the latter.

The Full Maximum Likelihood Estimation

The log-likelihood based on $\mathbf{d}$ is

$$l(\boldsymbol{\theta}) = \log\{f(\mathbf{d}|\boldsymbol{\theta})\} = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log\{\det(\Sigma_J)\} - \frac{1}{2}\mathbf{d}^T\Sigma_J^{-1}\mathbf{d} \qquad (5.9)$$

Maximum likelihood estimation can be carried out by maximizing (5.9) over the valid parameter space. For example, in the case of AR(1) defined in (5.2), we have $\sigma^2 > 0$, $0 < \alpha < 1$. However, we found that when the dimension of $\mathbf{d}$ increases, it becomes difficult to calculate the inverse and determinant of Σ_J. For this reason, we have investigated various pseudo-likelihood (PL) approaches (Besag, 1975, 1977). PL is a sub-optimal alternative to maximum likelihood but is often easy to implement and is useful when the maximum likelihood estimator is too hard to compute. It may be inefficient when spatial interactions are strong.

PL Estimation based on Pairs

There are many ways of implementing PL. For example, we may take the PL to be the product of the bivariate density functions of all distinct pairs of elements from vector $\mathbf{d}$, which is, for any index i and j, the pair $(d_i, d_j)^T$ distributed as the bivariate normal distribution

$$(d_i, d_j)^T \sim N_2(0, \Sigma_{[i,j]}).$$

The PL is

$$\prod_{i<j} f(d_i, d_j) = \prod_{i<j} \frac{1}{2\pi(\det(\Sigma_{[i,j]}))^{1/2}} \exp\Big(-(1/2)(d_i, d_j)\Sigma^{-1}_{[i,j]}(d_i, d_j)^T \Big). \quad (5.10)$$

In our case, when the data are highly correlated, the PL estimation based on pairs is not accurate enough.

PL Estimation Based on Large Blocks

In this thesis, we consider a different type of PL in which the data is split into a small number of large blocks: we split the vector $\mathbf{d}$ into k subvectors, where k is relatively small. Each subvector has $h = n/(2k)$ elements, denoted as $\mathbf{d}_i = (d_{i1}, d_{i2}, ..., d_{ih})^T \sim N_h(0, \Sigma_{Ji})$. For each block, we can write log-likelihood distribution as

$$l_i(\boldsymbol{\theta}) = \log\{f(\mathbf{d}_i|\boldsymbol{\theta})\} = const - \frac{1}{2}\log\{\det(\Sigma_{Ji})\} - \frac{1}{2}\mathbf{d}_i^T\Sigma_{Ji}^{-1}\mathbf{d}_i \quad (5.11)$$

and $\sum_{i=1}^{k} l_i(\boldsymbol{\theta})$ is the sum of these component log-likelihoods to be maximised. It is worth noting that this PL approach ignores correlations between blocks.

5.3.4 Identification of Parametric Structure

This parametric procedure in wavelet domain brings forward a problem of its own, which is the need to determine the parametric structure of the covariance matrix beforehand. A preliminary study, an idea we borrowed from the parametric procedure in time domain, may be used to satisfy this need. Before proceeding to follow the steps mentioned in § 5.3.2, we perform the following preliminary study:

Step 1 Start with the model $y_i = f_i + \epsilon_i$, obtain a preliminary estimate $\widehat{f_i}$ of f_i by using the level dependent "Universal" threshold;

Step 2 Obtain $\widehat{\epsilon_i} = y_i - \widehat{f_i}$, which can be regarded as the estimation of noise ϵ_i;

Step 3 Use standard time series model identification techniques to determine a suitable covariance structure.

This three-step preliminary study is similar to the steps of the parametric procedure in time domain. After identifying the parametric covariance structure, we can follow the procedure in § 5.3.2.

5.3.5 The Nonparametric Procedure

After estimating the covariance parameter vector $\boldsymbol{\theta}$, we can treat the $V = V(\boldsymbol{\theta})$ as known. Hence the unknown function f can be estimated using a suitable shrinkage or threshold method. In order to investigate this semi-parametric approach, we will concentrate on the generalized EBB model (4.3), with z now defined by $\widehat{\mathbf{d}}^T(\Sigma(\widehat{\boldsymbol{\theta}}))^{-1}\widehat{\mathbf{d}}$.

5.4 Simulation Study

In this section, we present the results of some simulations to illustrate the methods proposed above. As a comparison, the results of the level dependent "Universal" threshold methods (JS) mentioned by Johnstone and Silverman (1997), the semi-parametric method with the parametric procedure in time domain (TD), and the parametric procedure in wavelet domain (WD) with block sizes equal to 2, treated as two separate methods, will be presented here.

5.4.1 Specific Covariance Matrices

The types of correlated noise we will consider here are AR(1), AR(2) and MA(1). If the model where the noise comes from is known, we can write down the covariance matrix, see Cox and Miller (1968), Chatfield (1975). For the AR(1) process, the

covariance matrix with two parameters α and σ^2 is given in the following:

$$\sigma^2 V(\alpha) = \sigma^2 \begin{pmatrix} 1 & \alpha & \alpha^2 & \cdots & \alpha^{n-1} \\ \alpha & 1 & \alpha & \cdots & \alpha^{n-2} \\ \alpha^2 & \alpha & 1 & \cdots & \alpha^{n-3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \alpha^{n-1} & \alpha^{n-2} & \alpha^{n-3} & \cdots & 1 \end{pmatrix}.$$

For the AR(2) process, $\epsilon_t = \alpha_1 \epsilon_{t-1} + \alpha_2 \epsilon_{t-2} + \eta_t$, where $E(\eta_t) = 0$, $E(\eta_t \eta_s) = 0$ for $t \neq s$, and $E(\eta_t^2) = 1$. When the process is stationary (i.e. $\alpha_1 + \alpha_2 < 1$, $\alpha_2 - \alpha_1 < 1$ and $-1 < \alpha_2 < 1$), the elements of the covariance matrix $E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T) = \sigma^2 \Omega$ can be found from the variance

$$\sigma^2 = \frac{1 - \alpha_2}{(1 + \alpha_2)\{(1 - \alpha_2)^2 - \alpha_1^2\}}$$

and the autocorrelation matrix $\Omega = [r_{i,j}]$ with $r_{i,j} = \rho_{|i-j|}$ where $\rho_0 = 1$, $\rho_1 = \alpha_1(1 - \alpha_2)^{-1}$, $\rho_2 = \alpha_2 + \alpha_2^2(1 - \alpha_1)^{-1}$ and $\rho_i = \alpha_1 \rho_{i-1} + \alpha_2 \rho_{i-2}$ for $i > 2$. The inverse of Ω is given by

$$\Omega^{-1} = \begin{pmatrix} 1 & -\alpha_1 & -\alpha_2 & 0 & \cdots & 0 \\ -\alpha_1 & 1 + \alpha_1^2 & -\alpha_1 + \alpha_1\alpha_2 & -\alpha_2 & \cdots & 0 \\ -\alpha_2 & -\alpha_1 + \alpha_1\alpha_2 & 1 + \alpha_1^2 + \alpha_2^2 & -\alpha_1 + \alpha_1\alpha_2 & \cdots & 0 \\ 0 & -\alpha_2 & -\alpha_1 + \alpha_1\alpha_2 & 1 + \alpha_1^2 + \alpha_2^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \end{pmatrix}.$$

For the MA(1) process, the covariance matrix $\sigma^2 V(\beta)$ with two parameters β and σ^2 is given in the following:

$$\sigma^2 V(\beta) = \sigma^2 \begin{pmatrix} 1 + \beta^2 & \beta & 0 & 0 & \cdots & 0 \\ \beta & 1 + \beta^2 & \beta & 0 & \cdots & 0 \\ 0 & \beta & 1 + \beta^2 & \beta & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \\ 0 & 0 & 0 & 0 & \cdots & 1 + \beta^2 \end{pmatrix}.$$

Because of the way we normalise the noise, see (5.4), σ^2 in above three cases are actually equal to 1. In the following simulations, we do not estimate σ^2. If σ^2 is unknown, the methods described above can easily be extended to estimate σ^2 with the other parameters at the same time.

5.4.2 Simulation Results

Four signals, "HeaviSine", "Blocks", "Bumps" and "Doppler", first proposed in Donoho and Johnstone (1994, 1995) as test functions for wavelet estimators, are considered here. In order to compare with different methods, signal-to-noise ratio (SNR) equal to 7 and sample size equal to 1024 have been used for all signals.

Figures 5.6- 5.9 show the reconstructions of four test functions (HeaviSine, Bumps, Blocks and Doppler) from three types of noises: AR(1) ($\alpha = 0.7$), AR(2) ($\alpha_1 = 0.7$ and $\alpha_2 = -0.2$) and MA(1) ($\beta = 0.5$). Three methods, JS (Johnstone and Silverman, 1997), TD (the semi-parametric method with the parametric procedure in time domain) and WD (the semi-parametric method with the parametric procedure in wavelet domain) are used to denoise the noisy functions. Generally speaking, all three methods can give approximately noise-free reconstructions; compare Figure 5.6 with Figure 5.1. However, if we compare these three methods based on Figures 5.6- 5.9, we can see that TD and WD methods work better than JS method when the signal has a few discontinuous points (e.g. in the case of Bumps signal). It is worth noting that, in the simulation study, the correct ARMA model was assumed when implementing the TD and WD approaches. Thus, the TD and WD methods had an advantage that they would not have in a real data example, where the correlation structure would be unknown. Nevertheless, the results show conclusively that explicit modelling of the correlation structure can lead to substantially improved performance relative to the Johnstone and Silverman (1997) approach.

Table 5.2 gives the simulation results of AR(1), AR(2) and MA(1) noises. The $\hat{\alpha}$ for AR(1) case, $\hat{\alpha}_1$ and $\hat{\alpha}_2$ for AR(2) case and $\hat{\beta}$ for MA(1) case are the average values of 100 estimations of the parameters in each case. From MSEs of these three methods

with three noise types, we can see that the TD and WD methods are superior to the JS method. Especially, from Table 5.2 we can see that for rough signal, like Bumps, the MSEs of the TD and WD methods are more than ten times smaller than that of the JS method for three signals. The TD and WD methods are quite competitive although the PL estimation based on large blocks of the WD method is computationally intensive, especially when the order of the noise is high. However, considering the improvement of the average MSE and the widespread availability of high-powered computers, this cost is worthwhile.

Figures 5.10- 5.12 represent box plots of 100 estimations of the parameters in each noise type, which is added to Doppler signal. The boxes have lines at lower quantile, median and upper quantile of 100 estimations. The lines extending from each end of the boxes show the extent of the rest of the estimations. The plus symbol ("+") denotes estimates which are beyond the ends of the extension lines. Generally speaking, the estimations obtained by the WD method have smaller variance than those obtained by the TD method.

5.5 Conclusions and Further Work

1. A semi-parametric approach to the problem of estimating f in the presence of correlated noise ϵ has been explored. In the parametric part of this semi-parametric approach, two estimating procedures, a time domain approach and a wavelet domain approach, are used to estimate the parameters in the covariance structure of ϵ. Some simulation results examined the effect of combining this parametric approach with the empirical Bayes block (EBB) shrinkage method, and showed that this combined approach can successfully handle the correlated data situation.

2. Many of the methods developed for obtaining thresholding/shrinkage estimators of f in the standard case can be adapted to the case when the covariance matrix of ϵ is known. It is easy to combine the parametric part of this semi-

parametric approach with these existing methods, for example Abramovich *et al.* (2002).

3. In the semi-parametric approach there is a need to specify the parametric structure of the covariance matrix of ϵ beforehand. A preliminary study, which involves the use of traditional time series identification techniques, has been used to determine a suitable parametric structure.

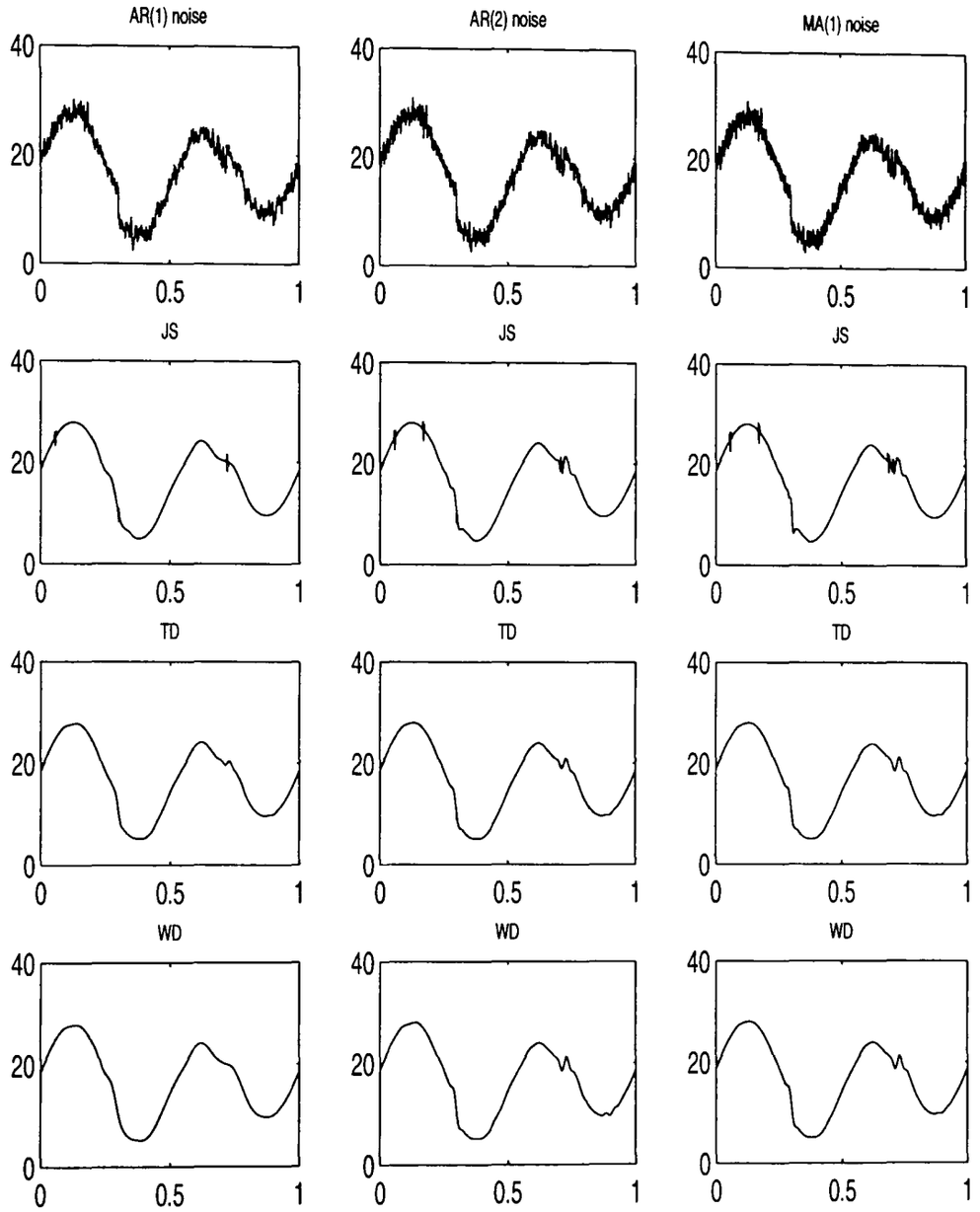

Figure 5.6: HeaviSine signal with three types of noises added, based on sample size n=1024 and SNR=7. The reconstructions are obtained using the JS (Johnstone and Silverman, 1997), TD (the semi-parametric method with the parametric procedure in time domain) and WD (the semi-parametric method with the parametric procedure in wavelet domain) procedures.

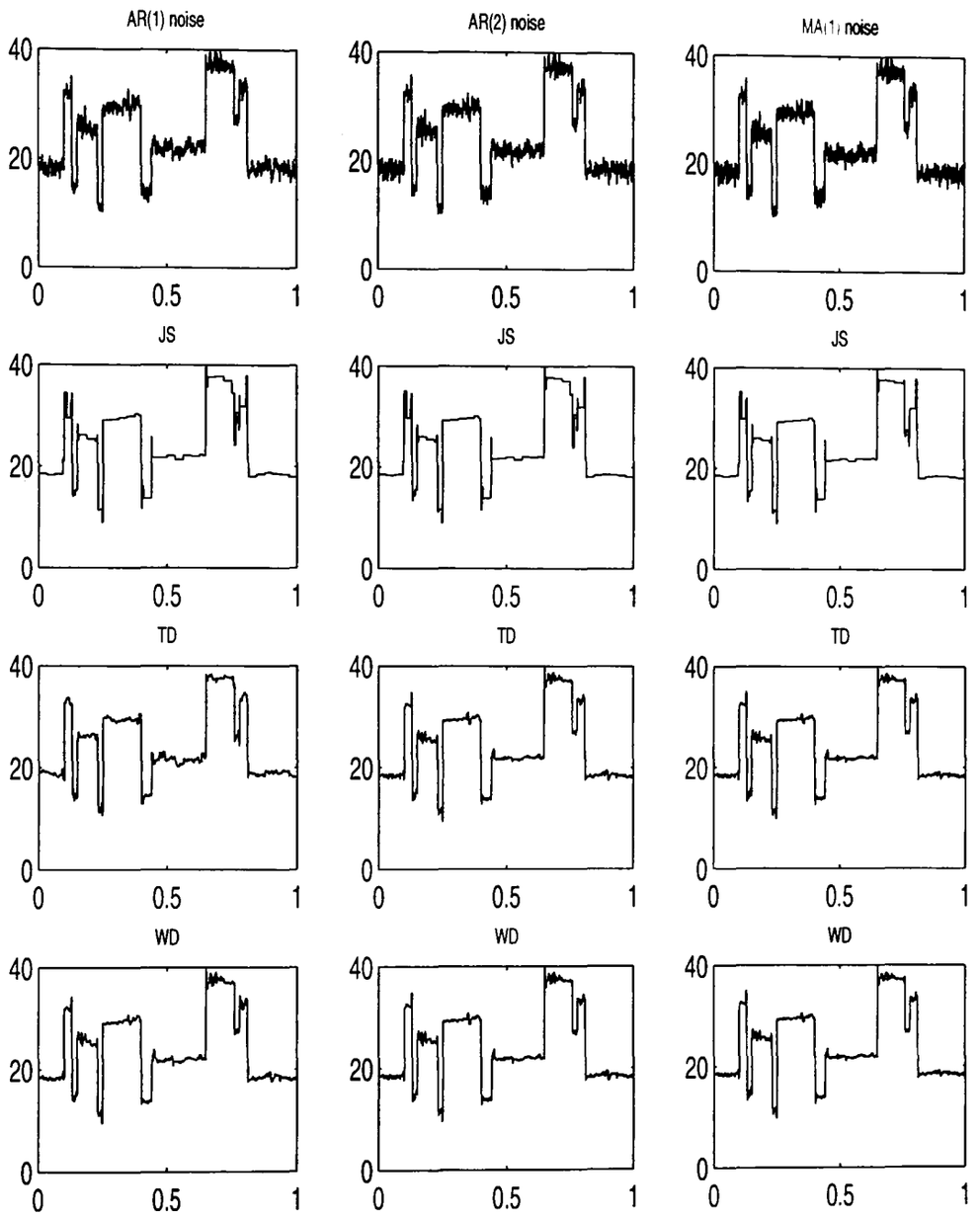

Figure 5.7: Blocks signal with three types of noises added, based on sample size n=1024 and SNR=7. The reconstructions are obtained using the JS (Johnstone and Silverman, 1997), TD (the semi-parametric method with the parametric procedure in time domain) and WD (the semi-parametric method with the parametric procedure in wavelet domain) procedures.

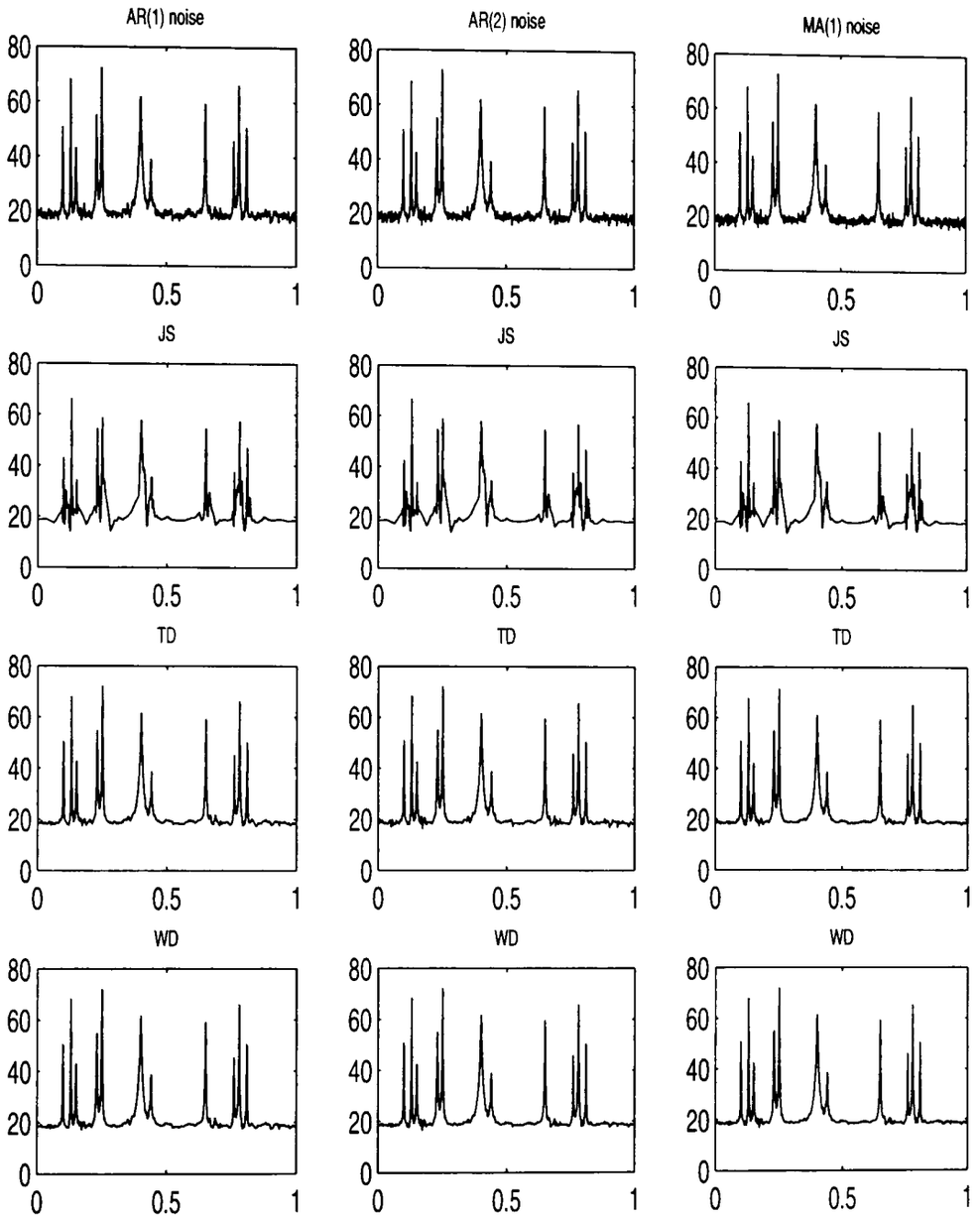

Figure 5.8: Bumps signal with three types of noises added, based on sample size n=1024 and SNR=7. The reconstructions are obtained using the JS (Johnstone and Silverman, 1997), TD (the semi-parametric method with the parametric procedure in time domain) and WD (the semi-parametric method with the parametric procedure in wavelet domain) procedures.

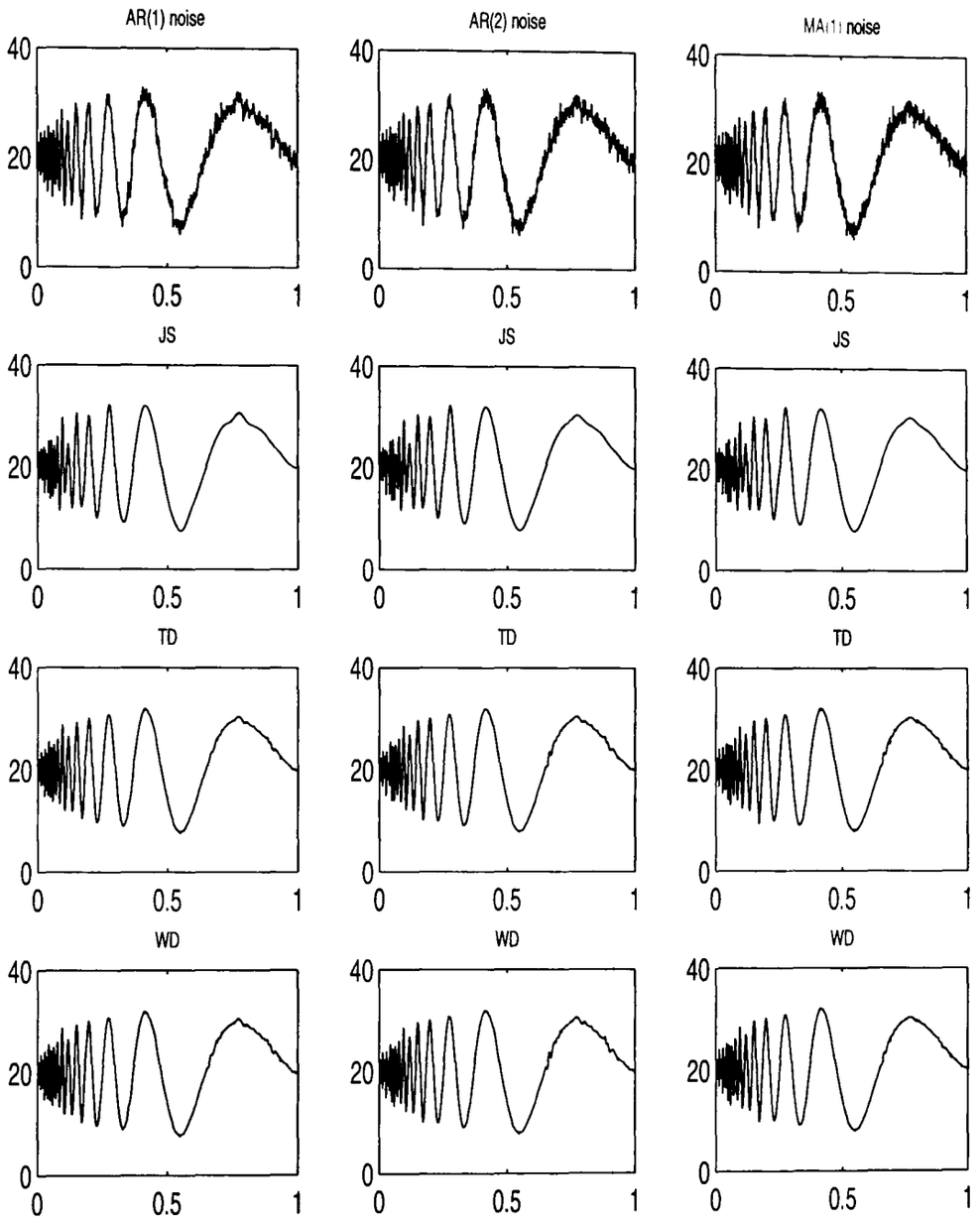

Figure 5.9: Doppler signal with three types of noises added, based on sample size n=1024 and SNR=7. The reconstructions are obtained using the JS (Johnstone and Silverman, 1997), TD (the semi-parametric method with the parametric procedure in time domain) and WD (the semi-parametric method with the parametric procedure in wavelet domain) procedures.

Signals	Methods	AR(1)		AR(2)			MA(1)	
		MSE	$\hat{\alpha}$	MSE	$\hat{\alpha}_1$	$\hat{\alpha}_2$	MSE	$\hat{\beta}$
Heavi-Sine	WD	0.2791	0.6905	0.1603	0.7272	-0.3436	0.1266	0.5529
	TD	0.2976	0.6045	0.1776	0.5967	-0.2374	0.1286	0.3944
	JS	0.3337	/	0.2316	/	/	0.1848	/
Bumps	WD	0.5567	0.6375	0.4411	0.6876	-0.3677	0.3618	0.4844
	TD	0.5493	0.7814	0.4816	0.5525	-0.2694	0.3576	0.6277
	JS	6.8436	/	6.948	/	/	6.9175	/
Blocks	WD	0.3795	0.7072	0.2816	0.7783	-0.3318	0.2281	0.5466
	TD	0.4102	0.608	0.2939	0.6006	-0.1909	0.2329	0.4424
	JS	0.465	/	0.3560	/	/	0.2827	/
Doppler	WD	0.3448	0.6781	0.2573	0.7004	-0.3419	0.2002	0.560
	TD	0.3732	0.6182	0.2932	0.5948	-0.2807	0.2035	0.4115
	JS	0.5463	/	0.4342	/	/	0.3747	/

Table 5.2: The comparison of three methods, JS (Johnstone and Silverman, 1997), TD (the semi-parametric method with the parametric procedure in time domain) and WD (the semi-parametric method with the parametric procedure in wavelet domain), under 100 simulation runs. MSE obtained for SNR=7 and sample sizes n=1024. The three noise types are AR(1) with $\alpha = 0.7$, AR(2) with $\alpha_1 = 0.7$ and $\alpha_2 = -0.2$ with MA(1) with $\beta = 0.5$

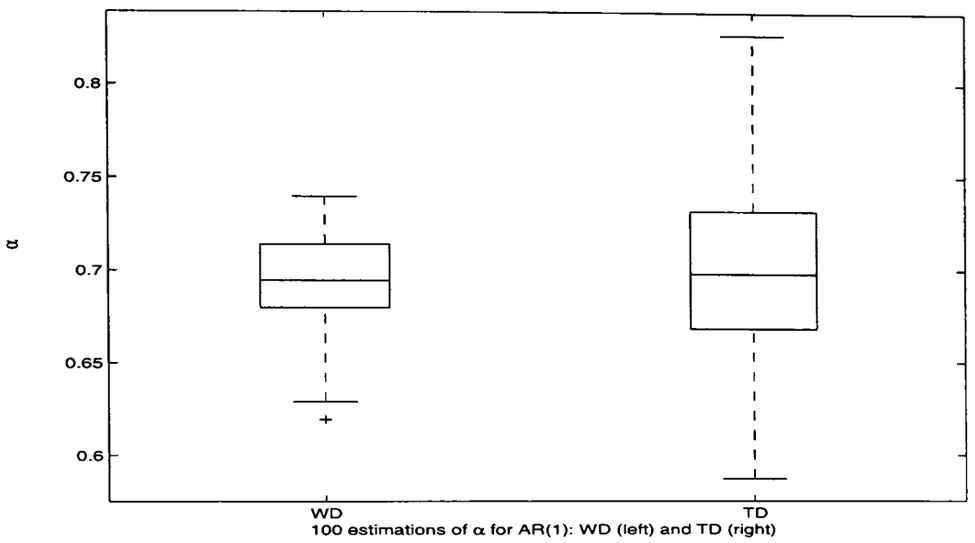

Figure 5.10: Box plots for 100 estimations of α of AR(1) noise added to Doppler signal. The true $\alpha = 0.7$, SNR=7 and sample size $n = 1024$.

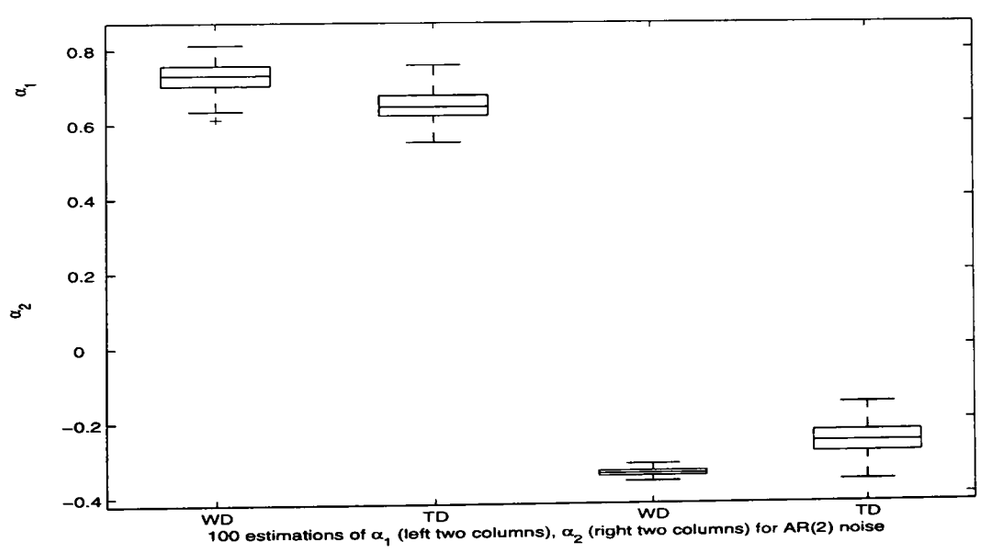

Figure 5.11: Box plots for 100 estimations of α_1 and α_2 of AR(2) noise added to Doppler signal. The true $\alpha_1 = 0.7$, $\alpha_2 = -0.2$, SNR=7 and sample size $n = 1024$.

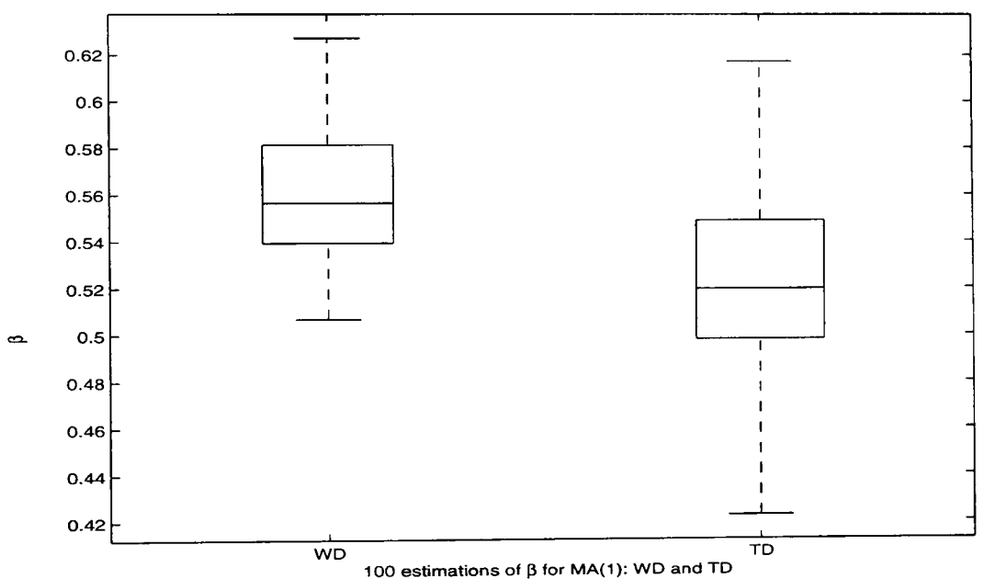

Figure 5.12: Box plots for 100 estimations of β of MA(1) noise added to Doppler signal. The true $\beta = 0.5$, SNR=7 and sample size $n = 1024$.

Appendix B: Review of Time Series Techniques

Most of the results in this appendix are from Brockwell and Davis (1991), where further details can be found.

A time series $\{x_t, t \in T\}$ is a realization of family of random variables $\{X(t), t \in T\}$, where T is a set of times at which observations are made. A time series is said to be strictly stationary if the joint distribution of $X(t_1), \ldots, X(t_n)$ is the same as the joint distribution of $X(t_1+\tau), \ldots, X(t_n+\tau)$ for all $t_1, \ldots, t_n, \tau$. This definition shows that the distribution of $X(t)$ must be the same for all t, so that

$$\begin{aligned}
\mu = \mu(t) &= E\{X(t)\}, && \text{mean} \\
\sigma^2 = \sigma^2(t) &= Var\{X(t)\}, && \text{variance} \\
\gamma(\tau) &= E\big[\{X(t)-\mu\}\{X(t+\tau)-\mu\}\big], && \text{autocovariance} \\
\rho(\tau) &= \gamma(\tau)/\gamma(0). && \text{autocorrelation function}
\end{aligned}$$

The sample autocorrelation function (ac.f) based on a set of observations of a time series is an important set of statistics for describing the time series. Given n observations $x_1, \ldots, x_n$ of a time series, the sample autocovariance function is defined as

$$\widehat{\gamma}_k = \frac{1}{n}\sum_{t=1}^{n-k}(x_t - \bar{x})(x_{t+k} - \bar{x}) \qquad 0 \le k < n$$

and $\widehat{\gamma}_k = \widehat{\gamma}_{-k}$, $-n \le k \le 0$, where $\bar{x}$ is the sample mean $n^{-1}\sum_{t=1}^{n} x_t$. The sample autocorrelation function is defined in terms of the sample autocovariance function as

$$\widehat{\rho}_k = \widehat{\gamma}_k/\widehat{\gamma}_0$$

for $k = 1, 2, \ldots, m$ where $m < n$.

The partial autocorrelation function (pac.f), like ac.f, conveys vital information regarding the dependence structure of a stationary process. Let $\{X(t)\}$ be a zero mean stationary process with autocovariance function $\gamma(\cdot)$ such that $\gamma(h) \to 0$ as

$h \to \infty$, and ac.f $\rho(\cdot)$. If we have an AR process, we have

$$\begin{pmatrix} \rho(0) & \rho(1) & \rho(2) & \cdots & \rho(k-1) \\ \rho(1) & \rho(0) & \rho(1) & \cdots & \rho(k-2) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho(k-1) & \rho(k-2) & \rho(k-3) & \cdots & \rho(0) \end{pmatrix} \begin{pmatrix} \phi_{k1} \\ \phi_{k2} \\ \vdots \\ \phi_{kk} \end{pmatrix} = \begin{pmatrix} \rho(1) \\ \rho(2) \\ \vdots \\ \rho(k) \end{pmatrix}. \quad (5.12)$$

where $k \geq 1$, then the pac.f $\zeta(k)$ of $\{X(t)\}$ at lag k is

$$\zeta(k) = \phi_{kk} \qquad k \geq 1,$$

where ϕ_{kk} is uniquely determined by (5.12).

The sample pac.f at lag k of $\{x_1, \ldots, x_n\}$ is defined, provided $x_i \neq x_j$ for some i and j, by

$$\widehat{\zeta}(k) = \widehat{\phi}_{kk} \qquad 1 \leq k \leq n,$$

where $\widehat{\phi}_{kk}$ is uniquely determined by (5.12) with each $\rho(j)$ replaced by the corresponding sample ac.f $\widehat{\rho}(j)$.

Definition B.1 *The autoregressive moving-average process of order p and q (denoted as ARMA(p,q)): a process $X_t, t = 0, \pm 1, \pm 2, \ldots$ is said to be an ARMA(p,q) process if X_t is stationary and for every t,*

$$X_t - \alpha_1 X_{t-1} - \cdots - \alpha_p X_{t-p} = Z_t + \beta_1 Z_{t-1} + \cdots + \beta_q Z_{t-q}, \quad (5.13)$$

where Z_t is the process which has zero mean and covariance function

$$\gamma(h) = \begin{cases} \sigma^2 & \text{if } h = 0 \\ 0 & \text{if } h \neq 0, \end{cases} \quad (5.14)$$

denoted $Z_t \sim \mathbf{WN}(0, \sigma^2)$.

In particular, a process is said to be an autoregressive process of order p (AR(p)) if $\beta_1 = \beta_2 = \cdots = \beta_q = 0$ and $X_t - \alpha_1 X_{t-1} - \cdots - \alpha_p X_{t-p} = Z_t$. A process is said to be a moving-average process of order q (MA(q)) if $\alpha_1 = \alpha_2 = \cdots = \alpha_p = 0$ and $X_t = Z_t + \beta_1 Z_{t-1} + \cdots + \beta_q Z_{t-q}$.

For an ARMA process, there are two related questions:

- What are the AR and MA orders of the process?
- How can we estimate the parameters of the process?

It is usually difficult to assess the order of an AR process from the sample ac.f alone. One way to determine the order of the AR process is the pac.f. The sample pac.f is estimated by fitting AR processes of successively higher order and taking $\hat{\zeta}(1) = \hat{\phi}_1$ when an AR(1) process is fitted, taking $\hat{\zeta}(2) = \hat{\phi}_2$ when an AR(2) process if fitted, and so on. If the sample pac.f values $\{\hat{\zeta}(k)\}$, $k \geq p$, lie outside the bounds $\pm 1.96n^{-1/2}$, the AR process "cuts off" at p so that the "correct" order is assessed.

If an MA process is appropriate for a given set of data, the order of the process is usually evident from the sample ac.f. The theoretical ac.f of an MA(q) process has a very simple form in that it "cuts off" at lag q:

$$\begin{cases} 1 & k = 0 \\ \sum_{i=0}^{q-k} \beta_i \beta_{i+k} / \sum_{i=0}^{q} \beta_i^2 & k = 1, \cdots, q \\ 0 & k > q \\ \rho(-k) & k < 0 \end{cases} \tag{5.15}$$

The parameters of the AR(p) or MA(q) process can be estimated by using the following propositions. Suppose we have observations $x_1, \ldots, x_n$ of a zero-mean stationary time series. Provided $\hat{\gamma}(0) > 0$, we can fit an autoregressive process of order $m < n$ to the data. The fitted AR(m) process is

$$X_t - \hat{\phi}_{m1} X_{t-1} - \ldots - \hat{\phi}_{mm} X_{t-m} = Z_t, \qquad \{Z_t\} \sim \mathbf{WN}(0, \hat{v}_m). \tag{5.16}$$

$\hat{\phi}_{m1}, \ldots, \hat{\phi}_{mm}$ and $\hat{v}_m$ can be obtained by the following proposition, see Brockwell and Davis (1991).

Proposition B.1 *The Durbin-Levinson algorithm for fitting autoregressive models: If $\hat{\gamma}(0) > 0$ then the fitted autoregressive models (5.16) for $m = 1, 2, \ldots, n-1$, can be determined recursively from the relations, $\hat{\phi}_{11} = \hat{\rho}(1)$, $\hat{v}_1 = \hat{\gamma}(0)[1 - \hat{\rho}^2(1)]$.*

$$\hat{\phi}_{mm} = \left\{ \hat{\gamma}(m) - \sum_{j=1}^{m-1} \hat{\phi}_{m-1,j} \hat{\gamma}(m-j) \right\} \Big/ \hat{v}_{m-1}, \tag{5.17}$$

$$\begin{pmatrix} \hat{\phi}_{m1} \\ \vdots \\ \hat{\phi}_{m,m-1} \end{pmatrix} = \hat{\boldsymbol{\phi}}_{m-1} - \hat{\phi}_{mm} \begin{pmatrix} \hat{\phi}_{m-1,m-1} \\ \vdots \\ \hat{\phi}_{m-1,1} \end{pmatrix}$$

and

$$\hat{v}_m = \hat{v}_{m-1}(1 - \hat{\phi}_{mm}^2). \tag{5.18}$$

Similar to fitting autoregressive models of orders $1, 2, \ldots,$ to the data $x_1, \ldots, x_n$ by applying the Durbin-Levinson algorithm to the sample autocovariances, we can also fit moving-average models,

$$X_t = Z_t + \hat{\theta}_{m1} Z_{t-1} + \cdots + \hat{\theta}_{mm} Z_{t-m}, \qquad \{Z_t\} \sim \mathbf{WN}(0, \hat{v}_m), \tag{5.19}$$

of orders $m = 1, 2, \ldots,$ by means of the innovations algorithm as follows (see Brockwell and Davis, 1991):

Proposition B.2 *The Innovation algorithm of moving-average models: If* $\hat{\gamma}(0) > 0$, *we have the innovation estimates* $\hat{\boldsymbol{\theta}}_m$, $\hat{v}_m$ *appearing in (5.19) for* $m = 1, 2, \ldots, n-1$, *by the recursion relations,* $\hat{v}_0 = \hat{\gamma}(0)$,

$$\hat{\theta}_{m,m-k} = \hat{v}_k^{-1} \Big\{ \hat{\gamma}(m-k) - \sum_{j=0}^{k-1} \hat{\theta}_{m,m-j} \hat{\theta}_{k,k-j} \hat{v}_j \Big\} \tag{5.20}$$

and

$$\hat{v}_m = \hat{v}_0 - \sum_{j=0}^{m-1} \hat{\theta}_{m,m-j}^2 \hat{v}_j. \tag{5.21}$$

Chapter 6

General Conclusions and Suggestions for Further Research

6.1 General Conclusions

This thesis has extended Bayesian machinery for wavelet shrinkage and thresholding by developing block shrinkage Bayesian methodology based upon the non-central χ^2 distribution. Following this theoretical work, an empirical Bayes block (EBB) shrinkage and thresholding procedure was developed.

In this Bayesian model, families of prior distributions have been chosen with careful consideration. These priors are sufficiently flexible to represent a variety of forms of prior knowledge and at the same time the theoretical calculation of the posterior distributions is relatively straightforward. Among these priors, the power prior shows advantages in both theoretical and numerical aspects. The posterior median with the power prior as an estimation rule is proved to be a shrinkage or a thresholding rule if certain mild conditions are satisfied.

Step (4.2) of the Bayesian block shrinkage approach is not a fully Bayesian procedure although the shrinkage of the sum of squares we proposed is fully Bayesian. However, our discussion of the shrinkage estimator in (4.2) and simulation results of the EBB procedure show that (4.2) is theoretically well-motivated and that the

proposed EBB procedure is competitive with existing methods. Furthermore, this approach can be easily extended to more general situations if the covariance matrix is known or can be effectively estimated.

The "quick and dirty" approach of §4.3, which is a completely data-based method for estimating the hyperparameters, provides a fast and accurate estimation procedure. An extensive review of previous work shows that the approach is competitive with alternative published methods.

Some existing standard shrinkage and thresholding methods and the EBB method proposed in the first part of this thesis are adapted to data with correlated noise. It is shown by numerical example that standard methods, which are based on the assumption that the noise is IID noise, will sometimes perform poorly if used in the correlated noise setting.

To address this problem, a semi-parametric approach is used for the identification of the covariance structure of correlated noise according to the data available. In the parametric part of this semi-parametric approach, estimation of the covariance structure in both the time domain and the wavelet domain are considered. Two estimation procedures, time series techniques and maximum likelihood estimation, are used to estimate the parameters in the covariance structure. The nonparametric part can be regarded as an application of the proposed EBB approach in a general situation where the covariance matrix of the noise is treated as a known matrix. The simulation results of this semi-parametric approach were compared with Johnstone and Silverman's (1997) approach and showed a significant improvement.

6.2 Suggestions for Further Work

It has already been noted that for the proposed EBB approach, the power prior and exponential prior for ρ have a heavy tail in both cases. It would be useful to investigate the theoretical properties of these wavelet estimations and determine to what extent a theory parallel to that developed by Johnstone and Silverman (2005)

for empirical Bayes in the standard framework can be developed in the non-central χ^2 framework considered here.

In this thesis, the EBB approach is developed using a block thresholding strategy to denoise the noisy data. As one specific application considered in § 4.9, the EBB approach has been applied to deal with the planer curve denoising problem. It would be interesting to use this approach to higher dimension planar curves, shape recognition and curve matching. In addition, the application of the EBB approach to fitting smooth curves in shape space is also envisioned; cf. Kume *et al.* (2004).

As pointed in § 4.2.1, if complex wavelets or multiwavelets are considered, the proposed EBB approach seems attractive and natural since it is convenient to base the shrinkage procedure on a suitable sum of squares. Although more work is needed to develop these application, we believe the proposed EBB approach can be successfully applied to these situations and obtain good practical performance.

In this thesis, only simple forms of time series model have been considered. It would be desirable to consider more general types of dependence and develop robust estimators of the covariance structure in correlated data settings. Furthermore, a challenging problem would be analysing the asymptotic properties of variances of robust estimations of the covariance parameters.

More generally, we could relax the assumption of Gaussian noise. If we consider non-Gaussian noise, extra steps in the wavelet transform procedure can be included to weaken the correlation of the wavelet coefficients of a noisy signal. For example: Donoho and Yu (1997) constructed a nonlinear multiresolution analysis based on a triad grid, which worked well in a non-Gaussian noise. The idea of this approach is to use the data at coarser levels to predict data at finer using median interpolation. This type of approach deserves further consideration.

Bibliography

[1] Abramovich, F., Bailey, T. C. and Sapatinas, T. (2000). Wavelet analysis and its statistical applications. *The Statistician*, **49**, 1-29.

[2] Abramovich, F., Besbeas, P. and Sapatinas, T. (2002). Empirical Bayes approach to block wavelet function estimation. *Computational Statistics and Data Analysis*, **39**, 435–451.

[3] Abramovich, F. and Sapatinas, T. (1999). Bayesian approach to wavelet decomposition and shrinkage. In *Bayesian Inference in Wavelet-Based Models* (Lecture Notes in Statistics, Vol. 141), edited by P. Müller and B. Vidakovic, 33–50. Springer–Verlag, New York.

[4] Abramovich, F., Sapatinas, T. and Silverman, B. W. (1998). Wavelet thresholding via a Bayesian approach. *J.R. Statist. Soc B*, **60**, 725–749.

[5] Antoniadis, A., Bigot, J. and Sapatinas, T. (2001). Wavelet estimator in nonparametric regression: a comparative simulation study. *J. Statist. Software*, **6**, 1–86.

[6] Barndorff-Nielson, O. E. and Cox, D. R. (1989). *Asymptotic Techniques for Use in Statistics*. Chapman and Hall, London.

[7] Bartlett, M. S. (1963). Statistical estimation of density functions. *Sankhya A*, **25**, 245–254.

[8] Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory.* John Wiley & Sons, New York.

[9] Besag, J. E. (1975). Statistical analysis of non-lattice data. *The Statistician*, **24**, 179–195.

[10] Besag, J. E. (1977). Efficiency of pseudolikelihood estimation for simple Gaussian fields. *Biometrika*, **64**, 616–618.

[11] Beylkin, G. (1992). On the representation of operators in bases of compactly supported wavelets. *SIAM Journal on Numerical Analysis*, **29(6)**, 1716–1740.

[12] Billingsley, P. (1968). *Convergence of Probability Measures*, John Wiley & Sons, New York.

[13] Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods*, Spring-Verlag, New York.

[14] Cai, T. T. (1996). Minimax wavelet estimation via block thresholding. *Technical Report*, Department of Statistics, Purde University.

[15] Cai, T. T. (1999). Adaptive wavelet estimation: a block thresholding and oracle inequality approach. *Ann. Statist.*, **27**, 898–924.

[16] Cai, T. T. and Silverman, B. W. (2001). Incorporating information on neighboring coefficients into wavelet estimation. *Sankhya B*, **63** , 127–148.

[17] Cai, T. T. (2002). On block thresholding in wavelet regression: adaptivity, block size, and threshold level. *Statist. Sinica*, **12**, 1241–1273.

[18] Chatfield, C. (1975). *The Analysis of Time Series: Theory and Practice.* Chapman and Hall, London.

[19] Chipman, H. A., Kolaczyk, E. D. and McCulloch, R.E. (1997). Adaptive Bayesian wavelet shrinkage. *J. Am. Statist. Assoc.*, **92**, 1413–1421.

[20] Chuang, G. C.-H. and Kuo, C.-C. J. (1996). Wavelet Descriptor of Planar Curves: Theory and Applications. *IEEE Transactions on Image Processing*, **5(1)**, 56-70.

[21] Clyde, M. and George, E. I. (1999). Empirical Bayes estimation in wavelet nonparametric regression. In *Bayesian Inference in Wavelet-Based Models* (Lecture Notes in Statistics, Vol. 141), edited by Mller, P. and Vidakovic, B. Oppenheim, 309–322. Springer–Verlag, New York.

[22] Clyde, M. and George, E. I. (2000). Flexible empirical Bayes estimation for wavelets. *J. R. Statist. Soc. B*, **62**, 681–698.

[23] Clyde, M., Parmigiani, G. and Vidakovic, B. (1998). Multiple shrinkage and subset selection in wavelets. *Biometrika*, **85**, 391–402.

[24] Coifman, R. and Donoho, D. (1995). Translation-invariant de-noising. *Lecture Notes in Statistics: Wavelets and Statistics*, **Vol. 103**, 125–150, Springer–Verlag, New York.

[25] Cox, D. R. and Miller, H. D. (1968). *The Theory of Stochastic Processes.* Chapman and Hall, London.

[26] Daniels, H. E. (1987). Tail probability approximations. *Int. Statist. Rev.*, **55**, 37–48.

[27] Daubechies, I. (1992). *Ten Lectures on Wavelets.* SIAM, Philadelphia.

[28] De Canditiis, D. and Vidakovic, B. (2004). Wavelet Bayesian block shrinkage via mixtures of normal-inverse gamma priors. *Journal of Computational and Graphical Statistics*, **13(2)**, 383–398.

[29] Donoho, D. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**, 425–455.

[30] Donoho, D. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.*, **90**, 1200–1224.

[31] Donoho, D., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1995). Wavelet shrinkage: asymptopia? (with discussion). *J.R. Statist. Soc. B*, **57**, 301–369.

[32] Donoho, D. and Yu, T. (1997). Nonlinear " wavelet transform" via median-interpolation. *Technical Report*, Statistics Department, Stanford University.

[33] Donoho, D. L. and Johnstone, I. M. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.*, **26**, 879-921.

[34] Downie, T. R. and Silverman, B. W. (1998). The discrete multiple wavelet transform and thresholding methods. *IEEE Trans. Sig. Proc.*, **46**, 2558–2561.

[35] Dudley, R. M. and Haughton, D. (2002). Asymptotic normality with small relative errors of posterior probabilities of half-spaces. *Ann. Statist.*, **30**, 1311-1344.

[36] Engel, J. (1990). Density estimation with Haar serie. *Statistics & Probability Letters*, **9**, 111–117.

[37] Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications.* Chapman and Hall, London.

[38] Frazier, M., Jawerth, B. and Weiss, G. (1991). Littlewood-Paley theory and the study of function spaces *CBMS Regional Conference Series*, **79**.

[39] Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approches to calculating marginal densities. *J. Amer. Statist. Assoc.*, **85**, 398–409.

[40] Gnedenko, B.V. and Kolmogorov, A. N. (1968). *Limit Distributions for Sums of Independent Random Variables (revised edition).* Addison Wesley, Reading, MA.

[41] Good, I. J. and Gaskins, R. A. (1971). Nonparametric roughness penalties for probability densities. *Biometrika*, **58**, 255–277.

[42] Graps, A. (1997). An introduction to wavelets. http://www.amara.com /IEEEwave/IEEEwavelet.html.

[43] Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models, A roughness penelty Approach* . Chapman and Hall, London.

[44] Hall, P (1992). *The Bootstrap and Edgeworth Expansion.* Springer-Verlag, New York.

[45] Hall, P., Penev, S., Kerkyacharian, G. and Picard, D. (1997). Numerical performance of block threshold wavelet estimators. *Statist. Comput.*, **7**, 115–124.

[46] Hall, P., Kerkyacharian, G. and Picard, D. (1998). Block threshold rules for curve estimation using kernel and wavelet methods. *Ann. Statist.*, **26**, 922–942.

[47] Hall, P., Kerkyacharian, G. and Picard, D. (1999). On the minimax optimality of block thresholded wavelet estimators. *Statist. Sinica*, **9**, 33–50.

[48] Hall, P. and Penev, S. (2001). Cross-validation for choosing resolution level for nonlinear wavelet curve estimators. *Bernoulli*, **7**, 317–341.

[49] Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

[50] Jensen, J. L. (1995). *Saddlepoint Approximations.* Oxford University Press. Oxford.

[51] Johnson, N. L. and Kotz, S. (1970). *Continuous Univariate Distributions-2.* John Wiley and Sons, New York.

[52] Johnstone, I. M. and Silverman, B. W. (1997) Wavelet threshold estimators for data with correlated noise. *J. R. Statist. Soc. B*, **59** , 319-351.

[53] Johnstone, I. M. and Silverman, B. W. (1998). Empirical Bayes approaches to mixture problems and wavelet regression. *Technical Report*, School of Mathematics, University of Bristol, Bristol.

[54] Johnstone, I. M. and Silverman, B. W. (2005). Empirical Bayes selection of wavelet thresholds. *Annals of Statistics*, **33(4)**, 1700–1752.

[55] Kingman, J. F. C. and Taylor, S. J. (1966) *Introduction to Measure with Probability*, Cambridge University Press, New York.

[56] Johnson, N. L. and Kotz, S. (1970). *Continuous Univariate Distributions, Vol. 1*. John Wiley and Sons, New York.

[57] Kovac, A. and Silverman, B. W. (2000). Extending the scope of wavelet regression methods by coefficient-dependent thresholing. *J. Am. Statist. Ass.*, **95**,172-183.

[58] Kume, A., Dryden, I. L., and Le, H. (2004). Shape space smoothing splines for planar landmark data. *Technical Report*, Division of Statistics, University of Nottingham, Nottingham.

[59] Lawton, W. (1993). Applications of complex valued wavelet transforms to subband decomposition. *IEEE Trans. Sig. Proc.*, **41**, 3566–3568.

[60] Lina, J-M. and Mayrand, M. (1995). Complex Daubechies wavelets. *Applied and Computational Harmonic Analysis*, **2**, 219–229.

[61] Lugannani, R. and Rice, S. O. (1980). Saddlepoint approximation for the distribution of the sum of independent random variables. *Adv. Appl. Prob.*, **12**, 475–490.

[62] Mallat, S. G. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Pattn. Anal. Mach. Intell.*, **11**, 674-693.

[63] Marron, J. S., Adak, S., Johnstone, I. M., Neumann, M. H. and Patil, P. (1998). Exact risk analysis of wavelet regression. *Journal of Computational and Graphical Statistics*, **7(3)**,278–309.

[64] McCullagh, P. (1987). *Tensor methods in statistics.* Chapman and Hall, London

[65] Meyer, Y. (1992). *Wavelets and Operators.* Cambridge University Press, Cambridge.

[66] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953). Equations of state calculations by fast computing machines. *J. Chem. Physik.*, **21**, 1087–1091.

[67] Morlet, J., Arens, G., Fourgeau, E. and Giard, D. (1982). Wave propagation and sampling theory. *Geophysics*, **47(2)**, 203–236.

[68] Morris, C. (1983). Parametric empirical Bayes inference: theory and applications (with discussion). *J. Amer. Statist. Assoc.*, **78**, 47–65.

[69] Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory.* John Wiley, New York.

[70] Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability Applied*, **10**, 186–190.

[71] Narula, S. C. and Desu, M. M. (1981). Algorithm AS 170: computation of probability and non-centrality parameter of a non-central chi-squared distribution. *Applied Statistics*, **30**, 349-352.

[72] Nason, G. P. (1996). Wavelet shrinkage using cross-validation. *J.R. Statist. Soc. B*, **58**, 463–479.

[73] Nason, G. P. (1999). Fast corss-validation choice of wavelet smoothness, primary resolution and threshold in wavelet shrinkage using the Kovac-Silverman

algorithm. *Technical Report*, School of Mathematics, University of Bristol, Bristol.

[74] Nason, G. P. and Silverman, B. W. (1995). The stationary wavelet transform and some statistical applications. In *Wavelets and Statistics* (Lecture Notes in Statistics, Vol. 103), edited by A. Antoniadis and G. Oppenheim, 281–299. Springer–Verlag, New York.

[75] O'Hagan, A. (1994). *Bayesian Inference.* In *Kendall's Advanced Theory of Statistics*, **2B** Arnold, a member of the Hodder Headline Group, London.

[76] Percival, D. B. and Walden, A. T. (2000). *Wavelet Methods for Time Series Analysis.* Cambridge University Press, Cambridge.

[77] Reinsch, C. (1967). Smoothing by spline functions. *Numer. Math.*, **10**, 177–183.

[78] Schwartz, S. C. (1967). Estimation of probability density by an orthogonal series. *The Annals of Mathematical Statistics*, **38**, 1261–1265.

[79] Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

[80] Silverman, B. W. (1985). Some aspects of the spline smoothing approach to nonparametric regression. *J.R. Statist. Soc. B*, **47**, 1–52.

[81] Shapior, J. M. (1993). Embedded image coding using zerotree of wavelet coefficients. *IEEE trans. on Signal Processing*, **441**, 3445–3462.

[82] Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, **9**, 1135–1151.

[83] Stone, C. J. (1977). Consistent nonparametric regression. *Applied Statistics*, **5**, 595–635.

[84] Tarter, M. E. and Lock, M. D. (1993). *Model-Free Curve Estimation*, Chapman and Hall, London.

[85] Vannucci, M. and Corradi, F. (1999). Covariance structure of wavelet coefficients: theory and models in a Bayesian perspective. *J.R. Statist. Soc B*, **61**, 971–986.

[86] Vidakovic, B. (1998). Non-linear wavelet shrinkage with Bayes rules and Bayes factors. *J. Am. Statist. Ass.*, **93**, 173-179.

[87] Vidakovic, B. and Müller, P. (1995). Wavelet shrinkage with affine Bayes rules with applications. *Technical Report*, Duke University.

[88] Vidakovic, B. and Ruggeri, F. (2001). BAMS method: theory and simulations. *Sankhya B: The Indian Journal of Statistics (Special Issue on Wavelet Methods)*, **63**, 234-249.

[89] Watson, G. S. (1964) Smooth regression analysis. *Sankhya A*, *26*, 359–372.

[90] Wood, A. T. A., Booth, J. G. and Butler, R. W. (1993). Saddlepoint approximations to the CDF of some statistics with nonnormal limit distributions. *J. Am. Statist. Ass.*, **88**, 370-376.