Wavelet Methods and Inverse Problems

Hassan Musallam S Aljohani

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

The University of Leeds
Department of Statistics

March 2017

The candidate confirms that the work submitted is his own and that appropriate credit has been given where reference has been made to the work of others.

This thesis is dedicated to my family.

Acknowledgements

This thesis would not have been possible without the help of my supervisor. Dr Robert Aykroyd has always been the main source of support and confidence in my study and I have relied on his constant encouragement and love throughout my academic career. His kindness and support for me will always be a great source of inspiration in my life. I thank my wife, Fadwa, whose unconditional love, support and courage helped me withstand hardship and make the best decisions in my life.

Abstract

Archaeological investigations are designed to acquire information without damaging the archaeological site. Magnetometry is one of the important techniques for producing a surface grid of readings, which can be used to infer underground features. The inversion of this data, to give a fitted model, is an inverse problem. This type of problem can be ill-posed or ill-conditioned, making the estimation of model parameters less stable or even impossible. More precisely, the relationship between archaeological data and parameters is expressed by a likelihood. It is not possible to use the standard regression estimate obtained through the likelihood, which means that no maximum likelihood estimate exists. Instead, various constraints can be added through a prior distribution with an estimate produced using the posterior distribution. Current approaches incorporate prior information describing smoothness, which is not always appropriate. The biggest challenge is that the reconstruction of an archaeological site as a single layer requires various physical features such as depth and extent to be assumed. By applying a smoothing prior in the analysis of stratigraphy data, however, these features are not easily estimated. Wavelet analysis has proved to be highly efficient at eliciting information from noisy data. Additionally, complicated signals can be explained by interpreting only a small number of wavelet coefficients. It is possible that a modelling approach, which attempts to describe an underlying function in terms of a multi-level wavelet representation will be an improvement on standard techniques. Further, a new method proposed uses an *elastic-net* based distribution as the prior. Two methods are used to solve the problem, one is based on one-stage estimation and the other is based on two stages. The one-stage considers two approaches a single prior for all wavelet resolution levels and a level-dependent prior, with separate priors at each resolution level. In a simulation study and a real data analysis, all these techniques are compared to several existing methods. It is shown that the methodology using a single prior provides good reconstruction, comparable even to several established wavelet methods that use mixture priors.

Contents

List of Tables

List of Figures

Chapter 1

Introduction and outline

This chapter has two purposes: (i) to introduce the main problem with a brief explanation, and (ii) to outline the thesis.

The principle aim of this thesis is to promote fundamental understanding of parameter estimation, inverse problems and wavelet transforms, regarding such key issues as ill-posedness, ill-conditioning, regularisation and reconstruction. This thesis contributes to the investigation of Bayesian modelling and shrinkage in the wavelet domain.

The focus is on the inverse problem of finding parameters given observed data. Estimating parameters might involve solving ordinary least squares (OLS) or maximum likelihood (ML). It is known that these methods often perform poorly in both estimation and interpretation for such problems (Zou and Hastie, 2005).

Wavelets are a tool used in mathematics for non-parametric function estimation, which has been of great interest in various statistical applications. There are three main reasons for using wavelets. The first is that wavelets often offer a sparse and localized decomposition appropriate for non-parametric functions with various degrees of smoothness (Reményi, 2012). The second is that it is often possible to interpret a complicated signal using only a few wavelet coefficients, instead of the full data. Finally, the use of the Haar wavelet might help to produce a step function reconstruction, which is appropriate for

many archaeological data analysis problems.

Markov chain Monte Carlo (MCMC) algorithms have attracted much attention over the last two decades. Moreover, statisticians have been increasingly drawn to the MCMC approach to simulate from complex or nonstandard multivariate distributions (Chib and Greenberg, 1995). In particular, it can be a powerful computational tool in Bayesian inference due to its conceptual simplicity and relative ease of implementation. The major limitation of Bayesian approaches is that finding the posterior distribution often requires high-dimensional integration in situations too complex for an analytical solution. The basic idea of the MCMC method is to construct a Markov chain to generate pseudo-random samples whose stationary distribution follows the target posterior distribution. Then the sample is used for parameter estimation.

In this thesis, a new method is proposed using an elastic-net based distribution to model wavelet coefficients. This includes the popular Laplace and Gaussian distributions as special cases. The prior parameters in the model are described by prior distributions. The posterior mean (PM) and maximum a posteriori (MAP) estimates are obtained and provide a reconstruction from archaeological data corrupted by levels of noise and blur. The proposed methodology is compared to several established wavelet methods through extensive simulations.

This thesis is structured as follows. Chapter 2 provides background to the archaeological problems, inverse problems and wavelets. Three methods of inverse problem solution namely ridge regression, Lasso and smooth regularization are considered. Some applications of shrinkage methods, such as SURE and cross-validation, are discussed. Inversion and thresholding algorithms are studied using simulation.

Chapter 3 provides an introduction to multi-resolution analysis and defines different wavelet transforms, such as non-decimated and unbalanced Haar transforms. Additionally, illustrative examples are discussed.

In Chapter 4, the unbalanced Haar and SureBlock thresholding rule are presented. Also, different wavelet transforms described in the previous Chapter are applied to data using

classical thresholding rules. Additionally, extensive simulations on standard test functions are used to find the best wavelet transformation method.

Chapter 5 discusses traditional Bayesian thresholding using single priors. Then Chapter 6 describes traditional mixture priors. Many existing methods are presented, such as adaptive Bayesian wavelet shrinkage and Bayesian adaptive multi-resolution shrinkage.

In Chapter 7, a two-stage wavelet-based estimation method is proposed. Additionally, wavelet-vaguelette and vaguelette-wavelet approaches are studied. Several established wavelet methods are demonstrated and compared using extensive simulations.

Chapter 8 presents a one-stage wavelet-based estimation method using different priors, such as the Laplace, the Gaussian, the elastic-net and the double Weibull distributions. A detailed simulation comparison is presented.

Chapter 9 contains an application to a real-world data set from archaeological stratigraphy. The features of earth cores are estimated using different priors, with the prior parameters also estimated.

In Chapter 10, a new method is proposed for wavelet coefficients in two dimensions. The prior parameters in the model are estimated using prior distributions. The posterior mean and the maximum a posteriori estimates are obtained for an extensive real archaeological problem.

The final summary and conclusions of the thesis are given in Chapter 11. The thesis is concluded by an Appendix and References.

Chapter 2

Background: Inverse Problems and Wavelets

2.1 Overview

This chapter is divided into four parts: the first part, containing only Section 2.2 gives background to archaeological prospecting, the second part, containing Sections 2.3 to 2.7, considers inverse problems, the third part, containing Sections 2.8 to 2.12 provides an introduction to wavelet methods. Finally, the fourth part, containing Sections 2.13 and 2.14, shows a comparison simulation and presents conclusions.

2.2 Archaeological prospecting

Archaeological prospecting aims to identify features buried at archaeological sites. Additionally, it refers to the discovery of cultural information from materials such as wood or ditches that have been covered by earth or sand. One of the key problems archaeologists face is how to collect all possible information about past human activity or habitation that is available at archaeological sites without destroying any evidence, and how to do

so in an efficient and accurate manner. Surface surveying is the chief modern methodology for collecting information (Ammerman and Feldman, 1978), compared to the older method of excavation, which involved digging one hole or many small holes. This strategy varies according to the way that archaeologists imagine the past and visualise what lies under the ground (Renfrew and Bahn, 2013). However, excavation is a simple method and a rudimentary technique for discovering past human activity. Nowadays, modern archaeological techniques are designed to minimise damage to the archaeological site. In particular, some of these methods rely on the indirect detection of magnetic susceptibility. During recent decades, magnetic prospecting has become an important tool for understanding, describing and classifying a wide range of features. Magnetic susceptibility plays a key role in understanding the subsurface and may also provide important information on soil composition (Le Borgne, 1955; Mullins, 1977). Magnetic susceptibility S is defined as the ratio

$$S = \frac{M}{F},$$

where M and F are the induced magnetization and magnetic field respectively. Since M and F have the same SI (International System of Units) units, S is a dimensionless number (Evans and Heller, 2003). Usually, its magnitude is of the order of 10^{-6} to 10^{-5}. Magnetic susceptibility is essentially controlled by a small group of iron-bearing minerals (Hanesch and Scholger, 2002; Karimi *et al.*, 2011; Leslie-Pelecky and Rieke, 1996). Just as materials and objects can be explained by their size, colour or chemical composition, they may also be defined by their magnetic properties (Dearing, 1994).

Magnetic surveying is one of a number of methods used in archaeological geophysics and it is a fast and dependable prospecting technique for detecting and mapping the distribution in the shallow subsurface (Scollar *et al.*, 1990). The Earth can be described as a large magnet with the north pole pointing south. The intensity of the Earth's magnetic field ranges from 30,000 to 60,000 nanoTesla, whereas local anomalies are around 10 nanoTesla (Caruso and Withanawasam, 1999; Le Borgne, 1955). The typical targets of magnetic surveys are the buried remnants of settlements and towns ranging in age from

Neolithic to medieval (Needham, 1985). Farming settlements and towns become the most frequent features in the landscape. These early settlements consist of groups of houses and other objects usually set amid fields enclosed by earth banks or stone walls (Lewis *et al.*, 1997). The spatial variation of magnetic surveying is impressive and complex patterns can be revealed from measurements, which are simple, fast and cheap, non-destructive and possible both in the laboratory and in the field. Furthermore, it can be combined with other physical quantities to allow reliable interpretation (Girault *et al.*, 2011).

There are different kinds of tools suitable for magnetic surveying, which have been constructed and used in the past. Examples include the free proton magnetometer, the fluxgate gradiometer, the alkali-vapour magnetometer, and the optically pumped magnetometer. The fluxgate gradiometer was the first electric magnetometer, which was used for military and geophysical exploration in the 1930s (Scollar *et al.*, 1990) but it remains the most popular piece of electronic equipment. It is made by aligning two fluxgate magnetometers above each other, and hence measures the vertical component of the earth's magnetic field, producing a zero reading in a constant field. It is carried vertically with the lower sensor at a height above the ground between 20-30 cm, whereas the distance between the two sensors is between 50-125 cm. It can be used to cover a single grid in 15-20 minutes. Figure 2.1 shows the magnetometer data from the Park, Guiting Power. In particular, this shows a diagonal linear ditch towards the top, a rectangular boundary ditch surrounding various collections of circular pits and post-holes (Allum, 1997).

The external flux of a magnet opposes the polarity induced in it by the earth's magnetic field. Figure 2.2 shows the inducing magnetic field; every positive anomaly is therefore accompanied by a negative one of lesser magnitude. In the northern hemisphere, where the angle of inclination is positive, that is a downward dip, the negative anomaly is greater to the northern side of the feature than to the south; the effect is reversed for negative inclinations in southern latitudes (Allum, 1997). Also, as the distance between the magnetometer and an object decreases then the magnetic readings move closer to zero.

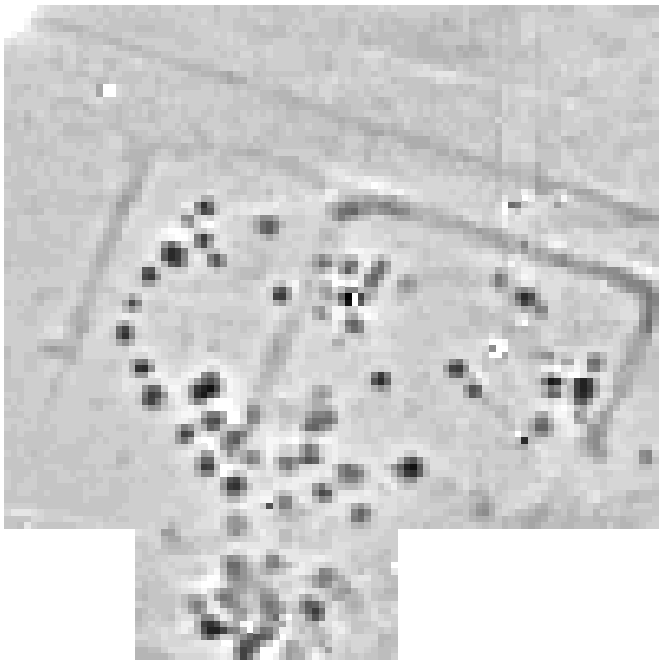


Figure 2.1: Data collected in 1994 from "The Park", at Guiting Power in Gloucestershire, using a fluxgate gradiometer (Allum *et al.*, 1999).

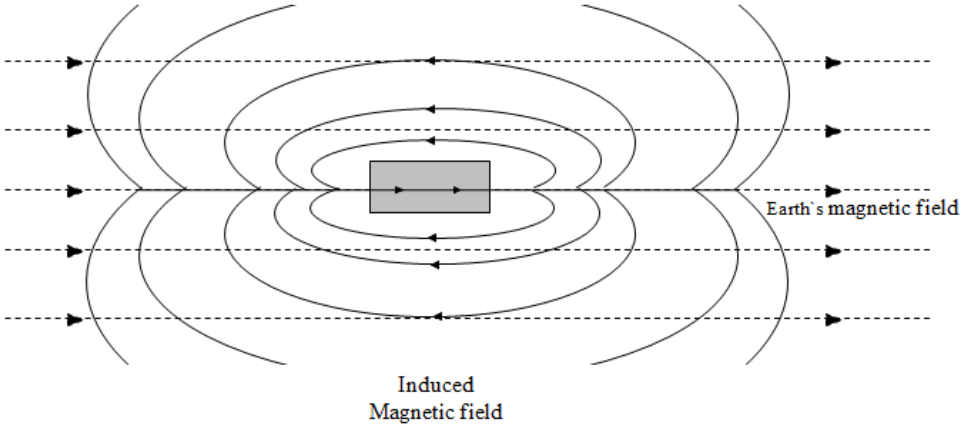


Figure 2.2: Diagram of the inducing magnetic field: the distinctive shape of a positive and a negative region occurs as a result of the magnetic susceptibility behaving as a bar magnet.

Type of soil	Magnetic susceptibility (10^{-6} emu)
Limestone, unbaked clays	10
Subsoils	10-100
Topsoils	100-100
Heated soils, fired clays, volcanic rocks	1000-5000

Table 2.1: Magnetic susceptibility of different kinds of soil where emu means electromagnetic unit (Tite and Linington, 1975).

Some minerals, such as iron oxide magnetite, are highly magnetic and are attracted to a nearby magnet, with some materials containing more magnetic minerals than others. For example, rocks with relatively high concentrations of magnetite, i.e. heated soils, fired clays and volcanic rocks, have much higher magnetic susceptibility values than limestone, some unbaked clays and sand, which do not show any visible attraction to a magnet. There are many results from studying English soils (Mullins, 1977; Tite and Linington, 1975), which are consistent with the above observations. Table 2.1 shows some typical values of the magnetic susceptibility of different soil types taken from archaeological sites across England. The term emu is short for "electromagnetic unit" and is not a unit in the conventional sense (Renfrew and Bahn, 2013).

Several methods are available for reconstruction of a true signal by inversion, such as regularization methods and Wiener Fourier method. For low-rank problems however, these prove to be unsatisfactory, since high frequency elements in the signal are smoothed but almost certainly exist in the recorded data due to noise (Allum *et al.*, 1999). Additionally, although these methods may reduce the broad spread of the peak, the reconstruction is still very smooth and it is therefore difficult to provide a sharp division between regions of different susceptibility.

2.3 Introduction to inverse problems

Inverse problems are universal in science and engineering, and have received a great deal of attention from scientists in areas such as geophysics, engineering and medicine. Moreover, inverse problems are a challenge in statistics and a number of methods have previously been proposed. There are two main types of inverse problems, linear inverse problems defined by

$$\mathbf{y} = \mathbf{H}\mathbf{f} + \boldsymbol{\epsilon}, \tag{2.1}$$

and non-linear inverse problems

$$\mathbf{y} = \phi(\mathbf{H}, \mathbf{f}) + \boldsymbol{\epsilon}, \tag{2.2}$$

depending on the type of relationship between response variable $\mathbf{y}_{n\times 1} = \{y_i : i = 1, 2, \ldots, n\}$ and independent variables $\mathbf{H}_{n\times m} = \{h_{i,j} : i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, m\}$. Also, $\boldsymbol{\epsilon} = \{\epsilon_i : i = 1, 2, \ldots, n\}$ are the measurement errors, and $\mathbf{f} = \{f_j : j = 1, 2, \ldots, m\}$ are the unknown model parameters. Note that our aim is not only to fit a model to allow the prediction of $\mathbf{y}$, but to interpret the estimates of $\mathbf{f}$. In each type of inverse problem estimation of the unknown parameter vector $\mathbf{f}_{m\times 1} = \{f_j : j = 1, 2, \ldots, m\}$ is not straight forward as either: (i) no solution exists; (ii) there are multiple solutions, or; (iii) the solution does not depend smoothly on the data, as small changes in the noise can lead to wildly different estimates – these properties define an ill-posed or ill-conditioned inverse problem (Hadamard, 2014). In the usual maximum likelihood approach for a linear problem, with normally distributed errors, the estimator of $\mathbf{f}$ is given by

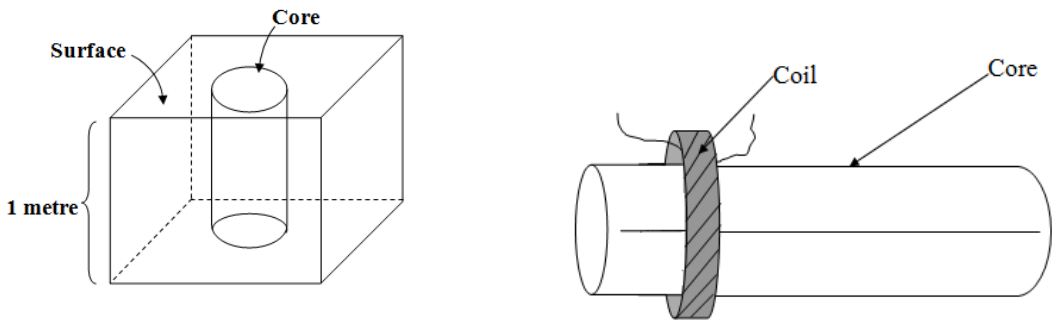
$$\widehat{\mathbf{f}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{y}.$$

With an inverse problem there are two possible reasons why the inverse cannot be found: the first reason is that m is greater than n, that is the number of parameters is larger than the number of observations – this is an ill-posed problem; the second reason is that even when m is less than n, there are still problems, due to collinearity, which is the condition where the independent variables are strongly correlated with each other – this is an ill-conditioned problem.

2.4 Inverse problems in archaeology

As motivation for this work, consider the following inverse problem, which occurs in archaeological geophysics; for an example, see Aykroyd and Al-Gezeri (2014). Consider an archaeological site, where all features above ground have been removed and it is no longer possible to see the location of buildings and ditches. However, the concentration of magnetic oxides of iron in the topsoil has been increased by human occupation of the site. Over time, and after the site has been abandoned, the topsoil fills ditches, post holes and wall trenches, creating an iron concentration contrast, which can be detected by magnetometer readings. In archaeological stratigraphy, a core sample of the ground is taken using a soil borer, which can penetrate over 1 metre deep; then a plastic cylinder is pushed into the soil and sealed on site to prevent the sample from drying out (Al-Gezeri, 2003).

On most occasions, no variation is observed between the layers in the core with regards to colour or texture. Nevertheless, the depth and vertical extent of occupation layers can still be estimated by measuring the magnetic properties of slices of the core at varying depths. An increase in the magnetic oxide concentration of the core slices leads to an increase in the magnetic susceptibility. Using a piece of equipment called a coil magnetometer, the local variation in susceptibility can be detected along the full length of the core. The plastic cylinder is positioned a small distance from one end of the coil; then it is moved in small steps of equal size, pausing between movements for readings to be made. Let the output readings be denoted by $\mathbf{y} = \{y_i : i = 1, 2, \ldots, n\}$. The first few measurements are recorded well before the core enters the coil and the last few well after it has emerged. Therefore, it may be assumed that initially and finally, the core has a negligible effect on the coil and hence the first few and last few readings are zero. Let the true magnetic susceptibilities be denoted by $\mathbf{f} = \{f_j : j = 1, 2, \ldots, m\}$. We might imagine that the core can be divided into parts, each part can be described by f_j where $j = 1, \ldots, m$. It is assumed that the first element, f_1, is at the centre of the coil when the first reading, y_1, is taken, and the last element, f_m, is there when the last reading, y_n, is taken.



(i) (ii)

Figure 2.3: Diagram showing removal of the core (i), and the geometry of core and coil (ii).

The susceptibility is indirectly measured by detecting small changes in the inductance of the coil as the core passes through. Although the detector coil is sensitive to the magnetic susceptibility along the full length of the core, it is much more sensitive to the part within the coil. Allum *et al.* (1999) show that a suitable approximation for the change in the inductance is

$$h = \frac{1}{4w}\left[\frac{d+w}{\sqrt{(r-a)^2+(d+w)^2}} - \frac{d-w}{\sqrt{(r-a)^2+(d-w)^2}}\right], \tag{2.3}$$

where a is the radius of the core, and r and $2w$ are the radius and the length of the coil respectively. With an extended core made of many elements, d, is replaced by d_{ij} in (2.3), and likewise h by h_{ij}, where d_{ij} is the distance along the coil axis, from the centre of the coil to the position of core element j at step i. The observed measurement, y_i, is then given by

$$y_i = \sum_{j=1}^{m} h_{ij} f_j + \epsilon_i, \quad \text{for} \quad i = 1, \ldots, n, \tag{2.4}$$

where $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, with ϵ_i and ϵ_j $(i \neq j)$ independent. Scollar (1970) states that the reason for including errors is to describe internal noise in the surveying instrument, machine-rounding errors and disturbance by superficial features such as small stones in

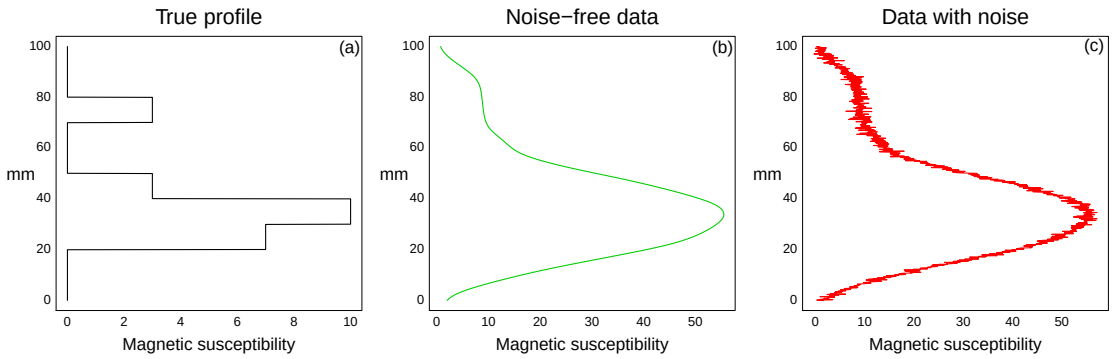


Figure 2.4: Simulated core: the true susceptibility (a) is made from blocks with known susceptibility; (b) shows the noise-free data corrupted by the matrix defined by (2.3), with parameters $r = 22, a = 18$, and $w = 17.5$; and (c) shows the observed data with noise.

the soil. This model can also be written

$$\mathbf{y} = \mathbf{H}\mathbf{f} + \boldsymbol{\epsilon}, \tag{2.5}$$

where $\mathbf{H}$ is known as the blur or kernel matrix with element h_{ij} and $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$.

As an illustration consider simulated data generated from a core, which is made from blocks of known susceptibility. Figure 2.4 (a) shows such a true magnetic susceptibility profile, (b) shows noise-free data, and (c) shows data with noise. Although the approximate location of the two main features can still be seen, the detail has been completely lost and it is no longer possible to see the start and end of each layer. From the data alone it would not be possible for the archaeologists to reliably determine the likely occupation of the site, nor to provide information to direct physical excavation.

2.5 Test functions

In this section two test functions will be described, which will be used as magnetic susceptibility profiles to generate simulated data. These will be used in later sections to illustrate various data analysis methods. The Blocks and the Bumps test functions (Donoho and

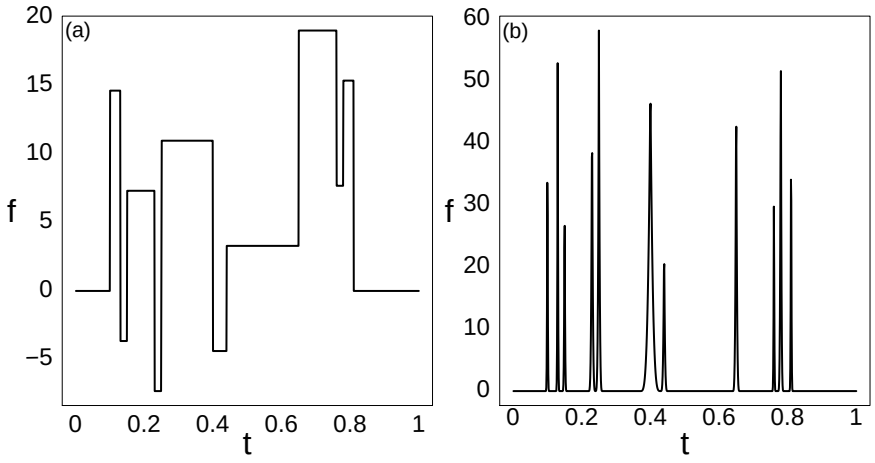


Figure 2.5: Plots of the Blocks (a) and the Bumps (b) test functions at $m = 1024$ equally spaced points.

Johnstone, 1994) are shown in Figure 2.5. The reason for choosing these functions is that we believe the magnetic susceptibility can be well described as piecewise constant blocks and we expect isolated highly magnetic objects to be well described by isolated Bumps.

To investigate the effect of blurring, a simple alternative to Equation (2.3) to build the kernel matrix, can be defined as

$$h_{ij} = C \exp\Big(- \frac{|i-j|}{k} \Big), \quad i = 1, \ldots, n, j = 1, \ldots, m, \tag{2.6}$$

where C and k are positive parameters. Here, $|i-j|$ is the distance along the x-axis, from the centre at step i to the position of element j. Changing one parameter value k, leads to a different point spread function and hence the amount of blurring is more easily controlled. From Chapter 2 to Chapter 8, the form of blur which is given in (2.6) will be used to produce simulated data.

Let $\mathbf{f} = \{f(j/m) : j = 1, 2, \ldots, m\}$ represent the value of the unknown magnetic susceptibility at a set of m equally spaced points. Suppose a set of noisy data $\mathbf{y} = \{y_i : i = 1, 2, \ldots, n\}$ are recorded at the same locations, hence $m = n$, then the model is given by

$$\mathbf{y} = \mathbf{f} + \boldsymbol{\epsilon}, \tag{2.7}$$

where the error $\boldsymbol{\epsilon}$ is a vector of random variables, such that $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I_n})$. Similarly,

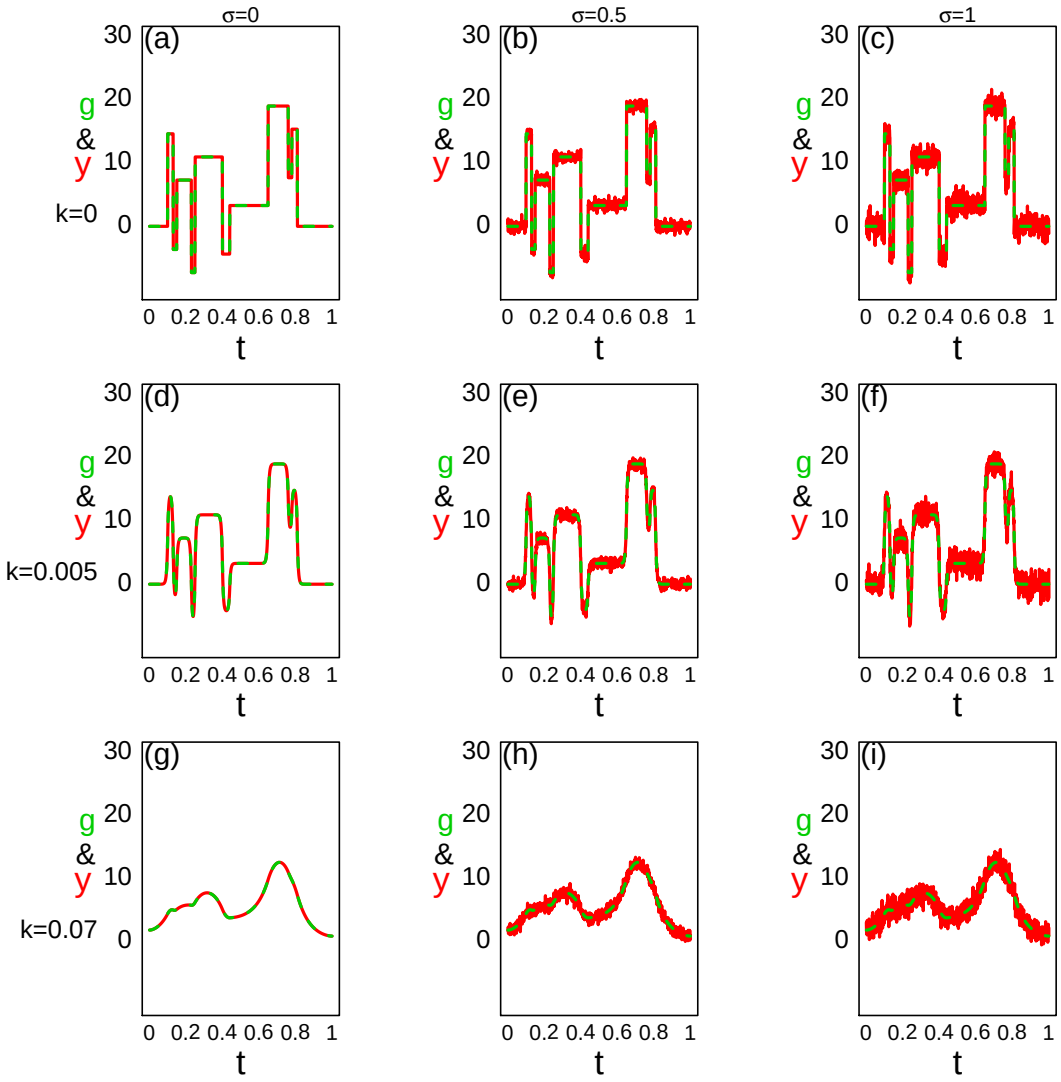


Figure 2.6: Plots of Blocks test function with different levels of blur as defined in (2.6), where $k = 0, 0.005, 0.07$ (rows), and noise, $\sigma = 0, 0.5, 1$ (columns), at $m = 1024$ equally spaced points.

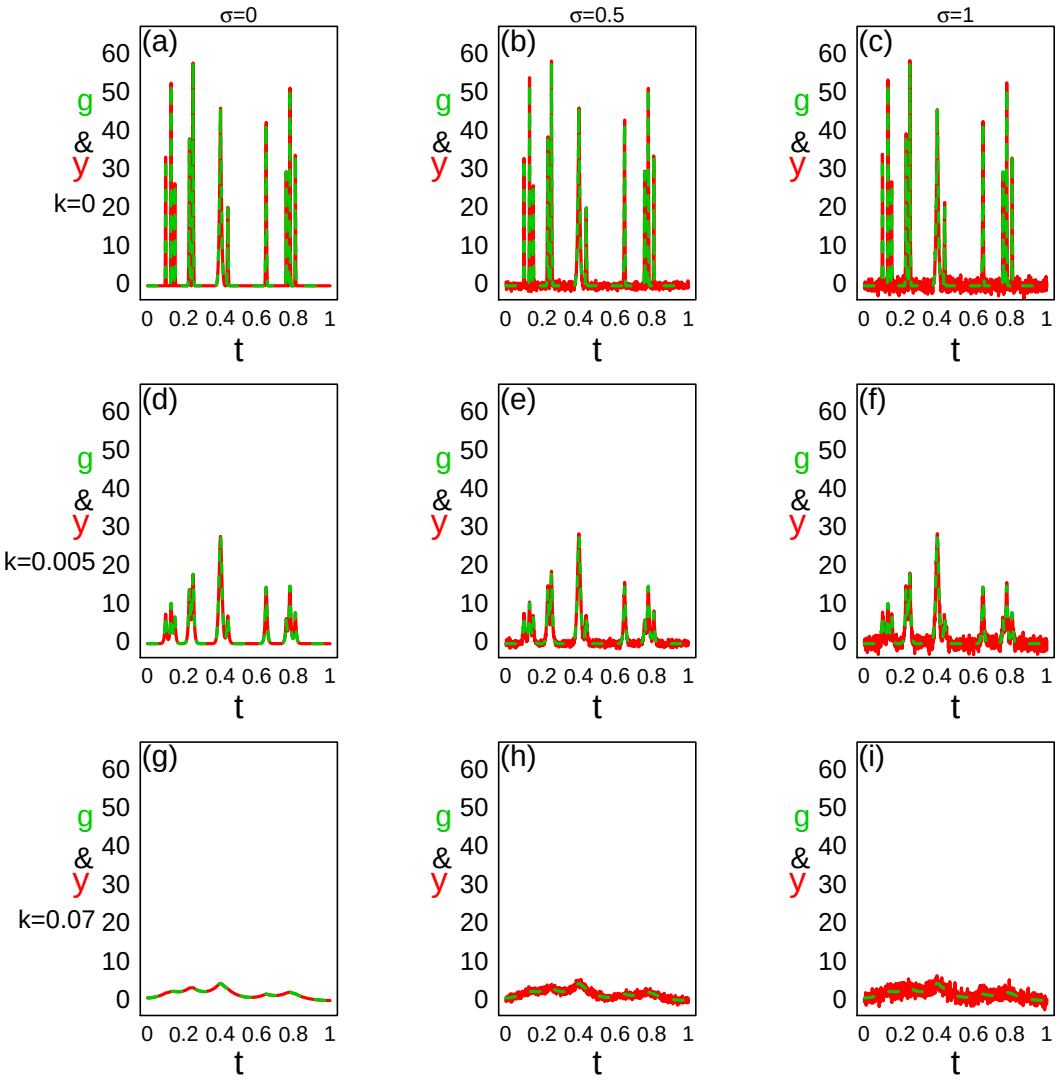


Figure 2.7: Plots of the Bumps test function with different levels of blur as defined in (2.6), where $k = 0, 0.005, 0.07$ (rows), and noise, $\sigma = 0, 0.5, 1$ (columns), at $m = 1024$ equally spaced points.

if the model is corrupted by noise and blur then the model is given by

$$\begin{aligned}\mathbf{y} &= \mathbf{H}\mathbf{f} + \boldsymbol{\epsilon} \\ &= \mathbf{g} + \boldsymbol{\epsilon},\end{aligned} \tag{2.8}$$

where $\mathbf{H}_{n\times m}$ is a given blur matrix and $\mathbf{g}_{n\times 1}$ is blurred noise-free data.

To investigate the effect of blur and noise, consider Figures 2.6 and 2.7, in particular $k = 0, 0.005, 0.07$ (rows) and $\sigma = 0, 0.5, 1$ (columns). For the Blocks test function, Figure 2.6, as the blur increases detail is lost and eventually only the vague appearance of general structure can be seen. Similarly, as the noise level is increased detail is hidden by the random variability. For the Bumps test function, Figure 2.7, as the blur increases the clusters of spikes merge and as the noise increases the remaining spikes become harder to distinguish from the random variability.

2.6 Inverse estimation

Considering the linear model (2.8), the log-likelihood is

$$\ell(\mathbf{f}) = -\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{H}\mathbf{f})^T(\mathbf{y} - \mathbf{H}\mathbf{f}) + \mathbf{C}, \tag{2.9}$$

and the maximum likelihood estimate of $\mathbf{f}$, is given by

$$\widehat{\mathbf{f}}^{\mathrm{ML}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{y}. \tag{2.10}$$

In many inverse problems, however, it is not possible to calculate the inverse $(\mathbf{H}^T\mathbf{H})^{-1}$, as the system has fewer equations than unknowns, that is $m > n$, or is ill-conditioned being nearly multicollinear. To solve this problem, additional constraints are introduced leading to a penalised log-likelihood

$$\ell_m(\mathbf{f}, \alpha) = -\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{H}\mathbf{f})^T(\mathbf{y} - \mathbf{H}\mathbf{f}) - \frac{1}{2}\kappa R(\mathbf{f}) + \mathbf{C}, \quad \kappa > 0, \tag{2.11}$$

where $R(\mathbf{f})$ is a penalty function with small values of $R(\mathbf{f})$ indicating preferred choices of $\mathbf{f}$. The parameter κ is chosen to balance the relative weight given to the likelihood

and penalty. Before moving on, it is worth noting that the penalised log-likelihood can be interpreted in a Bayesian setting as log-likelihood plus log-prior, hence the penalty function can be used to define a prior distribution and vice versa. Some examples of penalised likelihood will now be considered.

In ridge regression (Hoerl and Kennard, 1970) $R(\mathbf{f}) = \mathbf{f}^T\mathbf{f} = \sum_j^m f_j^2$ leading to the estimate

$$\widehat{\mathbf{f}}^{\text{Ridge}} = (\mathbf{H}^T\mathbf{H} + \Lambda\mathbf{I}_m)^{-1}\mathbf{H}^T\mathbf{y}, \tag{2.12}$$

where $\Lambda = \sigma^2\kappa$ and $\mathbf{I_m}$ is the $m \times m$ identity matrix, which can also be written as

$$\widehat{f}_i^{\text{Ridge}} = \frac{\widehat{f}_i^{\text{ML}}}{1+\Lambda}, \tag{2.13}$$

where $\widehat{f}_i^{\text{ML}}$ is defined in (2.10), when the matrix $\mathbf{H}$ is orthonormal (Filzmoser and Croux, 2002). To find the mean and variance of $\widehat{\mathbf{f}}^{\text{Ridge}}$, let $\mathbf{K} = \mathbf{H}^T\mathbf{H}$ so

$$\begin{aligned}
\widehat{\mathbf{f}}^{\text{Ridge}} &= (\mathbf{H}^T\mathbf{H} + \Lambda\mathbf{I}_m)^{-1}\mathbf{H}^T\mathbf{y} \\
&= (\mathbf{K} + \Lambda\mathbf{I}_m)^{-1}\mathbf{I}_m\mathbf{H}^T\mathbf{y} \\
&= (\mathbf{K}(\mathbf{I}_m + \Lambda\mathbf{K}^{-1}))^{-1}\mathbf{K}\mathbf{K}^{-1}\mathbf{H}^T\mathbf{y} \\
&= (\mathbf{I}_m + \Lambda\mathbf{K}^{-1})^{-1}\mathbf{K}^{-1}\mathbf{K}(\mathbf{K}^{-1}\mathbf{H}^T\mathbf{y}) \\
&= (\mathbf{I}_m + \Lambda\mathbf{K}^{-1})^{-1}\mathbf{I}_m((\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{y}) \\
&= (\mathbf{I}_m + \Lambda\mathbf{K}^{-1})^{-1}((\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{y}),
\end{aligned} \tag{2.14}$$

then the expectation of $\widehat{\mathbf{f}}$ is

$$\begin{aligned}
\mathbb{E}\{\widehat{\mathbf{f}}^{\text{Ridge}}\} &= \mathbb{E}\{(\mathbf{I}_m + \Lambda\mathbf{K}^{-1})^{-1}((\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{y})\} \\
&= \mathbb{E}\{(\mathbf{I}_m + \Lambda\mathbf{K}^{-1})^{-1}((\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{H}\mathbf{f} + \boldsymbol{\epsilon})\} \\
&= (\mathbf{I}_m + \Lambda(\mathbf{H}^T\mathbf{H})^{-1})^{-1}\mathbf{f}.
\end{aligned} \tag{2.15}$$

Notice that

$$\mathbb{E}\{\widehat{\mathbf{f}}^{\text{Ridge}}\} \neq \mathbf{f}, \tag{2.16}$$

and so the estimator $\widehat{\mathbf{f}}^{\text{Ridge}}$ is biased with bias given by

$$\begin{aligned}\mathbb{E}\{\widehat{\mathbf{f}}^{\text{Ridge}}\} - \mathbf{f} &= (\mathbf{I}_m + \Lambda(\mathbf{H}^T\mathbf{H})^{-1})^{-1}\mathbf{f} - \mathbf{f} \\ &= ((\mathbf{I}_m + \Lambda(\mathbf{H}^T\mathbf{H})^{-1})^{-1} - \mathbf{I}_m)\mathbf{f}. \end{aligned} \tag{2.17}$$

The variance of the ridge regression estimator is

$$\begin{aligned}\text{var}(\widehat{\mathbf{f}}^{\text{Ridge}}) &= \text{var}\{(\mathbf{H}^T\mathbf{H} + \Lambda\mathbf{I}_m)^{-1}\mathbf{H}^T\mathbf{y}\} \\ &= (\mathbf{H}^T\mathbf{H} + \Lambda\mathbf{I}_m)^{-1}\mathbf{H}^T\text{var}(\mathbf{y})((\mathbf{H}^T\mathbf{H} + \Lambda\mathbf{I}_m)^{-1}\mathbf{H}^T)^T \\ &= (\mathbf{H}^T\mathbf{H} + \Lambda\mathbf{I}_m)^{-1}\mathbf{H}^T\sigma^2\mathbf{I}_n\mathbf{H}((\mathbf{H}^T\mathbf{H} + \Lambda\mathbf{I}_m)^{-1})^T \\ &= \sigma^2(\mathbf{H}^T\mathbf{H} + \Lambda\mathbf{I}_m)^{-1}\mathbf{H}^T\mathbf{H}((\mathbf{H}^T\mathbf{H} + \Lambda\mathbf{I}_m)^{-1})^T. \end{aligned} \tag{2.18}$$

A second example of penalised likelihood is the Lasso with $R(\mathbf{f}) = ||\mathbf{f}||_1 = \sum |f_j|$, which was proposed by Tibshirani (1996). Although, in general, no closed form solution exists (Filzmoser and Croux, 2002), if the maximum likelihood solution, $\widehat{\mathbf{f}}^{\text{ML}}$, exists and the matrix $\mathbf{H}$ is orthonormal, then it is possible to write

$$\widehat{f}_i^{\text{Lasso}} = \begin{cases} \widehat{f}_i^{\text{ML}} - \Lambda, & \text{if } \widehat{f}_i^{\text{ML}} > \Lambda \\ \widehat{f}_i^{\text{ML}} + \Lambda, & \text{if } \widehat{f}_i^{\text{ML}} < -\Lambda \\ 0, & \text{otherwise,} \end{cases} \tag{2.19}$$

where $\widehat{f}_i^{\text{ML}}$ is defined in (2.10), and $\Lambda = \sigma^2\kappa$ is a parameter of the inversion method. Clearly, any maximum likelihood estimate smaller than Λ in magnitude will be set to zero and as Λ is increased, more and more values will be set to zero. From this it is clear that Lasso is related to thresholding, which will be described for wavelet coefficients in Chapter 4.

In a more general penalised likelihood approach $R(\mathbf{f}) = ||\mathfrak{R}\mathbf{f}||_2^2$ is proposed (Levine*et al.*, 1979), where the matrix $\mathfrak{R}$ can take different definitions, leading to the estimate

$$\widehat{\mathbf{f}} = (\mathbf{H}^T\mathbf{H} + \Lambda\mathfrak{R}^T\mathfrak{R})^{-1}\mathbf{H}^T\mathbf{y},$$

where $\Lambda = \sigma^2\kappa$. Two common choices of $\mathfrak{R}$, which can be derived, are based on a priori assumptions about the smoothness of the true function. If we believe that the function is

not different from a constant, then

$$R(\mathbf{f}) = \sum_{i=1}^{m-1} (f_i - f_{i+1})^2. \tag{2.20}$$

Equation (2.20) equals zero only when $\mathbf{f}$ is constant. Then, matrix $\mathfrak{R}_1$ can be defined as

$$\mathfrak{R}_1 = \begin{pmatrix} 1 & -1 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & -1 & 1 \end{pmatrix}_{(m-1)\times m}.$$

This leads to what is called first-order smoothing. If we believe that the function is linear, then

$$R(\mathbf{f}) = \sum_{i=1}^{m-2} (f_i - 2f_{i+1} + f_{i+2})^2. \tag{2.21}$$

Then, matrix $\mathfrak{R}_2$ can be defined as

$$\mathfrak{R}_2 = \begin{pmatrix} 1 & -2 & 1 & 0 & \dots & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 & -2 & 1 \end{pmatrix}_{(m-2)\times m},$$

which leads to second-order smoothing. A similar approach can lead to higher order smoothing and corresponding definitions of $\mathfrak{R}$. Hence, ridge regression is a special case of regularization, when $\mathfrak{R} = \mathbf{I}$.

2.7 A simulation comparison of inversion methods

All penalised likelihood estimators will now be evaluated and investigated. Here three methods for finding the inverse solution are considered: ridge regression, first order, and

second order smoothing. More precisely, the whole simulation and estimation procedure will be replicated $R = 1000$ times and the average mean squared-error (AMSE) calculated with

$$\text{AMSE} = \frac{1}{Rm} \sum_{i=1}^{R} \sum_{j=1}^{m} (\widehat{f}_j^i - f_j)^2, \tag{2.22}$$

where $\{\widehat{f}_j^i, j = 1, \ldots, m\}$ is the estimate of the true function from the i^{th} replicate. It is important to know that m is the length of $\mathbf{f}$ and R is the number of replicates. Algorithm 1 shows the main idea of the simulation. The simulated datasets consist of the standard test signals Blocks and Bumps (Donoho and Johnstone, 1994; Nason and Silverman, 1994) at $m = 128$ equally spaced points, multiplied by a blur matrix, and the value of blur defined in (2.6) by taking in turn k equal to 0.001, 0.005, and 0.01. Furthermore, datasets were corrupted by independent Gaussian noise with mean zero and standard deviation σ taken in turn as 0, 0.5, and 1. A range of smoothing parameter values Λ in the interval $[0, 3]$ were tested and results are presented. Figure 2.8 shows the boxplots of the mean squared-error (MSE), for the Blocks test function, with the minimum AMSE marked.

Algorithm 1: AMSE algorithm

Result: AMSE.

```
Let Λ ⊂ ℝ⁺.
for i=1 to length(Λ) do
    for j=1 to R do
        Generate ε ~ N_n(0, σ²I_n)
        y = Hf + ε
        Compute f̂ = I_*(y, H, Λ_i), I_*(·) is an inverse method
        Compute MSE
    end
    Compute the AMSE
end
```

Let $\widehat{\Lambda}_{\text{MAMSE}}$ be the value of Λ leading to the minimum AMSE. By studying the behaviour

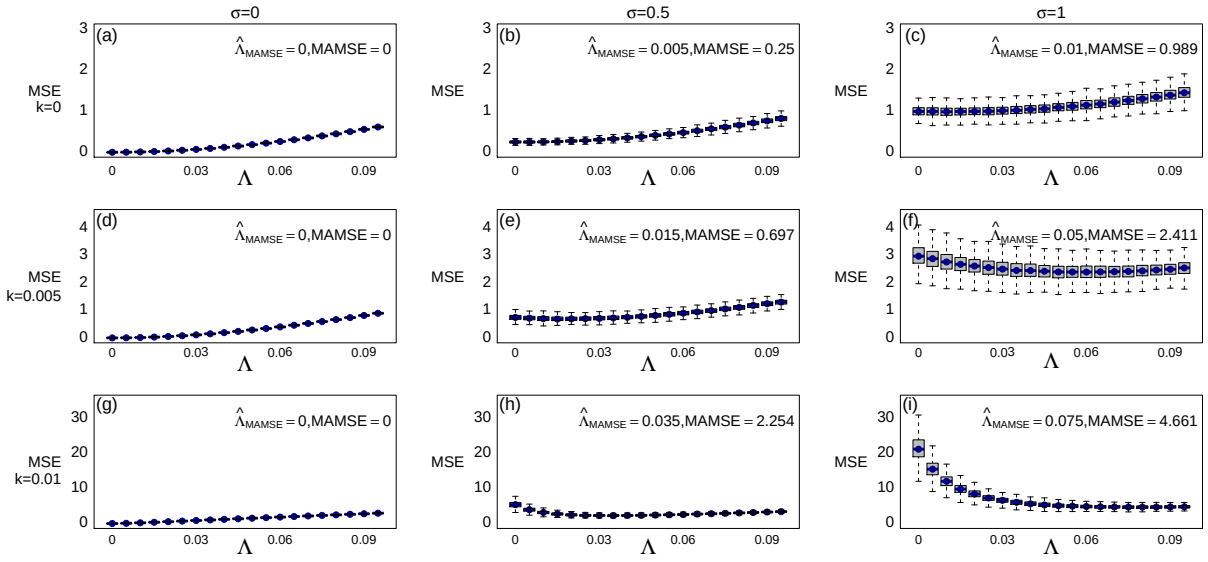


Figure 2.8: Boxplots of MSE as function of Λ for ridge regression using the Blocks test function with different blur, which is given in (2.6), and noise levels.

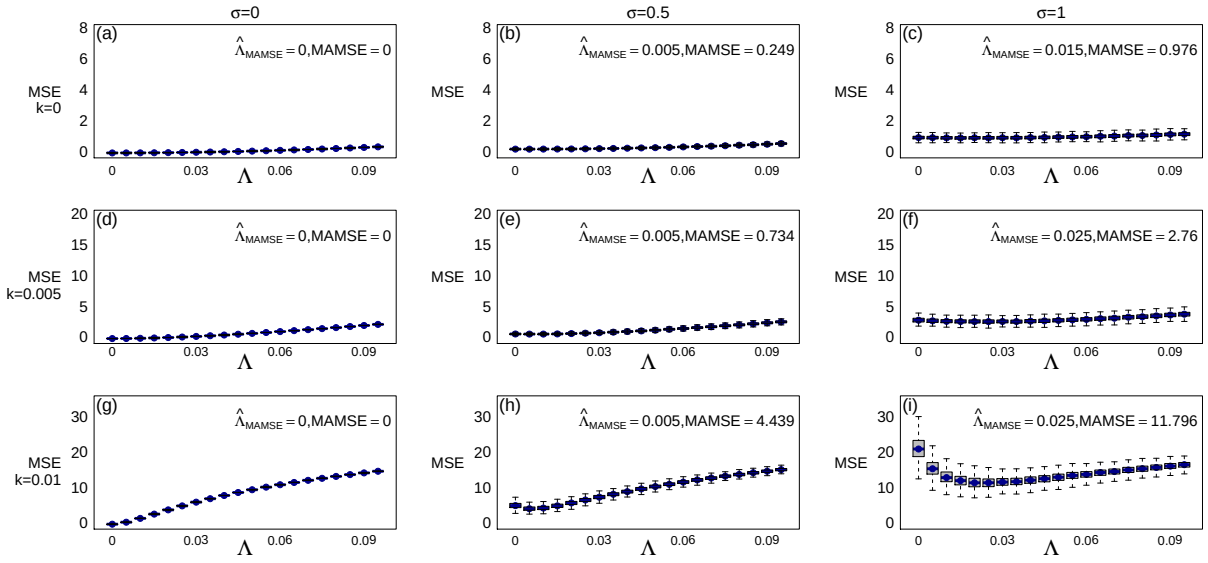


Figure 2.9: Boxplots of MSE results as function of Λ for ridge regression using the Bumps test function with different blur, which is given in (2.6), and noise levels.

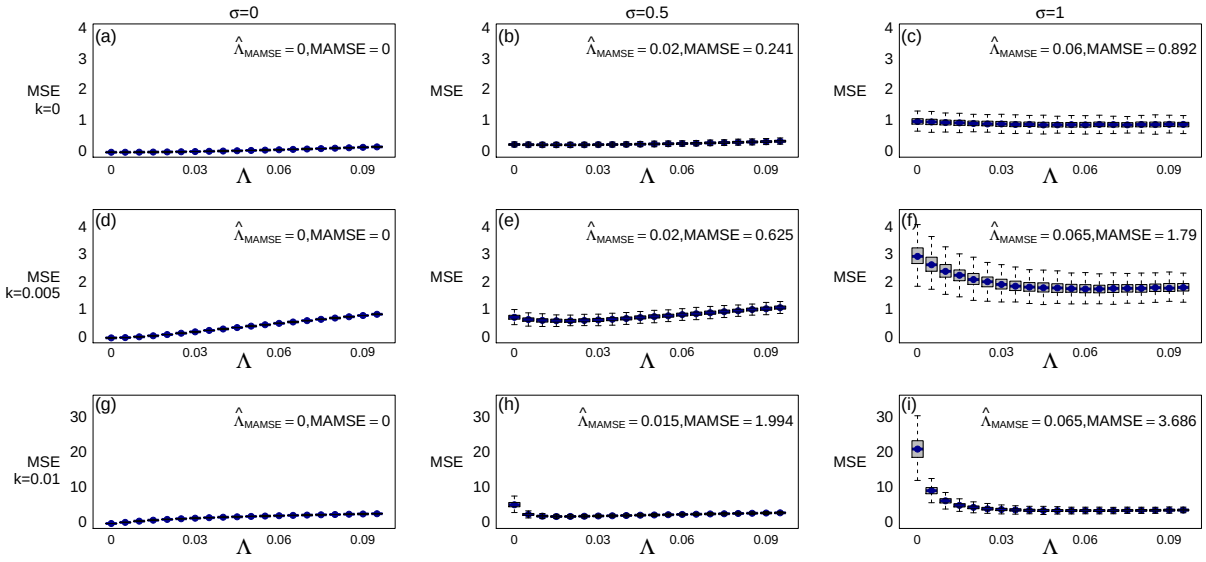


Figure 2.10: Boxplots of MSE results as function of Λ of first-order smoothing using the Blocks test function with different blur, which is given in (2.6), and noise levels.

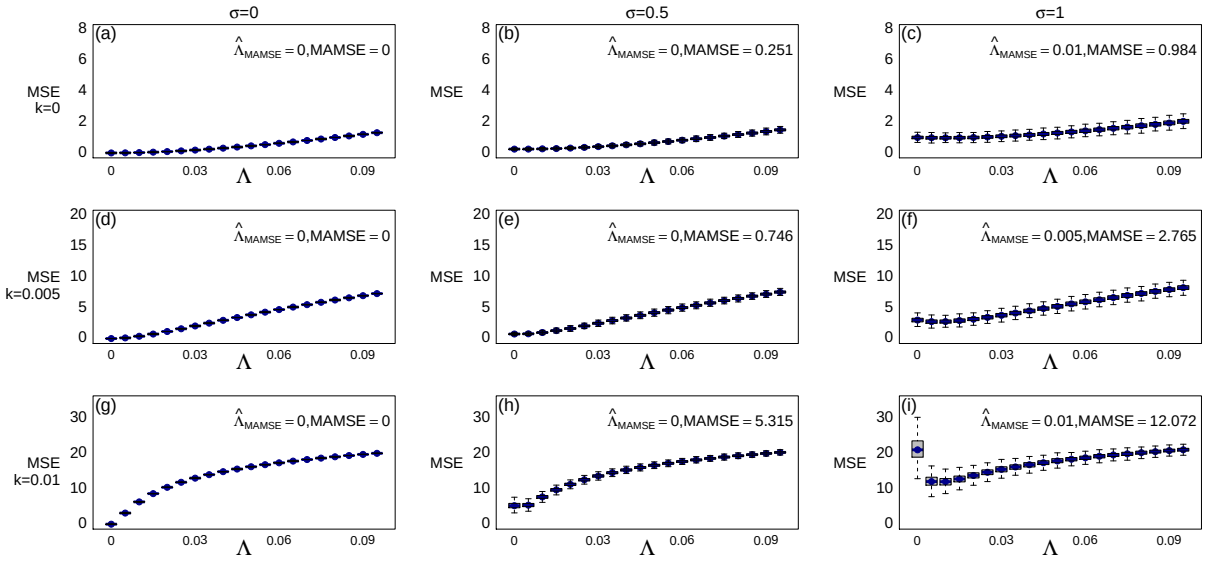


Figure 2.11: Boxplots of MSE results as function of Λ of first-order smoothing using the Bumps test function with different blur, which is given in (2.6), and noise levels.

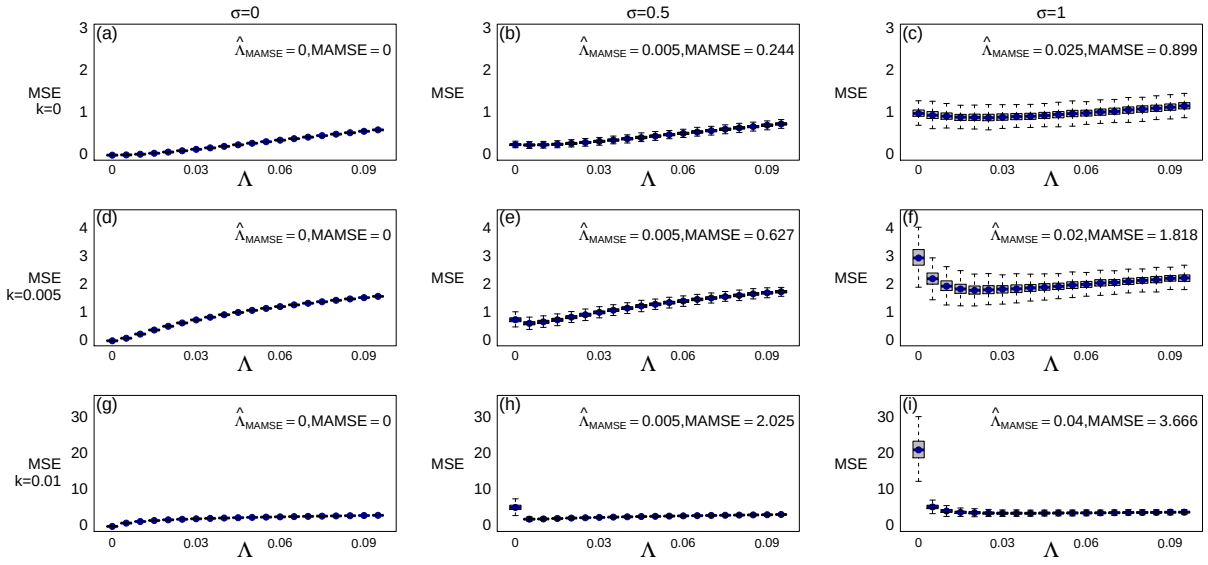


Figure 2.12: Boxplots of MSE results as function of Λ of second-order smoothing using the Blocks test function with different blur, which is given in (2.6), and noise levels.

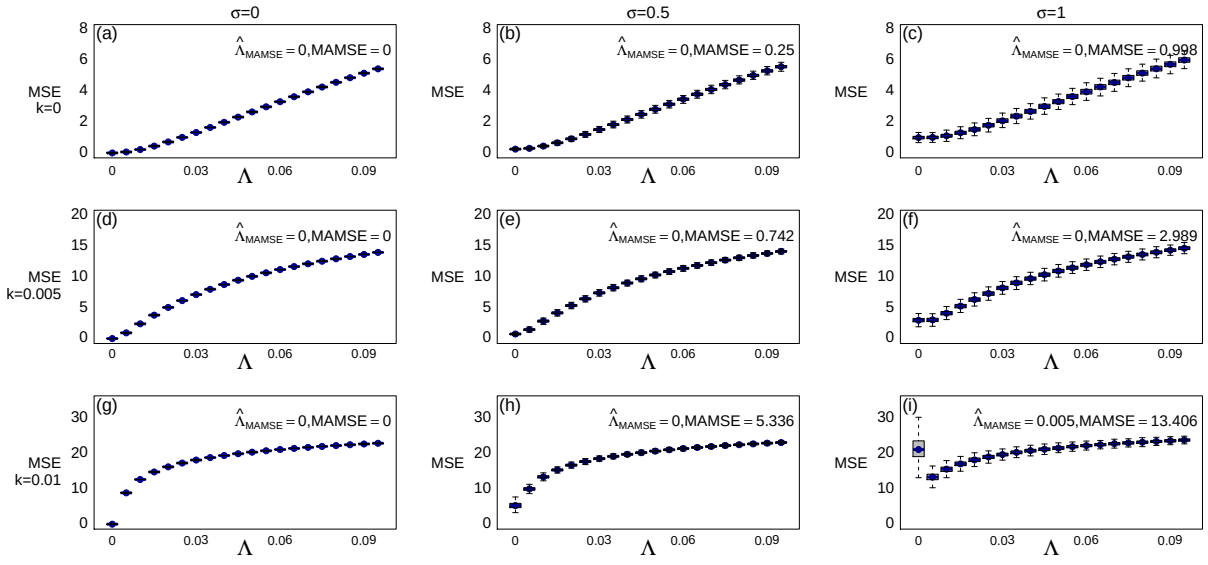


Figure 2.13: Boxplots of MSE results as function of Λ of second-order smoothing using the Bumps test function with different blur, which is given in (2.6), and noise levels.

of $\widehat{\Lambda}_{\text{MAMSE}}$, it can be seen that as the noise level σ increases so does the corresponding $\widehat{\Lambda}_{\text{MAMSE}}$ and generally $\widehat{\Lambda}_{\text{MAMSE}}$ increases slightly when the value of k increases.

Figures 2.8 and 2.9 show the MSE results using ridge regression estimates of the Blocks test function and the Bumps test function, based on data with different levels of blur and noise. The parameter $\widehat{\Lambda}_{\text{MAMSE}}$ and the MAMSE values increase as the level of noise and blur increase. Moreover, from Figures 2.8, 2.9, 2.10 2.11, 2.12 and 2.13, it is clear that ridge regression and first-order smoothing improve the MSE results slightly for $\sigma = 0.5$ and $\sigma = 1$, respectively. Also, Figures 2.10 and 2.12 show that MSE results seem to be similar in the case of the Blocks and Bumps test functions. For small levels of noise $\sigma = 0.5$ and blur $k = 0.005$, $\widehat{\Lambda}_{\text{MAMSE}}$ equals zero, so the underlying function, $\mathbf{f}$, can be estimated using ML.

2.8 Introduction to wavelets

Wavelets may be seen as "small waves". The term "wavelets" itself was coined in the geophysics literature by Morlet *et al.* (1982); see Daubechies (1992) and Nason (2010a). The underlying ideas behind the theory and application of wavelets can be found back in the early twentieth century. More recently, wavelets were re-introduced in the geophysics literature by Morlet *et al.* (1982).

Wavelet methods can be applied in many fields and applications, such as image analysis, radar, air acoustics, and endless other signal processing areas (Young, 1993). The wavelet transform can be thought of as a version of the Fourier transform (Sifuzzaman *et al.*, 2009). However, wavelets provide a sparse and localized decomposition appropriate for many non-parametric modelling situations. It can be explained in simple terms as describing a signal by a few large wavelet coefficients, which can be analysed, manipulated, or stored, and then used to transmit images or reconstruct the original signal (Hubbard, 1996).

There are many types of wavelets to choose from including smooth wavelets, compactly supported wavelets, wavelets with simple mathematical expressions and wavelets with

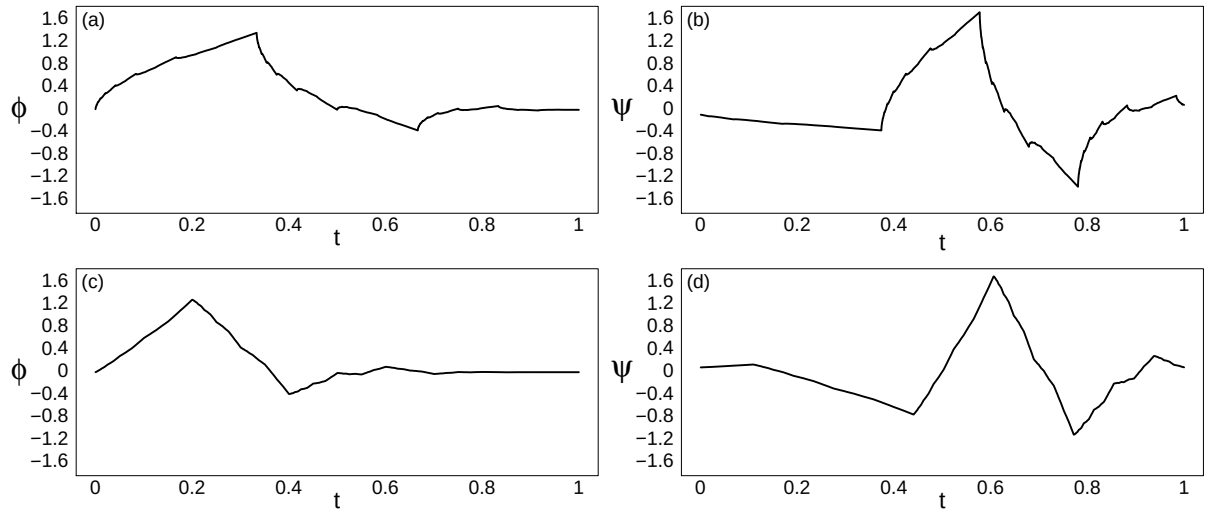


Figure 2.14: Wavelets from the Daubechies family: (a) the scaling function with the number of vanishing moments being $N = 2$; (b) the wavelet function with the number of vanishing moments being $N = 2$; (c) the scaling function with the number of vanishing moments being $N = 3$; and (d) the wavelet function with the number of vanishing moments being $N = 3$ of extremal phase wavelet family.

simple associated filters. However, the Haar wavelet will be the focus of study in this thesis. This provides a piecewise constant approximation of a function that is a representation of the function at different resolutions (Liu *et al.*, 2002).

By design the wavelet's usefulness rests on its ability to localize a process in time-frequency space. At high resolution levels, the wavelet is narrow, while at low resolutions the wavelet is stretched out. By moving from high to low resolution levels the wavelet is able to zoom in on process behaviour at a point in time or alternatively zoom out and reveal the general features of a signal (Jensen, 1995). Sobolu and Pusta (2010) summarized the structure, as "small bursts of high frequency wavelets followed by lower frequency waves or vice versa". As a result, the use of wavelet reconstruction helps to localize and identify such accumulations of small waves and thus leads to a better understanding of the reasons for these phenomena.

One approach to the wavelet method is to start with a set of orthonormal basis func-

tions generated by dilation and translation of a compactly supported scaling function (or father wavelet), ϕ, and a wavelet function (or mother wavelet), ψ, associated with an r-regular multiresolution analysis of $L^2(\mathbb{R})$. The notation $L^2(\mathbb{R})$ represents the space for all functions with well defined integral of the square of modulus of the function, where "L" signifies a Lebesque integral, "2" denotes the integral of the square of the modulus of the function, and "$\mathbb{R}$" states that the integration is over the set of real numbers. A variety of different wavelet families now exist that combine compact support with various degrees of smoothness and numbers of vanishing moments (Daubechies, 1992). The scaling function, $\phi_{j,l}(t)$, can be written as

$$\phi_{j,l}(t) = 2^{j/2}\phi(2^j t - l), \quad j \geq 0, \quad 0 \leq l \leq 2^j - 1; \tag{2.23}$$

a corresponding wavelet function, $\psi_{j,l}(t)$, is given by

$$\psi_{j,l}(t) = 2^{j/2}\psi(2^j t - l), \quad j \geq 0, \quad 0 \leq l \leq 2^j - 1, \tag{2.24}$$

where the integer variable j represents dilation, scale, level or resolution, and the integer variable l indicates translation, location in time or shift. A dilation, meaning a unit increase in j causes double the number of oscillations to occur within a set width. Whereas, a translation, meaning a unit increase in l, shifts $\psi(t)_{j,l}$ by 2^{-j}, and $\phi(t)_{j,l}$ by 2^{-j}. There are many possible choices for these father and mother wavelets, which from a suitable basis for $L^2(\mathbb{R})$ (Abramovich *et al.*, 2000).

The function $\psi(t)$, defined over the real axis must satisfy the two conditions

$$\int_{-\infty}^{\infty} \psi(t)dt = 0 \tag{2.25}$$

and

$$\int_{-\infty}^{\infty} \psi^2(t)dt = 1. \tag{2.26}$$

Hence, the non-zero values of $\psi(t)$ can be considered as being limited to a relatively small interval of time. While Equations (2.25) and (2.26) state that the function must vary away

from zero, they also state that any positive variations are matched by negative variations. Consequently, the function must resemble a wave or wavelet (Percival and Walden, 2006).

Daubechies (1988) designed two families of orthogonal wavelet bases, which combine the ideas of vanishing moments and compact support to produce wavelets with different degrees of smoothness. Function ψ is said to have N vanishing moments if $\int x^n \psi(x) dx = 0$ for $n = 0, 1, \ldots, N-1$. If the number of vanishing moments increases then the smoothness of the corresponding wavelet increases. The two wavelet families are known as the *Least-Asymmetric* and the *Extremal-Phase* and both are indexed in terms of the number of vanishing moments and hence smoothness (Vidakovic, 1999). Figure 2.14 shows the father and mother wavelets of the extremal phase wavelet family, with the number of vanishing moments being $N = 2$ and $N = 3$. As the number of vanishing moments increases, so does the smoothing of the corresponding wavelet.

A simple wavelet basis for $L^2(\mathbb{R})$ can be found from the Haar father wavelet $\phi(t)$. This was proposed by Alfred Haar in 1910 (Abramovich *et al.*, 2000; Haar, 1910) and is conveniently defined to have non-zero value on the interval $[0, 1]$ as

$$\phi(t) = \begin{cases} 1, & \text{if } t \in [0,1] \\ 0, & \text{otherwise.} \end{cases} \tag{2.27}$$

The Haar wavelet has only the "zeroth" vanishing moment resulting in a discontinuous wavelet function (Vidakovic, 1999). In general, the Haar scaling function can be written as

$$\phi_{j,l}(t) = \begin{cases} 2^{\frac{j}{2}}, & \text{if } t \in [2^{-j}l, 2^{-j}(l+1)] \\ 0, & \text{otherwise.} \end{cases} \tag{2.28}$$

Examples of the Haar scaling function, for the dilations $j = 0, 1, 2$ and translations $l = 0, 0, 3$ (black, red and green respectively) are shown in Figure 2.15 (a). More precisely, the value of j controls the width of the scaling function; as j increases the plot of $\phi(t)$ becomes narrower and the value increases significantly at the peak. Additionally, l controls the position; as l changes between 0 and $2^j - 1$ the location of the function moves left or right.

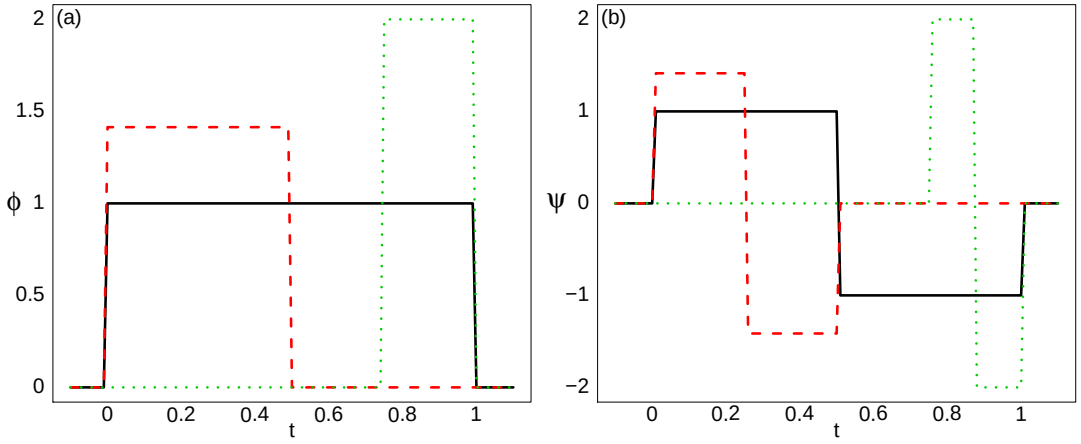


Figure 2.15: The Haar scaling and wavelet functions for various dilations and translations.

Nason (2010a) defined the father wavelet coefficients to be

$$c_{f\,j,l}(t) = \int_0^1 f(t)\phi_{j,l}(t)dt = \quad < f, \phi_{j,l}(t) >, \tag{2.29}$$

where $< \cdot >$ represents the inner product. The Haar mother wavelet function can be written on the interval $[0, 1)$ as

$$\psi(t) = \begin{cases} 1, & \text{if } t \in [0, \frac{1}{2}) \\ -1, & \text{if } t \in [\frac{1}{2}, 1) \\ 0, & \text{otherwise.} \end{cases} \tag{2.30}$$

This is a step function taking values 1 and -1 on $[0, \frac{1}{2})$ and $[\frac{1}{2}, 1)$, respectively. The Haar wavelet function is defined as

$$\psi_{j,l}(t) = \begin{cases} 2^{j/2}, & \text{if } t \in [\frac{l}{2^j}, \frac{l+\frac{1}{2}}{2^j}) \\ -2^{j/2}, & \text{if } t \in [\frac{l+\frac{1}{2}}{2^j}, \frac{l+1}{2^j}) \\ 0, & \text{otherwise.} \end{cases} \tag{2.31}$$

The Haar wavelet functions, for the dilations $j = 0, 1, 2$ and translations $l = 0, 0, 3$ (black, red and green respectively) are shown in Figure 2.15 (b). Again, j controls the width and l the position in the same way as for the scaling function. Abramovich *et al.* (1998) defined the wavelet coefficients to be

$$d_{f\,j,l}(t) = \int_0^1 f(t)\psi_{j,l}(t)dt = \quad < f, \psi_{j,l}(t) > . \tag{2.32}$$

The Equations (2.29) and (2.32) will be explained in Chapter 3.

2.9 Discrete wavelet transform

The purpose of this section is to briefly explain the discrete wavelet transform (DWT). More details on the DWT and other wavelet transforms will be given later, in Chapter 3. Consider an unknown signal $\mathbf{f} = \{f(t_i) : i = 1, 2, \ldots, m\}$ at a set of m equally spaced data points $t_i = i/m$. Suppose a set of noisy data $\mathbf{y} = \{y_i : i = 1, 2, \ldots, m\}$ are recorded at the same locations, then the model is given by

$$\mathbf{y} = \mathbf{f} + \boldsymbol{\epsilon}, \tag{2.33}$$

where the error $\boldsymbol{\epsilon} = \{\epsilon_i : i = 1, 2, \ldots, m\}$ are assumed independently $\boldsymbol{\epsilon} \sim N_m(\mathbf{0}, \sigma^2\mathbf{I}_m)$ distributed and $m = 2^J$, for some integer J. The unknown vector $\mathbf{f}$, can equivalently be described by its DWT

$$\mathbf{d_f} = \mathbf{W}\mathbf{f} = \{\mathrm{c_{f0,0}}, \mathrm{d_{f}}_{j,l} : j = 0, \ldots, J-1, \, l = 0, \ldots, 2^j - 1\}, \tag{2.34}$$

where $\mathbf{d_f}_{m\times 1}$ is a vector of wavelet coefficients containing both scaling coefficient (average) $\mathrm{c_{f0,0}}$ at level 0 and wavelet coefficients $\mathrm{d_{f}}_{j,l}$ from level 0 to level $J-1$ (Nason, 2010a). Suppose $\mathbf{W}$ is an orthogonal $m \times m$ matrix. For example, with the Haar basis and when $m = 2^3$ the matrix $\mathbf{W}$, is given by

$$\mathbf{W} = \begin{bmatrix} \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} \end{bmatrix}.$$

Since $\mathbf{W}$ is an orthogonal matrix $\mathbf{W}^T\mathbf{W} = \mathbf{I}$, it follows that

$$||\mathbf{d_y}||_2^2 = \mathbf{d_y}^T\mathbf{d_y} = (\mathbf{Wy})^T\mathbf{Wy} = \mathbf{y}^T(\mathbf{W}^T\mathbf{W})\mathbf{y} = \mathbf{y}^T\mathbf{y} = ||\mathbf{y}||_2^2,$$

where $||.||_2$ is the L_2 norm, and the length of the vector $\mathbf{d_y}$ is the same as that of the vector $\mathbf{y}$ (Nason, 2010a). The wavelet decomposition of $\mathbf{y}$, can be written as

$$\mathbf{d_y} = \mathbf{Wy} = \mathbf{W}(\mathbf{f} + \boldsymbol{\epsilon}) = \mathbf{Wf} + \mathbf{W}\boldsymbol{\epsilon} = \mathbf{d_f} + \boldsymbol{\eta}, \tag{2.35}$$

where $\mathbf{d_y}_{m\times 1}$ and $\mathbf{d_f}_{m\times 1}$ are vectors of the wavelet coefficients of $\mathbf{y}$ and $\mathbf{f}$ respectively. Thus, the model in (2.33) can be written equivalently as

$$\mathbf{d_y} = \mathbf{d_f} + \boldsymbol{\eta}. \tag{2.36}$$

The orthogonality of matrix $\mathbf{W}$ and normality of the noise vector $\boldsymbol{\epsilon}$ implies the noise vector $\boldsymbol{\eta}$ is also normal (Johnstone and Silverman, 1997) and the noise becomes spread across wavelet coefficients.

Similarly, if the model is corrupted with blurring then the model is given by

$$\mathbf{y} = \mathbf{Hf} + \boldsymbol{\epsilon}, \tag{2.37}$$

where $\mathbf{H}$ is a given $n \times m$ blur matrix and $\boldsymbol{\epsilon}$ is a vector of random variables, such that $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$. Then

$$\mathbf{d_y} = \mathbf{Wy} = \mathbf{W}(\mathbf{Hf} + \boldsymbol{\epsilon}) = \mathbf{WHf} + \mathbf{W}\boldsymbol{\epsilon} = \mathbf{d_g} + \boldsymbol{\eta}, \tag{2.38}$$

where $\mathbf{d_g}_{n\times 1}$ is a vector of wavelet coefficients containing the wavelet coefficients of $\mathbf{g} = \mathbf{Hf}$ and $\mathbf{f}_{m\times 1}$ is our signal of interest. Thus, the model in (2.38) can be written equivalently as

$$\mathbf{d_y} = \mathbf{d_g} + \boldsymbol{\eta}. \tag{2.39}$$

The main advantage of the DWT is that it is possible at express a signal at different levels of approximation as described in the next section.

2.10 Wavelet approximations

One extremely useful aspect of wavelets is that a given signal, $f(t)$, can be expressed at different levels of approximations in terms of a sum of wavelet coefficients and corresponding scaling and wavelet functions (Antoniadis *et al.*, 2001).

In particular, the p-level approximation of the function over the interval $[0, 1]$ can be written as

$$f^p(t) = \mathrm{c_{f0,0}}\phi_{0,0}(t) + \sum_{j=0}^{p}\sum_{l=0}^{2^j-1} \mathrm{d_{f}}_{j,l}\psi_{j,l}(t), \quad t \in [0,1]. \tag{2.40}$$

Alternatively, let the j^{th} component of the approximation be defined as

$$f_j(t) = \begin{cases} \mathrm{c_{f0,0}}\phi_{0,0}(t) + \mathrm{d_{f0,0}}\psi_{0,0}(t), & j = 0 \\ \sum_{l=0}^{2^j-1} \mathrm{d_{f}}_{j,l}\psi_{j,l}(t), & j = 1, 2, \ldots, J-1, \end{cases} \tag{2.41}$$

and the cumulative approximation up to resolution level p, defined as

$$f^p(t) = \sum_{j=0}^{p} f_j(t). \tag{2.42}$$

Figure 2.16 shows cumulative approximations of the Blocks test function with different numbers of wavelet functions. It can be seen that approximations contain information about the signal and as more components are added, the approximation becomes closer to the true function.

Figure 2.17 shows wavelet coefficients for the Blocks test function sampled at $m = 32$ equally spaced points. The black spikes represent the wavelet coefficients of the true function, the green spikes represent the wavelet coefficients of noise-free data, and the red spikes indicate the wavelet coefficients of the observed data with noise.

The important wavelet coefficients can typically be found in the lower resolution levels shown in Figure 2.17. As the blur increases the non-zero wavelet coefficients approach zero. In addition, as the level of noise increases, the number of non-zero wavelet coefficients in the lowest level increases.

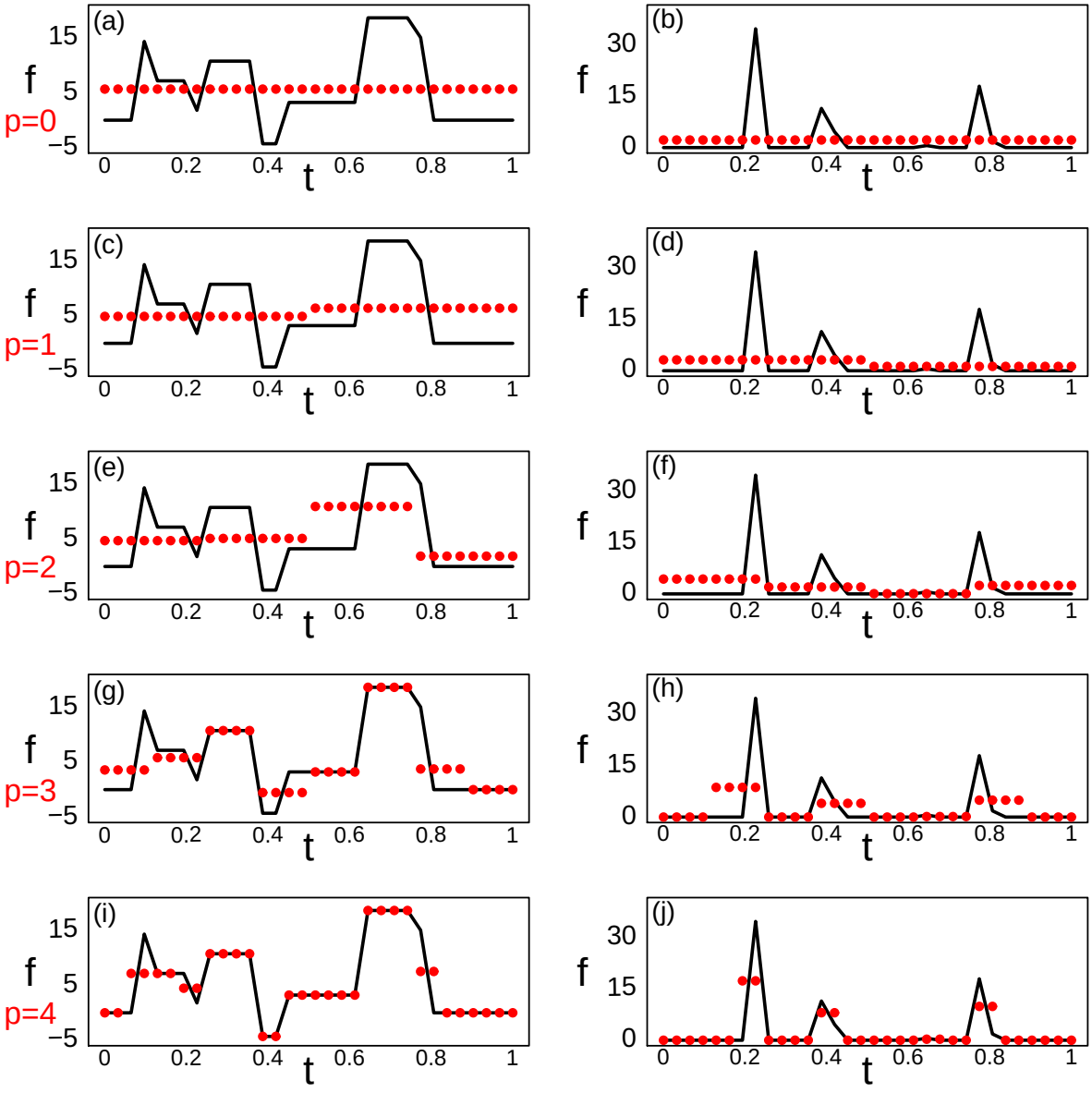


Figure 2.16: Cumulative approximations of Blocks and Bumps test functions, at $m = 32$ equally spaced points, at successive levels $p = 0, 1, 2, 3, 4$, with the data shown as points.

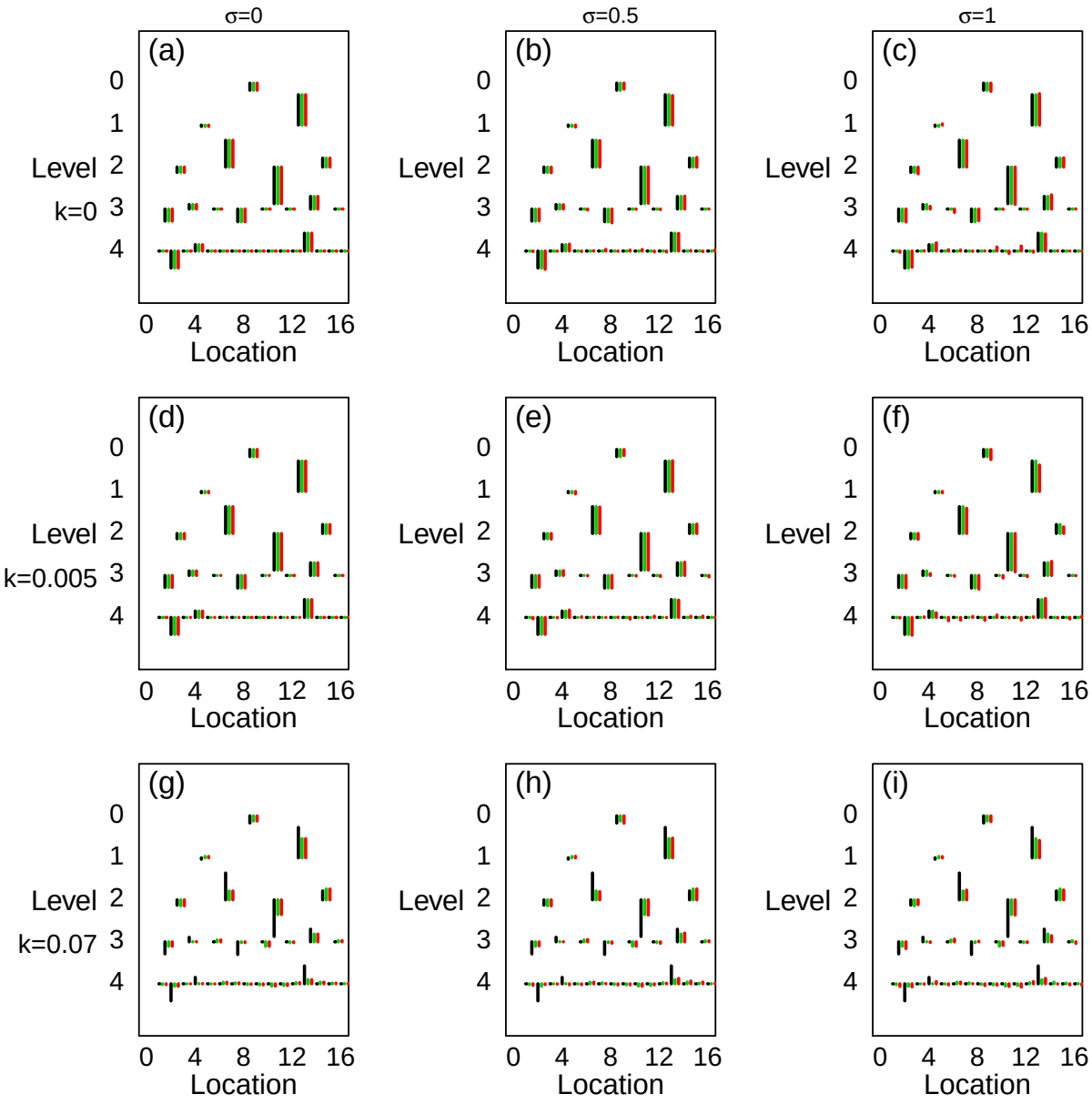


Figure 2.17: Wavelet tableaux of Blocks test function, for $k = 0, 0.005, 0.07$ (rows) and $\sigma = 0, 0.5, 1$ (columns) at $m = 32$ equally spaced points: the black spikes represent the wavelet coefficients of the true Blocks test function, the green spikes represent the wavelet coefficients of noise-free data, and the red spikes indicate the wavelet coefficients of the observed data with noise.

The main goal of wavelet methods is to keep important information about a signal by applying a shrinkage method to remove noise and then reconstructing the signal using the inverse transform as described in the next section.

2.11 Classical thresholding

Shrinkage can be summarised as a method for estimating signals from data corrupted by noise. It is a non-parametric technique used to estimate a function. Wavelet methods are orthogonal series based methods that use the concept of sparseness. Consequently, shrinkage of the empirical wavelet coefficients works best in problems where the underlying set of true coefficients is sparse. It is assumed that the majority of the wavelet coefficients are small, which are shrunk, and the remaining few are large, which are kept. By shrinking the empirical wavelet coefficients towards zero, the smaller ones may be reduced to negligible levels.

Flandrin (1992), Vidakovic and Ruggeri (2001) and Barber and Nason (2004) suggested that the coefficients $\mathrm{d}_{\mathrm{y}_{j,l}}$ can be considered independently and they omit the double index j, l and work with a "typical" wavelet coefficient, d_{y}, the same approach is followed here.

One natural way to obtain shrinkage estimates of the true coefficients is to use thresholding methods (Nason, 1995). The principle of a thresholding rule is to shrink or threshold wavelet coefficients towards zero. More precisely, those below a threshold are "killed" while the others are "kept", providing the effect of both reducing the noise and compressing the original data, whilst keeping a good quality of approximation. The biggest challenge in wavelet thresholding is finding a suitable threshold value (Raimondo, 2002). The hard and soft thresholding rules are given respectively by

$$\mathrm{T}^{\mathrm{H}}(\mathrm{d}_{\mathrm{y}}, \lambda) = \begin{cases} 0, & \text{if } |\mathrm{d}_{\mathrm{y}}| \leq \lambda \\ \mathrm{d}_{\mathrm{y}}, & \text{if } |\mathrm{d}_{\mathrm{y}}| > \lambda, \end{cases} \tag{2.43}$$

and

$$\mathrm{T}^{\mathrm{S}}(\mathrm{d_y}, \lambda) = \begin{cases} 0, & \text{if } |\mathrm{d_y}| \leq \lambda \\ \mathrm{d_y} - \lambda, & \text{if } \mathrm{d_y} > \lambda \\ \mathrm{d_y} + \lambda, & \text{if } \mathrm{d_y} < -\lambda. \end{cases} \tag{2.44}$$

Hard thresholding is a "kept" or "killed" method, while soft thresholding is a "shrunk" or "killed". Hard thresholding is a discontinuous function while soft is a continuous function.

Hence, an estimate of the function $\mathbf{g}$, using estimates of $\mathbf{d_g}$, is defined as

$$\widehat{\mathbf{g}} = \mathbf{W}^T \mathrm{T}^*(\mathbf{d_y}, \lambda), \tag{2.45}$$

where $\mathrm{T}^*(\cdot)$ is a thresholding rule applied to its argument element by element, hence, in model (2.36), $\widehat{\mathbf{g}}$ is equivalent to $\widehat{\mathbf{f}}$. Also, in the case of the model in (2.38), the resulting estimate of $\mathbf{f}$, is given

$$\widehat{\mathbf{f}} = \mathrm{I}_*(\mathbf{W}^T \mathrm{T}^*(\mathbf{d_y}, \lambda), \mathbf{H}, \Lambda), \tag{2.46}$$

where $\mathrm{I}_*(\cdot)$ is an inversion method. If regularised inversion and a hard thresholding rule are used then

$$\begin{aligned} \widehat{\mathbf{f}} &= \mathrm{I}_*(\mathbf{W}^T \mathrm{T}^*(\mathbf{d_y}, \lambda), \mathbf{H}, \Lambda) \\ &= (\mathbf{H}^T\mathbf{H} + \Lambda \mathfrak{R}^T \mathfrak{R})^{-1} \mathbf{H}^T \mathbf{W}^T \mathrm{T}^{\mathrm{H}}(\mathbf{d_y}, \lambda). \end{aligned} \tag{2.47}$$

Gao and Bruce (1997) introduced a thresholding rule to amend the drawbacks of both hard and soft thresholding. The thresholding rule is called firm thresholding, with rule

$$\mathrm{T}^{\mathrm{F}}(\mathrm{d_y}, \lambda_1, \lambda_2) = \begin{cases} 0, & \text{if } |\mathrm{d_y}| \leq \lambda_1 \\ \mathrm{sign}(\mathrm{d_y}) \frac{\lambda_2(|\mathrm{d_y}| - \lambda_1)}{\lambda_2 - \lambda_1}, & \text{if } \lambda_1 < |\mathrm{d_y}| \leq \lambda_2 \\ \mathrm{d_y}, & \text{if } |\mathrm{d_y}| > \lambda_2, \end{cases} \tag{2.48}$$

which is a "kept", "shrunk" or "killed". The rule is continuous, which means no jump. Note that if $\lambda_2 = \lambda_1$ then $\mathrm{T}^{\mathrm{F}}(\mathrm{d_y}, \lambda_1, \lambda_2) \longrightarrow \mathrm{T}^{\mathrm{H}}(\mathrm{d_y}, \lambda)$ and if $\lambda_2 \longrightarrow \infty$ then $\mathrm{T}^{\mathrm{F}}(\mathrm{d_y}, \lambda_1, \lambda_2) \longrightarrow \mathrm{T}^{\mathrm{S}}(\mathrm{d_y}, \lambda)$.

The firm thresholding rule requires two threshold values, this makes the procedure more computationally expensive. To remedy this, Gao (1998) considered a thresholding rule called non-negative garrote (G) thresholding with rule

$$\mathrm{T}^{\mathrm{G}}(\mathrm{d_y}, \lambda) = \begin{cases} 0, & \text{if } |\mathrm{d_y}| \leq \lambda \\ \mathrm{d_y} - \frac{\lambda^2}{\mathrm{d_y}}, & \text{if } |\mathrm{d_y}| > \lambda, \end{cases} \tag{2.49}$$

which is a "shrunk" or "killed" rule. The G thresholding rule is a continuous function. Gao (1998) and Antoniadis and Fan (2001) suggested the smoothly clipped absolute deviation (SCAD) thresholding rule

$$\mathrm{T}^{\mathrm{SCAD}}(\mathrm{d_y}, \lambda) = \begin{cases} \mathrm{sign}(\mathrm{d_y})\max(0, |\mathrm{d_y}| - \lambda), & \text{if } |\mathrm{d_y}| \leq 2\lambda \\ \frac{(c-1)|\mathrm{d_y}| - c\,\lambda\,\mathrm{sign}(\mathrm{d_y})}{c-2}, & \text{if } 2\lambda < |\mathrm{d_y}| \leq c\lambda \\ \mathrm{d_y}, & \text{if } |\mathrm{d_y}| > c\lambda, \end{cases} \tag{2.50}$$

which is a piecewise linear function and is also a "kept", "shrunk" or "killed" rule. Antoniadis and Fan (2001) suggested using the value $c = 3.7$.

Bruce and Gao (1996b) derived the bias, variance and risk of hard and soft thresholding as follows. Let $X \sim N(\theta, 1)$, where X represents a wavelet coefficient. The mean, variance and the risk function of the shrinkage estimator $\mathrm{T}(X)$ of θ, are defined as

$$\mathrm{M}_\lambda(\theta) = \mathrm{E}\mathrm{T}_\lambda(X), \tag{2.51}$$

$$\mathrm{V}_\lambda(\theta) = \mathrm{Var}\mathrm{T}_\lambda(X), \tag{2.52}$$

$$\mathrm{R}_\lambda(\theta) = \mathrm{V}_\lambda(X) + \mathrm{M}_\lambda(X)^2. \tag{2.53}$$

For more details see Bruce and Gao (1996b). The means for the hard and soft shrinkage rules are given by

$$\mathrm{M}^{\mathrm{H}}_\lambda(\theta) = \theta + \theta\Big(1 - \Phi(\lambda - \theta) - \Phi(\lambda + \theta)\Big) + \phi(\lambda - \theta) - \phi(\lambda + \theta), \tag{2.54}$$

$$\mathrm{M}^{\mathrm{S}}_\lambda(\theta) = \mathrm{M}^{\mathrm{H}}_\lambda(\theta) - \lambda\Big(\Phi(\lambda + \theta) - \Phi(\lambda - \theta)\Big), \tag{2.55}$$

variances by

$$\mathrm{V}_{\lambda}^{\mathrm{H}}(\theta) = (\theta^2+1)\Big(2-\phi(\lambda-\theta)-\Phi(\lambda+\theta)\Big) + (\lambda+\theta) \tag{2.56}$$
$$\times\, \phi(\lambda-\theta) + (\lambda-\theta)\phi(\lambda+\theta) - \mathrm{M}_{\lambda}^{\mathrm{H}}(\theta)^2,$$
$$\mathrm{V}_{\lambda}^{\mathrm{S}}(\theta) = \mathrm{V}_{\lambda}^{\mathrm{H}}(\theta) - \lambda v_1(\lambda,\theta) + v_1(\lambda,-\theta), \tag{2.57}$$

where Φ is the cumulative distribution function of the standard Gaussian probability distribution, ϕ is the standard Gaussian probability density function, and

$$v_1(\lambda,\theta) = \Big(1+\Phi(\lambda-\theta)-\Phi(\lambda+\theta)\Big)\Big((2\theta-\lambda)(1-\Phi(\lambda-\theta)+2\phi(\lambda-\theta))\Big), \tag{2.58}$$

and finally the risk is given by

$$\mathrm{R}_{\lambda}^{\mathrm{H}}(\theta) = 1 + (\theta^2-1)\Big(\Phi(\lambda-\theta)-\Phi(-\lambda-\theta)\Big) + (\lambda+\theta)\phi(\lambda+\theta) \tag{2.59}$$
$$+ (\lambda-\theta)\phi(\lambda-\theta),$$
$$\mathrm{R}_{\lambda}^{\mathrm{S}}(\theta) = 1 + \theta^2 + (\theta^2-\lambda^2-1)\Big(\Phi(\lambda-\theta)-\Phi(-\lambda-\theta)\Big) \tag{2.60}$$
$$- (\lambda-\theta)\phi(\lambda+\theta) - (\lambda+\theta)\phi(\lambda-\theta).$$

The G, F and SCAD thresholding rules are more complicated, thus numerical methods have been proposed to compute mean, variance and the risk. Figures 2.18 shows the result of numerical methods for computing mean, variance and the risk. Panels (u) and (v) show the risk of hard and soft thresholding. It can be seen that the risk of soft is larger than the risk of hard thresholding.

2.12 Optimal choice of λ

In the classical thresholding rules, the biggest challenge is to find an appropriate threshold value λ. Note that when $\lambda = 0$ all the coefficient are kept, while $\lambda = \infty$ means that all the coefficients are shrunk. The thresholding rule works better if the thresholding value is

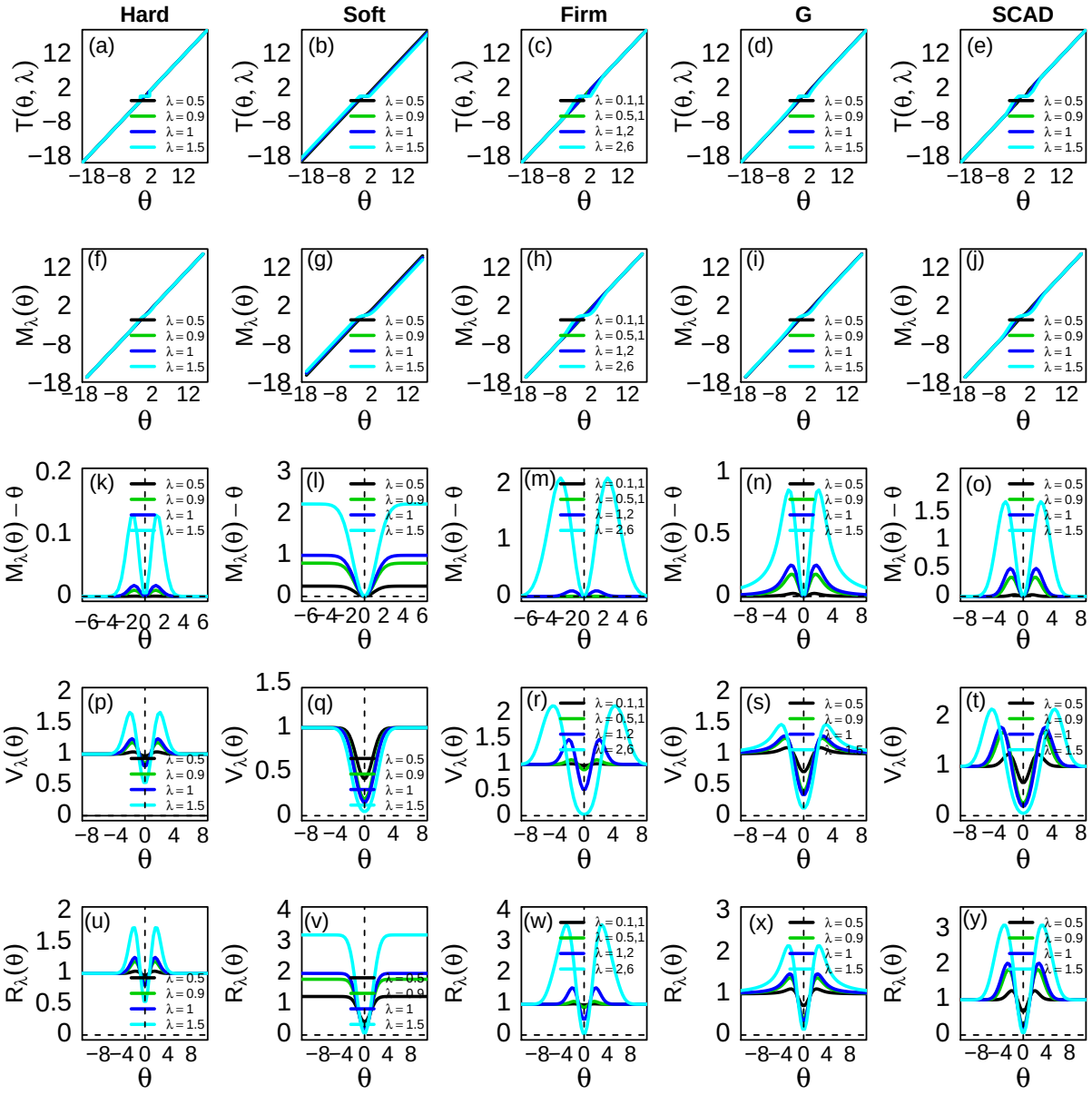


Figure 2.18: Plots of the rules, mean, bias, variance and risk for different classical thresholding rules.

well defined. In addition, different resolution levels might favour different values. Thus, Stein (1981) and Donoho and Johnstone (1994) considered several different methods of choosing the value of λ. In some rules, each individual wavelet coefficient is compared with a thresholding value, which is adaptive through term-by-term thresholding. A wavelet coefficient is retained if its magnitude is above the threshold level, and is thresholded, if its magnitude is under the thresholding level, which implies that no small coefficients will be let through the threshold (Cai, 1999).

In this chapter, three methods for choosing λ are studied and applied. Donoho and Johnstone (1994) introduced the universal threshold, which uses a threshold given by

$$\lambda_{\text{Univ}} = \widehat{\sigma}\sqrt{2\log(n)}, \tag{2.61}$$

where n is the number of data points. Nason (2010a) estimated the noise level using the finest-scale wavelet coefficients, as

$$\widehat{\sigma} = \text{MAD}[\mathbf{d_y}^{J-1}] = \frac{\text{MEDIAN}(|\mathbf{d_y}^{J-1} - \text{MEDIAN}(\mathbf{d_y}^{J-1})|)}{0.6745}. \tag{2.62}$$

Hence, σ is estimated using only the finest-scale wavelet coefficients $\mathbf{d_y}^{J-1}$. Even when there are a few large signal wavelet coefficients at that level then there is no significant effect on the MAD estimator (Nason, 2010a). An alternative method for estimating σ is to use the sample standard deviation of the finest-scale of wavelet coefficients (Cai and Zhou, 2009)

$$\widehat{\sigma} = \text{sd}[\mathbf{d_y}^{J-1}] = \sqrt{\frac{1}{\frac{n-2}{2}}\sum_{l=1}^{2^{J-1}}[\mathbf{d_y}_l^{J-1} - \bar{\mathbf{d}}_{\mathbf{y}_l}^{J-1}]^2}, \tag{2.63}$$

where sd denotes the standard deviation. A small simulation study involving 100 replications was carried out with results presented in Table 2.2 to show the difference between the estimators in (2.62) and (2.63) for an inverse problem.

Estimator	True σ	k		
		0.001	0.010	0.070
MAD[$\mathbf{d_y}^{J-1}$]	0.5	0.6701	0.9063	0.3704
sd[$\mathbf{d_y}^{J-1}$]	0.5	6.2741	1.6918	0.4273

Table 2.2: The results of using Equations in (2.62) and (2.63) for computing $\hat{\sigma}$, where the true variance of noise is $\sigma^2 = 0.25$ and different blur k.

The method of universal thresholding can be described by a five step procedure.

1. Transform the data into the wavelet domain using the DWT by calculating $\mathbf{d_y} = \mathbf{Wy}$.
2. Estimate the variance using the wavelet coefficients at finest-scale using Equation (2.62).
3. Estimate the value of the threshold using Equation (2.61).
4. Estimate the true wavelet coefficients using the thresholding rule.
5. Estimate the function using the inverse transform of the denoised wavelet coefficients.

The universal threshold, with high probability, ensures that every value for the wavelet transform, for which the underlying coefficient is exactly zero, will be estimated as zero. This is because, if $X_1, \ldots, X_n$ are normally distributed random variables with means 0 and variances $\sigma_1^2, \ldots, \sigma_n^2$, then

$$P\left(\max_{1\leq i\leq n}|X_i/\sigma_i| > \sqrt{2\log_2 n}\right) \to 0 \quad \text{as} \quad n \to \infty, \tag{2.64}$$

whether or not the variables are independent. The probability of X_i/σ_i exceeding the threshold $\sqrt{2\log_2 n}$ tends to zero (Johnstone and Silverman, 1997). However, the universal threshold causes over shrinkage, which means that too many true wavelet coefficients are deleted (Nason, 2010a).

The second method for estimating λ is Stein's Unbiased Risk Estimate (SURE), which was developed by Donoho and Johnstone (1995) and can be defined as level-dependent thresholding. This means that at each resolution level the estimated wavelet coefficients are compared to a separate threshold (Donoho and Johnstone, 1994). The value of the threshold λ_j is chosen to provide small risk, which means however that there is no explicit equation to compute λ directly. Stein (1981) gave the main idea to minimize the risk, which can be applied to each resolution level j, and is given by

$$\text{SURE}(\mathbf{d_y}, \lambda) = n - 2\#\{i : |\mathbf{d_y}_i| \leq \lambda\} + \sum_{i=1}^{n} \min(|\mathbf{d_y}_i|, \lambda), \tag{2.65}$$

where # denotes the number of elements in the set. Figure 2.19 shows that the SURE threshold is based on the number of wavelet coefficients, the magnitude of the coefficient and the value of the threshold. Figure 2.19 also shows that, for this example, the thresholding value λ is small and seems to be equal at each level. Thus

$$\lambda_{\text{S}} = \arg\min_{\lambda} \quad \text{SURE}(\mathbf{d_y}, \lambda). \tag{2.66}$$

The notation λ_S represents the value of the threshold at a particular resolution level using the SURE thresholding procedure. Donoho and Johnstone (1995) demonstrated that SURE thresholding can be found in $O(n \log_2(n))$ computational operations. However, the main problem with SURE thresholding is that the value of the threshold is always small.

The method of SURE thresholding can be described by a three step procedure:

1. Transform the data into the wavelet domain using the DWT by computing $\mathbf{d_y} = \mathbf{Wy}$.

2. Estimate the threshold for each resolution level using Equation (2.66), then estimate the wavelet coefficients using the thresholding rule.

3. Estimate the function using the inverse transform of the denoised wavelet coefficients.

The third method for choosing a value of threshold, λ, is called cross-validation, which aims to minimize the mean integrated squared-errorNason (1996). He suggested dropping

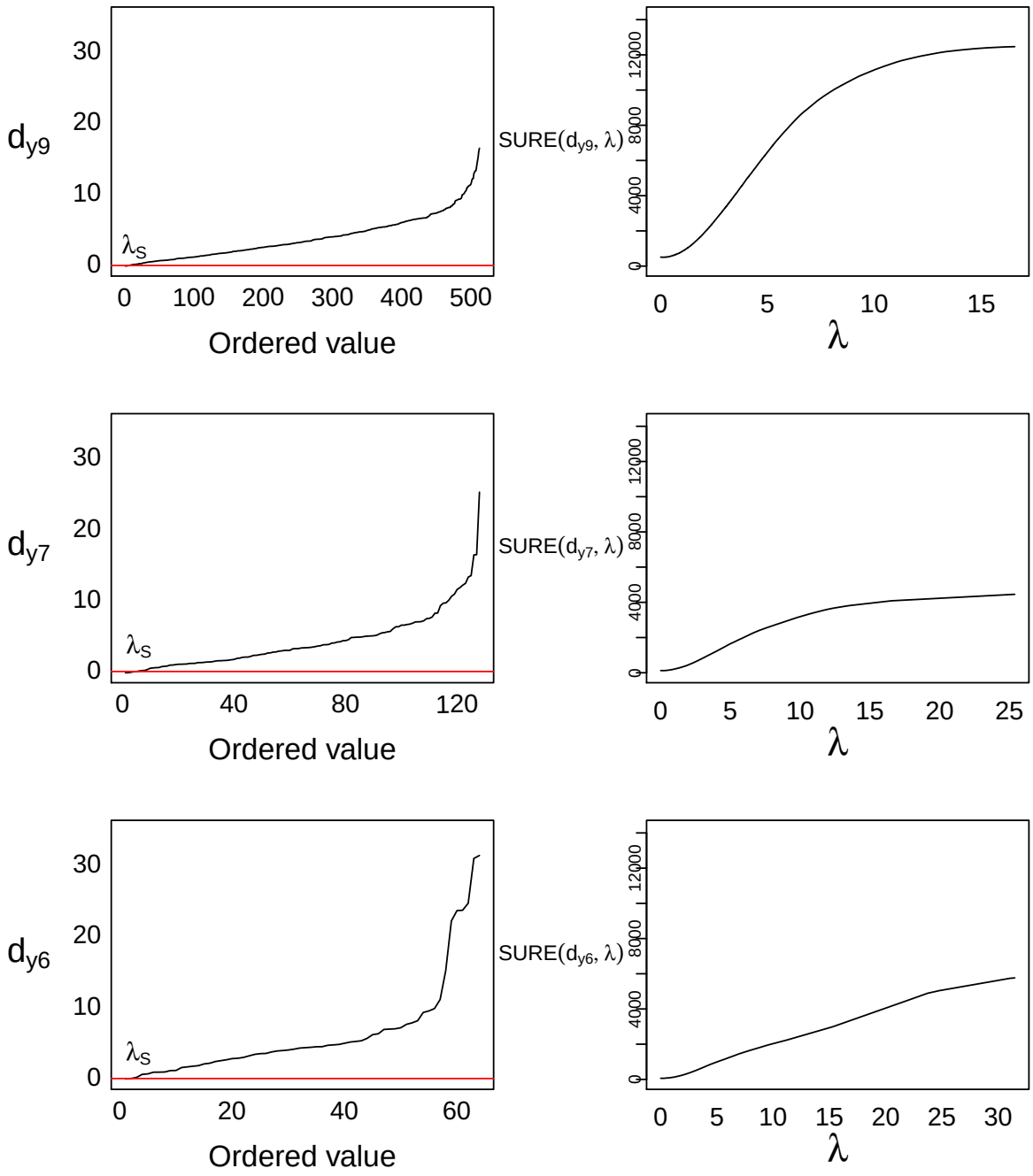


Figure 2.19: Illustrating the procedure of SURE threshold using the Block test function: the black lines in the left-hand columns denote the ordered wavelet coefficients, while; the red lines indicate the value of threshold, λ_j, at level j.

half the points of the data set, hence the length is still a power of two.
Given data, $\mathbf{y} = \{y_1, \ldots, y_n\}^T$, from the model in Equation (2.8), where n is of dyadic length (power of two), the first step is to remove all the odd-indexed values, which gives $\mathbf{y}^E = \{y_2, y_4, \ldots, y_{\frac{n}{2}}\}^T$, of length $\frac{n}{2}$, which is re-indexed from $i = 1, \ldots, \frac{n}{2}$. The estimate of the underlying signal is obtained using wavelet denoising on the even indexed sequence resulting in $\widehat{f}_\lambda(y^E)_i$. By assuming the dataset is periodic, then it can be written as

$$\bar{f}_\lambda(y^E)_i = \begin{cases} \frac{1}{2}\left(\widehat{f}_\lambda(y^E)_{i+1} + \widehat{f}_\lambda(y^E)_i\right), & \text{if} \quad i = 1, \ldots, \frac{n}{2} - 1 \\ \frac{1}{2}\left(\widehat{f}_\lambda(y^E)_i + \widehat{f}_\lambda(y^E)_1\right), & \text{if} \quad i = \frac{n}{2}. \end{cases} \tag{2.67}$$

The estimate $\widehat{f}_\lambda(y^O)_i$ is computed for the odd indexed points and the procedure above is repeated for $\bar{f}_\lambda(y^O)_i$. The estimate of the MISE, $\widehat{M}(\lambda)$, is then given by

$$\widehat{M}(\lambda) = \sum_{i=1}^{\frac{n}{2}} \left\{ \frac{1}{2}\left(\bar{f}_\lambda(y^E)_i - y_{2i-1}\right)^2 + \frac{1}{2}\left(\bar{f}_\lambda(y^O)_i - y_{2i}\right)^2 \right\}. \tag{2.68}$$

Nason (2010a) showed that when DWT is used $\widehat{M}(\lambda)$ can be found in $O(n)$ computational operations. Then, the threshold is given by

$$\lambda_{\text{CV}}\left(\frac{n}{2}\right) = \arg\min_{\lambda} \quad \widehat{M}(\lambda). \tag{2.69}$$

Nason (1996) also showed that the universal threshold for n data points is given by $\lambda_{\text{Univ}}(n) = \widehat{\sigma}\sqrt{2\log_2 n}$, hence the cross-validation threshold for n data points will be given by

$$\lambda_{\text{CV}}(n) = \left(1 - \frac{\log_2 2}{\log_2 n}\right)^{-\frac{1}{2}} \lambda_{\text{CV}}\left(\frac{n}{2}\right). \tag{2.70}$$

The correction in Equation (2.70) is applied to obtain the final stage of the cross-validation threshold estimate. The correction can be derived as follows.

$$\begin{aligned}
\sqrt{2\log_2 n} &= \sqrt{2\log_2 n}\frac{\sqrt{2\log_2 \frac{n}{2}}}{\sqrt{2\log_2 \frac{n}{2}}} \\
&= \left(\frac{2\log_2 n}{2\log_2 n - 2\log_2 2}\right)^{\frac{1}{2}} \sqrt{2\log_2 \frac{n}{2}} \\
&= \left(\frac{2\log_2 n - 2\log_2 2}{2\log_2 n}\right)^{-\frac{1}{2}} \sqrt{2\log_2 \frac{n}{2}} \\
&= \left(1 - \frac{2\log_2 2}{2\log_2 n}\right)^{-\frac{1}{2}} \sqrt{2\log_2 \frac{n}{2}} \\
&= \left(1 - \frac{\log_2 2}{\log_2 n}\right)^{-\frac{1}{2}} \sqrt{2\log_2 \frac{n}{2}}. \qquad (2.71)
\end{aligned}$$

The result in Equation (2.70) can be used to fix the value of the threshold for the whole set of wavelet coefficients. The method of cross-validation, can also be described by the following stepwise procedure:

1. Divide the data into two groups, even and odd.
2. For each group, transform the elements to the wavelet domain using the DWT by calculating $\mathbf{d}^E{}_{\mathbf{y}} = \mathbf{W}\mathbf{y}^E$ and $\mathbf{d}^O{}_{\mathbf{y}} = \mathbf{W}\mathbf{y}^O$.
3. Estimate the true wavelet coefficients using a threshold parameter λ, and with a particular thresholding rule to give $\mathrm{T}^*(\mathbf{d}^E_{\boldsymbol{y}}, \lambda)$ and $\mathrm{T}^*(\mathbf{d}^O_{\boldsymbol{y}}, \lambda)$, where $\mathrm{T}^*(.)$ is a thresholding rule.
4. Estimate $\widehat{\mathbf{f}}_{\lambda,y^E}$ and $\widehat{\mathbf{f}}_{\lambda,y^O}$ using the inverse transform, by calculating $\widehat{\mathbf{f}}_{\lambda,y^E} = \mathbf{W}^T\mathrm{T}^*(\mathbf{d}^E_{\boldsymbol{y}}, \lambda)$ and $\widehat{\mathbf{f}}_{\lambda,y^O} = \mathbf{W}^T\mathrm{T}^*(\mathbf{d}^O_{\boldsymbol{y}}, \lambda)$.
5. Compute the pairwise averages for each group, in Equation (2.67).
6. Compute the value of the MISE using (2.68).
7. Repeat the third, fourth, fifth and sixth steps for different values of the threshold.
8. Compute the value of the threshold using (2.69).
9. Correct the value of the threshold using (2.70).

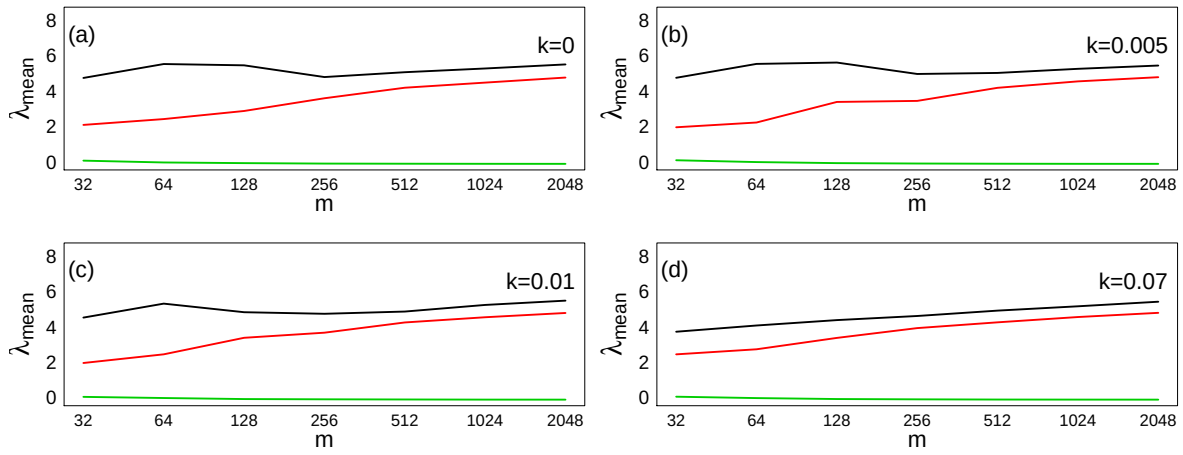


Figure 2.20: Average of λ over 1000 replications of universal compared to cross-validation and SURE methods: (a), (b), (c) and (d) show different levels of blur, which is given in (2.6), $k = 0, 0.005, 0.01, 0.07$, the black lines represent the value of $\widehat{\lambda}_{\text{Univ}}$, the red lines represent the value of $\widehat{\lambda}_{\text{CV}}$ using hard thresholding rule, and the green lines represent the value of $\widehat{\lambda}_{\text{S}}$, where Block test function datasets using $m = 32, 64, 128, 256, 512, 1024$ and 2048 equally spaced points.

A small simulation study of 1000 replications was carried out with the results in Figure 2.20. The simulated data sets consisted of the standard test signal Blocks, corrupted by independent Gaussian noise with $\sigma = 2$ and the values of blur k, which is given in (2.6), were taken as 0, 0.005, 0.01 and 0.07. In this simulation, the cross-validation algorithm uses the hard thresholding rule.

Figure 2.20 shows the average estimated value of threshold λ, using universal threshold, cross-validation and SURE, with different numbers of equally spaced points $m = 32, 64, 128, 256, 512, 1024$ and 2048, and different levels of blur, which is given in (2.6). For each replication the value of λ is computed, and the average of the values taken at the end of the process. In general, the average of 1000 replications for computing the value of threshold using universal threshold is always larger than the value of threshold using cross-validation with the hard thresholding rule and SURE thresholding, that is

$$\widehat{\lambda}_{\mathrm{S}} < \widehat{\lambda}_{\mathrm{CV}} < \widehat{\lambda}_{\mathrm{Univ}}, \tag{2.72}$$

here $\widehat{\lambda}_S$ is only computed for the finest resolution level.

There are numerous other methods for specifying a value for the threshold λ. One of them is known as the *false-discovery rate*. This multiple hypothesis testing approach was proposed by Benjamini and Hochberg (1995) and adapted by Abramovich and Benjamini (1996). For example, consider a single hypothesis

$$\mathrm{H}_0 : \theta = 0 \quad \text{verus} \quad \mathrm{H}_1 : \theta \neq 0,$$

and suppose a statistical test is performed at the significance level of p(Reject $\mathrm{H}_0|\mathrm{H}_0$ is true) = 0.05. If, say, the number of wavelet coefficients equals 127, which are to be independently tested at significance level of $q = 0.05$, then there are an expected $127 \times 0.05 \approx 7$ coefficients, which are not equal to zero. Abramovich and Benjamini (1996) suggested a situation where there are n hypotheses to be tested, where each null hypothesis is of the form $\mathrm{H}_0 : \theta = 0$. Also, there are n_1 null hypotheses, which are false. This means the corresponding coefficients should be included in the reconstruction. The other $n_0 = n - n_1$ coefficients are zero and then all the noisy versions should be set to zero. Also, define R to be the number of coefficients that are not set to zero by a given thresholding procedure. Of these R coefficients, S are kept, and V are kept in by mistake, such that $R = S + V$. The error is written as $Q = V/R$ and it represents the proportion of coefficients that should have been set to zero and therefore, the false Discovery Rate of Coefficients (FDRC) is defined to be the expectation of Q. Benjamini and Hochberg (1995) suggested maximising the number of included coefficients subject to controlling the FDRC to some level q. The method of FDRC thresholding can be described by a four step procedure:

- For each $\mathrm{d_y}^*$ calculate the two-sided p-value, p, testing $\mathrm{H}_0 : \mathrm{d_y} = 0$

$$p = 2(1 - \phi(|d_y|/\sigma)).$$

- Order the p's according to their size.
- Let i_0 be the largest i for which $p_{(i)} \leq (i/m)q$ and calculate

$$\lambda_{i_0} = \sigma\phi^{-1}(1 - p_{i_0}/2).$$

- Threshold all coefficients using threshold λ_{i_0}.

Benjamini and Hochberg (1995) proved that for the independent Gaussian noise model, assumed in Equation (2.33), the above procedure controls the FDRC at an (unknown) level $(m_o/m)q \leq q$, where m_o is the number of coefficients that are exactly zero and m is the number of tested hypotheses. Thus, using the above procedure will control the FDRC at a rate conservatively less than q.

An alternative approach to choosing a threshold was proposed by Ogden (1994) to develop two methods for thresholding, *selection thresholding* and *data-analytic thresholding*. Selection thresholding depends on hypothesis testing of coefficients level-by-level and provides a test statistic, that if large will encourage the user to include the largest (in absolute terms) coefficients into the reconstruction. Then the remainder of the wavelet coefficients continue to be tested. If the test statistic is not large enough (when compared to some critical value) then the threshold is set to be the absolute value of the largest remaining coefficient. Data-analytic thresholding depends on looking at plots of cumulative sums of the squares of the coefficients at a particular level. Wavelet coefficients are removed from the level if some test, based on Brownian bridge sampling, is significant. It continues by testing the remainder of the wavelet coefficients. The test tries to ascertain if the remaining coefficients are just white noise, by successively removing the larger coefficients until the test decides that the coefficients are indistinguishable from white noise.

An excellent critical overview and simulation study comparing different shrinkage methods, which gives results about optimality of classical thresholding, can be found in Fan and Li (2001), Antoniadis *et al.* (2001) and Katayama and Fujisawa (2016). For articles

focusing on SURE and cross-validation thresholding see Nason (1995), Nason (1996) and Altaher and Ismail (2010). This is mainly in the case of normal independent noise. Nason (1995) claimed that when correlated noise is used the cross-validation methods do not perform as well. Nason (1995) also stated that the universal threshold only detects the "large discontinuities" compared to the SURE and cross-validation methods. Abramovich *et al.* (1998) showed that the cross-validation method provides a good reconstruction for the Blocks test function. Furthermore, Katayama and Fujisawa (2016) studied classical thresholding rules and mentioned that soft thresholding performed worse than hard, SCAD and non-negative garrote thresholding.

There are several papers on classical thresholding estimation in the signal and image processing community. These papers usually use hard and soft thresholding rules with universal, SURE and cross-validation to specify the value of threshold λ.

2.13 Minimum mean squared-error for computing $\widehat{\lambda}_{\mathbf{MMSE}}$ and $\widehat{\Lambda}_{\mathbf{MMSE}}$

In this section, four procedures are applied to estimate $\mathbf{f}$ in model (2.8). All these approaches have two parameters, Λ for the inversion part and λ for the threshold value. These will be estimated by minimum mean squared-error (MMSE). Thus, let $\widehat{\lambda}_{\text{MMSE}}$ and $\widehat{\Lambda}_{\text{MMSE}}$ be the values of λ and Λ leading to the minimum MSE. In all cases coefficients in the lowest three resolution levels are left unchanged, that is $\lambda_j = 0$ for $j = 0, 1, 2$.

There are two approaches for estimating the value of threshold, one has a single parameter but uses $\lambda_j = 2^{-j/2}\lambda$, and $j = 3, 4, \ldots, J-1$ (Abramovich *et al.*, 1998) and the other is level-dependent, where $\boldsymbol{\lambda} = \{\lambda_3, \lambda_4, \ldots, \lambda_{J-1}\}$.

There are two fitting approaches, one is to compute $\widehat{\Lambda}_{\text{MMSE}}$ and $\widehat{\lambda}_{\text{MMSE}}$ separately (SE), the other is to compute the parameters together (TO). Furthermore, there are two modelling

methods used, one is to invert then threshold (IT) and the other is to threshold then inversion method is applied (TI). These can be defined as follows

$$\widehat{\mathbf{f}}_{\text{Reg}}^{\text{IT-SE}} = \arg\min_{\boldsymbol{\lambda}} \quad \mathbf{W}^T \mathrm{T}^*(\mathbf{W} \arg\min_{\Lambda} \quad \mathrm{I}_*(\mathbf{H}, \mathbf{y}, \Lambda), \boldsymbol{\lambda}), \tag{2.73}$$

$$\widehat{\mathbf{f}}_{\text{Reg}}^{\text{IT-TO}} = \arg\min_{\boldsymbol{\lambda},\Lambda} \quad \mathbf{W}^T \mathrm{T}^*(\mathbf{W} \mathrm{I}_*(\mathbf{H}, \mathbf{y}, \Lambda), \boldsymbol{\lambda}), \tag{2.74}$$

$$\widehat{\mathbf{f}}_{\text{Reg}}^{\text{TI-SE}} = \arg\min_{\Lambda} \quad \mathrm{I}_*(\arg\min_{\boldsymbol{\lambda}} \quad \mathbf{W}^T \mathrm{T}^*(\mathbf{W}\mathbf{y}, \boldsymbol{\lambda}), \mathbf{H}, \Lambda), \tag{2.75}$$

$$\widehat{\mathbf{f}}_{\text{Reg}}^{\text{TI-TO}} = \arg\min_{\Lambda,\boldsymbol{\lambda}} \quad \mathrm{I}_*(\mathbf{W}^T \mathrm{T}^*(\mathbf{W}\mathbf{y}, \boldsymbol{\lambda}), \mathbf{H}, \Lambda), \tag{2.76}$$

where $\mathrm{I}_*(\cdot)$ is an inversion method and $\mathrm{T}^*(\cdot)$ is a thresholding rule. Equation (2.73) indicates that the first step is to compute the value of $\widehat{\Lambda}_{\text{MMSE}}$ and then the second step is to compute the value of $\widehat{\boldsymbol{\lambda}}_{\text{MMSE}}$ by using the IT method separately. Equation (2.74) computes both values of $\widehat{\Lambda}_{\text{MMSE}}$ and $\widehat{\boldsymbol{\lambda}}_{\text{MMSE}}$ together by using the IT method together. Similarly, Equation (2.75) indicates that the first step is to compute the value of $\widehat{\boldsymbol{\lambda}}_{\text{MMSE}}$ and then the second step is to compute the value of $\widehat{\Lambda}_{\text{MMSE}}$ by using the TI method separately. Equation (2.76) gives the estimate of the values of $\widehat{\Lambda}_{\text{MMSE}}$ and $\widehat{\boldsymbol{\lambda}}_{\text{MMSE}}$ together by using the TI method together.

The method of MMSE, in Algorithm 2, can be described as; the first step is to start with the initial value of $\Lambda_0 \in \mathbb{R}^+$ and $\boldsymbol{\lambda}_0 \subset \mathbb{R}^+$, and the second step is to compute $\widehat{\mathbf{f}}$ using

$$\widehat{\mathbf{f}}_{\text{Reg}}^{\text{IT-TO}} = \mathbf{W}^T \mathrm{T}^*(\mathbf{W} \mathrm{I}_*(\mathbf{H}, \mathbf{y}, \Lambda_0), \boldsymbol{\lambda}_0),$$

where $\mathrm{I}_*(\cdot)$ is an inverse method and $\mathrm{T}^*(\cdot)$ is a thresholding rule. Then $\text{MSE} = \sum_{i=1}^{m} (\widehat{f}_i^{\text{IT-TO}} - f_i)^2$ is computed. At iteration k, propose a new parameter for $\Lambda^* = \Lambda^k + \epsilon$, with spread parameter, τ_1, chosen to achieve an acceptable convergence rate. If $\Lambda^* > 0$, then $\widehat{\mathbf{f}}$ is computed using

$$\widehat{\mathbf{f}}_{\text{Reg}}^{\text{IT-TO}^*} = \mathbf{W}^T \mathrm{T}^*(\mathbf{W} \mathrm{I}_*(\mathbf{H}, \mathbf{y}, \Lambda^*), \boldsymbol{\lambda}^{k-1}).$$

Then $\text{MSE}_{\text{new}} = \sum_{i=1}^{m} (\widehat{f}_i^{\text{IT-TO}^*} - f_i)^2$ is computed. If $\text{MSE}_{\text{new}} < \text{MSE}$ then the proposal is accepted $\Lambda^k = \Lambda^*$ and $\text{MSE} = \text{MSE}_{\text{new}}$. Otherwise the value is reset with $\Lambda^k = \Lambda^{k-1}$. Again, propose new parameters for $\lambda_3^* = \lambda_3^k + \epsilon, \lambda_4^* = \lambda_4^k + \epsilon, \ldots, \lambda_{J-1}^* = \lambda_{J-1}^k + \epsilon$, with spread parameter, $\boldsymbol{\tau}_2$, chosen to achieve an acceptable convergence rate. if $\lambda_j^* > 0$. Then

Algorithm 2: MMSE algorithms

Result: MMSE, $\widehat{\Lambda}_{\text{MMSE}}, \widehat{\lambda}_{\text{MMSE}}$

Starting with initial value $\Lambda_0 = 0.01, \boldsymbol{\lambda} = \{\lambda_{10}, \lambda_{20}, \ldots, \lambda_{J-1_0}\}$, $\tau_1 = 0.01$ and $\boldsymbol{\tau}_2 = \{0.01, 0.01, \ldots, 0.01, 0.01\}$.

Compute $\widehat{\mathbf{f}} = \mathbf{W}^T \mathrm{T}^{\mathrm{H}}(\mathbf{W}(\mathbf{H}^T\mathbf{H} + \Lambda_0 \mathfrak{R}^T\mathfrak{R})^{-1}\mathbf{H}^T\mathbf{y}, \boldsymbol{\lambda})$ and MSE=$\sum_i^m (\widehat{f}_i - f_i^{\text{true}})^2/M$.

for i=1 to R **do**

  for k=1 to M **do**

    Generate ϵ from a Gaussian distribution $N(0, \tau_1)$

    $\Lambda^* = \Lambda^k + \epsilon$

    if($\Lambda^* > 0$){

    Compute $\widehat{\mathbf{f}}^* = \mathbf{W}^T \mathrm{T}^{\mathrm{H}}(\mathbf{W}(\mathbf{H}^T\mathbf{H} + \Lambda^* \mathfrak{R}^T\mathfrak{R})^{-1}\mathbf{H}^T\mathbf{y}, \boldsymbol{\lambda})$ and MSE*

    if (MSE* <MSE^{k-1}) **then**

      Λ^k=Λ^*,

      MSE^k =MSE*

    else

      Λ^k=Λ^{k-1}

    end

    }

    Generate ϵ_3 from a Gaussian distribution $N(0, \tau_{2,1})$

    ${\lambda_3}^* = {\lambda_3}^k + \epsilon_3$

    Generate ϵ_4 from a Gaussian distribution $N(0, \tau_{2,2})$

    ${\lambda_4}^* = {\lambda_4}^k + \epsilon_4, \ldots$

    for j=3 to J-1 **do**

      if($\lambda_j^* > \mathbf{0}$){

      Compute $\widehat{\mathbf{f}}^* = \mathbf{W}^T \mathrm{T}^{\mathrm{H}}(\mathbf{W}(\mathbf{H}^T\mathbf{H} + \Lambda^{k-1} \mathfrak{R}^T\mathfrak{R})^{-1}\mathbf{H}^T\mathbf{y}, \{\lambda_3^{k-1}, \ldots, \lambda_j^*, \ldots, \lambda_{J-1}^{k-1}\})$ and MSE*

      if (MSE* <MSE^{k-1}) **then**

        $\lambda_j^k = \lambda_j^*$,

        MSE^k =MSE*

      else

        $\lambda_j^k = \lambda_j^{k-1}$

      end

      }

    end

  end

  Update τ_1 and $\boldsymbol{\tau}_2$

end

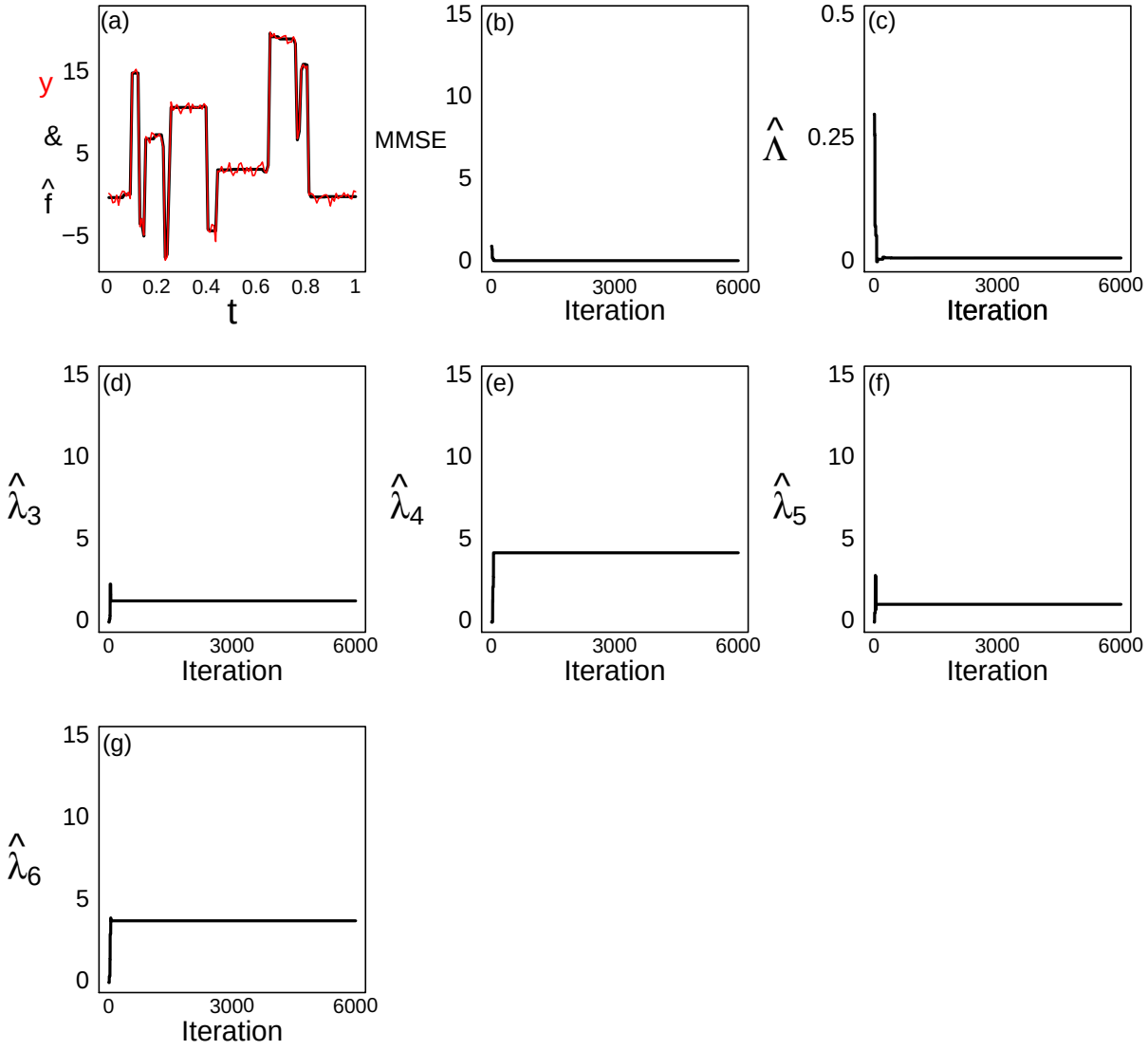


Figure 2.21: Plots of the monitoring minimum MSE algorithm using the IT method with the hard thresholding rule where $\boldsymbol{\lambda}$ and Λ are estimated together. The Blocks test function, at $m = 128$ equally spaced points is used and corrupted by level of noise and blur, which is given in (2.6), equal to 0.5 and 0.001, respectively: (a) the red line represents the true Blocks test function and the black line represents the result of the estimate at transient period of 6000 iterations; (b) minimum MSE is acceptable for the new value in each iteration; (c) acceptable Λ; (d) acceptable λ_3; (e) acceptable λ_4; (f) acceptable λ_5, and (d) acceptable λ_6.

for each level, $\widehat{\mathbf{f}}$ is computed using

$$\widehat{\mathbf{f}}_{\text{Reg}}^{\text{IT-TO}^*} = \mathbf{W}^T \mathrm{T}^*(\mathbf{W}\mathrm{I}_*(\mathbf{H}, \mathbf{y}, \Lambda^{k-1}), \{\lambda_3^{k-1}, \lambda_4^{k-1}, \ldots, \lambda_j^* \ldots, \lambda_{J-1}^{k-1}\}).$$

Similarly, $\text{MSE}_{\text{new}} = \sum_{i=1}^{m} (\widehat{f}_i^{\text{IT-TO}^*} - f_i)^2$ is computed. If $\text{MSE}_{\text{new}} < \text{MSE}$ then the proposal is accepted $\lambda_j^k = \lambda_j^*$ and $\text{MSE} = \text{MSE}_{\text{new}}$. Otherwise the value is reset with $\lambda_j^k = \lambda_j^{k-1}$.

The number of replications is equal to $R = 60$ and the number of iteration equals $M = 100$. This means that the total number of runs is equal to 6000. The same datasets are used for simulation and comparison. In other words, Λ and $\boldsymbol{\lambda}$ are chosen to minimise the mean squared-error, this is

$$\widehat{\boldsymbol{\theta}}_{\text{MMSE}} = \underset{\boldsymbol{\theta}}{\text{argmin}} \quad ||\mathbf{f} - \widehat{\mathbf{f}}_\theta||_2^2, \tag{2.77}$$

where $\widehat{\boldsymbol{\theta}}_{\text{MMSE}}$ is computed by two approaches the first is for level-dependent $\boldsymbol{\theta} = \{\Lambda, \boldsymbol{\lambda}\}$, and the second is to use $\boldsymbol{\lambda} = \{2^{-j/2}\lambda : j = 3, 4, \ldots, J-1, J = \log_2(n)\}$.

Figure 2.21 shows an example of monitoring the progress of Algorithm 2 using the IT method with the hard thresholding rule where $\boldsymbol{\lambda}$ and Λ are proposed together.

2.14 Comparison simulation

The purpose of this section is to evaluate and investigate whether universal, SURE or cross-validation methods are suitable for estimating the value of the threshold λ. First of all, one of the methods in (2.73), (2.74), (2.75) and (2.76) is chosen. The reason for choosing the MMSE approach is that, in practice, the true function $\mathbf{f}$ in (2.33) is unknown. However, simulated data, that is an artificial data set, can be used to reflect correct belief about the real data. The MMSE, described in Section 2.13, will be used to determine the optimal parameters Λ and $\boldsymbol{\lambda}$ and then these parameter values will be used to produce a reconstruction of the original function $\mathbf{f}$.

The simulated data sets consisted of the standard test signals Blocks and Bumps at $m = 128$ equally spaced points (Donoho and Johnstone, 1994; Nason and Silverman, 1994), multiplied by a blur matrix, which is given in (2.6), with k taken as 0.001, 0.005, and 0.01. Also, the dataset was corrupted by independent Gaussian noise with mean zero, and variance was taken as 0.5. Moreover, no thresholding was done below level 3. The first-order method in Section 2.6 is used to estimate $\mathbf{f}$.

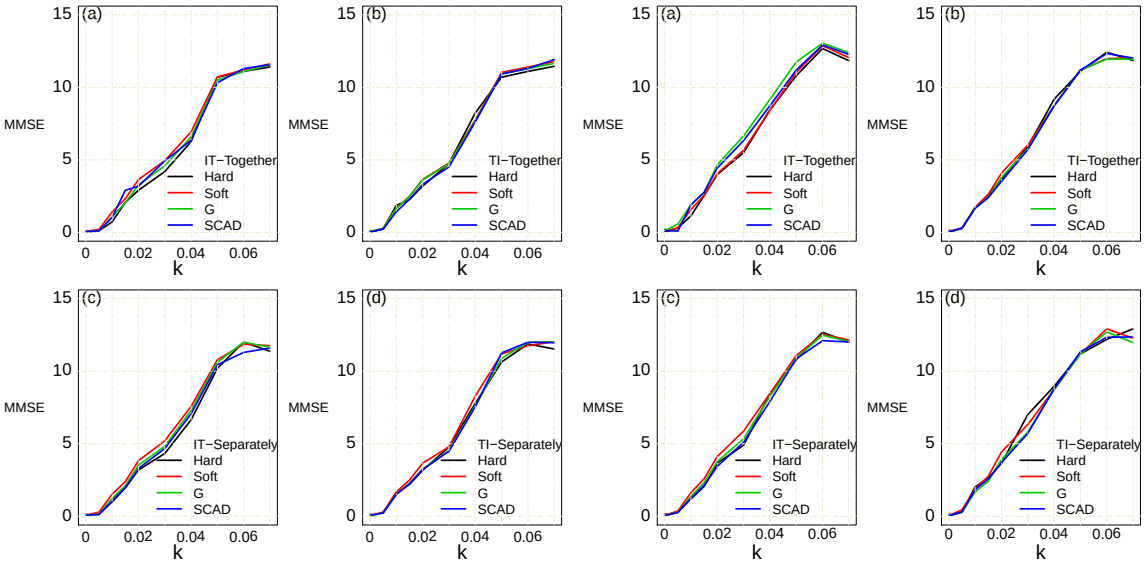
Figures 2.22 (i) and (ii) show the plots of MMSEs for different thresholding rules to recover the Blocks test function. In each plot the vertical axis is MMSE, and the horizontal axis is the blur k. In general, the method of IT in (2.74) provides a smaller MSE result than others in (2.73), (2.75) and (2.76). Also, the use of level-dependent method, λ_j, improves the MSE compared to use $\lambda_j = 2^{-j/2}\lambda$, see the plots 2.22 (i).

Similarly, Figures 2.23 (i) and 2.23 (ii) show the plots of MMSEs for different thresholding rules to recover the Bumps test function. In general, the IT methods in (2.74) provides a small MSE and level-dependent, λ_j, improves the MSE compared to use $\lambda_j = 2^{-j/2}\lambda$.

The second simulation is to evaluate and investigate whether universal threshold, SURE and cross-validation method are suitable for estimating an unknown function, where the IT methods and AMSE were used, as described in section 2.7. Figures 2.24 and 2.25 show the boxplots of MSE for different thresholding rules for estimating Blocks and Bumps test function. The plots of the boxplots depend on the minimum AMSE in (2.22).

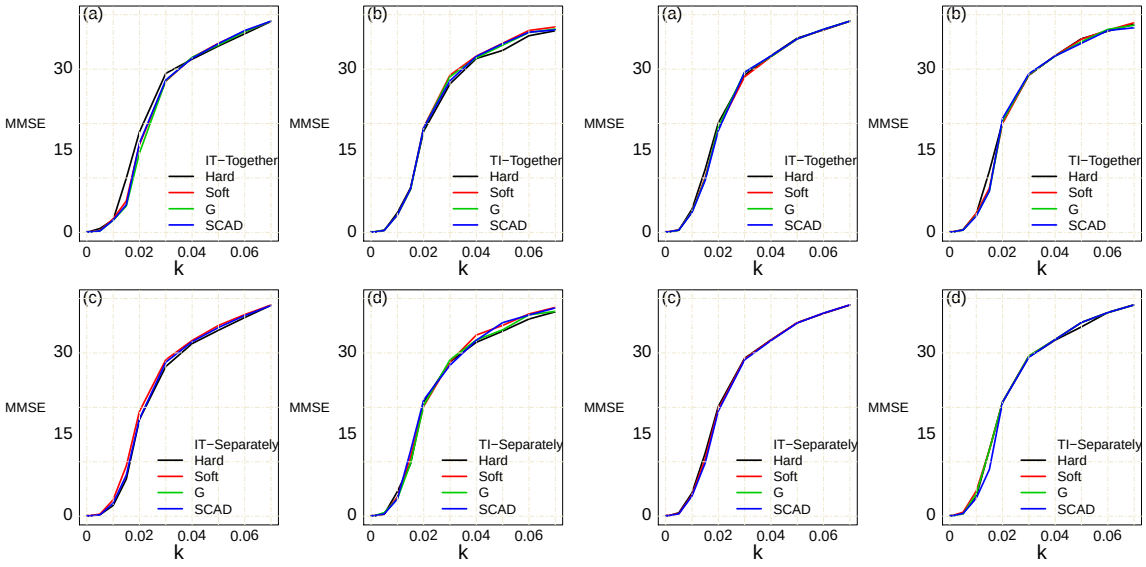
Figure 2.24 shows the boxplots of MSE for different thresholding rules for estimating Blocks test function using the IT method; the first row represents the universal thresholding method, the second row represents the cross-validation method and the third row represents the SURE method. For $k = 0.001$, the cross-validation method with the hard thresholding rule provides a smaller MSE than the universal and SURE methods. For $k = 0.005$ and 0.01, the cross-validation and universal thresholds with the hard rule improve the MSE.

Similarly, Figure 2.25 shows the boxplots of MSE for different thresholding rules for estimating Bumps using the IT method. For $k = 0.001$ the universal, SURE and cross-



(i) Level-dependent, $\boldsymbol{\lambda}$. (ii) Single parameter uses $\lambda_j = 2^{-j/2}\lambda$

Figure 2.22: Simulated Blocks test function at $m = 128$ equally spaced points: plots of minimum MSE with first-order smoothing with different values of blur, which is given in (2.6); (i) estimate λ_j for each resolution level $j = 3, 4, 5, 6$, and (ii) estimate λ_j for each resolution level $j = 3, 4, 5, 6$, where $\lambda_j = 2^{-j/2}\lambda$.



(i) Level-dependent, $\boldsymbol{\lambda}$. (ii) Single parameter uses $\lambda_j = 2^{-j/2}\lambda$.

Figure 2.23: Simulated Bumps test function at $m = 128$ equally spaced points: plots of minimum MSE with first-order smoothing with different values of blur, which is given in (2.6); (i) estimate λ_j for each resolution level $j = 3, 4, 5, 6$, and (ii) estimate λ_j for each resolution level $j = 3, 4, 5, 6$, where $\lambda_j = 2^{-j/2}\lambda$.

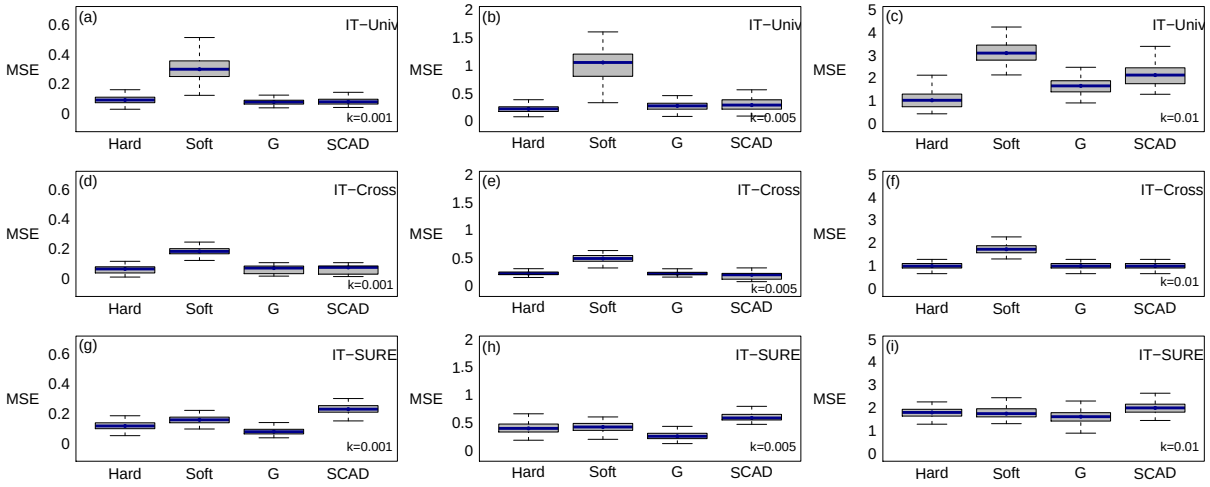


Figure 2.24: Boxplots of MSE results with first-order smoothing for estimating Blocks test function at $m = 128$ equally spaced points. Each column represents different blur, which is given in (2.6), $k = 0.001, 0.005, 0.01$ and each row represents different methods using the IT method.

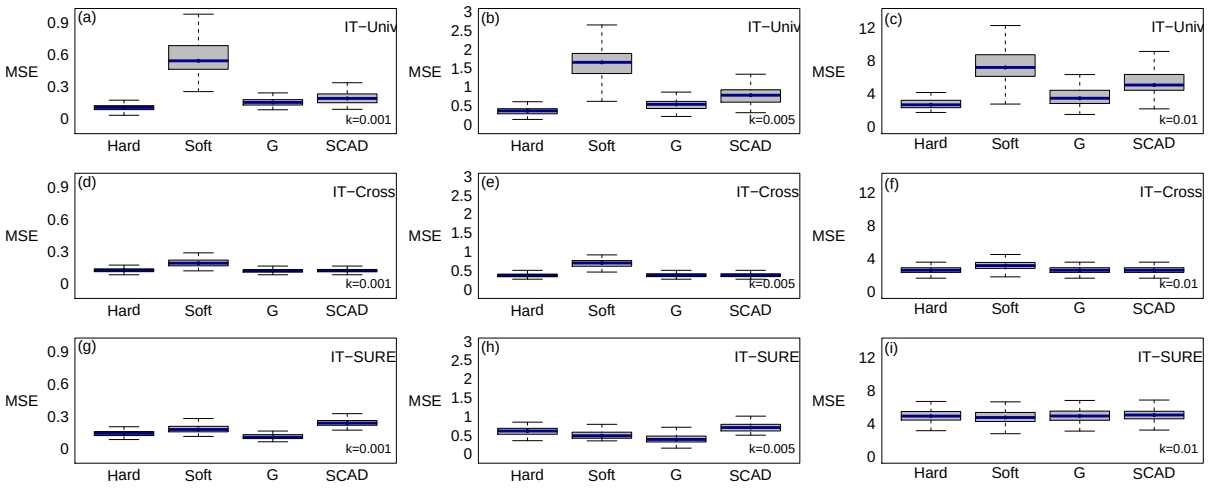


Figure 2.25: Boxplots of MSE results with first-order smoothing for estimating Bumps test function at $m = 128$ equally spaced points. Each column represents different blur, which is given in (2.6), $k = 0.001, 0.005, 0.01$ and each row represents different methods using the IT method.

validation methods with the hard rule, provide a small MSE. Finally, for $k = 0.005$ and 0.01 the cross-validation and the universal thresholding improve MSE.

2.15 Conclusions

This chapter has introduced different inversion and wavelet methods for deblurring and denoising. The inversion methods are used for estimating the underlying function $\mathbf{f}$. The main result is that for small noise with $\sigma = 0.5$ and small level of blur $k = 0.001, 0.005$ and 0.01 there is no need to use inversion. Which means that $\Lambda = 0$ and the underlying function can be estimated by ML in Equation (2.10). However, for large levels of noise $\sigma = 1$ and blur $k = 0.07$, the smoothing parameter Λ becomes bigger than 0. In general, ridge regression gives better result when $\sigma = 0.5$ and $\Lambda = 0.001, 0.005$ and 0.01. Also, for $\sigma = 1$ and level of blur $k = 0.001, 0.005$ and 0.01, first-order smoothing provides a small MSE.

We extended the classical thresholding rule with different methods for the choice of the value of λ. An investigation into applying a thresholding rule, with an inversion method, finds that the MSE improves. For example, the IT method with the cross-validation thresholding gives better results than is achieved by using only an inversion method. Extensive simulation studies were presented using MSE to compute $\widehat{\Lambda}_{\text{MMSE}}$ and $\widehat{\lambda}_{\text{MMSE}}$ with different approaches, such as separately and together, with different thresholding rules for estimating $\mathbf{f}$. It can be concluded that, the IT method works well and gives slightly better results than the TI method.

Chapter 3

Decimated and non-decimated wavelet transforms

3.1 Overview

This chapter is organized as follows: Section 3.2 gives an introduction, Section 3.3 explains multi-resolution analysis, Sections 3.4 to 3.6 consider the discrete wavelet transform, whilst Section 3.7 describes the non-decimated wavelet transform, and finally Section 3.9 gives conclusions.

3.2 Introduction

Recall from Equation (2.38), the model

$$\mathbf{y} = \mathbf{H}\mathbf{f} + \boldsymbol{\epsilon}, \tag{3.1}$$

where $\mathbf{H}$ is a given $n \times m$ blur matrix, $\mathbf{y}$ is an $n \times 1$ vector of data, $\mathbf{f}_{m\times 1}$ is our signal of interest and $\boldsymbol{\epsilon}$ is a vector of random variables, such that $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$ then

$$\mathbf{d_y} = \mathbf{W}\mathbf{y} = \mathbf{W}(\mathbf{H}\mathbf{f} + \boldsymbol{\epsilon}) = \mathbf{W}\mathbf{H}\mathbf{f} + \mathbf{W}\boldsymbol{\epsilon} = \mathbf{d_g} + \boldsymbol{\eta}, \tag{3.2}$$

where $\mathbf{d_g}$ is an $n \times 1$ vector, containing the wavelet coefficients of $\mathbf{g} = \mathbf{Hf}$. Thus, the model in (3.1) can be written equivalently as

$$\mathbf{d_y} = \mathbf{d_g} + \boldsymbol{\eta}, \tag{3.3}$$

where $\mathbf{d_y} = \mathbf{Wy}$ and $\mathbf{d_g} = \mathbf{Wg}$.

In order to understand how a wavelet decomposition can be computed and constructed, the principle of a multi-resolution analysis should be introduced and studied. The purpose of the multi-resolution analysis is to provide a "window" to look at the wavelet coefficients at a particular level. Mallat (1989) introduced the fast filtering algorithm, known as the *cascade algorithm*, which involves a recursive formula to produce wavelet coefficients. This process allows the computation of higher level coefficients from lower level coefficients and vice versa (Hubbard, 1996). An excellent critical overview and introduction to the discrete wavelet transform can be found in Mallat (1989), Vidakovic (1999), Nason (2010a) and Jerri (2011).

Several authors have considered the non-decimated wavelet transform, which have included many points of view and many problems. See, for example, Nason and Silverman (1995), Coifman and Donoho (1995), Pesquet *et al.* (1996) and Silverman (1999). Some recent results about theoretical properties and comparisons of the discrete wavelet transform and the non-decimated wavelet transform can be found in Gyaourova *et al.* (2002), Starck *et al.* (2007) and Vidya (2008).

The usual approach for wavelet thresholding is to choose a wavelet basis, compute the wavelet coefficients using either the (decimated) discrete wavelet transform or the non-decimated wavelet transform, and then use a thresholding rule. The reconstruction is typically found using the inverse transform. Extensions to the discrete wavelet transform or non-decimated wavelet transform, have been considered in several articles. Wavelet packets are proposed by Coifman and Wickerhauser (1992), which are a richer family that are more flexible in representing different types of signals than the standard wavelet transformation.

Full, introductory and mathematical accounts can be found in several texts such as Mallat (1989), Daubechies (1992), Meyer (1995),Wickerhauser (1994), Nason and Silverman (1995), Coifman and Donoho (1995), Silverman (1999) and Vidakovic (1999). In the next section, multi-resolution analysis is described and an example is included.

3.3 Multi-resolution analysis (MRA)

The main idea of wavelets is to adapt automatically to the different components of a signal or image, using a large window to look at long-lived components of low frequency and small windows to look at short-lived components of high frequency. This procedure is known as *Multi-resolution Analysis* (Hubbard, 1996). The concept of MRA is useful as it provides a natural framework for the understanding of wavelet bases (Mallat, 1989). Recall from Chapter 2, two functions are structured by Abramovich *et al.* (2000) follow the concept of wavelet which was adopted by Daubechies (1992), are given by

$$\phi_{j_0,l}(t) = 2^{j_0/2}\phi(2^{j_0}t - l), \quad j_0 \in \mathbb{Z}, \quad 0 \leq l \leq 2^{j_0} - 1; \tag{3.4}$$

and

$$\psi_{j,l}(t) = 2^{j/2}\psi(2^{j}t - l), \quad j = j_0, j_0 + 1, \ldots. \tag{3.5}$$

Daubechies (1992) defined the MRA as a sequence of closed subspaces $\mathbf{v}_j$ of $L^2(\mathbb{R})$, $j \in \mathbb{Z}$. These subspaces satisfy the following properties:

1. $\ldots \mathbf{v}_{-1} \subset \mathbf{v}_0 \subset \mathbf{v}_1 \ldots$,
2. $\bigcup_{j=-\infty}^{+\infty} \mathbf{v}_j = L^2(\mathbb{R})$, and
3. $\bigcap_{j=-\infty}^{+\infty} \mathbf{v}_j = \{0\}$.

The requirements (2) and (3) mean that the intersection of these subspaces is trivial and the union is dense in $L^2(\mathbb{R})$. The structure of the spaces in (1) is constructed so that all of the spaces are simply scaled versions of the central space, $\mathbf{v}_0$, more precisely,

$$f(2^j t) \in \mathbf{v}_j \text{ iff } f(t) \in \mathbf{v}_0. \tag{3.6}$$

The structure of (1) is also constructed such that there exists a scaling function, $\phi \in \mathbf{v}_0$, whose integer translates span, $\mathbf{v}_0$ (Vidakovic, 1999),

$$\mathbf{v}_0 = \{f \in L^2(\mathbb{R}) | f(t) = \sum_l c_l \phi(t-l)\}, \tag{3.7}$$

for some coefficients, c_l. Invariance of $\mathbf{v}_0$ under integer translations is also required, consequently

$$f \in \mathbf{v}_0 \Leftrightarrow f(\cdot - l) \in \mathbf{v}_0, \quad \forall l \in \mathbb{Z}. \tag{3.8}$$

Hence, (3.6) and (3.8) imply that

$$f \in \mathbf{v}_j \Leftrightarrow f(\cdot - 2^j l) \in \mathbf{v}_j, \quad \forall l \in \mathbb{Z}, \tag{3.9}$$

also, it is required that there exists $\phi \in \mathbf{v}_0$, such that

$$\{\phi_{0,l}, l \in \mathbb{Z}\} \quad \text{is an orthonormal basis in } \mathbf{v}_0, \tag{3.10}$$

where the scaling function, $\phi_{j,l}(t)$, has been dilated and translated, as previously defined in (2.23).

These requirements for an MRA can be used to construct a mother wavelet, ψ. The wavelet space, $\mathbf{w}_j$, is defined to denote the difference in space between $\mathbf{v}_{j+1}$ and $\mathbf{v}_j$. This means for every $j \in \mathbb{Z}$, there is an orthonormal space complementing $\mathbf{v}_j$ in $\mathbf{v}_{j+1}$, which can be written as

$$\mathbf{v}_{j+1} = \mathbf{v}_j \oplus \mathbf{w}_j.$$

As all of these subspaces are orthogonal by the requirements (2) and (3), this then implies that

$$\bigoplus_j \mathbf{w}_j = L^2(\mathbb{R}),$$

which is a decomposition of $L^2(\mathbb{R})$ into mutually orthogonal subspaces. Consequently, the spaces, $\mathbf{w}_j$, inherit the scaling property, which means,

$$f \in \mathbf{w}_0 \Leftrightarrow f(2^j \cdot) \in \mathbf{w}_j. \tag{3.11}$$

If it is now supposed that there is a MRA with scaling function, $\phi \in \mathbf{v}_0$, from (3.10) it is known that the integer translates $\phi(t-l)$ form is an orthogonal basis for $\mathbf{v}_0$ and it is true, from (3.9), that the half-integer translates, $\phi(2t-l)$, all lie in the space, $\mathbf{v}_1$, and are orthogonal (Broughton and Bryan, 2011). Also, by utilizing the fact that $\mathbf{v}_0 \subset \mathbf{v}_1$ the scaling function, $\phi(t)$, can be represented as a linear combination of functions from $\mathbf{v}_1$ (Vidakovic, 1999). From this, $\phi(t)$, can be defined by

$$\phi(t) = \sum_{l\in\mathbb{Z}} h_l \sqrt{2}\phi(2t-l), \tag{3.12}$$

where $\mathbf{h} = \{h_l : l \in \mathbb{Z}\}$ is a vector. Within the signal processing literature this is referred to as the *normalization property* (Vidakovic, 1999). The normalization property is

$$\sum_{l\in\mathbb{Z}} h_l = \sqrt{2}.$$

This property can be proven by considering the following

$$\begin{aligned}\int \phi(t)dt &= \sum_{l\in\mathbb{Z}} h_l\sqrt{2}\int \phi(2t-l)dt \\ &= \sum_{l\in\mathbb{Z}} h_l\frac{\sqrt{2}}{2}\int \phi(2t-l)d(2t-l) \\ &= \sum_{l\in\mathbb{Z}} h_l\frac{\sqrt{2}}{2}\int \phi(t)dt,\end{aligned} \tag{3.13}$$

as, by assumption, $\int \phi(t)dt \neq 0$. The result follows as

$$\sum_{l\in\mathbb{Z}} h_l = \sqrt{2}. \tag{3.14}$$

Moreover, the orthogonality property is, for any $n \in \mathbb{Z}$,

$$\sum_{l\in\mathbb{Z}} h_l h_{l-2n} = \delta_n, \tag{3.15}$$

where $\delta_n = \int \psi(t)\psi(t-n)$. The result (3.15) is proven by Vidakovic (1999), first noting the scaling function (3.12), it follows that

$$\begin{aligned}\psi(t)\psi(t-n) &= \sqrt{2}\sum_l h_l\psi(2t-l)\psi(t-n) \\ &= \sqrt{2}\sum_l h_l\psi(2t-l)\sum_m h_m\psi(2(t-n)-m).\end{aligned} \tag{3.16}$$

Integrating both sides in (3.16) shows that

$$\begin{aligned}\delta_n &= 2\sum_l h_l\left[\sum_m h_m \frac{1}{2}\int \psi(2t-l)\psi(2t-2n-m)d(2t)\right] \\ &= \sum_l h_l \sum_m h_m \delta_{l,2n+m} \\ &= \sum_l h_l h_{l-2n}. \end{aligned} \tag{3.17}$$

The last line is obtained by taking $m = l - 2n$ as $\delta_{l,2n+m} = \delta_{l,l} = \sum_l h_l h_l = 1$ when $l = 2n + m$ and 0 otherwise. An important special case is that when $n = 0$, (3.15) becomes

$$\sum_l h_l h_{l-0} = \sum_l h_l^2 = 1. \tag{3.18}$$

In order to better explore the properties of MRA subspaces and their corresponding bases, Vidakovic (1999) is followed where a move into the Fourier domain is undertaken. Initially, a function $m_0(\omega)$, is defined as

$$m_0(\omega) = \frac{1}{\sqrt{2}}\sum_{l\in\mathbb{Z}} h_l e^{-il\omega} = \frac{1}{\sqrt{2}}H(\omega),$$

where this function is referred to as the *transfer function* and it describes the behavior of the associated **h** filter in the Fourier domain. Note that m_0 is a 2π periodic function and that $\{h_l : l \in \mathbb{Z}\}$ are the Fourier coefficients of the function $H(\omega) = \sqrt{2}m_0(\omega) = \sum_{l\in\mathbb{Z}} h_l e^{-il\omega}$. Hence $m_0(0) = 1$. Thus, in the Fourier domain, (3.12) becomes

$$\Phi(w) = m_0\left(\frac{\omega}{2}\right)\Phi\left(\frac{\omega}{2}\right),$$

where $\Phi(w)$ is the Fourier transform of $\phi(t)$. Vidakovic (1999) proved that

$$\begin{aligned}\Phi(\omega) &= \int_{-\infty}^{\infty}\phi(t)e^{-iwt}dt \\ &= \sum_l \sqrt{2}h_k \int_{-\infty}^{\infty}\phi(2t-l)e^{-iwt}dt \\ &= \sum_l \frac{h_l}{\sqrt{2}}e^{-ilw/2}\int_{-\infty}^{\infty}\phi(2t-l)e^{-iw(2t-l)/2}d(2t-l) \\ &= \sum_l \frac{h_l}{\sqrt{2}}e^{-ilw/2}\Phi\left(\frac{\omega}{2}\right) \\ &= m_0\left(\frac{\omega}{2}\right)\Phi\left(\frac{\omega}{2}\right). \end{aligned} \tag{3.19}$$

Note that the scaling function can be written in the Fourier domain. In general, wherever a sequence of subspace exists satisfying the criteria for MRA, there exists, (though not unique), an orthonormal basis for $L^2(\mathbb{R})$,

$$\{\psi_{j,l}(t) = 2^{j/2}\psi(2^j t - l), \quad j, l \in \mathbb{Z}\}, \tag{3.20}$$

such that $\{\psi_{j,l}(t) : j - \text{fixed}, \quad l \in \mathbb{Z}\}$ is an orthogonal basis of the *difference space*, $\mathbf{w}_j = \mathbf{v}_{j+1} \ominus \mathbf{v}_j$. Since $\psi(t) \in \mathbf{v}_1$, and due to the containment of $\mathbf{w}_0 \subset \mathbf{v}_1$, Vidakovic (1999) showed that the wavelet function, $\psi(t)$, in (2.24) can be represented by

$$\psi(t) = \sum_{l \in \mathbb{Z}} g_l \sqrt{2}\psi(2t - l), \tag{3.21}$$

for some coefficients, g_l, where $l \in \mathbb{Z}$ and $\mathbf{g} = \{g_l : l \in \mathbb{Z}\}$ is a vector. Moreover, it becomes convenient to define a function, $m_1(\omega)$, to be the wavelet function in the Fourier domain:

$$m_1(\omega) = \frac{1}{\sqrt{2}} \sum_{l \in \mathbb{Z}} g_l e^{-il\omega}. \tag{3.22}$$

Then the Fourier transform of $\psi(t)$, is given by

$$\Psi(w) = m_0\left(\frac{\omega}{2}\right)\Psi\left(\frac{\omega}{2}\right), \tag{3.23}$$

by virtue of the fact that $\mathbf{v}_0$ and $\mathbf{w}_0$ are orthogonal via construction. Through computation of Fourier series, Vidakovic (1999) defined

$$m_1(\omega) = -e^{-i\omega}\overline{m_0(\omega + \pi)}. \tag{3.24}$$

A detailed proof of Equation (3.4) and this subsequent definition can be found in Härdle *et al.* (1998). This definition also leads to a standard relation between $\mathbf{h}$ and $\mathbf{g}$. By comparing this definition of $m_1(\omega)$ in (3.24) with that given in Equation (3.4), this gives

$$\begin{aligned} m_1(\omega) &= \frac{-e^{-i\omega}}{\sqrt{2}} \sum_l h_l e^{-i(\omega+\pi)l} \\ &= \frac{1}{\sqrt{2}} \sum_l h_l (-1)^{1-l} e^{-i\omega(1-l)} \\ &= \frac{1}{\sqrt{2}} \sum_l h_{1-n} (-1)^n e^{-i\omega n}, \end{aligned} \tag{3.25}$$

since, the second line is obtained from the Fourier properties, $\mathcal{F}(\omega) = -\mathcal{F}(\omega + \pi)$ and $\mathcal{F}(\omega)$ is 2π-periodic. Thus, the standard relation between $\mathbf{h}$ and $\mathbf{g}$, is given by

$$g_l = (-1)^l h_{1-l}. \tag{3.26}$$

In the literature related to signal processing, this relation is known as the *quadrature mirror relation*, with filters $\mathbf{h}$ and $\mathbf{g}$ subsequently referred to as the *quadrature mirror filters*.

As an example, consider the Haar scaling function given in (3.12). By inspection of simple graphs of two scaled Haar wavelets, $\phi(2t)$ and $\phi(2t+1)$, which stuck to each other, it can be concluded that the scaling equation is given by

$$\begin{aligned} \phi(t) &= \phi(2t) + \phi(2t-1) \\ &= \sqrt{2}h_0\phi(2t) + \sqrt{2}h_1\phi(2t-1) \\ &= \phi(t) + \phi(2t-1). \end{aligned} \tag{3.27}$$

Hence, the quadrature mirror relation is

$$h_0 = h_1 = \frac{1}{\sqrt{2}}.$$

The Haar scaling function can be written in the Fourier domain as

$$\begin{aligned} m_0(\omega) &= \frac{1}{\sqrt{2}} \sum_{l \in \mathbb{Z}} h_l e^{-il\omega} \\ &= \frac{1}{\sqrt{2}} \left(\frac{1}{\sqrt{2}} e^{-i\omega 0} \right) + \frac{1}{\sqrt{2}} \left(\frac{1}{\sqrt{2}} e^{-i\omega 1} \right) \\ &= \frac{1}{\sqrt{2}} \left(\frac{1}{\sqrt{2}} e^{0} \right) + \frac{1}{\sqrt{2}} \left(\frac{1}{\sqrt{2}} e^{-i\omega 1} \right) \\ &= \frac{1 + e^{-i\omega}}{2}, \end{aligned}$$

and

$$m_1(\omega) = e^{-i\omega}\overline{m_0(\omega + \pi)} = e^{-i\omega} \left(\frac{1}{2} - \frac{1}{2} e^{i\omega} \right) = \frac{1 - e^{-i\omega}}{2}.$$

Further

$$\Psi(w) = \frac{1 - e^{-i\omega/2}}{2} \Phi\left(\frac{\omega}{2} \right) = \frac{1}{2}\Phi\left(\frac{\omega}{2} \right) - \frac{1}{2}\Phi\left(\frac{\omega}{2} \right) e^{-i\omega/2}.$$

Applying the inverse Fourier transformation, Vidakovic (1999) proved that

$$\psi(t) = \phi(2t) - \phi(2t - l),$$

in the time domain. Using the transfer function in the Fourier domain, it can then be shown that

$$g_0 = -g_{-1} = \frac{1}{\sqrt{2}}.$$

Note that $h_0 + h_1 = \sqrt{2}$ and $h_0^2 + h_1^2 = 1$.

3.4 The cascade algorithm

The *cascade algorithm* process is known alternatively as the *pyramid algorithm* when using a recursive formula. This formula allows the computation of higher level coefficients from lower level coefficients and vice versa (Härdle *et al.*, 1998).

Mallat (1989) was the first to connect together wavelets, multi-resolution analyses and cascade algorithms in a formal way. The wavelet scaling and detail coefficients are estimated by filtering with **h** and **g**, which are defined in Section 3.3. The following explanation of the cascade algorithm can be found in Vidakovic (1999). Upon considering the problem of the wavelet decomposition, it is convenient to link the original signal with the coefficients from the space, $\mathbf{v}_J$, for some J. Suppose the two subspaces $\mathbf{v}_j$ and $\mathbf{w}_j$ are available at resolution level $j, j-1$ and so on. When the index in **v**-spaces is decreased, this is equivalent to coarsening the approximation to the data. Through simple substitution of indices in the scaling for $\phi(t)$ and $\psi(t)$ given in (3.12) and (3.21) respectively, one obtains

$$\phi_{j-1,l}(t) = \sum_{n\in\mathbb{Z}} h_{n-2l}\phi_{j,l}(t) \quad \text{and} \quad \psi_{j-1,l}(t) = \sum_{n\in\mathbb{Z}} g_{n-2l}\phi_{j,l}(t). \tag{3.28}$$

The relationships in (3.28) are fundamental to the procedure of the cascade algorithm. In order to start with this procedure one must first consider a multi-resolution analysis, which is explained in Section 3.3. It is usual to denote the wavelet coefficients associated with $\phi_{j,l}(t)$ and $\psi_{j,l}(t)$ by $\mathrm{c_f}_{j,l}$ and $\mathrm{d_f}_{j,l}$, which were defined in Chapter 2. Now, let $v_j \in \mathbf{v}_j$

and $w_j \in \mathbf{w}_j$, then $v_j(t)$, can be expressed by

$$\begin{aligned} v_j(t) &= \sum_l \mathrm{c_f}_{j,l}\phi(t)_{j,l} \\ &= \sum_l \mathrm{c_f}_{j-1,l}\phi(t)_{j-1,l} + \sum_l \mathrm{d_f}_{j-1,l}\psi(t)_{j-1,l} \\ &= v_{j-1}(t) + w_{j-1}(t), \end{aligned} \tag{3.29}$$

using the general scaling and wavelet function in (3.28). It can also be shown that

$$\begin{aligned} \mathrm{c_f}_{j-1,l} &=< v_j, \phi_{j-1,l} > \\ &=< v_j, \sum_l h_{l-2n}\phi(t)_{j,l} > \\ &= \sum_l h_{l-2n} < v_j, \phi(t)_{j,l} > \\ &= \sum_l h_{l-2n}\mathrm{c_f}_{j,l}. \end{aligned} \tag{3.30}$$

Similarly, $\mathrm{d_f}_{j-1,l} = \sum_l g_{l-2n}\mathrm{c_f}_{j,l}$. Coefficients in the next finer scale corresponding to $\mathbf{v}_j$ can be obtained from the coefficients corresponding to $\mathbf{v}_{j-1}$ and $\mathbf{w}_{j-1}$. This can be mathematically expressed by

$$\begin{aligned} \mathrm{c_f}_{j,l} &=< v_j, \phi_{j,l} > \\ &= \sum_l \mathrm{c_f}_{j-1,l} < \phi_{j-1,l}, \phi_{j,l} > + \sum_l \mathrm{d_f}_{j-1,l} < \psi_{j-1,l}, \phi_{j,l} > \\ &= \sum_l \mathrm{c_f}_{j-1,l} h_{l-2n} + \sum_l \mathrm{d_f}_{j-1,l} g_{l-2n}. \end{aligned} \tag{3.31}$$

The relation in (3.31) describes a single step in the reconstruction algorithm. Nason (2010a) and Vidakovic (1999) described the DWT of a signal $\mathbf{f}$ in (2.33), of size $n = 2^j$. Thus, the vector of wavelet coefficients, $\mathbf{d_f}$, of the form

$$\begin{aligned} \mathbf{d_f} &= \mathbf{W}\mathbf{f} \\ &= \left(\mathrm{c_f}_{00}, \mathrm{d_f}_{j-1,0}, \ldots, \mathrm{d_f}_{j-1,2^{j-1}-1}, \mathrm{d_f}_{j-2,0}, \ldots, \mathrm{d_f}_{j-2,2^{j-2}-1}, \ldots, \mathrm{d_f}_{1,0}, \mathrm{d_f}_{1,1}, \mathrm{d_f}_{0,0}\right)^T, \end{aligned} \tag{3.32}$$

where $\mathrm{c_f}_{0,0} =< v_0, \phi_{0,0} >$ and $\mathrm{d_f}_{0,0} =< w_0, \psi_{0,0} >$. The computational effort required to compute the DWT is $O(n)$ time, where n is the number of data points (Nason, 2010a). Figure 3.1 shows a pictorial example of the cascade algorithm decomposition for $m = 8$.

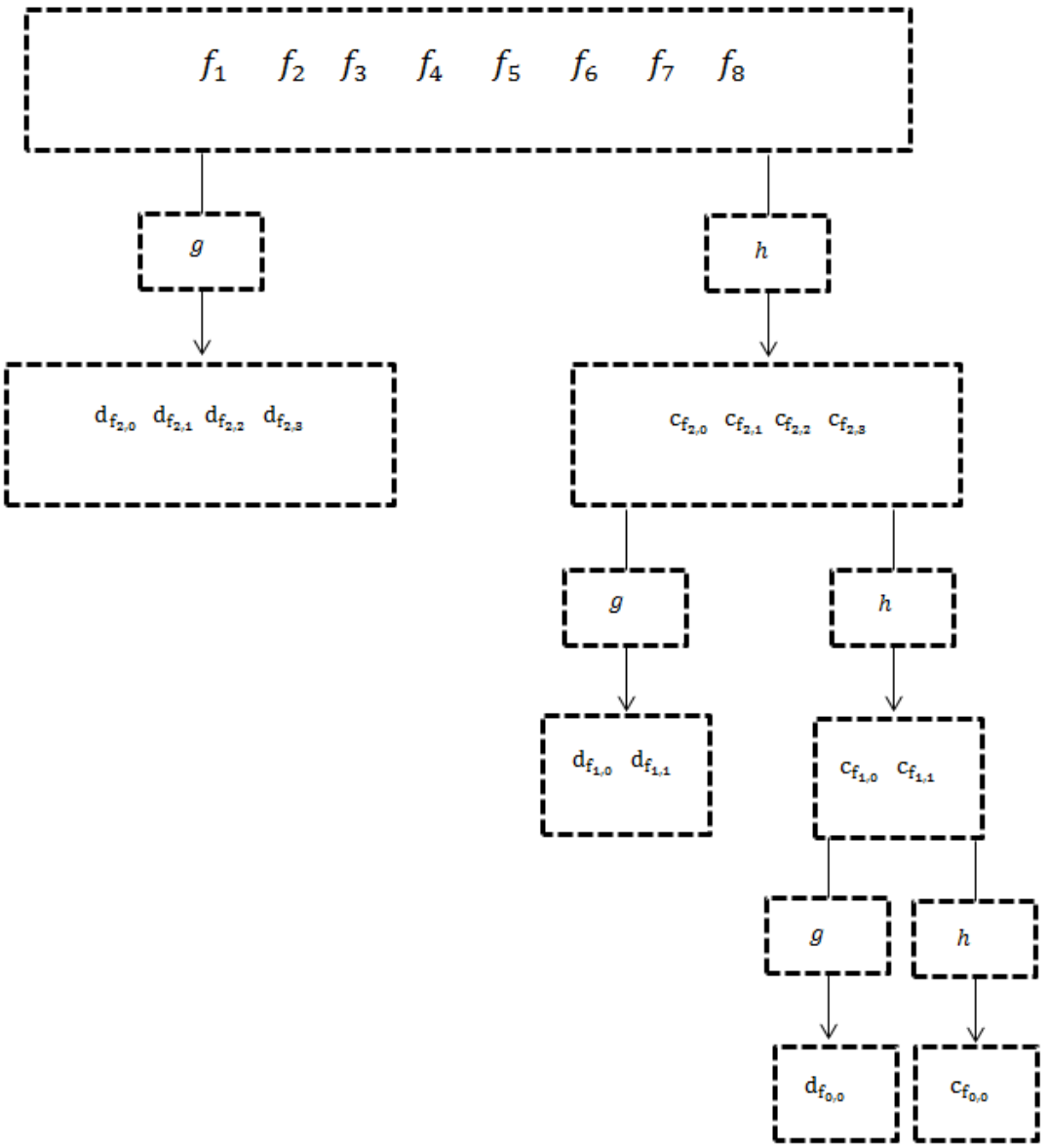


Figure 3.1: Schematic representation of the cascade algorithm decomposition for $m = 8$.

3.5 Numerical example

The purpose of this section is to provide a simple example of how the cascade algorithm works, as explained in Figure 3.1, taking a vector input, $\mathbf{f} = (1, 0, 0, 3, 2, 1, 8, 6)^T$. The Haar wavelet is applied to this data where an 8×8 matrix $\mathbf{W}$ is formed using the $\mathbf{h}$ and $\mathbf{g}$ filters of the Haar wavelet and is of the form

$$\mathbf{W} = \begin{bmatrix} \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} \end{bmatrix},$$

and produces a set of output coefficients that can be represented as

$$\begin{aligned} \mathbf{d_f} &= \mathbf{W}\mathbf{f} \\ &= \left(\mathrm{c_{f00}}, \mathrm{d_{f}}_{j-1,0}, \ldots, \mathrm{d_{f}}_{j-1,2^{j-1}-1}, \mathrm{d_{f}}_{j-2,0}, \ldots, \mathrm{d_{f}}_{j-2,2^{j-2}-1}, \ldots, \mathrm{d_{f1,0}}, \mathrm{d_{f1,1}}, \mathrm{d_{f0,0}} \right)^T \\ &= \left(\frac{21}{2\sqrt{2}}, \frac{1}{\sqrt{2}}, -\frac{3}{\sqrt{2}}, \frac{1}{\sqrt{2}}, \frac{2}{\sqrt{2}}, -\frac{2}{2}, -\frac{11}{2}, -\frac{13}{2\sqrt{2}} \right)^T. \end{aligned}$$

3.6 Inverse discrete wavelet transform

The inverse discrete wavelet transform (IDWT) is defined to recover a signal, $\mathbf{f}$, in (2.33) from the vector of the discrete wavelet coefficients, $\mathbf{d_f}$. The IDWT is computed using the matrix multiplication, as given in Section 3.5, this is

$$\mathbf{f} = \mathbf{W}^T \mathbf{d_f}.$$

Nason (2010a) showed how to obtain scaling and detail coefficients at the next finer scale as

$$\mathrm{c}_{\mathrm{f}\,j-1,l} = (\mathrm{c}_{\mathrm{f}\,j,2l} + \mathrm{c}_{\mathrm{f}\,j,2l+1})/\sqrt{2}, \tag{3.33}$$

$$\mathrm{d}_{\mathrm{f}\,j-1,l} = (\mathrm{c}_{\mathrm{f}\,j,2l} - \mathrm{c}_{\mathrm{f}\,j,2l+1})/\sqrt{2}. \tag{3.34}$$

One can obtain $\mathrm{c}_{\mathrm{f}\,j,2l}$ and $\mathrm{c}_{\mathrm{f}\,j,2l+1}$ by solving the Equations (3.33) and (3.34) to obtain the following formula

$$\mathrm{c}_{\mathrm{f}\,j,2l} = (\mathrm{c}_{\mathrm{f}\,j-1,l} + \mathrm{d}_{\mathrm{f}\,j-1,1})/\sqrt{2}, \tag{3.35}$$

$$\mathrm{d}_{\mathrm{f}\,j,2l+1} = (\mathrm{c}_{\mathrm{f}\,j-1,l} - \mathrm{d}_{\mathrm{f}\,j-1,1})/\sqrt{2}. \tag{3.36}$$

For general wavelet coefficients, Mallat (1989) showed that the inversion relation is given by

$$\mathrm{c}_{\mathrm{f}\,j,l} = \sum_n h_{l-2n}\mathrm{c}_{\mathrm{f}\,j-1,n} + \sum_n g_{l-2n}\mathrm{d}_{\mathrm{f}\,j-1,n}, \tag{3.37}$$

where **h** and **g** are again the quadrature mirror filters defined by (3.12) and (3.26), which are *exactly* the same as those used for computing the forward cascade algorithm (Nason, 2010a).

3.7 The non-decimated wavelet transform

The non-decimated (or stationary) wavelet transform (NDWT) was discussed by Pesquet *et al.* (1996) and is concerned with an extension of the standard discrete wavelet transform (Nason and Silverman, 1995). The basic idea of the non-decimated wavelet transform is to "fill in the gaps" between the coefficients in the standard wavelet transform. The NDWT is sometimes referred to as *cycle-spinning* or the maximum overlap wavelet transform within the literature. The NDWT leads to an "over-determined" or redundant representation of the original data (Nason and Silverman, 1995). However, the redundant basis provides a shift invariant denoising method, which simultaneously provide improvements in smoothness, in reconstruction and in squared-error performance (Coifman and Donoho,

1995). Moreover, the NDWT is obtained from the standard DWT by repeatedly padding out the **h** and **g** filters with alternate zeroes to double their length. This means that no decimation takes place and the non-decimated wavelet transform includes all the coefficients of the decimated wavelet transform. The general effect, depending on certain boundary conditions, is to yield an overdetermined transform with n coefficients at each of $\log_2 n$ levels of the transform. Contained within the NDWT is the result of the standard DWT for every choice of origin (Silverman, 1999).

As illustration, the Blocks test function is plotted in Figure 2.17, in Chapter 2. It can be seen that the data set has a large discontinuity. Also, the discontinuity can be identified clearly from the non-decimated wavelet coefficients, see Figure 3.2. The black spikes represent the wavelet coefficients of the true Blocks test function, the green spikes represent the wavelet coefficients of noise-free data corrupted by the matrix in (2.6), and the red spikes indicate the wavelet coefficients of the observed data with noise.

Again, the standard wavelet transform of the Blocks test function in Figure 2.17, does not describe the feature, of the Blocks test function, very clearly. This means that the discrete wavelet transform has a sampling rate that is essentially too low to give any clear picture of the data. On the other hand, the non-decimated wavelet transform, in Figure 3.2, shows the wavelet coefficients of Blocks test function where the point of discontinuity can be clearly identified. Also that, the amplitude of the oscillation, within any particular frequency level, increases and then dies, and that the region of high amplitude becomes closer to the singularity as the frequency band increases (Nason and Silverman, 1995). Note that each panel contains $32 \times \log_2 32 = 160$ wavelet coefficients, in Figure 3.2, where $n = 32$ is the length of data.

The main reason for considering the NDWT is that the standard wavelet transform sometimes displays *visual artifacts*. For example, the so-called pseudo-Gibbs phenomena can be created by wavelet denoising. The size of the artifacts are connected with the location of any discontinuities and these artifacts are all connected, in some way, with the accurate alignments between features in the original signal and basis elements. This can be

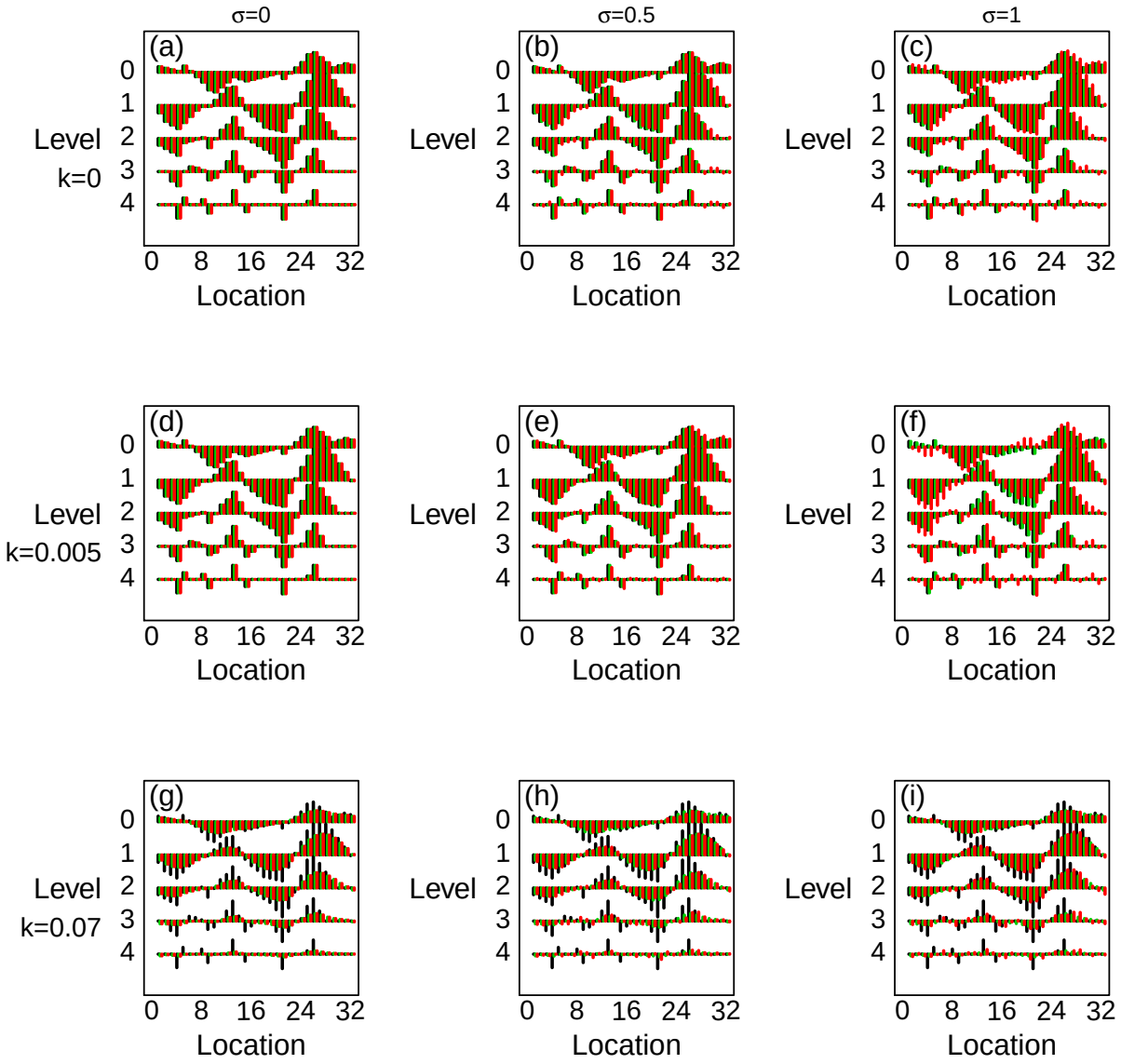


Figure 3.2: Wavelet tableaux of the Blocks test function, for $k = 0, 0.005, 0.07$ (rows), which is given in (2.6), and $\sigma = 0, 0.5, 1$ (columns), using the non-decimated wavelet transform: the black spikes represent the wavelet coefficients of the true Blocks test function; the green spikes represent the wavelet coefficients of noise-free data, and the red spikes indicate the wavelet coefficients of the observed data with noise.

attributed to the lack of the translation invariance of the wavelet basis. One approach to correcting these mis-alignments, between features in the signal and features in the basis, is to forcibly shift the signals so that features change position.

Coifman and Donoho (1995) proposed a method to deal with the lack of scale invariance, which is known as *cycle-spinning*, and acts to "average out" the translation dependence. The procedure of this method is to shift the data either right or left, for a range of different shifts. This shifted data is then denoised and the denoised data is then unshifted, producing results that are averaged to obtain a single reconstruction of the original data. This reconstruction is subject to far weaker Gibbs phenomena than the thresholding based on denoising using the standard DWT. Cycle-spinning over the range of all circulant shifts require $O(n \log_2(n))$ time.

It may well be that a given signal can be re-aligned to minimize artifacts, although there is no guarantee that this will always be the case, especially when a signal contains several discontinuities. These may also interfere with each other, that is, the best shift for one discontinuity in the signal may also be the worst shift for another discontinuity. Consequently, it becomes necessary to attempt to apply a *range of shifts*, and then *average* over several results (Coifman and Donoho, 1995). Inversion of the NDWT is more complicated than for the standard DWT because the NDWT algorithm is no longer a one-to-one transform. Hence, the DWT needs to be modified to yield the average basis inverse, providing the average of the DWT reconstructions overall choices of time origin. Both the NDWT and the average basis reconstruction are $O(n \log_2 n)$ time, where n is the number of data points. Thus, each algorithm takes the order, n, at each resolution level, and these computations take place at the order of $\log_2 n$ levels (Silverman, 1999).

3.8 Numerical example

The purpose of this section is to provide an example of the NDWT, taking a vector input, $\mathbf{f} = \{1, 1, 7, 9, 2, 8, 8, 6\}^T$. The Haar wavelet is applied to the data using the $\mathbf{h}$ and $\mathbf{g}$ filters,

defined by (3.12) and (3.26). The first set of the scaling functions can be computed

$$\begin{aligned}
\mathrm{c}_{\mathrm{f}20} &= \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right).(f_1, f_2) = \frac{2}{\sqrt{2}} & \mathrm{c}_{\mathrm{f}21} &= \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right).(f_2, f_3) = \frac{8}{\sqrt{2}} \\
\mathrm{c}_{\mathrm{f}22} &= \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right).(f_3, f_4) = \frac{16}{\sqrt{2}} & \mathrm{c}_{\mathrm{f}23} &= \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right).(f_4, f_5) = \frac{11}{\sqrt{2}} \\
\mathrm{c}_{\mathrm{f}24} &= \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right).(f_5, f_6) = \frac{10}{\sqrt{2}} & \mathrm{c}_{\mathrm{f}25} &= \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right).(f_6, f_7) = \frac{16}{\sqrt{2}} \\
\mathrm{c}_{\mathrm{f}26} &= \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right).(f_7, f_8) = \frac{14}{\sqrt{2}} & \mathrm{c}_{\mathrm{f}27} &= \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right).(f_8, f_1) = \frac{7}{\sqrt{2}}.
\end{aligned}$$

Similarly, the first set of detail coefficients can be computed

$$\begin{aligned}
\mathrm{d}_{\mathrm{f}20} &= \left(\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}\right).(f_1, f_2) = 0 & \mathrm{d}_{\mathrm{f}21} &= \left(\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}\right).(f_2, f_3) = -\frac{6}{\sqrt{2}} \\
\mathrm{d}_{\mathrm{f}22} &= \left(\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}\right).(f_3, f_4) = -\frac{2}{\sqrt{2}} & \mathrm{d}_{23} &= \left(\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}\right).(f_4, f_5) = \frac{7}{\sqrt{2}} \\
\mathrm{d}_{\mathrm{f}24} &= \left(\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}\right).(f_5, f_6) = -\frac{6}{\sqrt{2}} & \mathrm{d}_{\mathrm{f}25} &= \left(\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}\right).(f_6, f_7) = 0 \\
\mathrm{d}_{\mathrm{f}26} &= \left(\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}\right).(f_7, f_8) = \frac{2}{\sqrt{2}} & \mathrm{d}_{\mathrm{f}27} &= \left(\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}\right).(f_8, f_1) = \frac{5}{\sqrt{2}}.
\end{aligned}$$

From these, the scaling coefficients at the first level are

$$\begin{aligned}
\mathrm{c}_{\mathrm{f}10} &= \left(\frac{1}{\sqrt{2}}, 0, \frac{1}{\sqrt{2}}, 0\right).(\mathrm{c}_{\mathrm{f}20}, \mathrm{c}_{\mathrm{f}21}, \mathrm{c}_{\mathrm{f}22}, \mathrm{c}_{\mathrm{f}23}) = 9 \\
\mathrm{c}_{\mathrm{f}11} &= \left(\frac{1}{\sqrt{2}}, 0, \frac{1}{\sqrt{2}}, 0\right).(\mathrm{c}_{\mathrm{f}21}, \mathrm{c}_{\mathrm{f}22}, \mathrm{c}_{\mathrm{f}23}, \mathrm{c}_{\mathrm{f}24}) = \frac{19}{2} \\
\mathrm{c}_{\mathrm{f}12} &= \left(\frac{1}{\sqrt{2}}, 0, \frac{1}{\sqrt{2}}, 0\right).(\mathrm{c}_{\mathrm{f}22}, \mathrm{c}_{\mathrm{f}23}, \mathrm{c}_{24}, \mathrm{c}_{25}) = 13 \\
\mathrm{c}_{\mathrm{f}13} &= \left(\frac{1}{\sqrt{2}}, 0, \frac{1}{\sqrt{2}}, 0\right).(\mathrm{c}_{\mathrm{f}23}, \mathrm{c}_{\mathrm{f}24}, \mathrm{c}_{\mathrm{f}25}, \mathrm{c}_{\mathrm{f}26}) = \frac{27}{2} \\
\mathrm{c}_{\mathrm{f}14} &= \left(\frac{1}{\sqrt{2}}, 0, \frac{1}{\sqrt{2}}, 0\right).(\mathrm{c}_{\mathrm{f}24}, \mathrm{c}_{\mathrm{f}25}, \mathrm{c}_{\mathrm{f}26}, \mathrm{c}_{\mathrm{f}27}) = 12 \\
\mathrm{c}_{\mathrm{f}15} &= \left(\frac{1}{\sqrt{2}}, 0, \frac{1}{\sqrt{2}}, 0\right).(\mathrm{c}_{\mathrm{f}25}, \mathrm{c}_{\mathrm{f}26}, \mathrm{c}_{\mathrm{f}27}, \mathrm{c}_{\mathrm{f}20}) = \frac{23}{2} \\
\mathrm{c}_{\mathrm{f}16} &= \left(\frac{1}{\sqrt{2}}, 0, \frac{1}{\sqrt{2}}, 0\right).(\mathrm{c}_{\mathrm{f}26}, \mathrm{c}_{\mathrm{f}27}, \mathrm{c}_{\mathrm{f}20}, \mathrm{c}_{\mathrm{f}21}) = 8 \\
\mathrm{c}_{17} &= \left(\frac{1}{\sqrt{2}}, 0, \frac{1}{\sqrt{2}}, 0\right).(\mathrm{c}_{\mathrm{f}27}, \mathrm{c}_{\mathrm{f}20}, \mathrm{c}_{\mathrm{f}21}, \mathrm{c}_{\mathrm{f}22}) = \frac{15}{2}.
\end{aligned}$$

Similarly, from these scaling coefficients at the first level, the detail coefficients obtained

$$\begin{aligned}
\mathrm{d_{f10}} &= \left(\frac{1}{\sqrt{2}}, 0, \frac{-1}{\sqrt{2}}, 0\right).(\mathrm{c_{f20}}, \mathrm{c_{f21}}, \mathrm{c_{f22}}, \mathrm{c_{f23}}) = -7 \\
\mathrm{d_{f11}} &= \left(\frac{1}{\sqrt{2}}, 0, \frac{-1}{\sqrt{2}}, 0\right).(\mathrm{c_{f21}}, \mathrm{c_{f22}}, \mathrm{c_{f23}}, \mathrm{c_{f24}}) = -\frac{3}{2} \\
\mathrm{d_{f12}} &= \left(\frac{1}{\sqrt{2}}, 0, \frac{-1}{\sqrt{2}}, 0\right).(\mathrm{c_{f22}}, \mathrm{c_{f23}}, \mathrm{c_{f24}}, \mathrm{c_{f25}}) = 3 \\
\mathrm{d_{f13}} &= \left(\frac{1}{\sqrt{2}}, 0, \frac{-1}{\sqrt{2}}, 0\right).(\mathrm{c_{f23}}, \mathrm{c_{f24}}, \mathrm{c_{f25}}, \mathrm{c_{f26}}) = -\frac{5}{2} \\
\mathrm{d_{f14}} &= \left(\frac{1}{\sqrt{2}}, 0, \frac{-1}{\sqrt{2}}, 0\right).(\mathrm{c_{f24}}, \mathrm{c_{f25}}, \mathrm{c_{f26}}, \mathrm{c_{f27}}) = -2 \\
\mathrm{d_{f15}} &= \left(\frac{1}{\sqrt{2}}, 0, \frac{-1}{\sqrt{2}}, 0\right).(\mathrm{c_{f25}}, \mathrm{c_{f26}}, \mathrm{c_{f27}}, \mathrm{c_{f20}}) = \frac{9}{2} \\
\mathrm{d_{f16}} &= \left(\frac{1}{\sqrt{2}}, 0, \frac{-1}{\sqrt{2}}, 0\right).(\mathrm{c_{f26}}, \mathrm{c_{f27}}, \mathrm{c_{f20}}, \mathrm{c_{f21}}) = 6 \\
\mathrm{d_{f17}} &= \left(\frac{1}{\sqrt{2}}, 0, \frac{-1}{\sqrt{2}}, 0\right).(\mathrm{c_{f27}}, \mathrm{c_{f20}}, \mathrm{c_{f21}}, \mathrm{c_{f22}}) = -\frac{1}{2},
\end{aligned}$$

with, the scaling coefficients at the lowest level

$$\begin{aligned}
\mathrm{c_{f00}} &= \left(\frac{1}{\sqrt{2}}, 0, 0, 0, \frac{1}{\sqrt{2}}, 0, 0, 0\right).(\mathrm{c_{f10}}, \mathrm{c_{f11}}, \mathrm{c_{f12}}, \mathrm{c_{f13}}, \mathrm{c_{f14}}, \mathrm{c_{f15}}, \mathrm{c_{f16}}, \mathrm{c_{f17}}) = \frac{21}{\sqrt{2}} \\
\mathrm{c_{f01}} &= \left(\frac{1}{\sqrt{2}}, 0, 0, 0, \frac{1}{\sqrt{2}}, 0, 0, 0\right).(\mathrm{c_{f11}}, \mathrm{c_{f12}}, \mathrm{c_{f13}}, \mathrm{c_{f14}}, \mathrm{c_{f15}}, \mathrm{c_{f16}}, \mathrm{c_{f17}}, \mathrm{c_{f18}}) = \frac{21}{\sqrt{2}} \\
\mathrm{c_{f02}} &= \left(\frac{1}{\sqrt{2}}, 0, 0, 0, \frac{1}{\sqrt{2}}, 0, 0, 0\right).(\mathrm{c_{f12}}, \mathrm{c_{f13}}, \mathrm{c_{f14}}, \mathrm{c_{f15}}, \mathrm{c_{f16}}, \mathrm{c_{f17}}, \mathrm{c_{f17}}, \mathrm{c_{f10}}) = \frac{21}{\sqrt{2}} \\
\mathrm{c_{f03}} &= \left(\frac{1}{\sqrt{2}}, 0, 0, 0, \frac{1}{\sqrt{2}}, 0, 0, 0\right).(\mathrm{c_{f13}}, \mathrm{c_{f14}}, \mathrm{c_{f15}}, \mathrm{c_{f16}}, \mathrm{c_{f17}}, \mathrm{c_{f10}}, \mathrm{c_{f12}}, \mathrm{c_{f13}}) = \frac{21}{\sqrt{2}} \\
\mathrm{c_{f04}} &= \left(\frac{1}{\sqrt{2}}, 0, 0, 0, \frac{1}{\sqrt{2}}, 0, 0, 0\right).(\mathrm{c_{f14}}, \mathrm{c_{f15}}, \mathrm{c_{f16}}, \mathrm{c_{f17}}, \mathrm{c_{f10}}, \mathrm{c_{f11}}, \mathrm{c_{f12}}, \mathrm{c_{f13}}) = \frac{21}{\sqrt{2}} \\
\mathrm{c_{f05}} &= \left(\frac{1}{\sqrt{2}}, 0, 0, 0, \frac{1}{\sqrt{2}}, 0, 0, 0\right).(\mathrm{c_{f15}}, \mathrm{c_{f16}}, \mathrm{c_{f17}}, \mathrm{c_{f10}}, \mathrm{c_{f11}}, \mathrm{c_{f12}}, \mathrm{c_{f13}}, \mathrm{c_{f14}}) = \frac{21}{\sqrt{2}} \\
\mathrm{c_{f06}} &= \left(\frac{1}{\sqrt{2}}, 0, 0, 0, \frac{1}{\sqrt{2}}, 0, 0, 0\right).(\mathrm{c_{f16}}, \mathrm{c_{f17}}, \mathrm{c_{f10}}, \mathrm{c_{f11}}, \mathrm{c_{f12}}, \mathrm{c_{f13}}, \mathrm{c_{f14}}, \mathrm{c_{f15}}) = \frac{21}{\sqrt{2}} \\
\mathrm{c_{f07}} &= \left(\frac{1}{\sqrt{2}}, 0, 0, 0, \frac{1}{\sqrt{2}}, 0, 0, 0\right).(\mathrm{c_{f17}}, \mathrm{c_{f10}}, \mathrm{c_{f11}}, \mathrm{c_{f12}}, \mathrm{c_{f13}}, \mathrm{c_{f14}}, \mathrm{c_{f15}}, \mathrm{c_{f16}}) = \frac{21}{\sqrt{2}}.
\end{aligned}$$

Finally, the detail coefficients at the lowest level given by

$$
\begin{aligned}
\mathrm{d_{f00}} &= \left(\frac{1}{\sqrt{2}}, 0, 0, 0, \frac{1}{\sqrt{2}}, 0, 0, 0\right).(\mathrm{c_{f10}}, \mathrm{c_{f11}}, \mathrm{c_{f12}}, \mathrm{c_{f13}}, \mathrm{c_{f14}}, \mathrm{c_{f15}}, \mathrm{c_{f16}}, \mathrm{c_{f17}}) = -\frac{3}{\sqrt{2}} \\
\mathrm{d_{f01}} &= \left(\frac{1}{\sqrt{2}}, 0, 0, 0, \frac{1}{\sqrt{2}}, 0, 0, 0\right).(\mathrm{c_{f11}}, \mathrm{c_{f12}}, \mathrm{c_{f13}}, \mathrm{c_{f14}}, \mathrm{c_{f15}}, \mathrm{c_{f16}}, \mathrm{c_{f17}}, \mathrm{c_{f18}}) = -\frac{2}{\sqrt{2}} \\
\mathrm{d_{f02}} &= \left(\frac{1}{\sqrt{2}}, 0, 0, 0, \frac{1}{\sqrt{2}}, 0, 0, 0\right).(\mathrm{c_{f12}}, \mathrm{c_{f13}}, \mathrm{c_{f14}}, \mathrm{c_{f15}}, \mathrm{c_{f16}}, \mathrm{c_{f17}}, \mathrm{c_{f17}}, \mathrm{c_{f10}}) = \frac{5}{\sqrt{2}} \\
\mathrm{d_{f03}} &= \left(\frac{1}{\sqrt{2}}, 0, 0, 0, \frac{1}{\sqrt{2}}, 0, 0, 0\right).(\mathrm{c_{f13}}, \mathrm{c_{f14}}, \mathrm{c_{f15}}, \mathrm{c_{f16}}, \mathrm{c_{f17}}, \mathrm{c_{f10}}, \mathrm{c_{f12}}, \mathrm{c_{f13}}) = \frac{6}{\sqrt{2}} \\
\mathrm{d_{f04}} &= \left(\frac{1}{\sqrt{2}}, 0, 0, 0, \frac{1}{\sqrt{2}}, 0, 0, 0\right).(\mathrm{c_{f14}}, \mathrm{c_{f15}}, \mathrm{c_{f16}}, \mathrm{c_{f17}}, \mathrm{c_{f10}}, \mathrm{c_{f11}}, \mathrm{c_{f12}}, \mathrm{c_{f13}}) = \frac{3}{\sqrt{2}} \\
\mathrm{d_{f05}} &= \left(\frac{1}{\sqrt{2}}, 0, 0, 0, \frac{1}{\sqrt{2}}, 0, 0, 0\right).(\mathrm{c_{f15}}, \mathrm{c_{f16}}, \mathrm{c_{f17}}, \mathrm{c_{f10}}, \mathrm{c_{f11}}, \mathrm{c_{f12}}, \mathrm{c_{f13}}, \mathrm{c_{f14}}) = \frac{2}{\sqrt{2}} \\
\mathrm{d_{f06}} &= \left(\frac{1}{\sqrt{2}}, 0, 0, 0, \frac{1}{\sqrt{2}}, 0, 0, 0\right).(\mathrm{c_{f16}}, \mathrm{c_{f17}}, \mathrm{c_{f10}}, \mathrm{c_{f11}}, \mathrm{c_{f12}}, \mathrm{c_{f13}}, \mathrm{c_{f14}}, \mathrm{c_{f15}}) = -\frac{5}{\sqrt{2}} \\
\mathrm{d_{f07}} &= \left(\frac{1}{\sqrt{2}}, 0, 0, 0, \frac{1}{\sqrt{2}}, 0, 0, 0\right).(\mathrm{c_{f17}}, \mathrm{c_{f10}}, \mathrm{c_{f11}}, \mathrm{c_{f12}}, \mathrm{c_{f13}}, \mathrm{c_{f14}}, \mathrm{c_{f15}}, \mathrm{c_{f16}}) = -\frac{6}{\sqrt{2}}.
\end{aligned}
$$

3.9 Conclusions

This chapter gives a detailed description of the decimated and non-decimated wavelet transforms, including multi-resolution analysis and the cascade algorithm, with illustrative examples. For some of these bases the calculations are fast and for others they are slow. For example, the non-decimated wavelet transform (NDWT) basis requires $O(n \log_2 n)$ operations, whereas the standard discrete wavelet transform (DWT) basis requires only $O(n)$ operations. Comparison between the plots of the wavelet coefficients for DWT and NDWT was made. These show that the NDWT transform produces a better explanation of non-smooth functions because the NDWT is a richer family, more flexible and captures more information than the DWT when representing different types of signals. On the other hand, the DWT is easier to implement than the NDWT for estimating unknown functions.

Chapter 4

Wavelet transformation and Non-Bayesian thresholding

4.1 Overview

The chapter is organized as follows: Section 4.2 gives an introduction, Section 4.3 provides a description of wavelet packets, Section 4.4 gives a brief introduction to the methodology of complex-valued wavelet bases, whilst Sections 4.5 to 4.7 discusses the unbalanced Haar technique. Section 4.8 discusses SureBlock thresholding and Section 4.9 discusses a comparison simulation with Section 4.10 providing conclusions.

4.2 Introduction

In previous work, the methods mentioned involve term by term thresholding, which means to "kill" or "keep" the wavelet coefficients on the basis of their individual magnitudes. Hall *et al.* (1999) and Cai and Silverman (2001) provided a new thresholding rule by studying groups of wavelet coefficients.

Consider the model

$$\mathbf{y} = \mathbf{Hf} + \boldsymbol{\epsilon}, \tag{4.1}$$

where $\mathbf{H}$ is a given $n \times m$ blur matrix and $\boldsymbol{\epsilon}$ is a vector of random variables, such that $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$. Then

$$\mathbf{d_y} = \mathbf{Wy} = \mathbf{W}(\mathbf{Hf} + \boldsymbol{\epsilon}) = \mathbf{WHf} + \mathbf{W}\boldsymbol{\epsilon} = \mathbf{d_g} + \boldsymbol{\eta}, \tag{4.2}$$

where $\mathbf{d}_{\mathbf{g}_{n\times 1}}$ is a vector containing the wavelet coefficients of $\mathbf{g} = \mathbf{Hf}$, and $\mathbf{f}_{m\times 1}$ is our signal of interest. Thus, the model in (4.1) can be written equivalently as

$$\mathbf{d_y} = \mathbf{d_g} + \boldsymbol{\eta}, \tag{4.3}$$

where $\mathbf{d_y} = \mathbf{Wy}$ and $\mathbf{d_g} = \mathbf{Wg}$. The aim is to estimate the unknown function using the best method from Sections 2.14 and 2.15, IT-TO which is defined by

$$\widehat{\mathbf{f}}^{\text{IT-TO}}_{\text{Reg}_{\boldsymbol{\lambda},\Lambda}} = \mathbf{W}^T\mathrm{T}^*(\mathbf{W}\mathrm{I}_*(\mathbf{H}, \mathbf{y}, \Lambda), \boldsymbol{\lambda}),$$

where $\mathrm{I}_*(\cdot)$ is an inversion method, such as $\mathrm{I}_*(\mathbf{H}, \mathbf{y}, \Lambda) = (\mathbf{H}^T\mathbf{H} + \Lambda\mathfrak{R}^T\mathfrak{R})^{-1}\mathbf{H}^T\mathbf{y}$, with $\Lambda = \sigma^2\kappa$, the result of the inversion are transformed to wavelet coefficients followed by applying a thresholding rule, such as hard or soft. Finally, the unknown function is estimated using the inverse wavelet transform. The key point is to transform the observed data to wavelet coefficients. Three transformations will be introduced; wavelet packets, complex-valued and unbalanced Haar transforms, and block thresholding will be considered.

There is a wide range of articles considering complex wavelet bases, which includes Barber and Nason (2004). They modified the modulus of complex wavelet coefficients by a bivariate shrinkage rule leaving the phase undamaged. The authors showed that after taking the complex wavelet transform, the real and imaginary parts of the transformed noise become correlated. Also, multiple wavelet bases are used with more than one mother and father wavelet. Fryzlewicz (2007) has developed a new approach to wavelet methods, which involves using an unbalanced Haar transform, thresholding of the wavelet decomposition and then applying the inverse unbalanced Haar transform.

Early work by Hall *et al.* (1997) and Hall *et al.* (1999) used block thresholding by dividing the wavelet coefficients into groups, rather than individually, by noticing what wavelet coefficients say about the size of their nearby neighbours. The block thresholding method computes a near unbiased estimate of the sum of squares of the true wavelet coefficients in a group. Then, all of the wavelet coefficients within the group are kept or set to zero depending on the thresholding rule. There are several articles on block thresholding in the signal processing community, including Cai (1999), Abramovich *et al.* (2002), De Canditiis and Vidakovic (2004) and Cai and Zhou (2009). Cai and Zhou (2009) proposed the Bayesian block shrinkage (BBS) method and a data-driven approach to block thresholding. De Canditiis and Vidakovic (2004) defined the block thresholding by grouping wavelet coefficients at each resolution level in a block of a given size. A Bayesian model is defined on each block, by taking into account both the sparseness of the wavelet representations of a noiseless signal and the magnitude of the error affecting the sample.

4.3 The wavelet packet transform

Wavelet packets can be considerd as a generalization of wavelets and can also be expanded to produce a non-decimated version. In wavelet theory there are many such bases, and some of them are organized as *basis libraries* such as the *wavelet packet* library, which is described by Wickerhauser (1994). The wavelet packet basis is a rich family that is more flexible in representing different types of signals. This also means, however, that it provides a large number of wavelet coefficients.

In order to better explore the basis of the wavelet packet, Coifman and Wickerhauser (1992) started with the scaling ϕ and wavelet ψ functions and define $\mathrm{W}_0(t) = \phi(t)$ and $\mathrm{W}_1(t) = \psi(t)$. Then define the sequence of functions $\{\mathrm{W}_k(t)\}_{k=0}^{\infty}$ by

$$\mathrm{W}_{2n}(t) = \sqrt{2}\sum_{l} h_l \sqrt{2}\mathrm{W}_n(2t-l),$$

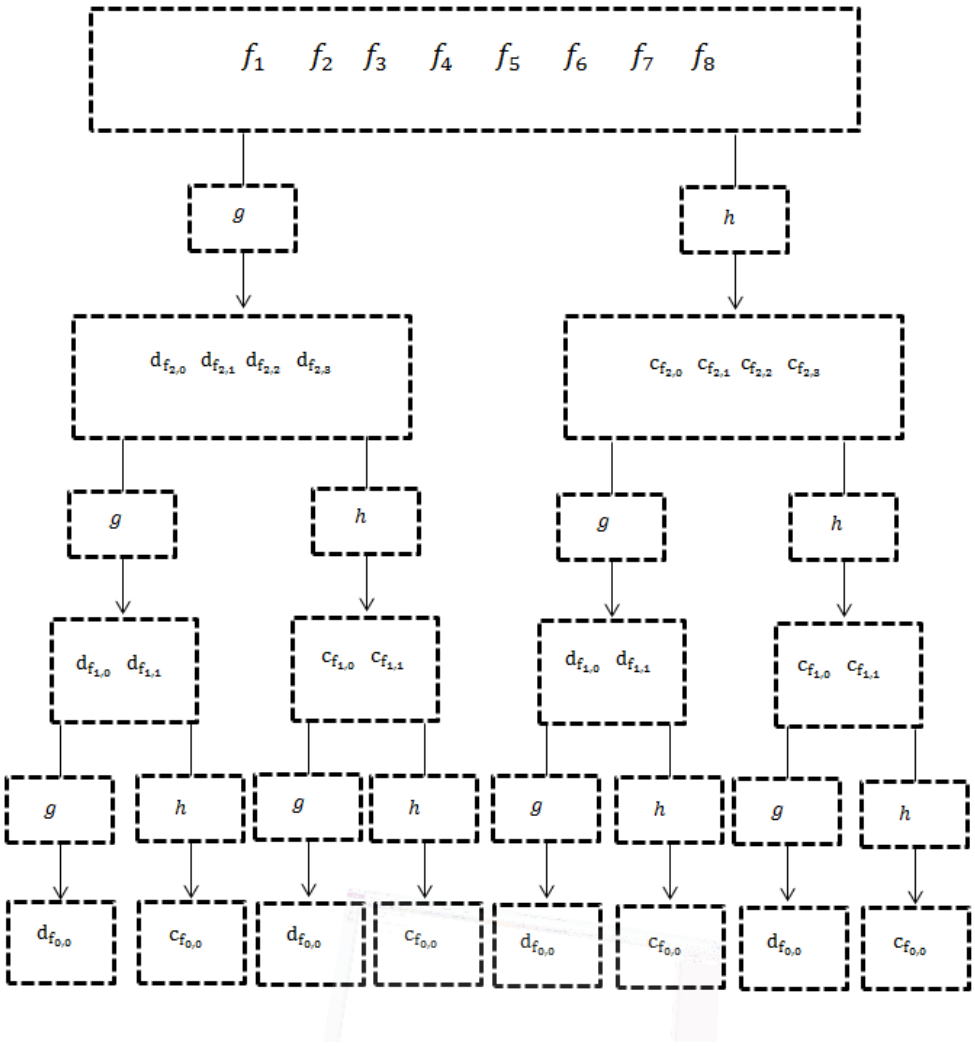


Figure 4.1: Schematic of the wavelet packet transforms on $m = 8$ points.

$$\mathrm{W}_{2n+1}(t) = \sqrt{2}\sum_{l} g_l \sqrt{2} W_n(2t - l).$$

They define a library of wavelet packets to be the collection of orthonormal bases comprised of dilated and translated versions of W_n, producing functions of the form $\mathrm{W}_n(2^j t - l)$, where $n \in \mathbb{N}$ is the number of oscillations and $j, l \in \mathbb{Z}$ are the scale and translation numbers. Wavelet packet coefficients can be computed in only $O(n \log_2 n)$ operations (Nason, 2010a). Figure 4.1 shows schematically the wavelet packet transform applied to 8 data points. The **h** and **g** filters carry out the smooth and detail operations as in the regular wavelet transform. The regular scaling function coefficients are labeled "c" and

the regular wavelet function coefficients are labeled "d" (Nason, 2010a).

4.4 Complex-valued wavelets

Shrinkage methods using Daubechies families of compactly supported real valued wavelets are explained by Daubechies (1988), where a function has compact support if it is zero outside a compact set. Lawton (1993) and Lina and Mayrand (1995) described the complex-valued Daubechies wavelets, known as $cDWs$. There are several articles on complex-valued Daubechies wavelets in the signal and image processing community, including, Lawton (1993), Lina and MacGibbon (1997), Lina (1997) and Lina *et al.* (1999). Sardy (2000), proposed an extension called "Waveshrink" to denoising signals, using the complex wavelet transform. Another article discussing complex wavelets is Clonda *et al.* (2004) who denoised images using complex-valued wavelets and a hierarchical Markov graphical model. Selesnick *et al.* (2005) proposed an alternative shrinkage method for complex wavelet known as the *dual-tree complex wavelet transform*. This approach utilises two real DWTs, the first gives the real part and the second gives the imaginary part of the transform. Barber and Nason (2004) proposed a method for denoising univariate functions using complex-valued wavelets.

Daubechies (1988) showed that when a wavelet has N vanishing moments, this means that all coefficients of any polynomial of degree N or less, will be exactly zero. Note that, when $\mathbf{f}$ is quite smooth but in some parts is interrupted by a discontinuity, then the wavelet coefficient "on the smooth part" will be small or even zero. Also, Daubechies (1988) showed that real-valued Daubechies wavelets are indexed by the number of vanishing moments. For a given N, there are 2^{N-1} solutions to the defining equations of Daubechies' wavelets. For example, in the case of $N = 3$, then there are 4 possible solutions. Two solutions are real and the other two are a complex-valued conjugate pair. Figure 4.2 shows the real and imaginary components of the wavelet coefficients for the Blocks test function sampled at $m = 32$ equally spaced points. The black spikes represent the real-valued component of the wavelet coefficients and the red spikes indicate the imaginary component

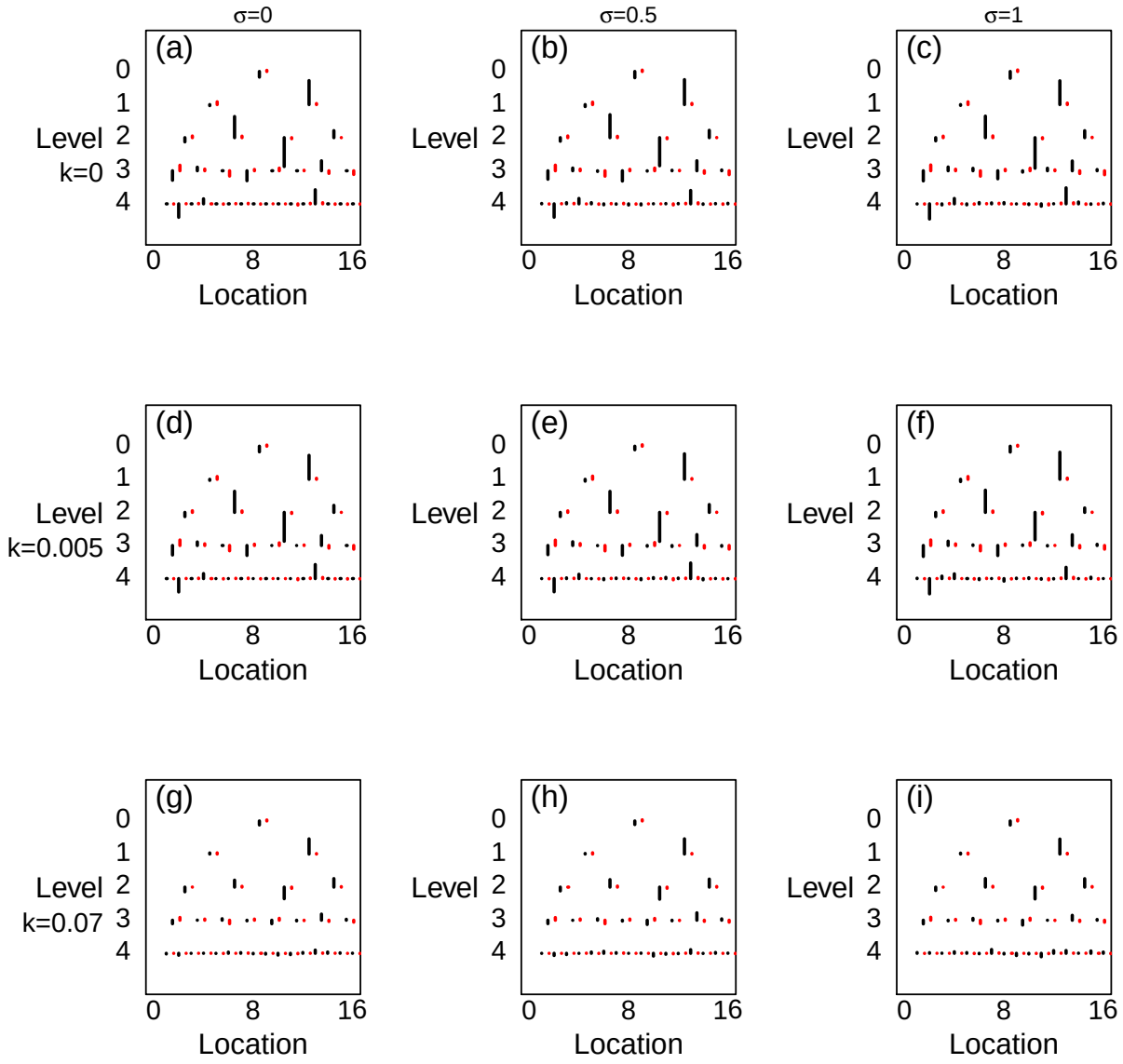


Figure 4.2: Plots of the wavelet tableaux of Blocks test function at $m = 32$ equally spaced points, for $k = 0, 0.005, 0.07$ (rows), which is given in (2.6), and $\sigma = 0, 0.5, 1$ (columns): the black spikes represent the real component of the wavelet coefficients, and the red spikes indicate the imaginary component of the wavelet coefficients (multiplied by 10^5) of the observed data.

of the wavelet coefficients (multiplied by 10^5) of the observed data with different levels of noise, $\sigma = 0, 0.5, 1$ and blur, $k = 0, 0.005, 0.07$. The wavelet tableaux shows that the wavelet coefficients are not purely real-valued. Figure 4.2 shows that as the blur increases the real-valued component of the wavelet coefficients and the imaginary component of the wavelet coefficients become closer to zero. Non-zero wavelet coefficients appeared dramatically for both in the highest resolution levels. In addition, as the level of noise increases, the number of non-zero wavelet coefficients for both in the lowest level increases.

Now, considering the inverse problem model in section 2.10, the discrete wavelet transform is given by

$$\mathbf{d_y} = \mathbf{Wy} = \mathbf{W}(\mathbf{Hf} + \boldsymbol{\epsilon}) = \mathbf{WHf} + \mathbf{W}\boldsymbol{\epsilon} = \mathbf{d_g} + \boldsymbol{\eta},$$

where $\mathbf{d}_{\mathbf{g}_{n\times 1}}$ is a vector containing the wavelet coefficients of $\mathbf{g} = \mathbf{Hf}$, $\mathbf{f}_{m\times 1}$ is our signal of interest and the DWT may represented by an $n \times n$ unitary matrix $\mathbf{W}$, constructed from $\psi(t)$, such that $\mathbf{d_y} = \mathbf{Wy}$. Barber and Nason (2004) noted that a unitary matrix is one where $\bar{\mathbf{W}}^T\mathbf{W} = \mathbf{W}\bar{\mathbf{W}}^T = \mathbf{I}_n$, the over bar $\bar{\cdot}$ denotes complex conjugation and $\mathbf{I}_n$ is the $n \times n$ identity matrix. The authors show that if the wavelet used is only real-valued then $\boldsymbol{\eta} = \mathbf{W}\boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$, because $\mathbf{W}$ is orthogonal. The individual components of $\boldsymbol{\eta}$, considered as *complex-valued* random variables, are uncorrelated. However, the authors also note that after taking the complex wavelet transform, the real and imaginary parts of the transform noise are normal real-valued random variables in their own right and can be strongly correlated. The authors demonstrate that

$$\text{cov}\{\mathbf{Re}(\boldsymbol{\eta}), \mathbf{Im}(\boldsymbol{\eta})\} = -\sigma^2\mathbf{Im}(\mathbf{WW}^T)/2 \tag{4.4}$$

$$\text{cov}\{\mathbf{Re}(\boldsymbol{\eta}), \mathbf{Re}(\boldsymbol{\eta})\} = \sigma^2\{\mathbf{I}_n + \mathbf{Re}(\mathbf{WW}^T)\}/2 \tag{4.5}$$

$$\text{cov}\{\mathbf{Im}(\boldsymbol{\eta}), \mathbf{Im}(\boldsymbol{\eta})\} = \sigma^2\{\mathbf{I}_n - \mathbf{Re}(\mathbf{WW}^T)\}/2. \tag{4.6}$$

The authors proved the above equations, that by letting $\mathbf{Re}(\boldsymbol{\eta}) = \frac{1}{2}(\boldsymbol{\eta} + \bar{\boldsymbol{\eta}})$ and $\mathbf{Im}(\boldsymbol{\eta}) =$

$\frac{1}{2i}(\boldsymbol{\eta} - \bar{\boldsymbol{\eta}})$, where $\bar{\boldsymbol{\eta}}$ denotes the complex conjugate of $\boldsymbol{\eta}$. Hence,

$$\begin{aligned} \text{var}(\boldsymbol{\eta}) &= \text{var}(\mathbf{W}\boldsymbol{\epsilon}) \\ &= \mathbf{W}\text{var}(\boldsymbol{\epsilon})\bar{\mathbf{W}}^T \\ &= \mathbf{W}\sigma^2\mathbf{I}_n\bar{\mathbf{W}}^T. \end{aligned} \tag{4.7}$$

A similar computation can then be used to find $\text{var}(\bar{\boldsymbol{\eta}})$ and to simplify the assumption $\mathbf{Re}(\boldsymbol{\eta})$ and $\mathbf{Im}(\boldsymbol{\eta})$, which are defined above

$$\mathbf{Re}(\mathbf{W}\mathbf{W}^T) = \frac{\mathbf{W}\mathbf{W}^T + \bar{\mathbf{W}}\bar{\mathbf{W}}^T}{2}, \quad \mathbf{Im}(\mathbf{W}\mathbf{W}^T) = \frac{\mathbf{W}\mathbf{W}^T - \bar{\mathbf{W}}\bar{\mathbf{W}}^T}{2i}.$$

Hence, $\mathbf{W}\bar{\mathbf{W}}^T = \bar{\mathbf{W}}\mathbf{W}^T = \mathbf{I}_n$, so

$$\begin{aligned} \text{cov}\{\mathbf{Re}(\boldsymbol{\eta}), \mathbf{Im}(\boldsymbol{\eta})\} &= \text{cov}\left\{\frac{1}{2}(\boldsymbol{\eta} + \bar{\boldsymbol{\eta}}), \frac{1}{2i}(\boldsymbol{\eta} - \bar{\boldsymbol{\eta}})\right\} \\ &= \frac{1}{4i}\Big[\text{cov}\{\boldsymbol{\eta}, \boldsymbol{\eta}\} - \text{cov}\{\boldsymbol{\eta}, \bar{\boldsymbol{\eta}}\} + \text{cov}\{\bar{\boldsymbol{\eta}}, \boldsymbol{\eta}\} - \text{cov}\{\bar{\boldsymbol{\eta}}, \bar{\boldsymbol{\eta}}\}\Big] \\ &= \frac{1}{4i}\Big[\text{var}\{\boldsymbol{\eta}\} - \text{cov}\{\boldsymbol{\eta}, \bar{\boldsymbol{\eta}}\} + \text{cov}\{\bar{\boldsymbol{\eta}}, \boldsymbol{\eta}\} - \text{var}\{\bar{\boldsymbol{\eta}}, \bar{\boldsymbol{\eta}}\}\Big] \\ &= \frac{\sigma^2}{4i}\Big[\mathbf{W}\bar{\mathbf{W}}^T - \mathbf{W}\mathbf{W}^T + \bar{\mathbf{W}}\bar{\mathbf{W}}^T - \bar{\mathbf{W}}\mathbf{W}^T\Big] \\ &= \frac{\sigma^2}{4i}\Big[\mathbf{I}_n - \mathbf{W}\mathbf{W}^T + \bar{\mathbf{W}}\bar{\mathbf{W}}^T - \mathbf{I}_n\Big] \\ &= \frac{\sigma^2}{4i}\Big[- 2i\mathbf{Im}(\mathbf{W}\mathbf{W}^T)\Big] \\ &= -\frac{\sigma^2}{2}\mathbf{Im}(\mathbf{W}\mathbf{W}^T). \end{aligned} \tag{4.8}$$

A similar calculation can be used to prove (4.5) and (4.6), see Barber and Nason (2004).

Barber and Nason (2004) mentioned that any given element, $\eta_{j,l}$, of the vector $\boldsymbol{\eta}$, has a complex normal distribution equivalent to a bivariate normal with a mean vector zero and covariance matrix $\sum_{j,l}$. As periodic transforms are being used, all the covariance matrices for a given resolution level are equal and the subscript "j" on the covariance

matrices can be omitted. Thus, the covariance matrix is given by

$$\sum_j = \begin{bmatrix} \text{cov}\{\mathbf{Re}(\boldsymbol{\eta}), \mathbf{Re}(\boldsymbol{\eta})\} & \text{cov}\{\mathbf{Re}(\boldsymbol{\eta}), \mathbf{Im}(\boldsymbol{\eta})\} \\ \text{cov}\{\mathbf{Re}(\boldsymbol{\eta}), \mathbf{Im}(\boldsymbol{\eta})\} & \text{cov}\{\mathbf{Im}(\boldsymbol{\eta}), \mathbf{Im}(\boldsymbol{\eta})\} \end{bmatrix}$$
$$= \begin{bmatrix} \sigma^2\{\mathbf{I}_n + \mathbf{Re}(\mathbf{W}_j\mathbf{W}_j^T)\}/2 & -\sigma^2\mathbf{Im}(\mathbf{W}_j\mathbf{W}_j^T)/2 \\ -\sigma^2\mathbf{Im}(\mathbf{W}_j\mathbf{W}_j^T)/2 & \sigma^2\{I_n - \mathbf{Re}(\mathbf{W}_j\mathbf{W}_j^T)\}/2 \end{bmatrix},$$

where the noise level of σ is assumed known. However, Barber and Nason (2004) estimated σ using the sum of the squared median absolute deviation of the finest level of the real part and the squared median absolute derivation of the finest level of the imaginary part.

Barber and Nason (2004) used an approach similar to the so-called "multiwavelet" scheme (DMWT), which was introduced by Downie and Silverman (1998), that has a basis with more than one mother and father wavelet. A discrete multiwavelet transform is actually similar to the DWT except that at each time-scale location there are L coefficients. For example, if $L = 2$ then the multiwavelet transform has two coefficients at each time-scale location, similar to the complex-wavelet transform. The main difference between two dimensional multiwavelets and complex-wavelets is that the two mother wavelets used in a multiwavelets transform are orthogonal, whereas the same is not true for the real and imaginary parts of the complex mother wavelets. In contrast to a multiwavelet transform, the inputs to the complex-wavelet transform are univariate and require no prefilter (Barber and Nason, 2004). Writing the complex-valued empirical wavelet coefficients, $\mathrm{d_y}^*$, as a column vector, recall that $\mathrm{d_y}^* \sim N_2(\mathrm{d_y}, \sum_j)$. For each $\mathrm{d_y}^*$, let a "thresholding statistic" be computed; $\mathrm{d_{yc}} = {\mathrm{d_y}^*}^T \sum_j^{-1} \mathrm{d_y}^*$, where if this value exceeds the threshold, $\lambda = 2\log n$, then the coefficient is retained, otherwise it is set to zero. This defines a hard-thresholding estimation rule

$$\widehat{\mathrm{T}}^{\mathrm{MH}}(\mathrm{d_{yc}}, \lambda) = \mathrm{d_{yc}}\mathbb{I}(\mathrm{d_{yc}} > \lambda),$$

where $\mathbb{I}(.)$ is the indicator function. The soft thresholding rule is also possible,

$$\widehat{\mathrm{T}}^{\mathrm{MS}}(\mathrm{d_{yc}}, \lambda) = \frac{{\mathrm{d_y}}^*}{|{\mathrm{d_y}}^*|}\max\{\mathrm{d_{yc}} - \lambda, 0\},$$

where "M" means that multiwavelet transform and thresholding is done on the $\mathrm{d_{yc}}$ scale. The estimate of $\mathbf{g}$ is then formed in the usual way, by inverting the DWT using the

estimated wavelet coefficients of the result of $\widehat{\mathrm{T}}^{\mathrm{MH}}(\mathrm{d_{yc}}, \lambda)$ or $\widehat{\mathrm{T}}^{\mathrm{MS}}(\mathrm{d_{yc}}, \lambda)$. Barber and Nason (2004) stated that, in general, the resulting signal estimate will not always be purely real-valued. The method of Barber and Nason (2004) focused solely upon recovering real-valued signals, so that any imaginary component in the estimate was considered to be an artifact and discarded. In practice, the imaginary part of the signal estimate was found to be negligible.

The approach using complex-valued wavelets was formulated, though this will not be further investigated in this thesis because it requires a wavelet basis with $N \geq 3$ vanishing moments. As the number of vanishing moments increase, so does the degree of smoothing of the corresponding wavelets basis – this leads to the magnitude of non-zero detail coefficients at the finest level increasing as well. On the other hand, the Haar wavelet has one vanishing moment, since any constant function when integrated against it will be zero, which provides detail coefficients close to zero at the finest level (Cohen and Wallace, 2012; Vidakovic, 1999).

4.5 The unbalanced Haar approach

The discrete unbalanced Haar (DUHT) method was proposed by Girardi and Sweldens (1997) and discussed by Fryzlewicz (2007). Fryzlewicz (2007) stated that the jumps in the technique do not necessarily occur in the middle of their support.

Fryzlewicz (2007) favoured the use of non-linear estimators, which are well known to offer superior theoretical and practical performance to linear estimators when the original function is uniformally smooth. There are numerous articles related to non-linear methods that produce piecewise constant reconstructions, such as those by Polzehl and Spokoiny (2000) and Davies and Kovac (2001). The former uses local averaging, where the local neighbourhood is chosen in a data-driven way, while the latter considers the problem of nonparametric regression, with emphasis on controlling the number of local extremes. Here, two methods are applied; the run method and the taut-string-multi-

resolution method. Fryzlewicz (2007) proposed an unbalanced Haar wavelet thresholding estimator with respect to any choice of an unbalanced Haar basis, and showed its mean-square consistency over a large class of function spaces. The estimator is mean-square consistent for a large class of functions, and its computational procedure is of the order of $O(n \log_2 n)$ operations. The following explanation of the DUH procedure can be found in Baek and Pipiras (2009).

The DUHT wavelet basis vectors were first studied in Girardi and Sweldens (1997) as an extended version of the classical Haar wavelet. The extension being that the break-point was permitted to occur anywhere within the support (Cho and Fryzlewicz, 2011). Fryzlewicz (2007) gave a description of the construction of the DUHT. Suppose that the domain of an observed data is indexed by $i = 1, \ldots, n$, where $n \geq 2$. Let $\psi_{0,1}$ be a vector, which is constructed from two vectors; the first is constant and positive for index $i = 1, \ldots, b_{0,1}$, provided $b_{0,1} \geq 2$; and the second is constant and negative for index $i = b_{0,1}+1, \ldots, n$, where $b_{0,1} < n$ and $n - b_{0,1} \geq 2$. The constants are chosen such that the vector, $\psi_{0,1}$, satisfies the conditions that (a) the sum of the elements of $\psi_{0,1}$, equals zero and (b) the square of the sum of the elements of $\psi_{0,1}$, equals 1. Thus, the UH function satisfies

$$\sum_t \psi_{0,1}(t) = 0 \quad \text{and} \quad \sum_t (\psi_{0,1}(t))^2 = 1.$$

The UH vector on the interval, $[s, e]$, is given by

$$\psi_{s,b,e}(t) = \left(\frac{1}{b-s+1} - \frac{1}{e-s+1}\right)^{1/2} \mathbb{I}_{s \leq t \leq b} - \left(\frac{1}{e-b} - \frac{1}{e-s+1}\right)^{1/2} \mathbb{I}_{b+1 \leq t \leq e}, \quad (4.9)$$

for $s \leq b \leq e$, where s denotes the starting-point, b denotes the break-point, e denotes the end-point and $\mathbf{t} = \{t_i : i = 1, \ldots, N\}$. The function, $\psi_{s,b,e}(t)$, generalises the usual Haar wavelet, where $e - s + 1$ is a power of 2, and b corresponds to a midpoint. Let $\mathbf{y}$ be a set with size N and

$$< \mathbf{y}, \psi_{s,b,e} >= \left(\frac{1}{b-s+1} - \frac{1}{e-s+1}\right)^{1/2} \sum_{i=s}^{b} y_i - \left(\frac{1}{e-b} - \frac{1}{e-s+1}\right)^{1/2} \sum_{i=s+1}^{e} y_i, \quad (4.10)$$

where $< \cdot, \cdot >$ is the inner product. The first break-point can be defined as

$$b_{0,1} = \arg\max_{b} < \mathbf{y}, \psi_{1,b,N} >, \tag{4.11}$$

where the first starting-point is $s_{0,1} = 1$ and the first end-point is $e_{0,1} = N$.

The process then repeats on the two vectors of the domain determined by $\psi_{0,1}$. So, a vector, $\psi_{1,1}$, is constructed and supported on $i = 1, \ldots, b_{0,1}$, with break-point, $b_{1,1}$, which is defined as

$$b_{1,1} = \arg\max_{b} < \mathbf{y}, \psi_{1,b,b_{0,1}} >, \tag{4.12}$$

where the starting-point is $s_{1,1} = 1$ and the end-point is $e_{1,1} = b_{0,1}$, and a vector, $\psi_{1,2}$, is constructed and supported on $i = b_{0,1} + 1, \ldots, n$, with break-point, $b_{1,2}$, which is defined as

$$b_{1,2} = \arg\max_{b} < \mathbf{y}, \psi_{b_{0,1}+1,b,N} >, \tag{4.13}$$

where the starting-point is $s_{1,2} = b_{0,1} + 1$ and the end-point is $e_{1,2} = N$. The procedure then continues in the same manner. Each vector, $\psi_{j,l}$, having two subvectors, $\psi_{j+1,2l-1}$ and $\psi_{j+1,2l}$. Hence, each vector, $\psi_{j,l}$, has a starting-point, a break-point and an end-point. Let $s_{j,l}$, $b_{j,l}$ and $e_{j,l}$ donate starting-point, break-point and end-point, respectively. Then, for $j \geq 0$ and $l \in \{1, \ldots, 2^j\}$, proceed recursively as follows: If $b_{j,l} - s_{j,l} \geq 1$, set $s_{j+1,2l-1} = s_{j,l}$ and $e_{j,1,2l-1} = b_{j,l}$; If $e_{j,l} - b_{j,l} \geq 2$, set $s_{j+1,2l} = b_{j,l} + 1$, $e_{j+1,2l} = e_{j,l}$. Thus, the break-point $r = 2l - 1$, or $r = 2l$, is given by

$$b_{j+1,r} = \arg\max_{b} | < \mathbf{y}, \psi_{s_{j+1,r},b,e_{j+1,r}} > |. \tag{4.14}$$

This procedure can be continued as long as possible. In particular, for fixed j, some of $s_{j,l}$, $b_{j,l}$ and $e_{j,l}$ may not be defined. Let also

$$\psi_{j,l} = \psi_{s_{j,1},b_{j,1},e_{j,1}}, \tag{4.15}$$

$$d_{j,1} = < \mathbf{y}, \psi_{j,l} > . \tag{4.16}$$

The above procedure is known as an *unbalanced Haar transform*, with a particular choice of break-points (4.14). If there is no break-point then UHT detail coefficients are set to

zero. Fryzlewicz (2007) proposed a backward stage-wise basis selection algorithm, which proceeds from the finest level to the coarsest level, attempting to concentrate as *little* power as possible at fine scales, which provides a similar impact: it concentrates the bulk of the power of the signal at the coarse level. The algorithm of the DUHT will be illustrated with calculations using the R package, `unbalhaar`, in Section 4.8. Baek and Pipiras (2009) outlined the estimation algorithm as follows:

Let $\mathbf{y}$ be a set with size n, $\mathbf{d_y}$ be the set of wavelet coefficients in the DUHT (in literature this is referred to DUHT($\mathbf{y}$)). A few of the wavelet coefficients can be disregarded in several ways. For example,

$$\mathrm{T}^{\mathrm{H}}(\mathrm{d_y}, \lambda) = \begin{cases} 0, & \text{if } |\mathrm{d_y}| \leq \lambda \\ \mathrm{d_y}, & \text{if } |\mathrm{d_y}| > \lambda, \end{cases} \tag{4.17}$$

corresponding to a hard thresholding rule. Then, the denoised vector $\widehat{\mathbf{f}}$ is obtained by taking the inverse DUHT of the coefficient, $\mathrm{T}^{\mathrm{H}}(\mathrm{d_y}, \lambda)$. The total computational load of the algorithm is of the order of $O(n \log_2 n)$ operations. This procedure can be used with other thresholding rules, such as soft, non-negative garrote, or SCAD thresholding. The method of unbalanced Haar thresholding can also be described by the following stepwise procedure:

- Use `best.unbal.haar` in the R package `unbalhaar` to transform $\mathbf{y}$.
- Each matrix is 5$\times$ number of DUHT coefficients at given a scale. Each column has a length of 5, which contains an unbalanced Haar coefficient in the following format: (1) an index of the coefficient; (2) the value of the wavelet coefficient; (3) a time-point where the corresponding DUHT vector starts; (4) the last time-point before the break-point of the DUHT vector; (5) the end-point of the DUHT vector.
- Estimate the scaling coefficient at level 0, $\mathrm{c_{y_0}}$, by calculating the average $\frac{\sum_{i=1}^{n} y_i}{n}$, where n is the length of $\mathbf{y}$.
- Use thresholding rules to thresh the wavelet coefficients in the second row for each tree.

- Estimate the function using the inverse transform of the denoised wavelet coefficients using `reconstr` in the R package `unbalhaar`.

4.6 SureBlock thresholding

The main idea of block thresholding involves specifying the length of the block, L, and the value of the threshold, λ. In particular, using the blockwise James-Stein estimator with block size L, and a threshold level chosen empirically by minimizing the SURE criterion in (2.66). Consider each possible block length, $1 \leq L \leq \sqrt{n}$ starting from $L = 1$, which means that the coefficients are thresholded individually. Also, let $m = n/L$ be the number of blocks, where n is the number of data values and L is size of the block. Let $\mathbf{d}_{\mathbf{y}b} = (\mathbf{d}_{\mathbf{y}(b-1)L+1}, \ldots, \mathbf{d}_{\mathbf{y}bL})$ represent observations in the b-th block, and $S_b^2 = ||\mathbf{d}_{\mathbf{y}b}||_2^2$ for $b = 1, 2, \ldots, m$, then

$$\mathrm{SURE}(\mathrm{d_y}, \lambda, L) = \sum_{b=1}^{m} \mathrm{SURE}(\mathrm{d_{y}}_b, \lambda, L), \tag{4.18}$$

where

$$\mathrm{SURE}(\mathrm{d_{y}}_b, \lambda, L) = L + \frac{\lambda^2 - 2\lambda(L-2)}{S_b^2}\mathbb{I}(S_b^2 > \lambda) + (S_b^2 - 2L)\mathbb{I}(S_b^2 \leq \lambda), \tag{4.19}$$

where $\mathbb{I}(.)$ is the indicator function. From this, (λ^*, L^*) represent the minimizer of SURE, defined as

$$(\lambda^*, L^*) = \underset{\max(L-2,0)\leq\lambda\leq\lambda^F, 1\leq L\leq n^v}{\arg\min} \mathrm{SURE}(\mathrm{d_y}, \lambda, L),$$

for a fixed $0 \leq v < 1$. Then, the SureBlock thresholding rule, if $C_{2^J} > \gamma_{2^J}$, is given by

$$\mathrm{T}(\mathrm{d_{y}}_b, \lambda^*, L^*) = \left(1 - \frac{\lambda^*}{S_b^2}\right)\mathrm{d_{y}}_b, \tag{4.20}$$

and if $C_{2^J} \leq \gamma_{2^J}$

$$\mathrm{T}(\mathrm{d_y}) = \left(1 - \frac{2\log_2 2^J}{(\mathrm{d_y})^2}\right)_+ \mathrm{d_y}, \tag{4.21}$$

where $\lambda \geq 0$ is the threshold and $a_+ = \max(a, 0)$. The hybrid method works as follows: Set

$$\lambda^F = 2L \log 2^J, \quad C_{2^J} = \frac{1}{2^J} \sum_{i=1}^{2^J} (\mathrm{d}_{\mathrm{y}_i} - 1), \quad \gamma_{2^J} = \frac{1}{\sqrt{2^J}} (J)^{3/2}.$$

This is called the "SureBlock" estimator and if $C_{2^J} \leq \gamma_{2^J}$, the estimator becomes similar to the block James-Stein estimator, with block size $L = 1$. In this case, the estimator is also called the non-negative garrote estimator. The procedure involves the following steps:

- Transform the data to the wavelet domain using the DWT.
- For each resolution level j, select the block size L_j^* and threshold level, λ_j^*, using

$$(\lambda^*, L^*) = \underset{\max(L-2,0) \leq \lambda \leq \lambda^F, 1 \leq L \leq n^v}{\arg\min} \mathrm{SURE}\left(\frac{\sqrt{2^J}}{\sigma} \mathrm{d}_{\mathrm{y}j}, \lambda, L \right),$$

  where $\mathrm{d}_{\mathrm{y}j}$ is the empirical wavelet coefficient vector, at resolution level j.

- The SureBlock estimator in (4.20) and (4.21) is used,

$$\frac{\sigma}{\sqrt{2^J}} \mathrm{T}\left(\frac{\sqrt{2^J}}{\sigma} \mathrm{d}_{\mathrm{y}b}, \lambda^*, L^* \right) \quad \text{or} \quad \frac{\sigma}{\sqrt{2^J}} \mathrm{T}\left(\frac{\sqrt{2^J}}{\sigma} \mathrm{d}_{\mathrm{y}} \right),$$

  where 2^J is the length of the data.

- The function at the sample points is estimated by the inverse transform of the denoised wavelet coefficients.

The value of the variance of the noise σ^2 is assumed to be unknown and computed using the estimator given in (2.62) and (2.63). However, instead it can be estimated using the maximum likelihood estimator for σ^2 from the data (Cai and Zhou, 2009).

4.7 Comparison simulation

The purpose of this section is to evaluate and investigate whether the standard discrete wavelet transform (DWT), the non-decimated wavelet transform (NDWT), the wavelet

packet transforms (WPT), and discrete unbalanced Haar (DUHT) method, are suitable for estimating an unknown vector $\mathbf{f}$.

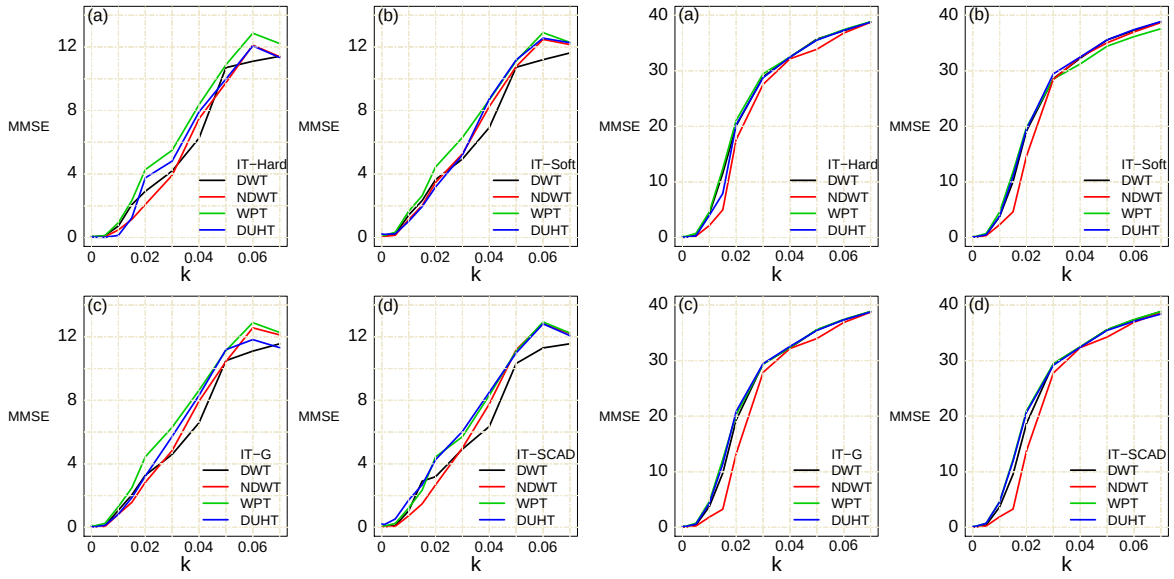
The simulated data sets consist of the standard test signals, at $m = 128$ equally spaced points, Blocks and Bumps (Donoho and Johnstone, 1994; Nason and Silverman, 1994), multiplied by different values of blur matrix, which is given in (2.6). The signal was corrupted by independent Gaussian noise, with mean zero, and variance taken as 0.5 at $m = 128$ equally spaced points. Also, thresholding was applied below level 3, the IT-TO method was used, as the best method from Chapter 2, and the value of the parameter λ is estimated at each level using the MMSE approach, as described in Section 2.13. Moreover, the first-order method in Section 2.6 is included, $\mathbf{f}$. The aim is to estimate an unknown vector $\mathbf{f}$ using the IT-TO, which is defined by

$$\widehat{\mathbf{f}}^{\text{IT-TO}}_{\text{Reg}_{\boldsymbol{\lambda},\Lambda}} = \mathbf{W}^T \mathrm{T}^*(\mathbf{W}\mathrm{I}_*(\mathbf{H}, \mathbf{y}, \Lambda), \boldsymbol{\lambda}),$$

where $\mathrm{I}_*(\cdot)$ is the result of an inversion method, such that $\mathrm{I}_*(\mathbf{H}, \mathbf{y}, \Lambda) = (\mathbf{H}^T\mathbf{H} + \Lambda\mathfrak{R}^T\mathfrak{R})^{-1}\mathbf{H}^T\mathbf{y}$, $\Lambda^* = \sigma^2\kappa$, $\mathfrak{R}$ was defined in Section 2.6 and $\mathbf{W}$ was defined in Section 2.9. The result of inversion are transformed to wavelet coefficients and then a thresholding rule is applied such as hard or soft. The total number of replications is equal to 6000.

Figure 4.3 shows the plots of MMSE, which is described in Section 2.13, using different wavelet transformations with different thresholding rules. In general, the NDWT and DUHT transform provide smaller MSE than the DWT and WPT transforms. More precisely, Figure 4.3 (i) shows the plots of MMSE for the DWT, NDWT, WPT, DUHT with different thresholding rules for estimating Blocks using the IT-TO method. The DUHT and NDWT improve the MSE. Similarly, Figure 4.3 (ii) shows the plots of MMSE for DWT, NDWT, WPT, DUHT with different thresholding rules for estimating Bumps using the IT-TO method. The NDWT improves MSE, which provides a smaller MSE than the DWT, WPT and DUHT.

Figure 4.4 shows the plots of the reconstruction using different bases, where the observations come from the Blocks test function at $m = 128$ equally spaced points, corrupted by the level of noise, $\sigma = 0.5$, and multiplied by the level of blur, which is given in



(i) Blocks (ii) Bumps

Figure 4.3: Simulated Bumps and Blocks test function: plots of minimum MSE results, which is described in Section 2.13, with the first-order method for estimating the original Blocks and Bumps test function with different values of blur.

(2.6), $k = 0.005$. No thresholding was done below level 3, the IT-TO method and hard thresholding rule were used.

The results are summarized in Tables 4.1 and 4.2, where bold numbers indicate the smallest MSE result for the Blocks test function, where the level of noise equals 0.5 and the level of blur equals $k = 0.005$.

The result can be summarized as follows: the DUHT method provides sharp edges and flat topped reconstructions, when MMSE equals 0.01. There is also a slight difference between the DWT and the NDWT for estimating unknown vector $\mathbf{f}$. Further, the NDWT provides a smaller MSE than the DWT.

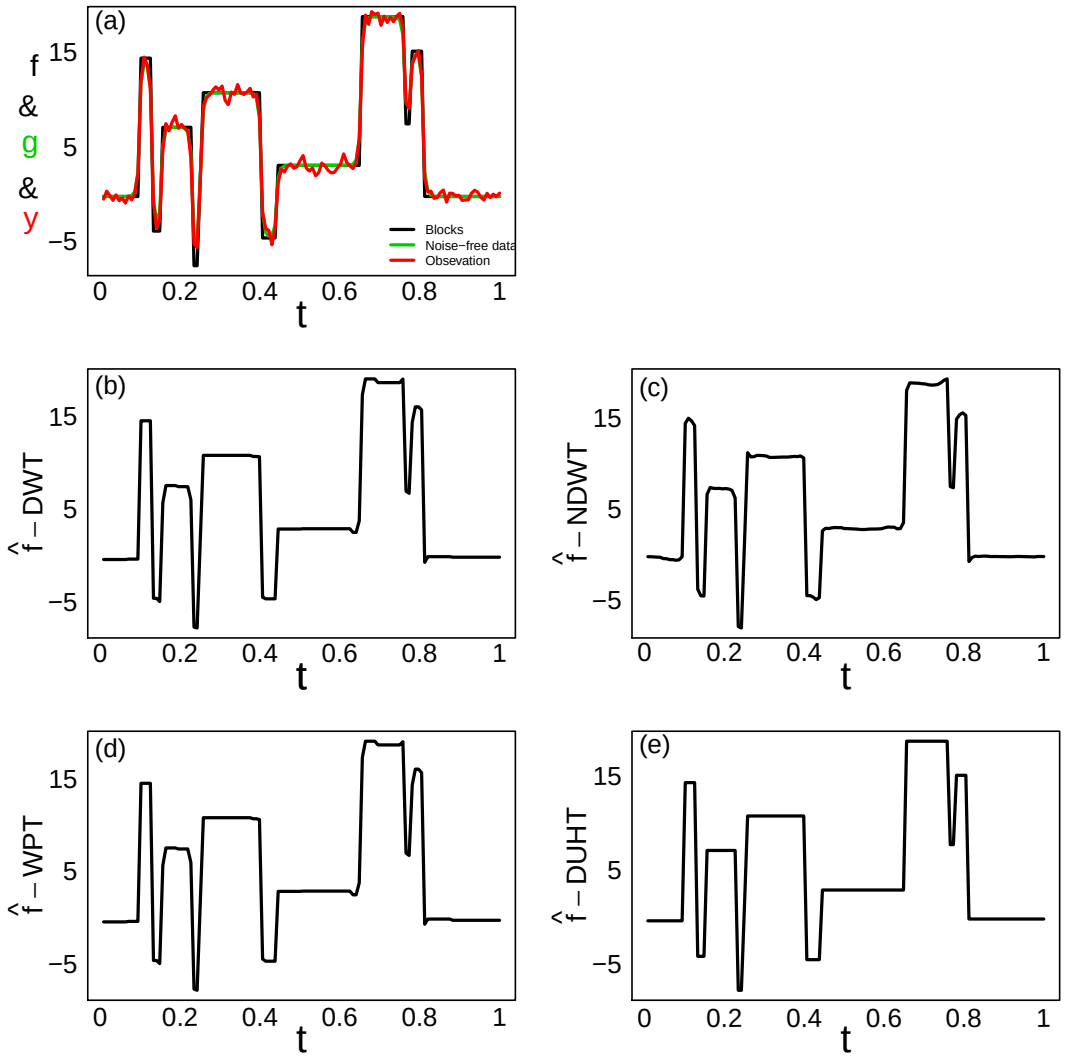


Figure 4.4: Plots of the reconstructions of hard thresholding with different transforms: (a) black line is made from Blocks test function at $m = 128$ equally spaced points, green line is made by multiplying the blocks by the blur, which is given in (2.6), $k = 0.005$ and red line shows the observed data with noise $\sigma = 0.5$, the other panels show reconstruction using; (b) DWT; (c) NDWT; (d) WPT; and (e) DUHT, where the parameters $\boldsymbol{\theta} = \{\Lambda, \boldsymbol{\lambda}\}$ are taken from Tables 4.1 and 4.2.

	MMSE	$\widehat{\Lambda}_{\text{MMSE}}$	$\widehat{\lambda}_{\text{MMSE}}$			
			Level 3	Level 4	Level 5	Level 6
DWT	0.107	0.002	2.7431	1.6065	1.8422	3.9842
NDWT	**0.09**	0.003	1.3011	3.6458	1.8356	2.5558
WPT	0.119	0.003	1.5769	6.5079	1.9297	3.8277

Table 4.1: The results of minimum MSE for different transforms with the hard thresholding rule to estimate parameters Λ and $\boldsymbol{\lambda}$, where λ is estimated for each resolution level, $j = 3, 4, 5, 6$.

	MMSE	$\widehat{\Lambda}_{\text{MMSE}}$	$\widehat{\lambda}_{\text{MMSE}}$						
			Matrix 1	Matrix 2	Matrix 3	Matrix 4	Matrix 5	Matrix 6	Matrix 7
			4.7167	5.5578	1.9846	3.1861	2.1646	4.1942	3.640
DUHT	**0.014**	0.017	Matrix 8	Matrix 9	Matrix 10	Matrix 11	Matrix 12	Matrix 13	Matrix 14
			4.4360	2.8838	2.9616	1.4343	1.3420	2.0731	2.2791

Table 4.2: The results of minimum MSE for the DUHT transform with the hard thresholding rule to estimate parameters Λ and $\boldsymbol{\lambda}$, where λ is estimated for each resolution matrix.

4.8 Conclusions

This chapter gives a detailed description of the wavelet packet transforms (WPT), complex-valued wavelets and the unbalanced Haar technique (DUHT), including non-Bayesian methods, such as SureBlock thresholding, with illustrative examples. Extensive simulation studies use the MMSE, which is described in Chapter 2, to study the standard discrete wavelet transform (DWT), the non-decimated wavelet transform (NDWT), WPT, DUHT, with different thresholding rules such as, hard, soft, non-negative garrote (G), and the smoothly clipped absolute deviation (SCAD) to estimate $\widehat{\Lambda}_{\text{MMSE}}$, and $\widehat{\lambda}_{\text{MMSE}}$. It can be concluded that for small noise, $\sigma = 0.5$, and small level of blur $k < 0.01$, the DUHT and NDWT algorithms work well applying to the Blocks test function.

Chapter 5

Bayesian thresholding using non-mixture priors

5.1 Overview

In this chapter, Section 5.2 gives an introduction to non-mixture priors. Section 5.3 applies the double Weibull distribution as a wavelet coefficient prior, then Section 5.4 is about the Gaussian distribution, Sections 5.5 and 5.6 consider the Laplace and elastic-net distributions as priors. Section 5.7 gives the larger posterior mode for different models and finally, Section 5.8 gives conclusions.

5.2 Introduction to non-mixture priors

Over the last two decades shrinkage techniques have been found to be an efficient tool, with several approaches being proposed to estimate an underlying function in the wavelet domain from a noisy sample. It is well known that maximum likelihood (ML) often does poorly in terms of both prediction and interpretation due to the maximum occurring when $\mathbf{y} = \mathbf{g}$. Penalization techniques have been proposed to improve ML, such as ridge regres-

sion and Lasso. In order to produce a reconstruction in the presence of noise, different prior models can be used to describe the wavelet coefficients of the underlying function. Empirical distributions of detail wavelet coefficients for most signals encountered in practical application are notably centered around and peaked at zero (Mallat, 1989). There are several available priors, such as Laplace, also known as the double exponential ($\mathcal{DE}$), Gaussian (N), elastic-net and double Weibull ($\mathcal{DW}$) distributions. Bayesian approaches can be constructed to mimic thresholding rules, where the large coefficients are slightly shrunk and the small coefficients are heavily shrunk (Ruggeri and Vidakovic, 2005). Reviews on early Bayesian approaches can be found in Vidakovic (1998a), Vidakovic (1998b) and Clyde and George (1999). The idea was developed substantially by Abramovich *et al.* (2000) and Ruggeri and Vidakovic (2005).

Considering the model

$$\mathbf{y} = \mathbf{Hf} + \boldsymbol{\epsilon}, \tag{5.1}$$

where $\mathbf{H}$ is a given $n \times m$ blur matrix, $\mathbf{y}_{n\times 1}$ and $\boldsymbol{\epsilon}$ is a vector of random variables, such that $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$ then

$$\mathbf{d_y} = \mathbf{Wy} = \mathbf{W}(\mathbf{Hf} + \boldsymbol{\epsilon}) = \mathbf{WHf} + \mathbf{W}\boldsymbol{\epsilon} = \mathbf{d_g} + \boldsymbol{\eta}, \tag{5.2}$$

where $\mathbf{d_g}_{n\times 1}$ is a vector of wavelet coefficients containing the wavelet coefficients of $\mathbf{g} = \mathbf{Hf}$, and $\mathbf{f}_{m\times 1}$ is our signal of interest. Thus, the model in (5.1) can be written equivalently as

$$\mathbf{d_y} = \mathbf{d_g} + \boldsymbol{\eta}, \tag{5.3}$$

where $\mathbf{d_y} = \mathbf{Wy}$ and $\mathbf{d_g} = \mathbf{Wg}$.

The procedure of estimating $\mathbf{d_g}$ from $\mathbf{d_y}$ is now considered. The posterior, including a prior distribution on the wavelet coefficient $\mathbf{d_g}$, is given by

$$\mathrm{p}(\mathbf{d_g}|\mathbf{d_y}) = \frac{\mathrm{p}(\mathbf{d_y}|\mathbf{d_g})\mathrm{p}(\mathbf{d_g})}{\mathrm{p}(\mathbf{d_y})}, \tag{5.4}$$

where $\mathbf{d_g}$ is the vector of model parameters, $\mathrm{p}(\mathbf{d_y}|\mathbf{d_g})$ is the likelihood function, and $\mathrm{p}(\mathbf{d_g})$ is the prior distribution. The posterior can be written as

$$\mathrm{p}(\mathbf{d_g}|\mathbf{d_y}) \propto \mathrm{p}(\mathbf{d_y}|\mathbf{d_g})\mathrm{p}(\mathbf{d_g}),$$

because the normalising constant has no information about the unknown parameters.

In this chapter, the best method in Chapter 2, which was the IT-TO method, is used, the posterior estimate, $\mathbf{d_f}$, is computed and then the unknown is estimated by

$$\begin{aligned}\widehat{\mathbf{f}}^{\text{IT-TO}}_{\text{Reg}_{\boldsymbol{\kappa},\Lambda}} &= \mathbf{W}^T\mathrm{p}(\mathbf{d_f}|\widehat{\mathbf{d}}_\mathbf{g}) \\ &= \mathbf{W}^T\mathrm{p}(\mathbf{d_f}|(\mathbf{H}^T\mathbf{H} + \Lambda\mathfrak{R}^T\mathfrak{R})^{-1}\mathbf{H}^T\mathbf{y}),\end{aligned} \tag{5.5}$$

where $\widehat{\mathbf{d}}_\mathbf{g} = (\mathbf{H}^T\mathbf{H} + \Lambda\mathfrak{R}^T\mathfrak{R})^{-1}\mathbf{H}^T\mathbf{y}$, where $\Lambda = \sigma^2\kappa$, $\mathfrak{R}$ was defined in Section 2.6 and $\mathbf{W}$ was defined in Section 2.9.

There are several papers on Bayesian wavelet estimation in the signal and image processing community. These papers usually specify a non-mixture prior on the wavelet coefficients and provide a Bayes estimator. The posterior mode is a popular choice, which is used for example by Huerta (2005) who used a multivariate Gaussian distribution, with vector mean, 0, and a covariance matrix, $\tau^2\sum$, to model wavelet coefficients. Cutillo *et al.* (2008) also proposed a hierarchical model with a Gaussian distribution as a prior to model wavelet coefficients.

The double Weibull distribution is discussed by Balakrishnan and Kocherlakota (1985), but it was Reményi and Vidakovic (2015) who first applied the double Weibull as a prior. They simulated from standard test functions and compared it with numerous existing methods. They showed that the double Weibull gives excellent results, even compared to different methods that use mixture priors (these will be introduced in Chapter 6). The reason behind using the double Weibull is that, in practice, the distribution for the wavelet coefficients has heavier tails than the Gaussian distribution.

The early work of Laplace has received some attention in recent literature, for example by Vidakovic and Ruggeri (2001), Meinshausen and Bühlmann (2006), and Zhao *et al.*

(2012), who explored a wavelet-based procedure with a Laplace penalty in functional linear regression. The reason for choosing Laplace is that this distribution has a heavier tails and is more peaked than the Gaussian density. A new method is proposed using an elastic-net distribution as a prior, which was first applied as a penalty (Zou and Hastie, 2005). The reason for introducing this prior is to balance the limitations of the Laplace and Gaussian distributions.

In this chapter, the aim is to test a simple model, and propose that a carefully selected non-mixture prior can compete with the performance of more complex mixture priors. The best non-mixture prior will then be compared with mixture models, this comparison will be discussed in Chapter 6.

5.3 The Double Weibull distribution

Cutillo *et al.* (2008) considered thresholding rules through the "larger posterior mode" principle. The main idea is to pick the mode of the posterior, which is the larger in absolute value. This approach is also discussed by Reményi and Vidakovic (2015), where two estimators are applied; the first is the posterior mean, which is a common choice in Bayesian methods, and the second is the larger posterior mode. Assuming an additive Gaussian error model, then the conditional distribution of the data given the truth, that is, the distribution for a single wavelet coefficient $\mathrm{d_y}$ given the corresponding wavelet coefficient $\mathrm{d_g}$, is given by

$$\mathrm{p}(\mathrm{d_y}|\mathrm{d_g},\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(\mathrm{d_y}-\mathrm{d_g})^2\right\}, \quad \mathrm{d_y},\mathrm{d_g}\in\mathbb{R};\sigma>0, \tag{5.6}$$

where σ^2 is the noise variance. A Double Weibull distribution is assumed as the prior on the wavelet coefficients $\mathrm{d_g}$, and is defined as

$$\mathrm{p}(\mathrm{d_g}) = \frac{c}{2\mathrm{b}}|\mathrm{d_g}|^{c-1}\exp\left\{-\frac{1}{\mathrm{b}}|\mathrm{d_g}|^{c}\right\}, \quad \mathrm{d_g}\in\mathbb{R};c>0,\mathrm{b}>0, \tag{5.7}$$

where b and c are the scale and shape parameters. The parameters b, c and σ^2 are assumed to be known; Reményi and Vidakovic (2015) suggested the value of c takes $0 < c < 1$,

because the double Weibull density approaches infinity as $|\mathrm{d_g}|$ approaches zero, which is in agreement with the summary of Mallat (1989), regarding the shape of the empirical distribution for wavelet coefficients. According to Balakrishnan and Kocherlakota (1985) a normal distribution as likelihood and double Weibull distribution as prior is integrable and finite.

The posterior mean (PM) of the single wavelet coefficient $\mathrm{d_g}$ given $\mathrm{d_y}$, can be written generally as

$$\mathrm{PM}(\mathrm{z}|\mathrm{d_y}) = \frac{\int_{-\infty}^{\infty} \mathrm{z}^{1/c} \exp\left\{ -\frac{1}{\mathrm{b}}|\mathrm{z}| \right\} \exp\left\{ -\frac{1}{2\sigma^2}(\mathrm{d_y} - \mathrm{z}^{1/c})^2 \right\} d\mathrm{z}}{\int_{-\infty}^{\infty} \exp\left\{ -\frac{1}{\mathrm{b}}|\mathrm{z}| \right\} \exp\left\{ -\frac{1}{2\sigma^2}(\mathrm{d_y} - \mathrm{z}^{1/c})^2 \right\} d\mathrm{z}}, \tag{5.8}$$

where $\mathrm{z} = |\mathrm{d_g}|^c$ (Reményi and Vidakovic, 2015). This is the so-called double Weibull wavelet shrinkage (DWWS) estimator. The posterior distribution for the wavelet coefficients $\mathrm{d_g}$ given $\mathrm{d_y}$, is

$$\mathrm{p}(\mathrm{d_g}|\mathrm{d_y}) \propto |\mathrm{d_g}|^{c-1} \exp\left\{ -\frac{1}{\mathrm{b}}|\mathrm{d_g}|^c \right\} \exp\left\{ -\frac{1}{2\sigma^2}(\mathrm{d_y} - \mathrm{d_g})^2 \right\}. \tag{5.9}$$

Figure 5.1 shows the posterior, defined in (5.9), for different observed wavelet coefficients, in particular for the wavelet coefficients $\mathbf{d_y} = (-3, -2, -1, 1, 2, 3)$. Clearly, the posterior is not symmetrical. If $|\mathrm{d_y}| \leq \sigma^2$, the posterior is unimodal with an (infinite) mode at zero. For $|\mathrm{d_y}| > \sigma^2$, the posterior is bimodal with a non-zero mode, which has the same sign as the coefficient $\mathrm{d_y}$ (Reményi and Vidakovic, 2015).

Now, the log-posterior is given by

$$\log \mathrm{p}(\mathrm{d_g}|\mathrm{d_y}) \propto -\frac{1}{2\sigma^2}(\mathrm{d_y} - \mathrm{d_g})^2 + (c-1)\log|\mathrm{d_g}| - \frac{1}{\mathrm{b}}|\mathrm{d_g}|^c, \tag{5.10}$$

with the maximum given as the solution of

$$-\frac{1}{\sigma^2}{\mathrm{d_g}}^2 + \frac{1}{\sigma^2}\mathrm{d_y}\mathrm{d_g} - \frac{c}{\mathrm{b}}|\mathrm{d_g}|^c + c - 1 = 0. \tag{5.11}$$

Reményi and Vidakovic (2015) gave a numerical algorithm to compute the larger posterior mode for the double Weibull wavelet shrinkage (DWWS-LPM)

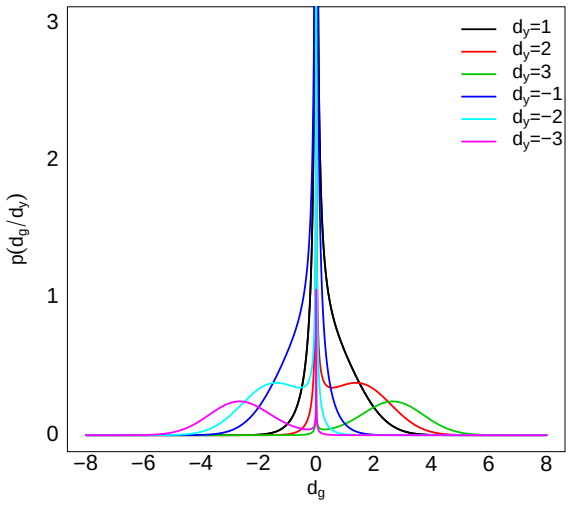


Figure 5.1: Posterior distribution in (5.9) for $c = 1/3$, $\sigma = 1$, and $\mathbf{d_y} = (-3, -2, -1, 1, 2, 3)$.

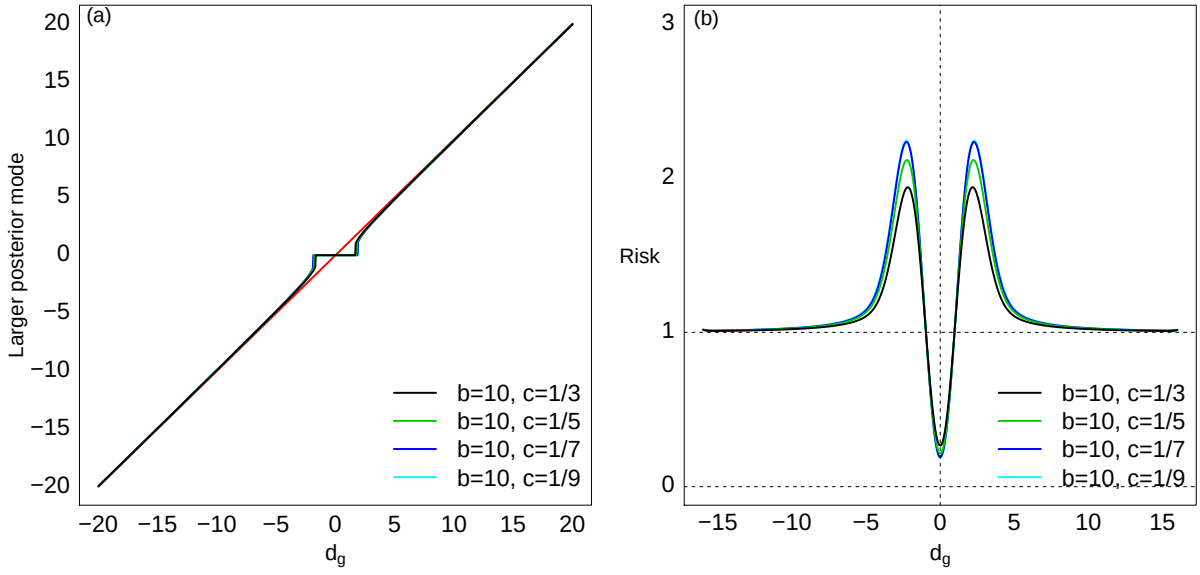


Figure 5.2: Plots of the shape of the larger posterior mode (a), Equation (5.14), and the exact risk (b), for Equation (5.14), using different values of the constant, c, with $\sigma = 1$ and b = 10.

- Let $\mathrm{z} = |\mathrm{d_g}|^c$ and for the fixed parameter $0 < c < 1$, Equation (5.11) can be modified. Thus, the solution is equivalent to the solution of a polynomial equation of the order $\frac{2}{c}$.
- Find the roots of the equation $-\frac{1}{\sigma^2}\mathrm{z}^{2/c} + \frac{1}{\sigma^2}|\mathrm{d_y}|\mathrm{z}^{1/c} - \frac{c}{\mathrm{b}}\mathrm{z} + c - 1 = 0$.
- Let the real roots be z^r.
- Let the complex roots be z^c.
- If all the roots are complex, $\mathrm{T}(\mathrm{d_y}) = 0$
- If there is more than one real root, take the largest.
- If real roots exist, $\mathrm{T}(\mathrm{d_y}) = \mathrm{sign}(\mathrm{d_y})[\max(\mathrm{z}^r)]^{1/c}$.

The larger posterior mode depends on three parameters, which have to be specified σ^2, c and b_j, where b_j is estimated for each level. The approach for estimating these parameters can be summarized as

- The prior parameter σ^2 represents the variance of the noise, which can be estimated using (2.62) or (2.63).
- The second parameter is the scale parameter, b_j, which represents the variance of the prior distribution at resolution level j. Since the model assumes independence of signal and error parts, we have

$$\widehat{\sigma}^2_{\mathbf{d_y}_j} = \left[\widehat{\mathrm{b}}_j^2 \Gamma\left(1 + \frac{2}{c}\right)\right]^{1/c} + \widehat{\sigma}^2, \tag{5.12}$$

  where $\widehat{\sigma}^2_{\mathbf{d_y}_j}$ is the variance of the wavelet coefficients, $\mathbf{d_y}$, at the j^{th} level and $\widehat{\sigma}^2$ is the variance of noise. Therefore, a reasonable estimator for b_j, is given by

$$\widehat{\mathrm{b}}_j = \max\left[0, \frac{\widehat{\sigma}^2_{\mathbf{d_y}_j} - \widehat{\sigma}^2}{\Gamma\left(1 + \frac{2}{c}\right)}\right]^{\frac{c}{2}}, \quad j_0 < j < J - 1. \tag{5.13}$$

  where $j_0 \in \mathbb{Z}$ and Antoniadis *et al.* (2001) suggested that $j_0 = 3$. In the case $\widehat{\sigma}^2_{\mathbf{d_y}_j} < \widehat{\sigma}^2$, the parameter $\widehat{\mathrm{b}}_j$ can be set to zero.

- The parameter c controls the shape of the prior distribution. Reményi and Vidakovic (2015) suggested the value of c is chosen from the interval $0 < c < 1$. Then the double Weibull density approaches infinity as $|\mathrm{d_g}|$ approaches zero. Figure 5.2 (b) displays the exact risk for different values of c, the prior variance for each resolution level was $\mathrm{b} = 10$, and variance of noise was $\sigma^2 = 1$. It is apparent that the area under the curve is small as c decreases. It can also be seen that when $c = 1/3$ and $c = 1/5$ the risk is smaller than $c = 1/9$. Reményi and Vidakovic (2015) suggested that $c = 1/3$. It provides a small risk for large wavelet coefficients.

For $c = 1/3$, Equation (5.11) can be written as

$$-\frac{1}{\sigma^2}\mathrm{d_g}^2 + \frac{1}{\sigma^2}\mathrm{d_y d_g} - \frac{1}{3\mathrm{b}}|\mathrm{d_g}|^{1/3} - \frac{2}{3} = 0,$$

and the larger posterior mode estimator becomes equivalent to solving the equation

$$-\frac{1}{\sigma^2}\mathrm{z}^6 + \mathrm{sign}(\mathrm{d_y})\frac{1}{\sigma^2}\mathrm{d_y z}^3 - \frac{1}{3b}\mathrm{z} - \frac{2}{3} = 0. \tag{5.14}$$

Figure 5.2 (a) shows that for $|\mathrm{d_y}| \leq \sigma^2$, the posterior mode is unique and equal to 0. On the other hand, for large values of $|\mathrm{d_y}| > \sigma^2$ there are two modes (Reményi and Vidakovic, 2015). Figure 5.2 (a) also shows the LPM rule for different values of the constant c with $\sigma = 1$ and $b = 10$ and it is apparent that the rule is thresholding because the rule is described as a heavily thresholding small coefficient in magnitude. The form in (5.14) will be used later to estimate $\mathbf{f}$. The next section, however, is about the Gaussian distribution, which is a common choice of prior on wavelet coefficients.

5.4 Gaussian distribution

The purpose of this section is to investigate and apply the Gaussian distribution as a prior. Previously, Cutillo *et al.* (2008) used a model with Gaussian distribution as a prior on wavelet coefficients. Assuming an additive Gaussian error model, the conditional distribution of the data given the truth, or equivalently, the distribution of the wavelet

coefficient $\mathrm{d_y}$ given the corresponding wavelet coefficient $\mathrm{d_g}$, is given by

$$\mathrm{p}(\mathrm{d_y}|\mathrm{d_g},\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(\mathrm{d_y}-\mathrm{d_g})^2\right\}, \quad \mathrm{d_y},\mathrm{d_g}\in\mathbb{R};\sigma>0, \tag{5.15}$$

where σ^2 is the noise variance, with the prior on the wavelet coefficients $\mathrm{d_g}$, defined as

$$\mathrm{p}(\mathrm{d_g}|\kappa) = \frac{\sqrt{\kappa}}{\sqrt{\pi}}\exp\left\{-\kappa{\mathrm{d_g}}^2\right\}, \quad \mathrm{d_g}\in\mathbb{R};\kappa>0, \tag{5.16}$$

where $N(0,\frac{1}{2\kappa})$ denotes the Gaussian with mean zero and variance $\frac{1}{2\kappa}$ (Gribble, 2001). Hence, the variances of the signal and noise $1/2\kappa$ and σ^2 are assumed known. The posterior density of the wavelet coefficients $\mathrm{d_g}$ given $\mathrm{d_y}$, can be written

$$\mathrm{p}(\mathrm{d_g}|\mathrm{d_y}) \propto \exp\left\{-\kappa{\mathrm{d_g}}^2\right\}\exp\left\{-\frac{1}{2\sigma^2}(\mathrm{d_y}-\mathrm{d_g})^2\right\}, \tag{5.17}$$

and the logarithm

$$\log\mathrm{p}(\mathrm{d_g}|\mathrm{d_y}) \propto -\kappa{\mathrm{d_g}}^2 - \frac{1}{2\sigma^2}(\mathrm{d_y}-\mathrm{d_g})^2, \tag{5.18}$$

with the maximum given as the solution of

$$-2\kappa\mathrm{d_g} + \frac{1}{\sigma^2}\mathrm{d_y} - \frac{1}{\sigma^2}\mathrm{d_g} = 0, \tag{5.19}$$

and hence the maximum a posteriori estimator (NNWS-MAP) using a normal distribution as the likelihood and the normal as a prior is given by

$$\widehat{\mathrm{d}}_\mathrm{g} = \frac{1}{(1+2\sigma^2\kappa)}\mathrm{d_y}. \tag{5.20}$$

The rule in (5.20) is shrinkage by a factor of $1/(1+2\sigma^2\kappa)$ and the posterior mean of $\mathrm{d_g}|\mathrm{d_y}$, is given by

$$\mathrm{PM}(\mathrm{d_y}) = \frac{1}{(1+2\sigma^2\kappa)}\mathrm{d_y}. \tag{5.21}$$

For more details, see Appendix (A).

The joint distribution of $\mathrm{d_y}$ and $\mathrm{d_g}$ is then

$$\begin{aligned}
\mathrm{p}(\mathrm{d_y},\mathrm{d_g}) &= \mathrm{p}(\mathrm{d_y}|\mathrm{d_g})\mathrm{p}(\mathrm{d_g}) \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)\exp\left\{-\frac{1}{2\sigma^2}(\mathrm{d_y}-\mathrm{d_g})^2\right\}\left(\sqrt{\frac{\kappa}{\pi}}\right)\exp\left\{-\kappa{\mathrm{d_g}}^2\right\} \\
&= \left(\frac{\sqrt{\kappa}}{\sqrt{2}\pi\sigma}\right)\exp\left\{-\frac{1}{2\sigma^2}(\mathrm{d_y}-\mathrm{d_g})^2\right\}\exp\left\{-\kappa{\mathrm{d_g}}^2\right\} \\
&= \left(\frac{\sqrt{\kappa}}{\sqrt{2}\pi\sigma}\right)\exp\left\{-\frac{1}{2\sigma^2}{\mathrm{d_y}}^2\right\}\exp\left\{-\frac{1}{2\sigma^2}({\mathrm{d_g}}^2-2\mathrm{d_y}\mathrm{d_g})\right\}\exp\left\{-\kappa{\mathrm{d_g}}^2\right\} \\
&= \left(\frac{\sqrt{\kappa}}{\sqrt{2}\pi\sigma}\right)\exp\left\{-\frac{1}{2\sigma^2}{\mathrm{d_y}}^2\right\}\exp\left\{-\frac{1}{2\sigma^2}({\mathrm{d_g}}^2(1+2\kappa\sigma^2)-2\mathrm{d_y}\mathrm{d_g})\right\} \\
&= \left(\frac{\sqrt{\kappa}}{\sqrt{2}\pi\sigma}\right)\exp\left\{-\frac{1}{2\sigma^2}{\mathrm{d_y}}^2\right\}\exp\left\{-\frac{1}{2\sigma^2}(1+2\kappa\sigma^2)({\mathrm{d_g}}^2-\mathrm{d_g}\frac{2\mathrm{d_y}}{1+2\kappa\sigma^2})\right\} \\
&= \left(\frac{\sqrt{\kappa}}{\sqrt{2}\pi\sigma}\right)\exp\left\{-\frac{1}{2\sigma^2}{\mathrm{d_y}}^2\right\}\exp\left\{-\frac{1}{2\sigma^2}(1+2\kappa\sigma^2)(\mathrm{d_g}-\frac{\mathrm{d_y}}{(1+2\kappa\sigma^2)})^2\right\} \\
&\quad\times\exp\left\{\frac{1}{2\sigma^2(1+2\kappa\sigma^2)}{\mathrm{d_y}}^2\right\},
\end{aligned} \tag{5.22}$$

where κ and σ^2 are assumed known. The marginal distribution of $\mathrm{d_y}$ becomes

$$\begin{aligned}
\mathrm{m}(\mathrm{d_y}) &= \int \mathrm{p}(\mathrm{d_y},\mathrm{d_g})d\mathrm{d_g} \\
&= \left(\frac{\sqrt{\kappa}}{\sqrt{2}\pi\sigma}\right)\exp\left\{-\frac{1}{2\sigma^2}{\mathrm{d_y}}^2\right\}\exp\left\{\frac{1}{2\sigma^2(1+2\kappa\sigma^2)}{\mathrm{d_y}}^2\right\} \\
&\quad\times\int\exp\left\{-\frac{(1+2\kappa\sigma^2)}{2\sigma^2}(\mathrm{d_g}-(\frac{\mathrm{d_y}}{2(1+2\kappa\sigma^2)}))^2\right\}d\mathrm{d_g} \\
&= \left(\frac{\sqrt{\kappa}}{\sqrt{2}\pi\sigma}\right)\exp\left\{-\frac{1}{2\sigma^2}{\mathrm{d_y}}^2\right\}\exp\left\{\frac{1}{2\sigma^2(1+2\kappa\sigma^2)}{\mathrm{d_y}}^2\right\}\frac{\sigma\sqrt{2\pi}}{\sqrt{1+2\kappa\sigma^2}} \\
&= \left(\frac{\sqrt{\kappa}}{\sqrt{2}\pi\sigma}\right)\frac{\sigma\sqrt{2\pi}}{\sqrt{1+2\kappa\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}{\mathrm{d_y}}^2\right\}\exp\left\{\frac{1}{2\sigma^2(1+2\kappa\sigma^2)}{\mathrm{d_y}}^2\right\}.
\end{aligned} \tag{5.23}$$

The distribution of $\mathrm{d_g}$, can be written as

$$\begin{aligned}
\mathrm{p}(\mathrm{d_g}|\mathrm{d_y}) &= \frac{\mathrm{p}(\mathrm{d_y},\mathrm{d_g})}{\mathrm{m}(\mathrm{d_y})} \\
&= \frac{\sqrt{1+2\kappa\sigma^2}}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{(1+2\kappa\sigma^2)}{2\sigma^2}(\mathrm{d_g}-\frac{\mathrm{d_y}}{2(1+2\kappa\sigma^2)})^2\right\}.
\end{aligned} \tag{5.24}$$

The given model depends on two prior parameters, σ^2 and κ, which have to be specified. The approach for estimating the parameters can be summarized as:

- σ^2 represents the variance of the noise level, which can be estimated using (2.62) or (2.63), to give $\widehat{\sigma}$.

- κ_j represents the variance of the prior distribution at resolution level j, which depends on the spread of the normal prior distribution and the error, so

$$\widehat{\sigma}^2_{\mathbf{d}_{\mathbf{y}_j}} = \frac{1}{2\kappa_j} + \widehat{\sigma}^2, \tag{5.25}$$

where $\widehat{\sigma}^2_{\mathbf{d}_{\mathbf{y}_j}}$ is the variance of the wavelet coefficient at the j^{th} level in the likelihood and $\frac{1}{2\kappa_j}$ is the variance of the signal. Then

$$\widehat{\kappa}_j = \frac{1}{2(\widehat{\sigma}^2_{\mathbf{d}_{\mathbf{y}_j}} - \widehat{\sigma}^2)}, \quad j_0 < j < \log_2(n) - 1, \tag{5.26}$$

where $j_0 \in \mathbb{Z}$ and Antoniadis *et al.* (2001) suggested that $j_0 = 3$. In the case $\widehat{\sigma}^2_{\mathbf{d}_{\mathbf{y}_j}} < \widehat{\sigma}^2$, the parameter $\widehat{\kappa}_j = 0$ can be set. So, the formula for $\widehat{\kappa}_j$, can be written as

$$\widehat{\kappa}_j = \max\left[0, \frac{1}{2(\widehat{\sigma}^2_{\mathbf{d}_{\mathbf{y}_j}} - \widehat{\sigma}^2)}\right], \quad j_0 < j < \log_2(n) - 1. \tag{5.27}$$

5.5 Laplace distribution

The Laplace prior was first used as a regularisation penalty by Tibshirani (1996) and is discussed by Vidakovic and Ruggeri (2001) and Johnstone and Silverman (2005a,b) in the wavelet domain. Considering an additive Gaussian error model for the conditional distribution of the data given the truth, or equivalently, the distribution of the wavelet coefficient $\mathrm{d_y}$ given the corresponding wavelet coefficient $\mathrm{d_g}$, is given by

$$\mathrm{p}(\mathrm{d_y}|\mathrm{d_g}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(\mathrm{d_y} - \mathrm{d_g})^2 \right\}, \quad \mathrm{d_y}, \mathrm{d_g} \in \mathbb{R}; \sigma > 0, \tag{5.28}$$

with the prior on the wavelet coefficients $\mathrm{d_g}$, defined as

$$\mathrm{p}(\mathrm{d_g}) = \frac{\kappa}{2} \exp\left\{ -\kappa|\mathrm{d_g}| \right\}, \quad \mathrm{d_g} \in \mathbb{R}; \kappa > 0, \tag{5.29}$$

where $\mathrm{p}(\mathrm{d_g}|\kappa)$ denotes the Laplace distribution with mean zero and variance $2/\kappa^2$. Hence, the variances of the signal and noise $2/\kappa^2$ and σ are known. The prior distribution of $\mathrm{d_g}$ includes prior parameter κ. Hence, for $\kappa = 1$, the Laplace density function tends to 0.5 as t approaches zero and is strictly decreasing as $|\mathrm{d_g}|$ increases. As κ increases, the value of the peak of the density function increases at $t = 0$.

The PM of the wavelet coefficients $\mathrm{d_g}$ given $\mathrm{d_y}$, can be written as

$$\begin{aligned}\mathrm{PM}(\mathrm{d_g}|\mathrm{d_y}) = &\left((\mathrm{d_y}+\sigma^2\kappa)\exp\left\{\frac{1}{2\sigma^2}(\mathrm{d_y}+\sigma^2\kappa)^2\right\}\bar{\Phi}\left(\frac{1}{\sigma}(\mathrm{d_y}+\sigma^2\kappa)\right)\right.\\ &\left.+(\mathrm{d_y}-\sigma^2\kappa)\exp\left\{\frac{1}{2\sigma^2}(\mathrm{d_y}-\sigma^2\kappa)^2\right\}\Phi\left(\frac{1}{\sigma}(\mathrm{d_y}-\sigma^2\kappa)\right)\right)\\ \Big/&\left(\exp\left\{\frac{1}{2\sigma^2}(\mathrm{d_y}+\sigma^2\kappa)^2\right\}\bar{\Phi}\left(\frac{1}{\sigma}(\mathrm{d_y}+\sigma^2\kappa)\right)-\exp\left\{\frac{1}{2\sigma^2}(\mathrm{d_y}-\sigma^2\kappa)^2\right\}\Phi\left(\frac{1}{\sigma}(\mathrm{d_y}-\sigma^2\kappa)\right)\right),\end{aligned} \tag{5.30}$$

where Φ is the standard Gaussian probability distribution function. In addition, consider that $\Phi\Big(-\frac{1}{\sigma}(\kappa\sigma^2+\mathrm{d_y})\Big) = \bar{\Phi}\Big(\frac{1}{\sigma}(\kappa\sigma^2+\mathrm{d_y})\Big)$, where $\bar{\Phi}$ is the complement of the standard Gaussian probability distribution function, that is $\Phi(-t) = \bar{\Phi}(t)$.

Consider the posterior density of the wavelet coefficients $\mathrm{d_g}$ given $\mathrm{d_y}$, written as

$$\mathrm{p}(\mathrm{d_g}|\mathrm{d_y}) \propto \exp\Big\{-\frac{1}{2\sigma^2}(\mathrm{d_y}-\mathrm{d_g})^2\Big\}\exp\Big\{-\kappa|\mathrm{d_g}|\Big\}, \tag{5.31}$$

then the maximum is given by the solution of

$$\frac{1}{\sigma^2}(\mathrm{d_y}-\mathrm{d_g}) - \kappa\frac{1}{|\mathrm{d_g}|}\mathrm{d_g} = 0, \tag{5.32}$$

which is equivalent to the equation

$$-\frac{1}{\sigma^2}\mathrm{d_g} - \kappa\mathrm{d_g} + \frac{1}{\sigma^2}\mathrm{d_y}|\mathrm{d_g}| = 0. \tag{5.33}$$

That is

$$\begin{aligned}-\frac{1}{\sigma^2}\mathrm{d_g} - (\kappa+\frac{\mathrm{d_y}}{\sigma^2})\mathrm{d_g} = 0, \quad &\text{if} \quad \mathrm{d_g} < 0\\ -\frac{1}{\sigma^2}{\mathrm{d_g}}^2 - (\kappa-\frac{d}{\sigma^2})\mathrm{d_g} = 0, \quad &\text{if} \quad \mathrm{d_g} > 0.\end{aligned} \tag{5.34}$$

So, the solution of (5.34) is

$$\widehat{\mathrm{d}}_{\mathrm{g}} = \begin{cases} \mathrm{d}_{\mathrm{y}} + \kappa\sigma^2, & \text{if } \mathrm{d}_{\mathrm{g}} < 0 \\ \mathrm{d}_{\mathrm{y}} - \kappa\sigma^2, & \text{if } \mathrm{d}_{\mathrm{g}} > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{5.35}$$

The maximum a posteriori estimator has three solutions; if the wavelet coefficient $|\mathrm{d}_{\mathrm{y}}| \leq \kappa\sigma^2$, then the solution is zero; when $\mathrm{d}_{\mathrm{y}} < -\kappa\sigma^2$, the solution is $\mathrm{d}_{\mathrm{y}} + \kappa\sigma^2$, and when $\mathrm{d}_{\mathrm{y}} > \kappa\sigma^2$ the solution is $\mathrm{d}_{\mathrm{y}} - \kappa\sigma^2$. So, the maximum a posteriori estimator (NDEWS-MAP) using the Gaussian distribution as the likelihood and the Laplace as the prior can be re-written as

$$\widehat{\mathrm{d}}_{\mathrm{g}} = \begin{cases} \mathrm{d}_{\mathrm{y}} + \kappa\sigma^2, & \text{if } \mathrm{d}_{\mathrm{y}} < -\sigma^2\kappa \\ \mathrm{d}_{\mathrm{y}} - \kappa\sigma^2, & \text{if } \mathrm{d}_{\mathrm{y}} > \sigma^2\kappa \\ 0, & \text{otherwise.} \end{cases} \tag{5.36}$$

The results in (5.36) show that the maximum a posteriori estimator keeps all the information if $\kappa = 0$, which means that only the likelihood is used and there is no shrinkage; for more detail see Hastie *et al.* (2009). Hence, there is no difference between Equations (2.19) and (5.36). The mean of the MAP is given by

$$\begin{aligned} \mathrm{E}_{\kappa}(\widehat{\mathrm{d}}_{\mathrm{g}}) = &-\sigma\phi(\frac{1}{\sigma}(\kappa\sigma^2 + \mathrm{d}_{\mathrm{y}})) + \mathrm{d}_{\mathrm{y}}\bar{\Phi}(\frac{1}{\sigma}(\kappa\sigma^2 + \mathrm{d}_{\mathrm{y}})) - \sigma^2\kappa\Phi(\frac{1}{\sigma}(\kappa\sigma^2 + \mathrm{d}_{\mathrm{y}})) \\ &+ \sigma\phi(\frac{1}{\sigma}(\kappa\sigma^2 - \mathrm{d}_{\mathrm{y}})) + \mathrm{d}_{\mathrm{y}}\bar{\Phi}(\frac{1}{\sigma}(\kappa\sigma^2 - \mathrm{d}_{\mathrm{y}})) + \sigma^2\kappa\Phi(\frac{1}{\sigma}(\kappa\sigma^2 - \mathrm{d}_{\mathrm{y}})). \end{aligned} \tag{5.37}$$

Then

$$\begin{aligned} &\mathrm{E}_{\kappa}(\widehat{\mathrm{d}}_{\mathrm{g}})^2 \\ &= -\sigma(\frac{1}{\sigma}(\kappa\sigma^2 + \mathrm{d}_{\mathrm{y}}))\phi(\frac{1}{\sigma}(\kappa\sigma^2 + \mathrm{d}_{\mathrm{y}})) + (\sigma - 2(\mathrm{d}_{\mathrm{y}} + \sigma^2\kappa) + (\mathrm{d}_{\mathrm{y}} + \sigma^2\kappa)^2)\bar{\Phi}(\frac{1}{\sigma}(\kappa\sigma^2 + \mathrm{d}_{\mathrm{y}})) \\ &\quad + \sigma(\frac{1}{\sigma}(\kappa\sigma^2 - \mathrm{d}_{\mathrm{y}}))\phi(\frac{1}{\sigma}(\kappa\sigma^2 - \mathrm{d}_{\mathrm{y}})) + (\sigma + 2(\mathrm{d}_{\mathrm{y}} - \sigma^2\kappa) + (\mathrm{d}_{\mathrm{y}} - \sigma^2\kappa)^2)\bar{\Phi}(\frac{1}{\sigma}(\kappa\sigma^2 - \mathrm{d}_{\mathrm{y}})), \end{aligned} \tag{5.38}$$

and, hence, the variance is

$$\mathrm{V}_{\kappa}(\widehat{\mathrm{d}}_{\mathrm{g}}) = \mathrm{E}_{\kappa}(\widehat{\mathrm{d}}_{\mathrm{g}})^2 - (\mathrm{E}_{\kappa}(\widehat{\mathrm{d}}_{\mathrm{g}}))^2. \tag{5.39}$$

For more details, see Appendix (B).

The joint distribution for $\mathrm{d_y}$ and $\mathrm{d_g}$ is then

$$
\begin{aligned}
&\mathrm{p}(\mathrm{d_y}, \mathrm{d_g}) = \mathrm{p}(\mathrm{d_y}|\mathrm{d_g})\mathrm{p}(\mathrm{d_g}) \\
=&\left(\frac{\kappa}{2\sqrt{2\pi\sigma^2}}\right) \exp\left\{-\frac{1}{2\sigma^2}(\mathrm{d_y}-\mathrm{d_g})^2\right\} \exp\{-\kappa|\mathrm{d_g}|\} \\
=&\left(\frac{\kappa}{2\sqrt{2\pi\sigma^2}}\right) \exp\left\{-\frac{1}{2\sigma^2}({\mathrm{d_g}}^2 - 2\mathrm{d_g}(-\mathrm{sign}(\mathrm{d_g})\sigma^2\kappa + \mathrm{d_y}) + {\mathrm{d_y}}^2)\right\} \\
=&\left(\frac{\kappa}{2\sqrt{2\pi\sigma^2}}\right) \exp\left\{-\frac{1}{2\sigma^2}(\mathrm{d_g} - (\mathrm{d_y} - \mathrm{sign}(\mathrm{d_g})\sigma^2\kappa)^2)\right\} - \exp\left\{-\frac{1}{2\sigma^2}(-(\mathrm{d_y} - \mathrm{sign}(\mathrm{d_g})\sigma^2\kappa)^2 + {\mathrm{d_y}}^2)\right\} \\
=&\left(\frac{\kappa}{2\sqrt{2\pi\sigma^2}}\right) \exp\left\{\frac{1}{2}\sigma^2\kappa\right\} \exp\{-\mathrm{sign}(\mathrm{d_g})\mathrm{d_y}\kappa\} \times \exp\left\{-\frac{1}{2\sigma^2}(\mathrm{d_g} - (\mathrm{d_y} - \mathrm{sign}(\mathrm{d_g})\sigma^2\kappa))^2\right\} \\
=&\begin{cases}
\dfrac{\kappa \exp\left\{\frac{1}{2}\sigma^2\kappa\right\} \exp\{-\mathrm{d_y}\kappa\}}{(2\sqrt{2\pi\sigma^2})} \exp\left\{-\frac{1}{2\sigma^2}(\mathrm{d_g} - (\mathrm{d_y} - \sigma^2\kappa))^2\right\}, & \text{if } \mathrm{d_g} \geq 0 \\
\dfrac{\kappa \exp\left\{\frac{1}{2}\sigma^2\kappa\right\} \exp\{\mathrm{d_y}\kappa\}}{(2\sqrt{2\pi\sigma^2})} \exp\left\{-\frac{1}{2\sigma^2}(\mathrm{d_g} - (\mathrm{d_y} + \sigma^2\kappa))^2\right\}, & \text{if } \mathrm{d_g} < 0.
\end{cases}
\end{aligned}
\tag{5.40}
$$

The marginal distribution for $\mathrm{d_y}$, is given by

$$
\begin{aligned}
&\mathrm{m}(\mathrm{d_y}) = \int \mathrm{p}(\mathrm{d_y}, \mathrm{d_g}|\sigma^2, \kappa) d\mathrm{d_g} \\
=&\left(\frac{\kappa}{2\sqrt{2\pi\sigma^2}}\right) \exp\left\{\frac{1}{2}\sigma^2\kappa\right\} \Bigg(\exp\{\mathrm{d_y}\kappa\} \int_{-\infty}^{0} \exp\left\{-\frac{1}{2\sigma^2}(\mathrm{d_g} - (\mathrm{d_y} + \sigma^2\kappa))^2\right\} d\mathrm{d_g} \\
&\quad + \exp\{-\mathrm{d_y}\kappa\} \int_{0}^{\infty} \exp\{-\frac{1}{2\sigma^2}(\mathrm{d_g} - (\mathrm{d_y} - \sigma^2\kappa))^2\} d\mathrm{d_g} \Bigg) \\
=&\left(\frac{\kappa}{2\sqrt{2\pi\sigma^2}}\right) \sqrt{2\pi\sigma^2} \exp\left\{\frac{1}{2}\sigma^2\kappa\right\} \Bigg(\exp\{\mathrm{d_y}\kappa\}\Phi(\frac{-\mathrm{d_y} - \sigma^2\kappa}{\sigma}) + \exp\{-\mathrm{d_y}\kappa\}\Phi(\frac{\mathrm{d_y} - \sigma^2\kappa}{\sigma}) \Bigg),
\end{aligned}
\tag{5.41}
$$

and the conditional distribution for $\mathrm{d_g}$ given $\mathrm{d_y}$, can be written as

$$\begin{aligned}
\mathrm{p}(\mathrm{d_g}|\mathrm{d_y}) &= \frac{\mathrm{p}(\mathrm{d_y},\mathrm{d_g})}{\mathrm{m}(\mathrm{d_y})} \\
&= \begin{cases} \frac{\exp\{-\mathrm{d_y}\kappa\}}{\exp\{-\mathrm{d_y}\kappa\}\Phi(\frac{\mathrm{d_y}-\sigma^2\kappa}{\sigma})\exp\{\mathrm{d_y}\kappa\}\Phi(\frac{-\mathrm{d_y}-\sigma^2\kappa}{\sigma})} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathrm{d_g}-(\mathrm{d_y}-\sigma^2\kappa))^2\right\}, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{if } \mathrm{d_g} \geq 0 \\ \frac{\exp\{\mathrm{d_y}\kappa\}}{\exp(-\mathrm{d_y}\kappa)\Phi(\frac{\mathrm{d_y}-\sigma^2\kappa}{\sigma})\exp\{\mathrm{d_y}\kappa\}\Phi(\frac{-\mathrm{d_y}-\sigma^2\kappa}{\sigma})} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathrm{d_g}-(\mathrm{d_y}+\sigma^2\kappa))^2\right\}, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{if } \mathrm{d_g} < 0, \end{cases}
\end{aligned} \tag{5.42}$$

where these results were derived by Pericchi and Smith (1992). The posterior distribution for the wavelet coefficients in (5.31) is shown in Figure (5.3) for $\kappa = 1$, $\sigma^2 = 1$, and $\mathbf{d_y} = \{-0.1, -2, 0.1, 2\}$. It is apparent that the shape of the posterior depends on the wavelet coefficient $\mathrm{d_y}$, κ and σ^2. If $|\mathrm{d_y}| \leq \kappa\sigma^2$, the posterior distribution is unimodal and the maximum of the posterior equals 0. In contrast, there are two values for posterior mode when $|\mathrm{d_y}| > \kappa\sigma^2$.

Figure 5.3 shows the posterior density for different values of the wavelet coefficients. For $|\mathrm{d_y}| \leq \kappa\sigma^2$, the posterior is symmetric and the solution for estimating the wavelet coefficients is zero, but for $|\mathrm{d_y}| > \kappa\sigma^2$, the posterior is not symmetric and the solution for estimating the wavelet coefficient is $\mathrm{d_y} \pm \kappa\sigma^2$, that is non-unique.

The model depends on two prior parameters, which are σ^2 and κ. The parameter κ represents the variance of the prior distribution, with the variance of the signal part being $2/\kappa^2$. The variance of the wavelet coefficients at the j^{th} level is $\sigma^2_{\mathbf{d_y}_j}$, which depends on the variance of the prior distribution and the error, then

$$\widehat{\kappa}_j = \max\left[0, \sqrt{\frac{2}{(\sigma^2_{\mathbf{d_y}_j} - \widehat{\sigma}^2)}}\right], \quad J_0 < j < J-1. \tag{5.43}$$

In the case of $\sigma^2_{\mathbf{d_y}_j} < \widehat{\sigma}^2$, the parameter $\widehat{\kappa}_j$ can be set to zero – this is equivalent to no thresholding.

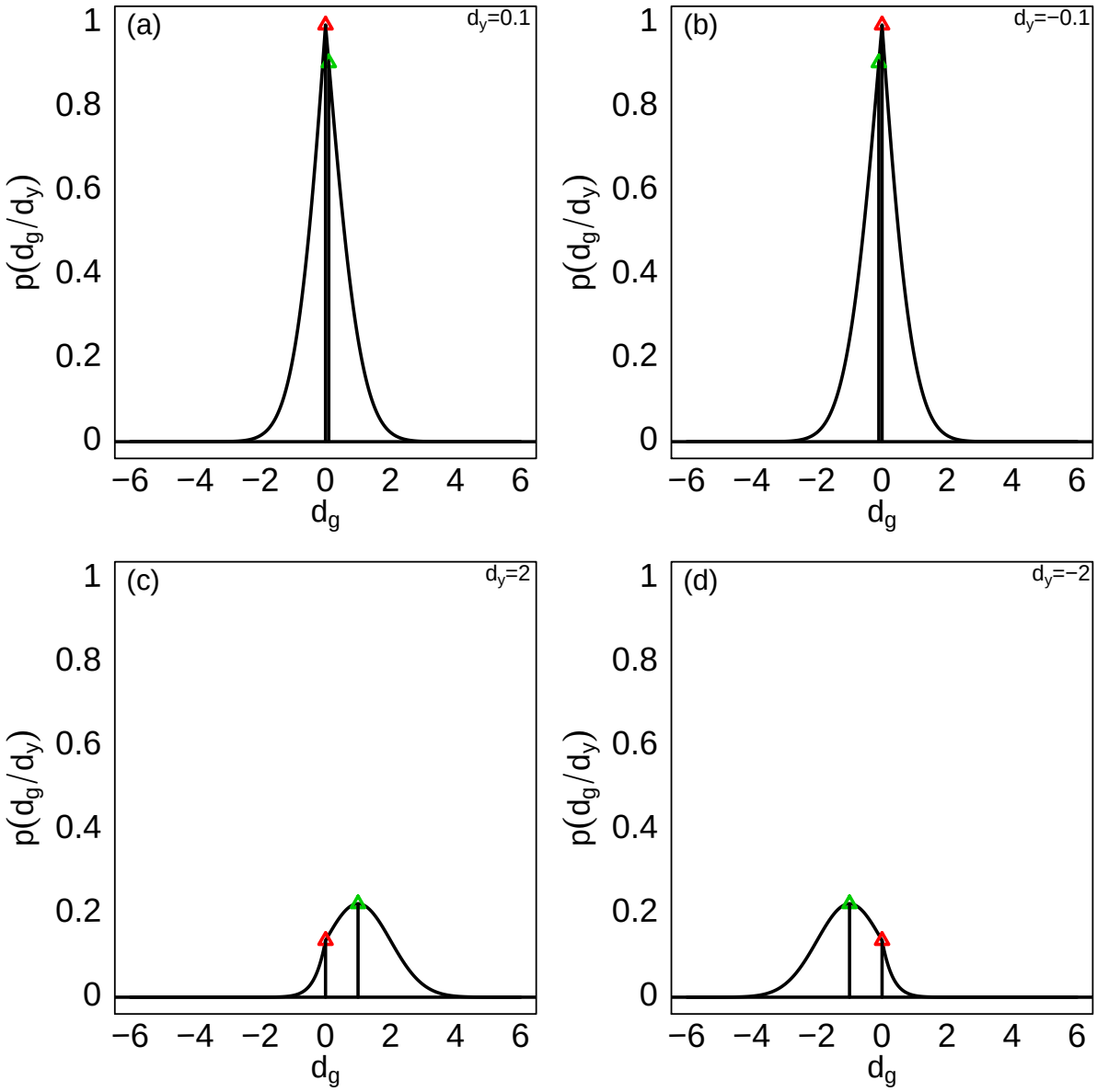


Figure 5.3: Plots of the posterior (5.31) and maximum a posteriori estimate (5.36) for different values of d_y; the red triangles are the maximum, when $|d_y| \leq \kappa\sigma^2$; and the green triangles are the maximum, when $|d_y| > \kappa\sigma^2$.

5.6 Elastic-net distribution

The purpose of this section is to introduce a new elastic-net distribution which combines the Laplace and Gaussian distributions to overcome the limitations of each.

The model for the conditional distribution of the data given the truth, or equivalently, the distribution of the wavelet coefficient $\mathrm{d_y}$ given the corresponding wavelet coefficient $\mathrm{d_g}$, is given by

$$\mathrm{p}(\mathrm{d_y}|\mathrm{d_g},\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(\mathrm{d_y}-\mathrm{d_g})^2\right\},\quad \mathrm{d_y},\mathrm{d_g}\in\mathbb{R};\sigma>0, \tag{5.44}$$

where $\mathrm{p}(\mathrm{d_y}|\mathrm{d_g},\sigma^2)$ is the likelihood and σ^2 is the noise variance, which is assumed known. The elastic-net prior of the wavelet coefficients $\mathrm{d_g}$, is defined as

$$\mathrm{p}(\mathrm{d_g})=\left(\frac{1}{Z(\kappa,\gamma)}\right)\exp\left\{-\kappa(\gamma{\mathrm{d_g}}^2+(1-\gamma)|\mathrm{d_g}|)\right\},\quad \mathrm{d_g}\in\mathbb{R};\kappa>0,0<\gamma<1, \tag{5.45}$$

where

$$Z(\kappa,\gamma)=\begin{cases}2/\kappa, & \gamma=0\\ \sqrt{\frac{4\pi}{\kappa\gamma}}\exp\left\{\frac{1}{4\gamma}\kappa(1-\gamma)^2\right\}\left(1-\Phi(\frac{\kappa(1-\gamma)}{\sqrt{2\kappa\gamma}})\right), & 0<\gamma<1\\ \sqrt{\pi/\kappa}, & \gamma=1.\end{cases} \tag{5.46}$$

In this section, the parameters, κ and γ of elastic-net will be assumed known. Note that for the limit values of γ, this reduces to the Gaussian case ($\gamma=1$) and the Laplace case ($\gamma=0$). The reason for finding the constant $Z(\kappa,\gamma)$, is to allow the prior parameters, κ and γ, to be estimated.

To prove the constant in the above distribution, consider the integrated,

$$Z(\kappa,\gamma)=\int\exp\left\{-(\kappa\gamma{\mathrm{d_g}}^2+\kappa(1-\gamma)|\mathrm{d_g}|)\right\}d\mathrm{d_g}, \tag{5.47}$$

then suppose $A = \kappa\gamma$ and $B = \kappa(1-\gamma)$, giving

$$\begin{aligned}
&\int \exp\left\{-(A{\mathrm{d_g}}^2 + B|\mathrm{d_g}|)\right\} d\mathrm{d_g} \\
&= \int_{-\infty}^{0} \exp\left\{-(A{\mathrm{d_g}}^2 - B\mathrm{d_g})\right\} d\mathrm{d_g} + \int_{0}^{\infty} \exp\left\{-(A{\mathrm{d_g}}^2 + B\mathrm{d_g})\right\} d\mathrm{d_g} \\
&= \int_{-\infty}^{0} \exp\left\{-A({\mathrm{d_g}}^2 - \frac{B}{A}\mathrm{d_g})\right\} d\mathrm{d_g} + \int_{0}^{\infty} \exp\left\{-A({\mathrm{d_g}}^2 + \frac{B}{A}\mathrm{d_g})\right\} d\mathrm{d_g} \\
&= \exp\left\{\frac{1}{4A}B^2\right\} \int_{-\infty}^{0} \exp\left\{-A(\mathrm{d_g} - \frac{B}{2A})^2\right\} d\mathrm{d_g} + \exp\left\{\frac{1}{4A}B^2\right\} \int_{0}^{\infty} \exp\left\{-A(\mathrm{d_g} + \frac{B}{2A})^2\right\} d\mathrm{d_g}.
\end{aligned} \tag{5.48}$$

Now, let $v = \sqrt{2A}(\mathrm{d_g} - \frac{B}{2A})$ and $u = \sqrt{2A}(\mathrm{d_g} + \frac{B}{2A})$, hence

$$\begin{aligned}
&\int \exp\left\{-(A{\mathrm{d_g}}^2 + B|\mathrm{d_g}|)\right\} d\mathrm{d_g} \\
&= \frac{1}{\sqrt{2A}} \exp\left\{A(\frac{B}{2A})^2\right\} \left(\int_{-\infty}^{-\sqrt{2A}\frac{B}{2A}} \exp\left\{-\frac{1}{2}v^2\right\} dv + \int_{\sqrt{2A}\frac{B}{2A}}^{\infty} \exp\left\{-\frac{1}{2}u^2\right\} du \right) \\
&= \frac{\sqrt{\pi}}{\sqrt{A}} \exp\left\{\frac{1}{4A}B^2\right\} \left(\Phi(-\sqrt{2A}\frac{B}{2A}) + (1 - \Phi(\sqrt{2A}\frac{B}{2A})) \right) \\
&= \frac{\sqrt{4\pi}}{\sqrt{A}} \exp\left\{\frac{1}{4A}B^2\right\} \left(1 - \Phi(\sqrt{2A}\frac{B}{2A}) \right).
\end{aligned} \tag{5.49}$$

So, that

$$\mathrm{p}(\mathrm{d_g}) = \frac{1}{\frac{\sqrt{4\pi}}{\sqrt{\kappa\gamma}} \exp\left\{\frac{(\kappa(1-\gamma))^2}{4\kappa\gamma}\right\} \left(1 - \Phi(\frac{\kappa(1-\gamma)}{\sqrt{2\kappa\gamma}}) \right)} \exp\left\{-\kappa(\gamma{\mathrm{d_g}}^2 + (1-\gamma)|\mathrm{d_g}|)\right\},$$

$$\mathrm{d_g} \in \mathbb{R}; \kappa > 0, 0 < \gamma < 1. \tag{5.50}$$

Consider the model, with likelihood in (5.44) and (5.50), the joint distribution of $\mathrm{d_y}$ and

$\mathrm{d_g}$ is then

$$
\begin{aligned}
&\mathrm{p}(\mathrm{d_y}, \mathrm{d_g}) \\
&= \begin{cases}
\dfrac{1}{\frac{2\sqrt{2}\pi\sigma}{\sqrt{\kappa\gamma}} \exp\left\{\frac{(\kappa(1-\gamma))^2}{4\kappa\gamma}\right\}\left(1-\Phi(\frac{\kappa(1-\gamma)}{\sqrt{2\kappa\gamma}})\right)} \exp\left\{-\frac{1}{2\sigma^2}{\mathrm{d_y}}^2\right\} \quad \exp\left\{\left(\frac{1+2\sigma^2\kappa\gamma}{2\sigma^2}\right)\left(\frac{\mathrm{d_y}-2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma}\right)^2\right\} \\
\qquad \times \exp\left\{-\left(\frac{(1+2\sigma^2\kappa\gamma)}{2\sigma^2}\right)\left(\mathrm{d_g}-\frac{\mathrm{d_y}-2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma}\right)^2\right\}, \qquad \text{if } \mathrm{d_g} \geq 0 \\
\dfrac{1}{\frac{2\sqrt{2}\pi\sigma}{\sqrt{\kappa\gamma}} \exp\left\{\frac{(\kappa(1-\gamma))^2}{4\kappa\gamma}\right\}\left(1-\Phi(\frac{\kappa(1-\gamma)}{\sqrt{2\kappa\gamma}})\right)} \exp\left\{-\frac{1}{2\sigma^2}{\mathrm{d_y}}^2\right\} \quad \exp\left\{\left(\frac{1+2\sigma^2\kappa\gamma}{2\sigma^2}\right)\left(\frac{\mathrm{d_y}+2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma}\right)^2\right\} \\
\qquad \times \exp\left\{-\left(\frac{(1+2\sigma^2\kappa\gamma)}{2\sigma^2}\right)\left(\mathrm{d_g}-\frac{\mathrm{d_y}+2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma}\right)^2\right\}, \qquad \text{if } \mathrm{d_g} < 0,
\end{cases}
\end{aligned}
\tag{5.51}
$$

the marginal distribution is given by

$$
\begin{aligned}
\mathrm{m}(\mathrm{d_y}) =& \frac{1}{\sqrt{\frac{8\pi^2\sigma^2}{\kappa\gamma}} \exp\left\{\frac{(\kappa(1-\gamma))^2}{4\kappa\gamma}\right\}\left(1-\Phi(\frac{\kappa(1-\gamma)}{\sqrt{2\kappa\gamma}})\right)} \exp\left\{-\frac{1}{2\sigma^2}{\mathrm{d_y}}^2\right\} \\
&\times \left(\exp\left\{\left(\frac{1+2\sigma^2\kappa\gamma}{2\sigma^2}\right)\left(\frac{\mathrm{d_y}+2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma}\right)^2\right\}\right. \\
&\times \sqrt{\frac{2\pi\sigma^2}{1+2\sigma^2\kappa\gamma}}\left(1-\Phi(\sqrt{\frac{1+2\sigma^2\kappa\gamma}{\sigma^2}}\frac{\mathrm{d_y}+2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma})\right) \\
&+ \exp\left\{\left(\frac{1+2\sigma^2\kappa\gamma}{2\sigma^2}\right)\left(\frac{\mathrm{d_y}-2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma}\right)^2\right\} \\
&\left.\times \sqrt{\frac{2\pi\sigma^2}{1+2\sigma^2\kappa\gamma}}\left(\Phi(\sqrt{\frac{1+2\sigma^2\kappa\gamma}{\sigma^2}}\frac{\mathrm{d_y}-2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma})\right)\right),
\end{aligned}
\tag{5.52}
$$

and the condition distribution of $\mathrm{d_g}$ given $\mathrm{d_y}$, can be written as

$$\mathrm{p}(\mathrm{d_g}|\mathrm{d_y}) = \frac{\mathrm{p}(\mathrm{d_y},\mathrm{d_g})}{\mathrm{m}(\mathrm{d_y})}$$

$$= \begin{cases} \dfrac{\exp\left\{\left(\frac{\mathrm{d_y}-2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma}\right)^2\right\}\exp\left\{-\left(\frac{1+2\sigma^2\kappa\gamma)}{2\sigma^2}\right\}\left(\mathrm{d_g}-\frac{\mathrm{d_y}-2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma}\right)^2\right)}{\sqrt{\frac{2\pi\sigma^2}{1+2\sigma^2\kappa\gamma}}\left(\exp\left\{\left(\frac{\mathrm{d_y}+2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma}\right)^2\right\}\left(1-\Phi(\sqrt{\frac{1+2\sigma^2\kappa\gamma}{\sigma^2}}\frac{\mathrm{d_y}+2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma})\right)+\exp\left\{\left(\frac{\mathrm{d_y}-2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma}\right)^2\right\}\left(\Phi(\sqrt{\frac{1+2\sigma^2\kappa\gamma}{\sigma^2}}\frac{\mathrm{d_y}-2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma})\right)\right)}, & \text{if } \mathrm{d_g}\geq 0 \\ \dfrac{\exp\left\{\left(\frac{\mathrm{d_y}+2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma}\right)^2\right\}\exp\left\{-\left(\frac{1+2\sigma^2\kappa\gamma)}{2\sigma^2}\right\}\left(\mathrm{d_g}-\frac{\mathrm{d_y}+2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma}\right)^2\right)}{\sqrt{\frac{2\pi\sigma^2}{1+2\sigma^2\kappa\gamma}}\left(\exp\left\{\left(\frac{\mathrm{d_y}+2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma}\right)^2\right\}\left(1-\Phi(\sqrt{\frac{1+2\sigma^2\kappa\gamma}{\sigma^2}}\frac{\mathrm{d_y}+2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma})\right)+\exp\left\{\left(\frac{\mathrm{d_y}-2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma}\right)^2\right\}\left(\Phi(\sqrt{\frac{1+2\sigma^2\kappa\gamma}{\sigma^2}}\frac{\mathrm{d_y}-2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma})\right)\right)}, & \text{if } \mathrm{d_g}<0. \end{cases} \tag{5.53}$$

Now, the posterior mean and the maximum a posteriori will be computed for the elastic-net prior. Suppose the posterior is given by

$$\mathrm{p}(\mathrm{d_g}|\mathrm{d_y}) \propto \exp\left\{\frac{-(\mathrm{d_y}-\mathrm{d_g})^2}{2\sigma^2}\right\}\exp\left\{-(\kappa\gamma{\mathrm{d_g}}^2+\kappa(1-\gamma)|\mathrm{d_g}|\right\}, \tag{5.54}$$

then the maximum is the solution of

$$\frac{(\mathrm{d_y}-\mathrm{d_g})}{\sigma^2}-\kappa(1-\gamma)\frac{\mathrm{d_g}}{|\mathrm{d_g}|}-2\kappa\gamma\mathrm{d_g}=0, \tag{5.55}$$

which is equivalent to the equation

$$-\left(\frac{1}{\sigma^2}+2\kappa\gamma\right){\mathrm{d_g}}^2-\kappa(1-\gamma)\mathrm{d_g}+\frac{\mathrm{d_y}}{\sigma^2}|\mathrm{d_g}|=0, \tag{5.56}$$

that is

$$\begin{aligned} -\left(\frac{1}{\sigma^2}+2\kappa\gamma\right){\mathrm{d_g}}^2-\left(\kappa(1-\gamma)+\frac{\mathrm{d_y}}{\sigma^2}\right)\mathrm{d_g}=0, \quad \text{if} \quad \mathrm{d_g}<0 \\ -\left(\frac{1}{\sigma^2}+2\kappa\gamma\right){\mathrm{d_g}}^2-\left(\kappa(1-\gamma)-\frac{\mathrm{d_y}}{\sigma^2}\right)\mathrm{d_g}=0, \quad \text{if} \quad \mathrm{d_g}>0. \end{aligned} \tag{5.57}$$

So, the solution of (5.57) is

$$\widehat{\mathrm{d}}_\mathrm{g} = \begin{cases} \frac{(\frac{\mathrm{d_y}}{\sigma^2}+\kappa(1-\gamma))}{(\frac{1}{\sigma^2}+2\kappa\gamma)}, & \text{if } \mathrm{d_g}<0 \\ \frac{(\frac{\mathrm{d_y}}{\sigma^2}-\kappa(1-\gamma))}{(\frac{1}{\sigma^2}+2\kappa\gamma)}, & \text{if } \mathrm{d_g}>0 \\ 0, & \text{otherwise.} \end{cases} \tag{5.58}$$

So, there are three cases: when the wavelet coefficient $|\mathrm{d_y}| \leq \sigma^2\kappa(1-\gamma)$, then the solution is zero, and when $|\mathrm{d_y}| > \sigma^2\kappa(1-\gamma)$, there are two solutions. In this case, the threshold is equal to $\sigma^2\kappa(1-\gamma)$. Then the solution in (5.58) can be re-written as

$$\widehat{\mathrm{d}}_{\mathrm{g}} = \widehat{\mathrm{T}}(\mathrm{d_y}, \kappa, \gamma) = \begin{cases} \frac{\mathrm{d_y}+\sigma^2\kappa(1-\gamma)}{1+2\kappa\gamma\sigma^2}, & \text{if } \mathrm{d_y} < -\sigma^2\kappa(1-\gamma) \\ \frac{\mathrm{d_y}-\sigma^2\kappa(1-\gamma)}{1+2\kappa\gamma\sigma^2}, & \text{if } \mathrm{d_y} > \sigma^2\kappa(1-\gamma) \\ 0, & \text{otherwise.} \end{cases} \tag{5.59}$$

Hence, if $\gamma = 1$ then the maximum a posteriori is given by

$$\widehat{\mathrm{d}}_{\mathrm{g}} = \widehat{\mathrm{T}}(\mathrm{d_y}, \kappa) = \begin{cases} \frac{\mathrm{d_y}}{1+2\kappa\sigma^2}, & \text{if } \mathrm{d_y} < 0 \\ \frac{\mathrm{d_y}}{1+2\kappa\sigma^2}, & \text{if } \mathrm{d_y} > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{5.60}$$

Similarly, if $\gamma = 0$ then the maximum a posteriori is given by

$$\widehat{\mathrm{d}}_{\mathrm{g}} = \widehat{\mathrm{T}}(\mathrm{d_y}, \kappa) = \begin{cases} \mathrm{d_y} + \sigma^2\kappa, & \text{if } \mathrm{d_y} < -\sigma^2\kappa \\ \mathrm{d_y} - \sigma^2\kappa, & \text{if } \mathrm{d_y} > \sigma^2\kappa \\ 0, & \text{otherwise.} \end{cases} \tag{5.61}$$

The posterior mean can be written as

$$\begin{aligned} \mathrm{PM}(\mathrm{d_g}|\mathrm{d_y}) = & \left(\exp\left\{ \left(\tfrac{\mathrm{d_y}+2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma} \right)^2 \right\} \left(-\tfrac{\sqrt{2\pi}}{2}\phi\left(\sqrt{2}\tfrac{(\mathrm{d_y}+2\sigma^2\kappa(1-\gamma))}{1+2\sigma^2\kappa\gamma} \right) \right.\right. \\ & \left. + \sqrt{\tfrac{2\pi}{2}} \left(\tfrac{\mathrm{d_y}+2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma} \right) \left(1 + \Phi\left(\tfrac{\mathrm{d_y}+2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma} \right) \right) \right) \\ & + \exp\left\{ \left(\tfrac{\mathrm{d_y}-2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma} \right)^2 \right\} \left(\tfrac{\sqrt{2\pi}}{2}\phi\left(\sqrt{2}\tfrac{(\mathrm{d_y}-2\sigma^2\kappa(1-\gamma))}{1+2\sigma^2\kappa\gamma} \right) \right. \\ & \left. \left. + \sqrt{\tfrac{2\pi}{2}} \left(\tfrac{\mathrm{d_y}-2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma} \right) \Phi\left(\tfrac{\mathrm{d_y}-2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma} \right) \right) \right) \\ & \Big/ \\ & \left(\exp\left\{ \left(\tfrac{1+2\sigma^2\kappa\gamma}{2\sigma^2} \right) \left(\tfrac{\mathrm{d_y}+2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma} \right)^2 \right\} \times \sqrt{\tfrac{2\pi\sigma^2}{1+2\sigma^2\kappa\gamma}} \left(1 - \Phi\left(\sqrt{\tfrac{1+2\sigma^2\kappa\gamma}{\sigma^2}} \tfrac{\mathrm{d_y}+2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma} \right) \right) \right. \\ & \left. + \exp\left\{ \left(\tfrac{1+2\sigma^2\kappa\gamma}{2\sigma^2} \right) \left(\tfrac{\mathrm{d_y}-2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma} \right)^2 \right\} \times \sqrt{\tfrac{2\pi\sigma^2}{1+2\sigma^2\kappa\gamma}} \left(\Phi\left(\sqrt{\tfrac{1+2\sigma^2\kappa\gamma}{\sigma^2}} \tfrac{\mathrm{d_y}-2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma} \right) \right) \right). \end{aligned} \tag{5.62}$$

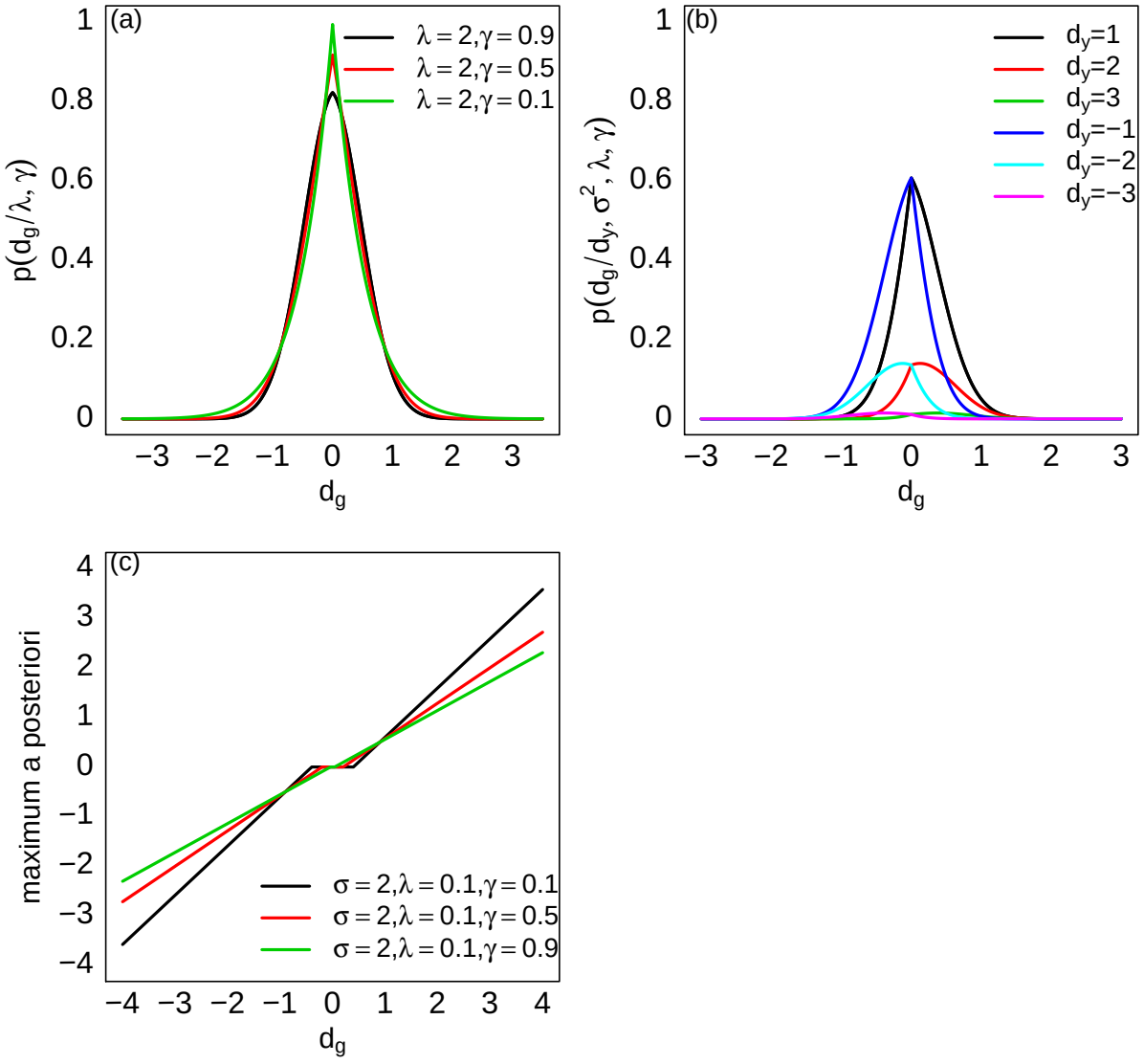


Figure 5.4: Plots of the elastic-net: (a) probability density function, Equation (5.50); (b) posterior mean, Equation (5.62) with $\sigma = 1$, $\kappa = 3$ and $\gamma = 0.5$; and (c) maximum a posteriori, Equation (5.59); as the parameter γ, in (c), decreases, the rule can be explained as thresholding rule.

The proof of (5.62) can be found in Appendix (C). Figure (5.4) shows; (a) the probability density function for different values of κ and γ; (b) the posterior in (5.62) for different wavelet coefficients, in particular for the wavelet coefficients $\mathbf{d_y} = \{-3, -2, -1, 1, 2, 3\}$; and (c) maximum a posteriori of (5.59) for different values of κ and γ. The maximum a posteriori estimator depends on three prior parameters, that is σ^2, κ and γ. The parameter σ^2 is the noise variance, which is estimated by the usual median absolute deviation proposed by Nason and Silverman (1994). The parameters κ and γ will be calculated using the MMSE approach, as described in Section 2.13. Figure 5.5 shows the plot of monitoring the MMSE in Algorithm 2 using the IT-TO methods with a MAP estimator using the elastic-net prior where κ, γ and Λ are estimated, whereas the parameters κ and γ are estimated for each resolution level. We denote this by

$$\widehat{\mathbf{f}}^{\text{IT-TO}}_{\text{Reg}_{\boldsymbol{\lambda},\boldsymbol{\gamma},\Lambda}} = \mathbf{W}^T\widehat{\mathbf{d}}_{\mathbf{f}}, \tag{5.63}$$

where $\widehat{\mathrm{d}}_{\mathrm{f}} = \widehat{\mathrm{T}}(\widehat{\mathrm{d}}_{\mathrm{g}}, \kappa, \gamma)$ is computed from (5.59) and $\widehat{\mathbf{d}}_{\mathbf{g}} = \mathbf{W}(\mathbf{H}^T\mathbf{H} + \Lambda\mathfrak{R}^T\mathfrak{R})^{-1}\mathbf{H}^T\mathbf{y}$, hence, the same MMSE approach, as described in Section 2.13 and $\mathfrak{R}$ is defined in Section 2.6, will be used.

5.7 "Larger posterior mode" method

Cutillo *et al.* (2008) proposed a thresholding rule, which always picks the mode of the posterior, which is the larger mode in absolute value. They assumed that the variance of the noise is known. Then for the model of a two normal distributions given by

$$\mathrm{p}(\mathrm{d_y}|\mathrm{d_g}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(\mathrm{d_y} - \mathrm{d_g})^2 \right\}, \quad \mathrm{d_y}, \mathrm{d_g} \in \mathbb{R}; \sigma > 0, \tag{5.64}$$

with the prior on the wavelet coefficients $\mathrm{d_g}$, defined as

$$\mathrm{p}(\mathrm{d_g}|\tau^2) = \frac{1}{\sqrt{2\pi\tau^2}} \exp\left\{ -\frac{1}{2\tau^2}\mathrm{d_g}^2 \right\}, \quad \mathrm{d_g} \in \mathbb{R}; \tau > 0, \tag{5.65}$$

and prior parameter on the prior variance given by

$$\mathrm{p}(\tau^2) = (\tau^2)^{-c}, \quad c > 0. \tag{5.66}$$

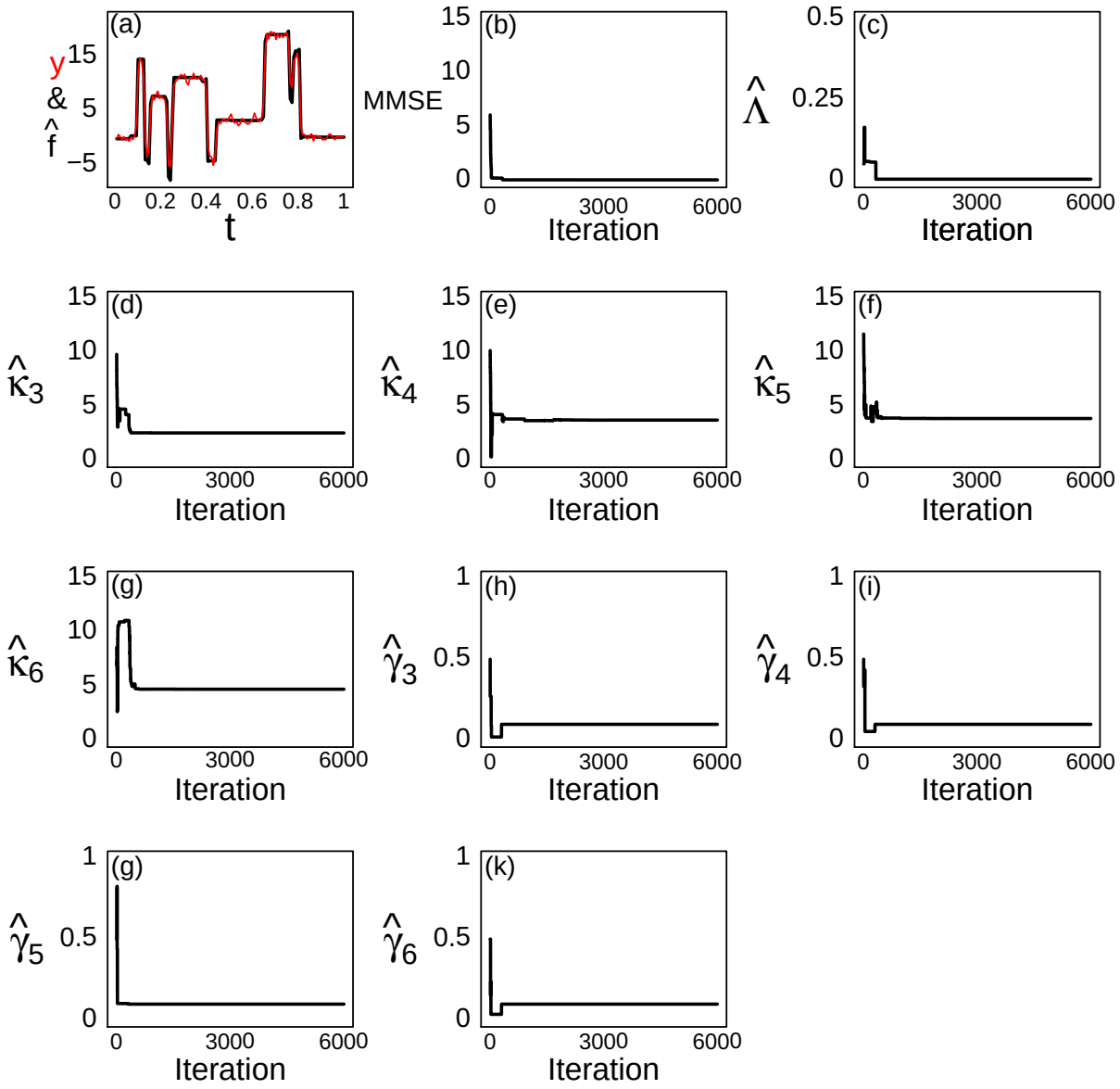


Figure 5.5: Plots of monitoring the minimum MSE algorithm, described in Section 2.13, using the IT-TO method with MAP rule in (5.59), where κ and γ are proposed for each level above 2. The Blocks test function, at $m = 128$ equally spaced points, is used and corrupted by levels of noise and blur, which is given in (2.6), that are equal to 0.5 and 0.005, respectively: (a) the red line represents the true Blocks test function and the black line represents the result of the estimate at a transient period of 6000 iterations; (b) the new value of MMSE is acceptable; (c) acceptable Λ; (d)-(g) acceptable of $\boldsymbol{\kappa}$ and (e)-(k) acceptable of $\boldsymbol{\gamma}$ at resolution levels $j = 3, 4, 5, 6$, respectively.

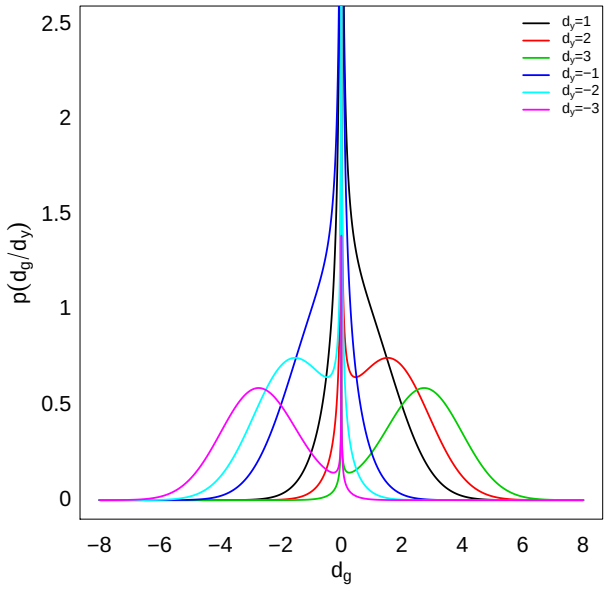


Figure 5.6: Plots of the posterior distribution in (5.68) for $\mathbf{d_y} = \{-3, -2, -1, 1, 2, 3\}$, $c = 3/4$ and $\sigma = 1$.

The joint distribution of $\mathrm{d_y}$ and $\mathrm{d_g}$, is given by

$$\begin{aligned} \mathrm{p}(\mathrm{d_g}, \mathrm{d_y}) &= \int_0^\infty \mathrm{p}(\mathrm{d_y}|\mathrm{d_g})\mathrm{p}(\mathrm{d_g}|\tau^2)\mathrm{p}(\tau^2)d\tau^2 \\ &= \frac{1}{2\pi\sigma}\exp\left\{-\frac{1}{2\sigma^2}(\mathrm{d_y}-\mathrm{d_g})^2\right\}\int_0^\infty \tau^{2-(c+1/2)}\exp\left\{-\frac{1}{2\tau^2}\mathrm{d_g}^2\right\}d\tau^2 \\ &= \frac{1}{2\pi\sigma}\exp\left\{-\frac{1}{2\sigma^2}(\mathrm{d_y}-\mathrm{d_g})^2\right\}\frac{2^{1/2-c}}{(\mathrm{d_g}^2)^{c-1/2}}\Gamma\left(c-\frac{1}{2}\right), \quad c>\frac{1}{2}. \end{aligned} \tag{5.67}$$

Then the posterior density of the wavelet coefficients $\mathrm{d_g}$ given $\mathrm{d_y}$, is given by

$$\mathrm{p}(\mathrm{d_g}|\mathrm{d_y}) \propto \exp\left\{-\frac{1}{2\sigma^2}(\mathrm{d_y}-\mathrm{d_g})^2\right\}|\mathrm{d_g}|^{-2c+1}. \tag{5.68}$$

Figure 5.6 displays the posterior distribution with $c = 3/4$ and $\sigma = 1$, for different values of the wavelet coefficient. For large absolute values of the wavelet coefficient, the posterior is bimodal with a non-zero mode, which has the same sign as the coefficient $\mathrm{d_y}$. For small absolute values of the wavelet coefficient, the posterior is unimodal with an infinite mode at zero (Cutillo *et al.*, 2008).

Then, the logarithm of the posterior density is

$$\log \mathrm{p}(\mathrm{d_g}|\mathrm{d_y}) \propto \frac{1}{2\sigma^2}(\mathrm{d_y}-\mathrm{d_g})^2 + (1-2c)\log \mathrm{d_g}, \tag{5.69}$$

with derivative

$$\mathrm{d_g}^2 - \mathrm{d_y}\mathrm{d_g} + \sigma^2(2c-1) = 0.$$

The solution of this quadratic equation is given by

$$\mathrm{d_g} = \frac{\mathrm{d_y} \pm \sqrt{\mathrm{d_y}^2 - 4\sigma^2(2c-1)}}{2}.$$

The roots exist if and only if $\mathrm{d_y}^2 \geq 4\sigma^2(2c-1)$, then $|\mathrm{d_y}| \geq 2\sigma\sqrt{2c-1} = \kappa$. If this condition is not satisfied, then the likelihood is decreasing in $|\mathrm{d_g}|$. So, the two normal wavelet shrinking rule (NNBWS-MAP) is given by

$$\widehat{\mathrm{d}}_\mathrm{g} = \begin{cases} \frac{\mathrm{d_y}+\sqrt{\mathrm{d_y}^2-4\sigma^2(2c-1)}}{2}, & \text{if } \mathrm{d_y} > \kappa \\ \frac{\mathrm{d_y}-\sqrt{\mathrm{d_y}^2-4\sigma^2(2c-1)}}{2}, & \text{if } \mathrm{d_y} < -\kappa \\ 0, & \text{otherwise.} \end{cases} \tag{5.70}$$

The second model assumes that the variance of the noise σ^2 is unknown. Cutillo *et al.* (2008) suggested that the variance of noise σ^2 is assigned an exponential prior by following Zellner (1996), leading to a double exponential marginal likelihood. The exponential distribution is the entropy maximizer among all distributions supported on $(0, 1)$ with a fixed first moment (Vidakovic, 1998a). Then the marginal likelihood is given by

$$\mathrm{p}(\mathrm{d_y}|\mathrm{d_g}, \mu) = \frac{\sqrt{2\mu}}{2}\exp\left\{-\sqrt{2\mu}|\mathrm{d_y}-\mathrm{d_g}|\right\}, \quad \mathrm{d_y}, \mathrm{d_g} \in \mathbb{R}; \mu > 0; \tag{5.71}$$

this form will be proven in Chapter 6. The prior on the wavelet coefficients $\mathrm{d_g}$, can be defined as

$$\mathrm{p}(\mathrm{d_g}|\tau^2) = \frac{1}{\sqrt{2\pi\tau^2}}\exp\left\{-\frac{\mathrm{d_g}^2}{2\tau^2}\right\}, \quad \mathrm{d_g} \in \mathbb{R}; \tau > 0, \tag{5.72}$$

and the prior parameter on the prior variance given by

$$\mathrm{p}(\tau^2|c) = (\tau^2)^{-c}, \quad c > 0. \tag{5.73}$$

Similarly, the resulting MAP rule, for the Laplace as likelihood and the normal as prior (DENWS-MAP), turns out to be hard thresholding (Cutillo *et al.*, 2008), with

$$\widehat{\mathrm{d}}_{\mathrm{g}} = \begin{cases} \mathrm{d_y}, & \text{if } |\mathrm{d_y}| > \kappa \\ 0, & \text{otherwise,} \end{cases} \tag{5.74}$$

where $\kappa = \frac{2c-1}{\sqrt{2\mu}}$, with $c = 0.75$. If $|\mathrm{d_y}| \leq \kappa$, there is only one solution 0, and when $|\mathrm{d_y}| > \kappa$, the solution is $\mathrm{d_y}$.

5.8 Comparison simulation

The purpose of this section is to evaluate and investigate whether the maximum a posteriori using normal distribution as likelihood and elastic-net as prior (elastic-net-MAP), the maximum a posteriori using double exponential distribution as likelihood and normal distribution as prior with parameter, which is described by prior parameter (DENWS-MAP), the maximum a posteriori of using normal distribution as likelihood and normal distribution as prior with parameter, which is described by prior parameter (NNBWS-MAP), the maximum a posteriori of using normal distribution as likelihood and normal distribution as prior (NNWS-MAP), the maximum a posteriori of using normal distribution as likelihood and double exponential distribution as prior (NDEWS-MAP) and the maximum a posteriori of using normal distribution as likelihood and double Weibull distribution as prior (DWWS-LPM), are suitable for estimating an unknown vector $\mathbf{f}$.

The simulated data sets consisted of the standard test signal Blocks (Donoho and Johnstone, 1994; Nason and Silverman, 1994), multiplied the blur matrix, which is given in (2.6), with $k = 0.005$. Also, it is corrupted by independent Gaussian noise, with mean zero and variance taken as 0.5, no thresholding was done below level 3, the IT-TO method was used, and the value of the parameter $\boldsymbol{\kappa}$ for level-dependent priors are considered.

	MMSE	$\widehat{\Lambda}_{\text{MMSE}}$	$\widehat{\kappa}_{\text{MMSE}}$			
			Level 3	Level 4	Level 5	Level 6
NDEWS-MAP	0.21	0.0002	0.1541	1.0016	0.7426	0.9625
NNWS-MAP	0.539	0.0113	0.0114	0.1239	0.2193	0.0436
DWWS-MAP	0.19	0.007	2.9314	5.0406	6.7344	7.3927
DENWS-MAP	**0.13**	0.006	1.1875	7.5869	1.6186	2.1219
NNBWS-MAP	0.133	0.0121	1.5769	6.5079	1.9297	3.8277

Table 5.1: The results of minimum MSE for estimating the Blocks test function using different priors with hard thresholding rule, where the parameter κ is estimated for each resolution level, $j = 3, 4, 5, 6$.

	MMSE	$\widehat{\Lambda}_{\text{MMSE}}$	$\widehat{\kappa}_{\text{MMSE}}$				$\widehat{\gamma}_{\text{MMSE}}$			
			Level 3	Level 4	Level 5	Level 6	Level 3	Level 4	Level 5	Level 6
Elastic-net-MAP	**0.096**	0.006	0.0319	4.1626	3.92528	11.3981	0.0596	0.0141	0.1313	0.1313

Table 5.2: The results of minimum MSE for estimating the Blocks test function using elastic-net prior with hard thresholding rule, where the parameters κ and γ are estimated for each resolution level, $j = 3, 4, 5, 6$.

Moreover, the first-order method in Section 2.6 is used to estimate $\mathbf{f}$. The number of replications is equal to 60 and the number of iterations equals 100. The MMSE approach, as described in Section 2.13, is used to obtain the prior parameters.

The results are summarized in Tables 5.1 and 5.2, where bold numbers indicate the smallest MSE result for the Blocks test function, where the level of noise is equal to $\sigma^2 = 0.5$ and the level of blur equals $k = 0.005$. It can be see that DENWS-MAP and elastic-net-MAP provide a smaller MSE than NDEWS-MAP, NNWS-MAP, DWWS-LPM and DENWS-MAP.

Figure 5.7 shows the plots of reconstructions using different methods. Figure 5.7 (b) displays the reconstruction obtained from the NDEWS-MAP defined in (5.36). Figure 5.7 (c) displays the result of the reconstruction for the maximum a posteriori estimator defined in (5.20). The Gaussian prior provides a reconstruction, which does not fully recover from noise. However, the sharp edges can be identified. Figure 5.7 (d) displays a

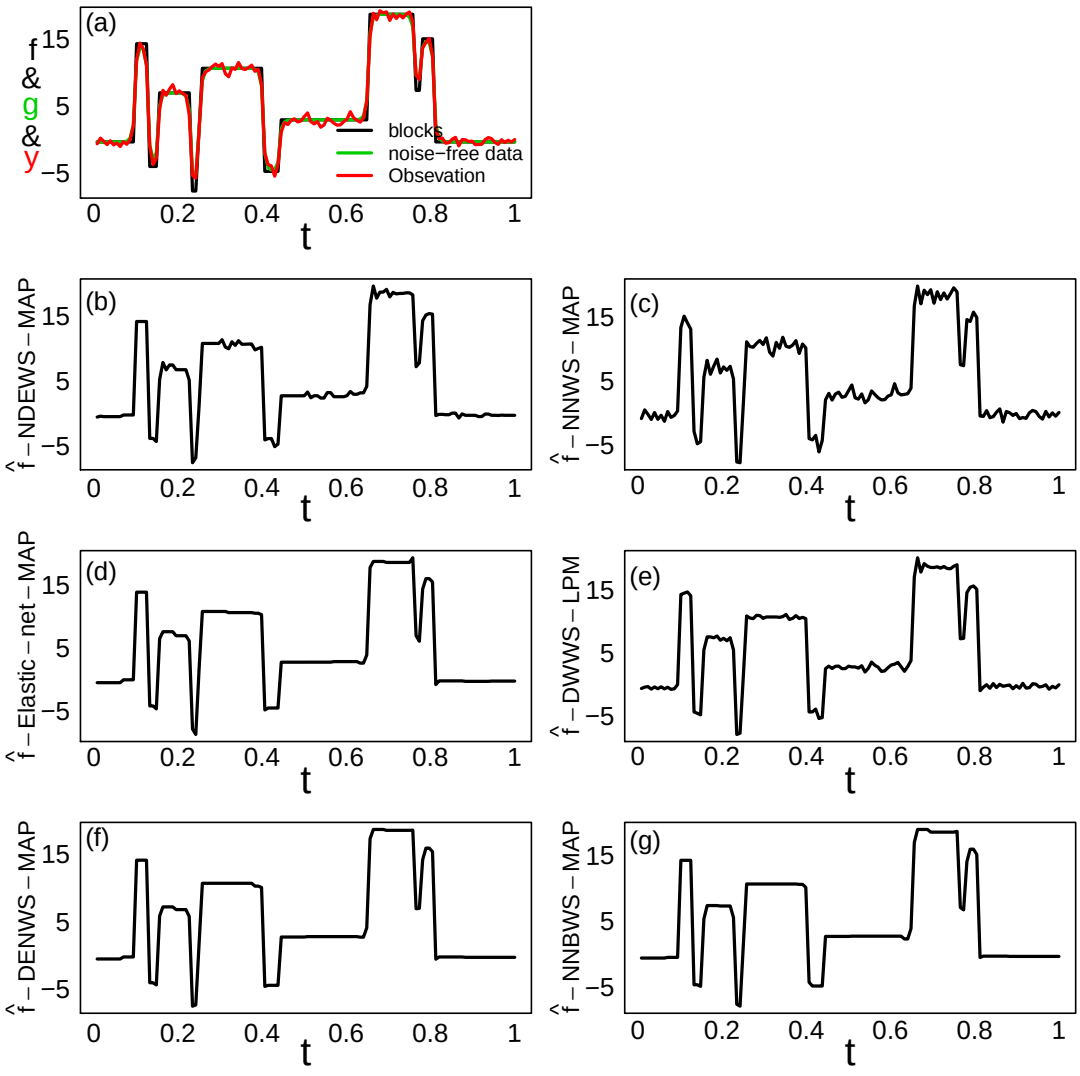


Figure 5.7: Plots of the reconstruction for estimating the unknown vector $\mathbf{f}$, where the parameter κ is estimated for each level, $j = 3, 4, 5, 6$: (a) the black line is the Blocks test function at $m = 128$ equally spaced points, the green line is made by multiplying the blocks by the blur, which is given in (2.6), $k = 0.005$ and the red line shows the observed data with noise $\sigma = 0.5$, the other panels show reconstruction using; (b) NDEWS-MAP; (c) NNWS-MAP; (d) elastic-net-MAP; (e) DWWS-LPM; (f) DENWS-MAP and (g) NNBWS-MAP.

reconstructed profile using the maximum a posteriori estimator and the elastic-net as the prior. It provides sharp edges and flat topped reconstructions. Additionally, it provides a smaller MSE than DENWS-MAP, NDEWS-MAP, NNWS-MAP and DWWS-LPM. Figure 5.7 (e) displays a reconstructed profile using $\mathcal{DW}$ as the prior, with the features of the reconstruction identified. Figure 5.7 (f) displays the result of the reconstruction for the maximum a posteriori estimator in (5.74), with the shape of the blocks are resolved. Note that Cutillo *et al.* (2008) showed that there are no differences between the hard thresholding rule and DENWS-MAP. Figure 5.7 (g) displays the result of the reconstruction for the maximum a posteriori estimator in (5.70). Also, the NNBWS-MAP thresholding provides sharp edges and flat topped reconstructions and the shape of the blocks are more clear than DWWS-MAP and NNWS-MAP and NDEWS-MAP.

5.9 Conclusions

Within this chapter, Bayesian thresholding, using non-mixture priors, was investigated and applied in the wavelet domain. Different priors were investigated, such as the Laplace, the Gaussian, the double Weibull and the elastic-net distributions. The variance of the noise σ^2 is assumed as unknown and assigned an exponential prior, leading to a double exponential marginal likelihood. So, there are two types of likelihoods that are considered. The first is the Gaussian when the variance of noise is known and the second is the double exponential when the variance of noise is unknown and assigned an exponential prior. Bayesian thresholding was applied to estimate an underlying function, $\mathbf{f}$. The choice of the elastic-net provides sharp edges and flat topped reconstructions. The elastic-net-MAP, DENWS-MAP and NNBWS-MAP methods provide better block shapes than the NNWS-MAP, NDEWS-MAP and the DWWS-LPM. It can be concluded that elastic-net-MAP gives an excellent reconstruction.

Chapter 6

Bayesian thresholding using mixture priors

6.1 Overview

Within this chapter, Section 6.2 gives an introduction, while Section 6.3 provides detail about adaptive Bayesian wavelet shrinkage, and Section 6.4 is about Bayesian adaptive multi-resolution shrinkage. Section 6.5 then discusses the BayesThresh method, Section 6.6 gives an empirical Bayes method, and finally Section 6.7 presents conclusions.

6.2 Introduction

There are many wavelet-based mixture priors suggested for wavelet coefficients. For example, Chipman *et al.* (1997) proposed Bayesian adaptive multi-resolution shrinkage and Vidakovic and Ruggeri (2001) introduced adaptive Bayesian wavelet shrinkage. The traditional Bayesian models consider a prior distribution on the wavelet coefficient $\mathrm{d_g}$, is given by

$$\pi(\mathrm{d_g}) = \gamma\delta(\mathrm{d_g} = 0) + (1-\gamma)\zeta(\mathrm{d_g}), \tag{6.1}$$

where δ is a point mass at zero, and ζ represents a heavy-tailed density, which is symmetric about zero. This type of model was considered by Mallat (1989), Abramovich *et al.* (1998), Vidakovic and Ruggeri (2001), Barber *et al.* (2002) and among others. Their reason can be summarized as thus: "For most of the signals and images encountered in practice, the empirical distribution of a typical detail wavelet coefficient is notably centered about zero and peaked at it." In other words, the above prior distribution was designed to capture the sparseness of the wavelet transform due to the empirical distribution for wavelet coefficients being centered on zero and peaked at it (Abramovich *et al.*, 1998). In addition, the parameter, γ $(0 < \gamma < 1)$, represents the probability of the wavelet coefficient $\mathrm{d_g}$ being exactly zero. This means that if γ is large then the wavelet coefficient is likely to be zero. For example, the finest-scale level should have $\gamma \approx 1$ so that most of the wavelet coefficients in that level are zero. More precisely, if the wavelet coefficient, $\mathrm{d_g}$, equals zero then it is modelled by the first part, $\delta(\mathrm{d_g} = 0)$ in (6.1), whilst non-zero wavelet coefficients are described by the second part. Each wavelet coefficient is either 0 with probability γ, or probability $1 - \gamma$ distributed as ζ.

Now, considering the model

$$\mathbf{y} = \mathbf{Hf} + \boldsymbol{\epsilon}, \tag{6.2}$$

where $\mathbf{H}$ is a given $n \times m$ blur matrix, $\mathbf{y}_{n\times 1}$ and $\boldsymbol{\epsilon}$, is a vector of random variables, such that $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$. Then

$$\mathbf{d_y} = \mathbf{Wy} = \mathbf{W}(\mathbf{Hf} + \boldsymbol{\epsilon}) = \mathbf{WHf} + \mathbf{W}\boldsymbol{\epsilon} = \mathbf{d_g} + \boldsymbol{\eta}, \tag{6.3}$$

where $\mathbf{d_g}_{n\times 1}$ is a vector of wavelet coefficients containing the wavelet coefficients of $\mathbf{g} = \mathbf{Hf}$ and $\mathbf{f}_{m\times 1}$ is our signal of interest. Thus, the model in (6.2) can be written equivalently as

$$\mathbf{d_y} = \mathbf{d_g} + \boldsymbol{\eta}, \tag{6.4}$$

where $\mathbf{d_y} = \mathbf{Wy}$ and $\mathbf{d_g} = \mathbf{Wg}$.

The procedure of estimating $\mathbf{d_g}$ from $\mathbf{d_y}$ is now considered. The posterior, including a prior distribution on the wavelet coefficient $\mathbf{d_g}$, is given by

$$\mathrm{p}(\mathbf{d_g}|\mathbf{d_y}) = \frac{\mathrm{p}(\mathbf{d_y}|\mathbf{d_g})\mathrm{p}(\mathbf{d_g})}{\mathrm{p}(\mathbf{d_y})}, \tag{6.5}$$

where $\mathbf{d_g}$ is the vector of model parameters, $\mathrm{p}(\mathbf{d_y}|\mathbf{d_g})$ is the likelihood function and $\mathrm{p}(\mathbf{d_g})$ is the prior distribution. The posterior can be written as

$$\mathrm{p}(\mathbf{d_g}|\mathbf{d_y}) \propto \mathrm{p}(\mathbf{d_y}|\mathbf{d_g})\mathrm{p}(\mathbf{d_g}),$$

because the normalising constant has no information about the unknown parameters.

In this chapter, the best method in Chapter 2, which was the IT-TO method, will be used, the posterior of estimate, $\mathbf{d_f}$, is computed then the estimate of $\mathbf{f}$, is given by

$$\begin{aligned}\widehat{\mathbf{f}}^{\text{IT-TO}}_{\text{Reg}_\Lambda} &= \mathbf{W}^T\mathrm{p}(\mathbf{d_f}|\widehat{\mathbf{d}}_\mathbf{g}) \\ &= \mathbf{W}^T\mathrm{p}(\mathbf{d_f}|(\mathbf{H}^T\mathbf{H} + \Lambda\mathfrak{R}^T\mathfrak{R})^{-1}\mathbf{H}^T\mathbf{y}), \end{aligned} \tag{6.6}$$

where $\widehat{\mathbf{d}}_\mathbf{g} = (\mathbf{H}^T\mathbf{H} + \Lambda\mathfrak{R}^T\mathfrak{R})^{-1}\mathbf{H}^T\mathbf{y}$, where $\Lambda = \sigma^2\kappa$, $\mathfrak{R}$ was defined in Section 2.6 and $\mathbf{W}$ was defined in Section 2.9. Posterior mean and posterior median will be applied to estimate an unknown vector $\mathbf{f}$.

In this chapter, the posterior mean using adaptive Bayesian wavelet shrinkage of Vidakovic and Ruggeri (2001) and the Bayesian adaptive multi-resolution shrinkage of Chipman *et al.* (1997) are used to shrink the wavelet coefficients of $\mathbf{d_g}$. Also the posterior median using BayesThresh from Abramovich *et al.* (1998) and the empirical Bayes approach of Johnstone and Silverman (2005a) are used to threshold the wavelet coefficients of $\mathbf{d_g}$. All these methods will be studied and investigated to estimate an underlying function and then the best method will used to build a reconstruction from the real data in Chapter 9.

Chipman *et al.* (1997) used a prior with two normal distributions. The first term is normal with small variance, $N(0, \tau_j^2)$, and it can replace a point mass at zero. The second term, $N(0, c^2\tau_j^2)$, is a spread distribution that models wavelet coefficients with large values. So, the factor $c \gg 1$ determines whether the wavelet coefficients are non-zero, arising from the $N(0, c^2\tau_j^2)$ distribution, or close to zero, arising from the $N(0, \tau_j^2)$ distribution. Also, the part $N(0, c^2\tau_j^2)$ is intended to describe a large wavelet coefficient, while $N(0, \tau_j^2)$ is intended to depict a small coefficient. This prior is discussed by Clyde *et al.* (1998) and Vidakovic (1998a).

The Bayesian adaptive multi-resolution shrinkage is proposed by Vidakovic and Ruggeri (2001). The author assumed the likelihood as an additive Gaussian error model, so that the wavelet coefficients from data, $\mathrm{d_y}$, are conditionally independent given the wavelet coefficients of the truth, $\mathrm{d_g}$, i.e. $\mathrm{d_y}|\mathrm{d_g}, \sigma^2 \sim N(\mathrm{d_g}, \sigma^2)$. The variance of noise σ^2 is assigned an exponential prior, so that the marginal likelihood of the wavelet coefficient is a double exponential distribution, $\mathcal{DE}$, with parameter $\sqrt{2\mu}$. The choice of the exponential prior can be additionally justified by its maxent property. The exponential distribution was first proposed by Zellner (1996) and is the entropy maximiser in the class of all distributions supported on $(0, \infty)$ with a fixed first moment (Ruggeri and Vidakovic, 2005). In this method, the $\mathcal{DE}$ is used as prior on the wavelet coefficients of $\mathrm{d_g}$; this choice of $\mathcal{DE}$ for the wavelet coefficients is a realistic model, because it describes the behavior of wavelet coefficients around zero, and it will account for heavy tails encountered in empirical distributions of wavelet coefficients (Cutillo *et al.*, 2008).

Abramovich *et al.* (1998) summarized that the traditional Bayes rule (posterior mean) corresponds to an $\mathbb{L}_2$-loss based on the wavelet coefficients. However, such a rule is not a thresholding rule but a shrinkage. For a rule to be a thresholding rule, it must not only shrink the wavelet coefficient towards zero but must also map actually to zero all the wavelet coefficients falling into an interval around zero. Thus, they suggest to use the posterior median in the context of wavelet shrinkage. Their method is known as *BayesThresh* and could be a thresholding rule, which is preferable to smooth shrinkage rules in many applications, such as model selection and data compression. The idea was developed, simulated and studied by Barber (2001) and Barber *et al.* (2002).

For other early examples of the Bayesian approach to wavelet regression see papers, such as Johnstone and Silverman (2005a,b) who presented the empirical Bayes methods for wavelet shrinkage. The parameters of the model are estimated by marginal maximum likelihood; therefore, the authors use the data to estimate parameters. Different level-dependent priors are considered, all of which are a mixture of point mass at zero and a heavy-tailed density. The two choices for the heavy-tailed density are the Cauchy and the Laplace priors.

6.3 Adaptive Bayesian wavelet shrinkage

Adaptive Bayesian wavelet shrinkage (ABWS) was proposed by Chipman *et al.* (1997). Consider an additive Gaussian error model for the conditional distribution of the data given the truth, so that, the distribution of the wavelet coefficient $\mathrm{d_y}$ given the corresponding wavelet coefficient $\mathrm{d_g}$, is given by

$$\mathrm{p}(\mathrm{d_y}|\mathrm{d_g}, \sigma^2) \sim N(\mathrm{d_g}, \sigma^2), \tag{6.7}$$

where σ^2 is noise variance, which is estimated by the usual median absolute deviation by Nason and Silverman (1994). Independent prior distributions on wavelet coefficients $\mathrm{d_g}$ are each defined as a mixture of two normal distributions

$$\mathrm{p}(\mathrm{d_g}|\gamma, c^2, \tau_j^2) \sim \gamma N(0, c^2\tau_j^2) + (1-\gamma)N(0, \tau_j^2), \tag{6.8}$$

with

$$\mathrm{p}(\gamma = 1) = 1 - \mathrm{p}(\gamma = 1) \equiv \mathrm{p}_j. \tag{6.9}$$

The p_j, c and τ_j are prior parameters to be chosen. Hence, the prior parameters p_j and τ_j depend on the resolution level j. Chipman *et al.* (1997) showed that each wavelet coefficient in resolution level j either follows the normal distribution, with mean zero and variance $c^2\tau_j^2$ or with probability $1-\gamma$ follows the normal distribution, with mean zero and variance τ_j^2. So, the factor c determines the variance of the first part, which means that c makes the normal distribution either narrow or wide. The parameter γ_j is the proportion of wavelet coefficients which are expected to be non-negligible at resolution level j (Ruggeri and Vidakovic, 2005). The posterior mean of wavelet coefficient $\mathrm{d_g}$ given the corresponding wavelet coefficient $\mathrm{d_y}$, is given by minimisation of the squared-error loss and has an explicit form,

$$\mathrm{PM}(\mathrm{d_g}|\mathrm{d_y}) = \left[\mathrm{p}(\gamma = 1|\mathrm{d_y})\frac{c^2\tau_j^2}{\sigma^2 + c^2\tau_j^2} + \mathrm{p}(\gamma = 0|\mathrm{d_y})\frac{\tau_j^2}{\sigma^2 + \tau_j^2}\right]\mathrm{d_y}, \tag{6.10}$$

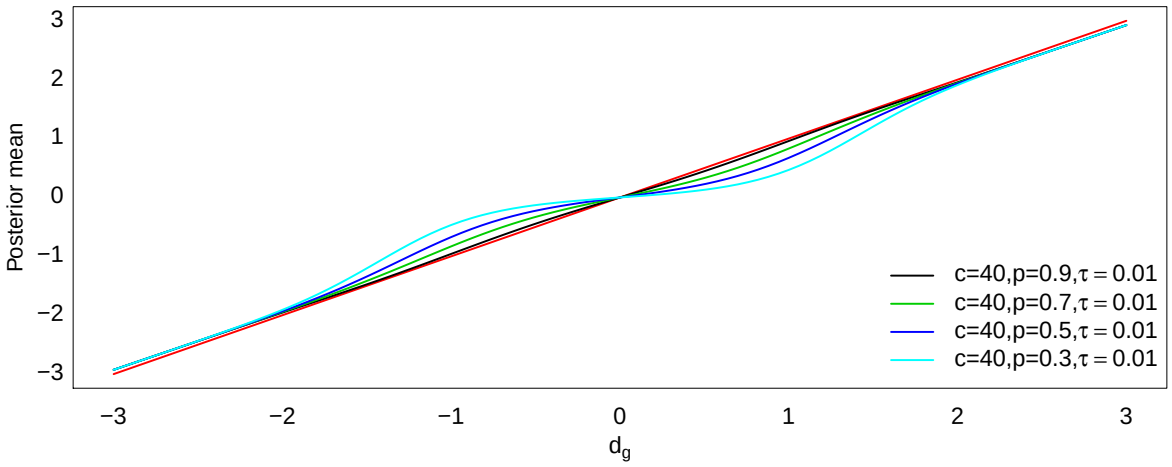


Figure 6.1: Plots of the posterior mean in (6.10) with different values of p, τ and c.

where,

$$\mathrm{p}(\gamma = 1|\mathrm{d_y}) = \frac{\frac{\mathrm{p}_j \mathrm{p}(\mathrm{d_y}|\gamma=1)}{(1-\mathrm{p}_j)\mathrm{p}(\mathrm{d_y}|\gamma=0)}}{\frac{\mathrm{p}_j \mathrm{p}(\mathrm{d_y}|\gamma_j=1)}{(1-\mathrm{p}_j)\mathrm{p}(\mathrm{d_y}|\gamma=0)} + 1}, \tag{6.11}$$

and,

$$\mathrm{p}(\gamma = 0|\mathrm{d_y}) = \frac{1}{\frac{\mathrm{p}_j \mathrm{p}(\mathrm{d_y}|\gamma=1)}{(1-\mathrm{p}_j)\mathrm{p}(\mathrm{d_y}|\gamma=0)} + 1}, \tag{6.12}$$

where $\mathrm{p}(\mathrm{d_y}|\gamma = 1) \sim N(0, \sigma^2 + c^2\tau_j^2)$ and $\mathrm{p}(\mathrm{d_y}|\gamma = 0) \sim N(0, \sigma^2 + \tau_j^2)$. The posterior mean of ABWS in (6.10) is plotted in Figure 6.1 for different values of γ, τ and c. The rule is described as heavily shrinking small coefficients in magnitude. Thus, the rule is a shrinkage rule. The first term in (6.8) is intended to explain the large wavelet coefficients, while $N(0, \tau_j^2)$ is intended to depict the small coefficients. The probability, p_j, gives the proportion of non-zero wavelet coefficients at resolution level j. More precisely, when the wavelet coefficient is small, this suggests that $\mathrm{d_g}$ is small and $\mathrm{p}(\gamma = 0|\mathrm{d_y})$ is large. Thus, the shrinkage approximately follows a straight line with the intercept at zero and the

slope $\frac{\tau_j^2}{\sigma^2+\tau_j^2}$. So, relatively small values of τ_j give the flat portion of the shrinkage function around zero. On the other hand, if the wavelet coefficient is large, this suggests that $\mathrm{d_g}$ is large and $\mathrm{p}(\gamma = 1|\mathrm{d_y})$ is large. Thus, the shrinkage approximately follows a straight line with the intercept at zero and the slope $\frac{c^2\tau_j^2}{\sigma^2+c^2\tau_j^2}$. The parameters c and τ_j determine the slopes of the two lines. Small values of p_j will increase the width of the interval about zero, where the shrinkage function clings to the line with the smaller slope. Given τ_j and p_j, increasing c will shorten the interval, in which the shrinkage function climbs from the line with the smaller slope up to the line with the larger slope. This is because c controls the two alternative components of the mixture (Chipman *et al.*, 1997).

6.4 Bayesian adaptive multi-resolution shrinkage

The Bayesian adaptive multi-resolution shrinkage (BAMS) method was introduced by Vidakovic and Ruggeri (2001). The variance of noise σ^2 is assigned an exponential prior $\sigma^2 \sim \mathcal{E}(\mu), \mu > 0$, to estimate the variance in the likelihood. This means that they assumed the variance of the noise is unknown and will be modelled by an exponential prior, leading to a Laplace marginal likelihood. Using an exponential distribution was first proposed by Zellner (1996) and is the entropy maximiser in the class of all distributions supported on $(0, \infty)$ with a fixed first moment (Ruggeri and Vidakovic, 2005). The reason for choosing a Laplace marginal likelihood is that it is a realistic model for wavelet coefficients. Indeed, if a histogram of wavelet coefficients for a signal is plotted, it resembles the Laplace distribution (Cutillo *et al.*, 2008). Thus, the marginal likelihood is given by

$$\mathrm{p}(\mathrm{d_y}|\mathrm{d_g}, \mu) = \frac{\sqrt{2\mu}}{2} \exp\left\{ -\sqrt{2\mu}|\mathrm{d_y} - \mathrm{d_g}| \right\}, \quad \mathrm{d_y}, \mathrm{d_g} \in \mathbb{R}; \mu > 0, \tag{6.13}$$

where the result in (6.13) was proven by Andrews and Mallows (1974) and Jeffrey and Zwillinger (2007); let $b = \frac{(\mathrm{d_y}-\mathrm{d_g})^2}{2}$, then

$$\begin{aligned}
\mathrm{p}(\mathrm{d_y}|\mathrm{d_g}) &= \int_0^\infty \mathrm{p}(\mathrm{d_y}|\mathrm{d_g},\sigma^2)\mathrm{p}(\sigma^2)d\sigma^2 \\
&= \int_0^\infty \frac{\mu}{\sqrt{2\pi\sigma^2}} \times \exp\left\{-\frac{1}{2\sigma^2}(\mathrm{d_y}-\mathrm{d_g})^2\right\} \times \exp\left\{-\mu\sigma^2\right\}d\sigma^2 \\
&= \frac{\mu}{\sqrt{2\pi}}\int_0^\infty \frac{1}{\sigma} \times \exp\left\{-(\frac{b}{\sigma^2}+\mu\sigma^2)\right\}d\sigma^2 = \frac{\sqrt{2}\mu}{\sqrt{\pi}}\int_0^\infty \exp\left\{-(\frac{b}{\sigma^2}+\mu\sigma^2)\right\}d\sigma^2 \\
&= \frac{\sqrt{2\mu}}{2}\exp\left\{-\frac{2\sqrt{(\mathrm{d_y}-\mathrm{d_g})^2}\sqrt{\mu}}{\sqrt{2}}\right\} = \frac{\sqrt{2\mu}}{2}\exp\left\{-\sqrt{2\mu}|d-\mathrm{d_g}|\right\}. \qquad (6.14)
\end{aligned}$$

The posterior mean under the prior $\zeta(\mathrm{d_g})$, is given by

$$\mathrm{PM}_\zeta(\mathrm{d_g}|\mathrm{d_y}) = \frac{\int \mathrm{d_g}\mathrm{p}(\mathrm{d_y}|\mathrm{d_g})\zeta(\mathrm{d_g})d\mathrm{d_g}}{\int \mathrm{p}(\mathrm{d_y}|\mathrm{d_g})\zeta(\mathrm{d_g})d\mathrm{d_g}}. \qquad (6.15)$$

Vidakovic and Ruggeri (2001) showed that the posterior mean of the marginal distribution for wavelet coefficients under the prior $\zeta(\mathrm{d_g})$, is given by

$$\begin{aligned}
\mathrm{m}_\zeta(\mathrm{d_y}) &= \int \mathcal{DE}(\mathrm{d_g},\frac{1}{\sqrt{2\mu}})\mathcal{DE}(0,\tau)d\mathrm{d_g} \\
&= \frac{\tau\exp\left\{-\frac{|\mathrm{d_y}|}{\tau}\right\} - \frac{1}{\sqrt{2\mu}}\exp\left\{-|\mathrm{d_y}|\sqrt{2\mu}\right\}}{2\tau^2-\frac{1}{\mu}}, \qquad \tau \neq \frac{1}{\sqrt{2\mu}}, \qquad (6.16)
\end{aligned}$$

where μ is the reciprocal of the mean for the prior (exponential) distribution on the variance, σ^2 (Vidakovic and Ruggeri, 2001). The marginal distribution corresponding to the model in (6.16) exhibits heavier tails, and is more peaked than the normal density. Vidakovic and Ruggeri (2001) proved that the posterior mean is given by

$$\mathrm{PM}_\zeta(\mathrm{d_g}|\mathrm{d_y}) = \frac{\tau\mathrm{d_y}(\tau^2-\frac{1}{2\mu})\exp\left\{-\frac{|\mathrm{d_y}|}{\tau}\right\} - \frac{\tau^2\left(\exp\left\{-|\mathrm{d_y}|\sqrt{2\mu}\right\}-\exp\left\{-\frac{|\mathrm{d_y}|}{\tau}\right\}\right)}{\mu}}{(\tau^2-\frac{1}{2\mu})\left(\tau\exp\left\{-\frac{|\mathrm{d_y}|}{\tau}\right\} - (\frac{1}{\sqrt{2\mu}})\exp\left\{-|\mathrm{d_y}|\sqrt{2\mu}\right\}\right)}. \qquad (6.17)$$

The marginal distribution under the prior in (6.1) is

$$\mathrm{m}_\pi(\mathrm{d_y}) = \gamma\int \mathcal{DE}(0,\frac{1}{\sqrt{2\mu}})d\mathrm{d_g} + (1-\gamma)\int \mathcal{DE}(\mathrm{d_g},\frac{1}{\sqrt{2\mu}})\mathcal{DE}(0,\tau)d\mathrm{d_g}, \qquad (6.18)$$

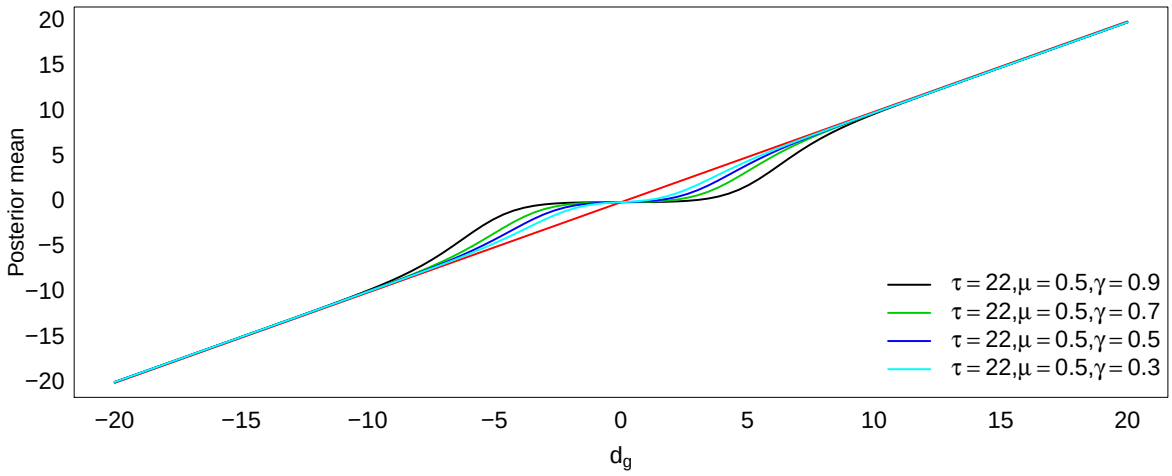


Figure 6.2: Plots of the posterior mean in (6.20) with different values of γ, τ and μ.

where the component $\zeta(\mathrm{d_g})$ reflects the belief that the non-zero wavelet coefficients are exponentially distributed. The prior in (6.1) leads to a rule which is a smooth approximation to a thresholding (Ruggeri and Vidakovic, 2005). Thus,

$$\mathrm{m}_\pi(\mathrm{d_g}) = \gamma \mathcal{DE}(0, \frac{1}{\sqrt{2\mu}}) + (1-\gamma)\mathrm{m}_\zeta(\mathrm{d_y}). \tag{6.19}$$

So, the posterior mean under the prior in (6.1) is given by

$$\mathrm{PM}_\pi(\mathrm{d_g}|\mathrm{d_y}) = \frac{(1-\gamma)\mathrm{m}_\zeta(\mathrm{d_y})\mathrm{PM}_\zeta(\mathrm{d_g}|\mathrm{d_y})}{\mathrm{m}_\pi(\mathrm{d_y})}, \qquad \tau \neq \frac{1}{\sqrt{2\mu}}. \tag{6.20}$$

The posterior mean of BAMS in (6.20) is plotted in Figure 6.3 with different values of γ, τ and c. The rule is a shrinkage rule. Also, the rule is described as heavily shrinking small coefficients. The proof of (6.20) can be found in Appendix (D).

The posterior mean in (6.20) contains three parameters, which are clearly defined. For more detail see Vidakovic and Ruggeri (2001).

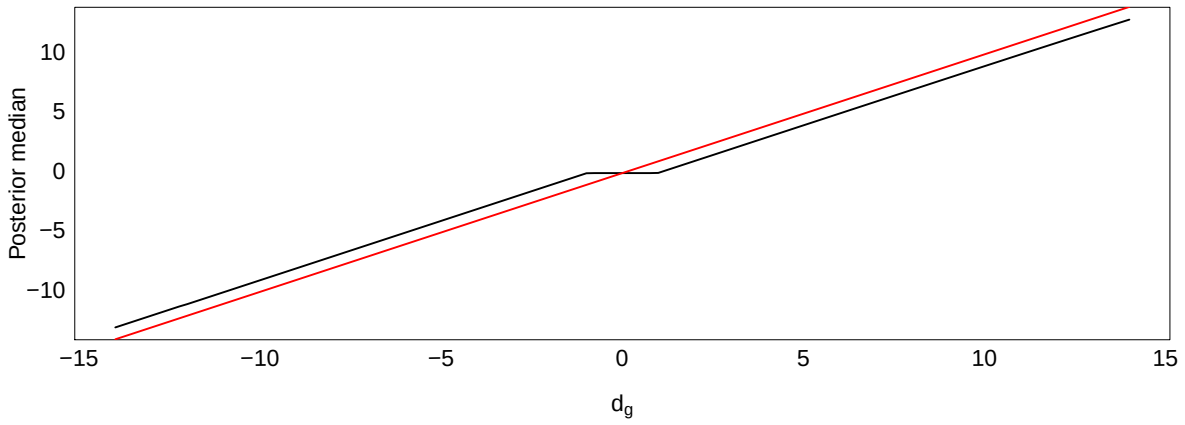


Figure 6.3: Plot of the posterior median in (6.24) with values of $\beta = 1$ and $\alpha = 0.5$.

6.5 BayesThresh

A common prior for wavelet coefficients $\mathbf{d_g}$ is proposed by Abramovich *et al.* (1998). The traditional Bayesian models as the prior distribution on the wavelet coefficient $\mathrm{d_g}$, is given by

$$\mathrm{d_g} \sim \mathrm{p}_j N(0, \tau_j^2) + (1 - \mathrm{p}_j)\delta(0), \tag{6.21}$$

where $\mathrm{p}_j \in \{0, 1\}$and $\delta(0)$ is point mass at zero. The binary random variable, p_j, determines whether the relevant coefficient is zero. When $\mathrm{p}_j = 0$, this implies that the wavelet coefficient arises from a point mass at zero. When $\mathrm{p}_j = 1$, this implies that the wavelet coefficient comes from $N(0, \tau_j^2)$. This prior is a limiting case of the ABMS prior (6.8). The parameters in the prior are defined to be $\tau_j^2 = 2^{-\alpha j} C_1$ and $\mathrm{p}_j = \min\{1, 2^{-\beta j} C_2\}$, where C_1 and C_2 are non-negative constants chosen empirically from the data, α and β being selected by the user (Abramovich *et al.*, 1998). The authors showed that the default choice $\alpha = 0.5$ and $\beta = 1$ is robust to varying degrees of smoothness. The posterior cumulative distribution for $\mathrm{d_g}$, given the observed value of $\mathrm{d_y}$, is given by

$$\mathrm{p}(\mathrm{d_g}|\mathrm{d_y}) = \frac{1}{1+\omega}\Phi\left\{\frac{\mathrm{d_g} - \mathrm{d_y}\tau_j^2/(\sigma^2 + \tau_j^2)}{\sigma\tau_j/\sqrt{(\sigma^2 + \tau_j^2)}}\right\} + \frac{\omega}{1-\omega}\mathbb{I}(\mathrm{d_g} \geq 0), \tag{6.22}$$

where $\mathbb{I}(.)$ is the indicator function and Φ is the standard normal cumulative distribution, and

$$\omega = \frac{1 - \mathrm{p}_j}{\mathrm{p}_j} \frac{\sqrt{\tau_j^2 + \sigma^2}}{\sigma} \exp\left\{ - \frac{\tau_j^2 d_y}{2\sigma^2(\tau_j^2 + \sigma^2)} \right\}. \tag{6.23}$$

Abramovich *et al.* (1998) used the posterior median of (6.22) as the point estimate for $\mathrm{d_g}$. The posterior cumulative distribution function in (6.22) and (6.23), has a jump at 0. The posterior median is 0 if $\omega \geq 1$ and also if $\omega < 1$ and

$$0.5(1 - \omega) \leq \Phi\left\{ - \frac{\mathrm{d_y}\tau_j}{\sigma\sqrt{\sigma^2 + \tau_j^2}} \right\} \leq 0.5(1 + \omega).$$

It is non-zero otherwise. The thresholding procedure can be written as

$$\mathrm{Med}(\mathrm{d_g}|\mathrm{d_y}) = \mathrm{sgn}(\mathrm{d_y})\max(0, o), \tag{6.24}$$

where

$$o = \frac{\tau_j^2}{\sigma^2 + \tau_j^2}|\mathrm{d_y}| - \frac{\tau_j \sigma}{(\sigma^2 + \tau_j^2)^{1/2}} \Phi^{-1}\left\{ \frac{1 + \min(\omega, 1)}{2} \right\}, \tag{6.25}$$

and Φ is the cumulative distribution function of a standard normal distribution. If the quantity o in (6.25) is negative, then $\mathrm{Med}(\mathrm{d_g}|\mathrm{d_y})$ is zero. More precisely, if a wavelet coefficient $\mathrm{d_y}$ falls in the interval $[-\lambda_j, \lambda_j]$, where λ_j is the value of threshold, which is described in (2.61), the estimate of the wavelet coefficient can be set to zero, $\mathrm{Med}(\mathrm{d_g}|\mathrm{d_y}) = 0$. For large wavelet coefficients, the BayesThresh method asymptotes to linear shrinkage by a factor of $\frac{\tau_j^2}{\sigma^2 + \tau_j^2}$, because the second part in form (6.25) become negligible as $|\mathrm{d_y}| \to \infty$ (Abramovich *et al.*, 1998). Thus, Figure (6.3) shows that the BayesThresh method in (6.24) may be described as slightly shrinking large coefficients and heavily thresholding small coefficients. The posterior median is therefore a level-dependent "kill" or shrinkage.

The BayesThresh method is implemented in the `WaveThresh` package for R (Nason, 2010b).

6.6 Empirical Bayes approach

Johnstone and Silverman (2005a) proposed a class of empirical Bayes (EB) methods for wavelet shrinkage. The parameters of the model are estimated by marginal maximum likelihood and the name "empirical Bayes" means that the parameters are estimated using the data, no prior information being used. They consider different level-dependent priors, all of which are mixtures of a point mass at zero and a heavy-tailed density. There are two choices for the heavy-tailed density considered, which are the Laplace and the Cauchy distributions. The model is given by

$$\mathbf{d_y} = \mathbf{d_g} + \boldsymbol{\eta}. \tag{6.26}$$

The orthogonality of the matrix of $\mathbf{W}$ and the normality of the noise vector $\boldsymbol{\epsilon}$ implies the noise vector $\boldsymbol{\eta}$ is also normal, as described in Section 2.9. The prior distribution of the parameter $\mathrm{d_g}$ is an independent prior distribution given by the mixture

$$\pi(\mathrm{d_g}) = (1-\omega)\delta_0(\mathrm{d_g}=0) + \omega\zeta(\mathrm{d_g}), \tag{6.27}$$

where $\delta(0)$ is point mass at zero and the non-zero part of the prior, ζ, is assumed to be a fixed unimodal symmetric density. The first term in (6.27) is intended to explain the small wavelet coefficients, while the second is intended to depict the large wavelet coefficients. By using the Laplace distribution, $\gamma_a(\mathrm{d_g}) = \frac{a}{2}\exp\{-a|\mathrm{d_g}|\}$, with the scale parameter $a > 0$, the marginal distribution for the wavelet coefficients $\mathrm{d_y}$, is given by

$$\mathrm{m}_\pi(\mathrm{d_y}) = (1-\omega)\varphi(\mathrm{d_y}) + \omega g(\mathrm{d_y}), \tag{6.28}$$

where φ denotes the standard normal density and

$$\begin{aligned} g(\mathrm{d_y}) &= \int_{-\infty}^{\infty} \gamma(\mathrm{d_g})_a \varphi(\mathrm{d_y}) d\mathrm{d_g} \\ &= \frac{a}{2}\exp\{\frac{a^2}{2}\}\Big[\exp\{-a\mathrm{d_y}\}\Phi(\mathrm{d_y}-a) + \exp\{a\mathrm{d_y}\}\bar{\Phi}(\mathrm{d_y}+a)\Big]. \end{aligned} \tag{6.29}$$

In the above equation $\Phi(\cdot)$ denotes the cumulative distribution of the standard normal and $\bar{\Phi}(\cdot) = 1 - \Phi(\cdot)$. The posterior distribution of $\mathrm{d_g}$, is given by

$$\mathrm{p}(\mathrm{d_g}|\mathrm{d_y}) = (1-\omega_{post})\delta_0(\mathrm{d_g}) + \omega_{post} p_1(\mathrm{d_g}|\mathrm{d_y}), \tag{6.30}$$

where the posterior probability ω_{post}, is given by

$$\omega_{post}(\mathrm{d_y}) = \frac{\omega g(\mathrm{d_y})}{[\omega g(\mathrm{d_y}) + (1-\omega)\phi(\mathrm{d_y})]}, \tag{6.31}$$

and

$$p_1(\mathrm{d_g}|\mathrm{d_y}) = \begin{cases} \frac{\exp\{a\mathrm{d_y}\}\phi(\mathrm{d_g}-\mathrm{d_y}-a)}{\exp\{-a\mathrm{d_y}\}\Phi(\mathrm{d_y}-a)+\exp\{a\mathrm{d_y}\}\bar{\Phi}(\mathrm{d_y}+a)}, & \mathrm{d_g} \leq 0 \\ \frac{\exp\{-a\mathrm{d_y}\}\phi(\mathrm{d_g}-\mathrm{d_y}+a)}{\exp\{-a\mathrm{d_y}\}\Phi(\mathrm{d_y}-a)+\exp\{a\mathrm{d_y}\}\bar{\Phi}(\mathrm{d_y}+a)}, & \mathrm{d_g} > 0, \end{cases} \tag{6.32}$$

which is the weighted sum of truncated normal distributions. Detailed derivations of $g(\mathrm{d_y})$ and $p_1(\mathrm{d_g}|\mathrm{d_y})$ are provided by Pericchi and Smith (1992). The posterior mean is given by

$$\mathrm{PM}(\mathrm{d_g}|\mathrm{d_y}) = \omega_{post}(\mathrm{d_y})\left[\mathrm{d_y} - \frac{a[\exp\{-a\mathrm{d_y}\}\Phi(\mathrm{d_y}-a) - \exp\{a\mathrm{d_y}\}\bar{\Phi}(\mathrm{d_y}+a)]}{\exp\{-a\mathrm{d_y}\}\Phi(\mathrm{d_y}-a) + \exp\{a\mathrm{d_y}\}\bar{\Phi}(\mathrm{d_y}+a)}\right]. \tag{6.33}$$

For $\mathrm{d_g} \geq 0$, we have

$$\begin{aligned} \bar{F}_1(\mathrm{d_g}|\mathrm{d_y}) &= 1 - F_1(\mathrm{d_g}|\mathrm{d_y}) = \int_{\mathrm{d_g}}^{\infty} p_1(\mathrm{d_g}|\mathrm{d_y}) d\mathrm{d_g} \\ &= \frac{\exp\{-a\mathrm{d_y}\}\bar{\Phi}(\mathrm{d_g}-\mathrm{d_y}+a)}{\exp\{-a\mathrm{d_y}\}\Phi(\mathrm{d_y}-a) + \exp\{a\mathrm{d_y}\}\bar{\Phi}(\mathrm{d_y}+a)}. \end{aligned} \tag{6.34}$$

The result in (6.34) can be written as

$$\begin{aligned} &\frac{\exp\{-a\mathrm{d_y}\}\bar{\Phi}(\mathrm{d_g}-\mathrm{d_y}+a)}{\exp\{-a\mathrm{d_y}\}\Phi(\mathrm{d_y}-a) + \exp\{a\mathrm{d_y}\}\bar{\Phi}(\mathrm{d_y}+a)} \\ &= \frac{\omega g(\mathrm{d_y}) + (1-\omega)\phi(\mathrm{d_y})}{2\omega g(\mathrm{d_y})} \\ &= a^{-1}\omega^{-1}\exp\{-\frac{1}{2}a^2\}\phi(\mathrm{d_y})\frac{1+\omega\beta(\mathrm{d_y})}{\exp\{-a\mathrm{d_y}\}\Phi(\mathrm{d_y}-a) + \exp\{a\mathrm{d_y}\}\bar{\Phi}(\mathrm{d_y}+a)}, \end{aligned} \tag{6.35}$$

where $\beta(\mathrm{d_y}, a) = \frac{1}{2}a\{\frac{\Phi(\mathrm{d_y}-a)}{\phi(\mathrm{d_y}-a)} + \frac{\bar{\Phi}(\mathrm{d_y}+a)}{\phi(\mathrm{d_y}+a)}\} - 1$. This leads to

$$\bar{\Phi}(\widehat{\mathrm{d}}_{\mathrm{g}} - \mathrm{d_y} + a) = a^{-1}\omega^{-1}\phi(\mathrm{d_y}-a)\{1+\omega\beta(\mathrm{d_y})\}.$$

For $\mathrm{d_g} > 0$ the posterior median of $\widehat{\mathrm{d}}_{\mathrm{g}}(\mathrm{d_y};\omega)$ of $\mathrm{d_g}$ given $\mathrm{d_y}$, can be found from the properties

$$\begin{aligned} \widehat{\mathrm{d}}_{\mathrm{g}}(\mathrm{d_y};\omega) &= 0 \qquad\qquad \text{if} \quad \omega_{post}(\mathrm{d_y})\bar{F}_1(0|\mathrm{d_y}) \leq \frac{1}{2} \\ \bar{F}_1(\widehat{\mathrm{d}}_{\mathrm{g}}(\mathrm{d_y};\omega)|\mathrm{d_y}) &= \{2\omega_{post}(\mathrm{d_y})\}^{-1} \qquad \text{otherwise.} \end{aligned} \tag{6.36}$$

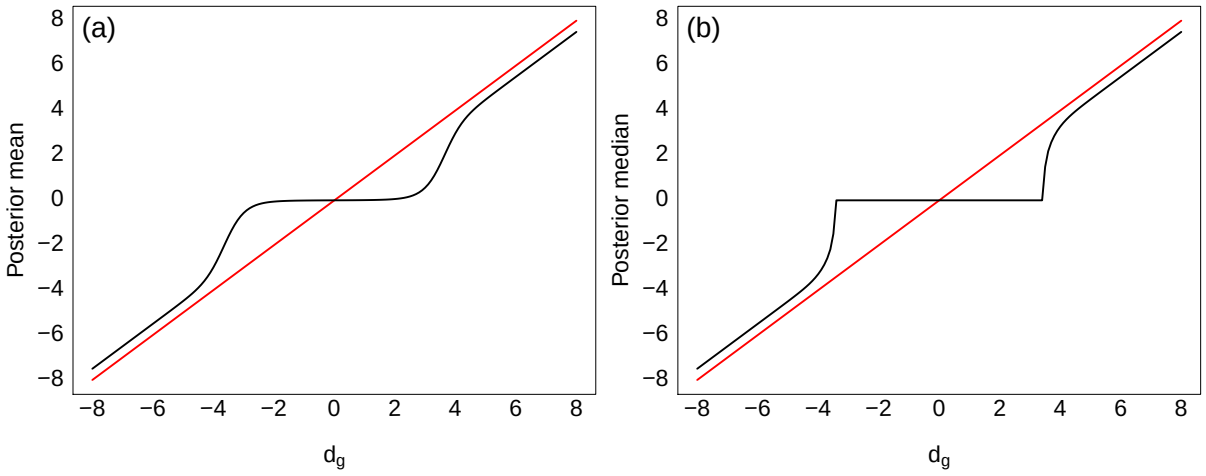


Figure 6.4: Plots of the posterior mean (a) and posterior median (b) using EB with the Laplace prior for $\omega = 0.02$ and $a = 0.5$.

Hence, if $\omega_{post}(\mathrm{d_y}) \leq \frac{1}{2}$ then the median is zero, and it is unnecessary to evaluate $\bar{F}_1(0|\mathrm{d_y})$. If $\mathrm{d_g} < 0$, the properties $-\widehat{\mathrm{d}}_{\mathrm{g}}(\mathrm{d_y};\omega) = \widehat{\mathrm{d}}_{\mathrm{g}}(-\mathrm{d_y};\omega)$ and $\bar{\Phi}^{-1}(\mathrm{d_y}) = -\Phi^{-1}(\mathrm{d_y})$ are used, and then

$$\widehat{\mathrm{d}}_{\mathrm{g}} = \mathrm{d_y} - a - \Phi^{-1}(\mathrm{z}), \tag{6.37}$$

where

$$\mathrm{z} = a^{-1}\phi(\mathrm{d_y} - a)\{\omega^{-1}\beta\mathrm{d_y}\}. \tag{6.38}$$

As $\mathrm{d_y} \to \infty$, the limiting value of z is $\frac{1}{2}$, and it is useful to use the approximation that $\phi(\mathrm{d_y} - a)$ is zero when $\mathrm{d_y}$ is large and $\beta(\mathrm{d_y})$ is infinite. If $\mathrm{d_y}$ is small, so that the value of z given by (6.38) is greater than 1, or $\mathrm{d_y} - a - \Phi^{-1}(\mathrm{z})$, then the posterior median will be equal to zero. Thus, the posterior median is given by

$$\widehat{\mathrm{d}}_{\mathrm{g}} = \max[0, \mathrm{d_y} - a - \Phi^{-1}\{\min(1, \mathrm{z})\}]. \tag{6.39}$$

The posterior mean and the posterior median in (6.33) and (6.39) are plotted in Figure 6.4 (a) and (b). The rule of the posterior mean can be described as slightly shrinking large and heavily shrinking small coefficients, and the posterior median can be described as slightly

shrinking large and heavily thresholding small coefficients (Reményi, 2012). Hence, as ω becomes close to 1, the posterior mean becomes smooth and the small coefficients slightly shrink. For more details and related theoretical results, the reader is referred to Johnstone and Silverman (2005b), and for more examples using the method, see Johnstone and Silverman (2005a). Calculations use the R package `EbayesThresh` of the empirical Bayes methods.

6.7 Comparison simulation

The purpose of this section is to evaluate and investigate whether Empirical Bayes (EB) and BayesThresh methods provide better block shapes than adaptive Bayesian wavelet shrinkage (ABWS) and Bayesian adaptive multi-resolution shrinkage (BAMS) when estimating an unknown vector $\mathbf{f}$.

The simulated datasets consisted of the standard test signal Blocks (Donoho and Johnstone, 1994; Nason and Silverman, 1994) at $m = 128$ equally spaced points, multiplied by the blur matrix, which is given in (2.6), with $k = 0.005$. Also, it was corrupted by independent Gaussian noise with the mean zero and the variance of noise taken as 0.5. No thresholding was done below level 3, the IT-TO method was used and the same datasets were used to simulate these methods. Moreover, the first-order method in Section 2.6 was used to estimate $\mathbf{f}$. The number of replications is equal to 60 and the number of iteration equals 100. Hence, MMSE is used to estimate the parameters, the approach of MMSE being described in Section 2.13.

Figure 6.5 shows the plots of reconstructions using different methods. Figure 6.5 (b) displays the reconstruction obtained from ABWS with MMSE=0.216, where it can be seen that the rule provides a reconstruction which does not fully recover the function from noise. Figure 6.5 (c) displays the result of the reconstruction obtained from BAMS, where it can be seen that the function is not fully recovered from the noise, although the MSE is slightly improved. Figure 6.5 (d) displays a reconstructed profile using BayesThresh,

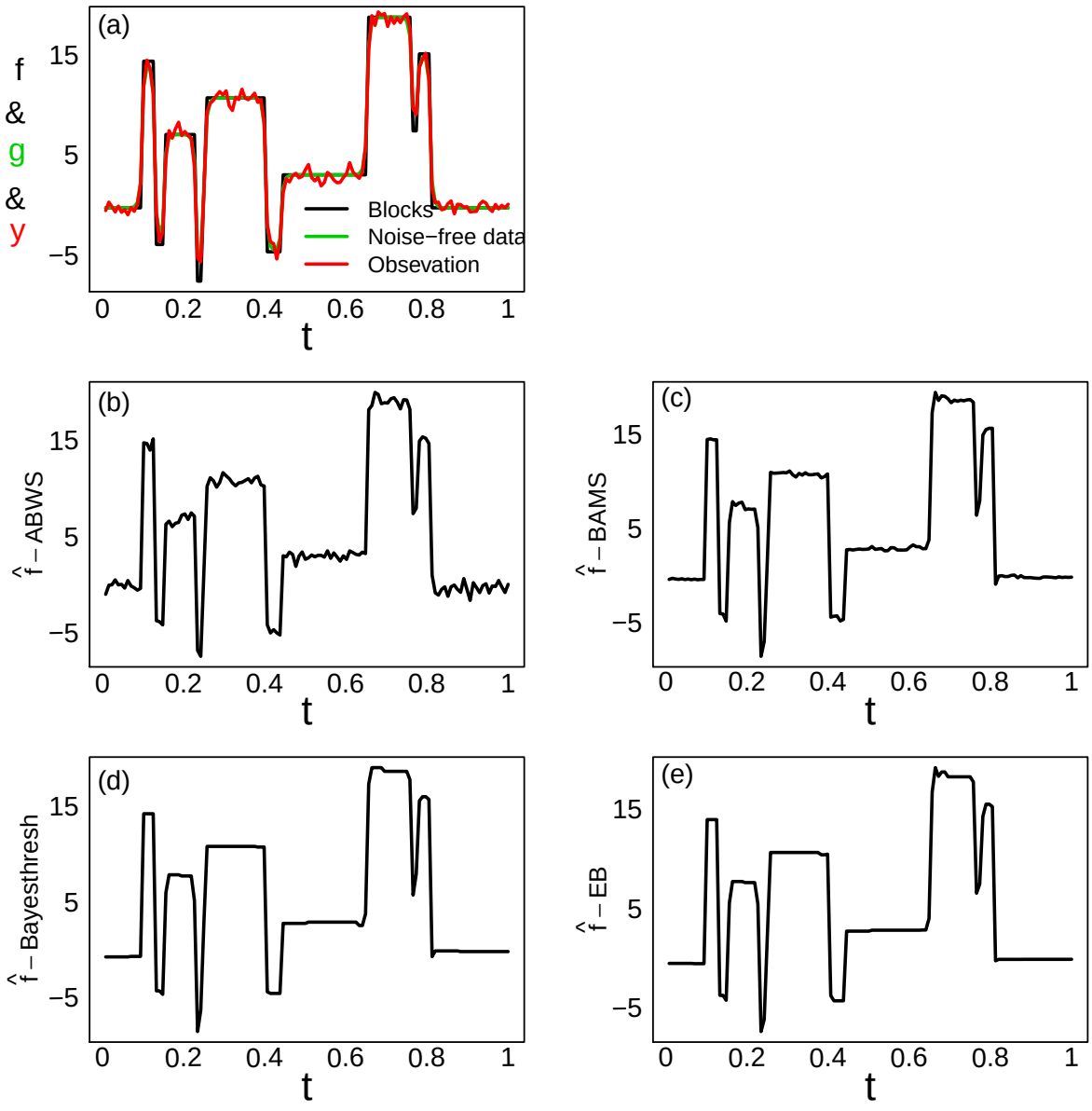


Figure 6.5: Plots of the reconstruction for estimating the unknown vector $\mathbf{f}$: (a) the black line is made from Blocks test function, the green line is made by multiplying the blocks by the blur, which is given in (2.6), $k = 0.005$ and the red line shows the observed data with noise $\sigma = 0.5$, the other panels show reconstruction using;(b) ABWS with $\Lambda_{\text{MMSE}} = 0.002$ and MMSE=0.216; (c) BAMS with $\Lambda_{\text{MMSE}} = 0.0002$ and MMSE= 0.15; (d) BayesThresh with $\Lambda_{\text{MMSE}} = 0.033$ and MMSE= 0.2; and (e) EB with $\Lambda_{\text{MMSE}} = 0.005$ and MMSE= 0.13.

providing sharp edges and flat topped reconstructions.

Figure 6.5 (e) displays the result of the reconstruction using the posterior median from the empirical Bayes method. It can be seen that the rule provides sharp edges and flat topped reconstructions. The MMSE equals 0.13.

6.8 Conclusions

In this chapter, Bayesian thresholding was investigated and a brief review of wavelet shrinkage provided. A range of shrinkage functions, by approaching the standard context from a Bayesian point of view, are studied and two estimation procedures are compared; the posterior mean, and the posterior median. An automatic technique has been studied whereby a set of level-dependent shrinkage functions may be chosen adaptively for a given dataset. In addition, the methods were illustrated on a dataset from the Blocks test function. It can be concluded that the Empirical Bayes (EB) and BayesThresh methods provided better block shapes than the adaptive Bayesian wavelet shrinkage (ABWS) and Bayesian adaptive multi-resolution shrinkage (BAMS) methods. Additionally, the empirical Bayes method performed well at both denoising and preserving the important features of the Blocks test function.

Chapter 7

Wavelet-based two-stage reconstruction

7.1 Overview

Within this Chapter, Section 7.2 gives an introduction, while 7.3 considers the wavelet-vaguelette decomposition, then Section 7.4 looks at vaguelette-wavelet decomposition. Section 7.5 discusses the procedure of wavelet-based two-stage method. Section 7.6 presents the results from a simulation study, and Section 7.7 provides the conclusions.

7.2 Introduction

Consider the aim of recovering an underlying vector $\mathbf{f}$ from the linear model

$$\mathbf{y} = \mathbf{H}\mathbf{f} + \boldsymbol{\epsilon}, \tag{7.1}$$

with data vector $\mathbf{y}_{n\times 1}$, known blur matrix $\mathbf{H}_{n\times m}$ and a vector of random variables $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$. Furthermore, this type of problem is referred to as ill-posed, as even when $n = m$ the somewhat naive estimate of $\widehat{\mathbf{f}} = \mathbf{H}^{-1}\mathbf{y}$ often fails to yield a reasonable solution

since $\mathbf{H}^{-1}$ is an unbounded linear operator. This means that even small amounts of noise in the data "blow up" when the straightforward inversion estimate is used. Such problems arise in a wide variety of practical scientific settings with different types of transformation $\mathbf{H}$, and are often referred to as *linear inverse problems* (Abramovich *et al.*, 2000).

Abramovich and Silverman (1998) stated that there are two methods which can be used to solve inverse problems. The first is to use a wavelet method, that was first introduced by Donoho (1995). The second is to use a properly chosen truncated singular-value decomposition (SVD) method. The level of approximation of a Fourier series depends on the number of terms in the sum of sines and cosines; it will not be appropriate if the function is not uniformly smooth. This means that if the function is of uniform smoothness, the reconstruction of $\mathbf{f}$, using the eigenfunctions of $\mathbf{H}^*\mathbf{H}$, where $\mathbf{H}^*$ is the adjoint of $\mathbf{H}$, is inefficient and the estimator does not perform well (Cai, 2002). To overcome this limitation wavelet-vaguelette and vaguelette-wavelet methods were proposed (Abramovich *et al.*, 2000). We first briefly review the wavelet-vaguelette decomposition approach and the vaguelette-wavelet decomposition approach of Abramovich and Silverman (1998). Then the problem of estimating the Blocks test function is used to illustrate the estimation procedure. The resulting MSE is plotted for different methods of inversion with different thresholding rules.

7.3 The wavelet-vaguelette decomposition

Kane *et al.* (2002) gave details of how the wavelet-vaguelette decomposition (WVD) can be applied to the inverse problem and Donoho (1995) and Kane *et al.* (2002) proposed an alternative methodology for solving linear inverse problems. The procedure is described as follows: define an orthogonal wavelet transform matrix $\mathbf{W}$, where each row is a discrete wavelet. The matrix $\mathbf{H}$ then operates on each individual wavelet to produce what is called a *vaguelette*

$$\mathbf{H}\mathbf{W}^T = \mathbf{V}^T\mathbf{\Gamma}, \tag{7.2}$$

where each row of $\mathbf{V}$ is a discrete vaguelette, which has been normalized to unit energy. Each normalization factor has been put on the diagonal of the diagonal matrix $\mathbf{\Gamma}$. Abramovich and Silverman (1998) stated that the method can be applied to inverse problems to provide an estimate of the vector $\mathbf{f}$. Kane *et al.* (2002) stated that the matrices $\mathbf{W}$ and $\mathbf{\Gamma}$ are always invertible, but $\mathbf{V}^T$ is invertible only if $\mathbf{H}$ is invertible, so consider

$$\mathbf{H}^{-1} = \mathbf{W}^T\mathbf{\Gamma}^{-1}(\mathbf{V}^T)^{-1}. \tag{7.3}$$

Hence, $(\mathbf{V}^T)^{-1} = \mathbf{\Gamma W H}^{-1}$ and thus,

$$\begin{aligned}\widehat{\mathbf{f}} &= \mathbf{H}^{-1}\mathbf{y} \\ &= \mathbf{W}^T\mathbf{\Gamma}^{-1}(\mathbf{V}^T)^{-1}\mathbf{y} \\ &= \mathbf{W}^T\mathbf{\Gamma}^{-1}(\mathbf{\Gamma W H}^{-1}\mathbf{y}).\end{aligned} \tag{7.4}$$

Consequently, Equation (7.4) can be explained as inversion, the vector, $\mathbf{f}$, estimated using the inverse transform of the wavelet coefficients of $\mathbf{\Gamma}^{-1}(\mathbf{\Gamma W H}^{-1}\mathbf{y})$. If the matrix, $\mathbf{H}$, is not square, then $\mathbf{H}^{-1}$ is replaced by $(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T$ and the ordinary least squares (OLS) estimator can be considered

$$\widehat{\mathbf{f}}^{\text{WVD}}_{\text{OLS}_\lambda} = \mathbf{W}^T\mathbf{\Gamma}^{-1}\text{T}(\mathbf{\Gamma W}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{y}), \tag{7.5}$$

where $\text{T}(\cdot)$ is a thresholding rule, and λ is the value of the threshold. If $\mathbf{H}^T\mathbf{H}$ is not invertible then the ridge regression estimator can be considered

$$\widehat{\mathbf{f}}^{\text{WVD}}_{\text{Ridge}_\lambda,\Lambda} = \mathbf{W}^T\mathbf{\Gamma}^{-1}\text{T}(\mathbf{\Gamma W}(\mathbf{H}^T\mathbf{H} - \Lambda\mathbf{I})^{-1}\mathbf{H}^T\mathbf{y}), \tag{7.6}$$

where $\Lambda = \sigma^2\kappa$ denotes a parameter of the inversion method, and λ indicates the value of the threshold (Kane *et al.*, 2002). Where ridge regression is a special case of regularization, when $\mathfrak{R} = \mathbf{I}$. So, for other regularization methods the estimate is given by

$$\widehat{\mathbf{f}}^{\text{WVD}}_{\text{Reg}_\lambda,\Lambda} = \mathbf{W}^T\mathbf{\Gamma}^{-1}\text{T}(\mathbf{\Gamma W}(\mathbf{H}^T\mathbf{H} - \Lambda\mathfrak{R}^T\mathfrak{R})^{-1}\mathbf{H}^T\mathbf{y}), \tag{7.7}$$

where $\Lambda = \sigma^2\kappa$, $\mathfrak{R}$ can have different definitions, such as the three common choices considered earlier in Section 2.6, and $\mathbf{W}$ was defined in Section 2.9.

Wavelet-vaguelettes tend to isolate the signal into a few large values, whilst the noise tends to be spread around equally. The result of thresholding of the small wavelet coefficients will be to remove the noise and leave the more interesting coherent features untouched.

7.4 The vaguelette-wavelet decomposition

Abramovich and Silverman (1998) introduced the vaguelette-wavelet decomposition (VWD) method, which can be viewed as a natural alternative to the previous use of the WVD. Consider the form in (7.3), which can be expressed as

$$\mathbf{WH} = \mathbf{\Gamma V},$$

for some operator matrix, $\mathbf{H}$, with

$$\mathbf{H}^{-1} = \mathbf{V}^{-1}\mathbf{\Gamma}^{-1}\mathbf{W}.$$

Hence, $\mathbf{V}^{-1} = \mathbf{H}^{-1}\mathbf{W\Gamma}$, so,

$$\begin{aligned}\mathbf{f} &= \mathbf{H}^{-1}\mathbf{y} \\ &= \mathbf{V}^{-1}\mathbf{\Gamma}^{-1}\mathbf{W}\mathbf{y} \\ &= \mathbf{H}^{-1}\mathbf{W}^T\mathbf{\Gamma}(\mathbf{\Gamma}^{-1}\mathbf{W}\mathbf{y}),\end{aligned} \tag{7.8}$$

and then the OLS estimator is given by

$$\widehat{\mathbf{f}}^{\mathrm{VWD}}_{\mathrm{OLS}_\lambda} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{W}^T\mathbf{\Gamma}\mathrm{T}(\mathbf{\Gamma}^{-1}\mathbf{W}\mathbf{y}). \tag{7.9}$$

But, if $\mathbf{H}^T\mathbf{H}$ is not invertible then ridge regression can be used

$$\widehat{\mathbf{f}}^{\mathrm{VWD}}_{\mathrm{Ridge}_\lambda,\Lambda} = (\mathbf{H}^T\mathbf{H} - \Lambda\mathbf{I})^{-1}\mathbf{H}^T\mathbf{W}^T\mathbf{\Gamma}\mathrm{T}(\mathbf{\Gamma}^{-1}\mathbf{W}\mathbf{y}), \tag{7.10}$$

where $\Lambda = \sigma^2\kappa$. Other regularization methods are given by

$$\widehat{\mathbf{f}}^{\mathrm{VWD}}_{\mathrm{Reg}_\lambda,\Lambda} = (\mathbf{H}^T\mathbf{H} - \Lambda\mathfrak{R}^T\mathfrak{R})^{-1}\mathbf{H}^T\mathbf{W}^T\mathbf{\Gamma}\mathrm{T}(\mathbf{\Gamma}^{-1}\mathbf{W}\mathbf{y}), \tag{7.11}$$

where $\Lambda = \sigma^2\kappa$, $\mathfrak{R}$ was defined in Section 2.6 and $\mathbf{W}$ was defined in Section 2.9. Abramovich and Silverman (1998) summarized the procedure of the VWD, as $\mathbf{f}$ is unknown itself, rather than $\mathbf{Hf}$. The main idea of VWD is to find the corresponding empirical vaguelette coefficients and then apply a suitable thresholding rule, such as hard thresholding, with the aim of deriving a vaguelette estimator of $\mathbf{\Gamma}^{-1}\mathbf{W}\mathbf{y}$, the function $\mathbf{f}$ is estimated using the inverse transform of $\mathbf{W}^T\mathbf{\Gamma}\mathrm{T}(\mathbf{\Gamma}^{-1}\mathbf{W}\mathbf{y})$ and then an inversion method is applied, such as ridge regression.

7.5 The procedure of wavelet-based two-stage method

The simulation procedure can be summarised as follows.

1. The first method is IT-TO, which can be explained as inversion and then the wavelet coefficients are thresholded. More precisely, the unknown vector $\mathbf{f}$, can be estimated by

$$\widehat{\mathbf{f}}^{\text{IT-TO}}_{\text{Reg}_{\boldsymbol{\lambda},\Lambda}} = \mathbf{W}^T\mathrm{T}(\mathbf{W}(\mathbf{H}^T\mathbf{H} - \Lambda\mathfrak{R}^T\mathfrak{R})^{-1}\mathbf{H}^T\mathbf{y}),$$

   where $\Lambda = \sigma^2\kappa$, $\mathfrak{R}$ and $\mathbf{W}$ were defined in Sections 2.6 and 2.9, $\mathrm{T}(\cdot)$ is the thresholding rule, and the parameters $\boldsymbol{\lambda}$ and Λ are estimated together using MMSE, as described in Section 2.13. There are many different choices of thresholding rule:

   (a) $\mathrm{T}(\cdot)$ can be chosen to be classical thresholding such as

$$\mathrm{T}^{\mathrm{H}}(\mathbf{W}(\mathbf{H}^T\mathbf{H} - \Lambda\mathfrak{R}^T\mathfrak{R})^{-1}\mathbf{H}^T\mathbf{y}),$$

   where $\mathrm{T}^{\mathrm{H}}(\cdot)$ represents the result of the hard thresholding rule.

   (b) $\mathrm{T}(\cdot)$ can be chosen as Bayesian thresholding using single or a mixture priors and then the IT-TO method will be applied, which can be assumed as $\widehat{\mathbf{d}}_{\mathbf{g}} = \mathbf{W}\mathrm{I}_*(\mathbf{H},\mathbf{y},\Lambda)$, with, for example, $\mathrm{I}_*(\mathbf{H},\mathbf{y},\Lambda) = (\mathbf{H}^T\mathbf{H}+\Lambda\mathfrak{R}^T\mathfrak{R})^{-1}\mathbf{H}^T\mathbf{y}$, where $\Lambda = \sigma^2\kappa$, and then the result of filtering, $\widehat{\mathbf{d}}_{\mathbf{g}} = \mathbf{W}(\mathbf{H}^T\mathbf{H} + \Lambda\mathfrak{R}^T\mathfrak{R})^{-1}\mathbf{H}^T\mathbf{y}$, is used to estimate $\mathbf{d}_{\mathbf{f}}$ and the estimate of $\mathbf{f}$, is given by

$$\widehat{\mathbf{f}}^{\text{IT-TO}}_{\text{Reg},\Lambda} = \mathbf{W}^T\mathrm{p}(\mathbf{d}_{\mathbf{f}}|\widehat{\mathbf{d}}_{\mathbf{g}}).$$

Note that, the values of the wavelet coefficients of $\mathbf{d_f}$ is assumed to follow the same distribution and are conditionally independent given the wavelet coefficients of $\widehat{\mathbf{d}}_{\mathbf{g}}$.

(c) The wavelet coefficients of $(\mathbf{H}^T\mathbf{H}-\Lambda\mathfrak{R}^T\mathfrak{R})^{-1}\mathbf{H}^T\mathbf{y}$ can be calculated using the unbalanced Haar transform, then

$$\widehat{\mathbf{f}}^{\text{IT-TO}}_{\text{Reg}_{\boldsymbol{\lambda},\Lambda}} = \text{IDUHT}\left(\text{T}\left(\text{DUHT}((\mathbf{H}^T\mathbf{H}-\Lambda\mathfrak{R}^T\mathfrak{R})^{-1}\mathbf{H}^T\mathbf{y})\right)\right),$$

where $\Lambda=\sigma^2\kappa$, IDUHT is the inversion of the unbalanced Haar transform and $\text{T}\left(\text{DUHT}((\mathbf{H}^T\mathbf{H}-\Lambda\mathfrak{R}^T\mathfrak{R})^{-1}\mathbf{H}^T\mathbf{y})\right)$ represents the result of thresholding, such as the hard thresholding rule.

2. The second method is TI-TO, which can be explained as thresholding the wavelet coefficients and then applying an inversion method. More precisely, the unknown vector $\mathbf{f}$, can be estimated by

$$\widehat{\mathbf{f}}^{\text{TI-TO}}_{\text{Reg}_{\boldsymbol{\lambda},\Lambda}} = (\mathbf{H}^T\mathbf{H}-\Lambda\mathfrak{R}^T\mathfrak{R})^{-1}\mathbf{H}^T\mathbf{W}^T\text{T}(\mathbf{W}\mathbf{y}),$$

where $\Lambda=\sigma^2\kappa$, $\mathfrak{R}$ and $\mathbf{W}$ were defined in Sections 2.6 and 2.9, $\text{T}(\cdot)$ is the thresholding rule, and the parameters $\boldsymbol{\lambda}$ and Λ are estimated together using MMSE, as described in Section 2.13. There are many different choices of thresholding rule:

(a) $\text{T}(\cdot)$ can be chosen to be classical thresholding such as

$$\text{T}^{\text{H}}(\mathbf{W}\mathbf{y}),$$

where $\text{T}^{\text{H}}(\cdot)$ represents the result of hard thresholding rule.

(b) $\text{T}(\cdot)$ can be chosen as Bayesian thresholding using a single or mixture priors and then the TI-TO method will be applied, so it is assumed that $\mathbf{d_y}=\mathbf{W}\mathbf{y}$ and then the result of thresholding, $\mathbf{d_y}=\mathbf{W}\mathbf{y}$ is used to estimate $\mathbf{d_g}$ and the estimation of $\mathbf{f}$, is given by

$$\widehat{\mathbf{f}}^{\text{TI-TO}}_{\text{Reg},\Lambda} = (\mathbf{H}^T\mathbf{H}-\Lambda\mathfrak{R}^T\mathfrak{R})^{-1}\mathbf{H}^T\mathbf{W}^T\text{p}(\mathbf{d_g}|\mathbf{d_y}).$$

Note that, the values of the wavelet coefficients of $\mathbf{d_g}$ is assumed to follow the same distribution and are conditionally independent given the wavelet coefficients of $\mathbf{d_y}$.

(c) The wavelet coefficients of $\mathbf{d_y}$ can be calculated using the unbalanced Haar transform, then

$$\widehat{\mathbf{f}}^{\text{TI-TO}}_{\text{Reg}_{\boldsymbol{\lambda},\Lambda}} = (\mathbf{H}^T\mathbf{H} - \Lambda\mathfrak{R}^T\mathfrak{R})^{-1}\mathbf{H}^T\text{IDUHT}\left(\text{T}\left(\text{DUHT}(\mathbf{y})\right)\right),$$

where IDUHT is the inversion of the unbalanced Haar transform and $\text{T}\left(\text{DUHT}(\mathbf{y})\right)$ represents the result of the hard thresholding rule.

3. The third method is WVD-TO, which can be explained as inversion and then thresholding. More precisely, the unknown vector $\mathbf{f}$, can be estimated by

$$\widehat{\mathbf{f}}^{\text{WVD-TO}}_{\text{Reg}_{\boldsymbol{\lambda},\Lambda}} = \mathbf{W}^T\boldsymbol{\Gamma}^{-1}\text{T}(\boldsymbol{\Gamma}\mathbf{W}(\mathbf{H}^T\mathbf{H} - \Lambda\mathfrak{R}^T\mathfrak{R})^{-1}\mathbf{H}^T\mathbf{y}),$$

where $\Lambda = \sigma^2\kappa$, $\mathfrak{R}$ and $\mathbf{W}$ were defined in Sections 2.6 and 2.9, $\text{T}(\cdot)$ is the thresholding rule, and the parameters $\boldsymbol{\lambda}$ and Λ are estimated together using MMSE, as described in Section 2.13. There are many different choices of thresholding rule:

(a) $\text{T}(\cdot)$ can be chosen to be classical thresholding such as

$$\text{T}^{\text{H}}(\boldsymbol{\Gamma}\mathbf{W}(\mathbf{H}^T\mathbf{H} - \Lambda\mathfrak{R}^T\mathfrak{R})^{-1}\mathbf{H}^T\mathbf{y}),$$

where $\text{T}^{\text{H}}(\cdot)$ represents the result of hard thresholding rule.

(b) $\text{T}(\cdot)$ can be chosen as Bayesian thresholding using a single or mixture of priors and then the WVD-TO method will be applied, so it is assumed that $\widehat{\mathbf{d}}_{\mathbf{g}} = \boldsymbol{\Gamma}\mathbf{W}\text{I}_*(\mathbf{H}, \mathbf{y}, \Lambda)$, such as $\widehat{\mathbf{d}}_{\mathbf{g}} = \boldsymbol{\Gamma}\mathbf{W}(\mathbf{H}^T\mathbf{H} + \Lambda\mathfrak{R}^T\mathfrak{R})^{-1}\mathbf{H}^T\mathbf{y}$, where $\Lambda = \sigma^2\kappa$, and then the result of filtering, $\widehat{\mathbf{d}}_{\mathbf{g}} = \boldsymbol{\Gamma}\mathbf{W}(\mathbf{H}^T\mathbf{H} + \Lambda\mathfrak{R}^T\mathfrak{R})^{-1}\mathbf{H}^T\mathbf{y}$, is used to estimate $\widehat{\mathbf{d}}_{\mathbf{g}}$, and the estimation of $\mathbf{f}$, is given by

$$\widehat{\mathbf{f}}^{\text{WVD-TO}}_{\text{Reg}_{\Lambda}} = \mathbf{W}^T\boldsymbol{\Gamma}^{-1}\text{p}(\mathbf{d_f}|\boldsymbol{\Gamma}\mathbf{W}(\mathbf{H}^T\mathbf{H} - \Lambda\mathfrak{R}^T\mathfrak{R})^{-1}\mathbf{H}^T\mathbf{y}).$$

Note that, the values of the wavelet coefficients of $\mathbf{d_f}$ is assumed to follow the same distribution and are conditionally independent given the wavelet coefficients of $\boldsymbol{\Gamma}\mathbf{W}(\mathbf{H}^T\mathbf{H} - \Lambda\mathfrak{R}^T\mathfrak{R})^{-1}\mathbf{H}^T\mathbf{y}$.

(c) The wavelet coefficients of $(\mathbf{H}^T\mathbf{H}-\Lambda\mathfrak{R}^T\mathfrak{R})^{-1}\mathbf{H}^T\mathbf{y}$ can be calculated using the unbalanced Haar transform, then

$$\widehat{\mathbf{f}}^{\text{WVD-TO}}_{\text{Reg}_{\boldsymbol{\lambda},\Lambda}} = \text{IDUHT}\,\boldsymbol{\Gamma}^{-1}(\text{T}\,(\boldsymbol{\Gamma}\text{DUHT}(\mathbf{H}^T\mathbf{H}-\Lambda\mathfrak{R}^T\mathfrak{R})^{-1}\mathbf{H}^T\mathbf{y})),$$

where IDUHT is the inversion of the unbalanced Haar transform and T ($\boldsymbol{\Gamma}$ DUHT $(\mathbf{H}^T\mathbf{H}-\Lambda\mathfrak{R}^T\mathfrak{R})^{-1}\mathbf{H}^T\mathbf{y}$) represents the result of the hard thresholding rule.

4. The fourth method is VWD-TO, which can be explained as thresholded the wavelet coefficients of $\boldsymbol{\Gamma}^{-1}\mathbf{W}\mathbf{y}$ and then applying an inversion method. More precisely, the unknown vector $\mathbf{f}$, can be estimated by

$$\widehat{\mathbf{f}}^{\text{VWD-TO}}_{\text{Reg}_{\boldsymbol{\lambda},\Lambda}} = (\mathbf{H}^T\mathbf{H}-\Lambda\mathfrak{R}^T\mathfrak{R})^{-1}\mathbf{H}^T\mathbf{W}^T\boldsymbol{\Gamma}\text{T}(\boldsymbol{\Gamma}^{-1}\mathbf{W}\mathbf{y}),$$

where $\Lambda=\sigma^2\kappa$, $\mathfrak{R}$ and $\mathbf{W}$ were defined in Sections 2.6 and 2.9, T$(\cdot)$ is thresholding rule, and the parameters $\boldsymbol{\lambda}$ and Λ are estimated together using MMSE, as described in Section 2.13. There are many different choices of thresholding rule:

(a) T$(\cdot)$ can be chosen to be classical thresholding such as

$$\text{T}^{\text{H}}(\boldsymbol{\Gamma}^{-1}\mathbf{W}\mathbf{y}),$$

where $\text{T}^{\text{H}}(\cdot)$ represents the result of hard thresholding rule.

(b) T$(\cdot)$ can be chosen as Bayesian thresholding using single or mixture of priors and then the VWD-TO method will be applied, so it is assumed that $\mathbf{d_y} = \boldsymbol{\Gamma}^{-1}\mathbf{W}\mathbf{y}$ and then the result of threshold, $\mathbf{d_y} = \boldsymbol{\Gamma}^{-1}\mathbf{W}\mathbf{y}$, is used to estimate $\mathbf{d_g}$ and the estimation of $\mathbf{f}$, is given by

$$\widehat{\mathbf{f}}^{\text{VWD-TO}}_{\text{Reg}_{\boldsymbol{\lambda},\Lambda}} = (\mathbf{H}^T\mathbf{H}-\Lambda\mathfrak{R}^T\mathfrak{R})^{-1}\mathbf{H}^T\mathbf{W}^T\boldsymbol{\Gamma}\text{p}(\mathbf{d_g}|\mathbf{d_y}).$$

Note that, the values of the wavelet coefficients of $\mathbf{d_g}$ is assumed to follow the same distribution and are conditionally independent given the wavelet coefficients of $\widehat{\mathbf{d}}_\mathbf{y}$.

(c) The wavelet coefficients of $\mathbf{d_y}$ can be calculated using the unbalanced Haar transform, then

$$\widehat{\mathbf{f}}^{\text{VWD-TO}}_{\text{Reg}_{\boldsymbol{\lambda},\Lambda}} = (\mathbf{H}^T\mathbf{H}-\Lambda\mathfrak{R}^T\mathfrak{R})^{-1}\mathbf{H}^T\,\text{IDUHT}\,\boldsymbol{\Gamma}\,(\text{T}\,(\boldsymbol{\Gamma}^{-1}\text{DUHT}(\mathbf{y}))),$$

where IDUHT is the inversion of the unbalanced Haar transform and T $(\mathbf{\Gamma}^{-1}\text{DUHT}(\mathbf{y}))$ represents the result of the hard thresholding rule.

7.6 Simulation and comparisons

This section is concerned with evaluating and investigating whether the IT-TO, TI-TO, VWD-TO and WVD-TO algorithms are suitable for estimating the vector $\mathbf{f}$ in (2.33). To examine the accuracy of the proposed methods, these algorithms were applied to the Blocks and Bumps test functions corrupted by noise and blur, which is given in (2.6), with $\sigma = 0.5$, $k = 0.005$ and the methods were combined with a variety of thresholding methods.

For each simulation 1000 independent runs were used to estimate the unknown vector $\mathbf{f}$. The estimation at each run was compared with the true values of $\mathbf{f}$ and the result was used to compute the MMSE using the approach described in Section 2.13.

The results of the MMSE are computed using the TI-TO, IT-TO, VWD-TO and WVD-TO methods for estimating the original Blocks test function and are shown in Figure 7.1, with the first-order method involved for estimating the unknown vector $\mathbf{f}$. In general, Figure 7.1 shows the WVD-TO method for estimating the original Blocks test function provides a smaller MSE than TI-TO, IT-TO or VWD-TO methods. More precisely, the hard thresholding rule, UH transform, NNBWS-MAP and DWWS-LPM with WVD-TO algorithms all improve the MSE compared to using the TI-TO, IT-TO or VWD-TO methods.

Similarly, the MMSE obtained using the TI-TO, IT-TO, VWD-TO and WVD-TO algorithms for estimating the original Bumps test function is shown in Figure 7.2, with the first-order method used for estimating $\mathbf{f}$. In general, Figure 7.2 shows the WVD-TO method for estimating the original Bumps test function provides a smaller MSE than TI-TO, IT-TO or VWD-TO methods. More precisely, the UH transform with hard, DENWS-MAP, NNBWS-MAP, DWWS-LPM and EB algorithms all improve the MSE

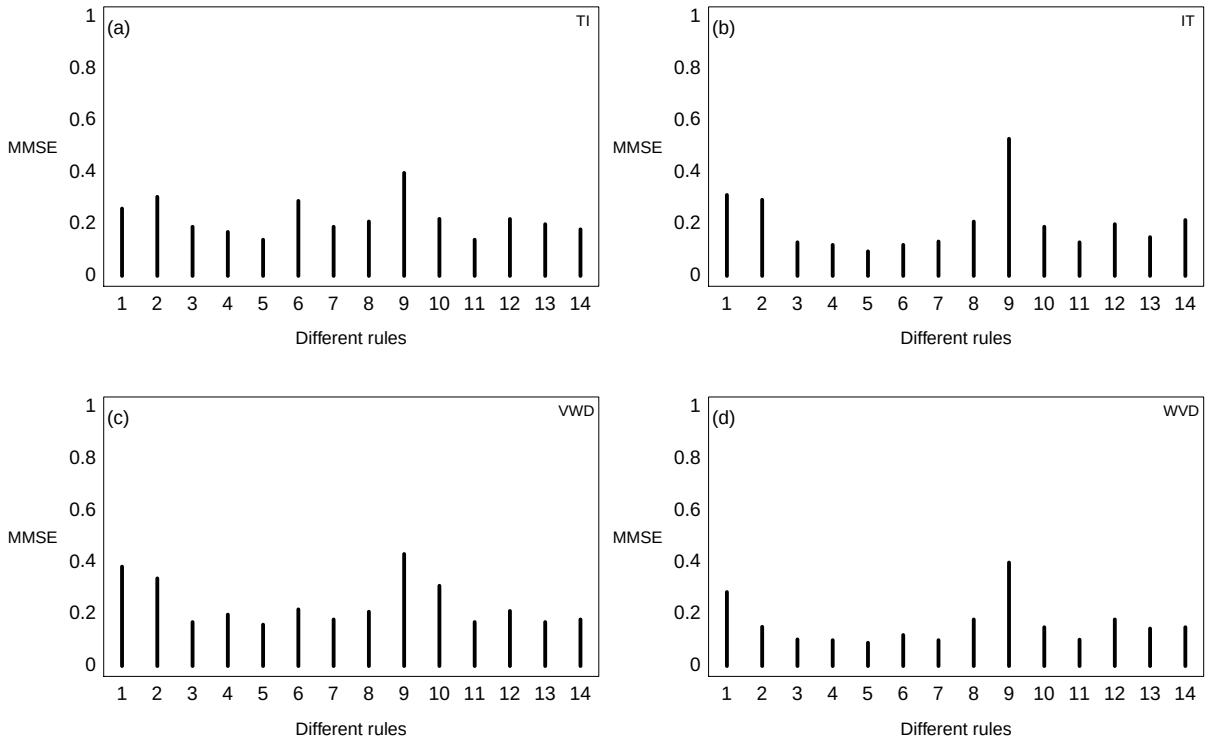


Figure 7.1: Plots of the MMSE results for all reconstructions of the Blocks test function using the first-order smoothing for various shrinking procedures: (1) SURE with hard; (2) cross-validation with hard; (3) hard; (4) BlockSure; (5) Unbalanced Haar transform with hard; (6) DENWS-MAP; (7) NNBWS-MAP; (8) NDEWS-MAP; (9) NNWS-MAP; (10) DWWS-LPM; (11) EB; (12) BayesThresh; (13) BAMS; and (14) ABWS.

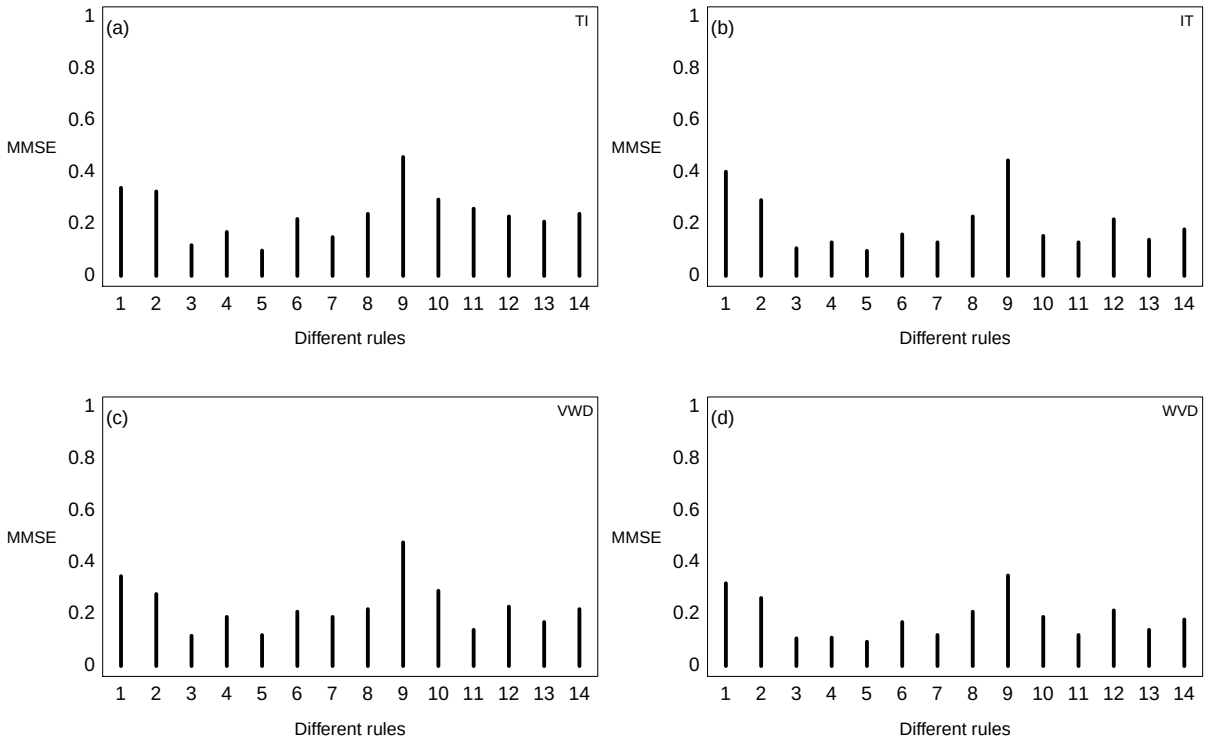


Figure 7.2: Plots of the MMSE results for all reconstructions of the Bumps test function using the first-order smoothing for various shrinking procedures: (1) SURE with hard; (2) cross-validation with hard; (3) hard; (4) BlockSure; (5) Unbalanced Haar transform with hard; (6) DENWS-MAP; (7) NNBWS-MAP; (8) NDEWS-MAP; (9) NNWS-MAP; (10) DWWS-LPM; (11) EB; (12) BayesThresh; (13) BAMS; and (14) ABWS.

compared to using TI-TO, IT-TO and VWD-TO methods.

7.7 Conclusions

In this chapter, TI-TO, IT-TO, VWD-TO and WVD-TO algorithms involving different thresholding rules of wavelet thresholding have been introduced and extended. The methodology of WVD-TO and VWD-TO algorithms are considered in Sections 7.3 and 7.4 for different inversion methods. The simulations considered in Section 7.6 show that the WVD-TO algorithm is comparable to or even slightly better than VWD-TO.

An investigation was carried out into how applying a thresholding rule affects the inversion method with the TI-TO, IT-TO, VWD-TO and WVD-TO methods. Extensive simulations have shown that MSE can be used to compare different methods. In general, the simulation shows that WVD-TO and IT-TO methods work well across two signal types, with the IT-TO method gives slightly better result than the TI-TO method, because the true signal is a step. It can be concluded that the WVD-TO and IT-TO methods improve the MSE compared to using TI-TO or VWD-TO methods.

Chapter 8

Wavelet-based one-stage reconstruction

8.1 Overview

Within this chapter, Section 8.2 contains an introduction, whilst Section 8.3 describes Bayesian modelling, thereafter Section 8.4 provides the numerical methods, and Section 8.5 gives the results from a simulation study. Finally, Section 8.6 presents the conclusions.

8.2 Introduction

Within this chapter a new model is considered, which depends on a statistical approach and uses a stochastic algorithm for the estimation of an underlying vector $\mathbf{f}$. In particular, the new method describes a curve in terms of wavelet coefficients. The vector is estimated, in a Bayesian framework, using a Markov chain Monte Carlo (MCMC) algorithm. Mathematically there are many different types of inverse problems, whereas in this thesis, inverse problems are divided into two types. In the case of the first type, the signal is only corrupted by white noise. In the case of the second, the signal is corrupted

by noise and there is also blurring. If there is enough information about the model then the relationships between the data and the parameters will be well defined. There are two relationships, which can be defined; the first relationship, between observations and parameters, is explained by the likelihood; and the second is the relationship between parameters, which is described by a prior (Aykroyd, 2015). Priors with a single component are commonly chosen, for example by Hoerl and Kennard (1970), who used the Gaussian prior, and Tibshirani (1996), who used the Laplace prior. For Bayesian wavelet-based modelling, Figueiredo and Nowak (2001) proposed the scale invariant term-by-term "Bayesian ABE" method, whereas Huerta (2005) proposed a multivariate Bayes wavelet shrinkage method that allows for correlations among wavelet coefficients corresponding to the same level of detail. Cutillo *et al.* (2008) proposed a list of shrinkage rules, which always pick the mode of the posterior that is "larger mode", in absolute value, and Reményi and Vidakovic (2015) proposed the double Weibull prior on the locations of wavelet coefficients. Some of these models are complicated and with prior parameters, that must also be modelled, which is called hierarchical modelling. Related ideas of hierarchical modelling have been used in wavelet methods, for example, Clyde *et al.* (1998), Cutillo *et al.* (2008), Clyde and George (1999), Clyde and George (2000), Aykroyd and Mardia (2003) and Reményi and Vidakovic (2015).

The resulting wavelet reconstruction problem is now more complicated, because the integration or summation required is too difficult, with numerical and MCMC methods often being used. Thus, the MCMC approach is an important tool for solving difficult computational problems and then the flexibility of the MCMC method allows estimation in complex cases. Moreover, this approach allows general investigation of the posterior distribution. The MCMC approach has been used for some wavelet-based problems, for example by Aykroyd and Mardia (2003) and Reményi (2012).

The MCMC approach has attracted much attention over the last three decades. Statisticians have been increasingly drawn to MCMC approach to simulate from complex and complicated distributions (Chib and Greenberg, 1995). It can be a powerful computational tool owing to its conceptual simplicity and relative ease of implementation. The

biggest challenge of the implementation of the Bayesian approach is that finding the posterior distribution often requires high-dimensional integration, which is too complex for an analytical solution. In general, the basic idea of the MCMC approach is to construct a Markov chain to generate pseudo-random samples, such that its stationary distribution follows the target posterior distribution. The use of the MCMC approach dates back to the early '50s (Metropolis *et al.*, 1953), and was extended and generalised by Hastings (1970). The ideas of the MCMC approach were developed by Geman and Geman (1984), who proposed the simulated annealing method, which is a stochastic optimization method. Besag *et al.* (1995) gave a good review of the theory and application, whereas Chib and Greenberg (1995) provided a tutorial exposition of the Metropolis-Hastings algorithm. Gelman (1996) improved the convergence monitoring process in various ways to more effectively use the information in the Markov chain simulation and Roberts *et al.* (1997) considered the problem of scaling the proposal distribution of a multidimensional random walk Metropolis-Hastings algorithm, in order to maximize the efficiency of the algorithm.

In Section 8.4, the Metropolis-Hastings (M-H) algorithm will be explained and used to recover the unknown vector $\mathbf{f}$ from noisy and blurred data on the wavelet domain, based on a Bayesian model using different prior distributions. There are two types of model proposed; one is based on only a single prior parameter for all wavelet coefficients; and the other is based on a level-dependent prior, which means that there are different smoothing parameters that depend on the resolution level.

8.3 Bayesian modelling

This section is divided into seven parts; the first part gives a general introduction about the Bayesian approach; the second part introduces the likelihood using wavelet coefficients; the third part explains a prior for the noise variance; the fourth part introduces the single priors of wavelet coefficients; the fifth part models κ, γ and b; the sixth part introduces the multiple priors of wavelet coefficients; the final part models κ, γ and b for each resolution

level.

General

The key elements in the Bayesian approach are the likelihood function and the prior distribution, and hence the resulting posterior distribution. The likelihood is the conditional distribution of the data given the unknown parameters, denoted as $\mathrm{p}(\mathbf{y}|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of model parameter and $\mathbf{y}$ is a set of noisy data. The prior distribution, denoted $\mathrm{p}(\boldsymbol{\theta})$, quantifies detailed expert knowledge or general beliefs prior to data collection.

For estimation, evidence from the data and from prior beliefs are brought together by combining the likelihood and prior distribution, using Bayes's Theorem, to form the posterior distribution, defined as

$$\mathrm{p}(\boldsymbol{\theta}|\mathbf{y}) = \frac{\mathrm{p}(\mathbf{y}|\boldsymbol{\theta})\mathrm{p}(\boldsymbol{\theta})}{\mathrm{p}(\mathbf{y})},$$

where $\mathrm{p}(\mathbf{y})$ is a normalizing constant. Note that since this usually involves a high-dimensional integral it will be unacceptably time-consuming to perform the calculation. Fortunately, the normalising constant contains no information about the unknowns and hence can be dropped, giving the key statement

$$\mathrm{p}(\boldsymbol{\theta}|\mathbf{y}) \propto \mathrm{p}(\mathbf{y}|\boldsymbol{\theta})\mathrm{p}(\boldsymbol{\theta}),$$

meaning that "posterior" is proportional to "likelihood" multiplied by "prior". This distribution incorporates evidence from the data and knowledge from the prior distribution allowing an estimation process which balances the two types of information.

When there are multiple groups of parameters, these parameters will be assumed to be independent and modelled separately. Hence, if $\boldsymbol{\theta}$ is made-up of two sub-sets, say, with $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, then the previous equation becomes

$$\mathrm{p}(\boldsymbol{\theta}|\mathbf{y}) = \mathrm{p}(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}|\mathbf{y}) \propto \mathrm{p}(\mathbf{y}|\boldsymbol{\theta_1}, \boldsymbol{\theta_2})\mathrm{p}(\boldsymbol{\theta_1})\mathrm{p}(\boldsymbol{\theta_2}).$$

In the Bayesian setting, the posterior distribution is the basis for estimation and hence a point estimate can be found, for example, using the value that corresponds to the maximum of the posterior distribution, called the maximum a posteriori estimator (MAP)

$$\widehat{\boldsymbol{\theta}}_{\text{MAP}} = \underset{\boldsymbol{\theta}}{\text{argmin}} \quad \text{p}(\boldsymbol{\theta}|\mathbf{y}).$$

Other point estimates can also be used, such as the posterior mean or the posterior median. In addition, the joint posterior distribution can be examined, for example to construct marginal posterior distributions, or to calculate Bayesian credible intervals.

Likelihood using wavelet coefficients

Consider the model

$$\mathbf{y} = \mathbf{f} + \boldsymbol{\epsilon},$$

where the error $\boldsymbol{\epsilon}$ is a vector of random variables, such that $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I_n})$, $\mathbf{f} = \{f(i/m) : i = 1, 2, \ldots, m\}$ is a vector of values of some unknown function at a set of m equally-spaced locations, and $\mathbf{y} = \{y_i : i = 1, 2, \ldots, m\}$ are observed data values recorded at the same locations.

Wavelets are a common choice for this type of non-parametric regression problem when noise removal or a multi-resolution analysis is required. Let $\mathbf{W}$ be an orthogonal matrix holding an appropriate (decimated) discrete wavelet basis, with Haar wavelets. The wavelet decomposition of the data $\mathbf{y}$, can be written as

$$\mathbf{d_y} = \mathbf{W}\mathbf{y} = \mathbf{W}(\mathbf{f} + \boldsymbol{\epsilon}) = \mathbf{W}\mathbf{f} + \mathbf{W}\boldsymbol{\epsilon} = \mathbf{d_f} + \boldsymbol{\eta},$$

where $\mathbf{d_y}$ and $\mathbf{d_f}$ are vectors of the wavelet coefficients of $\mathbf{y}$ and $\mathbf{f}$ respectively. Also, orthogonality of $\mathbf{W}$ and normality of the noise vector $\boldsymbol{\epsilon}$ implies that the noise vector $\boldsymbol{\eta}$ is also normal. This shows that noise in the measurements results in corresponding noise in the wavelet coefficients.

It is a common approach to say that fine level coefficients are the result of noise, with the signal being represented in a small number of low-level coefficient values. The method of

wavelet thresholding can then be used to set the small coefficients to zero, or shrinkage can be used to shrink the coefficient values closer to zero. A set of modified coefficient values, $\mathbf{d}^*_{\mathbf{y}}$, after thresholding or shrinkage can be used as an estimate of the wavelet coefficients of $\mathbf{f}$, that is $\widehat{\mathbf{d}}_{\mathbf{f}} = \mathbf{d}^*_{\mathbf{y}}$, with the resulting estimate of $\mathbf{f}$, defined as

$$\widehat{\mathbf{f}} = \mathbf{W}^T \mathbf{d}^*_{\mathbf{y}}.$$

This denoising method can also be given an interpretation in a Bayesian setting, which is the approach followed later.

The overall aim in a general linear inverse problem is also to estimate an unknown function from a finite set of measurements but these quantities are related through a convolution equation, such as

$$\mathbf{y} = \mathbf{g} + \boldsymbol{\epsilon} = \mathbf{H}\mathbf{f} + \boldsymbol{\epsilon}, \tag{8.1}$$

where $\mathbf{H}_{n\times m}$ is a given matrix, with elements $h_{i,j}$, and $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$, and $\boldsymbol{\epsilon}$ is some error vector as above. The matrix, $\mathbf{H}$, can be defined as a spread or transfer function. If the matrix, $\mathbf{H}$, is square and invertible then the solution is

$$\widehat{\mathbf{f}} = \mathbf{H}^{-1}\mathbf{y}.$$

Assuming an additive Gaussian error model, then the conditional distribution of the data given the truth is

$$\mathbf{y}|\mathbf{f}, \sigma^2 \sim N(\mathbf{H}\mathbf{f}, \sigma^2 \mathbf{I}_n),$$

with likelihood

$$\mathrm{p}(\mathbf{y}|\mathbf{f}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{2\sigma^2} ||\mathbf{y} - \mathbf{H}\mathbf{f}||_2^2 \right\}, \quad \mathbf{y} \subset \mathbb{R}^n, \mathbf{f} \subset \mathbb{R}^m; \sigma > 0, \tag{8.2}$$

where $||.||_2$ is the L_2 norm and $\mathbf{f} = \{f_j : j = 1, 2, \ldots, m\}$. It is often not possible to reliably estimate $\mathbf{f}$ using the likelihood alone and so previous approaches have used Bayesian modelling with smoothing prior distributions directly on the unknown $\mathbf{f}$, such as Allum *et al.* (1999).

The corresponding form of the likelihood, using wavelet coefficients, is given by

$$\mathrm{p}(\mathbf{y}|\mathbf{d_f},\sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{2\sigma^2}||\mathbf{y}-\mathbf{H}\mathbf{W}^T\mathbf{d_f}||_2^2\right\}, \quad \mathbf{y}\subset\mathbb{R}^n, \mathbf{d_f}\subset\mathbb{R}^m; \sigma>0, \tag{8.3}$$

where $||.||_2$ is the L_2 norm, $\mathbf{d_f}$ is a vector of values of wavelet coefficients of the unknown and hence $\mathbf{g} = \mathbf{H}\mathbf{W}^T\mathbf{d_f}$.

A prior model for the noise variance

In applications, the variance of the noise σ^2 is estimated from the finest level of detail coefficients in the wavelet decomposition and plugged into the shrinkage rule. Here, the methodology is generalised by specifying a prior distribution on the unknown variance. The variance will be modelled by an inverse gamma distribution with parameters a_0 and b_0, $\sigma^2 \sim \mathbf{inverse-gamma}(a_0, b_0)$, with density

$$\mathrm{p}(\sigma^2) = \frac{1}{\Gamma(a_0)} \frac{b_0^{a_0}}{(\sigma^2)^{(a_0+1)}} \exp\left\{ -\frac{b_0}{\sigma^2}\right\}, \quad \sigma^2 \geq 0; a_0, b_0 > 0. \tag{8.4}$$

This approach follows that of Gelman (2006) and Cutillo *et al.* (2008) for modelling a noise variance. The marginal likelihood of the wavelet coefficients is t, which models heavy tails of the empirical distributions of wavelet coefficients (Cutillo *et al.*, 2008). The parameters a_0 and b_0 can be fixed based on knowledge or information from separate calibration experiments. In particular, an expert might provide a mean, σ_0^2, and variance, τ_0^2, for σ^2, which correspond to $a_0 = \frac{(\sigma_0^2)^2}{\tau_0^2} + 2$ and $b_0 = \sigma_0^2(a_0 - 1)$. Although this approach is general, in this thesis a value for σ_0^2 computed from Equation (2.62) and $\tau_0^2 = 1$ have been used.

A single component prior for wavelet coefficients

In this section four prior distributions, $\mathrm{p}(\mathbf{d_f})$, on the wavelet coefficients, $\mathbf{d_f}$, will be applied; the first is the Gaussian prior; this prior is used by Hoerl and Kennard (1970)

and Cutillo *et al.* (2008); the second is the Laplace prior, which was first used as a regularisation penally by Tibshirani (1996); the third is elastic-net, which was first used as a regularisation penally by Zou and Hastie (2005); finally, the double Weibull prior is suggested by Reményi and Vidakovic (2015). As prior distributions for κ_j, γ_j and b_j, we chose a gamma prior on κ_j describing the strength of shrinkage of a wavelet coefficient. Mixing weight, γ_j, is described by a beta prior and an inverse gamma prior on the b_j describes the variance of the double Weibull distribution.

Gaussian distribution

An obvious choice of prior is the Gaussian distribution, $\mathrm{d_f} \sim N(0, \frac{1}{2\kappa})$, with density

$$\mathrm{p}(\mathbf{d_f}|\kappa) = \left(\sqrt{\frac{\kappa}{\pi}}\right)^m \exp\left\{ -\kappa||\mathbf{d_f}||_2^2 \right\}, \quad \mathbf{d_f} \subset \mathbb{R}^m; \kappa > 0, \tag{8.5}$$

where $||.||_2$ is the L_2 norm and κ is the shrinkage parameter. This approach follows that of Hoerl and Kennard (1970), and Cutillo *et al.* (2008) for modelling wavelet coefficients. The reason for this assumption is to create a shrinkage rule, which keeps the important information (Zou and Hastie, 2005).

Laplace distribution

The second choice is the Laplace distribution, which was introduced by Laplace (1774) and is used as a regularisation penalty by Tibshirani (1996), $\mathrm{d_f} \sim \mathcal{DE}(0, \frac{1}{\kappa})$, with density

$$\mathrm{p}(\mathbf{d_f}|\kappa) = \left(\frac{\kappa}{2}\right)^m \exp\left\{ -\kappa||\mathbf{d_f}||_1 \right\}, \quad \mathbf{d_f} \subset \mathbb{R}^m; \kappa > 0, \tag{8.6}$$

where $||.||_1$ is the L_1 norm. Moreover, this prior was used by Vidakovic and Ruggeri (2001) and Johnstone and Silverman (2005a) in the wavelet domain. The reason behind the choice of Laplace is that in practice, wavelet coefficients have heavier tails than a Gaussian distribution (Donoho and Johnstone, 1994).

Elastic-net distribution

As a convex combination of the Laplace and the Gaussian, a density based on the *elastic-net* function might be suitable. The main benefit of the elastic-net is that it might provide a better representation when the number of parameters, m, is bigger than the number of observations, n, (Zou and Hastie, 2005). Thus, the elastic-net is an automatic variable selection and a continuous shrinkage rule, (Zou and Hastie, 2005). The elastic-net has a parameter γ and for the limiting values of γ, this reduces to the Gaussian case ($\gamma = 1$) and the Laplace case ($\gamma = 0$). So, we believe that at the highest resolution level the value of γ should be close to zero and close to 1 for the lowest resolution level.

The elastic-net distribution for the wavelet coefficients $\mathbf{d_f}$ given λ and γ, can be defined as

$$\mathrm{p}(\mathbf{d_f}|\kappa,\gamma) = \left(\frac{1}{Z(\kappa,\gamma)}\right)^m \exp\left\{ -\kappa(\gamma||\mathbf{d_f}||_2^2 + (1-\gamma)||\mathbf{d_f}||_1)\right\},$$
$$\mathrm{d_f} \in \mathbb{R}; \kappa > 0, 0 < \gamma < 1, \tag{8.7}$$

where

$$Z(\kappa,\gamma) = \begin{cases} 2/\kappa, & \gamma = 0 \\ \sqrt{\frac{4\pi}{\kappa\gamma}} \exp\left\{\frac{1}{4\gamma}\kappa(1-\gamma)^2\right\}\left(1 - \Phi(\frac{\kappa(1-\gamma)}{\sqrt{2\kappa\gamma}})\right), & 0 < \gamma < 1 \\ \sqrt{\pi/\kappa}, & \gamma = 1. \end{cases} \tag{8.8}$$

Double Weibull distribution

The fourth choice is the double Weibull distribution, which is chosen for its ability to mimic the features of a prior consisting of a point mass at zero and heavy tailed part. The double Weibull distribution can be defined as

$$\mathrm{p}(\mathbf{d_f}|\mathrm{b},c) = \prod_1^m \frac{c}{2\mathrm{b}}|\mathbf{d_f}|^{c-1} \exp\left\{ -\frac{1}{\mathrm{b}}|\mathbf{d_f}|^c\right\}, \quad \mathbf{d_f} \subset \mathbb{R}^m; \mathrm{b} > 0, c > 0. \tag{8.9}$$

Reményi and Vidakovic (2015) suggested that the constant c should be equal to 1/3, since this value gives a small risk.

Prior parameters κ, γ and b

There are different ways to estimate the parameters of the Gaussian, the Laplace and the elastic-net priors. The first is to use minimum mean squared-error (MMSE) approach, which will be explained in Section 8.4, if the true $\mathbf{f}$ is known. For some problems there might be realistic simulated data that can be used to estimate the parameters κ and γ. Clearly, the simulated data should reflect correct belief about the true $\mathbf{f}$ otherwise an appropriate estimate will be produced.

The second method is to estimate the parameters along with the wavelet coefficients. Hence, each of the prior distributions has introduced additional parameters that must also be modelled, which will also be explained in Section 8.4. In all cases, the parameter κ will follow the gamma distribution, $\kappa \sim \mathbf{gamma}(a_1, b_1)$, with density

$$\mathrm{p}(\kappa) = \frac{1}{\Gamma(a_1)} b_1^{a_1} \kappa^{a_1 - 1} \exp\left\{ - b_1 \kappa \right\}, \quad \kappa \geq 0; a_1, b_1 > 0. \tag{8.10}$$

This approach follows that of Park and Casella (2008) and Kyung *et al.* (2010). As with σ^2, the parameters a_1 and b_1 can be fixed based on knowledge or information from separate calibration experiments. In particular, an expert might provide a mean, κ_1, and variance, τ_1^2, for κ, which correspond to $a_1 = \frac{{\kappa_1}^2}{\tau_1^2}$ and $b_1 = \frac{\kappa_1}{\tau_1^2}$. Although this approach is general, in this thesis a value for κ_1 computed from Equation (2.61) and $\tau_1^2 = 1$ has been used.

Finally, given that γ can only take values within the range $[0, 1]$, the beta distribution is a sensible choice for a prior model, $\gamma \sim \mathbf{Beta}(a_2, b_2)$, with density

$$\mathrm{p}(\gamma) = \frac{\Gamma(a_2 + b_2)}{\Gamma(a_2)\Gamma(b_2)} \gamma^{a_2 - 1} (1 - \gamma)^{b_2 - 1}, \quad 0 < \gamma < 1; a_2, b_2 > 0, \tag{8.11}$$

the parameters a_2 and b_2 fixed based on knowledge or information from separate calibration experiments. In particular, an expert might provide a mean, γ_2, and variance, τ_2^2, for γ, which correspond to $a_2 = \frac{b_2\, \gamma_2}{1 - \gamma_2}$ and $b_2 = \frac{(1-\gamma_2)^2 \gamma_2 - \tau_2^2 (1 - \gamma_2)}{\tau_2^2}$. Although this approach is general, in this thesis $\gamma_2 = 0.5$ and $\tau_2^2 = 0.01$ have been used, giving $a_1 = 12$ and $b_1 = 12$.

In the case of the double Weibull distribution, the parameter b will follow an inverse

gamma distribution, $\mathrm{b} \sim \mathbf{inverse-gamma}(a_3, b_3)$, with density

$$\mathrm{p(b)} = \frac{1}{\Gamma(a_3)} \frac{b_3^{a_3}}{\mathrm{b}^{(a_3+1)}} \exp\left\{ -\frac{b_3}{\mathrm{b}} \right\}, \quad \mathrm{b} \geq 0; a_3, b_3 > 0, \tag{8.12}$$

As with σ^2, the parameters a_3 and b_3 can be fixed based on knowledge or information from separate calibration experiments. In particular, an expert might provide a mean, b_0, and variance, τ_0^3, for b, which correspond to $a_3 = \frac{{\mathrm{b}_0}^2}{\tau_3^2} + 2$ and $b_3 = \mathrm{b}_0(a_0 - 1)$. Although this approach is general, in this thesis $\tau_3^2 = 1$ has been used and b_0 is computed from Equation (5.13) as described in Chapter 5, unless $\widehat{\sigma}^2 > \widehat{\sigma}_d^2$ when we set $\mathrm{b}_0 = 0.5$ (Gelman, 2006). Here $\widehat{\sigma}^2$ and $\widehat{\sigma}_d^2$ are the variances of noise and wavelet coefficient at finest resolution level, respectively.

Hence, there are multiple parameters, which are assumed independent and which have been modelled separately. So, the posterior distribution with Gaussian prior for the wavelet coefficients $\mathbf{d_f}$, σ^2 and κ given $\mathbf{y}$, is given by

$$\mathrm{p}(\mathbf{d_f}, \sigma^2, \kappa | \mathbf{y}) \propto \mathrm{p}(\mathbf{y} | \mathbf{d_f}, \sigma^2) \mathrm{p}(\mathbf{d_f} | \kappa) \mathrm{p}(\sigma^2) \mathrm{p}(\kappa), \tag{8.13}$$

and the posterior distribution for the wavelet coefficients $\mathbf{d_g}$, σ^2 and κ given $\mathbf{d_y}$, is given by

$$\mathrm{p}(\mathbf{d_g}, \sigma^2, \kappa | \mathbf{d_y}) \propto \mathrm{p}(\mathbf{d_y} | \mathbf{d_g}, \sigma^2) \mathrm{p}(\mathbf{d_g} | \kappa) \mathrm{p}(\sigma^2) \mathrm{p}(\kappa). \tag{8.14}$$

In the following equations, individual wavelet coefficients will be denoted $\mathrm{d_g}$ and $\mathrm{d_y}$, omitting the double index j, l to simplify notation.

If the wavelet coefficient $\mathrm{d_g}$, are assumed independently $N(0, \frac{1}{2\kappa})$ distributed, then the distribution of the wavelet coefficient $\mathrm{d_g}$ given the corresponding wavelet coefficient $\mathrm{d_y}$, becomes

$$\begin{aligned}
\mathrm{p}(\mathrm{d_g} | \mathrm{d_y}) \propto & \int_0^\infty \int_0^\infty \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) \exp\left\{ -\frac{1}{2\sigma^2} (\mathrm{d_y} - \mathrm{d_g})^2 \right\} \left(\sqrt{\frac{\kappa}{\pi}} \right) \exp\left\{ -\kappa {\mathrm{d_g}}^2 \right\} \\
& \times \frac{1}{\Gamma(a_0)} \frac{b_0^{a_0}}{(\sigma^2)^{(a_0+1)}} \exp\left\{ -\frac{b_0}{\sigma^2} \right\} \frac{1}{\Gamma(a_1)} b_1^{a_1} \kappa^{a_1 - 1} \exp\left\{ -b_1 \kappa \right\} d(\sigma^2) d\kappa \\
& \propto \frac{1}{[1 + \frac{1}{2b_0}(\mathrm{d_y} - \mathrm{d_g})^2]^{a_0 + \frac{1}{2}} [\frac{{\mathrm{d_g}}^2}{b_1} + 1]^{a_1 + \frac{1}{2}}}.
\end{aligned} \tag{8.15}$$

If the wavelet coefficients, $\mathrm{d_g}$, are assumed independently $\mathcal{DE}(0, \frac{1}{\kappa})$ distributed, then the posterior of wavelet coefficient $\mathrm{d_g}$ given the corresponding wavelet coefficient $\mathrm{d_y}$, becomes

$$\begin{aligned}\mathrm{p}(\mathrm{d_g}|\mathrm{d_y}) \propto & \int_0^\infty \int_0^\infty \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) \exp\left\{ -\frac{1}{2\sigma^2}(\mathrm{d_y}-\mathrm{d_g})^2 \right\} \left(\frac{\kappa}{2}\right) \exp\left\{ -\kappa|\mathrm{d_g}| \right\} \\ & \times \frac{1}{\Gamma(a_0)} \frac{b_0^{a_0}}{(\sigma^2)^{(a_0+1)}} \exp\left\{ -\frac{b_0}{\sigma^2} \right\} \frac{1}{\Gamma(a_1)} b_1^{a_1} \kappa^{a_1-1} \exp\left\{ -b_1\kappa \right\} d(\sigma^2) d\kappa \\ & \propto \frac{1}{[1+\frac{1}{2b_0}(\mathrm{d_y}-\mathrm{d_g})^2]^{a_0+\frac{1}{2}} [\frac{|\mathrm{d_g}|}{b_1}+1]^{a_1+1}}. \end{aligned} \tag{8.16}$$

If the wavelet coefficients, $\mathrm{d_g}$, are assumed independently $\mathcal{DW}(b, c)$ distributed, then the posterior of wavelet coefficient $\mathrm{d_g}$ given the corresponding wavelet coefficient $\mathrm{d_y}$, becomes

$$\begin{aligned}\mathrm{p}(\mathrm{d_g}|\mathrm{d_y}) \propto & \int_0^\infty \int_0^\infty \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) \exp\left\{ -\frac{1}{2\sigma^2}(\mathrm{d_y}-\mathrm{d_g})^2 \right\} \frac{c}{2b} |\mathrm{d_g}|^{c-1} \exp\left\{ -\frac{1}{b}|\mathrm{d_g}|^c \right\} \\ & \times \frac{1}{\Gamma(a_0)} \frac{b_0^{a_0}}{(\sigma^2)^{(a_0+1)}} \exp\left\{ -\frac{b_0}{\sigma^2} \right\} \frac{1}{\Gamma(a_3)} \frac{b_3^{a_3}}{b^{(a_3+1)}} \exp\left\{ -\frac{b_3}{b} \right\} d(\sigma^2) d(b) \\ & \propto \frac{|\mathrm{d_g}|^{c-1}}{[1+\frac{1}{2b_0}(\mathrm{d_y}-\mathrm{d_g})^2]^{a_0+\frac{1}{2}} [\frac{|\mathrm{d_g}|^c}{b_3}+1]^{a_3+1}}. \end{aligned} \tag{8.17}$$

The model with elastic-net distribution (8.17) is more complex so that the integration required is too complicated. The resulting wavelet reconstruction problem is also more complex, with numerical methods and the Metropolis-Hastings algorithm often being used.

Multiple-component priors for wavelet coefficients

As an extension, the various coefficients are grouped by wavelet resolution level with the obvious extensions to the definitions given in the previous section. For the wavelet coefficients $\mathbf{d_f}_j$, at level j, the Gaussian prior density function becomes

$$\mathrm{p}(\mathbf{d_f}_j|\kappa_j) = \left(\frac{\kappa_j}{\pi}\right)^{(2^j-1)/2} \exp\left\{ -\kappa_j \sum_{l=0}^{2^j-1} \mathrm{d}_{\mathrm{f}j,l}^2 \right\}, \quad \mathbf{d_f}_j \subset \mathbb{R}^{2^j-1}; \kappa_j > 0, \tag{8.18}$$

where $\boldsymbol{\kappa} = \{\kappa_j : j = 0, 1, \ldots, J-1\}$ with $J = \log_2(m)$ and $\mathbf{d_f}_j$ is a vector of values of wavelet coefficients at level j. The Laplace prior density function becomes

$$\mathrm{p}(\mathbf{d_f}_j|\kappa_j) = \left(\frac{\kappa_j}{2}\right)^{2^j-1} \exp\left\{ -\kappa_j \sum_{l=0}^{2^j-1} |\mathrm{d_f}_{j,l}| \right\}, \quad \mathbf{d_f}_j \subset \mathbb{R}^{2^{j-1}}; \kappa_j > 0. \tag{8.19}$$

The model based on the elastic-net has density function given by

$$\mathrm{p}(\mathbf{d_f}_j|\kappa_j, \gamma_j) = \left(\frac{1}{Z(\kappa_j, \gamma_j)}\right)^{2^j-1} \exp\left\{ -\kappa_j \sum_{l=0}^{2^j-1} \left(\gamma_j \mathrm{d}_{\mathrm{f}j,l}^2 + (1-\gamma_j)|\mathrm{d_f}_{j,l}|\right)\right\},$$
$$\mathbf{d_f}_j \subset \mathbb{R}^{2^{j-1}}; \kappa_j > 0, 0 < \gamma_j < 1, \tag{8.20}$$

where

$$Z(\kappa_j, \gamma_j) = \begin{cases} 2/\kappa_j, & \gamma_j = 0 \\ \sqrt{\frac{4\pi}{\kappa_j \gamma_j}} \exp\left\{\frac{\kappa_j(1-\gamma_j)^2}{4\gamma_j}\right\}\left(1 - \Phi(\frac{\kappa_j(1-\gamma_j)}{\sqrt{2\kappa_j\gamma_j}})\right), & 0 < \gamma_j < 1 \\ \sqrt{\pi/\kappa_j}, & \gamma_j = 1. \end{cases} \tag{8.21}$$

Finally, for the double Weibull distribution, the density function becomes

$$\mathrm{p}(\mathbf{d_f}_j|\mathrm{b}_j, c) = \prod_0^{2^j-1} \frac{c}{2\mathrm{b}_j} |\mathbf{d_f}_j|^{c-1} \exp\left\{ -\frac{1}{\mathrm{b}_j}|\mathbf{d_f}_j|^c \right\}, \quad \mathbf{d_f}_j \subset \mathbb{R}^{2^{j-1}}; \mathrm{b}_j > 0, c > 0. \tag{8.22}$$

Prior parameters κ_j, γ_j and b_j

The prior densities become

$$\mathrm{p}(\kappa_j) = \frac{1}{\Gamma(a_1)} b_1^{a_1} {\kappa_j}^{a_1-1} \exp\left\{ -b_1\kappa_j \right\}, \quad \kappa_j \geq 0, j = 0, 1, \ldots, J-1; a_1, b_1 > 0, \tag{8.23}$$

$$\mathrm{p}(\gamma_j) = \frac{\Gamma(a_2+b_2)}{\Gamma(a_2)\Gamma(b_2)} {\gamma_j}^{a_2-1}(1-\gamma_j)^{b_2-1}, \quad 0 < \gamma_j < 1, j = 0, 1, \ldots, J-1; a_2, b_2 > 0, \tag{8.24}$$

and

$$\mathrm{p}(\mathrm{b}_j) = \frac{1}{\Gamma(a_3)} \frac{b_3^{a_3}}{{\mathrm{b}_j}^{(a_3+1)}} \exp\left\{ -\frac{b_3}{\mathrm{b}_j} \right\}, \quad \mathrm{b}_j \geq 0, j = 0, 1, \ldots, J-1; a_3, b_3 > 0. \tag{8.25}$$

The hyper parameters a_1, b_1, a_2, b_2, a_3 and b_3 can be fixed for all levels at the same values as chosen in the single component prior.

There are multiple parameters, which are assumed independent and have been modelled separately. So, the posterior distribution with prior Gaussian for the wavelet coefficients $\mathbf{d_f}$, σ^2 and κ given $\mathbf{y}$, is given by

$$\mathrm{p}(\mathbf{d_f}, \sigma^2, \boldsymbol{\kappa}|\mathbf{y}) \propto \mathrm{p}(\mathbf{y}|\mathbf{d_f}, \sigma^2)\mathrm{p}(\mathbf{d_{f0}}|\kappa_0)\mathrm{p}(\mathbf{d_{f1}}|\kappa_1)\ldots\mathrm{p}(\mathbf{d_{f}}_{J-1}|\kappa_{J-1})\mathrm{p}(\sigma^2)\mathrm{p}(\kappa_0)\mathrm{p}(\kappa_0)\ldots\mathrm{p}(\kappa_{J-1}), \tag{8.26}$$

where $\mathbf{d_{f}}_j$ are the wavelet coefficients at resolution level j, $j = 0, 1, \ldots, J-1$.

8.4 Numerical methods

In this section, three estimation approaches will be explained; the first is the PM estimate; the second is MAP estimate and the third MAP estimation using prior parameter chosen by MMSE.

Metropolis-Hastings sampling

In this thesis, the estimation of the underlying $\mathbf{f}$ is based on the approximate posterior distribution computations using a standard Metropolis-Hastings (M-H) algorithm. This is a special case of the Markov chain Monte Carlo (MCMC) approach, whose use has become widespread in the general statistical literature. The M-H algorithm is the first example of a MCMC approach used for parameter estimation and was proposed by Metropolis *et al.* (1953) and subsequently generalized by Hastings (1970).

The use of such methods for parameter estimation, and general density exploration, is widespread; a review can be found in Robert and Casella (2011), and for theoretical details see Gamerman and Lopes (2006), Liu (2001) and Brooks *et al.* (2011). For general practical examples see the collection by Gilks *et al.* (1996).

Based on the various model definitions in the previous section the parameter vector will simply be referred to as $\boldsymbol{\theta}$, which will represent $(\mathbf{d_f}, \sigma^2, \kappa)$, $(\mathbf{d_f}, \sigma^2, \boldsymbol{\kappa})$, $(\mathbf{d_f}, \sigma^2, \kappa, \gamma)$ and $(\mathbf{d_f}, \sigma^2, \boldsymbol{\kappa}, \boldsymbol{\gamma})$, with m simply counting the total number of parameters.

The Markov chain can start at any feasible point in the parameter space, let this arbitrary value be denoted as $\boldsymbol{\theta}^0$. From this starting value a discrete time Markov chain is simulated to produce values $\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \ldots, \boldsymbol{\theta}^K$. The algorithm will now be defined, and is also summarised in Figure 8.1. Also, the algorithm of the MCMC approach is shown in Algorithm (3). Consider the Markov chain transition from state $\boldsymbol{\theta}^{k-1}$ at time $k-1$, to state $\boldsymbol{\theta}^k$ at

---

Set an initial value for $\boldsymbol{\theta}$, call this $\boldsymbol{\theta}^0$

Repeat the following steps for $k = 1, \ldots, K$

Repeat the following steps for $i = 1, \ldots, m$

Generate from a Gaussian distribution $N(0, \tau^2)$

Generate a propose new value $\theta^* = \theta_i^k + \epsilon$

Evaluate

$$\alpha = \alpha(\boldsymbol{\theta}^k | (\boldsymbol{\theta})^*) = \min\left\{1, \frac{\mathrm{p}(\theta_1^k, \ldots, \theta_{i-1}^k, \theta_i^*, \theta_{i+1}^{k-1}, \ldots, \theta_m^{k-1} | \mathbf{y})}{\mathrm{p}(\theta_1^k, \ldots, \theta_{i-1}^k, \theta_i^{k-1}, \theta_{i+1}^{k-1}, \ldots, \theta_m^{k-1} | \mathbf{y})}\right\}$$

Generate u from a uniform distribution, $U(0,1)$

If $\alpha > u$ then accept the proposal and set $\theta_i^k = \theta_i^*$, else $\theta_i^k = \theta_i^{k-1}$

End repeat

End repeat

Discard initial values and use remainder to make inference.

Figure 8.1: Single-variable random walk MCMC Algorithm (Aykroyd, 2015).

time k. This is one of the simplest schemes, which works well for many applications based on a random walk and uses separate single variable updates. That is, at each step only the value of a single variable is proposed and the proposal is a perturbation of the current value with spread parameters chosen to achieve an acceptable convergence rate. Suppose that a new value for θ_i is being proposed, then $\theta_i^k = \theta_i^{k-1} + \epsilon$, and an obvious choice is $\epsilon \sim N(0, \tau^2)$. This proposal is accepted, with probability

$$\min\left\{1, \frac{\mathrm{p}(\boldsymbol{\theta}^k | \mathbf{y})}{\mathrm{p}(\boldsymbol{\theta}^{k-1} | \mathbf{y})}\right\},$$

otherwise the value is reset with $\theta^k = \theta^{k-1}$.

The components of $\boldsymbol{\theta}$ are of the different types, $\mathbf{d_f}, \sigma^2$, κ and γ, each allowing different simplifications of the above acceptance probability. To explain this, each type will be considered separately.

$\mathbf{d_f}$ updates: Let the current set of wavelet coefficients be $\mathbf{d_f}$. A proposed new value for one of the wavelet coefficients $\mathrm{d_f}'$, is drawn from a proposal distribution, $q(\mathrm{d_f}'|\mathrm{d_f})$, where $\mathbf{d_f}'$ is the set of wavelet coefficients including the proposed new value $\mathrm{d_f}'$. Moreover, q is a normal distribution centred on the current coefficient value, with the spread parameter, τ_1^2, chosen to achieve acceptable convergence rates. The proposal is accepted, and the parameter values are updated accordingly, with probability

$$\min\left\{1, \frac{\mathrm{p}(\mathbf{d_f}', \sigma^2, \kappa, \gamma|\mathbf{y})q(\mathbf{d_f}'|\mathbf{d_f})}{\mathrm{p}(\mathbf{d_f}, \sigma^2, \kappa, \gamma|\mathbf{y})q(\mathbf{d_f}|\mathbf{d_f}')}\right\},$$

otherwise it is rejected and no change is made. Note that in our case the proposal distribution is symmetric, that is $q(\mathbf{d_f}'|\mathbf{d_f}) = q(\mathbf{d_f}|\mathbf{d_f}')$, hence the ratio of these terms remove in the above expression. Then the acceptance probability can be written as

$$\min\left\{1, \frac{\mathrm{p}(\mathbf{y}|\mathbf{d_f}', \sigma^2)\mathrm{p}(\mathbf{d_f}'|\kappa, \gamma)}{\mathrm{p}(\mathbf{y}|\mathbf{d_f}, \sigma^2)\mathrm{p}(\mathbf{d_f}|\kappa, \gamma)}\right\},$$

where each of the wavelet coefficients $\mathbf{d_f}$, considered in the same way.

σ^2 updates: A proposed new value of the variance of noise σ'^2 is drawn from normal distribution, centred on the current parameter value, with the spread parameter, τ_2^2, chosen to achieve an acceptable convergence rate. Here negative proposals are rejected and, if positive, the proposal is accepted, with probability

$$\min\left\{1, \frac{\mathrm{p}(\mathbf{y}|\mathbf{d_f}, \sigma'^2)\mathrm{p}(\sigma'^2)}{\mathrm{p}(\mathbf{y}|\mathbf{d_f}, \sigma^2)\mathrm{p}(\sigma^2)}\right\},$$

otherwise it is rejected and no change is made.

κ updates: A proposed new value of the elastic-net parameter κ' for κ is drawn from normal distribution centred on the current parameter value, with spread parameter, τ_3^2, chosen to achieve an acceptable convergence rate. Here negative proposals are rejected and, if positive the proposal is accepted, with probability

$$\min\left\{1, \frac{\mathrm{p}(\mathbf{d_f}|\kappa', \gamma)\mathrm{p}(\kappa')}{\mathrm{p}(\mathbf{d_f}|\kappa, \gamma)\mathrm{p}(\kappa)}\right\},$$

otherwise it is rejected and no change is made.

γ updates: A proposed new value of the elastic-net weight γ' for γ is drawn from uniform distribution on the interval [0,1] centred on the current parameter value, with the spread parameter, τ_4^2, chosen to achieve an acceptable convergence rate, with probability

$$\min\left\{1, \frac{\mathrm{p}(\mathbf{d_f}|\kappa, \gamma')\mathrm{p}(\gamma')}{\mathrm{p}(\mathbf{d_f}|\kappa, \gamma)\mathrm{p}(\gamma)}\right\},$$

otherwise it is rejected and no change is made.

It is important to realise that both low and high values of τ_1^2, τ_2^2, τ_3^2 and τ_4^2, lead to long transient periods and highly correlated samples and hence unreliable estimations (Aykroyd, 2015). A reasonable proposal variance can be chosen adaptively during the early burn-in period, and it has been proven theoretically that for a wide variety of high-dimensional problems an acceptance rate of 23.4% is optimal (Roberts *et al.*, 1997).

Although, the theoretical derivation is complicated, the statement and implementation of the algorithm is usually straightforward. If the algorithm is designed carefully, then as the iterations progress, the current parameter set does not depend on the starting values, and the subsequent values can be treated as a correlated sample from the posterior distribution. Key issues then become how to judge, when the initial transient behaviour has ended, when the chain is in equilibrium, and how many iterations to perform to have a sufficiently large sample for a reliable estimation. Aykroyd (2015) showed that checking

Markov chain paths and calculating sample autocorrelation functions might provide good estimation, where the paths should look "random" and the autocorrelation functions should be close to zero for all except small lags. Aykroyd and Mardia (2003) stated that the estimates produced will have an asymptotic variance $\text{var}(d_{j,l}) = \varrho\sigma^2/M$, according to dependence within the Markov chain, where M is the sample size, σ^2 is the sampling variance of $d_{j,l}$, ϱ the integrated autocorrelation time is given by $\varrho = \sum_{t=-\infty}^{\infty} \rho(t)$ and $\rho(t)$ is the autocorrelation function of the process. This variance is a factor ϱ times greater than would be the case with an independent sample. The sample size, M, will be chosen so that the Monte Carlo variance is less than 1% of the sampling variance of the estimator, that is choose M to satisfy $\text{var}(d_{j,l})/\sigma^2 = \varrho/M < 1/100$. The value of ϱ can be estimated using the truncated periodogram estimator $\hat{k} = \sum_{|t|\leq T} \hat{\rho}(t)$, with window width T chosen as the minimum integer such that $T \geq 3\hat{\varrho}$. A variety of more formal convergence diagnostics are available, see for example Raftery and Lewis (1995), Cowles and Carlin (1996) and Geyer (2011).

Once the sample has been generated from the posterior distribution, a number of possible estimators are available. Let $\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \ldots, \boldsymbol{\theta}^N$ be the MCMC sample collected after the equilibrium of the Markov chain has been declared, then the posterior mean and variances can be estimated by the sample mean and variance.

$$\widehat{\theta}_i = \frac{1}{k}\sum_{k=1}^{K} \theta_i^k, \qquad \widehat{\sigma}^2 = \frac{1}{K-1}\sum_{k=1}^{K}(\theta_i^k - \bar{\theta}_i)^2 \quad i = 1, 2, \ldots, m.$$

Alternatively, the posterior summaries can be computed using the median and the 95% credible interval

$$\widehat{\mathbf{d}}_{\mathbf{f}i} = \text{median}\{\mathbf{d}_{\mathbf{f}i}^k : k = 1, 2, \ldots, K\} = \mathbf{d}_{\mathbf{f}i}^{(50)} \quad \text{and} \quad (\mathbf{d}_{\mathbf{f}i}^{(2.5)}, \mathbf{d}_{\mathbf{f}i}^{(97.5)}) \quad \text{for} \quad i = 1, 2, \ldots, m, \tag{8.27}$$

where $\mathbf{d}_{\mathbf{f}}{}^{(\mathcal{P})}$ indicates the $\mathcal{P}$-percentile of the sample $\{\mathbf{d}_{\mathbf{f}}{}^k : k = 1, 2, \ldots, K\}$ (see for example Aykroyd and Mardia, 2003, and Aykroyd, 2015).

To calculate an approximate MAP estimate, the MCMC algorithm can be converted into a simulated annealing algorithm (Geman and Geman, 1984). In particular a temperature, T_k, is included, which decreases as the iterations progress, with $T_k = 2/\log(1+k)$ being one choice of annealing schedule. Hence, the acceptance ratio, α, is replaced by α^{T_k}. Note that, the MAP estimate is taken as the final iteration, $\widehat{\boldsymbol{\theta}}_{\text{MAP}} = \widehat{\boldsymbol{\theta}}^k$. The advantage of simulated annealing is that it provides an answer more quickly than a sampling algorithm. On the other hand, the disadvantage is that it does not produce a posterior sample for further investigation (Aykroyd, 2015). Algorithm 4, in this chapter, shows the main idea of the MAP estimation algorithm.

Minimum mean squared-error estimation (MMSE)

There are several different ways to estimate the prior parameters, one of which is MMSE. This process is recommended if a set of true training functions are available, which can be used to estimate the parameters κ, γ and σ^2. In archaeological problems, there are five simulated cores, that can be used to estimate the prior parameters, which will be explained in Chapter 9. The reason for choosing the MMSE is that, in practice, it is difficult to estimate the prior parameters. However, training on realistic simulated data can be used to assess beliefs about real data and the MMSE will be employed to determine the optimal parameters, from which these parameters can be used to reconstruct $\mathbf{f}$ from real data. Hence, there are no prior distributions for the prior parameters in the MMSE algorithm.

The idea of the MMSE algorithm is shown in Figure 8.2. The prior parameters κ, b or κ and γ are chosen to minimise the mean squared-error, this is

$$\widehat{\boldsymbol{\theta}}_{\text{MMSE}} = \underset{\boldsymbol{\theta}}{\text{argmin}} \quad ||\mathbf{f} - \widehat{\mathbf{f}}_{\theta}||_2^2, \tag{8.28}$$

where $\widehat{\mathbf{f}}_{\theta}$ is the MAP estimate of $\mathbf{f}$ using parameters $\widehat{\boldsymbol{\theta}}_{\text{MMSE}}$ with $\theta = \kappa$, $\boldsymbol{\theta} = (\kappa, \gamma)$ or

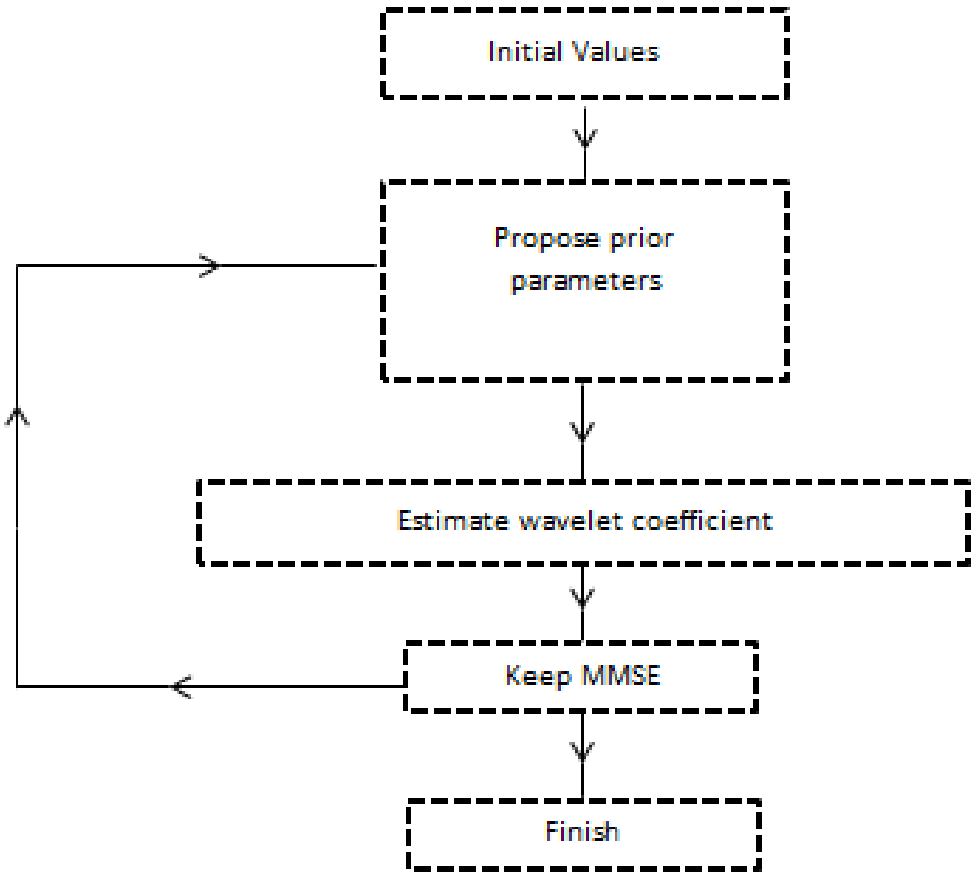


Figure 8.2: Diagram of the main idea for minimum MSE algorithm.

$\theta = b$. Also, for level-dependent prior $\boldsymbol{\theta} = \{\boldsymbol{\kappa}\}$, $\boldsymbol{\theta} = \{\boldsymbol{\kappa}, \boldsymbol{\gamma}\}$ or $\boldsymbol{\theta} = \{\mathbf{b}\}$. Hence, the true $\mathbf{f}$ is assumed to be known and at each run to estimate the function, which is computed using MAP estimate, the prior parameters are fixed, where the aim is to find the best prior parameters to estimate the function $\mathbf{f}$.

The MMSE algorithm is an alternative stochastic minimization method where, at each iteration, a random perturbation, θ', of the current best prior parameters θ^{k-1} is considered. The MAP estimate of the wavelet coefficient is found using θ' and the corresponding MSE is calculated. If the new MSE is less than that at the previous iteration then $\theta^k = \theta'$ otherwise $\theta^k = \theta^{k-1}$. Iteration continues until convergence.

Algorithm 3: MCMC sampling algorithm for level-dependent prior distributions

Result: Sample from posterior (MCMC)

Initialization $\mathbf{f} = \mathbf{0}$, $\boldsymbol{\theta}^0 = \{\theta_1, \theta_2, \ldots, \theta_{J-1}\}$, $\boldsymbol{\tau}_{11\times(J-1)} = \{0.01, 0.01 \ldots, 0.01\}$, $\boldsymbol{\tau}_{21\times(J-1)} = \{0.01, 0.01 \ldots, 0.01\}$ and $\mathbf{C} = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_{J-1}\}$, $\mathbf{c}_1 = \{a_1^0, b_1^0\}$ $\mathbf{c}_2 = \{a_2^0, b_2^0\} \ldots$.

Compute $\log \mathrm{p}(\mathbf{d_f}, \boldsymbol{\theta}|\mathbf{y})$, $\log \mathrm{p}(\theta_1|\mathbf{d}_{\mathbf{f}j}) = \log \mathrm{p}(\mathbf{d}_{\mathbf{f}j}|\theta_1) + \log \mathrm{p}(\theta_1|\mathbf{C}[1]) \ldots$

for k=1 to R **do**

  for j=0 to J-1 **do**

    $M = \mathrm{length}(\mathbf{d_f})$

    for i=1 to M **do**

      Compute $\log \mathrm{p}(\mathbf{d_f}, \theta[j]|\mathbf{y}) = \log \mathrm{p}(\mathbf{y}|\mathbf{d_f}) + \log \mathrm{p}(\mathbf{d}_{\mathbf{f}j}|\theta[j])$

      Generate ϵ_i from a Gaussian distribution $N(0, \boldsymbol{\tau}_1[r])$

      $\mathrm{d_f}^* = \mathrm{d_f}_i^k + \epsilon_i$

      Compute $\log \mathrm{p}(\mathbf{d_f}, \theta[j]|\mathbf{y})^* = \log \mathrm{p}(\mathbf{y}|\mathbf{d_f})^* + \log \mathrm{p}(\mathbf{d}_{\mathbf{f}j}|\theta[j])^*$

      Generate u from a uniform distribution, U(0,1)

      if $(\log \mathrm{p}(\mathbf{d_f}, \theta[j]|\mathbf{y})^* - \log \mathrm{p}(\mathbf{d_f}, \theta[j]|\mathbf{y})) > u$ **then**

        $\mathrm{d_f}_i^k = \mathrm{d_f}_i^*$

      else

        $\mathrm{d_f}_i^k = \mathrm{d_f}_i^{k-1}$

      end

    end

  end

  for j=0 to J-1 **do**

    Compute $\log \mathrm{p}(\boldsymbol{\theta}[j]|\mathbf{d}_{\mathbf{f}j})^k$

    Generate ϵ_j from a Gaussian distribution $N(0, \boldsymbol{\tau}_2[j])$

    $\boldsymbol{\theta}[j]^* = \boldsymbol{\theta}[j]^k + \epsilon_j^k$

    if$(\boldsymbol{\theta}[j]^* > \mathbf{0})\{$

    Compute $\log \mathrm{p}(\boldsymbol{\theta}[j]|\mathbf{d}_{\mathbf{f}j})^*$

    Generate u from a uniform distribution, U(0,1)

    if $(\log p(\boldsymbol{\theta}[j]|\mathbf{d}_{\mathbf{f}j})^* - \log p(\boldsymbol{\theta}[j]|\mathbf{d}_{\mathbf{f}j})^k) > u$ **then**

      $\boldsymbol{\theta}[j]^k = \boldsymbol{\theta}[j]^*$

    else

      $\boldsymbol{\theta}[j]^k = \boldsymbol{\theta}[j]^{k-1}$

    end

  end

  }

  Update $\boldsymbol{\tau}_1$ and $\boldsymbol{\tau}_2$

end

Algorithm 4: MAP algorithm for level-dependent prior distributions

Result: Sample from posterior (MCMC)

Initialization $\mathbf{f} = \mathbf{0}$, $\boldsymbol{\theta}^0 = \{\theta_1, \theta_2, \ldots, \theta_{J-1}\}$, $\boldsymbol{\tau}_{11\times(J-1)} = \{0.01, 0.01 \ldots, 0.01\}$, $\boldsymbol{\tau}_{21\times(J-1)} = \{0.01, 0.01 \ldots, 0.01\}$ and $\mathbf{C} = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_{J-1}\}$, $\mathbf{c}_1 = \{a_1^0, b_1^0\}$ $\mathbf{c}_2 = \{a_2^0, b_2^0\} \ldots$.

Compute $\log \mathrm{p}(\mathbf{d_f}, \boldsymbol{\theta}|\mathbf{y})$, $\log \mathrm{p}(\theta_1|\mathbf{d_f}_j) = \log \mathrm{p}(\mathbf{d_f}_j|\theta_1) + \log \mathrm{p}(\theta_1|\mathbf{C}[1]) \ldots$

for k=1 to R **do**

- $T_k = 2/\log(1+k)$
- **for** j=0 to J-1 **do**
  - $M = \mathrm{length}(\mathbf{d_f})$
  - **for** i=1 to M **do**
    - Compute $\log \mathrm{p}(\mathbf{d_f}, \theta[j]|\mathbf{y}) = \log \mathrm{p}(\mathbf{y}|\mathbf{d_f}) + \log \mathrm{p}(\mathbf{d_f}_j|\theta[j])$
    - Generate ϵ_i from a Gaussian distribution $N(0, \boldsymbol{\tau}_1[r])$
    - $\mathrm{d_f}^* = \mathrm{d_f}_i^k + \epsilon_i$
    - Compute $\log \mathrm{p}(\mathbf{d_f}, \theta[j]|\mathbf{y})^* = \log \mathrm{p}(\mathbf{y}|\mathbf{d_f})^* + \log \mathrm{p}(\mathbf{d_f}_j|\theta[j])^*$
    - Generate u from a uniform distribution, U(0,1)
    - **if** $(\log \mathrm{p}(\mathbf{d_f}, \theta[j]|\mathbf{y})^* - \log \mathrm{p}(\mathbf{d_f}, \theta[j]|\mathbf{y}))/T_k > u$ **then**
      - $\mathrm{d_f}_i^k = \mathrm{d_f}_i^*$
    - **else**
      - $\mathrm{d_f}_i^k = \mathrm{d_f}_i^{k-1}$
    - **end**
  - **end**
- **end**
- **for** j=0 to J-1 **do**
  - Compute $\log \mathrm{p}(\boldsymbol{\theta}[j]|\mathbf{d_f}_j)^k$
  - Generate ϵ_j from a Gaussian distribution $N(0, \boldsymbol{\tau}_2[j])$
  - $\boldsymbol{\theta}[j]^* = \boldsymbol{\theta}[j]^k + \epsilon_j^k$
  - if($\boldsymbol{\theta}[j]^* > \mathbf{0}$){
  - Compute $\log \mathrm{p}(\boldsymbol{\theta}[j]|\mathbf{d_f}_j)^*$
  - Generate u from a uniform distribution, U(0,1)
  - **if** $(\log p(\boldsymbol{\theta}[j]|\mathbf{d_f}_j)^* - \log p(\boldsymbol{\theta}[j]|\mathbf{d_f}_j)^k)/T_k > u$ **then**
    - $\boldsymbol{\theta}[j]^k = \boldsymbol{\theta}[j]^*$
  - **else**
    - $\boldsymbol{\theta}[j]^k = \boldsymbol{\theta}[j]^{k-1}$
  - **end**
- **end**
- }
- Update $\boldsymbol{\tau}_1$ and $\boldsymbol{\tau}_2$

end

Algorithm 5: MMSE algorithm for level-dependent prior distributions; step 7 to 27 represent MAP estimation of $\mathbf{d_f}$ for given $\boldsymbol{\theta}$.

Result: $\widehat{\mathbf{d}}_{\mathbf{f}}$, $\boldsymbol{\theta}_{\mathrm{MMSE}}$.

1 Initialization $\widehat{\mathbf{d}}_{\mathbf{f}}{}^{K}$, $\boldsymbol{\theta}^K = \{\theta_1^K, \theta_2^K, \ldots, \theta_{J-1}^K\}$ are the results of MAP, $\boldsymbol{\tau}_{1\,1\times(J-1)} = \{0.01, 0.01 \ldots, 0.01\}$, $\boldsymbol{\tau}_{2\,1\times(J-1)} = \{0.01, 0.01 \ldots, 0.01\}$ and Compute $\log \mathrm{p}(\mathbf{d_f}{}^K|\mathbf{y}, \boldsymbol{\theta}[j]^K)$, $\widehat{\mathbf{f}}^K = W^T\widehat{\mathbf{d}}_{\mathbf{f}}{}^K$, MSE=$\sum_i^m(\widehat{f}_i^K - f_i{}^{\mathrm{true}})^2/M$

2 **for** s=1 to N **do**

3 **for** j=0 to J-1 **do**

4 Generate ϵ_j from a Gaussian distribution $N(0, \boldsymbol{\tau}_1[j])$

5 $\boldsymbol{\theta}[j]^* = \boldsymbol{\theta}[j]^s + \epsilon_j^s$

6 **end**

7 if($\boldsymbol{\theta}^* > \mathbf{0}$){

8 **for** k=1 to R **do**

9 $T_k = 2/\log(1+k)$

10 **for** *j=0 to* $J-1$ **do**

11 $M = \mathrm{length}(\mathbf{d_f})$

12 **for** i=1 to M **do**

13 **for** i=1 to M **do**

14 $\mathrm{d_f}{}^* = \mathrm{d_f}{}_i^k + \epsilon$

15 Generate ϵ_j from a Gaussian distribution $N(0, \boldsymbol{\tau}_2[j])$

16 Compute $\log \mathrm{p}(\mathbf{d_f}|\mathbf{y}, \theta_1^*)^*$

17 Generate u from a uniform distribution, U(0,1)

18 **if** $(\log p(\mathbf{d_f}|\mathbf{y}, \boldsymbol{\theta}[j]^*)^* - \log p(\mathbf{d_f}|\mathbf{y}, \boldsymbol{\theta}[j]^s))/T_k > u$ **then**

19 $\mathrm{d_f}{}_i^{s(k)} = \mathrm{d_f}{}_i^*$

20 **else**

21 $\mathrm{d_f}{}_i^{s(k)} = \mathrm{d_f}{}_i^{s(k-1)}$

22 **end**

23 **end**

24 **end**

25 **end**

26 Update $\boldsymbol{\tau}_2$

27 **end**

28 Compute $\widehat{\mathbf{f}}^s = W^T\widehat{\mathbf{d}}_{\mathbf{f}}{}^s$, $\mathrm{MSE}^s = \sum_i^m(\widehat{f}_i^s - f_i{}^{\mathrm{true}})^2/M$

29 **if** ($\mathrm{MSE}^s < \mathrm{MSE}$) **then**

30 $\boldsymbol{\theta}^s = \boldsymbol{\theta}^*$, MSE= MSE^s

31 **else**

32 $\boldsymbol{\theta}^s = \boldsymbol{\theta}^{s-1}$

33 **end**

34 Update $\boldsymbol{\tau}_1$ }

35 **end**

8.5 Experiments

The purpose of this section is to evaluate and investigate whether the Gaussian, the Laplace, the elastic-net and the double Weibull priors are suitable for estimating the unknown $\mathbf{f}$. The proposed method is applied to the Blocks test function at $m = 128$ equally spaced points, and with the level of blur, which is given in (2.6), taken as $k = 0.005$, 0.01 and 0.07, and with level of noise taken as $\sigma = 0.5$. In addition, wavelet coefficients are computed using the (decimated) discrete wavelet basis with Haar wavelets. The benefit of using simulated data is that the quality of the solution can be investigated and any error attributed to the inversion procedure and not to mismodelling.

There are two methods applied to estimate the unknown $\mathbf{f}$; the first is to use the minimum mean squared-error; the second is to use a hierarchical Bayesian approach with prior distributions for σ^2, κ, γ and b, with inverse gamma, gamma, beta and inverse gamma distributions respectively. The hierarchical Bayesian approach is applied for the whole set of wavelet coefficients, or for level-dependent wavelet coefficients.

The total number of replications is $R = 10,000$ and the elastic-net prior is implemented to obtain the MAP estimates of the simulated data. The elastic-net with a single prior for whole data requires approximately 11.89 seconds, and 11.21 seconds for level-dependent priors, to process 1000 iterative updates of approximately 128 wavelet coefficients. In addition, the Gaussian, the Laplace and the double Weibull priors are implemented to obtain the MAP estimates of the simulated data. A single prior for whole data requires approximately 7.37 seconds, and 8.72 seconds for level-dependent priors, to process 1000 iterative updates of approximately 128 wavelet coefficients.

To investigate the proposed methods, three different techniques are used to obtain a reconstruction from observed data. The first is MMSE, which is described in Section 8.4, where the true function is assumed to be known. Each of the four priors, the Gaussian, the Laplace, the elastic-net and the double Weibull, are applied, for a range of parameter values where applicable. The parameters are altered heuristically by small increments

Prior	σ		MAP		
			k		
			0.005	0.010	0.070
Laplace	0.5	MMSE	0.0043	0.0205	5.2906
		$\widehat{\kappa}_{\text{MMSE}}$	0.0255	0.2588	3.0136
Elastic-net	0.5	MMSE	**0.004**	**0.0119**	**4.7830**
		$\widehat{\kappa}_{\text{MMSE}}$	1.0052	1.9014	3.0179
		$\widehat{\gamma}_{\text{MMSE}}$	0.1221	0.1630	0.3084
Gaussian	0.5	MMSE	0.0084	0.0371	5.1853
		$\widehat{\kappa}_{\text{MMSE}}$	0.0077	1.0863	2.0620
DW	0.5	MMSE	0.0048	0.0301	5.0823
		$\widehat{b}_{\text{MMSE}}$	0.8112	1.0540	3.9015

Table 8.1: Minimum MSE results to compare different priors for estimating the unknown vector $\mathbf{f}$. The Blocks test function, at $m = 128$ equally spaced points is used with different levels of blur, which is given in (2.6), k, and σ^2 =0.5

until the optimum, minimising the MSE, is located. The optimum parameter values for each of the priors and the resultant MSE are listed in Tables 8.1 and 8.2 with the minimum error highlighted in bold. Table 8.1 shows the results of MMSE by using MAP estimates where the prior parameters $\widehat{\kappa}$, $\widehat{\gamma}$ and $\widehat{b}$ are computed for whole data. Table 8.2 shows the results of MMSE where the prior parameters $\widehat{\kappa}_j$, $\widehat{\gamma}_j$ and $\widehat{b}_j$ are level-dependent. At each point in the trace the estimated value of the differential was compared with the true function $\mathbf{f}$, and the results were used to find the MSE over the whole signal, and then the set of prior parameters leading to the MMSE is collected. These results of estimating the prior parameters can be used to obtain a reconstruction using the corresponding Gaussian, Laplace, double Weibull or elastic-net priors.

Overall, for each of the blur levels tested, the elastic-net method outperformed all other methods and improves the estimation. The double Weibull prior performed well and this was the most computationally efficient of the single component priors models. Additionally, the double Weibull distribution provides better results than the Gaussian and Laplace, reducing the mean squared-error.

Prior	σ		k		
			0.005	0.010	0.070
Laplace	0.5	MMSE	0.0040	0.0136	3.2402
		$\widehat{\kappa}^j_{\text{MMSE}}$	$(0.0573)_6$ $(0.0301)_5$	$(0.0621)_6$ $(0.0496)_5$	$(0.1070)_6$ $(0.0823)_5$
			$(0.0154)_4$ $(0.0288)_3$	$(0.0314)_4$ $(0.0169)_3$	$(0.0536)_4$ $(0.0492)_3$
			$(0.0143)_2$ $(0.0101)_1$	$(0.0124)_2$ $(0.0087)_1$	$(0.0267)_2$ $(0.0189)_1$
			$(0.0071)_0$ $(0.0025)_{c_0}$	$(0.0062)_0$ $(0.0024)_{c_0}$	$(0.0133)_0$ $(0.0151)_{c_0}$
Elastic-net	0.5	MMSE	**0.0030**	**0.0105**	**2.9570**
		$\widehat{\kappa}^j_{\text{MMSE}}$	$(0.4133)_6$ $(0.2071)_5$	$(0.4785)_6$ $(0.3730)_5$	$(0.1549)_6$ $(0.0957)_5$
			$(0.0603)_4$ $(0.0239)_3$	$(0.0781)_4$ $(0.0318)_3$	$(0.0546)_4$ $(0.0177)_3$
			$(0.1033)_2$ $(0.0730)_1$	$(0.1196)_2$ $(0.0845)_1$	$(0.0387)_2$ $(0.0273)_1$
			$(0.0516)_0$ $(0.0181)_{c_0}$	$(0.0598)_0$ $(0.0265)_{c_0}$	$(0.0193)_0$ $(0.0125)_{c_0}$
		$\widehat{\gamma}^j_{\text{MMSE}}$	$(0.0717)_6$ $(0.0556)_5$	$(0.0345)_6$ $(0.0862)_5$	$(0.0505)_6$ $(0.0448)_5$
			$(0.4012)_4$ $(0.5238)_3$	$(0.6703)_4$ $(0.5751)_3$	$(0.7758)_4$ $(0.7241)_3$
			$(0.0179)_2$ $(0.0126)_1$	$(0.0086)_2$ $(0.0060)_1$	$(0.0126)_2$ $(0.0089)_1$
			$(0.0089)_0$ $(0.0189)_{c_0}$	$(0.0043)_0$ $(0.0076)_{c_0}$	$(0.0063)_0$ $(0.0072)_{c_0}$
Gaussian	0.5	MMSE	0.0032	0.0153	3.8037
		$\widehat{\kappa}^j_{\text{MMSE}}$	$(0.0316)_6$ $(0.0158)_5$	$(0.0417)_6$ $(0.0282)_5$	$(0.0561)_6$ $(0.0493)_5$
			$(0.0045)_4$ $(0.0032)_3$	$(0.0050)_4$ $(0.0016)_3$	$(0.0132)_4$ $(0.0085)_3$
			$(0.0079)_2$ $(0.0055)_1$	$(0.0104)_2$ $(0.0073)_1$	$(0.0140)_2$ $(0.0099)_1$
			$(0.0039)_0$ $(0.0061)_{c_0}$	$(0.0052)_0$ $(0.0071)_{c_0}$	$(0.0070)_0$ $(0.0017)_{c_0}$
DW	0.5	MMSE	0.0035	0.0131	3.8501
		$\widehat{b}^j_{\text{MMSE}}$	$(1.0940)_6$ $(1.7413)_5$	$(1.7801)_6$ $(1.4767)_5$	$(2.1910)_6$ $(2.9084)_5$
			$(1.8792)_4$ $(1.2885)_3$	$(1.6889)_4$ $(1.3267)_3$	$(2.2647)_4$ $(2.2492)_3$
			$(1.2735)_2$ $(1.1933)_1$	$(1.4450)_2$ $(1.3146)_1$	$(2.5477)_2$ $(2.3873)_1$
			$(1.1367)_0$ $(1.0393)_{c_0}$	$(1.2225)_0$ $(1.0680)_{c_0}$	$(2.2738)_0$ $(2.0831)_{c_0}$

Table 8.2: Minimum MSE result of simulation to compare the thresholding methods for estimating the unknown vector $\mathbf{f}$. Prior parameters are estimated at each level j. The Blocks test function, at $m = 128$ equally spaced points, is used with different levels of k and with σ^2 equal to 0.5.

The second technique, to obtain a reconstruction, is to use the MCMC algorithm 3, which is described in this chapter, to obtain the posterior mean (PM) estimator. In this procedure, prior parameters are assigned prior distributions. The procedure is described in Section 8.4. However, it is difficult to apply the prior simulation parameter estimation method without any real information about the true susceptibility. In general, the basic idea of MCMC approach is to apply a Markov Chain to generate pseudo-random samples such that its stationary distribution to follow a target probability distribution to provide the desired posterior distribution. So, MCMC result for $\mathbf{d_f}$ is a matrix, where the numbers of columns represents length of data and the number of rows represents the number of replications after equilibrium. Suppose the number of replications after equilibrium is equal to 1000. The MCMC results after equilibrium for prior parameters κ, γ and σ^2 are vectors of length 1000. Then the average of the samples can be calculated to compute $\widehat{\mathbf{d}}_\mathbf{f}$, $\widehat{\kappa}$, $\widehat{\gamma}$ and $\widehat{\sigma}^2$. Figure 8.3 shows the results of using the average of MCMC estimates. Also, quantiles are used to obtain the credible intervals from the sample of MCMC.

The third technique, to obtain a reconstruction, is to use MAP by annealing the MCMC algorithm, see Section 8.4. In this procedure, the prior parameters are described by the prior distributions. However, the reconstruction was made by taking the result of the final iteration. Figure 8.4 shows the reconstructions of MAP estimates and the credible intervals of these reconstructions were obtained from MCMC results.

Figure 8.4 and 8.3 show the reconstruction for the Blocks test function using MAP and PM, respectively. Overall, for each of the blur levels tested, the elastic-net and the double Weibull priors outperformed all other priors considered, including the Laplace and the Gaussian priors. Also, it can be seen that the reconstructions using a single prior for whole data fluctuate slightly. That is because there is one value of the threshold for the whole set of wavelet coefficients. The main interest is that MAP and PM using the elastic-net and the double Weibull priors give excellent reconstructions.

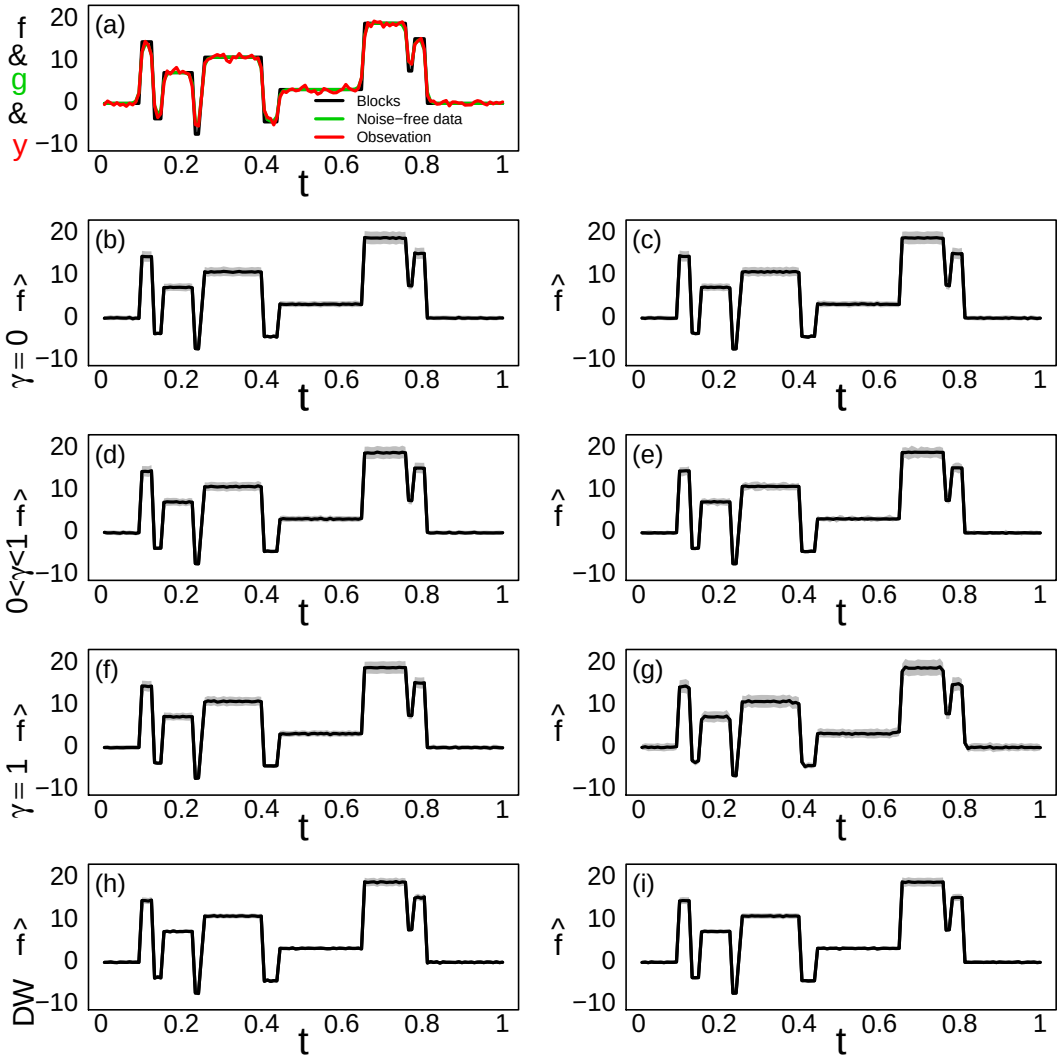


Figure 8.3: Plots of the reconstructions using the PM estimate to estimate the unknown vector $\mathbf{f}$. Blocks test function at $m = 128$ equally spaced points is used as the true function, plotted using a black line, green line is made by multiplying the blocks by the blur, which is given in (2.6), $k = 0.005$ and the red line shows the observed data with noise $\sigma^2 = 0.5$: (a) true Blocks test function, noise-free data and observations with large measurement error; (b) Single Laplace prior; (c) level-dependent Laplace priors; (d) Single elastic-net prior; (e) level-dependent elastic-net priors; (f) Single Gaussian prior; (g) level-dependent Gaussian priors; (h) Single double Weibull prior; and (i) level-dependent double Weibull priors.

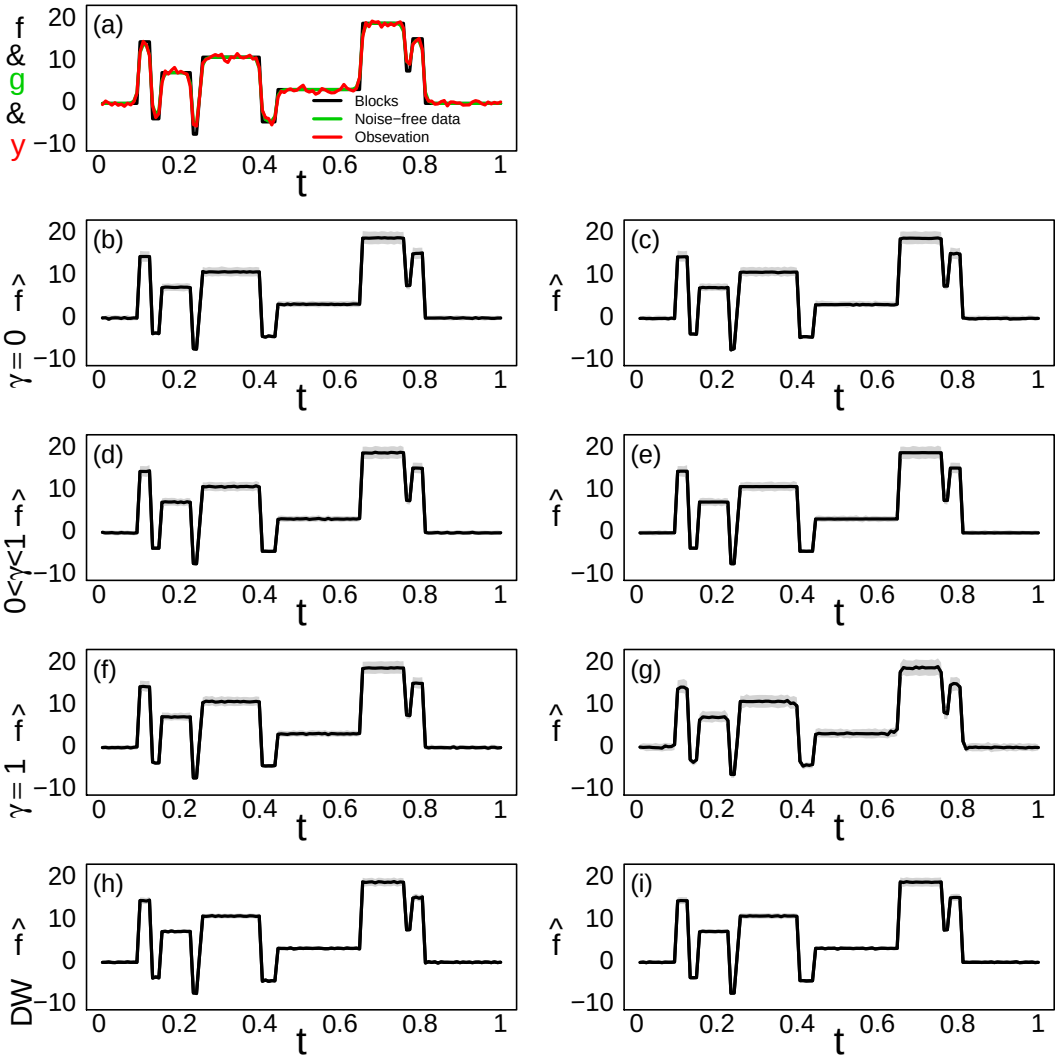


Figure 8.4: Plots of the reconstructions using the MAP estimate to estimate the unknown vector $\mathbf{f}$. Blocks test function at $m = 128$ equally spaced points is used as the true function, plotted as a black line, the green line is made by multiplying the blocks by the blur, which is given in (2.6), $k = 0.005$ and the red line shows the observed data with noise $\sigma^2 = 0.5$: (a) true Blocks test function, noise-free data and observations with large measurement error; (b) Single Laplace prior; (c) level-dependent Laplace priors; (d) Single elastic-net prior; (e) level-dependent elastic-net priors; (f) Single Gaussian prior; (g) level-dependent Gaussian priors; (h) Single double Weibull prior; and (i) level-dependent double Weibull priors.

8.6 Conclusions

Within this chapter the Gaussian, the Laplace, the elastic-net and the double Weibull distributions were introduced and used as prior models for wavelet coefficients. The approach has also been successfully applied to the Blocks test function, including parameter estimation using prior distributions for the whole wavelet coefficient vector and level-dependent priors. Two estimators were used, the PM estimate extracted from the MCMC sample; and the MAP estimate. In addition, the parameters σ, κ, γ and b are estimated by the minimum mean squared-error method.

Comparing the different priors, the elastic-net prior requires approximately 2 minutes to process 10,000 iterations of 128 wavelet coefficients. Also, the Gaussian, the Laplace and the double Weibull priors require approximately $1\frac{1}{3}$ minutes. The MAP estimates of the Blocks test function provides excellent reconstruction. The method is successfully applied to produce good reconstructions of the true feature profiles.

Overall, it was found that the level-dependent prior provides better reconstruction results than when using a single value of the parameters κ, γ and b for the whole wavelet coefficient vector. Also, the elastic-net prior distribution gives good reconstructions, which are close to the true feature profile, even when estimating the prior parameters.

A major limitation of performing a two-stage inversion and noise reduction, as described in Chapter 7, is that it is not possible to estimate standard errors, and hence it cannot be used to compute credible intervals, as well as being difficult to estimate the prior parameters. Furthermore, a single step reconstruction using the MCMC algorithm is more reliable than a two-stage reconstruction as it is possible to test many candidate priors in the estimation, and credible intervals can be computed.

In conclusion the best estimates are generated by a prior distribution incorporating the elastic-net prior. This is, not just merely in terms of MSE but because it gives reconstructions of the shape of the vector $\mathbf{f}$.

Chapter 9

Application to 1D archaeological stratigraphy

9.1 Overview

This chapter is organised as follows: Section 9.2 provides an introduction to estimation from simulated core data, whereas Section 9.3 describes the estimation of prior parameters, whilst Section 9.4 defines an application to real data, and then Section 9.5 gives the estimation of the susceptibility of real data. Finally, Section 9.6 provides conclusions.

9.2 Introduction to real and simulated core data

Real data

The cores were extracted from the 'Park', Guiting Power, which is a late iron-age farmstead. The experiment consisted of burning to the ground a wooden funeral pyre containing the corpse of a sheep followed by covering the burnt area with topsoil. Five cores were removed from the pyre region of the site, four from the main area of burning and one

from the periphery, and then taken to the laboratory for analysis, where each core was passed through the detector coil and the observed data recorded as described by Allum *et al.* (1999) and Aykroyd and Al-Gezeri (2014).

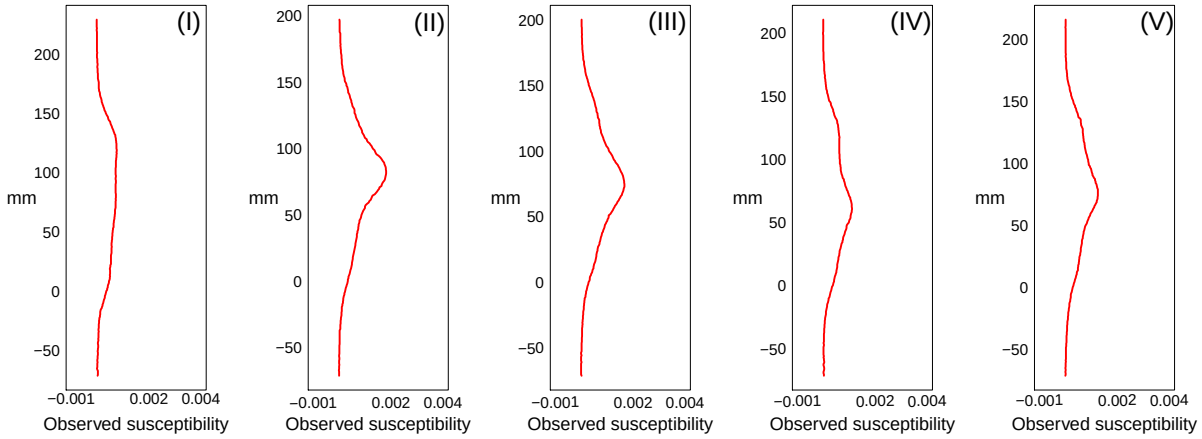


Figure 9.1: Plots of the real data from the 'Park', Guiting Power, where (I) shows data taken from the periphery around main area and (II)-(V) show data removed from the main area of burning (Allum *et al.*, 1999).

From Figure 9.1 it can be seen that the data from the periphery (I) has low magnetic susceptibility and there is no peak to be seen. It can also be seen that dataset (II) is similar to dataset (III) and dataset (IV) is similar to dataset (V). Additionally, it can be seen that the cores removed from the main area of burning (II)-(V) have higher magnetic susceptibility than the data taken from the periphery (I). Thus, the main area of burning has much higher magnetic susceptibility than the periphery.

Simulated data

Within this section, parameters are estimated from five simulated cores created by Aykroyd and Al-Gezeri (2014). In the real data, the features of the truth are usually not available. Instead it is reasonable to use simulated datasets, where the properties and features are known. Therefore five simulated cores will be studied and analysed and the simu-

lated datasets will be used to estimate the parameters Λ, κ, λ and γ using minimum mean squared-error (MMSE), which will be used in reconstruction from the real data, where MMSE is described in Sections 2.13 and 8.4. Since the true susceptibility profiles of simulated data are known, the accuracy of the statistical estimation process can be assessed.

Before starting the analysis, consider the following data collection method, as described in Section 2.4; the plastic cylinder, containing the core, is positioned a small distance from one end of the detector coil; then it is moved in small steps, pausing between movements for readings to be made. Therefore, it may be assumed that, initially and finally, the core produces no effect and hence the first few and last few readings are zero. However, as the core draws near the electrical coil, the magnetic readings increase.

Let the output readings be the data $\mathbf{y} = \{y_i, i = 1, 2, \ldots, n\}$. The data represents observations of the susceptibility over the length of the core. Let the magnetic susceptibilities be $\mathbf{f} = \{f_j, j = 1, 2, \ldots, m\}$. Also, we believe that the true values of $\mathbf{f}$ are larger than zero or equal to zero. The observed measurement y_i is then given by the convolution

$$y_i = \sum_{j=1}^{m} h_{ij} f_j + \epsilon_i, \quad \text{for} \quad i = 1, \ldots, n, \quad -\infty < y_i < \infty, \quad f_i \geq 0, \tag{9.1}$$

where $\epsilon_i \sim N(0, \sigma^2)$, with ϵ_i, ϵ_j $(i \neq j)$ independent and h_{ij} as defined in Equation (2.3).

Figure 9.2 shows a diagram of a simple core along with a corresponding susceptibility profile. Assume that the archaeology occupies a single layer, and that the site surface is at the top of the diagram. In practice, all the features of the core, such as depth, extent and susceptibility are unknown. The value d_1 represents the distance before the core enters the coil and the susceptibility over d_1 is assumed to be exactly zero. The value d_2 is assumed to be the first part of the core which enters the coil; it has susceptibility, x_B, which represents a background susceptibility. As the core passes through, the second part of the core, of length d_3, represents an archaeological feature with susceptibility x_F. There is a second background part, which is of length d_4 and has susceptibility x_B. Finally, d_5 represents the last distance after the core has emerged, and has zero susceptibility before the data recording stops; for more detail see Aykroyd and Al-Gezeri (2014).

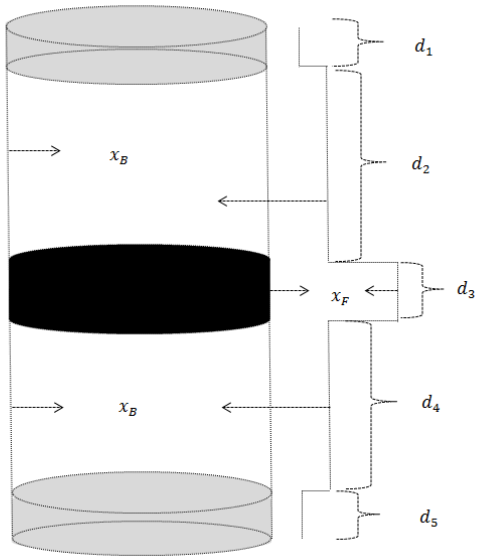


Figure 9.2: Diagram of the extracted core and corresponding susceptibility profile, where x_B represents a background susceptibility, x_F represents archaeological feature susceptibility, d_1 and d_5 represent the distance before and after the core enters and emerges respectively, d_2 and d_4 represent the distances when the magnetic susceptibility of background is recorded and d_3 represents the distance when magnetic susceptibility of archaeological feature is recorded (Aykroyd and Al-Gezeri, 2014).

Core	Distances					Susceptibility	
						Feature	Background
	d_1 (cm)	d_2 (cm)	d_3 (cm)	d_4 (cm)	d_5 (cm)	x_F SI x10^{-3}	x_B SI x10^{-3}
1	80	30	70	140	80	1.1	0.2
2	80	30	70	140	80	1.5	0.2
3	80	30	80	130	80	1.8	0.2
4	80	30	160	50	80	1.1	0.2
5	80	30	150	60	80	1.0	0.2

Table 9.1: True values of length of core and feature susceptibility for five simulated cores (Aykroyd and Al-Gezeri, 2014).

Table 9.1 shows the features of the five simulated cores generated by Aykroyd and Al-Gezeri (2014). They assume that the overall length of the data is 4.0 m, with a soil core length of 2.4 m, with the background susceptibility fixed at 0.2 (SI x10^{-3}), and the depth fixed at 30 cm, corresponding to the modern topsoil deposited after the archaeology was leveled to the prevailing ground level.

It can be seen, from Table 9.1, that the first three cores have a small extent, whilst the fourth and fifth cores have a large extent. The parameters d_1 and d_5 correspond to the readings when no core is in the detector, meaning before the core enters and after it emerges. All cores (1)-(5) have a similar distance, d_1 and d_5, of zero susceptibility, equal to 80 cm. The value d_2 represents the depth of the feature layer, whereas d_3 represents the extent of the feature, and d_4 represents the depth of background below the feature layer. The feature and background susceptibility are the parameters x_F and x_B respectively. The key parameters are d_2, d_3 and x_F, whereas d_1 and d_2 can be measured with high reliability, the background susceptibility, x_B, can be measured by taking a separate sample, and d_4 can be calculated by subtraction, given d_2, d_3 and the overall length of the core (Aykroyd and Al-Gezeri, 2014).

Figure 9.3 shows the plots of the five simulated cores represented in Table 9.1. It also shows the observations; the black lines show the true values and the red lines show the observations. These observations represent the true data corrupted by noise, and there is also blurring. In addition, it can be seen that the observations follow smooth curves and the features cannot be detected well from these datasets. This means that it is difficult to accurately compute the features, background susceptibility, and the depth and extent.

It is believed that the real susceptibility profile is a step function. Hence the simulated cores reflect this belief, and therefore can be used to estimate the prior parameters Λ, κ, λ, γ.

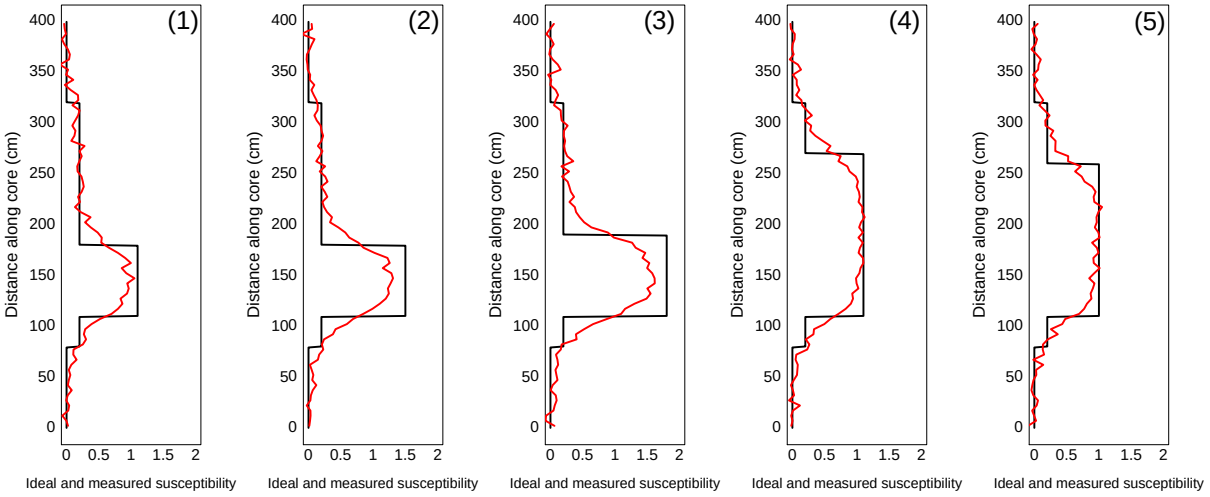


Figure 9.3: Plots of the five simulated cores: the black lines show the true susceptibility; and the red lines show the observations (Aykroyd and Al-Gezeri, 2014).

9.3 Minimum mean squared-error prior parameter estimation

General

In this section, two methods, one-stage and two-stage, will be used to estimate θ or $\boldsymbol{\theta}$, using MMSE. More precisely, different approaches will be applied for estimating parameters such as $\theta = \{\Lambda\}$, $\theta = \{\kappa\}$, $\boldsymbol{\theta} = \{\Lambda, \lambda\}$, $\boldsymbol{\theta} = \{\kappa, \gamma\}$, or $\boldsymbol{\theta} = \{\Lambda, \kappa\}$, and for level-dependent priors $\boldsymbol{\theta} = \{\boldsymbol{\kappa}\}$, $\boldsymbol{\theta} = \{\Lambda, \boldsymbol{\lambda}\}$, $\boldsymbol{\theta} = \{\Lambda, \boldsymbol{\kappa}\}$ or $\boldsymbol{\theta} = \{\boldsymbol{\kappa}, \boldsymbol{\gamma}\}$. The first is to use the two-stage method to estimate $\boldsymbol{\theta}$, where the parameter λ denotes the value of the threshold and the parameter Λ represents the value of the smoothing parameter in the inversion method. The second process involves using the one-stage method to estimate $\boldsymbol{\theta}$, the parameters κ and γ represent prior knowledge. All explanations related to the main idea of two-stage estimation were provided in Section 7.5 and Algorithm 2, in Section 2.13, was used to build a reconstruction using the two-stage process. Algorithm 5 in Section 8.4 explains the one-stage process.

Two-stage estimation

The purpose of this subsection is to estimate parameters from simulated cores using the two-stage methods, such as IT-TO and WVD-TO with classical thresholding rules, DWWS-LPM, DUHT, BlockSure and EB (posterior median).

First-order method						
Rule	Min	Core				
		1	2	3	4	5
BlockSure (IT-TO)	MMSE	0.016	0.036	0.040	**0.017**	0.013
	$\widehat{\Lambda}_{\text{MMSE}}$	0.160	0.074	0.076	0.377	0.127
EB (IT-TO)	MMSE	**0.015**	**0.035**	**0.039**	**0.017**	**0.012**
	$\widehat{\Lambda}_{\text{MMSE}}$	0.145	0.062	0.064	0.343	0.120

Table 9.2: The IT-TO results for minimum MSE, described in Section 2.13, using BlockSure and EB methods to estimate parameter Λ for different core samples. The bold font represents the smallest MSE.

First-order method						
Rule	Min	Core				
		1	2	3	4	5
Hard (IT-TO)	MMSE	**0.012**	**0.026**	0.031	0.016	0.011
	$\widehat{\Lambda}_{\text{MMSE}}$	0.066	0.017	0.018	0.313	0.074
	$\widehat{\boldsymbol{\lambda}}_{\text{MMSE}}$	$(0.0828)_7\ (3.1084)_6\ (0.1564)_5$ $(0.3649)_4\ (1.1324)_3$	$(0.0853)_7\ (0.6735)_6\ (0.2377)_5$ $(0.6943)_4\ (1.2433)_3$	$(0.7956)_7\ (0.1940)_6\ (0.4896)_5$ $(0.9767)_4\ (0.7154)_3$	$(0.3377)_7\ (0.0652)_6\ (0.1556)_5$ $(0.2475)_4\ (0.2679)_3$	$(0.1339)_7\ (3.1359)_6\ (0.2207)_5$ $(0.3073)_4\ (1.1083)_3$
Soft (IT-TO)	MMSE	0.0150	0.033	0.037	**0.015**	0.011
	$\widehat{\Lambda}_{\text{MMSE}}$	0.0816	0.012	0.052	0.201	0.083
	$\widehat{\boldsymbol{\lambda}}_{\text{MMSE}}$	$(0.1910)_7\ (0.7457)_6\ (0.0367)_5$ $(0.2776)_4\ (0.0937)_3$	$(0.0826)_7\ (0.0483)_6\ (0.2377)_5$ $(0.2270)_4\ (0.5618)_3$	$(0.5077)_7\ (0.0023)_6\ (0.0977)_5$ $(0.0553)_4\ (0.3712)_3$	$(0.1164)_7\ (0.0002)_6\ (0.0547)_5$ $(0.0021)_4\ (0.3106)_3$	$(0.1325)_7\ (0.1278)_6\ (0.0352)_5$ $(0.0709)_4\ (0.3712)_3$
G (IT-TO)	MMSE	0.013	0.031	0.033	**0.015**	0.011
	$\widehat{\Lambda}_{\text{MMSE}}$	0.073	0.014	0.0440	0.179	0.097
	$\widehat{\boldsymbol{\lambda}}_{\text{MMSE}}$	$(0.2150)_7\ (0.0405)_6\ (0.0950)_5$ $(0.7285)_4\ (0.2311)_3$	$(1.2662)_7\ (0.3365)_6\ (0.1550)_5$ $(0.4034)_4\ (0.7708)_3$	$(0.2930)_7\ (0.0446)_6\ (0.1768)_5$ $(0.3204)_4\ (0.8998)_3$	$(0.1177)_7\ (0.0180)_6\ (0.0973)_5$ $(0.1432)_4\ (0.6818)_3$	$(0.1987)_7\ (0.0323)_6\ (0.0888)_5$ $(0.1563)_4\ (0.6380)_3$
SCAD (IT-TO)	MMSE	0.014	0.030	**0.030**	0.016	**0.010**
	$\widehat{\Lambda}_{\text{MMSE}}$	0.092	0.021	0.031	0.188	0.086
	$\widehat{\boldsymbol{\lambda}}_{\text{MMSE}}$	$(0.1980)_7\ (0.0208)_6\ (0.0780)_5$ $(0.1674)_4\ (0.1610)_3$	$(0.2031)_7\ (0.5807)_6\ (0.1187)_5$ $(0.2269)_4\ (0.5364)_3$	$(1.2376)_7\ (0.0486)_6\ (0.1515)_5$ $(0.3656)_4\ (0.8242)_3$	$(0.1938)_7\ (0.0186)_6\ (0.0653)_5$ $(0.1506)_4\ (0.1559)_3$	$(0.5024)_7\ (0.5547)_6\ (0.0934)_5$ $(0.1455)_4\ (0.5222)_3$
DWWS-LPM (IT-TO)	MMSE	0.016	0.036	0.040	0.017	0.013
	$\widehat{\Lambda}_{\text{MMSE}}$	0.153	0.073	0.068	0.378	0.127
	$\widehat{\boldsymbol{\kappa}}_{\text{MMSE}}$	$(0.0398)_7\ (3.5852)_6\ (3.0636)_5$ $(2.0229)_4\ (3.4294)_3$	$(0.0191)_7\ (0.0165)_6\ (2.1915)_5$ $(2.3145)_4\ (1.9441)_3$	$(0.9249)_7\ (1.0565)_6\ (0.0021)_5$ $(1.3797)_4\ (1.6046)_3$	$(0.0299)_7\ (1.1329)_6\ (2.6393)_5$ $(1.7428)_4\ (1.6564)_3$	$(0.0217)_7\ (0.0100)_6\ (2.1673)_5$ $(1.9383)_4\ (1.3538)_3$

Table 9.3: The IT-TO results for minimum MSE, described in Section 2.13, of the different rules to estimate parameters Λ, $\boldsymbol{\kappa}$ and $\boldsymbol{\lambda}$ for different core samples. The bold font represents the smallest MSE.

For each core, the parameters θ or $\boldsymbol{\theta}$ are estimated using the MMSE method, explained in

First-order method-DUHT													
Core	MMSE	$\widehat{\Lambda}_{\text{MMSE}}$	$\widehat{\boldsymbol{\lambda}}_{\text{MMSE}}$										
1	0.016	0.179	$(0.0117)_1$	$(0.0005)_2$	$(0.0403)_3$	$(0.0563)_4$	$(0.0083)_5$	$(0.0176)_6$	$(0.0542)_7$	$(0.0791)_8$	$(0.0109)_9$	$(0.0237)_{10}$	$(0.0044)_{11}$
			$(0.0377)_{12}$	$(0.0141)_{13}$	$(0.0058)_{14}$	$(0.0065)_{15}$	$(0.0121)_{16}$	$(0.0097)_{17}$	$(0.0004)_{18}$	$(0.0100)_{19}$	$(0.0028)_{20}$	$(0.0039)_{21}$	-
2	0.036	0.074	$(0.0004)_1$	$(0.0090)_2$	$(0.0179)_3$	$(0.0356)_4$	$(0.0116)_5$	$(0.0298)_6$	$(0.0435)_7$	$(0.0009)_8$	$(0.0344)_9$	$(0.0003)_{10}$	$(0.0180)_{11}$
			$(0.0245)_{12}$	$(0.0045)_{13}$	$(0.0257)_{14}$	$(0.0298)_{15}$	$(0.0415)_{16}$	$(0.0217)_{17}$	$(0.0254)_{18}$	$(0.0037)_{19}$	$(0.0484)_{20}$	$(0.0034)_{21}$	-
3	0.040	0.077	$(0.0157)_1$	$(0.0043)_2$	$(0.0291)_3$	$(0.0150)_4$	$(0.0055)_5$	$(0.0273)_6$	$(0.0106)_7$	$(0.0143)_8$	$(0.0230)_9$	$(0.0375)_{10}$	$(0.0053)_{11}$
			$(0.0074)_{12}$	$(0.0257)_{13}$	$(0.0062)_{14}$	$(0.0144)_{15}$	$(0.0222)_{16}$	$(0.0051)_{17}$	$(0.0019)_{18}$	$(0.0101)_{19}$	$(0.0079)_{20}$	$(0.0117)_{21}$	-
4	0.017	0.399	$(0.0275)_1$	$(0.0164)_2$	$(0.0187)_3$	$(0.0142)_4$	$(0.0371)_5$	$(0.0161)_6$	$(0.0359)_7$	$(0.0256)_8$	$(0.0104)_9$	$(0.0110)_{10}$	$(0.0108)_{11}$
			$(0.0454)_{12}$	$(0.0511)_{13}$	$(0.0696)_{14}$	$(0.0103)_{15}$	$(0.0203)_{16}$	$(0.0773)_{17}$	$(0.0382)_{18}$	$(0.0127)_{19}$	$(0.0533)_{20}$	$(0.0239)_{21}$	-
5	0.013	0.123	$(0.0418)_1$	$(0.0119)_2$	$(0.0416)_3$	$(0.0082)_4$	$(0.0310)_5$	$(0.0177)_6$	$(0.0218)_7$	$(0.0124)_8$	$(0.0508)_9$	$(0.0328)_{10}$	$(0.0185)_{11}$
			$(0.0358)_{12}$	$(0.0476)_{13}$	$(0.0799)_{14}$	$(0.0045)_{15}$	$(0.0431)_{16}$	$(0.0891)_{17}$	$(0.0318)_{18}$	$(0.0165)_{19}$	$(0.0454)_{20}$	$(0.0049)_{21}$	-

Table 9.4: The IT-TO results for minimum MSE, described in Section 2.13, using DUHT method to estimate parameters Λ and $\boldsymbol{\lambda}$ for different core samples.

First-order method						
Rule	Min	Core				
		1	2	3	4	5
BlockSure (WVD-TO)	MMSE	**0.016**	**0.036**	**0.040**	**0.017**	**0.013**
	$\widehat{\Lambda}_{\text{MMSE}}$	0.164	0.074	0.078	0.331	0.130
EB (WVD-TO)	MMSE	**0.016**	**0.036**	**0.040**	**0.017**	**0.013**
	$\widehat{\Lambda}_{\text{MMSE}}$	0.159	0.073	0.075	0.375	0.123

Table 9.5: The WVD-TO results for minimum MSE, described in Section 2.13, using the BlockSure and the EB (posterior median) methods to estimate parameter Λ for different core samples. The bold font represents the smallest MSE.

Algorithm 2 and described in Section 2.13. However, in some methods only Λ is estimated, such as the EB (posterior median) and the BlockSure methods, which are applied as plug-in methods. Moreover, the first-order method, that is defined in Section 2.6, is involved to estimate unknown vector $\mathbf{f}$. The procedure is shown in Algorithm 2, which described in Section 2.13. The total number of iterations equals 6000, where for each iteration, θ or $\boldsymbol{\theta}$ are proposed.

The results of the IT-TO method are summarised in Tables 9.3, 9.2 and 9.4, where bold numbers indicate the smallest MSE result for each core. It can be concluded that classical, EB (posterior median), DWWS-MAP and DUHT methods improve MMSE.

Tables 9.6, 9.5 and 9.7 shows the WVD-TO results for minimum MSE. In practice, the

First-order method							
Rule	Min	Core					
		1	2	3	4	5	
Hard (WVD-TO)	MMSE	0.015	**0.034**	0.041	0.017	0.014	
	$\widehat{\Lambda}_{\text{MMSE}}$	0.136	0.015	0.020	0.254	0.115	
	$\widehat{\boldsymbol{\lambda}}_{\text{MMSE}}$	$(0.0010)_7$ $(0.0002)_6$ $(0.0001)_5$ $(0.0005)_4$ $(0.0028)_3$	$(0.0034)_7$ $(0.0265)_6$ $(0.0091)_5$ $(0.0006)_4$ $(0.0084)_3$	$(0.4676)_7$ $(0.0923)_6$ $(0.1449)_5$ $(0.0045)_4$ $(0.0809)_3$	$(0.0027)_7$ $(0.0009)_6$ $(0.0001)_5$ $(0.0029)_4$ $(0.0042)_3$	$(0.0185)_7$ $(0.0459)_6$ $(0.0184)_5$ $(0.0013)_4$ $(0.0109)_3$	
Soft (WVD-TO)	MMSE	0.017	0.039	0.044	0.017	0.012	
	$\widehat{\Lambda}_{\text{MMSE}}$	0.080	0.031	0.053	0.190	0.092	
	$\widehat{\boldsymbol{\lambda}}_{\text{MMSE}}$	$(0.0027)_7$ $(0.0109)_6$ $(0.0311)_5$ $(0.0004)_4$ $(0.0096)_3$	$(0.0018)_7$ $(0.0037)_6$ $(0.0027)_5$ $(0.0001)_4$ $(0.0023)_3$	$(0.0065)_7$ $(0.0100)_6$ $(0.0681)_5$ $(0.0003)_4$ $(0.0154)_3$	$(0.0108)_7$ $(0.0353)_6$ $(0.0676)_5$ $(0.0001)_4$ $(0.0138)_3$	$(0.0102)_7$ $(0.0067)_6$ $(0.0208)_5$ $(0.0003)_4$ $(0.0159)_3$	
G (WVD-TO)	MMSE	0.015	0.037	0.042	**0.016**	**0.011**	
	$\widehat{\Lambda}_{\text{MMSE}}$	0.069	0.023	0.048	0.142	0.093	
	$\widehat{\boldsymbol{\lambda}}_{\text{MMSE}}$	$(0.0443)_7$ $(0.0676)_6$ $(0.4312)_5$ $(0.0012)_4$ $(0.0243)_3$	$(0.0214)_7$ $(0.0082)_6$ $(0.0099)_5$ $(0.0004)_4$ $(0.0042)_3$	$(0.0545)_7$ $(0.1302)_6$ $(0.4709)_5$ $(0.0014)_4$ $(0.0258)_3$	$(0.0383)_7$ $(0.0440)_6$ $(0.4587)_5$ $0.0007)_4$ $(0.0237)_3$	$(0.0432)_7$ $(0.0797)_6$ $(0.4042)_5$ $(0.0008)_4$ $(0.0225)_3$	
SCAD (WVD-TO)	MMSE	**0.011**	0.036	**0.040**	**0.016**	**0.011**	
	$\widehat{\Lambda}_{\text{MMSE}}$	0.084	0.032	0.029	0.1316	0.084	
	$\widehat{\boldsymbol{\lambda}}_{\text{MMSE}}$	$(0.0105)_7$ $(0.0036)_6$ $(0.0130)_5$ $(0.0007)_4$ $(0.0180)_3$	$(0.1233)_7$ $(0.0233)_6$ $(0.1750)_5$ $(0.0002)_4$ $(0.0040)_3$	$(0.0450)_7$ $(0.0562)_6$ $(0.0390)_5$ $(0.0019)_4$ $(0.0277)_3$	$(0.0163)_7$ $(0.0098)_6$ $(0.0281)_5$ $(0.0009)_4$ $(0.0172)_3$	$(0.0105)_7$ $(0.0036)_6$ $(0.0130)_5$ $(0.0007)_4$ $(0.0180)_3$	
DWWS-LPM (WVD-TO)	MMSE	0.023	0.052	0.077	0.026	0.019	
	$\widehat{\Lambda}_{\text{MMSE}}$	0.014	0.148	0.015	0.978	0.858	
	$\widehat{\kappa}_{\text{MMSE}}$	$(0.1739)_7$ $(0.8747)_6$ $(0.7507)_5$ $(1.3664)_4$ $(0.7532)_3$	$(0.0987)_7$ $(0.0936)_6$ $(0.0753)_5$ $(0.0768)_4$ $(0.0504)_3$	$(0.4269)_7$ $(3.0793)_6$ $(0.8431)_5$ $(1.0637)_4$ $(0.0985)_3$	$(1.3801)_7$ $(2.6195)_6$ $(1.0560)_5$ $(0.9831)_4$ $(0.0148)_3$	$(1.1256)_7$ $(2.1688)_6$ $(0.7714)_5$ $(0.3932_4$ $(0.0086)_3$	

Table 9.6: The WVD-TO results for minimum MSE, described in Section 2.13, of the different rules to estimate parameters Λ, $\boldsymbol{\kappa}$ and $\boldsymbol{\lambda}$ for different core samples. The bold font represents the smallest MSE.

First-order method-DUHT												
Core	MMSE	$\widehat{\Lambda}_{\text{MMSE}}$	$\widehat{\boldsymbol{\lambda}}_{\text{MMSE}}$									
1	0.013	0.047	$(0.0005)_1$	$(0.0001)_2$	$(0.0008)_3$	$(0.0010)_4$	$(0.0004)_5$	$(0.0015)_6$	$(0.0013)_7$	$(0.0007)_8$	$(0.0007)_9$	$(0.0014)_{10}$	$(0.0607)_{11}$
			$(0.0439)_{12}$	$(0.0072)_{13}$	$(0.0070)_{14}$	$(0.0009)_{15}$	$(0.0566)_{16}$	$(0.0209)_{17}$	$(0.0614)_{18}$	$(0.0079)_{19}$	$(0.0146)_{20}$	$(0.0504)_{21}$	-
2	0.021	0.0072	$(0.0016)_1$	$(0.0005)_2$	$(0.0001)_3$	$(0.0008)_4$	$(0.0010)_5$	$(0.0004)_6$	$(0.0015)_7$	$(0.0013)_8$	$(0.0007)_9$	$(0.0007)_{10}$	$(0.0745)_{11}$
			$(0.1548)_{12}$	$(0.1021)_{13}$	$(0.0224)_{14}$	$(0.0027)_{15}$	$(0.0003)_{16}$	$(0.1654)_{17}$	$(0.0394)_{18}$	$(0.1071)_{19}$	$(0.0358)_{20}$	$(0.0313)_{21}$	-
3	0.037	0.042	$(6\times10^{-5})_1$	$(2\times10^{-5})_2$	$(2\times10^{-5})_3$	$(1\times10^{-6})_4$	$(1\times10^{-5})_5$	$(4\times10^{-6})_6$	$(1\times10^{-5})_7$	$(4\times10^{-5})_8$	$(9\times10^{-6})_9$	$(1\times10^{-5})_{10}$	$(7\times10^{-6})_{11}$
			$(8\times10^{-6})_{12}$	$(4\times10^{-6})_{13}$	$(2\times10^{-5})_{14}$	$(3\times10^{-6})_{15}$	$(1\times10^{-5})_{16}$	$(1\times10^{-6})_{17}$	$(1\times10^{-5})_{18}$	$(1\times10^{-5})_{19}$	$(7\times10^{-6})_{20}$	$(5\times10^{-6})_{21}$	-
4	0.002	0.063	$(0.0002)_1$	$(0.0006)_2$	$(0.0012)_3$	$(0.0011)_4$	$(0.0007)_5$	$(0.0001)_6$	$(0.0011)_7$	$(1\times10^{-5})_8$	$(0.0003)_9$	$(0.0008)_{10}$	$(0.0030)_{11}$
			$(0.0075)_{12}$	$(0.0099)_{13}$	$(0.0024)_{14}$	$(0.0035)_{15}$	$(0.0158)_{16}$	$(0.0022)_{17}$	$(0.0069)_{18}$	$(0.0110)_{19}$	$(0.0003)_{20}$	$(0.0069)_{21}$	-
5	0.005	0.727	$(4\times10^{-5})_1$	$(2\times10^{-5})_2$	$(1\times10^{-5})_3$	$(6\times10^{-6})_4$	$(2\times10^{-5})_5$	$(1\times10^{-5})_6$	$(3\times10^{-5})_7$	$(2\times10^{-6})_8$	$(1\times10^{-5})_9$	$(8\times10^{-6})_{10}$	$(1\times10^{-5})_{11}$
			$(3\times10^{-6})_{12}$	$(4\times10^{-7})_{13}$	$(1\times10^{-5})_{14}$	$(1\times10^{-5})_{15}$	$(2\times10^{-5})_{16}$	$(1\times10^{-5})_{17}$	$(1\times10^{-5})_{18}$	$(4\times10^{-6})_{19}$	$(4\times10^{-6})_{20}$	$(3\times10^{-6})_{21}$	-

Table 9.7: The WVD-TO results for minimum MSE, described in Section 2.13, using the DUHT method to estimate parameters Λ and $\boldsymbol{\lambda}$ for different core samples.

WVD-TO as the DUHT method provides a smaller MSE than the other thresholding rules, which are shown in Tables (9.6) and (9.5).

One-stage approach

The purpose of this subsection is to apply the method of MMSE that is described in Section 8.4 and estimate the parameters, θ or $\boldsymbol{\theta}$, using five datasets obtained from five simulated cores. The minimum mean squared-error approach, based on the MAP estimate, is applied to produce estimates of the prior parameters, θ or $\boldsymbol{\theta}$. Algorithm in 5, described in Chapter 8, explains the one-stage procedure used to obtain the prior parameters from simulated data. The total number of runs equals $256 \times 6000 \times 500 = 7.68 \times 10^8$, where the parameters θ or $\boldsymbol{\theta}$ are proposed 500 times and the estimated susceptibilities were compared with the true susceptibilities of the five cores. The number of iterations and replications equal 256 and 6000, respectively. The results are summarised in Table 9.8, where bold numbers indicate the smallest MSE, and θ and $\boldsymbol{\theta}$, are estimated using a single prior for different core samples.

Prior		Core				
		1	2	3	4	5
Laplace	MMSE	0.0119	0.0189	0.0227	0.0129	0.0095
	$\widehat{\kappa}_{\text{MMSE}}$	17.1029	14.1360	12.8202	13.0016	15.2445
Elastic-net	MMSE	**0.0033**	**0.0069**	**0.0114**	**0.0032**	**0.0027**
	$\widehat{\kappa}_{\text{MMSE}}$	14.1230	15.6785	14.0001	14.9020	17.3445
	$\widehat{\gamma}_{\text{MMSE}}$	0.9295	0.9238	0.9189	0.9091	0.9057
Gaussian	MMSE	0.0036	0.0075	0.0118	0.0035	0.0028
	$\widehat{\kappa}_{\text{MMSE}}$	12.0093	11.8091	13.1721	15.1775	14.1102

Table 9.8: The results for minimum MSE of the different priors used to estimate parameters κ and γ for different core samples, where MMSE is described in Section 8.4. The bold font represents the smallest MSE.

Prior		Core				
		1	2	3	4	5
Laplace	MMSE	0.0057	0.0121	0.0157	**0.0051**	0.0050
	$\widetilde{\kappa}^j_{\text{MMSE}}$	$(14.3925)_7\ (10.1719)_6$	$(14.3285)_7\ (10.1864)_6$	$(14.3830)_7\ (10.1049)_6$	$(14.3157)_7\ (10.1060)_6$	$(14.3739)_7\ (10.1281)_6$
		$(7.0751)_5\ (5.0374)_4$	$(7.07533)_5\ (5.0285)_4$	$(7.0740)_5\ (5.0121)_4$	$(7.0755)_5\ (5.0361)_4$	$(7.0735)_5\ (5.0116)_4$
		$(3.5290)_3\ (2.4741)_2$	$(3.5280)_3\ (2.4461)_2$	$(3.5530)_3\ (2.4514)_2$	$(3.5417)_3\ (2.4812)_2$	$(3.5458)_3\ (2.4974)_2$
		$(1.7704)_1\ (0.8995)_0$	$(1.7221)_1\ (0.8955)_0$	$(1.7160)_1\ (0.8989)_0$	$(1.7143)_1\ (0.8947)_0$	$(1.7601)_1\ (0.8983)_0$
		$(0.8313)_{c_0}$	$(0.8313)_{c_0}$	$(0.8375)_{c_0}$	$(0.8312)_{c_0}$	$(0.8688)_{c_0}$
Elastic-net	MMSE	**0.0042**	**0.0116**	**0.0124**	0.0058	**0.0043**
	$\widetilde{\kappa}^j_{\text{MMSE}}$	$(14.1014)_7\ (9.9034)_6$	$(14.3630)_7\ (10.1241)_6$	$(11.5352)_7\ (8.1523)_6$	$(21.3526)_7\ (15.2704)_6$	$(14.3477)_7\ (10.1168)_6$
		$(7.0549)_5\ (4.9340)_4$	$(7.1497)_5\ (5.0158)_4$	$(5.7756)_5\ (3.9917)_4$	$(10.6840)_5\ (7.4735)_4$	$(7.1175)_5\ (5.0162)_4$
		$(3.4456)_3\ (2.3920)_2$	$(3.4892)_3\ (2.2505)_2$	$(2.7549)_3\ (1.8531)_2$	$(5.4285)_3\ (3.6669)_2$	$(3.5476)_3\ (2.4377)_2$
		$(1.7235)_1\ (0.8813)_0$	$(1.6943)_1\ (0.8989)_0$	$(1.3051)_1\ (0.7209)_0$	$(2.4289)_1\ (0.3345)_0$	$(1.6494)_1\ (0.8961)_0$
		$(0.8375)_{c_0}$	$(0.8375)_{c_0}$	$(0.7875)_{c_0}$	$(0.3537)_{c_0}$	$(0.8062)_{c_0}$
	$\widetilde{\gamma}^j_{\text{MMSE}}$	$(0.1686)_7\ (0.2593)_6$	$(0.0246)_7\ (0.1381)_6$	$(0.0938)_7\ (0.1036)_6$	$(0.0983)_7\ (0.1315)_6$	$(0.1462)_7\ (0.1727)_6$
		$(0.4047)_5\ (0.6713)_4$	$(0.3915)_5\ (0.5125)_4$	$(0.3224)_5\ (0.6186)_4$	$(0.2674)_5\ (0.4534)_4$	$(0.58427)_5\ (0.7218)_4$
		$(0.7253)_3\ (0.8212)_2$	$(0.9246)_3\ (0.8001)_2$	$(0.4274)_3\ (0.6194)_2$	$(0.7041)_3\ (0.6709)_2$	$(0.5312)_3\ (0.5568)_2$
		$(0.9218)_1\ (0.0105)_0$	$(0.8992)_1\ (0.0015)_0$	$(0.9452)_1\ (0.0058)_0$	$(0.9802)_1\ (0.0061)_0$	$(0.8719)_1\ (0.0091)_0$
		$(0.0103)_{c_0}$	$(0.0017)_{c_0}$	$(0.0052)_{c_0}$	$(0.0064)_{c_0}$	$(0.0097)_{c_0}$
Gaussian	MMSE	0.0087	0.0212	0.0220	0.0066	0.0050
	$\widetilde{\kappa}^j_{\text{MMSE}}$	$(12.8746)_7\ (9.0687)_6$	$(10.7532)_7\ (7.5573)_6$	$(14.2840)_7\ (10.0657)_6$	$(14.2870)_7\ (10.0624)_6$	$(14.2201)_7\ (10.4309)_6$
		$(6.3501)_5\ (4.5513)_4$	$(5.3519)_5\ (3.7564)_4$	$(7.0562)_5\ (4.9908)_4$	$(7.0751)_5\ (5.0512)_4$	$(7.0741)_5\ (5.0153)_4$
		$(3.1504)_3\ (2.1523)_2$	$(2.6509)_3\ (1.6519)_2$	$(3.4408)_3\ (2.2274)_2$	$(3.5387)_3\ (2.4063)_2$	$(3.5331)_3\ (2.4325)_2$
		$(1.5532)_1\ (0.8046)_0$	$(1.2501)_1\ (0.6720)_0$	$(1.6214)_1\ (0.8927)_0$	$(1.5237)_1\ (0.8929)_0$	$(1.5025)_1\ (0.8887)_0$
		$(0.8625)_{c_0}$	$(0.6275)_{c_0}$	$(0.8271)_{c_0}$	$(0.89375)_{c_0}$	$(0.8563)_{c_0}$

Table 9.9: The results for minimum MSE of the different priors using level-dependent prior procedure to estimate parameters $\boldsymbol{\theta}$ for different core samples, where minimum MSE is described in Section 8.4. The bold font represents the smallest MSE.

There are three priors considered corresponding to the elastic-net with different values of γ, the Laplace and the Gaussian prior. The elastic-net provides excellent results and a smaller MMSE than Gaussian and Laplace. The Gaussian prior provides results, in the MSE sense, similar to the elastic-net, whereas the Laplace prior provides a larger MSE than the elastic-net and the Gaussian priors. Overall, for each of the simulated cores considered, the elastic-net prior outperformed each of the other priors (Gaussian and Laplace). Additionally, the elastic-net prior performed optimally and this was the most computationally efficient of the single component priors.

Table 9.9 shows that the results of the MMSE calculations for level-dependent priors.

Overall, the elastic-net prior provides a reconstruction which is close to the true susceptibility profile of the simulated core. It can be concluded that the one-stage method provides a somewhat better performance for estimating the underlying test signals than the two-stage method.

9.4 Reconstruction for the real data

General

In this chapter, there are three procedures to estimate the true susceptibility profiles. The first is based on two-stage estimation using the values of parameters θ or $\boldsymbol{\theta}$ taken from column (2) of Tables 9.3, 9.2, 9.4, 9.6, 9.5 and 9.7. See Figures 9.4 and 9.5 for results.

The second procedure is based on the one-stage approach where the values of parameters θ or $\boldsymbol{\theta}$ are fixed and taken from column (4) in Tables 9.8 and 9.9 and used to obtain PM and MAP reconstructions. See Figures 9.8, 9.9, 9.6 and 9.7. In the one-stage procedure, a value for the variance of the noise σ^2 is estimated from the finest level of detail coefficients in the wavelet decomposition as shown in Equation (2.63).

The third approach is based on the hierarchical model where the parameters θ or $\boldsymbol{\theta}$ and the variance of the noise σ^2 are described by prior distributions. Figures 9.10, 9.12, 9.11 and 9.13 show the reconstructions from real data. It can be concluded that the hierarchical model with elastic-net provide excellent results.

IT-TO and WVD-TO reconstructions from MMSE results

In the case of the two-stage method, the values of parameters $\boldsymbol{\theta}$ or θ are obtained from column (2) in Tables 9.3, 9.2, 9.4, 9.6, 9.5 and 9.7. Using the values in the other columns created smoothing and undesirable reconstructions. The results using the parameter values from core (2) are shown as they give better reconstructions.

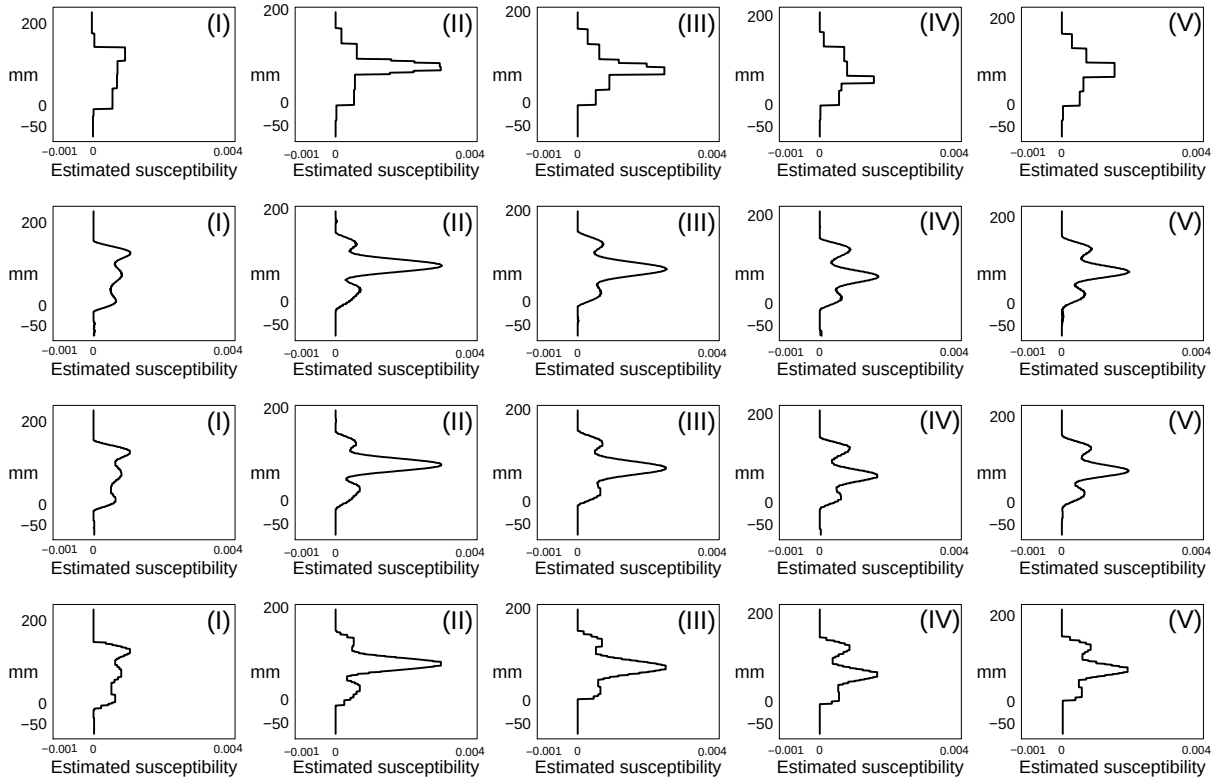


Figure 9.4: Plots of the reconstructions using the IT-TO method for estimating pyre cores: the first row represents the reconstruction using the hard rule; the second row represents the reconstruction using the DWWS-LPM method; the third row represents the reconstruction using the EB (posterior median) method; and the fourth row represents reconstruction using the DUHT method, where first-order method is used.

Four techniques of thresholding are chosen: the hard, the DWWS-LPM, the EB and the DUHT methods. The reason for choosing these thresholding methods is that they have been shown to provide a step function reconstruction and improved MSE. Also, different types of transform are chosen, such as the DWT and the DUHT. Additionally, different priors are involved such as single prior (DWWS-LPM) and mixture prior (EB).

Figures 9.4 and 9.5 show reconstructions of the five cores using the two-stage method with the IT-TO and the WVD-TO procedures, respectively. The panels are organised as follows; the first row shows the reconstructions using the hard rule; the second row shows the reconstructions using the DWWS-LPM; the third row shows the reconstruction using

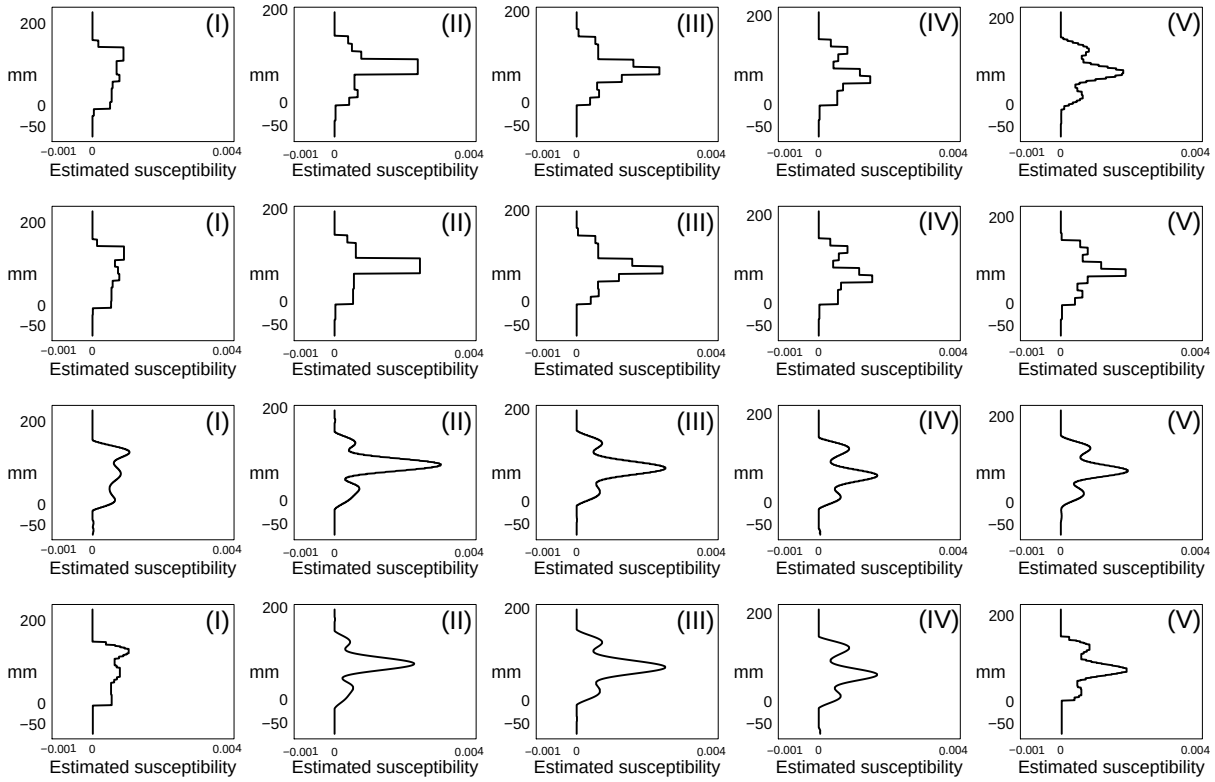


Figure 9.5: Plots of the reconstructions using the WVD-TO method for estimating pyre cores: the first row represents the reconstruction using the hard rule; the second row represents the reconstruction using the DWWS-LPM; the third row represents the reconstruction using the EB; and the fourth row represents reconstruction using the DUHT with first-order smoothing method.

the EB; and the fourth row shows reconstructions using the DUHT with the first-order method. And the columns represent the reconstructions of different real data, which are taken from the 'Park', Guiting Power.

It can be seen from Figure 9.4 that the hard thresholding rule provides sharp edges and flat topped reconstructions, and the features of the cores can be seen clearly. Figure 9.5 shows reconstructions of the cores using the two-stage method with the WVD-TO procedure, where it can be seen that the hard and the DWWS-LPM rules provide sharp edges and flat topped reconstructions except the hard reconstruction of pyre (V).

In general, the IT-TO involving the hard thresholding rule, the WVD-TO with the hard thresholding rule and the DWWS-LPM method, provide good shape to the reconstruction. Moreover, reconstruction with the WVD-TO involving the hard thresholding rule, and the DWWS-LPM method gives the main blocks clearly. A major limitation of the two-stage algorithm is that it is not possible to estimate the standard errors and then to compute the credible interval.

PM and MAP reconstruction using MMSE parameters

The aim of this subsection is to show the reconstructions obtained from the PM and the MAP estimates where the prior parameters are taken from columns (4) in Tables 9.8 and 9.9 and the total number of runs are equal to $256 \times 6000 = 15.36 \times 10^5$.

In the case of the one-stage method, the procedure of MMSE, which is described in Section 8.4, provides 60 reconstructions for each pyre, with total of 300 reconstructions for estimating five pyre cores, where the values of parameters $\theta = \{\kappa\}$, $\boldsymbol{\theta} = \{\kappa, \gamma\}$, $\boldsymbol{\theta} = \{\boldsymbol{\kappa}\}$ or $\boldsymbol{\theta} = \{\boldsymbol{\kappa}, \boldsymbol{\gamma}\}$ are obtained from the MMSE approach. Additionally, these prior parameters are applied in the one-stage estimation to produce a reconstruction. The results using the parameter values from core (4) are shown, as they give better reconstructions.

Figures 9.6, 9.7, 9.8 and 9.9 show the PM and the MAP estimates. The panels are organised as follows: the first row represents reconstruction using the Laplace prior, the second row represents reconstruction using the elastic-net prior, while the third row represents reconstruction using the Gaussian prior.

Figure 9.6 and 9.7 show PM reconstructions with single parameters κ and γ from column (4) in Table 9.8, whereas the level-dependent parameter values are taken from column (4) in Table 9.9.

In general, the Laplace prior provides a step reconstruction, but it also suggests that pyre (I) has three peaks and the middle block is clearer than the elastic-net prior and than the Gaussian reconstructions. For estimating pyre (II), the shape of the block is

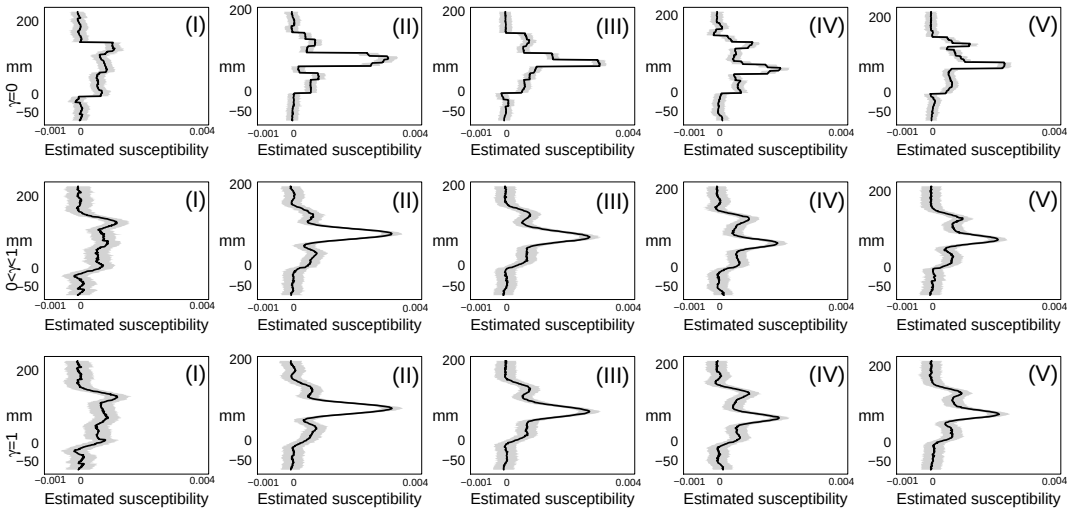


Figure 9.6: Plots of the reconstructions using the PM procedure, which is described in Section 8.4, for estimating pyre cores: parameters κ and γ are fixed; the first row represents the reconstruction using the Laplace; the second row represents the reconstruction using the elastic-net; and the third row represents the reconstruction using the Gaussian prior.

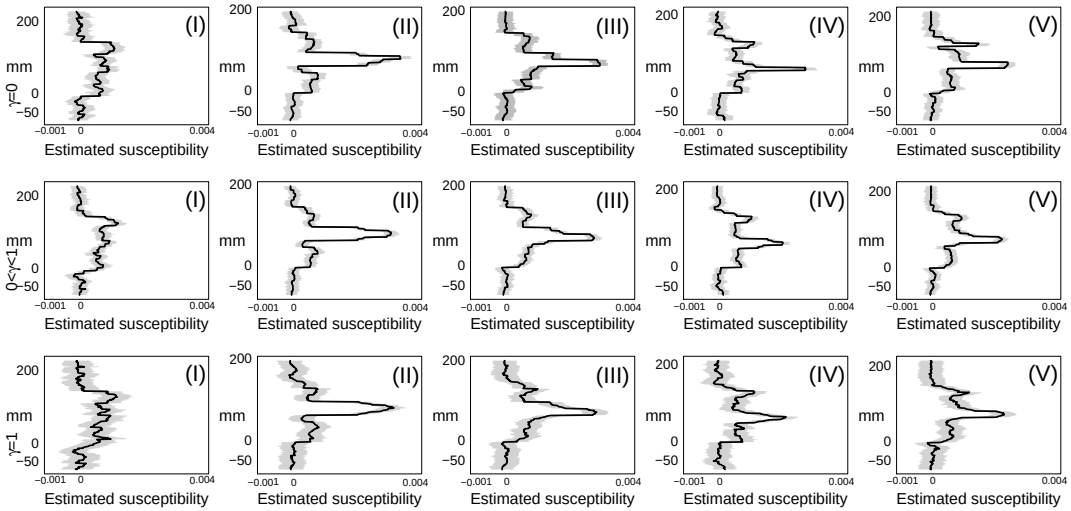


Figure 9.7: Plots of the reconstructions using the PM procedure, which is described in Section 8.4, for estimating pyre cores: level-dependent prior is used and parameters $\boldsymbol{\kappa}$ and $\boldsymbol{\gamma}$ are fixed; the first row represents the reconstruction using the Laplace; the second row represents the reconstruction using the elastic-net; and the third row represents the reconstruction using the Gaussian.

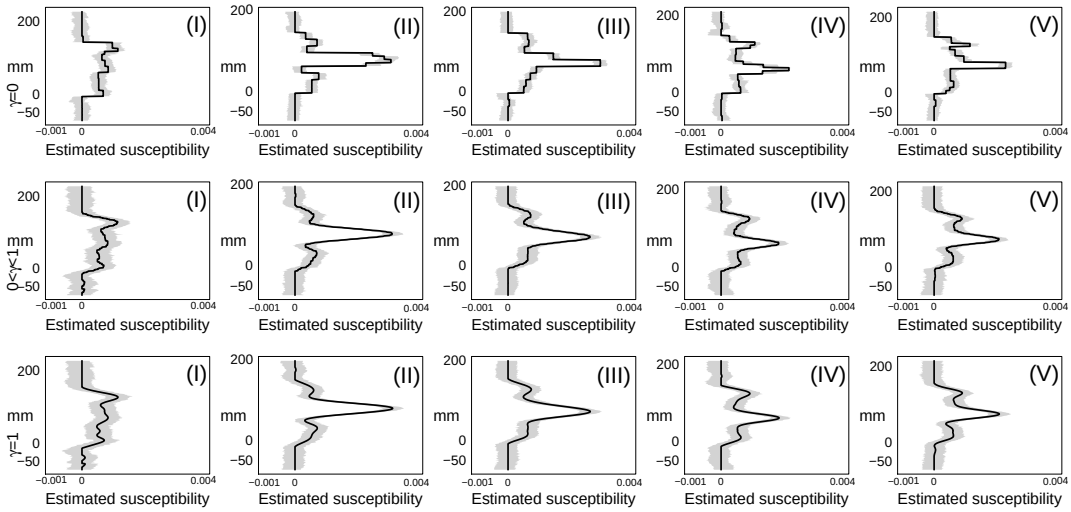


Figure 9.8: Plots of the reconstructions using the MAP procedure, which is described in Section 8.4, for estimating pyre cores: a parameters κ and γ are fixed; the first row represents the reconstruction using Laplace; the second row represents the reconstruction using elastic-net; and the third row represents the reconstruction using Gaussian.

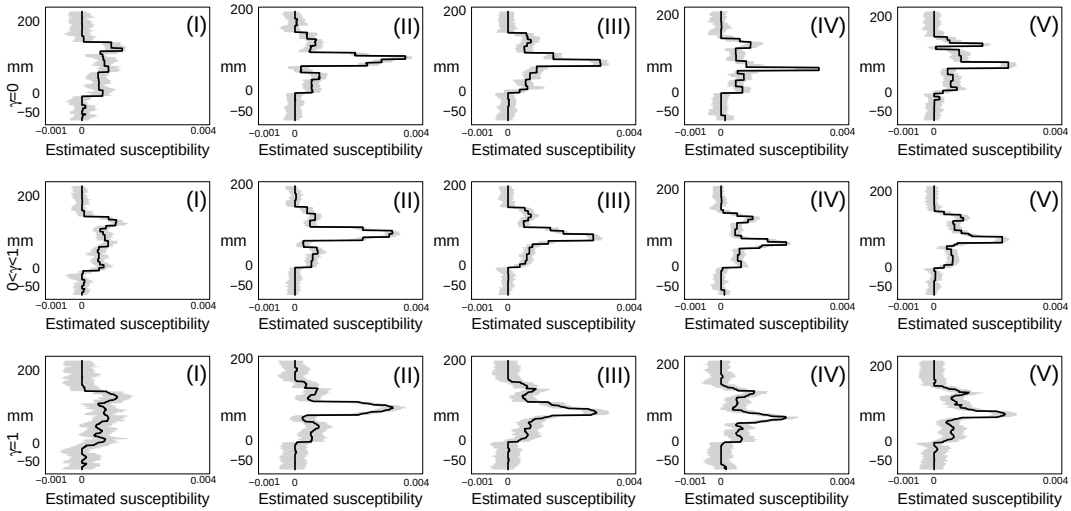


Figure 9.9: Plots of the reconstructions using the MAP procedure, which is described in Section 8.4, for estimating pyre cores: level-dependent prior is used and parameters $\boldsymbol{\kappa}$ and $\boldsymbol{\gamma}$ are fixed; the first row represents the reconstruction using Laplace; the second row represents the reconstruction using elastic-net; and the third row represents the reconstruction using Gaussian.

not fully resolved from the Laplace prior, and the results of the elastic-net prior and the Gaussian prior reconstructions are less satisfactory than the reconstruction obtained from the Laplace prior. The PM estimates with the Laplace prior for estimating pyre (III), (IV) and (V) provide excellent results and the features can be clearly identified.

Figure 9.8 shows the MAP reconstructions with single parameters, κ and γ. These parameters are chosen from column (4) in Table 9.8. Similarly, the Laplace prior provides sharp edges and flat top reconstructions. Additionally, it suggests that pyre (I) has three peaks and the middle block is clearer than the elastic-net prior and the Gaussian reconstructions. For estimating pyre (II) the shape of the block is not fully resolved from Laplace prior and the results are less satisfactory than the elastic-net and the Gaussian reconstructions. The MAP estimation with the Gaussian prior creates sloping sides around the peak and the features cannot be identified.

Figure 9.9 shows the MAP reconstructions using a level-dependent prior with parameters from column (4) in Table 9.9. The shape of the reconstruction is a step function and the features of the reconstruction with the Gaussian prior can be identified. The reconstructions with the elastic-net and the Laplace are better than with the Gaussian prior when using single prior parameters for all wavelet coefficients. The reconstructions with the elastic-net and the Laplace also suggest that pyre (I) has three peaks and the middle block is clearer than the Gaussian reconstruction. When estimating pyre (II) the shape of the block is not fully resolved and the results are again less satisfactory with the Gaussian prior, although for all reconstructions the start and the end point can be detected. Finally, the MAP estimators for estimating pyre (III), (IV) and (V) provide excellent results and the features can be identified. The Gaussian prior with level-dependent prior provides a step reconstruction. On the other hand, the Gaussian prior with single prior parameters for κ and γ provide a smooth reconstruction. Overall, the MAP estimates with a level-dependent prior provide a good shape and flat top for the reconstructions.

It can be concluded that Laplace and elastic-net priors provide a step reconstruction and the features can be identified and it is difficult to detect the feature of pyre from the

reconstructions of Gaussian prior. Additionally, it is not necessary to use parameters estimated from datasets to obtain a good reconstruction.

9.5 PM and MAP reconstruction with simultaneous prior parameter estimation

Estimation using single prior for all wavelet coefficients

The purpose of this subsection is to apply the proposed method, introduced in Section 8.3, to real data. The wavelet coefficients and the prior parameters are estimated using (i) a single prior distribution for all wavelet coefficients and (ii) level-dependent prior distributions. The total number of runs is equal to $256 \times 6000 = 15.36 \times 10^5$, where the number of iterations equals 256 and the number of replications equals 6000.

In general, the PM estimates provide a good shape and flat topped reconstructions. The main features of the reconstructions with all priors can be seen clearly. Also, the reconstructions with the elastic-net and the Laplace suggest that pyre (I) has three peaks and the middle block is more clear than the Gaussian prior reconstruction. For pyre (II) the shape of the blocks is fully resolved and the start and end points are identical. The MAP estimates for pyre (III), (IV) and (V) provide excellent results and the features can be identified well.

Figure 9.10 shows the PM reconstructions including estimation of the parameters κ and γ. The panels are organised as follows: the first row shows the reconstructions using the Laplace prior, the second row shows the reconstructions using the elastic-net while the third row shows the reconstructions using the Gaussian prior. The credible intervals are also plotted using the sample quantiles, as explained in Chapter 8.

Similarly, Figure 9.12 shows the PM reconstructions using a level-dependent wavelet coefficients prior. PM estimates provide a good shape and flat top for the reconstructions.

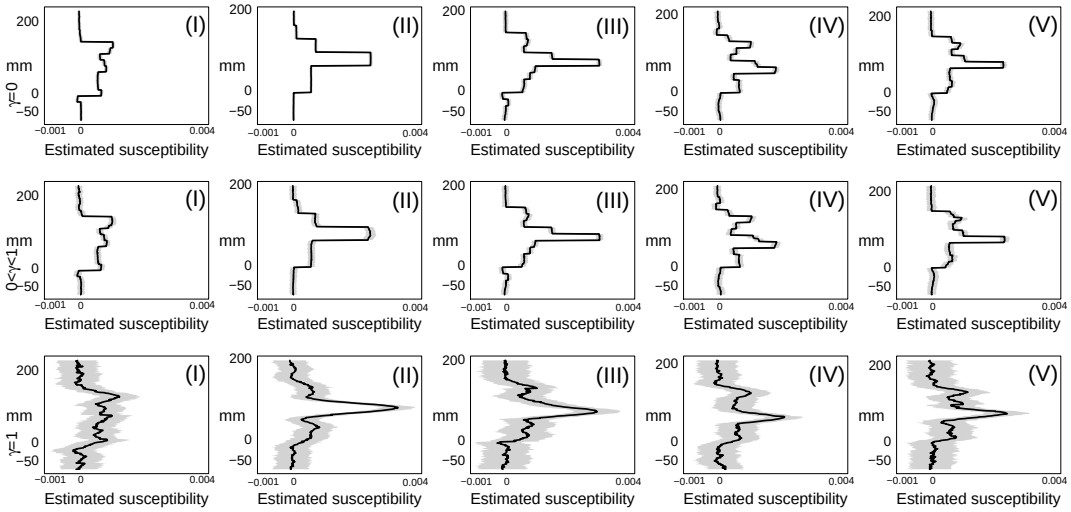


Figure 9.10: Plots of the reconstructions using the PM procedure, which is described in Section 8.4, for estimating pyre cores: parameters κ and γ are estimated; the first row represents the reconstruction using the Laplace; the second row represents the reconstruction using the elastic-net; and the third row represents the reconstruction using the Gaussian.

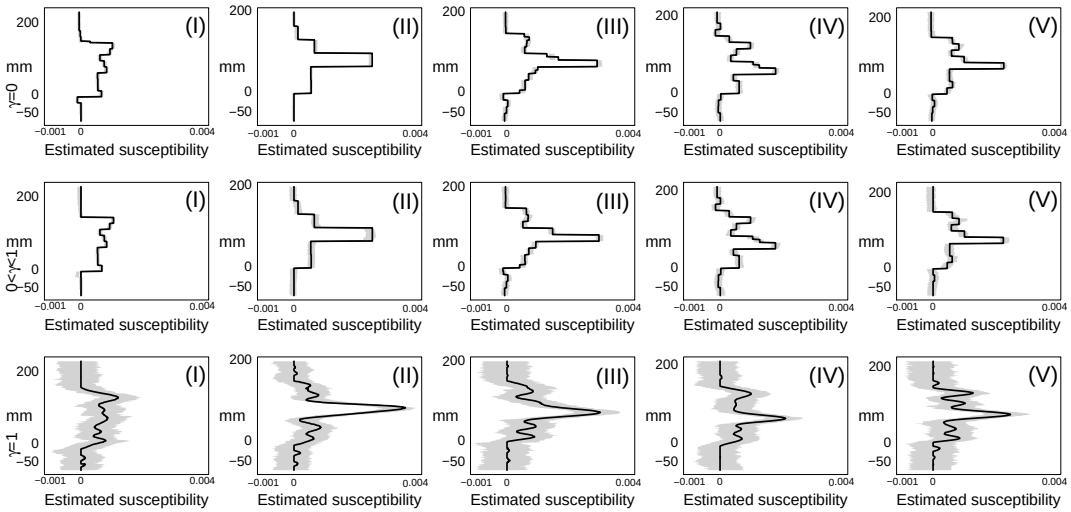


Figure 9.11: Plots of the reconstructions using the MAP procedure, which is described in Section 8.4, for estimating pyre cores: parameters κ and γ are estimated; the first row represents the reconstruction using the Laplace; the second row represents the reconstruction using the elastic-net; and the third row represents the reconstruction using the Gaussian.

The main features of the reconstructions of elastic-net, Laplace and Gaussian can be identified.

Figure 9.11 shows the MAP reconstructions where the parameters κ and γ are estimated using a single prior and the credible intervals are plotted by using the sample quantiles. The MAP estimates provide a good shape for the reconstructions, and it can be seen that the shape of the reconstruction obtained from the estimated single prior parameters is a step function.

Also, the reconstruction with the Gaussian prior is smoother than with the elastic-net and the Laplace prior. However, the reconstruction is smooth due to the jumps being estimated by a series of small steps and, as a result, the shape of the block is not fully resolved. The main features of the reconstructions with the elastic-net and the Laplace prior can be identified.

The reconstructions of the elastic-net and the Laplace also suggest that pyre (I) has three peaks with the middle block clearer than the Gaussian reconstructions. For pyre (II) the shape of the block is fully resolved and for all reconstructions the start and end point can be identified. Finally, the MAP estimates for pyre (III), (IV) and (V) provide excellent results and the main features can be identified.

Estimation using level-dependent coefficient priors

Figure 9.13 shows the MAP reconstructions using level-dependent priors where there is a separate prior for each wavelet coefficient level. The MAP estimates give better performance, where the shapes of the reconstructions obtained are step functions, even, when the reconstructions use the Gaussian prior.

The main features of the reconstructions using the elastic-net prior, the Laplace and the Gaussian priors can also be detected. The reconstructions of the elastic-net and Laplace also suggest that pyre (I) has three peaks, with the middle block clearer than the Gaussian reconstructions. For pyre (II) the shape of the block is fully resolved. For all

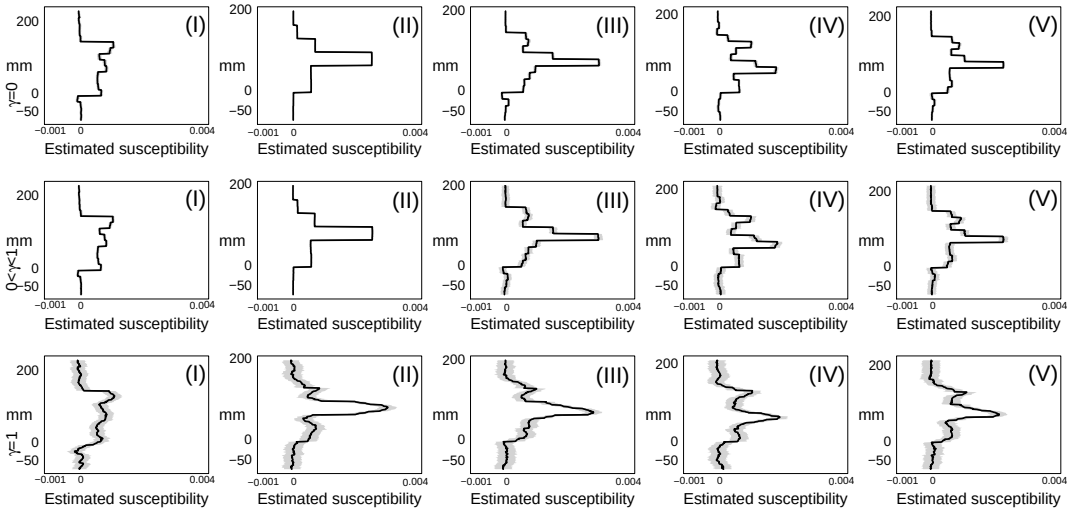


Figure 9.12: Plots of the reconstructions using the PM procedure, which is described in Section 8.4, for estimating pyre cores: level-dependent priors are used; the first row represents the reconstruction using the Laplace; the second row represents the reconstruction using the elastic-net; and the third row represents the reconstruction using the Gaussian.

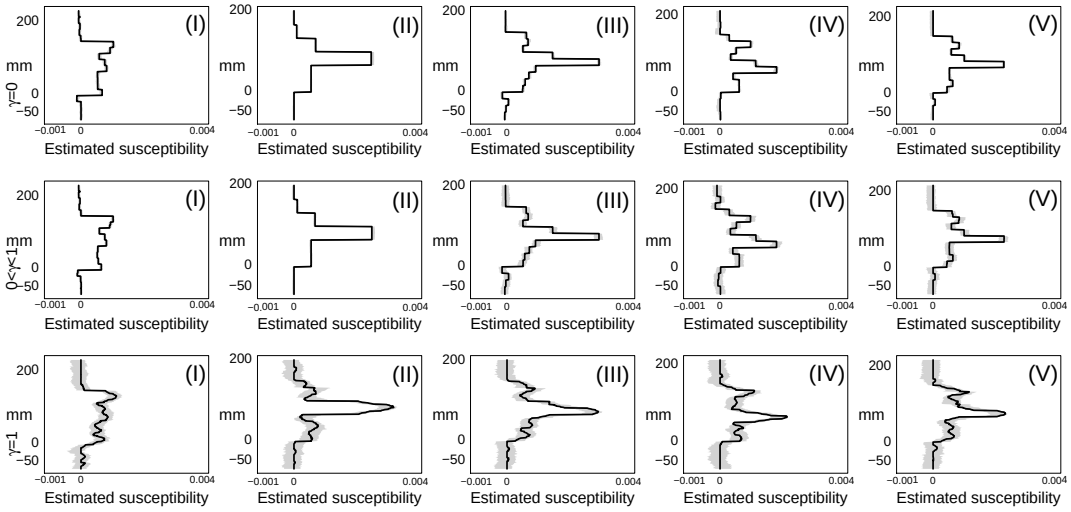


Figure 9.13: Plots of the reconstructions using the MAP procedure, which is described in Section 8.4, for estimating pyre cores: level-dependent priors are used; the first row represents the reconstruction using the Laplace; the second row represents the reconstruction using the elastic-net; and the third row represents the reconstruction using the Gaussian.

reconstructions the start and end point can also be identified. Finally, MAP estimates for pyre (III), (IV) and (V), which is described in Section 8.4, provide excellent results and the features can be identified.

9.6 Summary of main features from real data

The individual reconstructed profiles were inspected to identity the main features numerically. Results are summarized in Tables 9.10, 9.11, 9.12, 9.13 and 9.14 showing the main features of pyre cores (I), (II), (III), (IV) and (V) based on the MAP results for the elastic-net, with a level-dependent wavelet coefficient prior. It can be seen that the length of all pyre cores, except (II), are equal to 144 mm the main area of burning has magnetic susceptibility between 0.3 (SI x10^{-3}) and 0.632 (SI x10^{-3}), and magnetic susceptibility of the region of the site has magnetic susceptibility between 0.133 (SI x10^{-3}) and 2.57 (SI x10^{-3}). The distance d_1, which represents the recordings before the plastic cylinder enters the detector, is 65 mm $< d_1 <$ 69 mm. The distance $\tilde{d}$, that represents the last distance after the core has emerged, is 34 mm $< \tilde{d} <$ 82 mm. It can be seen that all pyre cores have the same extent.

		d_1 (mm)	d_2 (mm)	d_3 (mm)	d_4 (mm)	d_5 (mm)	d_6 (mm)	d_7 (mm)	d_8 (mm)	d_9 (mm)	d_{10} (mm)	length of pyre (mm)	variance SI x10^{-3}
		65	16	32	16	16	16	16	16	16	82		
magnetic	x_F SI x10^{-3}	-	0.688	-	-	0.845	0.768	0.624	0.944	1.07	-	144	0.0305
susceptibility	x_B SI x10^{-3}	-	-	0.553	0.543	-	-	-	-	-	-		
	SI x10^{-3}	0	-	-	-	-	-	-	-	-	0		

Table 9.10: Estimating length parameters of the core and feature susceptibility for pyre (I): d_1 represents the recording start, before the plastic cylinder enters. The distances d_3 and d_4 represent the second and third parts of the core and they have susceptibility, which represents a background susceptibility x_B. The distances d_2, d_5, d_6, d_7, d_8 and d_9 represent the first, fourth, fifth, sixth, seventh and eighth parts of the core respectively and they have susceptibility which represents an archaeological feature with susceptibility x_F, and d_{10} represents the last distance after the core has emerged.

		d_1 (mm)	d_2 (mm)	d_3 (mm)	d_4 (mm)	d_5 (mm)	d_6 (mm)	d_7 (mm)	d_8 (mm)	length of pyre (mm)	variance SI x10^{-3}
		66	16	16	32	32	32	32	34		
magnetic	x_F SI x10^{-3}	-	-	-	-	2.57	0.666	-	-		
susceptibility	x_B SI x10^{-3}	-	0.571	0.548	0.548	-	-	-	-	128	0.723
	SI x10^{-3}	0	-	-	-	-	-	0	0		

Table 9.11: Estimating length parameters of the core and feature susceptibility for pyre (II): in d_7, it is difficult to judge magnetic susceptibility as a background susceptibility or negligible magnetic susceptibility; d_1 represents the recording start, before the plastic cylinder enters. The distances d_2, d_3 and d_4 represent the first, second, third and eighth parts of the core respectively and they have susceptibility which represents a background susceptibility. The distances d_5 and d_6 represent the fourth and fifth parts of core and they have susceptibility which represents an archaeological feature with susceptibility x_F, and d_8 represents the last distance after the core has emerged.

		d_1 (mm)	d_2 (mm)	d_3 (mm)	d_4 (mm)	d_5 (mm)	d_6 (mm)	d_7 (mm)	d_8 (mm)	d_9 (mm)	d_{10} (mm)	d_{11} (mm)	d_{12} (mm)	length of pyre (mm)	variance SI x10^{-3}
		67	8	8	16	16	16	16	16	16	16	16	52		
magnetic	x_F SI x10^{-3}	-	-	-	-	0.716	0.946	3.22	1.49	-	0.695	0.636	-		
susceptibility	x_B SI x10^{-3}	-	0.416	0.597	0.587	-	-	-	-	0.520	-	-	-	144	0.0532
	SI x10^{-3}	0	-	-	-	-	-	-	-	-	-	-	0		

Table 9.12: Estimating length parameters of the core and feature susceptibility for pyre (III): d_1 represents the recording start, before the plastic cylinder enters. The distances d_2, d_3, d_4 and d_9 represent the first, second, third and eighth parts of the core respectively and they have susceptibility which represents a background susceptibility x_B. The distances d_5, d_6, d_7, d_8, d_{10} and d_{11} represent the fourth, fifth, sixth, seventh, ninth and tenth parts of core and they have susceptibility which represents an archaeological feature with susceptibility x_F, and d_{12} represents the last distance after the core has emerged.

		d_1 (mm)	d_2 (mm)	d_3 (mm)	d_4 (mm)	d_5 (mm)	d_6 (mm)	d_7 (mm)	d_8 (mm)	d_9 (mm)	d_{10} (mm)	d_{11} (mm)	length of pyre (mm)	variance SI x10^{-3}
		69	32	16	16	8	8	16	16	16	16	60		
magnetic	x_F SI x10^{-3}	-	0.628	-	1.82	1.30	1.07	-	0.538	1.00	-	-		
susceptibility	x_B SI x10^{-3}	-	-	0.438	-	-	-	0.353	-	-	0.300	-	144	0.0786
	SI x10^{-3}	0	-	-	-	-	-	-	-	-	-	0		

Table 9.13: Estimating length parameters of the core and feature susceptibility for pyre (IV): d_1 represents the recording start, before the plastic cylinder enters. The distances d_3, d_7 and d_{10} represent the second, sixth and ninth parts of the core respectively and they have susceptibility, which represent a background susceptibility x_B. The distances d_2, d_4, d_5, d_6, d_8 and d_9 represent the first, third, fourth, fifth, seventh and eighth parts of the core respectively and they have susceptibility which represents an archaeological feature with susceptibility x_F, and d_{11} represents the last distance after the core has emerged.

		d_1 (mm)	d_2 (mm)	d_3 (mm)	d_4 (mm)	d_5 (mm)	d_6 (mm)	d_7 (mm)	d_8 (mm)	d_9 (mm)	d_{10} (mm)	length of pyre (mm)	variance SI x10^{-3}
		69	16	16	32	16	16	16	16	16	65		
magnetic	x_F SI x10^{-3}	-	-	0.628	-	2.30	1.03	-	0.853	-	-		
susceptibility	x_B SI x10^{-3}	-	0.462	-	0.547	-	-	0.600	-	0.632	-	144	0.0725
	SI x10^{-3}	0	-	-	-	-	-	-	-	-	0		

Table 9.14: Estimating length parameters of the core and feature susceptibility for pyre (V): d_1 represents the recording start, before the plastic cylinder enters. The distances d_2, d_4, d_7 and d_9 represent the first, third, sixth and eighth parts of the core respectively and they have susceptibility, which represent a background susceptibility x_B. The distances d_3, d_5, d_6 and d_8 represent the second, fourth, fifth and seventh parts of the core respectively and they have susceptibility which represent an archaeological feature with susceptibility x_F, and d_{10} represents the last distance after the core has emerged.

Cores (I), (III), (IV) and (V) show a length of core equal to 144 mm, whereas pyre (II) shows length equal to 128 mm.

The magnetic susceptibility of pyre (II), before the plastic cylinder has emerged, is lower than pyre cores (I), (III), (IV) and (V). Note that, it is possible that core (II) shrank due to drying out after collection or that there was a problem while recording the sample.

9.7 Conclusions

An observation of the susceptibility profile alone cannot uniquely define the various physical features of the true susceptibility. It is not easy to estimate the parameters from highly correlated data and using simple methods might provide an overestimate or underestimate that depends on the type of method and the level of the smoothing occurring at sharp changes in the susceptibility profile. This impact cannot be overcome by tuning the value of the smoothing parameter Λ and changing the method, instead it is a property of the problem being considered.

One-stage and two-stage estimates clearly provide better estimation than using only an inversion method, due to producing sharp edges and flat-topped reconstructions. In contrast, inversion methods involving smoothing create sloping sides around the peak rather than the sharp vertical edge that are expected by archaeologists. Although a variety of methods exist, two-stage approaches are widely used for filtering and thresholding. The locations of clearly separated blocks of susceptibility are then predicted quite well, but their shapes are smooth with no sharp edges. Furthermore, decreasing the level of the smoothing parameter Λ does not help as the reconstruction is affected by noise. An alternative solution to the problem of smoothing would be to consider the one-stage estimators, which improve the sharpness of the edges. In particular, the one-stage method provides excellent results where the main features can be identified clearly.

The procedures for estimation can be divided into three types. Firstly, those where the parameters $\boldsymbol{\theta}$ or θ are estimated from simulated data and the resulting estimates are used for estimating the susceptibility from real data, where the two-stage method is used. The second is similar, except one-stage reconstruction is used. The third, is to use a fully Bayesian approach, where the prior distributions are included and estimation and reconstruction are preformed simultaneously.

In this chapter, the prior is chosen as the elastic-net, the Gaussian and the Laplace. In addition, the method developed provides a general framework for reconstruction where

sharp edges are believed to be important.

Finally, the method was illustrated on real data from an experiment to investigate the effect of burning on magnetic susceptibility. The method was also extended to the elastic-net, the Laplace and the Gaussian with one prior and level-dependent priors for the wavelet coefficients, which involved a hierarchical model. By using these methods, the local changes in susceptibility can be more clearly distinguished.

Chapter 10

Application to 2D magnetometry data

10.1 Overview

This chapter is organised as follows: Section 10.2 provides an introduction, Section 10.3 explains wavelet methods for two dimensional data, Section 10.4 discusses estimation of σ^2, Section 10.5 gives an introduction to point spread function, whilst Section 10.6 describes archaeological images, Section 10.7 considers Bayesian modelling of images, Sections 10.8 and 10.9 give the experimental results on simulated images, and apply the methods to 2D real data. Finally, Section 10.10 gives the conclusions.

10.2 Introduction

Within this chapter, a new model is considered that depends on a statistical approach and uses a stochastic algorithm for estimation of an image corrupted by noise and blur. In particular, the proposed method describes an image in terms of matrices of wavelet coefficients. The true image is estimated, in a Bayesian framework, using a Markov chain

Monte Carlo algorithm (MCMC).

There are a wide range of articles applying MCMC algorithms to images. To name a few, Besag (1983), who made the first recommendations for the use of prior information in image processing, Geman and Geman (1984) set out the idea for pixel-prior distributions and Qian and Titterington (1991) used the multidimensional Markov chain as a model for texture and two types of stochastic models for texture to describe an image Markov mesh (MM) and the Markov random field (MRF). Aykroyd (1998) used the multidimensional Markov chain as a model for texture and investigated homogeneous and inhomogeneous Gaussian random fields. Cross and Jain (1983) also used Markov random fields as texture models, with binomial model taken as the basic model for the analysis. Each point in the texture image has a binomial distribution with the probability parameters controlled by the values of its neighbors.

There are several books and papers on the use of 2D-dimensional wavelets in the image processing community. Denoising is a popular area of study, for example Bruce and Gao (1996a) presented four wavelet applications: digital image compression; noise removal, time-frequency analysis and the speeding up and improvement of classification algorithms. Nason (2010a) explained the procedure of thresholding and gives an example of denoising an image. Other articles in image processing using wavelet coefficients, include Antonini *et al.* (1992), who used two procedures to process an image; one is based on a wavelet transform and the other based on Shannon's rate distortion theory.

In the remainder of this chapter, a two dimensional inverse problem is considered. Section 10.3 explains wavelet methods for two dimensional data, this section also defines an orthogonal matrix for two dimensional data. Section 10.4 discusses estimation of the variance of noise for an image. In Section 10.5, the point spread function will be explained. Section 10.6 describes archaeological data. Section 10.7 shows that there are two types of prior model proposed. The first depends on only a single prior parameter for all wavelet coefficients, and the other is based on level-dependent priors. Different probability distributions will also be applied. The proposed methodology is compared in

extensive simulations and application to real data. Sections 10.8 and 10.9 give simple examples of experimental result on simulated images to investigate and discover any error attributed to the inversion procedure and not by mismodelling.

10.3 Wavelet coefficients for 2D images

As the results in Chapter 9 showed, the proposed techniques provide sharp edges and flat topped reconstructions. These methods will now be adapted for use on two dimensional magnetometry data. In this chapter only the PM and the MAP estimates, using the MCMC approach, will also be used to provide reconstructions.

Two dimensional wavelets are used in applications involving images, matrices, and other two dimensional data. The properties which make wavelets attractive for analysis of one dimensional functions hold for two dimensional functions as well. In particular, the two dimensional wavelets are constructed by taking the tensor product of a *horizontal* 1-D wavelet and a *vertical* 1-D wavelet. This procedure leads to four different types of 2-D wavelets

$$\Phi(x,y) = \phi_h(x) \times \phi_v(y), \tag{10.1}$$

$$\Psi^v(x,y) = \psi_h(x) \times \phi_v(y), \tag{10.2}$$

$$\Psi^h(x,y) = \phi_h(x) \times \psi_v(y), \tag{10.3}$$

$$\Psi^d(x,y) = \psi_h(x) \times \psi_v(y), \tag{10.4}$$

where (x,y) is a point in the unit square, $\phi_h(x)$ is the horizontal scaling, $\phi_v(y)$ is the vertical scaling, $\psi_h(x)$ is the horizontal wavelets and $\psi_v(y)$ is the vertical wavelets functions. Hence, the 2-D wavelet family has one father function $\Phi(x,y)$ and three mother wavelet functions $\Psi^v(x,y), \Psi^h(x,y)$, and $\Psi^d(x,y)$. These capture the detail in the vertical, horizontal, and diagonal directions, respectively, whereas $\Phi(x,y)$ captures the smooth part of the image. This can be simplified by stating that two dimensional wavelets have one scaling function and three wavelet functions. A two dimensional function, $F(x,y)$, can

be expressed as a series expansion in terms of three wavelet functions and one scaling function. In particular, the j-level approximation for a $M \times N$ discrete image, $\mathbf{F}_{M\times N}$, can be written as

$$\begin{aligned} F(x,y) \approx & \sum_{m=1}^{2^j}\sum_{n=1}^{2^j} \mathrm{s}_{j,m,n}\Phi_{j,m,n}(x,y) \\ & + \sum_{j}^{J}\sum_{m=1}^{2^j}\sum_{n=1}^{2^j} \mathrm{d}^{v}_{j,m,n}\Psi^{v}_{j,m,n}(x,y) \\ & + \sum_{j}^{J}\sum_{m=1}^{2^j}\sum_{n=1}^{2^j} \mathrm{d}^{h}_{j,m,n}\Psi^{h}_{j,m,n}(x,y) \\ & + \sum_{j}^{J}\sum_{m=1}^{2^j}\sum_{n=1}^{2^j} \mathrm{d}^{d}_{j,m,n}\Psi^{d}_{j,m,n}(x,y). \end{aligned} \tag{10.5}$$

Hence, $F(x,y)$ can be expressed in terms of a sum of wavelet coefficients and corresponding scaling and wavelet functions. Furthermore, Bruce and Gao (1996a) showed that two dimensional basis functions are generated from one father wavelet Φ and three mother wavelets Ψ^h, Ψ^v, and Ψ^d, by the scaling and translation as follows:

$$\begin{aligned} \Phi_{j,m,n}(x,y) =& 2^{-j}\Phi(2^{-j}x-m, 2^{-j}y-n), \quad 1 \le m < 2^j, 1 \le n < 2^j, \quad 1 \le j \le J, \\ \Psi^{v}_{j,m,n}(x,y) =& 2^{-j}\Phi^{v}(2^{-j}x-m, 2^{-j}y-n), \quad 1 \le m < 2^j, 1 \le n < 2^j, \quad 1 \le j \le J, \\ \Phi^{h}_{j,m,n}(x,y) =& 2^{-j}\Phi^{h}(2^{-j}x-m, 2^{-j}y-n), \quad 1 \le m < 2^j, 1 \le n < 2^j, \quad 1 \le j \le J, \text{and} \\ \Phi^{d}_{j,m,n}(x,y) =& 2^{-j}\Phi^{d}(2^{-j}x-m, 2^{-j}y-n), \quad 1 \le m < 2^j, 1 \le n < 2^j, \quad 1 \le j \le J, \end{aligned}$$

where j represents the resolution, $j = 0, 1, \ldots, J-1$, $m = 1, 2, \ldots, M$, and $n = 1, 2, \ldots, N$.

Moreover, the 2D wavelet transform coefficient is given approximately by the integrals

$$\begin{aligned}
\mathrm{s}_{j,m,n} &\approx \int\int \phi_{j,m,n}(x,y)F(x,y)dxdy,\\
\mathrm{d}^v_{j,m,n} &\approx \int\int \psi^v_{j,m,n}(x,y)F(x,y)dxdy,\\
\mathrm{d}^h_{j,m,n} &\approx \int\int \psi^h_{j,m,n}(x,y)F(x,y)dxdy,\\
\mathrm{d}^d_{j,m,n} &\approx \int\int \psi^d_{j,m,n}(x,y)F(x,y)dxdy,
\end{aligned}$$

Figure 10.1 shows the main idea of the computation of the wavelet transform of a 2D image. In one dimension, we scale the coefficients with $\frac{1}{\sqrt{2}}$, which is shown in Chapter 3. In this chapter, the normalization factors become integer powers of two. The procedure of computing the wavelet transform of a 2D image can be summarised as follows

- To calculate the wavelet coefficients at level $J-1$
  - Let $\mathbf{P}$ denote a 4×4 matrix, given by

$$\mathbf{P} = \frac{1}{2}\begin{pmatrix} 1 & -1 & 1 & -1\\ 1 & -1 & -1 & 1\\ 1 & 1 & -1 & -1\\ 1 & 1 & 1 & 1 \end{pmatrix}.$$

  - Let $\mathbf{v}_{J-1,1}$ denotes a vector with a set of elements $\mathrm{a}_{1,1}, \mathrm{a}_{2,1}, \mathrm{a}_{1,2}, \mathrm{a}_{2,2}$, from the original image and then the wavelet coefficients $\mathrm{d}^h_{J-1,1,1}, \mathrm{d}^d_{J-1,1,1}, \mathrm{d}^v_{J-1,1,1}$ and $\mathrm{s}_{J-1,1,1}$, are given by

$$\{\mathrm{d}^h_{J-1,1,1}, \mathrm{d}^d_{J-1,1,1}, \mathrm{d}^v_{J-1,1,1}, \mathrm{s}_{J-1,1,1}\} = \mathbf{P}\mathbf{v}_{J-1,1}.$$

  - Next take $\mathbf{v}_{1,2} = \{\mathrm{a}_{1,3}, \mathrm{a}_{2,3}, \mathrm{a}_{1,4}, \mathrm{a}_{2,4}\}$ taken from the original image to compute the wavelet coefficients $\mathrm{d}^h_{J-1,1,2}, \mathrm{d}^d_{J-1,1,2}, \mathrm{d}^v_{J-1,1,2}$ and $\mathrm{s}_{J-1,1,2}$, by

$$\{\mathrm{d}^h_{J-1,1,2}, \mathrm{d}^d_{J-1,1,2}, \mathrm{d}^v_{J-1,1,2}, \mathrm{s}_{J-1,1,2}\} = \mathbf{P}\mathbf{v}_{J-1,2}.$$

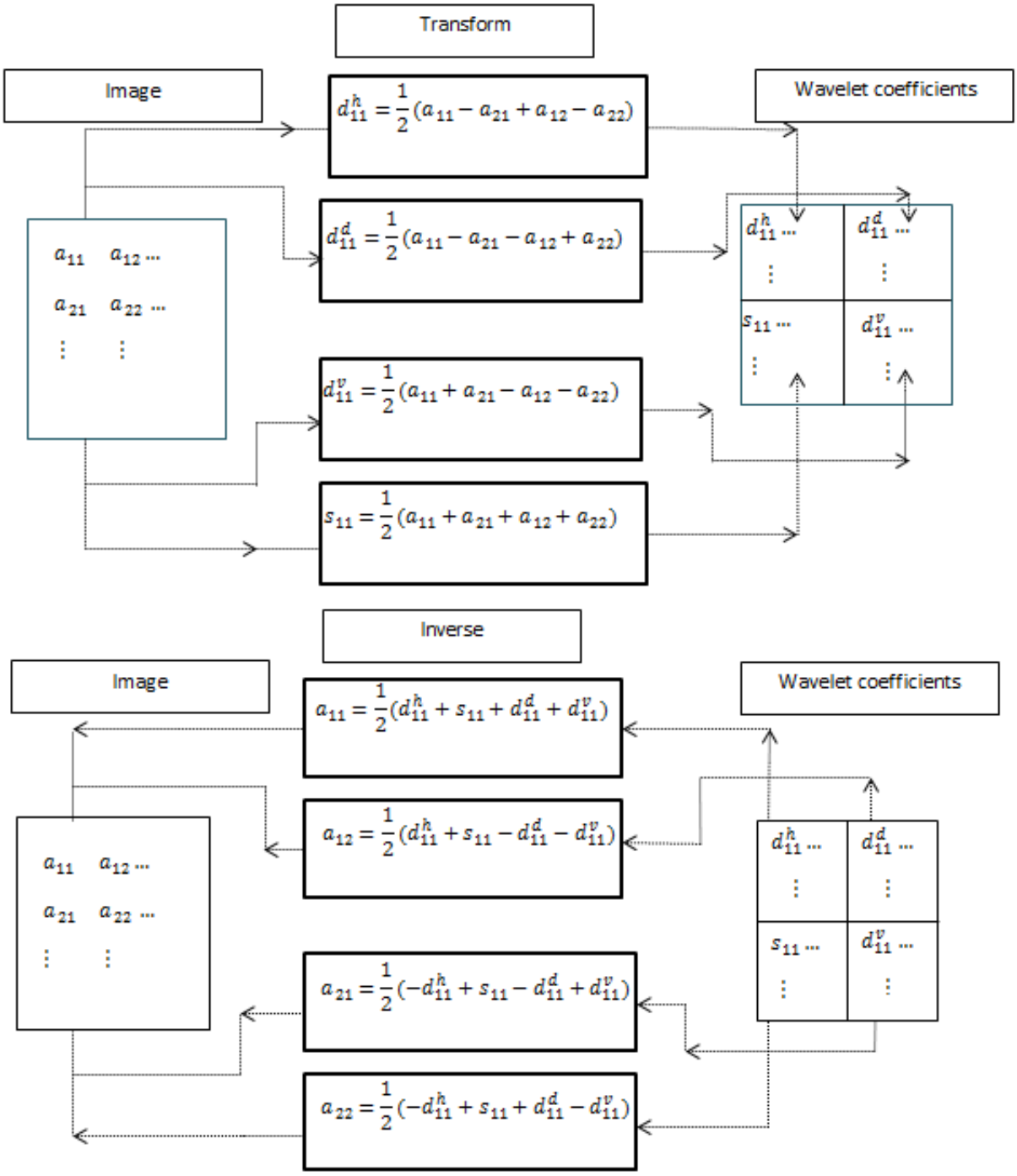


Figure 10.1: Diagram of the 2D Haar wavelet transform: a_{11}, a_{12}, a_{21} and a_{22} are a set of elements taken from the original image, and d_{11}^h, d_{11}^d, d_{11}^v and s_{11} are the wavelet coefficients at level $J-1$.

To calculate the wavelet coefficients at level $J-2$, the same procedure is used for the average of wavelet coefficients. For example, let $\mathbf{v}_{J-2,1} = \{\mathrm{s}_{J-1,1,3}, \mathrm{s}_{J-1,2,3}, \mathrm{s}_{J-1,1,4}, \mathrm{s}_{J-1,2,4}\}$ from the original image to compute the wavelet coefficients $\mathrm{d}^h_{J-2,1,2}, \mathrm{d}^d_{J-2,1,2}, \mathrm{d}^v_{J-2,1,2}$ and $\mathrm{s}_{J-2,1,2}$, by

$$\{\mathrm{d}^h_{J-2,1,2}, \mathrm{d}^d_{J-2,1,2}, \mathrm{d}^v_{J-2,1,2}, \mathrm{s}_{J-2,1,2}\} = \mathbf{P}\mathbf{v}_{J-2,1}.$$

Finally, let the wavelet coefficients $\mathbf{D}_{J-1}$ at resolution level $J-1$, be defined as

$$\mathbf{D}_{J-1} = \begin{pmatrix} \mathrm{d}^h_{1,1} & \mathrm{d}^h_{1,2} & \cdots & \mathrm{d}^h_{1,(\frac{M}{2}-1)} & \mathrm{d}^h_{1,\frac{M}{2}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathrm{d}^h_{\frac{M}{2},1} & \mathrm{d}^h_{\frac{M}{2},2} & \cdots & \mathrm{d}^h_{\frac{M}{2},(\frac{M}{2}-1)} & \mathrm{d}^h_{\frac{M}{2},\frac{M}{2}} \\ \mathrm{d}^d_{1,(\frac{M}{2}+1)} & \mathrm{d}^d_{1,(\frac{M}{2}+2)} & \cdots & \mathrm{d}^d_{1,(M-1)} & \mathrm{d}^d_{1,M} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathrm{d}^d_{\frac{M}{2},(\frac{M}{2}+1)} & \mathrm{d}^d_{\frac{M}{2},(\frac{M}{2}+2)} & \cdots & \mathrm{d}^d_{\frac{M}{2},(M-1)} & \mathrm{d}^d_{\frac{M}{2},M} \\ \mathrm{d}^v_{(\frac{M}{2}+1),(\frac{M}{2}+1)} & \mathrm{d}^v_{(\frac{M}{2}+1),(\frac{M}{2}+2)} & \cdots & \mathrm{d}^v_{(\frac{M}{2}+1),(M-1)} & \mathrm{d}^v_{1,M} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathrm{d}^v_{M,(\frac{M}{2}+1)} & \mathrm{d}^v_{M,(\frac{M}{2}+2)} & \cdots & \mathrm{d}^v_{M,(M-1)} & \mathrm{d}^v_{M,M} \end{pmatrix}, \tag{10.6}$$

where $N = M$ and $\mathbf{D}_{J-1}$ contains all the detail wavelet coefficients at level resolution $J-1$. Similarly,

$$\mathbf{D}_{J-2} = \begin{pmatrix} \mathrm{d}^h_{(\frac{M}{2}+1),1} & \mathrm{d}^h_{(\frac{M}{2}+1),2} & \cdots & \mathrm{d}^h_{(\frac{M}{2}+1),(\frac{M}{4}-1)} & \mathrm{d}^h_{(\frac{M}{2}+1),\frac{M}{4}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathrm{d}^h_{\frac{3M}{4},1} & \mathrm{d}^h_{\frac{3M}{4},2} & \cdots & \mathrm{d}^h_{\frac{3M}{4},(\frac{M}{4}-1)} & \mathrm{d}^h_{\frac{3M}{4},\frac{M}{4}} \\ \mathrm{d}^d_{(\frac{M}{2}+1),(\frac{M}{4}+1)} & \mathrm{d}^d_{(\frac{M}{2}+1),(\frac{M}{4}+2)} & \cdots & \mathrm{d}^d_{(\frac{M}{2}+1),(\frac{M}{2}-1)} & \mathrm{d}^d_{(\frac{M}{2}+1),\frac{M}{2}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathrm{d}^d_{\frac{3M}{4},(\frac{M}{4}+1)} & \mathrm{d}^d_{\frac{3M}{4},(\frac{M}{4}+2)} & \cdots & \mathrm{d}^d_{\frac{3M}{4},(\frac{M}{2}-1)} & \mathrm{d}^d_{\frac{3M}{4},\frac{M}{2}} \\ \mathrm{d}^v_{(\frac{3M}{4}+1),(\frac{M}{4}+1)} & \mathrm{d}^v_{(\frac{3M}{4}+1),(\frac{M}{4}+2)} & \cdots & \mathrm{d}^v_{(\frac{3M}{4}+1),(\frac{M}{2}-1)} & d^v_{(\frac{3M}{4}+1),\frac{M}{2}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathrm{d}^v_{M,(\frac{3M}{4}+1)} & d^v_{M,(\frac{3M}{4}+2)} & \cdots & \mathrm{d}^v_{M,(\frac{M}{2}-1)} & \mathrm{d}^v_{M,\frac{M}{2}} \end{pmatrix}. \tag{10.7}$$

Finally, $\mathbf{D}_0$ contains three elements $\mathrm{d}^h_{M-1,1}$, $\mathrm{d}^v_{M,2}$ and $\mathrm{d}^d_{M-1,2}$ at level 1. Figure 10.2 illustrates the multi-resolution scheme with several levels of wavelet transform.

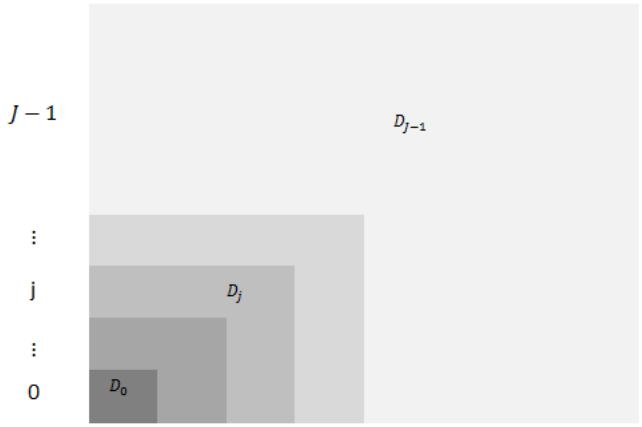


Figure 10.2: Diagram of the multi-resolution scheme with several levels of wavelet transform.

10.4 Estimation of σ^2

Nason (2010a) estimated the noise level in one dimension using the finest-scale wavelet coefficients. Now, the noise level in two dimensions, using the finest-scale wavelet coefficients, can be defined as

$$\widehat{\sigma} = sd[\mathbf{D}_{J-1}] = \sqrt{\frac{1}{\frac{3M^2-4}{4}} \sum_{M} \sum_{M} [d_{ij} - \bar{d}]^2}, \tag{10.8}$$

$\mathbf{D}_{J-1}$ is defined in (10.6), with elements d_{ij} and $\bar{d}$ is the mean of the wavelet coefficients in $\mathbf{D}_{J-1}$. This form suggests that the noise level σ is computed using the sample standard deviation of the finest-scale of wavelet coefficients, for more information see Nason (2010a).

As an alternative, it is assumed that the variance of the noise is unknown and is modelled by an inverse gamma distribution with parameters, a_0 and b_0, that is $\sigma^2 \sim \mathbf{inverse-gamma}(a_0, b_0)$, with density

$$\mathrm{p}(\sigma^2) = \frac{b_0^{a_0}}{\Gamma(a_0)} \sigma^{2(-a_0-1)} \exp\left\{ -\frac{b_0}{\sigma^2} \right\}, \quad \sigma^2 \geq 0; a_0, b_0 > 0. \tag{10.9}$$

This approach follows that of Gelman (2006) and Cutillo *et al.* (2008) for modelling a variance parameter. The parameters a_0 and b_0 can be fixed based on knowledge or

information from separate calibration experiments. In particular, an expert might provide a mean, σ_0^2, and variance, τ_0^2, for σ^2, that correspond to $a_0 = \frac{\sigma_0^{2\,2}}{\tau_0^2} + 2$ and $b_0 = \sigma_0^2(a_0 - 1)$. Although this approach is general, in this thesis a value for σ_0^2 computed from Equation (10.8) and $\tau_0^2 = 1$ has been used.

10.5 Point spread function

The following is a summary of the derivation and explanation that appears in Aykroyd *et al.* (2001). Consider the effect of a small rectangular block in the earth's magnetic field called an anomaly in archaeology. Suppose that the opposite corners of the block are situated at points with co-ordinates (x_1, y_1, z_1) and (x_2, y_2, z_2), where the x-axis points north, y-axis east and z-axis points vertically downwards. Then the vertical component of the anomaly due to the block, expressed in SI units, at a point with co-ordinates (x, y, z) is

$$\triangle Z(x,y,z) = \frac{\mu_0 \mathcal{K}\mathcal{H}}{4\pi}\left\{ [\triangle Z^{(1)} - \triangle Z^{(2)} - \triangle Z^{(3)}]_{\zeta=z-z_2}^{\zeta=z-z_1} \right\}$$

where μ_0 is the magnetic permeability of a vacuum ($4\pi \times 10^{-7}$ Henrys per metre), $\mathcal{K}$ is the susceptibility of the prism and $\mathcal{H}$ is the magnitude of the earth's magnetic field. The three separate contributions are

$$\begin{aligned}
\triangle Z^{(1)} &= \left[-\sin I \tan^{-1}\left(\frac{\zeta\eta}{\zeta(\xi^2+\eta^2+\zeta^2)^{1/2}} \right) \right]_{\xi=x-x_2,\eta=y-y_2}^{\xi=x-x_1,\eta=y-y_1} \\
\triangle Z^{(2)} &= \left[\frac{1}{2}\cos I \cos\theta \mathrm{In}\left(\frac{(\xi^2+\eta^2+\zeta^2)^{1/2}+\eta}{(\xi^2+\eta^2+\zeta^2)^{1/2}-\eta} \right) \right]_{\xi=x-x_2,\eta=y-y_2}^{\xi=x-x_1,\eta=y-y_1} \\
\triangle Z^{(3)} &= \left[\frac{1}{2}\cos I \sin\theta \mathrm{In}\left(\frac{(\xi^2+\eta^2+\zeta^2)^{1/2}+\xi}{(\xi^2+\eta^2+\zeta^2)^{1/2}-\xi} \right) \right]_{\xi=x-x_2,\eta=y-y_2}^{\xi=x-x_1,\eta=y-y_1}
\end{aligned}$$

where I is the inclination of the earth's magnetic field and θ is the angle between the direction of magnetic north and the x-axis of the survey co-ordinate system.

Since the anomaly is smaller by several orders of magnitude than the local magnetic field, it is standard practice to record the difference between simultaneous readings from two sensors. Usually, one sensor is mounted vertically above the other, typically at a distance

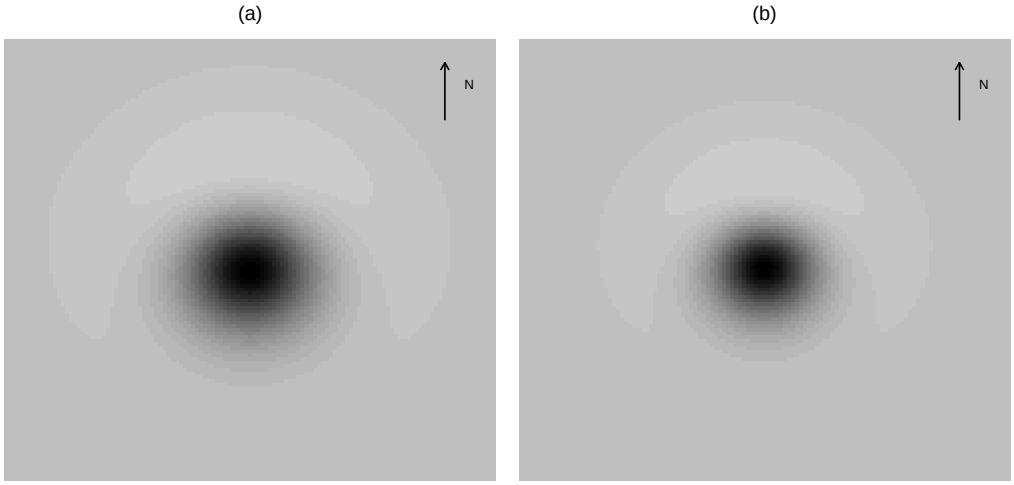


Figure 10.3: Plots of the spread function for depths: (a) 1 m and (b) 0.5 m, below the ground surface for a block with a vertical extent of 0.5 m.

of one half-metre or one metre. A magnetometer with two sensors used in this manner is known as a gradiometer. The recorded reading due to the block is then:

$$h(x, y) = \triangle Z(x, y, z_A) - \triangle Z(x, y, z_B) \tag{10.10}$$

where z_A is the vertical co-ordinate of the upper sensor and z_B is that of the lower sensor. Both vertical heights are normally held constant throughout the survey.

Clearly, the form of magnetic anomaly defined in Equation (10.10) is considerably more complicated than the point spread function usually encountered in imaging applications. Its precise form depends on the latitude and longitude of the archaeological site on the earth's surface, on the geometry of the gradiometer, and on the physical properties of the sensors. The spread function, or magnetic anomaly, for a single 1 m^3 prism with a susceptibility of 10^{-3}SI, buried 1 m below the surface at a typical location in the British Isles is shown in Figure 10.3. Each plot has longitudinal symmetry and the distance of the peak from the centre increases as the depth increases. The value and the shape of the point spread function depends on the values of its parameters. Using different parameter value leads to a different point spread function.

Figure 10.4 shows curves representing sections in north-south direction across the centre

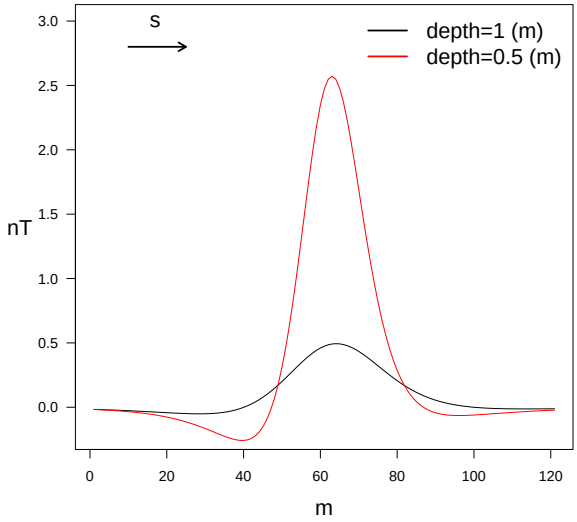


Figure 10.4: Plots of cross-sections through the centre for different depths in the north-south direction.

for different depths. Also, it shows an inverse relationship between the curve peak and depth. In the north-south direction a larger proportion of the curve is positive compared to the cross-section in the east-west direction, which is always positive. Furthermore, the curve is symmetrical in the east-west direction.

10.6 Archaeological images

Suppose that readings are recorded across a site at co-ordinates (as, ar), where $s = 1, 2, \ldots, M$, $r = 1, 2, \ldots, N$ and a is the sampling interval. Also, suppose that the subsurface features are divided into a rectangular grid of blocks, and that $f_{s',r'}$ is the susceptibility of the block whose centre is located at co-ordinate $(b_{s'}, b_{r'})$, where $s' = 1, 2, \ldots, M'$, $r' = 1, 2, \ldots, N'$ and b is the width of each block. Then, the influence of the block (s', r') at location (s, r), is

$$h_{s'r',sr} = h(as - b_{s'}, ar - b_{r'}),$$

where $h(x, y)$ is the spread function defined in Equation 10.10, with the edges of the block aligned to co-ordinate axes. The observed value $y_{s,r}$ at location (s, r) is the superposition of the influences of all the prisms, leading to the model

$$E[y_{s,r}] = \sum_{s'=1}^{M'} \sum_{r'=1}^{N'} h_{s'r',sr} f_{s',r'}, \quad s = 1, 2, \ldots, M, r = 1, 2, \ldots, N, \tag{10.11}$$

where $h_{s'r',sr}$ is the point spread function for the susceptibility due to a block centered at (s', r'), observed on the site surface at location (s, r). Then, these are corrupted by Gaussian error with mean 0 and variance σ^2. Consequently, the observations can be described by a Gaussian distribution with mean $E[y_{s,r}]$ and variance σ^2 (Allum, 1997; Aykroyd and Al-Gezeri, 2014).

10.7 Bayesian modelling of images

In Chapter 8 and 9, the proposed methods have been studied, a generalisation of the original one dimensional approach to a two dimensional image will now be considered.

Consider $\mathbf{F} = \{f_{s',r'} : s' = 1, 2, \ldots, M', r' = 1, 2, \ldots, N'\}$ that is a square matrix of values of some unknown function, and that $\mathbf{Y} = \{y_{s,r} : s = 1, 2, \ldots, M, r = 1, 2, \ldots, N, \}$ is a matrix of values of observed data in two dimensions. Also, let $\mathbf{G} = \mathbf{HF} = \{g_{s,r} : s = 1, 2, 3, \ldots, M, r = 1, 2, \ldots, N\}$, where $\mathbf{H}$ is the point spread function. If necessary, extra columns and rows of zero can be added so that M and N are powers of 2, to allow the following wavelet-based approach. Let $\mathbf{W}$ be an orthogonal matrix holding an appropriate discrete wavelet basis. The wavelet decomposition of the data $\mathbf{Y}$, $\mathbf{F}$ and $\mathbf{G}$, can be written as

$$\mathbf{D_Y} = \mathbf{WY}, \quad \mathbf{D_F} = \mathbf{WF}, \quad \mathbf{D_G} = \mathbf{WHF},$$

where $\mathbf{D_Y}$, $\mathbf{D_F}$ and $\mathbf{D_G}$ are matrices of the wavelet coefficients of $\mathbf{Y}$, $\mathbf{F}$ and $\mathbf{G}$ respectively.

The corresponding form of the likelihood, using wavelet coefficients, is given by

$$\mathrm{p}(\mathbf{Y}|\mathbf{D_F}) = \frac{1}{(2\pi\sigma^2)^{M^2/2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{s}^{M}\sum_{r}^{M} (y_{s,r} - g_{s,r})^2 \right\}, \quad y_{s,r}, g_{s,r} \in \mathbb{R}, ; \sigma^2 > 0, \tag{10.12}$$

where $\mathbf{G} = \mathbf{H}\mathbf{W}^T\mathbf{D_F} = \{g_{s,r} : s, r = 1, 2, \ldots, M\}$.

A single component prior for wavelet coefficients

Three choices for the prior distribution $\mathrm{p}(\mathbf{D_F})$ on the wavelet coefficients $\mathbf{D_F}$, will be considered. The most usual prior in the literature is the Gaussian distribution, with density

$$\mathrm{p}(\mathbf{D_F}|\kappa) = \left(\sqrt{\frac{\kappa}{\pi}}\right)^{M^2} \exp\left\{ -\kappa \sum_{j}^{J-1}\sum_{s'}^{M}\sum_{r'}^{M} d_{F\,j,s',r'}^2 \right\}, \quad d_{F\,j,s',r'} \in \mathbb{R}; \kappa > 0. \tag{10.13}$$

This approach follows that of Gribble (2001) for the modelling of wavelet coefficients.

The second choice for the prior distribution $\mathrm{p}(\mathbf{D_F})$ on the wavelet coefficients $\mathbf{D_F}$, that might be a better choice is the Laplace distribution, with density

$$\mathrm{p}(\mathbf{D_F}|\kappa) = \left(\frac{\kappa}{2}\right)^{M^2} \exp\left\{ -\kappa \sum_{j}^{J-1}\sum_{s'}^{M}\sum_{r'}^{M} |d_{F\,j,s',r'}| \right\}, \quad d_{F\,j,s',r'} \in \mathbb{R}; \kappa > 0, \tag{10.14}$$

where $|.|$ is the absolute value. This approach follows that of Vidakovic and Ruggeri (2001) and others for the modelling of wavelet coefficients.

The third choice is the elastic-net (Hastie *et al.*, 2009; Zou and Hastie, 2005), which is a compromise between these two prior distributions. The corresponding prior can be written as

$$\mathrm{p}(\mathbf{D_F}|\kappa,\gamma) = \left(\frac{1}{Z(\kappa,\gamma)}\right)^{M^2} \exp\left\{ -\kappa \sum_{j}^{J-1}\sum_{s'}^{M}\sum_{r'}^{M} (\gamma d_{F\,j,s',r'}^2 + (1-\gamma)|d_{F\,j,s',r'}|) \right\},$$
$$d_{F\,j,s',r'} \in \mathbb{R}; \kappa > 0, 0 < \gamma < 1, \tag{10.15}$$

where

$$Z(\kappa,\gamma) = \begin{cases} 2/\kappa, & \gamma = 0 \\ \sqrt{\frac{4\pi}{\kappa\gamma}} \exp\left\{\frac{\kappa(1-\gamma)^2}{4\gamma}\right\}\left(1 - \Phi(\frac{\kappa(1-\gamma)}{\sqrt{2\kappa\gamma}})\right), & 0 < \gamma < 1 \\ \sqrt{\pi/\kappa}, & \gamma = 1. \end{cases} \tag{10.16}$$

Note that each of these priors has introduced additional parameters, κ or κ and γ, which will also be modelled. In all cases, the parameter κ will also be modelled by a gamma distribution, $\kappa \sim \mathbf{gamma}(a_1, b_1)$, with density

$$\mathrm{p}(\kappa) = \frac{1}{\Gamma(a_1)} b_1^{a_1} \kappa^{(a_1-1)} \exp\left\{ - b_1\kappa \right\}, \quad \kappa \geq 0; a_1, b_1 > 0, \tag{10.17}$$

As with σ^2, the parameters a_1 and b_1 can also be fixed based on knowledge or information from separate calibration experiments. In Chapter 8, examples were given of one approach to the choice of prior parameter value for the various prior distributions.

Finally, the parameter γ takes value within the range $[0, 1]$, hence a beta distribution is a sensible choice of prior model, $\gamma \sim \mathbf{Beta}(a_2, b_2)$, with density

$$\mathrm{p}(\gamma) = \frac{\Gamma(a_2 + b_2)}{\Gamma(a_2)\Gamma(b_2)} \gamma^{a_2-1}(1-\gamma)^{b_2-1}, \quad 0 < \gamma < 1; a_2, b_2 > 0. \tag{10.18}$$

As with κ, the parameters a_2 and b_2 can also be fixed, based on knowledge or information from separate calibration experiments. In Chapter 8, examples were given one approach to the choice of prior parameter value for the various prior distributions.

Multiple prior models for wavelet coefficients

As an extension, the parameters κ and γ are to be grouped by wavelet resolution level with the obvious extensions to the definitions given in the previous section. For the wavelet coefficients at level j, $\mathbf{D}_{\mathbf{F}j}$, the Gaussian prior density function becomes

$$\mathrm{p}(\mathbf{D}_{\mathbf{F}j}|\kappa_j) = \left(\sqrt{\frac{\kappa_j}{\pi}}\right)^{3(2^{2j})} \exp\left\{ - \kappa_j \sum_{s'}^{2^j} \sum_{r'}^{2^j} d_{F\,j,s',r'}^2 \right\}, \quad d_{F\,j,s',r'} \in \mathbb{R}; \kappa_j > 0, \tag{10.19}$$

where, $\boldsymbol{\kappa} = \{\kappa_j : j = 0, 1, \ldots, J-1\}$, with $J = \log_2(m)$ and d_{Fj} are the level j wavelet coefficients. The power 3 appears because there are three sets of wavelet coefficients for horizontal, vertical and diagonal directions. The Laplace prior density function becomes

$$\mathrm{p}(\mathbf{D}_{\mathbf{F}j}|\kappa_j) = \left(\frac{\kappa_j}{2}\right)^{3(2^{2j})} \exp\left\{ -\kappa_j \sum_{s'}^{2^j}\sum_{r'}^{2^j} |d_{Fj,s',r'}| \right\}, \quad d_{Fj,s',r'} \in \mathbb{R}; \kappa_j > 0, \tag{10.20}$$

and finally, the elastic-net based model, with density

$$\mathrm{p}(\mathbf{D}_{\mathbf{F}j}|\kappa_j) = \left(\frac{1}{Z(\kappa_j,\gamma_j)}\right)^{3(2^{2j})} \exp\left\{ -\kappa_j \sum_{s'}^{2^j}\sum_{r'}^{2^j} (\gamma_j d_{Fj,s',r'}^2 + (1-\gamma_j)|d_{Fj,s',r'}|) \right\},$$
$$d_{Fj,s',r'} \in \mathbb{R}; \kappa_j > 0, 0 < \gamma_j < 1, \tag{10.21}$$

where,

$$Z(\kappa_j,\gamma_j) = \begin{cases} 2/\kappa_j, & \gamma_j = 0 \\ \sqrt{\frac{4\pi}{\kappa_j\gamma_j}} \exp\left\{\frac{\kappa_j(1-\gamma_j)^2}{4\gamma_j}\right\}\left(1 - \Phi(\frac{\kappa_j(1-\gamma_j)}{\sqrt{2\kappa_j\gamma_j}})\right), & 0 < \gamma_j < 1 \\ \sqrt{\pi/\kappa_j}, & \gamma_j = 1, \end{cases} \tag{10.22}$$

for $j = 0, 1, \ldots, J-1$ with $J = \log_2(m)$. The prior densities become

$$\mathrm{p}(\kappa_j) = \frac{1}{\Gamma(a_1)} b_1^{a_1} (\kappa_j)^{(a_1+1)} \exp\left\{ -b_1\kappa_j \right\}, \quad \kappa_j \geq 0, j = 0, 1, \ldots, J-1; a_1, b_1 > 0, \tag{10.23}$$

and

$$\mathrm{p}(\gamma_j) = \frac{\Gamma(a_2+b_2)}{\Gamma(a_2)\Gamma(b_2)} (\gamma_j)^{a_2-1}(1-\gamma_j)^{b_2-1}, \quad 0 < \gamma_j < 1, j = 0, 1, \ldots, J-1; a_2, b_2 > 0. \tag{10.24}$$

Similarly, the hyper parameters a_1, b_1, a_2, b_2, a_3 and b_3 can be fixed for all levels at the same values as chosen in the single component prior (see section 8.3.5).

10.8 Simulation experiments

The purpose of this section is to apply the proposed methods, introduced in Section 10.7, to simulated datasets. The benefit of using simulated datasets is that the properties and

features of artificial data are known. Then, the proposed methods are studied and assessed for their suitability. More precisely, the quality of the solution can be investigated and any error attributed to the inversion procedure and not mismodelling. In this chapter, PM and MAP estimation calculations similar to those in Algorithm 4 and 3, will be applied.

Two simulated datasets are chosen as representative examples. The calculated spread function models correspond to a fluxgate gradiometer with its sensors at a separation of 0.7 m and the lower sensor positioned 0.2 m above the ground. The simulated datasets have a uniform feature susceptibility of 2.6×10^{-3}**SI** and zero background susceptibility. Furthermore, the depth of the archaeological layer beneath the surface and the vertical extent are both fixed at 0.5 m. The magnetic flux density of the Earth's field is nT=48000. In the calculations, the blur function is truncated to produce a spread matrix of 19×19.

The additive Gaussian noise has a standard deviation, $\sigma = 1$. The true location of a block feature is shown in Figure 10.5 (a) displayed on a 10 m $\times$10 m grid with 1 m $\times$1 m blocks. The response of the magnetometer to this feature is calculated over the same area at 1 m intervals as shown in Figure 10.5 (b). The magnitudes of the true features in 10.5 (a) are equal to 2.6×10^{-3}**SI** and zero for background. Hence, the impact of spread function shifts the apparent location of the feature. This means that the feature in the simulated data seems to be located further south than it actually is and a slight negative anomaly occurs to its north side. However, there is no shift in the east-west direction because the spread function has longitudinal symmetry (Allum, 1997).

The MCMC approach through the Metropolis-Hastings algorithm can be applied to estimate the features from these simulated data. The wavelet coefficients of a true function is estimated with the Haar basis. The total run length is equal to $16 \times 16 \times 4000 = 1024000$. The observed image, $\mathbf{Y}$, is extended by zero elements to use the wavelet transform and the number of update iterations is equal to 4000. The parameters σ^2, κ and γ are described by the prior distributions, as discussed in Sections 10.6 and 10.7. Two estimators are used to estimate the true features of the susceptibility profile on a 10 m $\times$10 m reconstruction grid with 1 m spacing. One is based on the MAP estimate and the other is based on the

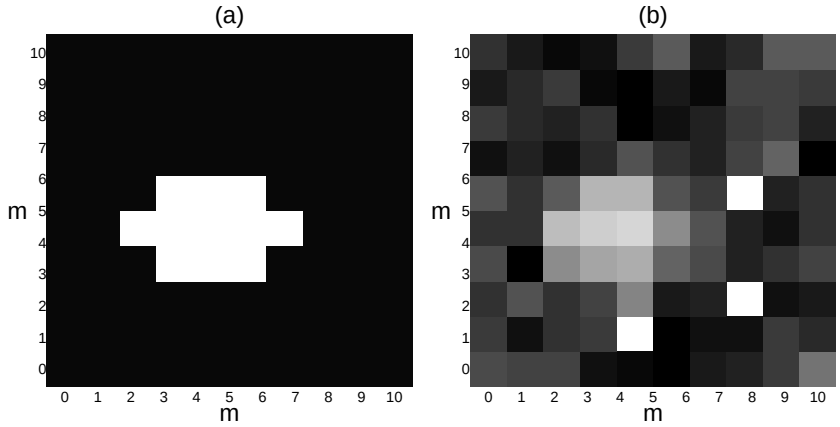


Figure 10.5: Plots of the box image: (a) true image on a 1 m ×1 m blocks; and (b) simulated data on a 1 m ×1 m blocks.

PM estimate.

The reconstruction of a single Laplace prior, using the MAP estimate is shown in Figure 10.6 (a) on a 10 m ×10 m reconstruction grid with 1 m spacing. The general shapes and edges are close to the true susceptibility surface. It is also of interest to see the reconstruction of the Laplace for a level-dependent prior. The MAP estimate is shown in Figure 10.6 (b) on a 10 m ×10 m reconstruction grid with 1 m spacing. It is clear that the features on the left-hand side of the reconstructions, of the single and multiple-component priors, are not fully resolved.

The reconstructions using the elastic-net prior for single and multiple prior parameters are shown in Figure 10.6 (c) and (d) on a 10 m ×10 m reconstruction grid with 1 m spacing. The reconstruction for a level-dependent prior is close to the true susceptibility surface.

The reconstructions with a single component Gaussian prior and a level-dependent prior are shown in Figures 10.6 (e) and 10.6 (f), respectively. The general of using the Laplace and the elastic-net priors, the shapes and edges of the reconstructions are much better than using the Gaussian prior. This is because the reconstructions show a slight variation in the background intensity. In general, the shape and location of the feature are very accurate. The edges of the feature have a well defined step for its entire boundary.

The MSE results using the MAP estimates are summarized in Table 10.1, where bold numbers indicate the smallest MSE result for each case. The definition of the mean squared-error in two dimensions is given by

$$\text{MSE} = \frac{1}{M^2}\sum_{s'}^{M}\sum_{r'}^{M}(f_{s'r'} - \widehat{f}_{s'r'})^2.$$

The single component Gaussian prior provides a smaller MSE than the Laplace and the elastic-net priors, whereas the elastic-net for the level-dependent model provides a smaller MSE than the Laplace and the Gaussian. The MAP parameters for single component and level-dependent prior are summarized in Tables 10.2 and 10.3. It is worth noting that the values of $\widehat{\kappa}$ for two dimensional data are larger than those for one dimensional data.

The reconstructions of the PM estimates are shown in Figure (10.7), on a 10 m $\times$10 m reconstruction grid with 1 m spacing. The reconstructions clearly show the edges and shapes, although it can be seen that there is a lack of contrast between the edges of the reconstructed feature and the background. Additionally, the reconstruction of Gaussian prior using the PM estimate shows that the background intensity is less variable.

The MSE results, using the PM estimates, are summarized in Table 10.4, where bold numbers indicate the smallest MSE result for each case. It can also be seen that the single component Laplace prior gives a smaller MSE than elastic-net and Gaussian priors, whereas the elastic-net for level-dependent prior provides a smaller MSE than the Laplace and the Gaussian priors.

Table 10.2 shows estimates of parameters using the elastic-net prior. The variance estimate is $\widehat{\sigma}^2 = 0.8518$, where the true variance is equal to 1. The value of $\widehat{\kappa} \approx 899$ and the value of $\widehat{\gamma} = 0.2106$. Table 10.3 shows the parameter estimates for the elastic-net level-dependent prior: $\widehat{\kappa}_3 \approx 960$, $\widehat{\kappa}_2 \approx 240$, $\widehat{\kappa}_1 = 240$, $\widehat{\kappa}_0 \approx 169$ and $\widehat{\kappa}_{s_0} = 120$. Also, $\widehat{\gamma}_3 = 0.0981$, $\widehat{\gamma}_2 = 0.3997$, $\widehat{\gamma}_1 = 0.0245$, $\widehat{\gamma}_0 = 0.0173$ and $\widehat{\gamma}_{s_0} = 0.0122$.

Another example is shown in Figure 10.8 (a), on a 20 m $\times$20 m reconstruction grid with 0.5 m spacing. The response of the magnetometer to this feature is shown in Figure 10.8 (b). The additive Gaussian noise has a standard deviation, $\sigma^2 = 1$, hence the data is very noisy and it is not easy to see the feature. The parameters σ^2, κ and γ are

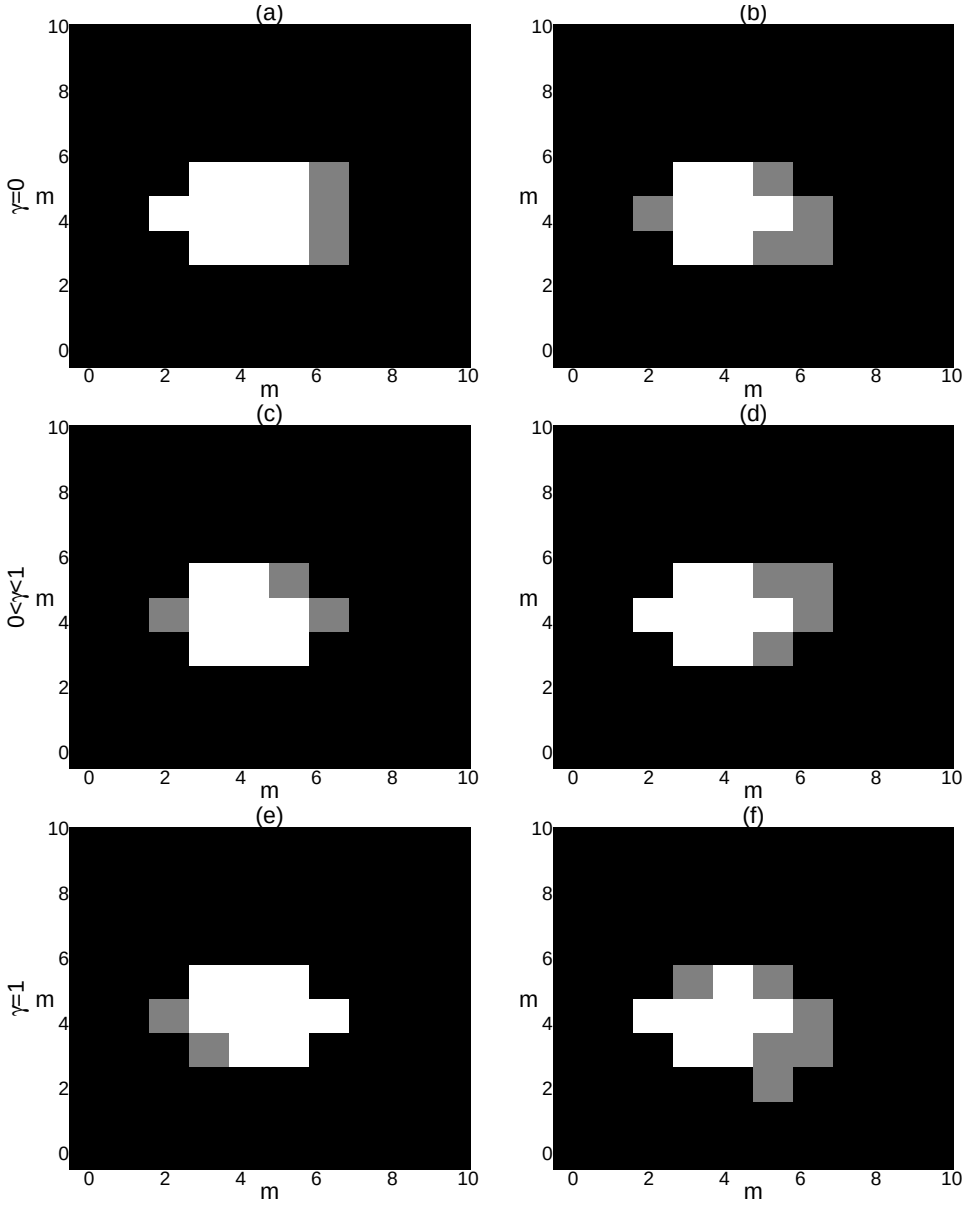


Figure 10.6: Plots of the MAP reconstruction for different priors: (a) single Laplace; (b) level-dependent Laplace; (c) single elastic-net; (d) level-dependent elastic-net; (e) single Gaussian; and (f) level-dependent Gaussian.

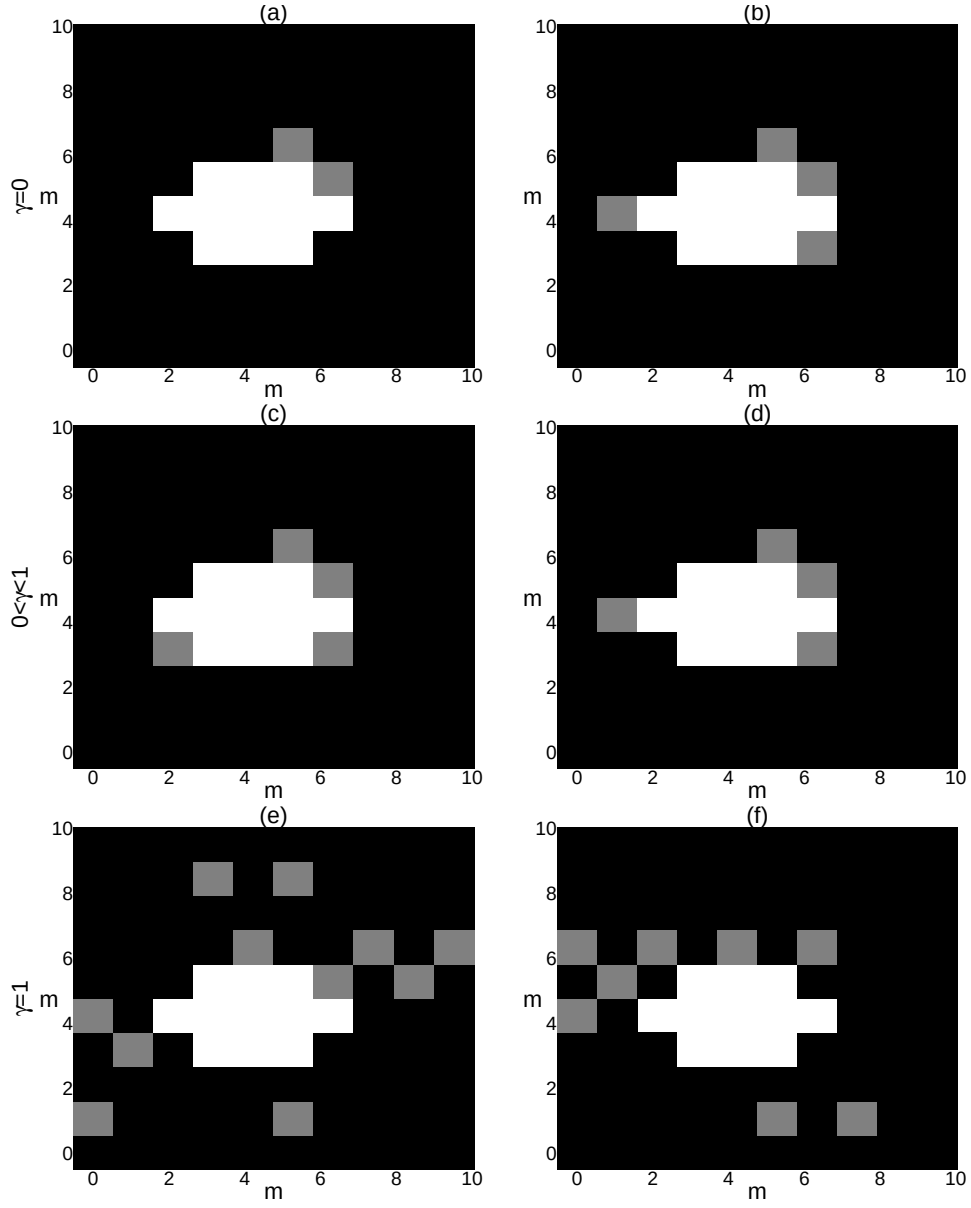


Figure 10.7: Plots of the PM reconstruction for different priors: (a) single Laplace; (b) level-dependent Laplace; (c) single elastic-net; (d) level-dependent elastic-net; (e) single Gaussian; and (f) level-dependent Gaussian.

Prior	MSE	
	Single component $\times 10^{-8}$	Level-dependent $\times 10^{-8}$
Laplace	3.07	3.12
Elastic-net	2.61	**3.08**
Gaussian	**2.28**	9.11

Table 10.1: MSE results using the MAP estimates for susceptibility. The bold font represents the smallest MSE.

Prior	$\widehat{\gamma}$	$\widehat{\kappa}$	$\widehat{\sigma}^2$
Laplace	0	831.90	0.8252
Elastic-net	0.2106	899.90	0.8518
Gaussian	1	883.90	0.8708

Table 10.2: MAP parameter estimates for single priors.

Prior	$\widehat{\gamma}$	$\widehat{\kappa}$	$\widehat{\sigma}^2$
Laplace	0	$(639.87)_3$ $(160.03)_2$ $(113.11)_1$ $(97.18)_0$ $(56.55)_{s_0}$	0.8719
Elastic-net	$(0.0981)_3$ $(0.3997)_2$ $(0.0245)_1$ $(0.0173)_0$ $(0.0122)_{s_0}$	$(960.01)_3$ $(240.02)_2$ $(240.00)_1$ $(169.70)_0$ $(120.00)_{s_0}$	0.8638
Gaussian	1	$(639.90)_3$ $(160.01)_2$ $(159.97)_1$ $(113.11)_0$ $(79.98)_{s_0}$	0.7621

Table 10.3: MAP parameter estimates for level-dependent priors.

Prior	MSE	
	Single component $\times 10^{-8}$	Level-dependent $\times 10^{-8}$
Laplace	**3.15**	3.78
Elastic-net	3.84	**3.77**
Gaussian	6.68	7.33

Table 10.4: MSE results using the PM estimates for susceptibility. The bold font represents the smallest MSE.

Prior	$\widehat{\gamma}$	$\widehat{\kappa}$	$\widehat{\sigma}^2$
Laplace	0	901.90	0.8782
Elastic-net	0.2196	979.90	0.7893
Gaussian	1	841.90	0.8843

Table 10.5: PM parameter estimates for single priors.

Prior	$\widehat{\gamma}$	$\widehat{\kappa}$	$\widehat{\sigma}^2$
Laplace	0	$(959.50)_3$ $(240.02)_2$ $(139.87)_1$ $(119.61)_0$ $(119.93)_{s_0}$	0.8455
Elastic-net	$(0.06705)_3$ $(0.5722)_2$ $(0.0467)_1$ $(0.0418)_0$ $(0.0080)_{s_0}$	$(1079.6)_3$ $(820)_2$ $(869.90)_1$ $(490.84)_0$ $(131.91)_{s_0}$	0.8644
Gaussian	1	$(960.01)_3$ $(240.05)_2$ $(240.01)_1$ $(169.70)_0$ $(90.00)_{s_0}$	0.8074

Table 10.6: PM parameter estimates for level-dependent priors.

estimated using the prior distributions. The reconstructions with MAP estimation are shown in Figure 10.9, it can also be seen that the location and general shape of the feature estimated well. However, the reconstructions with the Gaussian prior show that the background intensity is less variable, whereas the reconstructions with the Laplace prior shows that the edge between the feature and the background is not fully resolved. The general shape of the feature has been reproduced fairly accurately but the susceptibility is slightly underestimated over some of the reconstructions and a slight variation is also evident in the background intensity. Additionally, it can be seen that extra features appear in the reconstruction with the Gaussian prior, although the overall shape and location of the feature corresponds very well to the true feature. The total run length is equal to $32 \times 32 \times 5000 = 512 \times 10^4$ and the number of update iterations is equal to 5000.

The MSE results using the MAP estimates are summarized in Table 10.7, where bold numbers indicate the smallest MSE result for each case. It can be seen that the single component elastic-net prior provides a smaller MSE than the Laplace and the Gaussian for single component and level-dependent priors.

The PM estimates are shown in Figure 10.10. The location and general shape of the feature are well estimated, although the reconstructions with the Gaussian prior show that the background intensity is variable. The reconstructions with the single and the level-dependent for Laplace and the elastic-net priors show that the edges between the feature and the background are not fully resolved, although the general shape of the feature has been reproduced fairly accurately.

The MSE results using the PM estimates are summarized in Table 10.10, where bold numbers indicate the smallest MSE result for each case. It can be seen that the single elastic-net prior provides a smaller MSE than the Laplace and the Gaussian for single component and level-dependent prior Laplace provides a smaller MSE than elastic-net and the Gaussian priors.

The MAP and the PM parameter estimates for the reconstruction on a 20 m $\times$20 m reconstruction grid with 0.5 m spacing, using single component and level-dependent prior,

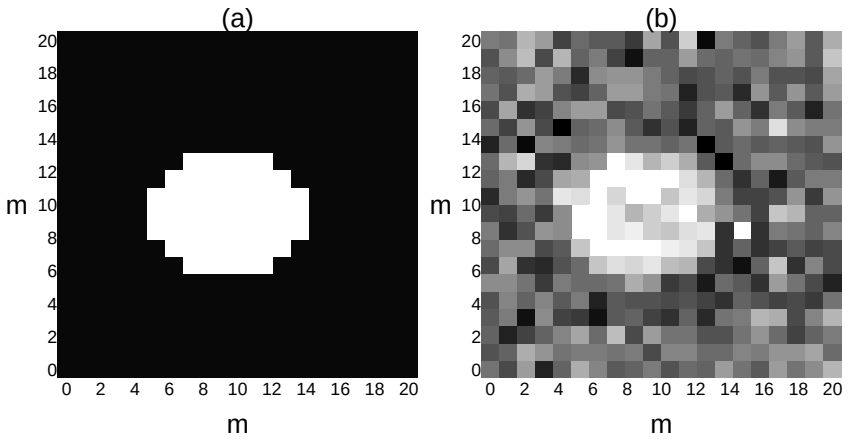


Figure 10.8: Plots of image: (a) true image; and (b) simulated data.

are summarized in Tables 10.8, 10.9, 10.11 and 10.12. The MAP estimates using the single elastic-net prior shows that the value of $\widehat{\kappa} \approx 960$ and $\widehat{\gamma} = 0.2106$. The value of the variance is $\widehat{\sigma}^2 = 0.6638$. For the level-dependent prior, it can be seen that the values of $\widehat{\kappa}_4 \approx 1238$, $\kappa_3 \approx 958$, $\kappa_2 \approx 240$, $\kappa_1 \approx 379$, $\kappa_0 \approx 218$ and $\kappa_{s_0} \approx 154$. The values of $\widehat{\gamma}_4 = 0.0532$, $\widehat{\gamma}_3 = 0.1438$, $\widehat{\gamma}_2 = 0.2566$, $\widehat{\gamma}_1 = 0.0133$, $\widehat{\gamma}_0 = 0.0094$ and $\widehat{\gamma}_{s_0} = 0.0065$.

Prior	MSE	
	Single component $\times 10^{-3}$	Level-dependent $\times 10^{-3}$
Laplace	5.40	**5.6**
Elastic-net	**5.10**	5.90
Gaussian	6.30	6.00

Table 10.7: MSE results MAP estimates for susceptibility. The bold font represents the smallest MSE.

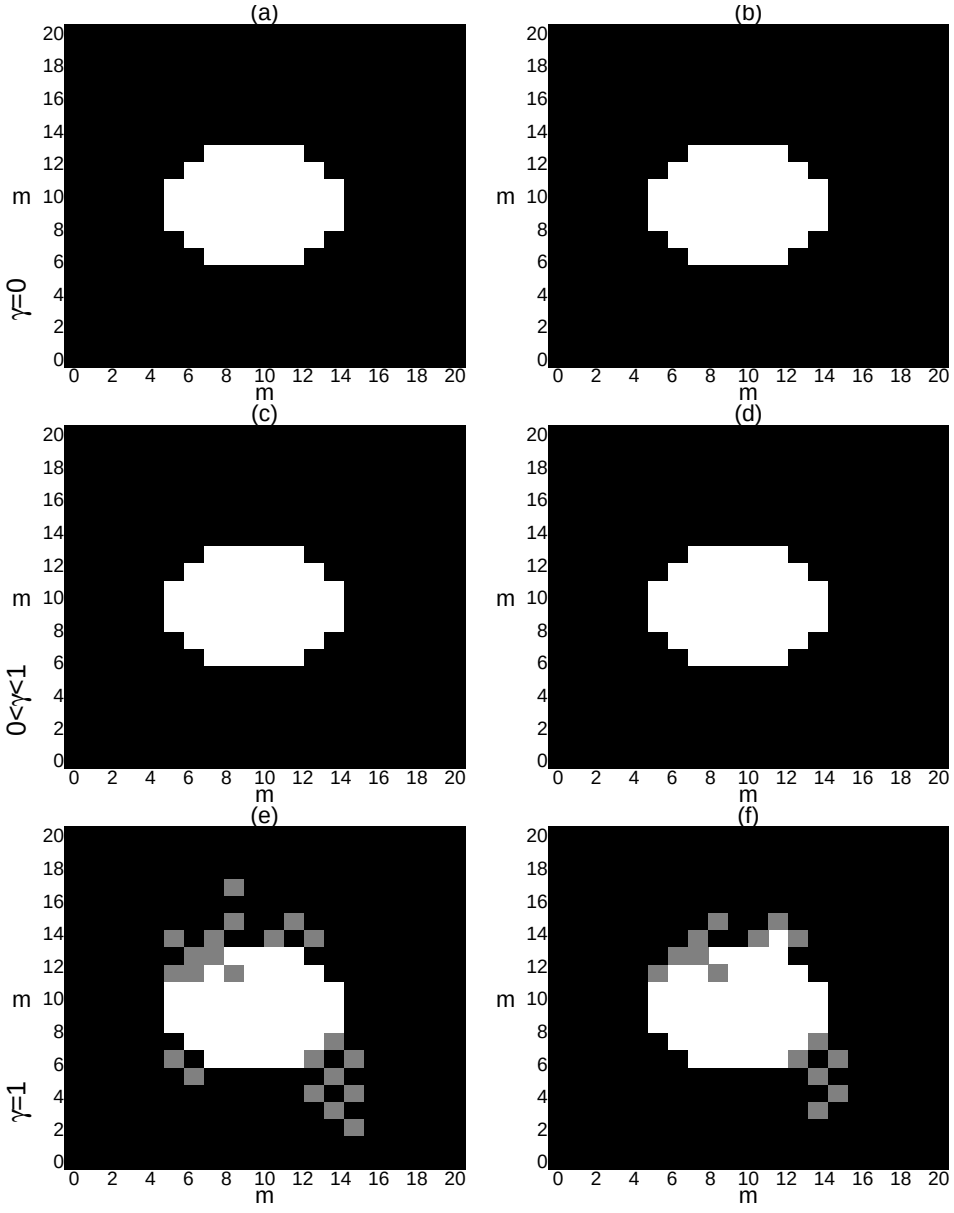


Figure 10.9: Plots of the MAP reconstruction using different prior models: (a) single Laplace; (b) level-dependent Laplace; (c) single elastic-net; (d) level-dependent elastic-net; (e) single Gaussian; and (f) level-dependent Gaussian.

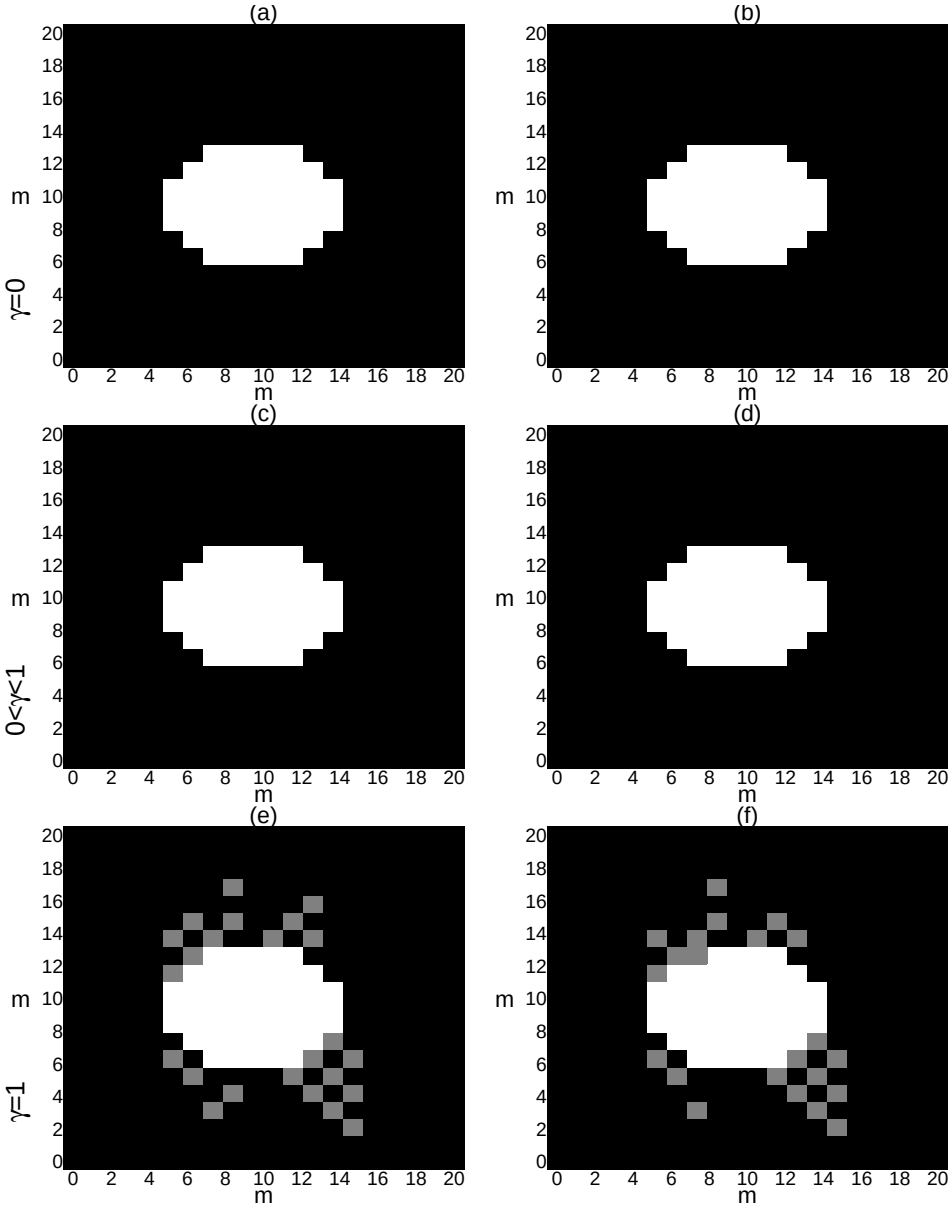


Figure 10.10: Plots of the PM reconstruction using different prior models: (a) single Laplace; (b) level-dependent Laplace; (c) single elastic-net; (d) level-dependent elastic-net; (e) single Gaussian; and (f) level-dependent Gaussian.

Prior	$\widehat{\gamma}$	$\widehat{\kappa}$	$\widehat{\sigma}$
Laplace	0	839.99	0.7725
Elastic-net	0.2106	960.00	0.6638
Gaussian	1	845.01	0.6318

Table 10.8: MAP parameter estimates for single priors.

Prior	$\widehat{\gamma}$	$\widehat{\kappa}$	$\widehat{\sigma}$
Laplace	0	$(1024.03)_4$ $(256.01)_3$ $(63.94)_2$ $(55.51)_1$ $(64.19)_0$ $(65.25)_{s_0}$	0.7222
Elastic-net	$(0.0532)_4$ $(0.1438)_3$ $(0.2566)_2$ $(0.0133)_1$ $(0.0094)_0$ $(0.0065)_{s_0}$	$(1238.81)_4$ $(958.21)_3$ $(240.13)_2$ $(379.70)_1$ $(218.99)_0$ $(154.85)_{s_0}$	0.6763
Gaussian	1	$(1312.00)_4$ $(927.99)_3$ $(231.99)_2$ $(228.01)_1$ $(131.93)_0$ $(164.81)_{s_0}$	0.7116

Table 10.9: MAP parameter estimates for level-dependent priors.

Prior	MSE	
	Single component $\times 10^{-8}$	Level-dependent $\times 10^{-8}$
Laplace	8.53	**22.00**
Elastic-net	**4.30**	38.00
Gaussian	37.00	35.00

Table 10.10: MSE results using the PM estimates for susceptibility. The bold font represents the smallest MSE.

Prior	$\widehat{\gamma}$	$\widehat{\kappa}$	$\widehat{\sigma}$
Laplace	0	839.98	0.7529
Elastic-net	0.2104	960.00	0.6353
Gaussian	1	959.99	0.5116

Table 10.11: PM parameter estimates for single priors.

Prior	$\widehat{\gamma}$	$\widehat{\kappa}$	$\widehat{\sigma}$
Laplace	0	$(841.12)_4$ $(676.07)_3$ $(209.66)_2$	0.7123
		$(269.01)_1$ $(119.51)_0$ $(84.50)_{s_0}$	
Elastic-net	$(0.0434)_4$ $(0.0289)_3$ $(0.1941)_2$	$(839.85)_4$ $(660.18)_3(239.66)_2$	0.7008
	$(0.0072)_1$ $(0.0051)_0$ $(0.0036)_{s_0}$	$(105.04)_1$ $(116.70)_0(82.52)_{s_0}$	
Gaussian	1	$(1312.00)_4$ $(928.00)_3(231.99)_2$	0.6799
		$(115.95)_1$ $(112.02)_0(82.28)_{s_0}$	

Table 10.12: PM parameter estimates for level-dependent priors.

In keeping with the conclusions of the previous experiments, the best reconstruction are obtained when the Laplace prior and the elastic-net prior are used. This is because the shape and location of the estimated feature corresponds very well to those of the true feature. However, it can be seen that the edges, when using the Laplace prior, are not resolved and the reconstruction of the elastic-net shows slight variation in the background intensity. In this case the Gaussian prior and level-dependent Laplace prior provide a good reconstruction.

10.9 Real data application

The estimation procedure is used to analyse the magnetometer data from the Park, Guiting Power as shown in Figure 10.11. In particular, this shows a diagonal linear ditch towards the top, and a rectangular boundary ditch surrounding various collections of

circular pits and post-holes. It is usually assumed that the features are buried at the same depth from the modern site surface, but that they have different susceptibilities and extents. The assumption of constant depth is not unreasonable, as ancient structures are often levelled to the prevailing ground level and any pits or ditches are in-filled to the same level. At later stages the whole site is uniformly covered with topsoil, leading to the common modern surface level. The strength of the surface magnetic readings depend on the depth of the features.

There are three different priors considered, elastic-net, Laplace and Gaussian. The estimated wavelet coefficients, using Haar wavelet, are shown in Figure 10.12. The wavelet coefficients are very sparse compared with the magnetometer data. Figure 10.12 (a) shows that most of the wavelet coefficients are close to zero and the range of the plot of the wavelet coefficients is between -2 and $6nT$. The darkness of the pixels correspond to the magnitude of the wavelet coefficients (Bruce and Gao, 1996a). It can be seen that the large wavelet coefficients tend to represent the important features in the image such as walls, pits and ditches.

Two estimators are applied to estimate the true features of the susceptibility profile on a 20 m $\times$20 m reconstruction grid with 0.5 m spacing. The calculated spread function models correspond to a fluxgate gradiometer with its sensors at a separation of 0.7 m and the lower sensor positioned 0.2 m above the ground. Further, the depth of the archaeological layer beneath the surface and the vertical extent are both fixed at 0.5 m. The magnetic flux density of the Earth's field is $nT = 48000$. The total run length is equal to $32 \times 32 \times 8000 = 8192 \times 10^3$ and the number of update iterations is equal to 8000. The MAP estimates of whole area are shown in Figures 10.13, 10.14 and 10.15.

The MAP estimates, using the single Laplace prior and the level-dependent prior, shown in Figure 10.13, reveal the main pits and ditches, although, some features have disappeared.

The MAP estimates using single elastic-net prior and level-dependent prior, shown in Figure 10.14, also reveal the main pits and ditches, although again, some features have disappeared.

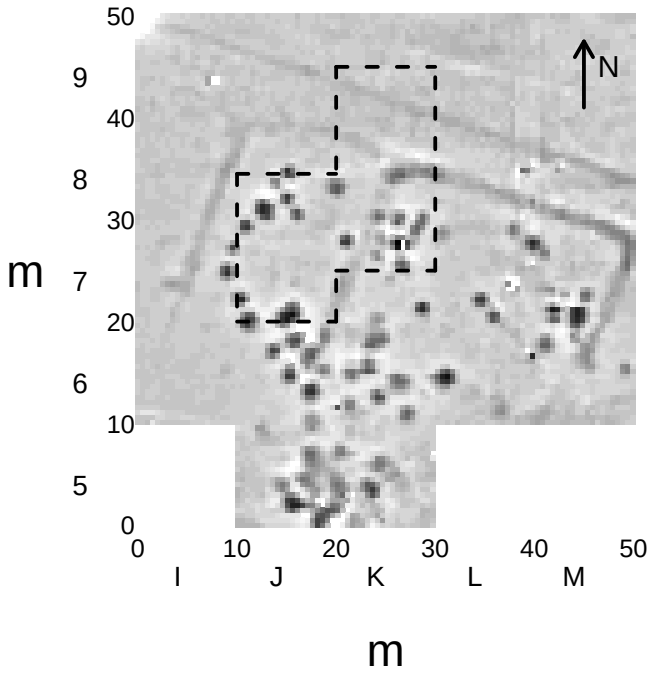


Figure 10.11: Data from "The Park", Guiting Power in Gloucestershire, collected using a fluxgate gradiometer, in 1994, at 0.5 m intervals, where I, J, K, L, M, 5, 6, 7, 8 and 9 are row and column labels to use as region references (Allum, 1997).

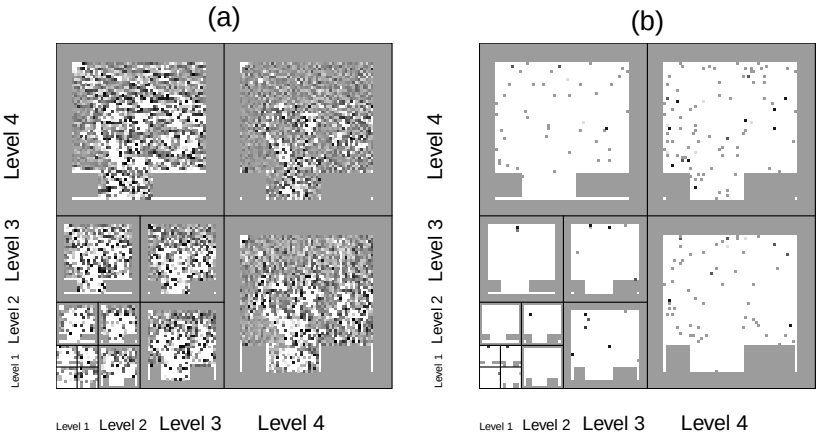


Figure 10.12: Plots of the wavelet transform with four resolution levels: plot of the wavelet coefficients in intervals (a) (-2, 3); and (b) (0.02, 0.03).

Feature Number	Feature Type	Vertical extent (m)	Average susceptibility ($\mathbf{SI} \times 10^{-3}$)	Description
1	pit	1.60	1.1	relatively low organic content
2	pit	1.50	1.0	relatively low organic content
3	pit	0.70	2.2	not overtly organic yet high sus
4	pit	1.05	2.4	enriched with organic material
5	pit	0.45	2.4	highly organic in lower half
6	pit	0.80	1.8	organic
7	pit	0.70	1.5	generally non-organic
8	pit	0.95	2.6	moderately organic
9	ditch	0.70	1.1	highly organic in upper third
10	ditch	1.00	1.3	highly organic in upper half
11	ditch	0.50	0.4	distinctly non-organic
12	gully	0.60	0.8	distinctly non-organic
13	ditch	0.70	1.5	highly organic in upper third
14	post hole	?	1.3	not recorded
15	post hole	?	0.8	not recorded
16	post hole	?	08	not recorded
17	post hole	?	0.7	not recorded

Table 10.13: Summary of Guiting Power excavation (Allum, 1997).

Data grid	J6	J7	J8	K6	K7	K6
$\widehat{\sigma}^2$	2.7377	2.4494	2.4905	2.6010	2.7493	1.4852

Table 10.14: Estimation of σ from different grids.

The MAP estimates, using the single Gaussian prior and the level-dependent prior, shown in Figure 10.15. The reconstructions show that the features are not fully defined. In general, the reconstruction shows variation in the background intensity and the excavation shows that the pits are distributed across 8L and 9L grids. The reconstruction of the Guiting Power shows that J5 and K5 regions are not clear, but more complex, see Figure 10.21. Also, there are some grids with high levels of magnetic susceptibility, such as J5, J7, J8, K5, K6, K7 and L7.

The PM estimates are shown in Figures 10.16, 10.17 and 10.18. The location and general shape of the features are estimated well, although the reconstructions with the Gaussian prior show that the background intensity is variable. The reconstructions with the single Laplace prior, the single and level-dependent elastic-net priors show that the features of susceptibility are clearly defined and the general shape of the features have been reproduced fairly accurately.

The main area of interest is that enclosed by the dotted line, which was also excavated in 1994 enabling a qualitative assessment of the estimation. An archaeologists impression of this is shown in Figure 10.19. The vertical extent of each of the excavated features is listed in Table 10.16. The range of susceptibility varies from 0.4×10^{-3}**SI** to 2.4×10^{-3}**SI** for organic and non-organic material. The selected grid contains a part from grid J7, J8, K7, K8 and a part from grid K9. The features of the selected grids are buried at the same distance from the site surface but have different susceptibilities and extents. The MAP estimates of the area of interest are obtained with the new technique and the features are now evident (see Figure 10.20). It is easy to see the cluster of post holes in grid J7, using the elastic-net prior (see Figure 10.20). The cluster of post holes in grid J7, incorporating features 14 to 17, have disappeared for several reasons, one reason being that these holes have low magnetic susceptibility values between 0.7×10^{-3}SI to and 1.3×10^{-3}SI. Another

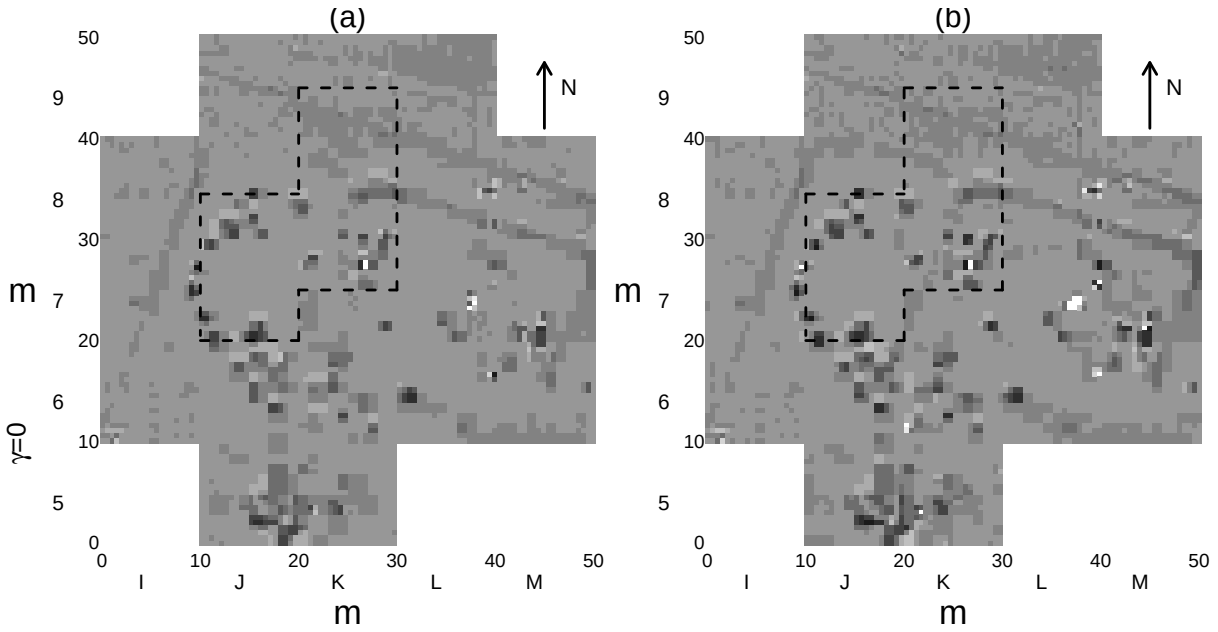


Figure 10.13: Reconstruction of the Park, Guiting Power using MAP estimation: (a) single Laplace; and (b) level-dependent Laplace priors.

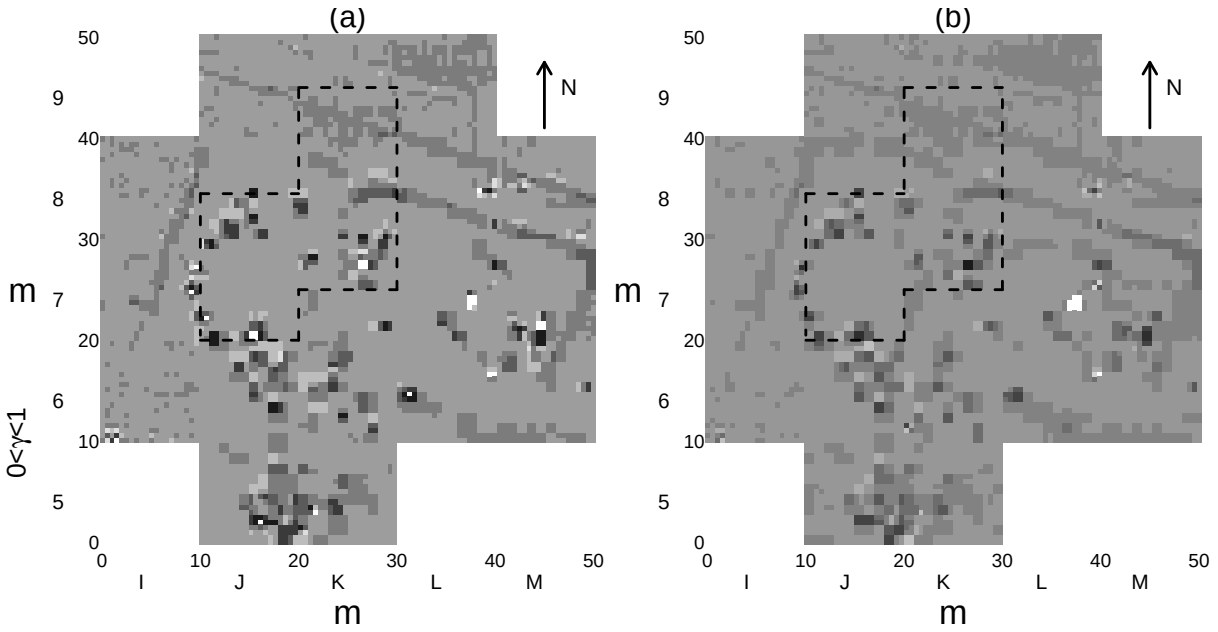


Figure 10.14: Reconstruction of the Park, Guiting Power using MAP estimation: (a) single elastic-net; and (b) level-dependent elastic-net priors.

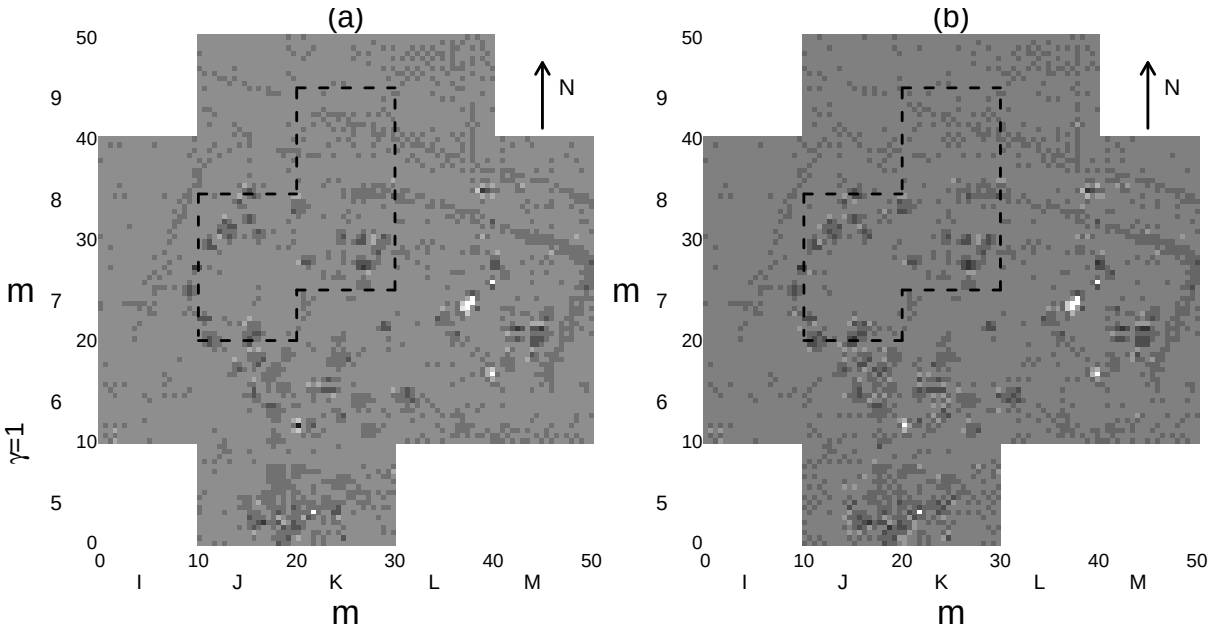


Figure 10.15: Reconstruction of the Park, Guiting Power using MAP estimation: (a) single Gaussian; and (b) level-dependent Gaussian priors.

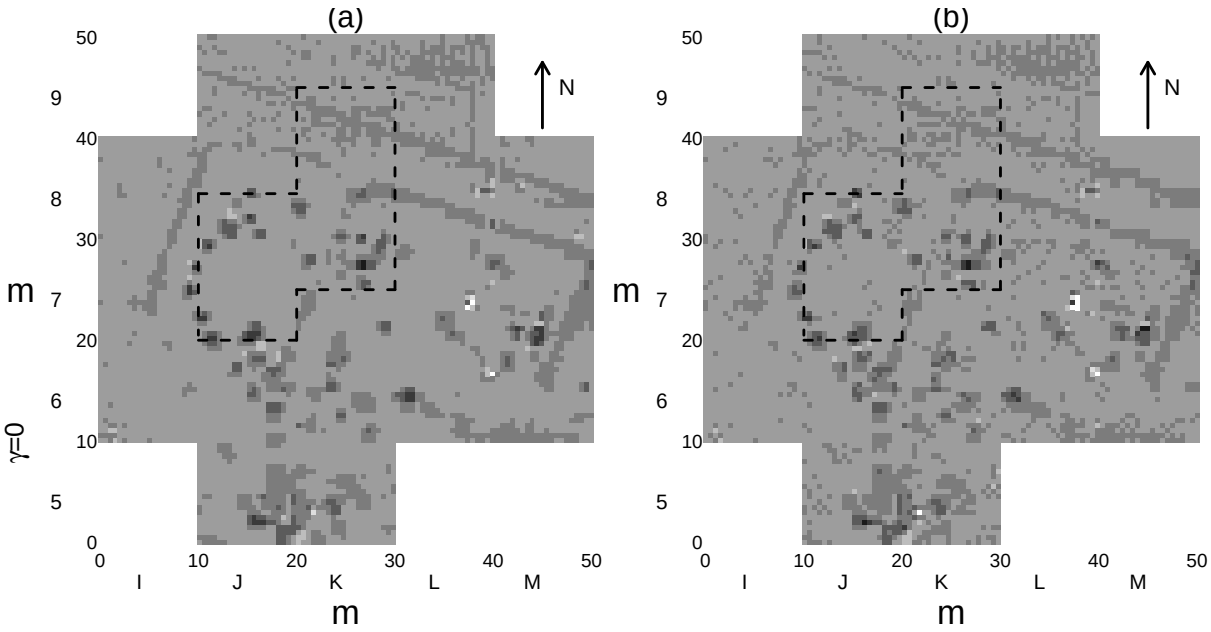


Figure 10.16: Reconstruction of the Park, Guiting Power using the PM estimate: (a) single Laplace; and (b) level-dependent Laplace priors.

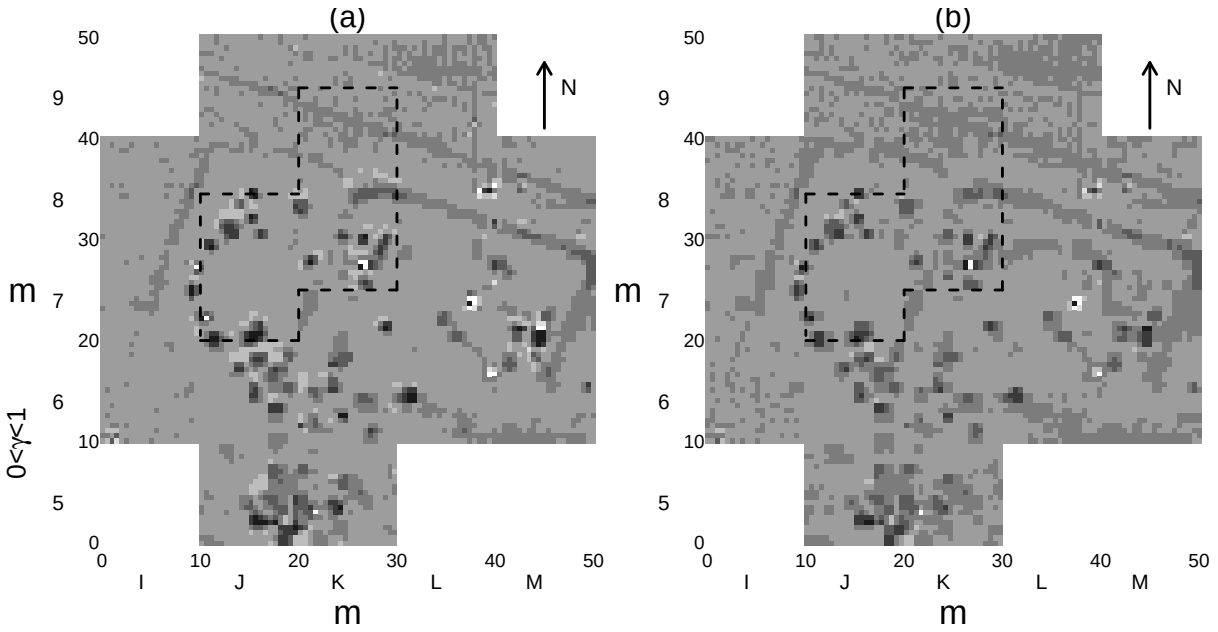


Figure 10.17: Reconstruction of the Park, Guiting Power using the PM estimate: (a) single elastic-net; and (b) level-dependent elastic-net priors.

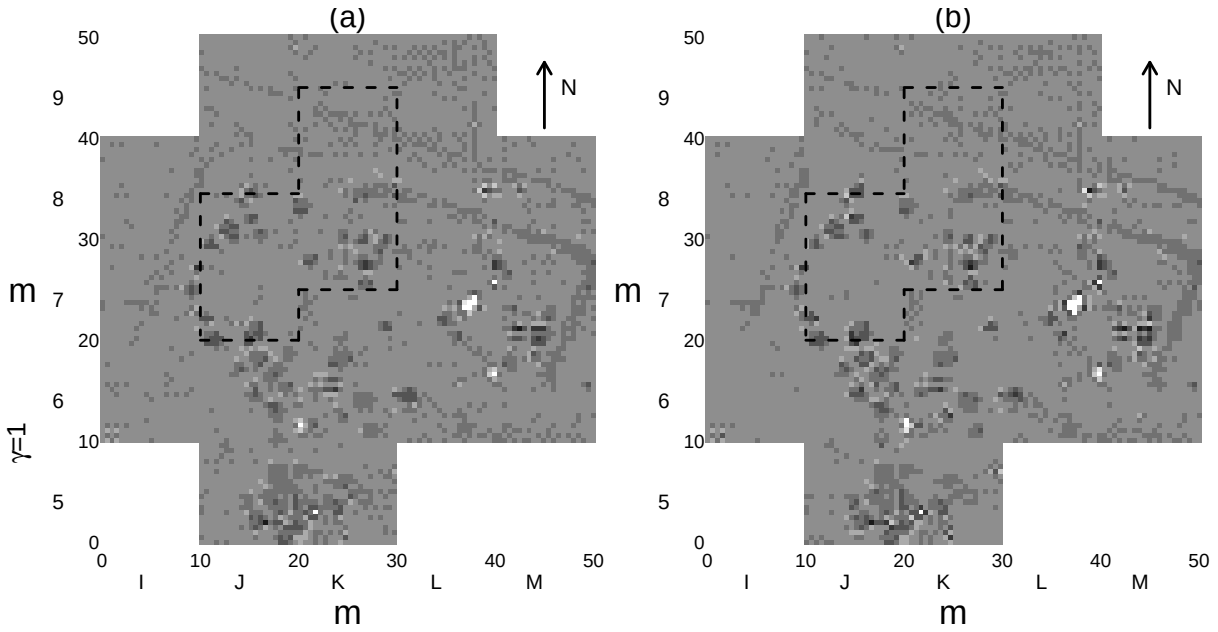


Figure 10.18: Reconstruction of the Park, Guiting Power using the PM estimate: (a) single Gaussian; and (b) level-dependent Gaussian priors.

Grid	Prior	$\widehat{\gamma}$	$\widehat{\kappa}$	$\widehat{\sigma}$
J5	Laplace	0	1023	2.5056
	Elastic-net	0.2195	1024	2.5070
	Gaussian	1	1280	2.5290
J8	Laplace	0	1023	2.4841
	Elastic-net	0.2195	1024	2.4915
	Gaussian	1	1280	2.4975
K9	Laplace	0	1023	0.5104
	Elastic-net	0.2194	1023	0.5110
	Gaussian	1	1279	1.4968

Table 10.15: MAP parameter estimates for single priors.

is that they are very small and placed in the grid of between two to six pixels (Allum, 1997).

Figure 10.21 shows a complex site with overlapping features, including some pits and a part of a ditch, whereas it can also be seen to have different susceptibilities and extents. Some of the data grids contain both strong and weak anomalies covering the range of high negative and high positive data values, such as J5, J8, K5 and L7. Additionally, it is not easy to estimate the true susceptibility, when the features are close.

The parameter values are listed in Table 10.15. It can be seen that some features having disappeared, when the elastic-net for single prior is used, the features of the selected grid are clearly defined. Finally, the feature in the eastern half of data from grid J5 is joined with the western half of data from grid K5. The variance σ^2 is estimated for different grids, which summarised in Table 10.14.

Grid	Prior	$\widehat{\gamma}_j$	$\widehat{\kappa}_j$	$\widehat{\sigma}$
	Laplace	0	$(865.78)_4$ $(480.94)_3$ $(381.85)_2$	2.5456
			$(210.29)_1$ $(129.89)_0$ $(86.61)_{s0}$	
J5	Elastic-net	$(0.0630)_4$ $(0.0578)_3$ $(0.0440)_2$	$(1023)_4$ $(255.99)_3$ $(163.98)_2$	2.5433
		$(0.0461)_1$ $(0.0077)_0$ $(0.0020)_{s0}$	$(115.99)_1$ $(61.739)_0$ $(58.12)_{s0}$	
	Gaussian	1	$(565.68)_4$ $(399.99)_3$ $(282.84)_2$	2.5455
			$(200)_1$ $(65.24)_0$ $(66.82)_{s0}$	
	Laplace	0	$(565.68)_4$ $(400)_3$ $(282.84)_2$	2.5151
			$(200)_1$ $(65.18)_0$ $(61.11)_{s0}$	
J8	Elastic-net	$(0.0565)_4$ $(0.0669)_3$ $(0.0513)_2$	$(1023)_4$ $(256)_3$ $(163.99)_2$	2.5090
		$(0.0471)_1$ $(0.0338)_0$ $(0.0166)_{s0}$	$(114.98)_1$ $(61.66)_0$ $(57.07)_{s0}$	
	Gaussian	1	$(771.01)_4$ $(361.95)_3$ $(390.18)_2$	2.5145
			$(181.69)_1$ $(55.88)_0$ $(51.22)_{s0}$	
	Laplace	0	$(767.18)_4$ $(419)_3$ $(279.04)_2$	1.5426
			$(205.40)_1$ $(63.25)_0$ $(57)_{s0}$	
K9	Elastic-net	$(0.0635)_4$ $(0.0547)_3$ $(0.0526)_2$	$(1023)_4$ $(255.99)_3$ $(164)_2$	0.5405
		$(0.0513)_1$ $(0.0109)_0$ $(0.0382)_{s0}$	$(115.9)_1$ $(81.61)_0$ $(60.49)_{s0}$	
	Gaussian	1	$(565.68)_4$ $(401)_3$ $(268.10)_2$	0.5426
			$(191.79)_1$ $(55.54)_0$ $(60.93)_{s0}$	

Table 10.16: MAP parameter estimates for level-dependent priors.

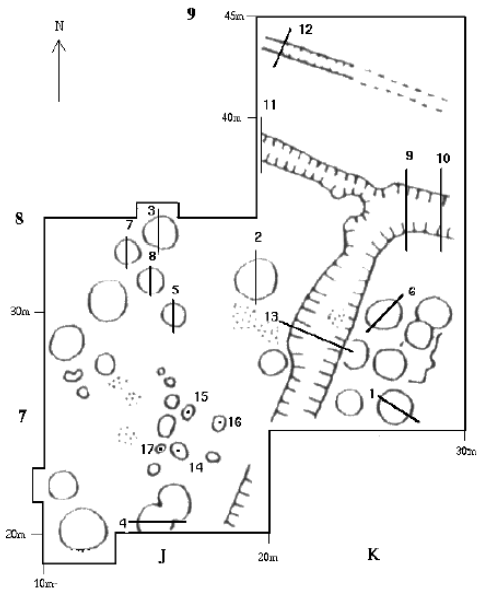


Figure 10.19: Archaeologists impression of the 1994 excavation of the Park, Guiting Power (Allum, 1997).

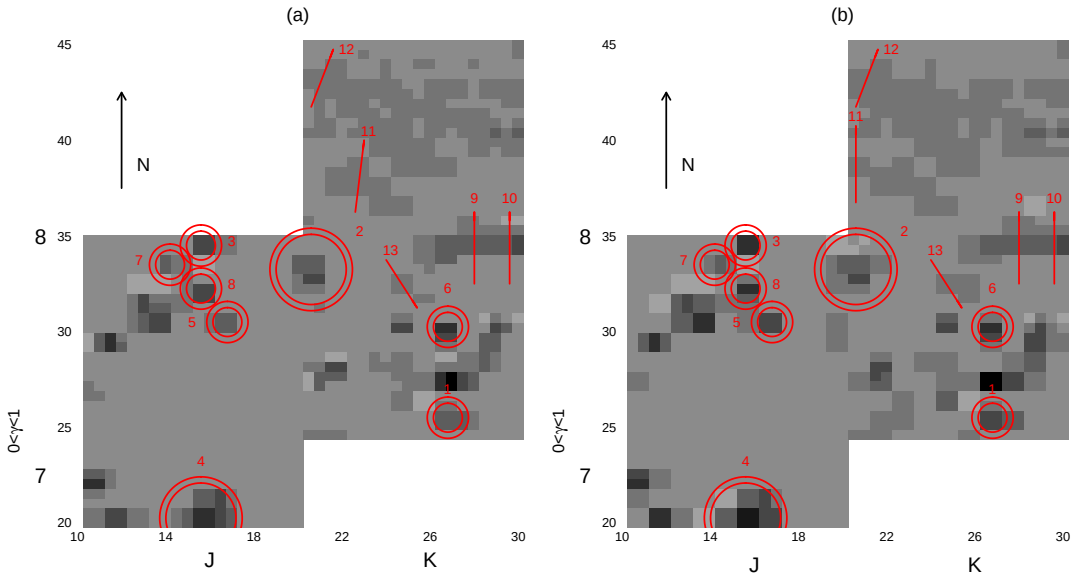


Figure 10.20: Reconstruction of the Park, Guiting Power using MAP estimation: (a) single elastic-net; and (b) level-dependent elastic-net priors, where the red circles and lines represent the features identified by comparing with Figure 10.19.

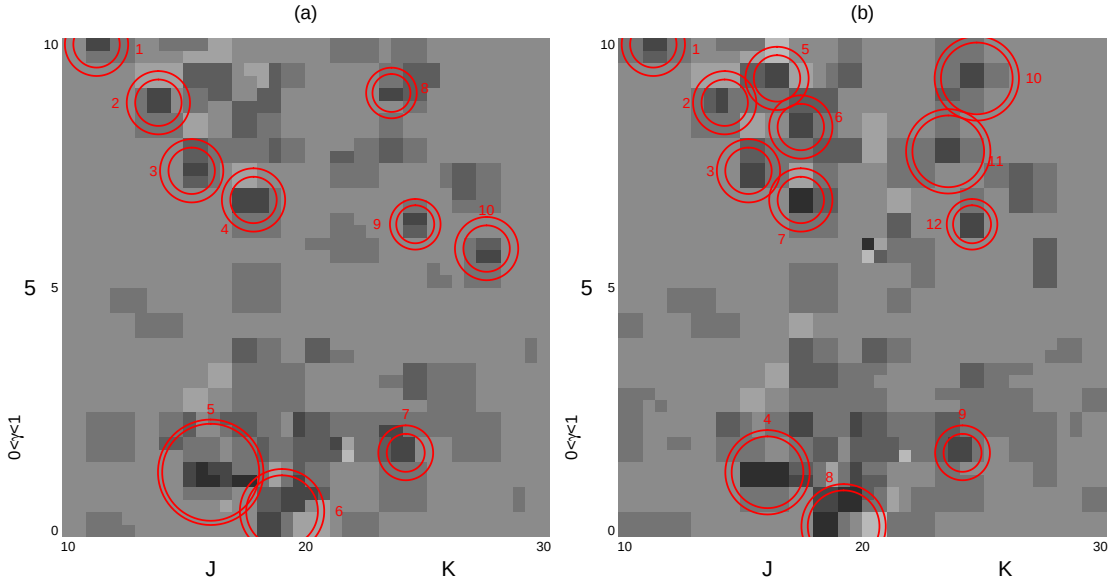


Figure 10.21: Reconstruction of the Park, Guiting Power using MAP estimation: (a) single elastic-net; and (b) level-dependent elastic-net priors, where the red circles represent the features that have much higher magnetic susceptibility than other features in regions in J5 and K5.

10.10 Conclusions

The Bayesian approach to estimate magnetic susceptibility has been successfully adapted for different priors, whereby a hierarchical model, using MCMC through the Metropolis-Hastings algorithm, was formulated.

Wavelet analysis was used to represent the real data, whereby wavelet coefficients were estimated using different priors for applications in one dimensional data and two dimensional magnetometer data. This thesis shows that the magnetic susceptibilities are well estimated.

The reconstruction of magnetic susceptibility and location of the feature using different priors are excellent, except with the Gaussian prior. The features are estimated by removing the noise and blur. Although single and level-dependent priors show that the features are clearly defined, some features disappear when a single prior is used because there is only one value of the threshold for the whole set of wavelet coefficients.

Experiments on simulated datasets show that the main aspects of the feature are well produced. However, reconstructions using the single Gaussian prior show greater variation in the background intensity.

Finally, the estimated features of the Guiting Power site show that the proposed method is capable of correctly locating features on a real site. Most features are well estimated and can easily be identified. However, in some subregions the variances are large, which suggest that it can be difficult to estimate susceptibility and prior parameters simultaneously.

Chapter 11

Final Summary and Conclusions

11.1 Overview

The purpose of this study was to review and develop statistical techniques for the estimation and analysis of inverse problems, with particular application to two types of archaeological magnetometry data. The proposed methods work well and give reconstructions of the archaeological site, which are in agreement with archaeologists' expectations. The magnetic susceptibility is modelled using a hierarchical model. The proposed inversion method determines the magnetic susceptibility profile, or surface, given the recorded data by estimating underlying wavelet coefficients. This thesis is divided into four parts: The first part gives background (Chapter 2); the second part investigates a two-stage method (Chapter 3 to 7); the third part presents a one-stage method (Chapter 8); in the final part, applications to real archaeological magnetometry data analysis are considered (Chapter 9 and 10). This chapter will now give the main conclusions and recommendations for future work.

11.2 Summary

This section presents a summary and discusses the main results obtained by applying different methods to simulated and real data.

The thesis begins with a brief introduction to the magnetic prospecting methods. The data collection is explained briefly, and a discussion on the levels of noise and blur in the modelling process is included. The methodology was motivated by an assumption that the Blocks and Bumps test functions described the underlying susceptibility profile well. We start by introducing wavelet methods and expressing a function as an approximation in terms of wavelet coefficients. The inversion and thresholding methods are then explained, with an emphasis on their specific application to the given experiment.

Several established wavelet denoising methods are discussed and explained, such as classical thresholding rules. The minimum mean squared-error (MMSE) method for estimating the prior parameters, which described in Section 2.13, is applied to different cases and is demonstrated by extensive simulation. We evaluate and investigate the estimation of prior parameters, assuming a single parameter for all wavelet levels and for the level-dependent case. Two types of estimation of the prior parameters are considered based on two-stage, or one-stage approaches. Different wavelet transforms are discussed and applied with different thresholding rules, which are also demonstrated using extensive simulation.

Other Bayesian methods are considered in Chapters 5 and 6. Generalised wavelet methods are introduced in Chapter 7, which are the vaguelette-wavelet decomposition (VWD) and the wavelet-vaguelette decomposition (WVD). It is shown that, WVD with the discrete unbalanced Haar transform (DUHT), the "Larger" posterior mode of double Weibull wavelet shrinkage, the empirical Bayes approach, with posterior median, and the Block-Sure methods work well for denoising, but not for high levels of blurring.

The MCMC algorithms are described in Chapter 8. They are used to estimate the wavelet coefficient of magnetic susceptibility, given the recorded data and prior information. The method proves able to estimate susceptibility values for the one dimensional stratigraphy

problem. It is then applied to a two dimensional magnetometry problem and is again able to produce good reconstructions of the susceptibility surfaces.

The elastic-net prior was applied as a new prior distribution, and was developed specifically to encourage regions of constant susceptibility, which separately sharp discontinuities. Using simulated data, it was shown that the benefit of using the elastic-net prior, compared to the Laplace and Gaussian distributions, is that substantial changes in the estimated susceptibility profile were made by one large step rather than a series of small ones and that constant regions appear with low variability in the estimate.

The final two chapters describe an introduction to archaeological magnetometry data for one dimensional and two dimensional data. The proposed methods are then applied to real archaeological data, which show the new model to be an improvement on the existing model, reducing the mean squared-error (MMSE), in particular with level-dependent priors. The new prior distribution produced regions with flat tops and sharp edges. As well as susceptibility, three prior parameters $(\sigma^2, \kappa, \gamma)$ were also modelled and estimated for both stratigraphy and magnetometry data.

Using simulated data, the maximum a posteriori estimate was shown to be a more practical choice than the posterior mean estimate. As a result, that was the method proposed for estimation from real magnetometry data. It was also demonstrated that it produces good reconstructions of the susceptibility profiles, while direct inversion methods, such as ML estimates, are unsuitable for many reasons, such as the level of the noise and blur.

It is important to note that the proposed methods are not restricted to archaeological magnetometry data but can be applied to observed data obtained from other types of inverse problem.

11.3 Further work

There is a lot of possible future work. The following are some suggestions.

A new method has been identified, which has many applications in imaging. The model is especially relevant to medical images resulting from tomographic investigations that were the original motivation for the Bayesian approach to be applied to the archaeological problems.

There are many problems in the statistical field. For example, the multi-resolution structure of the wavelet coefficients can be used to improve the MCMC algorithm. Suppose the first 500 iterations are used to estimate the wavelet coefficients at resolution level $j = 3$, then the next 500 to estimate the wavelet coefficients at resolution level $j = 4$, etc. This type of construction should make the MCMC algorithm faster.

Within Chapter 3, we applied different thresholding rules with different wavelet transforms and demonstrated that the DUHT algorithm and the non-decimated wavelet transform (NDWT) improve the MSE for estimating the underlying function. Also, MCMC algorithms with DWT was applied and the method developed to provide a general framework for reconstruction where sharp edges are believed to be important. So, the question is: How can the MCMC algorithm, with unbalanced Haar and non-decimated wavelet transforms, be applied to provide a reconstruction in one-stage?

Within Chapter 5, the Laplace prior and the Gaussian prior were discussed with plug-in methods for prior parameters. The reason for using a plug-in method is that it is easy and quicker to apply. Thus, the question is: How can Gaussian distribution as a likelihood and elastic-net distribution as a prior be written as a plug-in method?

Finally, within Chapter 10, a promising topic for future research would be to apply segmentation to the reconstruction of 2D magnetometry data and it would be interesting to apply the same approaches to produce a reconstruction from other type of data, such as single-photon emission computed tomography (SPECT) data.

Appendix

A

In this part, the PM for Gaussian distribution will computing, as given the result in Section 5.4. Vidakovic and Ruggeri (2001) suggested that the coefficients $\mathrm{d_{y}}_{j,k}$ can be considered independently, since the wavelet transformations are decorrelating and they omit the double index j, k and work with a "typical" wavelet coefficient $\mathrm{d_y}$. Additionally the wavelet coefficients, $\mathrm{d_y}$, are modelled via a density $\mathrm{p}(\mathrm{d_y}|\mathrm{d_g})$ where $\mathrm{d_g}$ is the single part.

The posterior mean, as an estimator of $\mathrm{d_g}$, has the following form

$$\mathrm{PM}(\mathrm{d_g}|\mathrm{d_y}) = \frac{\int \mathrm{d_g p}(\mathrm{d_y}|\mathrm{d_g})\pi(\mathrm{d_g})d\mathrm{d_g}}{\mathrm{m}(\mathrm{d_y})} = \frac{\int \mathrm{d_g p}(\mathrm{d_y}|\mathrm{d_g})\pi(\mathrm{d_g})d\mathrm{d_g}}{\int \mathrm{p}(\mathrm{d_y}|\mathrm{d_g})\pi(\mathrm{d_g})d\mathrm{d_g}}, \tag{1}$$

where $\mathrm{PM}(\mathrm{d_g}|\mathrm{d_y})$ is the posterior mean of $\mathrm{d_g}|\mathrm{d_y}$, $\mathrm{p}(\mathrm{d_y}|\mathrm{d_g})\pi(\mathrm{d_g})$ is the likelihood, π is the prior and $\mathrm{m}(\mathrm{d_y})$ is the marginal distribution. Assuming $\mathrm{d_y} > 0$. The expectation of the

marginal of $\mathrm{d_y}$, is given by

$$\begin{aligned}
\mathrm{E(m(d_y))} &= \int \mathrm{d_g} N(\mathrm{d_g}, \sigma^2) N(0, \frac{1}{2\lambda}) d\mathrm{d_g} \\
&= \frac{\sqrt{2\lambda}}{2\pi\sigma} \int \mathrm{d_g} \exp\left\{ -\frac{(\mathrm{d_y}-\mathrm{d_g})^2}{2\sigma^2} \right\} \exp\left\{ -\lambda {\mathrm{d_g}}^2 \right\} d\mathrm{d_g} \\
&= \frac{\sqrt{2\lambda}\exp\left\{ -\frac{1}{2\sigma^2}{\mathrm{d_y}}^2 + \frac{2\sigma^2{\mathrm{d_y}}^2}{(1+2\lambda\sigma^2)} \right\}}{2\pi\sigma} \int_{-\infty}^{\infty} \mathrm{d_g} \exp\left\{ -\frac{(\mathrm{d_g} - \frac{\mathrm{d_y}}{1+2\lambda\sigma^2})^2}{\frac{2\sigma^2}{(1+2\lambda\sigma^2)}} \right\} d\mathrm{d_g} \\
&= \frac{\sqrt{2\lambda}}{2\pi\sqrt{1+2\lambda\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}{\mathrm{d_y}}^2 + \frac{1}{(1+2\lambda\sigma^2)} 2\sigma^2 {\mathrm{d_y}}^2 \right\} \\
&\qquad \times \int_{-\infty}^{\infty} (\frac{\sigma \mathrm{y}}{\sqrt{1+2\lambda\sigma^2}} + \frac{\mathrm{d_y}}{1+2\lambda\sigma^2}) \exp\left\{ -\frac{1}{2}\mathrm{y}^2 \right\} d\mathrm{y} \\
&= \frac{\sqrt{\pi\lambda}\mathrm{d_y}}{\pi(1+2\lambda\sigma^2)\sqrt{1+2\lambda\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}{\mathrm{d_y}}^2 + \frac{1}{(1+2\lambda\sigma^2)} 2\sigma^2 {\mathrm{d_y}}^2 \right\}, \qquad (2)
\end{aligned}$$

where $\mathrm{y} = \frac{\mathrm{d_g} - \frac{\mathrm{d_y}}{1+2\lambda\sigma^2}}{\frac{\sigma}{\sqrt{1+2\lambda\sigma^2}}}$. Assuming $\mathrm{d_y} > 0$. The marginal distribution of $\mathrm{d_y}$, can be written as

$$\begin{aligned}
\mathrm{m(d_y)} &= \int N(\mathrm{d_g}, \sigma^2) N(0, \frac{1}{2\lambda}) d\mathrm{d_g} \\
&= \frac{\sqrt{2\lambda}}{2\pi\sigma} \int \mathrm{d_g} \exp\left\{ -\frac{(\mathrm{d_y}-\mathrm{d_g})^2}{2\sigma^2} \right\} \exp\left\{ -\lambda {\mathrm{d_g}}^2 \right\} d\mathrm{d_g} \\
&= \frac{\sqrt{2\lambda}\exp\left\{ -\frac{1}{2\sigma^2}{\mathrm{d_y}}^2 + \frac{2\sigma^2{\mathrm{d_y}}^2}{(1+2\lambda\sigma^2)} \right\}}{2\pi\sigma} \int_{-\infty}^{\infty} \exp\left\{ -\frac{(\mathrm{d_g} - \frac{\mathrm{d_y}}{1+2\lambda\sigma^2})^2}{\frac{2\sigma^2}{(1+2\lambda\sigma^2)}} \right\} d\mathrm{d_g} \\
&= \frac{\sqrt{2\lambda}}{2\pi\sqrt{1+2\lambda\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}{\mathrm{d_y}}^2 + \frac{1}{(1+2\lambda\sigma^2)} 2\sigma^2 {\mathrm{d_y}}^2 \right\} \int_{-\infty}^{\infty} \exp\left\{ -\frac{1}{2}\mathrm{y}^2 \right\} d\mathrm{y} \\
&= \frac{\sqrt{\pi\lambda}}{\pi\sqrt{1+2\lambda\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}{\mathrm{d_y}}^2 + \frac{1}{(1+2\lambda\sigma^2)} 2\sigma^2 {\mathrm{d_y}}^2 \right\}. \qquad (3)
\end{aligned}$$

Then, the posterior mean of $\mathrm{d_g}|\mathrm{d_y}$, is given by

$$\mathrm{PM(d_y)} = \left(\frac{\mathrm{d_y}}{1+2\lambda\sigma^2} \right). \qquad (4)$$

B

In this part, the PM for Laplace distribution will computing, as given in section 5.5. Additionally the wavelet coefficients, $\mathrm{d_y}$, are modelled via a density $\mathrm{p}(\mathrm{d_y}|\mathrm{d_g})$ where $\mathrm{d_g}$ is the single part.

The posterior mean is given by

$$\mathrm{PM}(\mathrm{d_g}|\mathrm{d_y}) = \frac{\int \mathrm{d_g p}(\mathrm{d_y}|\mathrm{d_g})\pi(\mathrm{d_g})d\mathrm{d_g}}{\mathrm{m}(\mathrm{d_y})} = \frac{\int \mathrm{d_g p}(\mathrm{d_y}|\mathrm{d_g})\pi(\mathrm{d_g})d\mathrm{d_g}}{\int \mathrm{p}(\mathrm{d_y}|\mathrm{d_g})\pi(\mathrm{d_g})d\mathrm{d_g}}, \tag{5}$$

where $\mathrm{PM}(\mathrm{d_g}|\mathrm{d_y})$ is the posterior mean of $\mathrm{d_g}|\mathrm{d_y}$, $\mathrm{p}(\mathrm{d_y}|\mathrm{d_g})\pi(\mathrm{d_g})$ is the likelihood, π is the prior and $\mathrm{m}(\mathrm{d_y})$ is the marginal distribution. Assuming $\mathrm{d_y} > 0$. The expectation of the marginal of $\mathrm{d_y}$

$$\begin{aligned}
\mathrm{E}(\mathrm{m}(\mathrm{d_y})) &= \int \mathrm{d_g p}(\mathrm{d_y}|\mathrm{d_g})\pi(\mathrm{d_g})d\mathrm{d_g} \\
&= \int \frac{\lambda \mathrm{d_g}}{2\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(\mathrm{d_y}-\mathrm{d_g})^2}{2\sigma^2}\right\} \exp\left\{ -\lambda|\mathrm{d_g}|\right\} d\mathrm{d_g} \\
&= \frac{\lambda \exp\left\{ -\frac{1}{2\sigma^2}{\mathrm{d_y}}^2\right\}}{2\sqrt{2\pi\sigma^2}} \left[\exp\left\{ \frac{(\mathrm{d_y}+\sigma^2\lambda)^2}{2\sigma^2}\right\} \right. \\
&\times \int_{-\infty}^{0} \mathrm{d_g} \exp\left\{ -\frac{(\mathrm{d_g}-(\mathrm{d_y}+\sigma^2\lambda))^2}{2\sigma^2}\right\} d\mathrm{d_g} + \exp\left\{ \frac{(\mathrm{d_y}-\sigma^2\lambda)^2}{2\sigma^2}\right\} \\
&\times \left. \int_{0}^{\infty} \mathrm{d_g} \exp\left\{ -\frac{(\mathrm{d_g}-(\mathrm{d_y}-\sigma^2\lambda))^2}{2\sigma^2}\right\} d\mathrm{d_g} \right]
\end{aligned}$$

$$
\begin{aligned}
=&\frac{\lambda \exp\left\{-\frac{1}{2\sigma^2}\mathrm{d_y}^2\right\}}{2\sqrt{2\pi}}\Bigg[\exp\left\{\frac{(\mathrm{d_y}+\sigma^2\lambda)^2}{2\sigma^2}\right\}\int_{-\infty}^{-\frac{(\mathrm{d_y}+\sigma^2\lambda)}{\sigma}}(\sigma \mathrm{y}+(\mathrm{d_y}+\sigma^2\lambda))\exp\left\{-\frac{1}{2}\mathrm{y}^2\right\}d\mathrm{y}\\
&+\exp\left\{\frac{(\mathrm{d_y}-\sigma^2\lambda)^2}{2\sigma^2}\right\}\int_{-\frac{(\mathrm{d_y}-\sigma^2\lambda)}{\sigma}}^{\infty}(\sigma \mathrm{y}+(\mathrm{d_y}-\sigma^2\lambda))\exp\left\{-\frac{1}{2}\mathrm{y}^2\right\}d\mathrm{y}\Bigg]\\
=&\frac{\lambda \exp\left\{-\frac{1}{2\sigma^2}\mathrm{d_y}^2\right\}}{2}\Bigg[(\mathrm{d_y}+\sigma^2\lambda)\exp\left\{\frac{(\mathrm{d_y}+\sigma^2\lambda)^2}{2\sigma^2}\right\}\bar{\Phi}(\frac{\mathrm{d_y}+\sigma^2\lambda}{\sigma})\\
&+(\mathrm{d_y}-\sigma^2\lambda)\exp\left\{\frac{(\mathrm{d_y}-\sigma^2\lambda)^2}{2\sigma^2}\right\}\Phi(\frac{\mathrm{d_y}-\sigma^2\lambda}{\sigma})\Bigg], \qquad (6)
\end{aligned}
$$

where $y=\frac{(\mathrm{d_y}-\sigma^2\lambda)}{\sigma}$. The marginal of $\mathrm{d_y}$, is given by

$$
\begin{aligned}
\mathrm{m}(\mathrm{d_y}) =& \int N(\mathrm{d_g},\sigma^2)\mathcal{DE}(0,\lambda)d\mathrm{d_g}\\
=& \int \frac{\lambda}{2\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(\mathrm{d_y}-\mathrm{d_g})^2}{2\sigma^2}\right\}\exp\left\{-\lambda|\mathrm{d_g}|\right\}d\mathrm{d_g}\\
=&\frac{\lambda \exp\left\{-\frac{1}{2\sigma^2}\mathrm{d_y}^2\right\}}{2\sqrt{2\pi\sigma^2}}\Bigg[\exp\left\{\frac{(\mathrm{d_y}+\sigma^2\lambda)^2}{2\sigma^2}\right\}\\
&\times\int_{-\infty}^{0}\exp\left\{-\frac{(\mathrm{d_g}-(\mathrm{d_y}+\sigma^2\lambda))^2}{2\sigma^2}\right\}d\mathrm{d_g}+\exp\left\{\frac{(\mathrm{d_y}-\sigma^2\lambda)^2}{2\sigma^2}\right\}\\
&\times\int_{0}^{\infty}\exp\left\{-\frac{(\mathrm{d_g}-(\mathrm{d_y}-\sigma^2\lambda))^2}{2\sigma^2}\right\}d\mathrm{d_g}\Bigg]\\
=&\frac{\sigma\lambda \exp\left\{-\frac{1}{2\sigma^2}\mathrm{d_y}^2\right\}}{2\sqrt{2\pi\sigma^2}}\Bigg[\exp\left\{\frac{(\mathrm{d_y}+\sigma^2\lambda)^2}{2\sigma^2}\right\}\\
&\times\int_{-\infty}^{-\frac{(\mathrm{d_y}+\sigma^2\lambda)}{\sigma}}\exp\left\{-\frac{1}{2}\mathrm{y}^2\right\}d\mathrm{y}+\exp\left\{\frac{(\mathrm{d_y}-\sigma^2\lambda)^2}{2\sigma^2}\right\}\\
&\times\int_{-\frac{(\mathrm{d_y}-\sigma^2\lambda)}{\sigma}}^{\infty}\exp\left\{-\frac{1}{2}\mathrm{y}^2\right\}d\mathrm{y}\Bigg]\\
=&\frac{\lambda \exp\left\{-\frac{1}{2\sigma^2}\mathrm{d_y}^2\right\}}{2}\Bigg[\exp\left\{\frac{(\mathrm{d_y}+\sigma^2\lambda)^2}{2\sigma^2}\right\}\Phi\bigg(-\frac{(\mathrm{d_y}+\sigma^2\lambda)}{\sigma}\bigg)\\
&+\exp\left\{\frac{(\mathrm{d_y}-\sigma^2\lambda)^2}{2\sigma^2}\right\}\tilde{\Phi}\bigg(-\frac{(\mathrm{d_y}-\sigma^2\lambda)}{\sigma}\bigg)\Bigg]
\end{aligned}
$$

$$=\frac{\lambda \exp\left\{-\frac{1}{2\sigma^2}\mathrm{d_y}^2\right\}}{2}\left[\exp\left\{\frac{(\mathrm{d_y}+\sigma^2\lambda)^2}{2\sigma^2}\right\}\widehat{\Phi}\left(-\frac{(\mathrm{d_y}+\sigma^2\lambda)}{\sigma}\right)\right.$$
$$\left.-\exp\left\{\frac{(\mathrm{d_y}-\sigma^2\lambda)^2}{2\sigma^2}\right\}\Phi\left(\frac{(d-\sigma^2\lambda)}{\sigma}\right)\right], \tag{7}$$

where Φ is the cumulative standard normal. The posterior mean of $\mathrm{d_g}|\mathrm{d_y}$, is given by

$$\begin{aligned}\mathrm{PM}(\mathrm{d_g}|\mathrm{d_y}) &= \frac{\int \mathrm{d_g p}(\mathrm{d_y}|\mathrm{d_g})\pi(\mathrm{d_g})d\mathrm{d_g}}{\int \mathrm{p}(\mathrm{d_y}|\mathrm{d_g})\pi(\mathrm{d_g})d\mathrm{d_g}}\\
&=\left[(\mathrm{d_y}+\sigma^2\lambda)\exp\left\{\frac{(\mathrm{d_y}+\sigma^2\lambda)^2}{2\sigma^2}\right\}\bar{\Phi}(\frac{\mathrm{d_y}+\sigma^2\lambda}{\sigma})\right.\\
&\quad\left.+(\mathrm{d_y}-\sigma^2\lambda)\exp\left\{\frac{(\mathrm{d_y}-\sigma^2\lambda)^2}{2\sigma^2}\right\}\Phi(\frac{\mathrm{d_y}-\sigma^2\lambda}{\sigma})\right]\\
&\quad\Big/\left[\exp\left\{\frac{(\mathrm{d_y}+\sigma^2\lambda)^2}{2\sigma^2}\right\}\bar{\Phi}\left(-\frac{(\mathrm{d_y}+\sigma^2\lambda)}{\sigma}\right)\right.\\
&\quad\left.-\exp\left\{\frac{(\mathrm{d_y}-\sigma^2\lambda)^2}{2\sigma^2}\right\}\Phi\left(\frac{(\mathrm{d_y}-\sigma^2\lambda)}{\sigma}\right)\right].\end{aligned} \tag{8}$$

In this part, the mean and variance of soft will prove, as given in Section 2.11. Let $\mathrm{d_y} \sim N(\mathrm{d_g}, \sigma^2)$, where $\mathrm{d_y}$ represents the wavelet coefficients and $\mathrm{d_g}$ represents the mean of normal. The mean, variance and the risk function of the shrinkage estimator of $\mathrm{d_g}$ can be written under shrinkage function $\mathrm{T}_\lambda(\mathrm{d_y})$ and threshold $\lambda = \pm\lambda\sigma^2$ by following Chib and Greenberg (1995):

$$\mathrm{E}_\lambda(\mathrm{d_g}) = \mathrm{E}(\mathrm{T}_\lambda(\mathrm{d_y})), \tag{9}$$

$$\mathrm{V}_\lambda(\mathrm{d_g}) = \mathrm{V}(\mathrm{T}_\lambda(\mathrm{d_y})), \tag{10}$$

$$\mathrm{R}_\lambda(\mathrm{d_g}) = \mathrm{E}(\mathrm{T}_\lambda(\mathrm{d_y}) - \mathrm{d_g})^2 = \mathrm{V}_\lambda(\mathrm{d_g}) + (\mathrm{E}_\lambda(\mathrm{d_g}))^2. \tag{11}$$

First of all the mean of soft is given by

$$\mathrm{E}_\lambda(\mathrm{d_g}) = \int_{-\infty}^{-\lambda\sigma^2} \frac{(\mathrm{d_y}+\lambda\sigma^2)}{\sqrt{2\pi\sigma^2}}\exp\left\{\frac{-(\mathrm{d_y}-\mathrm{d_g})^2}{2\sigma^2}\right\}d\mathrm{d_y} + \int_{\lambda\sigma^2}^{\infty}\frac{(\mathrm{d_y}-\lambda\sigma^2)}{\sqrt{2\pi\sigma^2}}\exp\left\{\frac{-(\mathrm{d_y}-\mathrm{d_g})^2}{2\sigma^2}\right\}d\mathrm{d_y}. \tag{12}$$

Then setting $\mathrm{x} = \frac{\mathrm{d_y}-\mathrm{d_g}}{\sigma}$ this implies that $\sigma \mathrm{x} + \mathrm{d_g} = \mathrm{d_y}$ and $\sigma d\mathrm{x} = d\mathrm{d_y}$ and

$$
\begin{aligned}
\mathrm{E}_\lambda(\mathrm{d_g}) &= \int_{-\infty}^{\frac{-\lambda\sigma^2-\mathrm{d_g}}{\sigma}} (\sigma\mathrm{x} + \mathrm{d_g} + \lambda\sigma^2) \frac{\exp\left\{-\frac{\mathrm{x}^2}{2}\right\}}{\sqrt{2\pi}} d\mathrm{x} \\
&\quad + \int_{\frac{\lambda\sigma^2-\mathrm{d_g}}{\sigma}}^{\infty} (\sigma\mathrm{x} + \mathrm{d_g} - \lambda\sigma^2) \frac{\exp\left\{-\frac{\mathrm{x}^2}{2}\right\}}{\sqrt{2\pi}} d\mathrm{x} \\
&= \int_{-\infty}^{\frac{-\lambda\sigma^2-\mathrm{d_g}}{\sigma}} (\sigma\mathrm{x} + \mathrm{d_g}) \frac{\exp\left\{-\frac{\mathrm{x}^2}{2}\right\}}{\sqrt{2\pi}} d\mathrm{x} + \int_{-\infty}^{\frac{-\lambda\sigma^2-\mathrm{d_g}}{\sigma}} \sigma^2\lambda \frac{\exp\left\{-\frac{\mathrm{x}^2}{2}\right\}}{\sqrt{2\pi}} d\mathrm{x} \\
&\quad + \int_{\frac{\lambda\sigma^2-\mathrm{d_g}}{\sigma}}^{\infty} (\sigma\mathrm{x} + \mathrm{d_g}) \frac{\exp\left\{-\frac{\mathrm{x}^2}{2}\right\}}{\sqrt{2\pi}} d\mathrm{x} - \int_{\frac{\lambda\sigma^2-\mathrm{d_g}}{\sigma}}^{\infty} \sigma^2\lambda \frac{\exp\left\{-\frac{\mathrm{x}^2}{2}\right\}}{\sqrt{2\pi}} d\mathrm{x} \\
&= -\sigma\phi(\mathrm{x})\big|_{-\infty}^{-\frac{\lambda\sigma^2-\mathrm{d_g}}{\sigma}} + \mathrm{d_g}\Phi(\frac{-\lambda\sigma^2-\mathrm{d_g}}{\sigma}) + \sigma 2\lambda\Phi(\frac{-\lambda\sigma^2-\mathrm{d_g}}{\sigma}) \\
&\quad - \sigma\phi(\mathrm{x})\big|_{\frac{\lambda\sigma^2-\mathrm{d_g}}{\sigma}}^{\infty} + \mathrm{d_g}\bar{\Phi}\left(\frac{\lambda\sigma^2-\mathrm{d_g}}{\sigma}\right) + \sigma^2\lambda\Phi\left(\frac{\lambda\sigma^2-\mathrm{d_g}}{\sigma}\right) \\
&= -\sigma\phi\left(\frac{\lambda\sigma^2+\mathrm{d_g}}{\sigma}\right) + \mathrm{d_g}\Phi\left(\frac{-\lambda\sigma^2-\mathrm{d_g}}{\sigma}\right) + \sigma^2\lambda\Phi\left(\frac{-\lambda\sigma^2-\mathrm{d_g}}{\sigma}\right) \\
&\quad + \sigma\phi\left(\frac{\lambda\sigma^2-\mathrm{d_g}}{\sigma}\right) + \mathrm{d_g}\bar{\Phi}\left(\frac{\lambda\sigma^2-\mathrm{d_g}}{\sigma}\right) + \sigma^2\lambda\Phi\left(\frac{\lambda\sigma^2-\mathrm{d_g}}{\sigma}\right) \\
&= -\sigma\phi\left(\frac{\lambda\sigma^2+\mathrm{d_g}}{\sigma}\right) + \mathrm{d_g}\bar{\Phi}\left(\frac{\lambda\sigma^2+\mathrm{d_g}}{\sigma}\right) - \sigma^2\lambda\Phi\left(\frac{\lambda\sigma^2+\mathrm{d_g}}{\sigma}\right) \\
&\quad + \sigma\phi\left(\frac{\lambda\sigma^2-\mathrm{d_g}}{\sigma}\right) + \mathrm{d_g}\bar{\Phi}\left(\frac{\lambda\sigma^2-\mathrm{d_g}}{\sigma}\right) + \sigma^2\lambda\Phi\left(\frac{\lambda\sigma^2-\mathrm{d_g}}{\sigma}\right). \qquad (13)
\end{aligned}
$$

Note that $\Phi\left(\frac{-\lambda\sigma^2-\mathrm{d_g}}{\sigma}\right) = \Phi\left(\frac{\lambda\sigma^2+\mathrm{d_g}}{\sigma}\right)$. Then, the $\mathrm{E}_\lambda(\mathrm{d_g})^2$, is given by

$$
\mathrm{E}_\lambda(\mathrm{d_g})^2 = \int_{-\infty}^{-\lambda\sigma^2} \frac{(\mathrm{d_y}+\lambda\sigma^2)^2}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(\mathrm{d_y}-\mathrm{d_g})^2}{2\sigma^2}\right\} d\mathrm{d_y} + \int_{\lambda\sigma^2}^{\infty} \frac{(Y-\lambda\sigma^2)^2}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(\mathrm{d_y}-\mathrm{d_g})^2}{2\sigma^2}\right\} d\mathrm{d_y}. \qquad (14)
$$

$$\mathrm{E}_\lambda(\mathrm{d_g})^2 = \int_{-\infty}^{\frac{-\lambda\sigma^2-\mathrm{d_g}}{\sigma}} (\sigma \mathrm{x} + \sigma \mathrm{d_g} + \lambda\sigma^2)^2 \frac{\exp\left\{-\frac{\mathrm{x}^2}{2}\right\}}{\sqrt{2\pi}} d\mathrm{x} \tag{15}$$

$$+ \int_{\frac{\lambda\sigma^2-\mathrm{d_g}}{\sigma}}^{\infty} (\sigma \mathrm{x} + \mathrm{d_g} - \lambda\sigma^2)^2 \frac{\exp\left\{-\frac{\mathrm{x}^2}{2}\right\}}{\sqrt{2\pi}} d\mathrm{x}$$

$$= \int_{-\infty}^{\frac{-\lambda\sigma^2-\mathrm{d_g}}{\sigma}} (\sigma \mathrm{x}^2 + 2\mathrm{x}(\mathrm{d_g} + \sigma^2\lambda) + (\mathrm{d_g} + \sigma^2\lambda)^2) \frac{\exp\left\{-\frac{\mathrm{x}^2}{2}\right\}}{\sqrt{2\pi}} d\mathrm{x}$$

$$+ \int_{\frac{\lambda\sigma^2-\mathrm{d_g}}{\sigma}}^{\infty} (\sigma \mathrm{x}^2 + 2\mathrm{x}(\mathrm{d_g} - \sigma^2\lambda) + (\mathrm{d_g} - \sigma^2\lambda)^2) \frac{\exp\left\{-\frac{\mathrm{x}^2}{2}\right\}}{\sqrt{2\pi}} d\mathrm{x}$$

$$= -\sigma \mathrm{x}\phi(\mathrm{x})\Big|_{-\infty}^{-\frac{\lambda\sigma^2-\mathrm{d_g}}{\sigma}} + \sigma\Phi(\mathrm{x})\Big|_{-\infty}^{-\frac{\lambda\sigma^2-\mathrm{d_g}}{\sigma}} - 2(\mathrm{d_g} + \sigma^2\lambda)\phi(\mathrm{x})\Big|_{-\infty}^{-\frac{\lambda\sigma^2-\mathrm{d_g}}{\sigma}} + (\mathrm{d_g} + \sigma^2\lambda)^2\phi(\mathrm{x})\Big|_{-\infty}^{-\frac{\lambda\sigma^2-\mathrm{d_g}}{\sigma}}$$

$$- \sigma \mathrm{x}\phi(\mathrm{x})\Big|_{\frac{\lambda\sigma^2-\mathrm{d_g}}{\sigma}}^{-\infty} + \sigma\Phi(\mathrm{x})\Big|_{\frac{\lambda\sigma^2-\mathrm{d_g}}{\sigma}}^{-\infty} - 2(\mathrm{d_g} + \sigma^2\lambda)\phi(\mathrm{x})\Big|_{\frac{\lambda\sigma^2-\mathrm{d_g}}{\sigma}}^{-\infty} + (\mathrm{d_g} + \sigma^2\lambda)^2\phi(\mathrm{x})\Big|_{\frac{\lambda\sigma^2-\mathrm{d_g}}{\sigma}}^{-\infty}$$

$$= -\sigma\left(\frac{\lambda\sigma^2 + \mathrm{d_g}}{\sigma}\right)\phi\left(\frac{\lambda\sigma^2 + \mathrm{d_g}}{\sigma}\right) + \sigma\Phi\left(\frac{-(\lambda\sigma^2 + \mathrm{d_g})}{\sigma}\right) - 2(\mathrm{d_g} + \sigma^2\lambda)\Phi\left(\frac{-(\lambda\sigma^2 + \mathrm{d_g})}{\sigma}\right)$$

$$+ (\mathrm{d_g} + \sigma^2\lambda)^2\Phi\left(\frac{-(\lambda\sigma^2 + \mathrm{d_g})}{\sigma}\right)$$

$$+ \sigma\left(\frac{\lambda\sigma^2 - \mathrm{d_g}}{\sigma}\right)\bar{\Phi}\left(\frac{\lambda\sigma^2 - \mathrm{d_g}}{\sigma}\right) + \sigma\bar{\Phi}\left(\frac{\lambda\sigma^2 - \mathrm{d_g}}{\sigma}\right) + 2(\mathrm{d_g} - \sigma^2\lambda)\bar{\Phi}\left(\frac{\lambda\sigma^2 - \mathrm{d_g}}{\sigma}\right)$$

$$+ (\mathrm{d_g} - \sigma^2\lambda)^2\bar{\Phi}\left(\frac{\lambda\sigma^2 - \mathrm{d_g}}{\sigma}\right)$$

$$= -\sigma(\frac{\lambda\sigma^2 + \mathrm{d_g}}{\sigma})\phi\left(\frac{\lambda\sigma^2 + \mathrm{d_g}}{\sigma}\right) + \sigma\Phi\left(\frac{(\lambda\sigma^2 + \mathrm{d_g})}{\sigma}\right) - 2(\mathrm{d_g} + \sigma^2\lambda)\left(1 - \Phi\left(\frac{\lambda\sigma^2 + \mathrm{d_g}}{\sigma}\right)\right)$$

$$+ (\mathrm{d_g} + \sigma^2\lambda)^2\left(1 - \Phi\left(\frac{(\lambda\sigma^2 + \mathrm{d_g})}{\sigma}\right)\right)$$

$$+ \sigma\left(\frac{\lambda\sigma^2 - \mathrm{d_g}}{\sigma}\right)\bar{\Phi}\left(\frac{\lambda\sigma^2 - \mathrm{d_g}}{\sigma}\right) + \sigma\bar{\Phi}\left(\frac{\lambda\sigma^2 - \mathrm{d_g}}{\sigma}\right) + 2(\mathrm{d_g} - \sigma^2\lambda)\bar{\Phi}\left(\frac{\lambda\sigma^2 - \mathrm{d_g}}{\sigma}\right)$$

$$+ (\mathrm{d_g} - \sigma^2\lambda)^2\bar{\Phi}\left(\frac{\lambda\sigma^2 - \mathrm{d_g}}{\sigma}\right)$$

$$= -\sigma\left(\frac{\lambda\sigma^2 + \mathrm{d_g}}{\sigma}\right)\phi\left(\frac{\lambda\sigma^2 + \mathrm{d_g}}{\sigma}\right) + \left(\sigma - 2(\mathrm{d_g} + \sigma^2\lambda) + (\mathrm{d_g} + \sigma^2\lambda)^2\right)\bar{\Phi}\left(\frac{\lambda\sigma^2 + \mathrm{d_g}}{\sigma}\right)$$

$$+ \sigma\left(\frac{\lambda\sigma^2 - \mathrm{d_g}}{\sigma}\right)\phi\left(\frac{\lambda\sigma^2 - \mathrm{d_g}}{\sigma}\right) + \left(\sigma + 2(\mathrm{d_g} - \sigma^2\lambda) + (\mathrm{d_g} - \sigma^2\lambda)^2\right)\bar{\Phi}\left(\frac{\lambda\sigma^2 - \mathrm{d_g}}{\sigma}\right). \tag{16}$$

Then, the variance is given by

$$
\begin{aligned}
\mathrm{V}_\lambda(\mathrm{d_g}) &= \mathrm{E}(\mathrm{T}_\lambda(\mathrm{d_y}))^2 - (\mathrm{E}(\mathrm{T}_\lambda(\mathrm{d_y})))^2 \\
&= -\sigma\left(\frac{\lambda\sigma^2+\mathrm{d_g}}{\sigma}\right)\phi\left(\frac{\lambda\sigma^2+\mathrm{d_g}}{\sigma}\right) + (\sigma - 2(\mathrm{d_g}+\sigma^2\lambda) + (\mathrm{d_g}+\sigma^2\lambda)^2)\bar{\Phi}\left(\frac{\lambda\sigma^2+\mathrm{d_g}}{\sigma}\right) \\
&\quad + \sigma\left(\frac{\lambda\sigma^2-\mathrm{d_g}}{\sigma}\right)\phi\left(\frac{\lambda\sigma^2-\mathrm{d_g}}{\sigma}\right) + (\sigma + 2(\mathrm{d_g}-\sigma^2\lambda) + (\mathrm{d_g}-\sigma^2\lambda)^2)\bar{\Phi}\left(\frac{\lambda\sigma^2-\mathrm{d_g}}{\sigma}\right) \\
&\quad - \left(- \sigma\phi\left(\frac{\lambda\sigma^2+\theta}{\sigma}\right) + \theta\bar{\Phi}\left(\frac{\lambda\sigma^2+\theta}{\sigma}\right) - \sigma^2\lambda\Phi\left(\frac{\lambda\sigma^2+\mathrm{d_g}}{\sigma}\right)\right. \\
&\quad \left. + \sigma\phi\left(\frac{\lambda\sigma^2-\mathrm{d_g}}{\sigma}\right) + \mathrm{d_g}\bar{\Phi}\left(\frac{\lambda\sigma^2-\mathrm{d_g}}{\sigma}\right) + \sigma^2\lambda\Phi\left(\frac{\lambda\sigma^2-\mathrm{d_g}}{\sigma}\right)\right)^2. \qquad (17)
\end{aligned}
$$

C

In this part, the mean and variance of the elastic-net density, which is mentioned in Section 5.6, will be computed and the mean is given by

$$
\begin{aligned}
\mathrm{E}(\mathrm{d_f}) &= \int \mathrm{d_f} p(\mathrm{d_f}|\kappa,\gamma) d\mathrm{d_f} \\
&= \int_{-\infty}^{\infty} \mathrm{d_f} \frac{1}{\frac{2\sqrt{\pi}}{\sqrt{A}} \exp\left\{\frac{1}{4A}B^2\right\}\left(1-\Phi(\frac{\sqrt{2A}B}{2A})\right)} \exp\left\{-(A{\mathrm{d_f}}^2 + B|\mathrm{d_f}|)\right\} d\mathrm{d_f} \\
&= \frac{1}{\frac{2\sqrt{\pi}}{\sqrt{A}} \exp\left\{\frac{1}{4A}B^2\right\}\left(1-\Phi(\frac{\sqrt{2A}B}{2A})\right)} \Bigg(\int_{-\infty}^{0} \mathrm{d_f} \exp\left\{-(A{\mathrm{d_f}}^2 + B|\mathrm{d_f}|)\right\} d\mathrm{d_f} \\
&\qquad + \int_{0}^{\infty} \mathrm{d_f} \exp\left\{-(A{\mathrm{d_f}}^2 + B|\mathrm{d_f}|)\right\} d\mathrm{d_f} \Bigg) \\
&= \frac{1}{\frac{2\sqrt{\pi}}{\sqrt{A}} \exp\left\{\frac{1}{4A}B^2\right\}\left(1-\Phi(\frac{\sqrt{2A}B}{2A})\right)} \Bigg(\int_{-\infty}^{0} \mathrm{d_f} \exp\left\{-A({\mathrm{d_f}}^2 - \frac{B}{A}\mathrm{d_f})\right\} d\mathrm{d_f} \\
&\qquad + \int_{0}^{\infty} \mathrm{d_f} \exp\left\{-A({\mathrm{d_f}}^2 + \frac{B}{A}\mathrm{d_f})\right\} d\mathrm{d_f} \Bigg) \\
&= \frac{1}{\frac{2\sqrt{2\pi A}}{\sqrt{A}} \exp\left\{\frac{1}{4A}B^2\right\}\left(1-\Phi(\sqrt{2A}\frac{B}{2A})\right)} \exp\left\{\frac{1}{4A}B^2\right\} \\
&\qquad \Bigg(\int_{-\infty}^{-\sqrt{2A}\frac{B}{2A}} (\frac{y}{\sqrt{2A}} + \frac{B}{2A}) \exp\left\{-\frac{1}{2}y^2\right\} dy + \int_{\sqrt{2A}\frac{B}{2A}}^{\infty} (\frac{y}{\sqrt{2A}} - \frac{B}{2A}) \exp\left\{-\frac{1}{2}y^2\right\} dy \Bigg)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{\frac{2\sqrt{\pi}}{\sqrt{A}} \exp\left\{\frac{1}{4A}B^2\right\}\left(1-\Phi(\sqrt{2A}\frac{B}{2A})\right)} \frac{1}{\sqrt{2A}} \exp\left\{\frac{1}{4A}B^2\right\} \\
&\quad \times \left[\left(\int_{-\infty}^{-\sqrt{2A}\frac{B}{2A}} \frac{y}{\sqrt{2A}} \exp\left\{-\frac{1}{2}y^2\right\}dy + \frac{B}{2A}\int_{-\infty}^{-\sqrt{2A}\frac{B}{2A}} \exp\left\{-\frac{1}{2}y^2\right\}dy\right)\right. \\
&\quad \left. + \left(\int_{\sqrt{2A}\frac{B}{2A}}^{\infty} \frac{y}{\sqrt{2A}} \exp\left\{-\frac{1}{2}y^2\right\}dy - \frac{B}{2A}\int_{\sqrt{2A}\frac{B}{2A}}^{\infty} \exp\left\{-\frac{1}{2}y^2\right\}dy\right)\right] \\
&= \frac{1}{\left(1-\Phi(\frac{\kappa(1-\gamma)}{\sqrt{2\kappa\gamma}})\right)}\left(\frac{-1}{\sqrt{2\kappa\gamma}}\phi(\frac{\kappa(1-\gamma)}{\sqrt{2\kappa\gamma}}) + \frac{(1-\gamma)}{2\gamma}\left(1-\Phi(\frac{\kappa(1-\gamma)}{\sqrt{2\kappa\gamma}})\right)\right. \\
&\quad \left. + \frac{1}{\sqrt{2\kappa\gamma}}\phi(\frac{\kappa(1-\gamma)}{\sqrt{2\kappa\gamma}}) - \frac{(1-\gamma)}{2\gamma}\left(1-\Phi(\frac{\kappa(1-\gamma)}{\sqrt{2\kappa\gamma}})\right)\right) = 0, \qquad (18)
\end{aligned}
$$

and the $\mathrm{E}({\mathrm{d_f}}^2)$, is given by

$$
\begin{aligned}
\mathrm{E}({\mathrm{d_f}}^2) &= \int {\mathrm{d_f}}^2 \mathrm{p}(\mathrm{d_f}|\kappa,\gamma) d\mathrm{d_f} \\
&= \int_{-\infty}^{\infty} {\mathrm{d_f}}^2 \frac{1}{\frac{2\sqrt{\pi}}{\sqrt{A}} \exp\left\{\frac{1}{4A}B^2\right\}\left(1-\Phi(\frac{\sqrt{2A}B}{2A})\right)} \exp\left\{-(A{\mathrm{d_f}}^2 + B|\mathrm{d_f}|)\right\} d\mathrm{d_f} \\
&= \frac{1}{\frac{2\sqrt{\pi}}{\sqrt{A}} \exp\left\{\frac{1}{4A}B^2\right\}\left(1-\Phi(\frac{\sqrt{2A}B}{2A})\right)} \left(\int_{-\infty}^{0} {\mathrm{d_f}}^2 \exp\left\{-(A{\mathrm{d_f}}^2 + B|\mathrm{d_f}|)\right\} d\mathrm{d_f}\right. \\
&\quad \left. + \int_{0}^{\infty} {\mathrm{d_f}}^2 \exp\left\{-(A{\mathrm{d_f}}^2 + B|\mathrm{d_f}|)\right\} d\mathrm{d_f}\right) \\
&= \frac{1}{\frac{2\sqrt{\pi}}{\sqrt{A}} \exp\left\{\frac{1}{4A}B^2\right\}\left(1-\Phi(\frac{\sqrt{2A}B}{2A})\right)} \left(\int_{-\infty}^{0} {\mathrm{d_f}}^2 \exp\left\{-A({\mathrm{d_f}}^2 - \frac{B}{A}\mathrm{d_f})\right\} d\mathrm{d_f}\right. \\
&\quad \left. + \int_{0}^{\infty} {\mathrm{d_f}}^2 \exp\left\{-A({\mathrm{d_f}}^2 + \frac{B}{A}\mathrm{d_f})\right\} d\mathrm{d_f}\right) \\
&= \frac{1}{\frac{2\sqrt{2\pi A}}{\sqrt{A}} \exp\left\{\frac{1}{4A}B^2\right\}\left(1-\Phi(\sqrt{2A}\frac{B}{2A})\right)} \exp\left\{\frac{1}{4A}B^2\right\} \\
&\quad \left(\int_{-\infty}^{-\sqrt{2A}\frac{B}{2A}} (\frac{y}{\sqrt{2A}} + \frac{B}{2A})^2 \exp\left\{-\frac{1}{2}y^2\right\}dy + \int_{\sqrt{2A}\frac{B}{2A}}^{\infty} (\frac{y}{\sqrt{2A}} - \frac{B}{2A})^2 \exp\left\{-\frac{1}{2}y^2\right\}dy\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{\frac{2\sqrt{2\pi A}}{\sqrt{A}} \exp\left\{\frac{1}{4A}B^2\right\}\left(1-\Phi(\sqrt{2A}\frac{B}{2A})\right)} \exp\left\{\frac{1}{4A}B^2\right\} \\
&\quad \left(\int_{-\infty}^{-\sqrt{2A}\frac{B}{2A}} \left(\frac{y^2}{2A} + 2\frac{B}{2A\sqrt{2A}}y + (\frac{B}{2A})^2\right)\exp\left\{-\frac{1}{2}y^2\right\}dy \right. \\
&\quad \left. + \int_{\sqrt{2A}\frac{B}{2A}}^{\infty} \left(\frac{y^2}{2A} - 2\frac{B}{2A\sqrt{2A}}y + (\frac{B}{2A})^2\right)\exp\left\{-\frac{1}{2}y^2\right\}dy\right) \\
&= \frac{2}{\frac{2\sqrt{2\pi A}}{\sqrt{A}} \exp\left\{\frac{1}{4A}B^2\right\}\left(1-\Phi(\sqrt{2A}\frac{B}{2A})\right)} \exp\left\{\frac{1}{4A}B^2\right\} \\
&\quad \times \left(\frac{\sqrt{2\pi}}{2A}\left(\frac{\sqrt{2A}B}{2A}\phi(\frac{\sqrt{2A}B}{2A}) + \left(1-\Phi(\frac{\sqrt{2A}B}{2A})\right)\right) + \sqrt{2\pi}(\frac{B}{2A})^2\left(1-\Phi(\frac{\sqrt{2A}B}{2A})\right)\right).
\end{aligned} \tag{19}
$$

So, the variance is given by

$$
\begin{aligned}
\mathrm{V}(\mathrm{d_f}) &= \mathrm{E}(\mathrm{d_f}^2) - (\mathrm{E}(\mathrm{d_f}))^2 \\
&= \mathrm{E}(\mathrm{d_f}^2) - 0 \\
&= \frac{2}{\frac{2\sqrt{2\pi A}}{\sqrt{A}} \exp\left\{\frac{1}{4A}B^2\right\}\left(1-\Phi(\sqrt{2A}\frac{B}{2A})\right)} \exp\left\{\frac{1}{4A}B^2\right\} \\
&\quad \times \left(\frac{\sqrt{2\pi}}{2A}\left(\frac{\sqrt{2A}B}{2A}\phi(\frac{\sqrt{2A}B}{2A}) + \left(1-\Phi(\frac{\sqrt{2A}B}{2A})\right)\right) + \sqrt{2\pi}(\frac{B}{2A})^2\left(1-\Phi(\frac{\sqrt{2A}B}{2A})\right)\right).
\end{aligned} \tag{20}
$$

The joint distribution of $\mathrm{p}(\mathrm{d_y}, \mathrm{d_g})$, is given by

$$
\begin{aligned}
\mathrm{p}(\mathrm{d_y}, \mathrm{d_g}) = &\frac{1}{\frac{2\sqrt{\pi}}{\sqrt{\kappa\gamma}} \exp\left\{\frac{(\kappa(1-\gamma))^2}{4\kappa\gamma}\right\}\left(1-\Phi\left(\frac{\kappa(1-\gamma)}{\sqrt{2\kappa\gamma}}\right)\right)} \frac{1}{\sqrt{2\pi\sigma^2}} \\
&\times \exp\left\{-\frac{1}{2\sigma^2}(\mathrm{d_y}-\mathrm{d_g})^2\right\} \exp\left\{-\kappa(\gamma{\mathrm{d_g}}^2 + (1-\gamma)|\mathrm{d_g}|)\right\}
\end{aligned}
$$

$$
\begin{aligned}
=&\frac{1}{\frac{2\sqrt{\pi}}{\sqrt{\kappa\gamma}}\exp\left\{\frac{(\kappa(1-\gamma))^2}{4\kappa\gamma}\right\}\left(1-\Phi\left(\frac{\kappa(1-\gamma)}{\sqrt{2\kappa\gamma}}\right)\right)}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}{\mathrm{d_y}}^2\right\}\\
&\times\exp\left\{-\frac{1}{2\sigma^2}({\mathrm{d_g}}^2(1+2\sigma^2\kappa\gamma)-2\mathrm{d_g}(\mathrm{d_y}-\sigma^2\kappa(1-\gamma)\mathrm{sign}(\mathrm{d_g})))\right\}\\
=&\frac{1}{\frac{2\pi\sqrt{2}\sigma}{\sqrt{\kappa\gamma}}\exp\left\{\frac{(\kappa(1-\gamma))^2}{4\kappa\gamma}\right\}\left(1-\Phi\left(\frac{\kappa(1-\gamma)}{\sqrt{2\kappa\gamma}}\right)\right)}\exp\left\{-\frac{1}{2\sigma^2}{\mathrm{d_y}}^2\right\}\\
&\times\exp\left\{-\left(\frac{1+2\sigma^2\kappa\gamma}{2\sigma^2}\right)\left({\mathrm{d_g}}^2-2\mathrm{d_g}\frac{(\mathrm{d_y}-2\sigma^2\kappa(1-\gamma)\mathrm{sign}(\mathrm{d_g}))}{(1+2\sigma^2\kappa\gamma)}\right)\right\}\\
=&\frac{1}{\frac{2\pi\sqrt{2}\sigma}{\sqrt{\kappa\gamma}}\exp\left\{\frac{(\kappa(1-\gamma))^2}{4\kappa\gamma}\right\}\left(1-\Phi\left(\frac{\kappa(1-\gamma)}{\sqrt{2\kappa\gamma}}\right)\right)}\exp\left\{-\frac{1}{2\sigma^2}{\mathrm{d_y}}^2\right\}\\
&\times\exp\left\{\left(\frac{1+2\sigma^2\kappa\gamma}{2\sigma^2}\right)\left(\frac{\mathrm{d_y}-2\sigma^2\kappa(1-\gamma)\mathrm{sign}(\mathrm{d_g})}{1+2\sigma^2\kappa\gamma}\right)^2\right\}\\
&\times\exp\left\{-\left(\frac{1+2\sigma^2\kappa\gamma)}{2\sigma^2}\right\}\left(\mathrm{d_g}-\frac{\mathrm{d_y}-2\sigma^2\kappa(1-\gamma)\mathrm{sign}(\mathrm{d_g})}{1+2\sigma^2\kappa\gamma}\right)^2\right)\\
=&\begin{cases}
\frac{1}{\frac{2\pi\sqrt{2}\sigma}{\sqrt{\kappa\gamma}}\exp\left\{\frac{(\kappa(1-\gamma))^2}{4\kappa\gamma}\right\}\left(1-\Phi\left(\frac{\kappa(1-\gamma)}{\sqrt{2\kappa\gamma}}\right)\right)}\exp\left\{-\frac{1}{2\sigma^2}{\mathrm{d_y}}^2\right\}\quad\exp\left\{\left(\frac{1+2\sigma^2\kappa\gamma}{2\sigma^2}\right)\left(\frac{\mathrm{d_y}-2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma}\right)^2\right\}\\
\qquad\times\exp\left\{-\left(\frac{(1+2\sigma^2\kappa\gamma)}{2\sigma^2}\right)\left(\mathrm{d_g}-\frac{\mathrm{d_y}-2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma}\right)^2\right\}, & \text{if } \mathrm{d_g}\geq 0\\
\frac{1}{\frac{2\pi\sqrt{2}\sigma}{\sqrt{\kappa\gamma}}\exp\left\{\frac{(\kappa(1-\gamma))^2}{4\kappa\gamma}\right\}\left(1-\Phi\left(\frac{\kappa(1-\gamma)}{\sqrt{2\kappa\gamma}}\right)\right)}\exp\left\{-\frac{1}{2\sigma^2}{\mathrm{d_y}}^2\right\}\quad\exp\left\{\left(\frac{1+2\sigma^2\kappa\gamma}{2\sigma^2}\right)\left(\frac{\mathrm{d_y}+2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma}\right)^2\right\}\\
\qquad\times\exp\left\{-\left(\frac{(1+2\sigma^2\kappa\gamma)}{2\sigma^2}\right)\left(\mathrm{d_g}-\frac{\mathrm{d_y}+2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma}\right)^2\right\}, & \text{if } \mathrm{d_g}<0,
\end{cases}
\end{aligned}
\tag{21}
$$

the marginal distribution is given by

$$
\begin{aligned}
\mathrm{m}(\mathrm{d_y})=&\int \mathrm{p}(\mathrm{d_y},\mathrm{d_g})d\mathrm{d_g}\\
=&\frac{1}{\frac{2\pi\sqrt{2}\sigma}{\sqrt{\kappa\gamma}}\exp\left\{\frac{(\kappa(1-\gamma))^2}{4\kappa\gamma}\right\}\left(1-\Phi(\frac{\kappa(1-\gamma)}{\sqrt{2\kappa\gamma}})\right)}\exp\left\{-\frac{1}{2\sigma^2}{\mathrm{d_y}}^2\right\}
\end{aligned}
$$

$$
\begin{aligned}
&\times \left(\exp\left\{ \left(\frac{1+2\sigma^2\kappa\gamma}{2\sigma^2} \right) \left(\frac{\mathrm{d_y} + 2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma} \right)^2 \right\} \right. \\
&\times \int_{-\infty}^{0} \exp\left\{ -\left(\frac{1+2\sigma^2\kappa\gamma}{2\sigma^2} \right) \left(\mathrm{d_g} - \frac{\mathrm{d_y} + 2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma} \right)^2 \right\} d\mathrm{d_g} \\
&+ \exp\left\{ \left(\frac{1+2\sigma^2\kappa\gamma}{2\sigma^2} \right) \left(\frac{\mathrm{d_y} - 2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma} \right)^2 \right\} \\
&\times \left. \int_{0}^{\infty} \exp\left\{ -\left(\frac{1+2\sigma^2\kappa\gamma}{2\sigma^2} \right) \left(\mathrm{d_g} - \frac{\mathrm{d_y} - 2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma} \right)^2 \right\} d\mathrm{d_g} \right) \\
=&\frac{1}{\frac{2\pi\sqrt{2}\sigma}{\sqrt{\kappa\gamma}} \exp\left\{ \frac{(\kappa(1-\gamma))^2}{4\kappa\gamma} \right\} \left(1 - \Phi\left(\frac{\kappa(1-\gamma)}{\sqrt{2\kappa\gamma}} \right) \right)} \exp\left\{ -\frac{1}{2\sigma^2} {\mathrm{d_y}}^2 \right\} \\
&\times \left(\exp\left\{ \left(\frac{1+2\sigma^2\kappa\gamma}{2\sigma^2} \right) \left(\frac{\mathrm{d_y} + 2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma} \right)^2 \right\} \right. \\
&\times \sqrt{\frac{2\pi\sigma^2}{1+2\sigma^2\kappa\gamma}} \left(1 - \Phi\left(\sqrt{\frac{1+2\sigma^2\kappa\gamma}{\sigma^2}} \frac{\mathrm{d_y} + 2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma} \right) \right) \\
&+ \exp\left\{ \left(\frac{1+2\sigma^2\kappa\gamma}{2\sigma^2} \right) \left(\frac{\mathrm{d_y} - 2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma} \right)^2 \right\} \\
&\times \left. \sqrt{\frac{2\pi\sigma^2}{1+2\sigma^2\kappa\gamma}} \left(\Phi\left(\sqrt{\frac{1+2\sigma^2\kappa\gamma}{\sigma^2}} \frac{\mathrm{d_y} - 2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma} \right) \right) \right),
\end{aligned}
\tag{22}
$$

and the distribution of $\mathrm{d_g}$, can be written as

$$
\mathrm{p}(\mathrm{d_g}|\mathrm{d_y}) = \frac{\mathrm{p}(\mathrm{d_y}, \mathrm{d_g})}{\mathrm{m}(\mathrm{d_y})}
$$

$$
= \begin{cases}
\dfrac{\exp\left\{ \left(\frac{\mathrm{d_y} - 2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma} \right)^2 \right\} \exp\left\{ -\left(\frac{1+2\sigma^2\kappa\gamma)}{2\sigma^2} \right\} \left(\mathrm{d_g} - \frac{\mathrm{d_y} - 2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma} \right)^2 \right)}{\sqrt{\frac{2\pi\sigma^2}{1+2\sigma^2\kappa\gamma}} \left(\exp\left\{ \left(\frac{\mathrm{d_y} + 2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma} \right)^2 \right\} \left(1 - \Phi\left(\sqrt{\frac{1+2\sigma^2\kappa\gamma}{\sigma^2}} \frac{\mathrm{d_y} + 2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma} \right) \right) + \exp\left\{ \left(\frac{\mathrm{d_y} - 2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma} \right)^2 \right\} \left(\Phi\left(\sqrt{\frac{1+2\sigma^2\kappa\gamma}{\sigma^2}} \frac{\mathrm{d_y} - 2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma} \right) \right) \right)}, & \text{if } \mathrm{d_g} \geq 0 \\
\dfrac{\exp\left\{ \left(\frac{\mathrm{d_y} + 2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma} \right)^2 \right\} \exp\left\{ -\left(\frac{1+2\sigma^2\kappa\gamma)}{2\sigma^2} \right\} \left(\mathrm{d_g} - \frac{\mathrm{d_y} + 2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma} \right)^2 \right)}{\sqrt{\frac{2\pi\sigma^2}{1+2\sigma^2\kappa\gamma}} \left(\exp\left\{ \left(\frac{\mathrm{d_y} + 2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma} \right)^2 \right\} \left(1 - \Phi\left(\sqrt{\frac{1+2\sigma^2\kappa\gamma}{\sigma^2}} \frac{\mathrm{d_y} + 2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma} \right) \right) + \exp\left\{ \left(\frac{\mathrm{d_y} - 2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma} \right)^2 \right\} \left(\Phi\left(\sqrt{\frac{1+2\sigma^2\kappa\gamma}{\sigma^2}} \frac{\mathrm{d_y} - 2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma} \right) \right) \right)}, & \text{if } \mathrm{d_g} < 0.
\end{cases}
\tag{23}
$$

Now, the posterior mean PM can be computed as

$$\begin{aligned}
\int \mathrm{d_g p(d_g|d_y)} d\mathrm{d_g} =& \frac{1}{\frac{2\pi\sqrt{2}\sigma}{\sqrt{\kappa\gamma}} \exp\left\{\frac{(\kappa(1-\gamma))^2}{4\kappa\gamma}\right\}\left(1-\Phi\left(\frac{\kappa(1-\gamma)}{\sqrt{2\kappa\gamma}}\right)\right)} \exp\left\{-\frac{1}{2\sigma^2}\mathrm{d_y}^2\right\} \\
&\times \left(\exp\left\{\left(\frac{1+2\sigma^2\kappa\gamma}{2\sigma^2}\right)\left(\frac{\mathrm{d_y}+2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma}\right)^2\right\}\right. \\
&\times \int_{-\infty}^{0} \mathrm{d_g} \exp\left\{-\left(\frac{1+2\sigma^2\kappa\gamma}{2\sigma^2}\right)\right. \\
&\left.\times \left(\mathrm{d_g}-\frac{\mathrm{d_y}+2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma}\right)^2\right\} d\mathrm{d_g} \\
&+ \exp\left\{\left(\frac{1+2\sigma^2\kappa\gamma}{2\sigma^2}\right)\left(\frac{\mathrm{d_y}-2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma}\right)^2\right\} \\
&\times \int_{0}^{\infty} \mathrm{d_g} \exp\left\{-\left(\frac{1+2\sigma^2\kappa\gamma}{2\sigma^2}\right)\right. \\
&\left.\left.\times \left(\mathrm{d_g}-\frac{\mathrm{d_y}-2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma}\right)^2\right\} d\mathrm{d_g}\right) \\
=& \frac{1}{\frac{2\pi\sqrt{2}\sigma}{\sqrt{\kappa\gamma}} \exp\left\{\frac{(\kappa(1-\gamma))^2}{4\kappa\gamma}\right\}\left(1-\Phi\left(\frac{\kappa(1-\gamma)}{\sqrt{2\kappa\gamma}}\right)\right)} \exp\left\{-\frac{1}{2\sigma^2}\mathrm{d_y}^2\right\} \\
&\times \exp\left\{\frac{1+2\sigma^2\kappa\gamma}{2\sigma^2}\right\}\left(\exp\left\{\left(\frac{\mathrm{d_y}+2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma}\right)^2\right\}\right. \\
&\times \int_{-\infty}^{0} \mathrm{d_g} \exp\left\{-\left(\mathrm{d_g}-\frac{\mathrm{d_y}+2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma}\right)^2\right\} d\mathrm{d_g} \\
&+ \exp\left\{\left(\frac{\mathrm{d_y}-2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma}\right)^2\right\} \\
&\times \left.\int_{0}^{\infty} \mathrm{d_g} \exp\left\{-\left(\mathrm{d_g}-\frac{\mathrm{d_y}-2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma}\right)^2\right\} d\mathrm{d_g}\right).
\end{aligned}$$

Then let $y = \sqrt{2}\frac{(\mathrm{d_y}+2\sigma^2\kappa(1-\gamma))}{1+2\sigma^2\kappa\gamma}$, given

$$\begin{aligned}
&\int \mathrm{d_g p(d_g|d_y)} d\mathrm{d_g} \\
&\qquad = \frac{1}{\frac{2\pi\sqrt{2}\sigma}{\sqrt{\kappa\gamma}} \exp\left\{\frac{(\kappa(1-\gamma))^2}{4\kappa\gamma}\right\}\left(1-\Phi\left(\frac{\kappa(1-\gamma)}{\sqrt{2\kappa\gamma}}\right)\right)} \exp\left\{-\frac{1}{2\sigma^2}\mathrm{d_y}^2\right\}
\end{aligned}$$

$$
\begin{aligned}
&\times \exp\left\{\frac{1+2\sigma^2\kappa\gamma}{2\sigma^2}\right\}\Bigg(\exp\left\{\left(\frac{\mathrm{d_y}+2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma}\right)^2\right\}\\
&\times \int_{-\infty}^{-\sqrt{2}\frac{(\mathrm{d_y}+2\sigma^2\kappa(1-\gamma))}{1+2\sigma^2\kappa\gamma}} \frac{1}{\sqrt{2}}\left(\frac{y}{\sqrt{2}}+\frac{\mathrm{d_y}+2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma}\right)\\
&\times \exp\left\{-\frac{1}{2}y^2\right\}dy + \exp\left\{\left(\frac{\mathrm{d_y}-2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma}\right)^2\right\}\\
&\times \int_{-\sqrt{2}\frac{(\mathrm{d_y}-2\sigma^2\kappa(1-\gamma))}{1+2\sigma^2\kappa\gamma}}^{\infty} \frac{1}{\sqrt{2}}\left(\frac{y}{\sqrt{2}}+\frac{\mathrm{d_y}-2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma}\right)\\
&\times \exp\left\{-\frac{1}{2}y^2\right\}dy\Bigg)\\
=\ &\frac{1}{\frac{2\pi\sqrt{2}\sigma}{\sqrt{\kappa\gamma}}\exp\left\{\frac{(\kappa(1-\gamma))^2}{4\kappa\gamma}\right\}\left(1-\Phi(\frac{\kappa(1-\gamma)}{\sqrt{2\kappa\gamma}})\right)}\exp\left\{-\frac{1}{2\sigma^2}{\mathrm{d_y}}^2\right\}\exp\left\{\frac{1+2\sigma^2\kappa\gamma}{2\sigma^2}\right\}\\
&\times\Bigg(\exp\left\{\left(\frac{\mathrm{d_y}+2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma}\right)^2\right\}\Bigg(-\frac{\sqrt{2\pi}}{2}\phi\left(\sqrt{2}\frac{(\mathrm{d_y}+2\sigma^2\kappa(1-\gamma))}{1+2\sigma^2\kappa\gamma}\right)\\
&+\sqrt{\frac{2\pi}{2}}\left(\frac{\mathrm{d_y}+2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma}\right)\left(1+\Phi\left(\frac{\mathrm{d_y}+2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma}\right)\right)\Bigg)\\
&+\exp\left\{\left(\frac{\mathrm{d_y}-2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma}\right)^2\right\}\left(\frac{\sqrt{2\pi}}{2}\phi\left(\sqrt{2}\frac{(\mathrm{d_y}-2\sigma^2\kappa(1-\gamma))}{1+2\sigma^2\kappa\gamma}\right)\right.\\
&\left.+\sqrt{\frac{2\pi}{2}}\left(\frac{\mathrm{d_y}-2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma}\right)\Phi\left(\frac{\mathrm{d_y}-2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma}\right)\right)\Bigg). \qquad (24)
\end{aligned}
$$

Thus, the posterior mean can be computed using (22) and (24), as

$$
\begin{aligned}
\mathrm{PM}(\mathrm{d_g}|\mathrm{d_y}) =&\frac{\int \mathrm{d_g}\mathrm{p}(\mathrm{d_g}|\mathrm{d_y})d\mathrm{d_g}}{\int \mathrm{p}(\mathrm{d_g}|\mathrm{d_y})d\mathrm{d_g}}\\
=\ &\Bigg(\exp\left\{\left(\frac{\mathrm{d_y}+2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma}\right)^2\right\}\Bigg(-\frac{\sqrt{2\pi}}{2}\phi\left(\sqrt{2}\frac{(\mathrm{d_y}+2\sigma^2\kappa(1-\gamma))}{1+2\sigma^2\kappa\gamma}\right)\\
&+\sqrt{\frac{2\pi}{2}}\left(\frac{\mathrm{d_y}+2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma}\right)\left(1+\Phi\left(\frac{\mathrm{d_y}+2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma}\right)\right)\Bigg)\\
&+\exp\left\{\left(\frac{\mathrm{d_y}-2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma}\right)^2\right\}\left(\frac{\sqrt{2\pi}}{2}\phi\left(\sqrt{2}\frac{(\mathrm{d_y}-2\sigma^2\kappa(1-\gamma))}{1+2\sigma^2\kappa\gamma}\right)\right.\\
&\left.+\sqrt{\frac{2\pi}{2}}\left(\frac{\mathrm{d_y}-2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma}\right)\Phi\left(\frac{\mathrm{d_y}-2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma}\right)\right)\Bigg)
\end{aligned}
$$

$$
\Bigg/
$$

$$
\begin{aligned}
&\left(\exp\left\{ \left(\frac{1+2\sigma^2\kappa\gamma}{2\sigma^2} \right) \left(\frac{\mathrm{d_y} + 2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma} \right)^2 \right\} \right. \\
&\times \sqrt{\frac{2\pi\sigma^2}{1+2\sigma^2\kappa\gamma}} \left(1 - \Phi\left(\sqrt{\frac{1+2\sigma^2\kappa\gamma}{\sigma^2}} \frac{\mathrm{d_y} + 2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma} \right) \right) \\
&+ \exp\left\{ \left(\frac{1+2\sigma^2\kappa\gamma}{2\sigma^2} \right) \left(\frac{\mathrm{d_y} - 2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma} \right)^2 \right\} \\
&\left. \times \sqrt{\frac{2\pi\sigma^2}{1+2\sigma^2\kappa\gamma}} \left(\Phi\left(\sqrt{\frac{1+2\sigma^2\kappa\gamma}{\sigma^2}} \frac{\mathrm{d_y} - 2\sigma^2\kappa(1-\gamma)}{1+2\sigma^2\kappa\gamma} \right) \right) \right)
\end{aligned}
$$

. (25)

D

The posterior mean PM of the Bayesian adaptive multi-resolution shrinkage (BAMS) method will be computed, Vidakovic and Ruggeri, (2001) state that the marginal distribution for $\mathrm{d_y}$ is

$$\begin{aligned}
\mathrm{m}(\mathrm{d_y}) &= \int \mathrm{p}(\mathrm{d_y}|\mathrm{d_g})\zeta(\mathrm{d_g})d\mathrm{d_g} \\
&= \int \mathcal{DE}(\mathrm{d_g}, \frac{1}{\sqrt{2\mu}})\mathcal{DE}(0,\tau)d\mathrm{d_g} \\
&= \frac{\sqrt{2\mu}}{4\tau}[\int_{-\infty}^{0} \exp\{-\sqrt{2\mu}(\mathrm{d_y}-\mathrm{d_g}) + \mathrm{d_g}/\tau\}d\mathrm{d_g} \\
&\quad + \int_{0}^{d} \exp\{-\sqrt{2\mu}(\mathrm{d_y}-\mathrm{d_g}) - \mathrm{d_g}/\tau\}d\mathrm{d_g} \\
&\quad + \int_{0}^{\infty} \exp\{\sqrt{2\mu}(\mathrm{d_y}-\mathrm{d_g}) - \mathrm{d_g}/\tau\}d\mathrm{d_g}] \\
&= \frac{\sqrt{2\mu}}{4\tau}[\exp\{-\sqrt{2\mu}\mathrm{d_y}\}\int_{-\infty}^{0} \exp\{\sqrt{2\mu}\mathrm{d_g} + \mathrm{d_g}/\tau\}d\mathrm{d_g} + \exp\{-\sqrt{2\mu}\mathrm{d_y}\} \\
&\quad \times \int_{0}^{\mathrm{d_y}} \exp\{\sqrt{2\mu}\mathrm{d_g} - \mathrm{d_g}/\tau\}d\mathrm{d_g} \\
&\quad + \exp\{\sqrt{2\mu}\mathrm{d_y}\}\int_{0}^{\infty} \exp\{-\sqrt{2\mu}\mathrm{d_g} - \mathrm{d_g}/\tau\}d\mathrm{d_g}] \\
&= \frac{t\exp\{-|\mathrm{d_y}|/\tau\} - \frac{1}{\sqrt{2\mu}}\exp\{-|\mathrm{d_y}|\sqrt{2\mu}\}}{2\tau^2 - 1/\mu}, \quad \tau \neq \frac{1}{\sqrt{2\mu}}.
\end{aligned} \tag{26}$$

To prove that step by step

$$\begin{aligned}
\mathrm{m}(\mathrm{d_y}) &= \int \mathrm{P}(\mathrm{d_y}|\mathrm{d_g})\zeta(\mathrm{d_g})d\mathrm{d_g} \\
&= \int \mathcal{DE}(\mathrm{d_g}, \frac{1}{\sqrt{2\mu}})\mathcal{DE}(0,\tau^2)d\mathrm{d_g}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\sqrt{2\mu}}{4\tau}\Big[\int_{-\infty}^{0}\exp\{-\sqrt{2\mu}(\mathrm{d_y}-\mathrm{d_g})+\mathrm{d_g}/\tau\}d\mathrm{d_g}+\int_{0}^{\mathrm{d_y}}\exp\{-\sqrt{2\mu}(\mathrm{d_y}-\mathrm{d_g})-\mathrm{d_g}/\tau\}d\mathrm{d_g}\\
&\quad+\int_{0}^{\infty}\exp\{\sqrt{2\mu}(\mathrm{d_y}-\mathrm{d_g})-\mathrm{d_g}/\tau\}d\mathrm{d_g}\Big]\\
&= \frac{\sqrt{2\mu}}{4\tau}\Big[\exp\{-\sqrt{2\mu}\mathrm{d_y}\}\\
&\quad\times\int_{-\infty}^{0}\exp\{\sqrt{2\mu}\mathrm{d_g}+\mathrm{d_g}/\tau\}d\mathrm{d_g}+\exp\{-\sqrt{2\mu}\mathrm{d_y}\}\int_{0}^{\mathrm{d_y}}\exp\{\sqrt{2\mu}\mathrm{d_g}-\mathrm{d_g}/\tau\}d\mathrm{d_g}\\
&\quad+\exp\{\sqrt{2\mu}\mathrm{d_y}\}\int_{0}^{\infty}\exp\{-\sqrt{2\mu}\mathrm{d_g}-\mathrm{d_g}/\tau\}d\mathrm{d_g}\Big]\\
&= \frac{\sqrt{2\mu}}{4\tau}\Big[\frac{\exp\{-\sqrt{2\mu}\mathrm{d_y}\}}{\sqrt{2\mu}+1/\tau}\exp\{\mathrm{d_g}(\sqrt{2\mu}+1/\tau)\}|_{-\infty}^{0}+\frac{\exp\{-\sqrt{2\mu}\mathrm{d_y}\}}{\sqrt{2\mu}-1/\tau}\exp\{\mathrm{d_g}(\sqrt{2\mu}-1/\tau)\}|_{0}^{\mathrm{d_y}}\\
&\quad+\frac{-\exp\{\sqrt{2\mu}\mathrm{d_y}\}}{\sqrt{2\mu}+1/\tau}\exp\{-\mathrm{d_g}(\sqrt{2\mu}+1/\tau)\}|_{\mathrm{d_y}}^{\infty}\Big]\\
&= \frac{\sqrt{2\mu}}{4\tau}\Big[\frac{\exp\{-\sqrt{2\mu}\mathrm{d_y}\}}{\sqrt{2\mu}+1/\tau}+\frac{\exp\{-\sqrt{2\mu}\mathrm{d_y}\}}{\sqrt{2\mu}-1/\tau}(\exp\{\mathrm{d_y}(\sqrt{2\mu}-1/\tau)\}-1)\\
&\quad+\frac{\exp\{\sqrt{2\mu}\mathrm{d_y}\}}{\sqrt{2\mu}+1/\tau}\exp\{-\mathrm{d_y}(\sqrt{2\mu}+1/\tau)\}\Big]\\
&= \frac{\sqrt{2\mu}}{4\tau}\Big[\exp\{-\mathrm{d_y}\sqrt{2\mu}\}(\frac{1}{\sqrt{2\mu}+1/\tau}-\frac{1}{\sqrt{2\mu}-1/\tau})\\
&\quad+\exp\{-\mathrm{d_y}/\tau\}(\frac{1}{\sqrt{2\mu}-1/\tau}+\frac{1}{\sqrt{2\mu}+1/\tau})\Big]\\
&= \frac{\sqrt{2\mu}}{4\tau}\Big[\frac{2\sqrt{2\mu}\exp\{-\mathrm{d_y}/\tau\}-2/\tau\exp\{-\mathrm{d_y}\sqrt{2\mu}\}}{2\mu+1/\tau^2}\Big]\\
&= \frac{4\mu\exp\{-\mathrm{d_y}/\tau\}-\frac{2\sqrt{2\mu}}{\tau}\exp\{-\mathrm{d_y}\sqrt{2\mu}\}}{8\tau\mu+\frac{4}{\tau}}\\
&= \frac{t\exp\{-|\mathrm{d_y}|/\tau\}-\frac{1}{\sqrt{2\mu}}\exp\{-|\mathrm{d_y}|\sqrt{2\mu}\}}{2\tau^2-1/\mu},\quad \tau\neq\frac{1}{\sqrt{2\mu}}.
\end{aligned}
$$

The posterior mean of $\mathrm{d_g}|\mathrm{d_y}$, is given by

$$
\begin{aligned}
\mathrm{PM}(\mathrm{d_g}|\mathrm{d_y}) &= \frac{\int \mathrm{d_g}\mathrm{p}(\mathrm{d_y}|\mathrm{d_g})\pi(\mathrm{d_g})d\mathrm{d_g}}{\int \mathrm{p}(\mathrm{d_y}|\mathrm{d_g})\pi(\mathrm{d_g})d\mathrm{d_g}}\\
&= \frac{\int \mathrm{d_g}\mathrm{p}(\mathrm{d_y}|\mathrm{d_g})\pi(\mathrm{d_g})d\mathrm{d_g}}{\mathrm{m}(\mathrm{d_g})}
\end{aligned}
$$

$$= \frac{\frac{\sqrt{2\mu}}{4\tau}\Big[\int_{-\infty}^{0} \mathrm{d_g} \exp\{-\sqrt{2\mu}(\mathrm{d_y}-\mathrm{d_g})+\mathrm{d_g}/\tau\} d\mathrm{d_g}}{\frac{t\exp\{-|\mathrm{d_y}|/\tau\}-\frac{1}{\sqrt{2\mu}}\exp\{-|\mathrm{d_y}|\sqrt{2\mu}\}}{2\tau^2-1/\mu}} \tag{27}$$

$$+ \frac{\int_{0}^{\mathrm{d_y}} \mathrm{d_g} \exp\{-\sqrt{2\mu}(\mathrm{d_y}-\mathrm{d_g})-\mathrm{d_g}/\tau\} d\mathrm{d_g} + \int_{\mathrm{d_g}}^{\infty} \mathrm{d_g} \exp\{\sqrt{2\mu}(\mathrm{d_y}-\mathrm{d_g})-\mathrm{d_g}/\tau\} d\mathrm{d_g}\Big]}{\frac{t\exp\{-|\mathrm{d_y}|/\tau\}-\frac{1}{\sqrt{2\mu}}\exp\{-|\mathrm{d_y}|\sqrt{2\mu}\}}{2\tau^2-1/\mu}}$$

$$= \frac{\tau(\tau^2-1/(2\mu))\mathrm{d_y}\exp\{-|\mathrm{d_y}|/\tau\} - \tau^2(\exp\{-|\mathrm{d_y}|\sqrt{2\mu}\}-\exp\{-|\mathrm{d_y}|/\tau\})/\mu}{(\tau^2-1/2\mu)(\tau\exp\{-|\mathrm{d_y}|/\tau\}-(1/\sqrt{2\mu})\exp\{-|\mathrm{d_y}|\sqrt{2\mu}\})}, \quad \tau \neq \frac{1}{\sqrt{2\mu}}. \tag{28}$$

To prove this step by step

$$\mathrm{PM}(\mathrm{d_g}|\mathrm{d_y}) = \frac{\int \mathrm{d_g} \mathrm{p}(\mathrm{d_y}|\mathrm{d_g})\pi(\mathrm{d_g}) d\mathrm{d_g}}{\int \mathrm{p}(\mathrm{d_y}|\mathrm{d_g})\pi(\mathrm{d_g}) d\mathrm{d_g}} = \frac{\int \mathrm{d_g} \mathrm{p}(\mathrm{d_y}|\mathrm{d_g})\pi(\mathrm{d_g}) d\mathrm{d_g}}{\mathrm{m}(\mathrm{d_g})}$$

$$= \frac{\frac{\sqrt{2\mu}}{4\tau}[\int_{-\infty}^{0} \mathrm{d_g} \exp\{-\sqrt{2\mu}(\mathrm{d_y}-\mathrm{d_g})+\mathrm{d_g}/\tau\} d\mathrm{d_g}}{\frac{t\exp\{-|\mathrm{d_g}|/\tau\}-\frac{1}{\sqrt{2\mu}}\exp\{-|\mathrm{d_y}|\sqrt{2\mu}\}}{2\tau^2-1/\mu}}$$

$$+ \frac{\int_{0}^{\mathrm{d_y}} \mathrm{d_g} \exp\{-\sqrt{2\mu}(\mathrm{d_y}-\mathrm{d_g})-\mathrm{d_g}/\tau\} d\mathrm{d_g} + \int_{\mathrm{d_y}}^{\infty} \mathrm{d_g} \exp\{\sqrt{2\mu}(\mathrm{d_y}-\mathrm{d_g})-\mathrm{d_g}/\tau\} d\mathrm{d_g}]}{\frac{t\exp\{-|\mathrm{d_g}|/\tau\}-\frac{1}{\sqrt{2\mu}}\exp\{-|\mathrm{d_y}|\sqrt{2\mu}\}}{2\tau^2-1/\mu}}$$

$$= \frac{\frac{\sqrt{2\mu}}{4\tau}[\exp\{-\sqrt{2\mu}\mathrm{d_y}\}\int_{-\infty}^{0} \mathrm{d_g} \exp\{\sqrt{2\mu}\mathrm{d_g}+\mathrm{d_g}/\tau\} d\mathrm{d_g}}{\frac{t\exp\{-|\mathrm{d_y}|/\tau\}-\frac{1}{\sqrt{2\mu}}\exp\{-|\mathrm{d_y}|\sqrt{2\mu}\}}{2\tau^2-1/\mu}}$$

$$+ \frac{\exp\{-\sqrt{2\mu}\mathrm{d_y}\}\int_{0}^{\mathrm{d_g}} \mathrm{d_g} \exp\{\sqrt{2\mu}\mathrm{d_g}-\frac{\mathrm{d_g}}{\tau}\} d\mathrm{d_g} + \exp\{\sqrt{2\mu}\mathrm{d_y}\}\int_{\mathrm{d_y}}^{\infty} \mathrm{d_g} \exp\{-\sqrt{2\mu}\mathrm{d_g}-\frac{\mathrm{d_g}}{\tau}\} d\mathrm{d_g}]}{\frac{t\exp\{-|\mathrm{d_y}|/\tau\}-\frac{1}{\sqrt{2\mu}}\exp\{-|\mathrm{d_y}|\sqrt{2\mu}\}}{2\tau^2-1/\mu}}$$

$$= \frac{\frac{\sqrt{2\mu}}{4\tau}[-\exp\{-\sqrt{2\mu}\mathrm{d_y}\}\int_{-\infty}^{0} \frac{1}{\sqrt{2\mu}+1/\tau} \exp\{\sqrt{2\mu}\mathrm{d_g}+\mathrm{d_g}/\tau\} d\mathrm{d_g}}{\frac{t\exp\{-|\mathrm{d_y}|/\tau\}-\frac{1}{\sqrt{2\mu}}\exp\{-|\mathrm{d_y}|\sqrt{2\mu}\}}{2\tau^2-1/\mu}}$$

$$+ \frac{\frac{\exp\{-\sqrt{2\mu}\mathrm{d_y}\}}{\sqrt{2\mu}-1/\tau}(\mathrm{d_y}\exp\{\sqrt{2\mu}\mathrm{d_y}-\mathrm{d_y}/\tau\} - \int_{0}^{\mathrm{d_y}} \exp\{\sqrt{2\mu}\mathrm{d_g}-\mathrm{d_g}/\tau\} d\mathrm{d_g})]}{\frac{t\exp\{-|\mathrm{d_y}|/\tau\}-\frac{1}{\sqrt{2\mu}}\exp\{-|\mathrm{d_y}|\sqrt{2\mu}\}}{2\tau^2-1/\mu}}$$

$$+ \frac{\frac{\sqrt{2\mu}}{4\tau}\frac{\exp\{\sqrt{2\mu}\mathrm{d_y}\}}{\sqrt{2\mu}+1/\tau}(\mathrm{d_y}\exp\{-\sqrt{2\mu}\mathrm{d_y}-\mathrm{d_y}/\tau\} + \int_{\mathrm{d_y}}^{\infty} \exp\{-\sqrt{2\mu}\mathrm{d_g}-\mathrm{d_g}/\tau\} d\mathrm{d_g})}{\frac{t\exp\{-|\mathrm{d_y}|/\tau\}-\frac{1}{\sqrt{2\mu}}\exp\{-|\mathrm{d_y}|\sqrt{2\mu}\}}{2\tau^2-1/\mu}}$$

$$= \frac{\frac{\sqrt{2\mu}}{4\tau}\big[\frac{-\exp\{-\mathrm{d_y}\sqrt{2\mu}\}}{(\sqrt{2\mu}+1/\tau)^2} + \frac{\mathrm{d_y}\exp\{-\mathrm{d_y}/\tau\}}{\sqrt{2\mu}-1/\tau} - \frac{\exp\{-\mathrm{d_y}/\tau\}}{(\sqrt{2\mu}-1/\tau)^2} + \frac{\exp\{-\mathrm{d_y}\sqrt{2\mu}\}}{(\sqrt{2\mu}-1/\tau)^2} + \frac{d\exp\{-\mathrm{d_y}/\tau\}}{(\sqrt{2\mu}+1/\tau)} + \frac{\exp\{-\mathrm{d_y}/\tau\}}{(\sqrt{2\mu}+1/\tau)^2}\big]}{\frac{t\exp\{-|\mathrm{d_y}|/\tau\}-\frac{1}{\sqrt{2\mu}}\exp\{-|\mathrm{d_y}|\sqrt{2\mu}\}}{2\tau^2-1/\mu}}$$

$$= \frac{(2\tau^2-\frac{1}{\mu})[2\mu\exp\{-\mathrm{d_y}\sqrt{2\mu}\}+\mu\tau(2\mu-\frac{1}{\tau^2})\mathrm{d_y}\exp\{-\mathrm{d_y}/\tau\}-2\mu\exp\{-\mathrm{d_y}/\tau\}]}{\tau^2(2\mu-\frac{1}{\tau^2})^2(t\exp\{-|\mathrm{d_y}|/\tau\}-\frac{1}{\sqrt{2\mu}}\exp\{-|\mathrm{d_y}|\sqrt{2\mu}\})}$$

$$= \frac{\frac{\tau^2}{\mu}(2\mu-\frac{1}{\tau^2})[2\mu\exp\{-d\sqrt{2\mu}\}+\mu\tau(2\mu-\frac{1}{\tau^2})\mathrm{d_y}\exp\{-\mathrm{d_y}/\tau\}-2\mu\exp\{-\mathrm{d_y}/\tau\}]}{\tau^2(2\mu-\frac{1}{\tau^2})^2(t\exp\{-|\mathrm{d_y}|/\tau\}-\frac{1}{\sqrt{2\mu}}\exp\{-|\mathrm{d_y}|\sqrt{2\mu}\})}$$

$$= \frac{2\exp\{-\mathrm{d_y}\sqrt{2\mu}\}+\tau(2\mu-\frac{1}{\tau^2})\mathrm{d_y}\exp\{-\mathrm{d_y}/\tau\}-2\exp\{-\mathrm{d_y}/\tau\}}{(2\mu-\frac{1}{\tau^2})(t\exp\{-|\mathrm{d_y}|/\tau\}-\frac{1}{\sqrt{2\mu}}\exp\{-|\mathrm{d_y}|\sqrt{2\mu}\})}$$

$$= \frac{2\exp\{-\mathrm{d_y}\sqrt{2\mu}\}+\frac{2\tau\mu}{\tau^2}(\tau^2-\frac{1}{2\mu})\mathrm{d_y}\exp\{-\mathrm{d_y}/\tau\}-2\exp\{-\mathrm{d_y}/\tau\}}{\frac{2\mu}{\tau^2}(\tau^2-\frac{1}{2\mu})(t\exp\{-|\mathrm{d_y}|/\tau\}-\frac{1}{\sqrt{2\mu}}\exp\{-|\mathrm{d_y}|\sqrt{2\mu}\})}$$

$$= \frac{\frac{2\mu}{\tau}(\tau^2-\frac{1}{2\mu})\mathrm{d_y}\exp\{-\mathrm{d_y}/\tau\}+2\exp\{-\mathrm{d_y}\sqrt{2\mu}\}-2\exp\{-\mathrm{d_y}/\tau\}}{\frac{2\mu}{\tau^2}(\tau^2-\frac{1}{2\mu})(t\exp\{-|\mathrm{d_y}|/\tau\}-\frac{1}{\sqrt{2\mu}}\exp\{-|\mathrm{d_y}|\sqrt{2\mu}\})}$$

$$= \frac{2\mu\tau^2(\tau^2-\frac{1}{2\mu})\mathrm{d_y}\exp\{-\mathrm{d_y}/\tau\}+2\exp\{-\mathrm{d_y}\sqrt{2\mu}\}-2\exp\{-\mathrm{d_y}/\tau\}}{2\mu\tau(\tau^2-\frac{1}{2\mu})(t\exp\{-|\mathrm{d_y}|/\tau\}-\frac{1}{\sqrt{2\mu}}\exp\{-|\mathrm{d_y}|\sqrt{2\mu}\})}$$

$$= \frac{\tau(\tau^2-\frac{1}{2\mu})\mathrm{d_y}\exp\{-\mathrm{d_y}/\tau\}+\frac{\tau}{\mu}\exp\{-\mathrm{d_y}\sqrt{2\mu}\}-\frac{\tau}{\mu}\exp\{-\mathrm{d_y}/\tau\}}{(\tau^2-\frac{1}{2\mu})(\tau\exp\{-|\mathrm{d_y}|/\tau\}-\frac{1}{\sqrt{2\mu}}\exp\{-|\mathrm{d_y}|\sqrt{2\mu}\})}$$

$$= \frac{\tau(\tau^2-1/(2\mu))\mathrm{d_y}\exp\{-|\mathrm{d_y}|/\tau\}+\tau^2(\exp\{-|\mathrm{d_y}|\sqrt{2\mu}\}-\exp\{-|\mathrm{d_y}|/\tau\})/\mu}{(\tau^2-1/2\mu)(\tau\exp\{-|\mathrm{d_y}|/\tau\}-(1/\sqrt{2\mu})\exp\{-|\mathrm{d_y}|\sqrt{2\mu}\})}, \quad \tau \neq \frac{1}{\sqrt{2\mu}}.$$

Bibliography

Abramovich, F., Bailey, T. C., and Sapatinas, T. (2000). Wavelet analysis and its statistical applications. *Journal of the Royal Statistical Society, Series D* , 49(1), pp.1–29.

Abramovich, F. and Benjamini, Y. (1996). Adaptive thresholding of wavelet coefficients. *Computational Statistics and Data Analysis*, 22(4), pp.351–361.

Abramovich, F., Besbeas, P., and Sapatinas, T. (2002). Empirical Bayes approach to block wavelet function estimation. *Computational Statistics and Data Analysis*, 39(4), pp.435–451.

Abramovich, F., Sapatinas, T., and Silverman, B. W. (1998). Wavelet thresholding via a Bayesian approach. *Journal of the Royal Statistical Society, Series B*, 60(4), pp.725–749.

Abramovich, F. and Silverman, B. W. (1998). Wavelet decomposition approaches to statistical inverse problems. *Biometrika*, 85(1), pp.115–129.

Al-Gezeri, S. M. (2003). *Modelling of multivariate data in archaeological geophysics*. PhD thesis, Department of Statistics, University of Leeds, United Kingdom.

Allum, G. T. (1997). *A statistical approach to inverse data problems in archaeological geophysics*. PhD thesis, Department of Statistics, University of Leeds, United Kingdom.

Allum, G. T., Aykroyd, R. G., and Haigh, J. (1999). Empirical Bayes estimation for archaeological stratigraphy. *Journal of the Royal Statistical Society, Series C*, 48(1), pp.1–14.

Altaher, A. M. and Ismail, M. T. (2010). A comparison of some thresholding selection methods for wavelet regression. *World Academy of Science, Engineering and Technology*, 62(1), pp.119–125.

Ammerman, A. J. and Feldman, M. W. (1978). Replicated collection of site surfaces. *American Antiquity*, 43, pp.734–740.

Andrews, D. F. and Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society, Series B*, pp.99–102.

Antoniadis, A., Bigot, J., and Sapatinas, T. (2001). Wavelet estimators in nonparametric regression: A comparative simulation study. *Journal of Statistical Software*, 6, pp.1–83.

Antoniadis, A. and Fan, J. (2001). Regularization of wavelet approximations. *Journal of the American Statistical Association*, 96(455), pp.939–967.

Antonini, M., Barlaud, M., Mathieu, P., and Daubechies, I. (1992). Image coding using wavelet transform. *IEEE Transactions on Image Processing*, 1(2), pp.205–220.

Aykroyd, R. G. (1998). Bayesian estimation for homogeneous and inhomogeneous Gaussian random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(5), pp.533–539.

Aykroyd, R. G. (2015). Statistical image reconstruction. In: Wang, M., ed. *Industrial tomography: Systems and applications*. United Kingdom: Woodhead Publishing, pp.401-424.

Aykroyd, R. G. and Al-Gezeri, S. M. (2014). 3D modeling and depth estimation in archaeological geophysics. *Chilean Journal of Statistics, Chilean Statistical Society*, 5(1), pp.19–35.

Aykroyd, R. G., Haigh, J. G. B., and Allum, G. T. (2001). Bayesian methods applied to survey data from archeological magnetometry. *Journal of the American Statistical Association*, 96(453), pp.64–76.

Aykroyd, R. G. and Mardia, K. V. (2003). A wavelet approach to shape analysis for spinal curves. *Journal of Applied Statistics*, 30(6), pp.605–623.

Baek, C. and Pipiras, V. (2009). Long range dependence, unbalanced Haar wavelet transformation and changes in local mean level. *International Journal of Wavelets, Multiresolution and Information Processing*, 7(1), pp.23–58.

Balakrishnan, N. and Kocherlakota, S. (1985). On the Double Weibull distribution: Order statistics and estimation. *Sankhyā: The Indian Journal of Statistics, Series B*, 47, pp.161–178.

Barber, S. (2001). *Simulating from the posterior density of Bayesian wavelet regression estimates*. Bristol: University of Bristol. [2014]. Available from: https://www1.maths.leeds.ac.uk/ stuart/research/pdf/simwb.pdf.

Barber, S. and Nason, G. P. (2004). Real nonparametric regression using complex wavelets. *Journal of the Royal Statistical Society, Series B*, 66(4), pp.927–939.

Barber, S., Nason, G. P., and Silverman, B. W. (2002). Posterior probability intervals for wavelet thresholding. *Journal of the Royal Statistical Society, Series B*, 64(2), pp.189–205.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B*, 57(1), pp.289–300.

Besag, J. (1983). Discussion of paper by p. switzer. *Bulletin of the International Statistical Institute*, 50(3), pp.422–425.

Besag, J., Green, P. J., Higdon, D., and Mengersen, K. (1995). Bayesian computation and stochastic systems. *Statistical Science*, 10(1), pp.3–41.

Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. (2011). *Handbook of Markov Chain Monte Carlo*. United Kingdom: Chapman and Hall/CRC.

Broughton, S. A. and Bryan, K. M. (2011). *Discrete Fourier analysis and wavelets: applications to signal and image processing*. United States: John Wiley and Sons.

Bruce, A. and Gao, H.-Y. (1996a). *Applied wavelet analysis with S-plus*. New York: Springer-Verlag. Inc.

Bruce, A. G. and Gao, H.-Y. (1996b). Understanding waveshrink: Variance and bias estimation. *Biometrika*, 83(4), pp.727–745.

Cai, T. T. (1999). Adaptive wavelet estimation: A block thresholding and oracle inequality approach. *Annals of Statistics*, 27(3), pp.898–924.

Cai, T. T. (2002). On adaptive wavelet estimation of a derivative and other related linear inverse problems. *Journal of Statistical Planning and Inference*, 108(1), pp.329–349.

Cai, T. T. and Silverman, B. W. (2001). Incorporating information on neighboring coefficients into wavelet estimation. *Sankhyā: The Indian Journal of Statistics, Series B*, 63, pp.127–148.

Cai, T. T. and Zhou, H. H. (2009). A data-driven block thresholding approach to wavelet estimation. *The Annals of Statistics*, 37(2), pp.569–595.

Caruso, M. J. and Withanawasam, L. S. (1999). Vehicle detection and compass applications using AMR magnetic sensors. In: *Sensors expo proceedings*. US: Honeywell Inc. Solid State Electronics Center. pp.477–489.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49(4), pp.327–335.

Chipman, H. A., Kolaczyk, E. D., and McCulloch, R. E. (1997). Adaptive Bayesian wavelet shrinkage. *Journal of the American Statistical Association*, 92(440), pp.1413–1421.

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

Clonda, D., Lina, J.-M., and Goulard, B. (2004). Complex Daubechies wavelets: Properties and statistical image modelling. *Signal Processing*, 84(1), pp.1–23.

Clyde, M. A. and George, E. I. (1999). Empirical Bayes estimation in wavelet nonparametric regression, In: Muller, P. and Vidakovic, B. eds. *Bayesian inference in wavelet-based models, lecture notes in statistic*. Springer-Verlag, New York, pp.309–322.

Clyde, M. A. and George, E. I. (2000). Flexible empirical Bayes estimation for wavelets. *Journal of the Royal Statistical Society, Series B*, 62(4), pp.681–698.

Clyde, M. A., Parmigiani, G., and Vidakovic, B. (1998). Multiple shrinkage and subset selection in wavelets. *Biometrika*, 85(2), pp.391–401.

Cohen, M. F. and Wallace, J. R. (2012). *Radiosity and realistic image synthesis*. United States: Academic Press Professional.

Coifman, R. R. and Donoho, D. L. (1995). Translation-invariant de-noising, In: Antoniadis, A. and Oppenheim, G. eds. *Wavelets and statistics*. New York: Springer-Verlag. 103, pp.125–150.

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

Cowles, M. K. and Carlin, B. P. (1996). Markov Chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91(434), pp.883–904.

Cross, G. R. and Jain, A. K. (1983). Markov random field texture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(1), pp.25–39.

Cutillo, L., Jung, Y. Y., Ruggeri, F., and Vidakovic, B. (2008). Larger posterior mode wavelet thresholding and applications. *Journal of Statistical Planning and Inference*, 138(12), pp.3758–3773.

Daubechies, I. (1988). Orthonormal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics*, 41(7), pp.909–996.

Daubechies, I. (1992). *Ten lectures on wavelets*. United States: Society for Industrial and Applied Mathematics Philadelphia.

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

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), pp.383-398.

Dearing, J. (1994). Environmental magnetic susceptibility: Using the bartington ms2 system. *Kenilworth*, United Kingdom: Chi Publishing, page pp.104.

Donoho, D. L. (1995). Nonlinear solution of linear inverse problems by wavelet–vaguelette decomposition. *Applied and Computational Harmonic Analysis*, 2(2), pp.101–126.

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432), pp.1200–1224.

Donoho, D. L. and Johnstone, J. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3), pp.425–455.

Downie, T. R. and Silverman, B. W. (1998). The discrete multiple wavelet transform and thresholding methods. *IEEE Transactions on Signal Processing*, 46(9), pp.2558–2561.

Evans, M. and Heller, F. (2003). *Environmental magnetism: Principles and applications of enviromagnetics*. New York: Academic Press.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456), pp.1348–1360.

Figueiredo, M. A. T. and Nowak, R. D. (2001). Wavelet-based image estimation: An empirical bayes approach using jeffrey's noninformative prior. *IEEE Transactions on Image Processing*, 10(9), pp.1322–1331.

Filzmoser, P. and Croux, C. (2002). A projection algorithm for regression with collinearity.

In: Jajuga, K., Sokolowski, A., Bock, H. eds. *Classification, clustering, and data analysis*. Berlin: Springer-Verlag, pp.227–234.

Flandrin, P. (1992). Wavelet analysis and synthesis of fractional brownian motion. *IEEE Transactions on Information Theory*, 38(2), pp.910–917.

Fryzlewicz, P. (2007). Unbalanced Haar technique for nonparametric function estimation. *Journal of the American Statistical Association*, 102(480), pp.1318–1327.

Gamerman, D. and Lopes, H. F. (2006). *Markov Chain Monte Carlo: Stochastic simulation for Bayesian inference*. 2nd ed. United Kingdom: Chapman and Hall/CRC.

Gao, H.-Y. (1998). Wavelet shrinkage denoising using the non-negative garrote. *Journal of Computational and Graphical Statistics*, 7(4), pp.469–488.

Gao, H.-Y. and Bruce, A. G. (1997). Waveshrink with firm shrinkage. *Statistica Sinica*, 7(4), pp.855–874.

Gelman, A. (1996). Inference and monitoring convergence. *The Annals of Applied Probability*, 7(1), pp.131–143.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *International Society for Bayesian Analysis*, 1(3), pp.515–534.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6), pp.721–741.

Geyer, C. (2011). Introduction to Markov Chain Monte Carlo. *Handbook of Markov Chain Monte Carlo*, pp.3–48.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996). Introducing Markov Chain Monte Carlo. *Markov Chain Monte Carlo in Practice*, London: Chapman and Hall/CRC.

Girardi, M. and Sweldens, W. (1997). A new class of unbalanced Haar wavelets that form

an unconditional basis for Lp on general measure spaces. *Journal of Fourier Analysis and Applications*, 3(4), pp.457–474.

Girault, F., Poitou, C., Perrier, F., Koirala, B. P., and Bhattarai, M. (2011). Soil characterization using patterns of magnetic susceptibility versus effective radium concentration. *Natural Hazards and Earth System Sciences, Copernicus GmbH*, 11(8), pp.2285–2293.

Gribble, S. D. (2001). Robustness in complex systems. In: *Proceedings of the Eighth Workshop Hot Topics in Operating Systems on 20-22 May*, pp.21–26.

Gyaourova, A., Kamath, C., and Fodor, I. K. (2002). Undecimated wavelet transforms for image de-noising. Reno: University of Nevada. [2014]. Available from: https://computation.llnl.gov/casc/sapphire/pubs/150931.pdf.

Haar, A. (1910). Zur theorie der orthogonalen funktionensysteme. *Mathematische Annalen*, 69(3), pp.331–371.

Hadamard, J. (2014). *Lectures on Cauchy's problem in linear partial differential equations*. New York: Dover publications, Inc. Courier Corporation.

Hall, P., Kerkyacharian, G., and Picard, D. (1999). On the minimax optimality of block thresholded wavelet estimators. *Statistica Sinica*, 9(1), pp.33–49.

Hall, P., Penev, S., Kerkyacharian, G., and Picard, D. (1997). Numerical performance of block thresholded wavelet estimators. *Statistics and Computing*, 7(2), pp.115–124.

Hanesch, M. and Scholger, R. (2002). Mapping of heavy metal loadings in soils by means of magnetic susceptibility measurements. *Environmental Geology*, 42(8), pp.857–870.

Härdle, W., Kerkyacharian, G., Picard, D., and Tsybakov, A. (1998). *Wavelets, approximation, and statistical applications*. United Kingdom: Springer Science and Business Media.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning*, Springer series in statistics. Berlin: Springer.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov Chains and their applications. *Biometrika*, 57(10), pp.97–109.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), pp.55–67.

Hubbard, B. B. (1996). *The world according to wavelets: The story of a mathematical technique in the making*. United States: A K Peters. Ltd.

Huerta, G. (2005). Multivariate bayes wavelet shrinkage and applications. *Journal of Applied Statistics*, 32(5), pp.529–542.

Jeffrey, A. and Zwillinger, D. ed. (2007). *Table of Integrals, Series, and Products*. New York: Academic Press. Academic Press, New York.

Jensen, M. J. (1995). Ordinary least squares estimate of the fractional differencing parameter using wavelets as derived from smoothing kernels. Southern Illinois University, Carbondale.

Jerri, A. J. (2011). *Introduction to wavelets*. New York: Sampling Publishing.

Johnstone, I. M. and Silverman, B. W. (1997). Wavelet threshold estimators for data with correlated noise. *Journal of the Royal Statistical Society, Series B*, 59(2), pp.319–351.

Johnstone, I. M. and Silverman, B. W. (2005a). Ebayesthresh: R and S-plus programs for empirical Bayes thresholding. *Journal of Statistical Software*, 12(8), pp.1–38.

Johnstone, I. M. and Silverman, B. W. (2005b). Empirical Bayes selection of wavelet thresholds. *Annals of Statistics*, 33(4), pp.1700–1752.

Kane, J., Herrmann, F., and Toksoz, M. N. (2002). *Wavelet domain geophysical inversion*. Technical report, Massachusetts Institute of Technology. Earth Resources Laboratory. [2014]. Available from: http://www-eaps.mit.edu/erl/research/report1/pdf2002/KANE2.pdf.

Karimi, R., Ayoubi, S., Jalalian, A., Sheikh-Hosseini, A. R., and Afyuni, M. (2011).

Relationships between magnetic susceptibility and heavy metals in urban topsoils in the arid region of Isfahan, central Iran. *Journal of Applied Geophysics*, 74(1), pp.1–7.

Katayama, S. and Fujisawa, H. (2016). Sparse and robust linear regression: An optimization algorithm and its statistical properties. *Statistica Sinica*.

Kyung, M., Gill, J., Ghosh, M., Casella, G.,*et al.* (2010). Penalized regression, standard errors, and bayesian Lasso. *Bayesian Analysis*, 5(2), pp.369–411.

Laplace, P. S. (1774). Mémoire sur la probabilité des causes par les évènements. *Mémoires de Mathématique et de Physique*, 6, pp.621-656.

Lawton, W. (1993). Applications of complex valued wavelet transforms to subband decomposition. *IEEE Transactions on Signal Processing*, 41(12), pp.3566–3568.

Le Borgne, E. (1955). Susceptibilité magnétique anormale du sol superficiel. *Annales De Géophysique, Elsevier Science*, pp.399–419.

Leslie-Pelecky, D. L. and Rieke, R. D. (1996). Magnetic properties of nanostructured materials. *Chemistry of Materials, ACS Publications*, 8(8), pp.1770–1783.

Levine, H. A.*et al.* (1979). Review: AN Tikhonov and VY Arsenin, solutions of ill-posed problems. *Bulletin of the American Mathematical Society*, 1(3), pp.521–524.

Lewis, C., Mitchell-Fox, P., and Dyer, C. (1997). *Village, hamlet and field: Changing medieval settlements in central England*. United Kingdom: Windgather Press.

Lina, J.-M. (1997). Image processing with complex daubechies wavelets. *Journal of Mathematical Imaging and Vision*, 7(3), pp.211–223.

Lina, J.-M. and MacGibbon, B. (1997). Nonlinear shrinkage estimation with complex Daubechies wavelets. In: *Proceedings of SPIE Wavelet applications in signal and image orocessing*, 3169, pp.67–79.

Lina, J.-M. and Mayrand, M. (1995). Complex Daubechies wavelets. *Applied and Computational Harmonic Analysis*, 2(3), pp.219–229.

Lina, J.-M., Turcotte, P., and Goulard, B. (1999). Complex dyadic multiresolution analyses. *Advances in Imaging and Electron Physics*, 109, pp.163–197.

Liu, J., Billings, S. A., Zhu, Z. Q., and Shen, J. (2002). Enhanced frequency analysis using wavelets. *International Journal of Control*, 75(15), pp.1145–1158.

Liu, J. S. (2001). *Monte Carlo strategies in scientific computing*. United Kingdom: Springer Science and Business Media.

Mallat, S. G. (1989). A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7), pp.674–693.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, 34(3), pp.1436–1462.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6), pp.1087–1092.

Meyer, Y. (1995). *Wavelets and operators*. United Kingdom: Cambridge University Press.

Morlet, J., Arens, G., Fourgeau, E., and Glard, D. (1982). Wave propagation and sampling theory-Part I: Complex signal and scattering in multilayered media. *Geophysics*, Society of Exploration Geophysicists, 47(2), pp.203–221.

Mullins, C. E. (1977). Magnetic susceptibility of the soil and its significance in soil science–a review. *Journal of Soil Science*, 28(2), pp.223–246.

Nason, G. P. (1995). Choice of the threshold parameter in wavelet function estimation. In: Antoniadis, A. and Oppenheim, G. eds. *Wavelets and statistics*. New York: Springer-Verlag. 103, pp.261–280.

Nason, G. P. (1996). Wavelet shrinkage using cross-validation. *Journal of the Royal Statistical Society, Series B*, 58(2), 463–479.

Nason, G. P. (2010a). *Wavelet methods in statistics with R*. New York: Springer.

Nason, G. P. (2010b). *Wavethresh 4.5. Software. Department of Mathematics*, United Kingdom: University of Bristol. [2014]. Available from http://www.stats.bris.ac.uk/ wavethresh.

Nason, G. P. and Silverman, B. W. (1994). The discrete wavelet transform in S. *Journal of Computational and Graphical Statistics*, 3(2), pp.163–191.

Nason, G. P. and Silverman, B. W. (1995). The stationary wavelet transform and some statistical applications In: Antoniadis, A. and Oppenheim, G. eds. *Wavelets and statistics: lecture notes in statistics*. New York: Springer-Verlag. 103, pp.281–299.

Needham, S. P. (1985). Neolithic and Bronze Age settlement on the buried floodplains of Runnymede. *Oxford Journal of Archaeology*, 4(2), pp.125–137.

Ogden, R. T. (1994). *Wavelet thresholding in nonparametric regression with change-point applications*. PhD thesis, M University, Texas A.

Park, T. and Casella, G. (2008). The bayesian Lasso. *Journal of the American Statistical Association*, 103(482), pp.681–686.

Percival, D. B. and Walden, A. T. (2006). *Wavelet methods for time series analysis*. New York: Cambridge University Press.

Pericchi, L. R. and Smith, A. F. M. (1992). Exact and approximate posterior moments for a normal location parameter. *Journal of the Royal Statistical Society, Series B*, 54(3), pp.793–804.

Pesquet, J.-C., Krim, H., and Carfantan, H. (1996). Time-invariant orthonormal wavelet representations. *IEEE Transactions on Signal Processing*, 44(8), pp.1964–1970.

Polzehl, J. and Spokoiny, V. G. (2000). Adaptive weights smoothing with applications to image restoration. *Journal of the Royal Statistical Society, Series B*, 62(2), pp.335–354.

Qian, W. and Titterington, D. M. (1991). Multidimensional Markov Chain models for image textures. *Journal of the Royal Statistical Society, Series B*, 53(3), pp.661–674.

Raftery, A. E. and Lewis, A. E. (1995). The number of iterations, convergence diagnostics

and generic Metroplis algorithms. In: Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. eds. *Practical Markov Chain Monte Carlo*. Chapman and Hall.

Raimondo, M. (2002). Wavelet shrinkage via peaks over threshold. *Inter-Stat*, 5(1), pp.1–19.

Reményi, N. (2012). Contributions to Bayesian wavelet shrinkage. PhD thesis, Industrial and Systems Engineering, Georgia Insitute of Technolog.

Reményi, N. and Vidakovic, B. (2015). Wavelet shrinkage with Double Weibull prior. *Communications in Statistics-Simulation and Computation*, 44(1), pp.88–104.

Renfrew, C. and Bahn, P. (2013). *Archaeology: The key concepts*. London and New York: Routledge.

Robert, C. and Casella, G. (2011). A short history of Markov Chain Monte Carlo: Subjective recollections from incomplete data. *Statistical Science*, 26(1), pp.102–115.

Roberts, G. O., Gelman, A., Gilks, W. R.,*et al.* (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, 7(1), pp.110–120.

Ruggeri, F. and Vidakovic, B. (2005). Bayesian modeling in the wavelet domain. *Handbook of Statistics: Bayesian Thinking, Modeling and Computation*, 25, pp.315–338.

Sardy, S. (2000). Minimax threshold for denoising complex signals with waveshrink. *IEEE Transactions on Signal Processing*, 48(4), pp.1023–1028.

Scollar, I. (1970). Magnetic methods of archaeological prospecting-advances in instrumentation and evaluation techniques. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 269(1193), pp.109–119.

Scollar, I., Tabbagh, A., Hesse, A., and Herzog, I. (1990). *Archaeological prospecting and remote sensing*. United Kingdom: Cambridge University Press.

Selesnick, I. W., Baraniuk, R. G., and Kingsbury, N. C. (2005). The dual-tree complex wavelet transform. *IEEE Signal Processing Magazine*, 22(6), pp.123–151.

Sifuzzaman, M., Islam, M. R., and Ali, M. Z. (2009). Application of wavelet transform and its advantages compared to Fourier transform. *Journal of Physical Sciences*.

Silverman, B. W. (1999). Wavelets in statistics: beyond the standard assumptions. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 357(1760), pp.2459–2473.

Sobolu, R. and Pusta, D. (2010). Wavelet Methods in Nonparametric Regression Based on Experimental Data. *Bulletin of University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca. Horticulture*, 66(2), pp.718–725.

Starck, J.-L., Fadili, J., and Murtagh, F. (2007). The undecimated wavelet decomposition and its reconstruction. *Transactions on Image Processing*, 16(2), pp.297–309.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6), pp.1135–1151.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58, 267–288.

Tite, M. S. and Linington, R. E. (1975). Effect of climate on the magnetic susceptibility of soils.

Vidakovic, B. (1998a). Nonlinear wavelet shrinkage with Bayes rules and Bayes factors. *Journal of the American Statistical Association*, 93(441), pp.173–179.

Vidakovic, B. (1998b). Wavelet-based nonparametric bayes methods. In: Dey, M. and Sinha. eds. *Practical nonparametric and semiparametric Bayesian statistics, lecture notes in statistics*. New York: Springer-Verlag. Inc. 133, pp.133–155.

Vidakovic, B. (1999). *Statistical Modeling by Wavelets*. United States: John Wiley and Sons, 503.

Vidakovic, B. and Ruggeri, F. (2001). Bams method: Theory and simulations. *Sankhyā: The Indian Journal of Statistics, Series B*, 63(2), pp.234-249.

Vidya, V. (2008). Non-decimated wavelet shrinkage algorithm for image denoising based

on inter-scale correlation. *Future generation communication and networking symposia. Second international conference on (2008)*. 3, pp.90–93.

Wickerhauser, M. V. (1994). *Adapted wavelet analysis from theory to software*. United States: AK Peters. Ltd.

Young, R. K. (1993). *Wavelet theory and its applications*. United Kingdom: Springer Science and Business Media.

Zellner, A. (1996). Bayesian method of moments (BMOM) analysis of mean and regression models. In: *Modelling and prediction honoring seymour geisser*. New York: Springer. pp.61–72.

Zhao, Y., Ogden, R. T., and Reiss, P. T. (2012). Wavelet-based Lasso in functional linear regression. *Journal of Computational and Graphical Statistics*, 21(3), pp.600–617.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67(2), pp.301–320.