Wavelet Estimators in Nonparametric Regression: A Comparative Simulation Study

Anestis Antoniadis, *Jeremie Bigot*
Laboratoire IMAG-LMC,
University Joseph Fourier,
BP 53,
38041 Grenoble Cedex 9,
France.

and

Theofanis Sapatinas,
Department of Mathematics and Statistics,
University of Cyprus,
P.O. Box 20537,
CY 1678 Nicosia,
Cyprus.

Journal of Statistical Software, **6**, Issue 6, 1–83 (2001).
(available from `http://www.jstatsoft.org/v06/i06/updates/jssv06i06.pdf`)

Abstract

Wavelet analysis has been found to be a powerful tool for the nonparametric estimation of spatially-variable objects. We discuss in detail wavelet methods in nonparametric regression, where the data are modelled as observations of a signal contaminated with additive Gaussian noise, and provide an extensive review of the vast literature of wavelet shrinkage and wavelet thresholding estimators developed to denoise such data. These estimators arise from a wide range of classical and empirical Bayes methods treating individual or blocks of wavelet coefficients either globally or in a level-dependent fashion. We compare various estimators in an extensive simulation study on a variety of sample sizes, test functions, signal-to-noise ratios and wavelet filters. Because there is no single criterion that can adequately summarise the behaviour of an estimator, we use various criteria to measure performance in finite sample situations. Insight into the performance of these estimators is obtained from graphical outputs and numerical tables. In order to provide some hints of how these estimators should be used to analyse real-life data sets, a detailed practical step-by-step illustration of a wavelet denoising analysis on electrical consumption is provided. `Matlab` codes are provided so that all figures and tables in this paper can be reproduced.

Some key words: EM ALGORITHM; EMPIRICAL BAYES; MONTE CARLO EXPERIMENTS; NONPARAMETRIC REGRESSION; SMOOTHING METHODS; SHRINKAGE; THRESHOLDING; WAVELETS.

1 INTRODUCTION

Nonparametric regression has been a fundamental tool in data analysis over the past two decades and is still an expanding area of ongoing research. The goal is to recover an unknown function, say g, based on sampled data that are contaminated with additive Gaussian noise. Only very general assumptions about g are made such as that it belongs to a certain class of *smooth* functions. Nonparametric regression (or denoising) techniques provide a very effective and simple way of finding structure in data sets without the imposition of a parametric regression model (as in linear or polynomial regression for example). However, nonparametric and parametric regression models should not be viewed as mutually exclusive competitors. In many cases, a nonparametric regression estimate will suggest a simple parametric model, while in other cases it will be clear that the underlying regression function is too complicated and no reasonable parametric model would be adequate.

During the 1980s and 1990s, the literature was inundated by hundreds (and perhaps thousands) of papers regarding various *linear* (on a fixed spatial scale) estimators to the nonparametric regression problem. Some of the more popular are those based on *kernel functions*, *smoothing splines* and *orthogonal series*. Each of these approaches has its own particular strengths and weaknesses. We refer, for example, to the monographs of Härdle (1990), Green & Silverman (1994), Wand & Jones (1995), Fan & Gijbels (1996) and Eubank (1999) for extensive bibliographies. For linear smoothers, asymptotic results are easily obtained. Usually, if g is sufficiently smooth (for example if g belongs to a Sobolev space of regularity s), then the *mean integrated squared error* (which is usually considered to measure asymptotic performance) converges to zero at a rate n^{-r} as the sample size n increases. This is known as the *optimal* asymptotic rate and for all reasonable linear smoothers, r is the same (for example, $r = 2s/(2s+1)$ for the Sobolev case).

Unfortunately, the asymptotics are not particularly helpful to practitioners faced with finite data sets when deciding what type of smoother to use. As a public service, Breiman & Peters (1992) designed a useful simulation study comparing some popular linear smoothing methods found in the literature on a variety of sample sizes, regression functions and signal-to-noise ratios using a number of different criterion to measure performance. As expected, they found that no linear smoother uniformly dominates in all aspects. However, valuable conclusions were drawn from these simulations about the small sample behaviour of the various linear smoothers and their computational complexity.

The fixed spatial scale estimators mentioned above assume that g belongs to some given set of functions $\{g_\theta;\ \theta \in \Theta\}$ where Θ is an infinite dimensional parameter set. It turns out that the prior knowledge of the family $\{g_\theta;\ \theta \in \Theta\}$ influences both the construction of the estimators and their performances. Roughly speaking, the larger the family the larger the risk, revealing a major drawback of the preceding estimators. The problem of *spatially adaptive estimation* concerns whether one can, *without prior knowledge* of the family $\{g_\theta;\ \theta \in \Theta\}$, built estimators that achieve the optimal asymptotic rates on some privileged subsets of the parameter set Θ. Over the last decade, various *nonlinear* (spatially adaptive) estimators in nonparametric regression have been proposed. The most popular are variable-bandwidth kernel methods, classification and regression trees, and adaptive regression splines. Although some of these methods achieve the optimal asymptotic rates, they can be computationally intensive and they are usually designed to denoise *regular* functions.

During the 1990s, the nonparametric regression literature was dominated by (nonlinear) *wavelet shrinkage* and *wavelet thresholding* estimators. These estimators are a new subset of an old class of nonparametric regression estimators, namely orthogonal series methods. Moreover, these estimators are easily implemented through fast algorithms so they are very appealing in practical situations. Donoho & Johnstone (1994) and Donoho, Johnstone, Kerkyacharian & Picard (1995) have introduced nonlinear wavelet estimators in nonparametric regression through thresholding which typically amounts to term-by-term assessment of estimates of coefficients in the empirical wavelet expansion of the unknown function g. If an estimate of a coefficient is sufficiently large in absolute value – that is, if it exceeds a predetermined threshold – then the corresponding term in the empirical wavelet expansion is retained (or shrunk toward to zero by an amount equal to the threshold); otherwise it is omitted.

In particular, these papers study nonparametric regression from a *minimax* viewpoint, using some important function classes not previously considered in statistics when linear smoothers were studied (and have also provided new viewpoints for understanding other nonparametric smoothers as well). These function classes model the notion of *different amounts of smoothness in different locations* more effectively than the usual smooth classes. In other words, these function classes contain function spaces like the Hölder and Sobolev spaces of regular functions, as well as spaces of *irregular* functions such as those of 'bounded variation'. These considerations are not simply esoteric, but are of statistical importance, since these classes of functions typically arise in practical applications such as the processing of speech, electrocardiogram or seismic signals. (Mathematically, all these function spaces can be formalized in terms of the so-called Besov or Triebel spaces; see, for example, Meyer (1992), Wojtaszczyk (1997), Donoho & Johnstone (1998) and Härdle, Kerkyacharian, Picard & Tsybakov (1998).) Although these contributions describe performance in terms of convergence rates that are achieved over large function classes, concise accounts of mean squared error for single functions also exist and have been discussed in detail by Hall & Patil (1996a, 1996b).

To date, the study of these methods has been mostly asymptotic in character. In particular, it has been shown that nonlinear wavelet thresholding estimators are asymptotically near optimal or optimal while traditional linear estimators are suboptimal for estimation over particular classes of the Besov or Triebel spaces (see, for example, Delyon & Juditsky, 1996; Donoho & Johnstone, 1998; Abramovich, Benjamini, Donoho & Johnstone, 2000). As with any asymptotic result, there remain doubts as to how well the asymptotics describe small sample behaviour. In other words, how large the sample size should be before the asymptotic theory applies is a very important question. To shed some light on this question, Marron, Adak, Johnstone, Newmann & Patil (1998) applied the tool of exact risk analysis to understand the small sample behaviour of the two wavelet thresholding estimators introduced by Donoho & Johnstone (1994), namely the *minimax* and *universal* wavelet estimators, and thus to check directly the conclusions suggested by asymptotics. Also, their analysis provide insight as to why the viewpoints and conclusions of Donoho-Johnstone (convergence rates achieved uniformly over large function classes) differ from those of Hall-Patil (convergence rates achieved for single functions). Bruce & Gao (1996), Gao & Bruce (1997), Gao (1998) and Antoniadis & Fan (2001) have also developed analytical tools to understand the finite sample behaviour of minimax wavelet estimators based on various thresholding rules.

Since the seminal papers by Donoho & Johnstone (1994) and Donoho, Johnstone, Kerkyacharian & Picard (1995), various alternative *data-adaptive* wavelet thresholding estimators have been developed. For example, Donoho & Johnstone (1995) proposed the *SureShrink* estimator based on minimizing Stein's unbiased risk estimate; Weyrich & Warhola (1995a, 1995b), Nason (1996) and Jansen, Malfait & Bultheel (1997) have considered estimators based on *cross-validation* approaches to choosing the thresholding parameter; while Abramovich & Benjamini (1995, 1996) and Ogden & Parzen (1996a, 1996b) considered thresholding as a *multiple hypotheses* testing procedure. Hall, Penev, Kerkyacharian & Picard (1997), Hall, Kerkyacharian & Picard (1998, 1999), Cai (1999), Efromovich (1999, 2000) and Cai & Silverman (2001) suggested further modifications of the basic thresholding by considering wavelet *block* thresholding estimators meaning that the wavelet coefficients are thresholded in blocks rather than term-by-term. Some of these alternative wavelet thresholding estimators possess near optimal asymptotic properties. Moreover, it has been shown that wavelet block thresholding estimators have excellent mean squared error performances relative to wavelet term-by-term thresholding estimators in finite sample situations.

Various Bayesian approaches for nonlinear wavelet thresholding and nonlinear *wavelet shrinkage* estimators have also recently been proposed. To fix terminology, a *shrinkage* rule shrinks wavelet coefficients to zero, whilst a *thresholding* rule in addition sets actually to zero all coefficients below a certain level (as in the classical approach). These estimators have been shown to be effective and it is argued that they are less *ad-hoc* than the classical proposals discussed above. In the Bayesian approach a prior distribution is imposed on the wavelet coefficients. The prior model is designed to capture the sparseness of wavelet expansions common to most applications. Then the function is estimated by applying a suitable Bayesian rule to the resulting posterior distribution of the wavelet coefficients. Different choices of loss function lead to different Bayesian rules and hence to different nonlinear wavelet shrinkage and wavelet thresholding rules. Such wavelet estimators have been discussed, for example, by Chipman, Kolaczyk & McCulloch (1997), Abramovich, Sapatinas & Silverman (1998), Clyde, Parmigiani & Vidakovic (1998), Crouse, Nowak & Baraniuk (1998), Johnstone & Silverman (1998), Vidakovic (1998a), Clyde & George (1999, 2000), Vannucci & Corradi (1999), Huang & Cressie (2000), Huang and Lu (2000), Vidakovic & Ruggeri (2001) and Angelini, De Canditiis & Leblanc (2003). Moreover, it has been shown that Bayesian wavelet shrinkage and thresholding estimators outperform the classical wavelet term-by-term thresholding estimators in terms of mean squared error in finite sample situations. Recently, Abramovich, Besbeas & Sapatinas (2002) have considered Bayesian wavelet block shrinkage and block thresholding estimators, and have shown that they outperform existing classical block thresholding estimators in terms of mean squared error in finite sample situations.

Extensive reviews and descriptions of the various classical and Bayesian wavelet shrinkage and wavelet thresholding estimators are given in the books by Ogden (1997), Vidakovic (1999) and Percival and Walden (2000), in the papers appeared in the edited volume by Müller & Vidakovic (1999), and in the review papers by Antoniadis (1997), Vidakovic (1998b) and Abramovich, Bailey and Sapatinas (2000). It is evident that, although a number of wavelet estimators has been compared to get insight as to which ones are to be *preferred against the others*, a more detailed study involving recent classical and Bayesian wavelet methods is required in the development towards *high-performance* wavelet estimators. Furthermore, with the increased applicability of these estimators in nonparametric regression, their finite sample properties become even more important.

Therefore, as a public service, similar to one given by Breiman & Peters (2000) for linear smoothers, in this paper we design an extensive simulation study to compare most of the above mentioned wavelet estimators on a variety of sample sizes, test functions, signal-to-noise ratios and wavelet filters. Because there is no single criterion that can adequately summarise the behaviour of an estimator, various criteria (including the traditional mean squared error) are used to measure performance. Insight into the performance of these estimators in finite sample situations is obtained from graphical outputs and numerical tables.

In particular, from the classical wavelet thresholding estimators we consider: minimax estimators, the universal estimator, SureShrink estimators, a translation invariant estimator (a variant of the universal estimator), the two-fold cross-validation estimator, multiple hypotheses testing estimators, a nonoverlapping block estimator and an overlapping block estimator. A linear wavelet estimator (extending spline smoothing estimation methods) is also considered.

From the Bayesian wavelet shrinkage and wavelet thresholding estimators we consider: posterior mean estimators, posterior median estimators, a hypothesis testing estimator, a deterministic/stochastic decomposition estimator, a nonparametric mixed-effect estimator and nonoverlapping block estimators. The elicitation of the hyperparameters of the above estimators are resolved in detail by employing empirical Bayes methods that attempt to determine the hyperparameters of the prior distributions from the data being analysed.

The plan of this paper is as follows. Section 2 recalls some known results about wavelet series and the discrete wavelet transform. Section 3 briefly discusses the nonlinear wavelet approach to nonparametric regression. In Section 4 the different classical and empirical Bayes wavelet estimators used in the simulation study are presented. The description of the actual simulation study is given in Section 5. Section 6 summarises and discusses the results of the simulations. The overall conclusions of our simulation comparison is presented in Section 7. Section 8 explains in some detail which files the user should download and how to start running them. In order to provide some hints of how the functions should be used to analyse real-life data sets, a detailed practical step-by-step illustration of a wavelet denoising analysis on electrical consumption is provided. Finally, following the principle of reproducible research as advocated by Buckheit & Donoho (1995), `Matlab` routines and a description of software implementing these routines (so that all figures and tables in this paper can be reproduced) are available at `http://www-lmc.imag.fr/SMS/software/GaussianWaveDen.html` or `http://www.ucy.ac.cy/∼fanis/links/software.html` .

2 WAVELET SERIES AND THE DISCRETE WAVELET TRANSFORM

In this section we give a brief overview of some relevant material on the wavelet series expansion and a fast wavelet transform that we need later.

2.1 THE WAVELET SERIES EXPANSION

The term *wavelets* is used to refer to a set of orthonormal basis functions generated by dilation and translation of a compactly supported *scaling function* (or *father wavelet*), ϕ, and a *mother* *wavelet*, ψ, associated with an r-regular multiresolution analysis of $L^2(\mathbb{R})$. A variety of different *wavelet families* now exist that combine compact support with various degrees of smoothness and numbers of vanishing moments (see Daubechies (1992)), and these are now the most intensively used wavelet families in practical applications in statistics. Hence, many types of functions encountered in practice can be sparsely (i.e. parsimoniously) and uniquely represented in terms of a wavelet series. Wavelet bases are therefore not only useful by virtue of their special structure, but they may also be (and have been!) applied in a wide variety of contexts.

For simplicity in exposition, we shall assume that we are working with periodised wavelet bases on $[0,1]$ (see, for example, Mallat (1999), Section 7.5.1), letting

$$\phi^{\mathrm{p}}_{jk}(t) = \sum_{l\in\mathbb{Z}} \phi_{jk}(t-l) \quad \text{and} \quad \psi^{\mathrm{p}}_{jk}(t) = \sum_{l\in\mathbb{Z}} \psi_{jk}(t-l), \quad \text{for} \quad t\in[0,1],$$

where

$$\phi_{jk}(t) = 2^{j/2}\phi(2^j t - k) \quad \text{and} \quad \psi_{jk}(t) = 2^{j/2}\psi(2^j t - k).$$

For any given *primary resolution* level $j_0 \geq 0$, the collection

$$\{\phi^{\mathrm{p}}_{j_0 k},\ k = 0,1,\ldots,2^{j_0}-1;\ \psi^{\mathrm{p}}_{jk},\ j \geq j_0 \geq 0,\ k = 0,1,\ldots,2^j-1\}$$

is then an orthonormal basis of $L^2([0,1])$. The superscript "p" will be suppressed from the notation for convenience.

Despite the poor behaviour of periodic wavelets near the boundaries (they create high amplitude wavelet coefficients in the neighborhood of the boundaries when the analysed function is not periodic) they are commonly used because the numerical implementation is particularly simple. Also, as Johnstone (1994) has pointed out, this computational simplification affects only a fixed number of wavelet coefficients at each resolution level and does not affect the qualitative phenomena that we wish to present. The idea underlying such an approach is to express any function $g \in L^2([0,1])$ in the form

$$g(t) = \sum_{k=0}^{2^{j_0}-1} \alpha_{j_0 k}\phi_{j_0 k}(t) + \sum_{j=j_0}^{\infty}\sum_{k=0}^{2^j-1} \beta_{jk}\psi_{jk}(t), \quad j_0 \geq 0, \quad t\in[0,1],$$

where

$$\alpha_{j_0 k} = \langle g, \phi_{j_0 k}\rangle = \int_0^1 g(t)\phi_{j_0 k}(t)\,dt, \quad j_0 \geq 0, \quad k = 0,1,\ldots,2^{j_0}-1$$

and

$$\beta_{jk} = \langle g, \psi_{jk}\rangle = \int_0^1 g(t)\psi_{jk}(t)\,dt, \quad j \geq j_0 \geq 0, \quad k = 0,1,\ldots,2^j-1.$$

For detailed expositions of the mathematical aspects of wavelets we refer to, for example, Daubechies (1992), Meyer (1992), Wojtaszczyk (1997) and Mallat (1999).

2.2 THE DISCRETE WAVELET TRANSFORM

In statistical settings we are more usually concerned with discretely sampled, rather than continuous, functions. It is then the wavelet analogy to the discrete Fourier transform which is of primary interest and this is referred to as the discrete wavelet transform (DWT). Given a vector of function values $\mathbf{g} = (g(t_1), ..., g(t_n))'$ at equally spaced points t_i, the discrete wavelet transform of $\mathbf{g}$ is given by

$$\mathbf{d} = W\mathbf{g},$$

where $\mathbf{d}$ is an $n \times 1$ vector comprising both discrete scaling coefficients, c_{j_0k}, and discrete wavelet coefficients, d_{jk}, and W is an orthogonal $n \times n$ matrix associated with the orthonormal wavelet basis chosen. The c_{j_0k} and d_{jk} are related to their continuous counterparts α_{j_0k} and β_{jk} (with an approximation error of order n^{-1}) via the relationships

$$c_{j_0k} \approx \sqrt{n}\, \alpha_{j_0k} \quad \text{and} \quad d_{jk} \approx \sqrt{n}\, \beta_{jk}.$$

The factor $\sqrt{n}$ arises because of the difference between the continuous and discrete orthonormality conditions. This root factor is unfortunate but both the definition of the DWT and the wavelet coefficients are now fixed by convention, hence the different notation used to distinguish between the discrete wavelet coefficients and their continuous counterpart. Note that, because of orthogonality of W, the inverse DWT (IDWT) is simply given by

$$\mathbf{g} = W'\mathbf{d},$$

where W' denotes the transpose of W.

If $n = 2^J$ for some positive integer J, the DWT and IDWT may be performed through a computationally fast algorithm developed by Mallat (1989) that requires only order n operations. In this case, for a given j_0 and under periodic boundary conditions, the DWT of $\mathbf{g}$ results in an n-dimensional vector $\mathbf{d}$ comprising both discrete scaling coefficients $c_{j_0k},\ k = 0, ..., 2^{j_0} - 1$ and discrete wavelet coefficients $d_{jk},\ j = j_0, ..., J-1;\ k = 0, ..., 2^j - 1$.

We do not provide technical details here of the order n DWT algorithm mentioned above. Essentially the algorithm is a fast hierarchical scheme for deriving the required inner products which at each step involves the action of low and high pass filters, followed by a decimation (selection of every even member of a sequence). The IDWT may be similarly obtained in terms of related filtering operations. For excellent accounts of the DWT and IDWT in terms of filter operators we refer to Nason & Silverman (1995), Strang & Nguyen (1996), or Burrus, Gonipath & Guo (1998).

3 THE WAVELET APPROACH TO NONPARAMETRIC REGRESSION

Consider the standard univariate nonparametric regression setting

$$y_i = g(t_i) + \sigma \, \epsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where ϵ_i are independent $N(0,1)$ random variables and the noise level σ may, or may not, be known. The goal is to recover the underlying function g from the observations $\mathbf{y} = (y_1, \ldots, y_n)'$ without assuming any particular parametric structure on its form.

In what follows we assume, without loss of generality, that the sample points t_i are within the unit interval $[0,1]$. For simplicity, we also assume that the sample points are equally spaced, i.e. $t_i = i/n$, and that the sample size n is a power of two: $n = 2^J$ for some positive integer J. These assumptions allow us to perform both the DWT and the IWDT using Mallat's (1989) fast algorithm.

Remark 3.1 *It should be noted that for non-equispaced or random designs, or sample sizes which are not a power of two, or data contaminated with correlated noise, modifications are needed to the standard wavelet-based estimation procedures that will be discussing in Section 4. We refer, for example, to Deylon & Juditsky (1995), Neumann & Spokoiny (1995), Wang (1996), Hall & Turlach (1997), Johnstone & Silverman (1997), Antoniadis, Grégoire & Vial (1997), Antoniadis & Pham (1998), Cai & Brown (1998, 1999), Kovac & Silverman (2000), von Sachs & MacGibbon (2000), Nason (2002) and Angelini, De Canditiis & Leblanc (2003). In our simulation comparison in Section 5, we have not included these latter methods but rather concentrated on the standard nonparametric regression setting. Although a more general comparison will be valuable it is, however, outside the scope of this paper.*

One of the basic approaches to nonparametric regression is to consider the unknown function g expanded as a generalised Fourier series and then to estimate the generalised Fourier coefficients from the data. The original nonparametric problem is thus transformed to a parametric one, although the potential number of parameters is infinite. An appropriate choice of basis for the expansion is therefore a key point in relation to the efficiency of such an approach. A 'good' basis should be parsimonious in the sense that a large set of possible response functions can be approximated well by only few terms of the generalized Fourier expansion employed. As already discussed, wavelet series allow a parsimonious expansion for a wide variety of functions, including inhomogeneous cases. It is therefore natural to consider applying the generalized Fourier series approach using a wavelet series.

Due to the orthogonality of the matrix W, the DWT of white noise is also an array of independent $N(0,1)$ random variables, so from (1) it follows that

$$\hat{c}_{j_0k} = c_{j_0k} + \sigma\, \epsilon_{jk}, \quad k = 0, 1, \ldots, 2^{j_0} - 1, \tag{2}$$

$$\hat{d}_{jk} = d_{jk} + \sigma\, \epsilon_{jk}, \quad j = j_0, \ldots, J-1, \quad k = 0, \ldots, 2^j - 1, \tag{3}$$

where $\hat{c}_{j_0k}$ and $\hat{d}_{jk}$ are respectively the *empirical scaling* and the *empirical wavelet* coefficients of the noisy data $\mathbf{y}$, and ϵ_{jk} are independent $N(0,1)$ random variables. The sparseness of the wavelet expansion makes it reasonable to assume that essentially only a few 'large' d_{jk} contain information about the underlying function g, while 'small' d_{jk} can be attributed to the noise which uniformly contaminates all wavelet coefficients. If we can decide which are the 'significant' large wavelet coefficients, then we can retain them and set all others equal to zero, thus obtaining an approximate wavelet representation of the underlying function g. It is also advisable to keep the scaling coefficients c_{j_0k}, the coefficients on the lower *coarse* levels, intact because they represent 'low-frequency' terms that usually contain important components about the underlying function g.

Finally, we mention that the primary resolution level j_0 that we have used throughout our simulations was chosen to be $j_0 = [\log_2(\log(n))] + 1$, following the asymptotic considerations given in Chapter 10 of Härdle, Kerkyacharian, Picard & Tsybakov (1998).

3.1 THE CLASSICAL APPROACH TO WAVELET THRESHOLDING

A wavelet based linear approach, extending simply spline smoothing estimation methods as described by Wahba (1990), is the one suggested by Antoniadis (1996) and independently by Amato & Vuza (1997). Of non-threshold type, this method is appropriate for estimating relatively regular functions. Assuming that the smoothness index s of the function g to be recovered is known, the resulting estimator is obtained by estimating the scaling coefficients c_{j_0k} by their empirical counterparts $\hat{c}_{j_0k}$ and by estimating the wavelet coefficients d_{jk} via a linear shrinkage

$$\tilde{d}_{jk} = \frac{\hat{d}_{jk}}{1 + \lambda 2^{2js}},$$

where $\lambda > 0$ is a smoothing parameter. The parameter λ is chosen by cross-validation in Amato & Vuza (1997), while the choice of λ in Antoniadis (1996) is based on risk minimization and depends on a preliminary consistent estimator of the noise level σ. The above linear wavelet method *is not* designed to handle spatially inhomogeneous functions with *low* regularity. For such functions one usually relies upon nonlinear wavelet (thresholding or shrinkage) methods.

Donoho & Johnstone (1994, 1995, 1998) and Donoho, Johnstone, Kerkyacharian & Picard (1995) proposed a *nonlinear* wavelet estimator of g based on reconstruction by keeping the empirical scaling coefficients $\hat{c}_{j_0k}$ in (2) intact and from a more judicious selection of the empirical wavelet coefficients $\hat{d}_{jk}$ in (3). They suggested the extraction of the significant wavelet coefficients by *thresholding* in which wavelet coefficients are set to zero if their absolute value is below a certain threshold level, $\lambda \geq 0$, whose choice we discuss in more detail in Section 4.1. Under this scheme we obtain thresholded wavelet coefficients using either the *hard* or *soft* thresholding rule given respectively by

$$\delta_\lambda^{\mathrm{H}}(\hat{d}_{jk}) = \begin{cases} 0 & \text{if } |\hat{d}_{jk}| \leq \lambda \\ \hat{d}_{jk} & \text{if } |\hat{d}_{jk}| > \lambda \end{cases} \tag{4}$$

and

$$\delta_\lambda^{\mathrm{S}}(\hat{d}_{jk}) = \begin{cases} 0 & \text{if } |\hat{d}_{jk}| \leq \lambda \\ \hat{d}_{jk} - \lambda & \text{if } \hat{d}_{jk} > \lambda \\ \hat{d}_{jk} + \lambda & \text{if } \hat{d}_{jk} < -\lambda. \end{cases} \tag{5}$$

Thresholding allows the data itself to decide which wavelet coefficients are significant; hard thresholding (a discontinuous function) is a 'keep' or 'kill' rule, while soft thresholding (a continuous function) is a 'shrink' or 'kill' rule.

Bruce & Gao (1996) and Marron, Adak, Johnstone, Newmann & Patil (1998) have shown that simple threshold values with hard thresholding results in larger variance in the function estimate, while the same threshold values with soft thresholding shift the estimated coefficients by an amount of λ even when $|\hat{d}_{jk}|$ stand way out of noise level, creating unnecessary bias when the true coefficients are large. Also, due to its discontinuity, hard thresholding can be unstable, that is, sensitive to small changes in the data.

To remedy the drawbacks of both hard and soft thresholding rules, Gao & Bruce (1997) and considered the *firm* threshold thresholding

$$\delta_{\lambda_1,\lambda_2}^{\mathrm{F}}(\hat{d}_{jk}) = \begin{cases} 0 & \text{if } |\hat{d}_{jk}| \leq \lambda_1 \\ \mathrm{sign}(\hat{d}_{jk}) \frac{\lambda_2(|\hat{d}_{jk}| - \lambda_1)}{\lambda_2 - \lambda_1} & \text{if } \lambda_1 < |\hat{d}_{jk}| \leq \lambda_2 \\ \hat{d}_{jk} & \text{if } |\hat{d}_{jk}| > \lambda_2 \end{cases} \tag{6}$$

which is a "keep" or "shrink" or "kill" rule (a continuous function).

The resulting wavelet thresholding estimators offer, in small samples, advantages over both hard thresholding (generally smaller mean squared error and less sensitivity to small perturbations in the data) and soft thresholding (generally smaller bias and overall mean squared error) rules. For values of $|\hat{d}_{jk}|$ near the lower threshold λ_1, $\delta_{\lambda_1,\lambda_2}^{\mathrm{F}}(\hat{d}_{jk})$ behaves like $\delta_{\lambda_1}^{\mathrm{S}}(\hat{d}_{jk})$. For values of $|\hat{d}_{jk}|$ above the upper threshold λ_2, $\delta_{\lambda_1,\lambda_2}^{\mathrm{F}}(\hat{d}_{jk})$ behaves like $\delta_{\lambda_2}^{\mathrm{H}}(\hat{d}_{jk})$. Note that the hard thresholding and soft thresholding rules are limiting cases of (6) with $\lambda_1 = \lambda_2$ and $\lambda_2 = \infty$ respectively.

Note that firm thresholding has a drawback in that it requires two threshold values (one for 'keep' or 'shrink' and another for 'shrink' or 'kill'), thus making the estimation procedure for the threshold values more computationally expensive. To overcome this drawback, Gao (1998) considered the *nonnegative garrote* thresholding

$$\delta_\lambda^{\mathrm{G}}(\hat{d}_{jk}) = \begin{cases} 0 & \text{if } |\hat{d}_{jk}| \leq \lambda \\ \hat{d}_{jk} - \frac{\lambda^2}{\hat{d}_{jk}} & \text{if } |\hat{d}_{jk}| > \lambda \end{cases} \tag{7}$$

which is a "shrink" or "kill" rule (a continuous function). The resulting wavelet thresholding estimators offer, in small samples, advantages over both hard thresholding and soft thresholding rules that is comparable to the firm thresholding rule, while the latter requires two threshold values.

In the same spirit to that in Gao (1998), Antoniadis & Fan (2001) suggested the *SCAD* thresholding rule

$$\delta_\lambda^{\mathrm{SCAD}}(\hat{d}_{jk}) = \begin{cases} \operatorname{sign}(\hat{d}_{jk}) \max(0, |\hat{d}_{jk}| - \lambda) & \text{if } |\hat{d}_{jk}| \leq 2\lambda \\ \frac{(\alpha-1)\hat{d}_{jk} - a\lambda \operatorname{sign}(\hat{d}_{jk})}{\alpha - 2} & \text{if } 2\lambda < |\hat{d}_{jk}| \leq \alpha\lambda \\ \hat{d}_{jk} & \text{if } |\hat{d}_{jk}| > \alpha\lambda \end{cases} \tag{8}$$

which is also a "keep" or "shrink" or "kill" rule (a piecewise linear function). It does not over penalize large values of $|\hat{d}_{jk}|$ and hence does not create excessive bias when the wavelet coefficients are large. Antoniadis & Fan (2001) have recommended to use the value of $\alpha = 3.7$ based on a Bayesian argument.

Remark 3.2 *In our simulation comparison in Section 5, we have only considered hard, soft and SCAD thresholding rules since the firm and the nonnegative garrote thresholding rules have been extensively studied by Gao & Bruce (1997) and Gao (1998) – see also Vidakovic (1999) for a discussion on other types of thresholding rules. Moreover, hard and soft thresholding rules are the most commonly used (if not the only ones!) among the various classical wavelet thresholding estimators considered in practice that we will be discussing in Section 4.1.*

The thresholded wavelet coefficients obtained by applying any of the thresholding rules δ_λ, given in (4)–(7), are used to obtain a selective reconstruction of the response function g. The resulting estimate can be written as

$$\hat{g}_\lambda(t) = \sum_{k=0}^{2^{j_0}-1} \frac{\hat{c}_{j_0 k}}{\sqrt{n}} \, \phi_{j_0 k}(t) + \sum_{j=j_0}^{J-1} \sum_{k=0}^{2^j-1} \frac{\delta_\lambda(\hat{d}_{jk})}{\sqrt{n}} \, \psi_{jk}(t). \tag{9}$$

Often, we are interested in estimating the unknown response function at the observed data-points. In this case the vector $\hat{\mathbf{g}}_\lambda$ of the corresponding estimates can be derived by simply performing the IDWT of $\{\hat{c}_{j_0k},\ \delta_\lambda(\hat{d}_{jk})\}$ and the resulting three-step selective reconstruction estimation procedure can be summarized by the following diagram

$$\mathbf{y} \overset{\text{DWT}}{\longrightarrow} \left\{\hat{c}_{j_0k},\ \hat{d}_{jk}\right\} \overset{\text{thresholding}}{\longrightarrow} \left\{\hat{c}_{j_0k},\ \delta_\lambda(\hat{d}_{jk})\right\} \overset{\text{IDWT}}{\longrightarrow} \hat{\mathbf{g}}_\lambda. \tag{10}$$

As mentioned in Section 2.2, for the equally spaced design and sample size $n = 2^J$ assumed throughout this section, both DWT and IDWT in (10) can be performed by a fast algorithm of order n, and so the whole process is computationally very efficient.

3.2 The Bayesian approach to wavelet shrinkage and thresholding

Recall from (2) that the empirical scaling coefficients $\{\hat{c}_{j_0k} : k = 0, \ldots, 2^{j_0} - 1\}$ conditionally on c_{j_0k} and σ^2 are independently distributed as

$$\hat{c}_{j_0k} \mid c_{j_0k}, \sigma^2 \sim N(c_{j_0k}, \sigma^2). \tag{11}$$

Similarly, from (3), the empirical wavelet coefficients $\{\hat{d}_{jk} : j = j_0, \ldots, J-1;\ k = 0, \ldots, 2^j - 1\}$ conditionally on d_{jk} and σ^2 are independently distributed as

$$\hat{d}_{jk} \mid d_{jk}, \sigma^2 \sim N(d_{jk}, \sigma^2). \tag{12}$$

In the Bayesian approach a prior model is imposed on the function's wavelet coefficients, designed to capture the sparseness of the wavelet expansion common to most applications. It is assumed that in the prior model the wavelet coefficients d_{jk} are *mutually independent* random variables (and independent of the empirical wavelet coefficients $\hat{d}_{jk}$). A popular prior model for each wavelet coefficient d_{jk} is a scale mixture of two distributions, one mixture component corresponding to *negligible* coefficients, the other to *significant* coefficients. Usually, a scale mixture of two normal distributions or a mixture of one normal distribution and a point mass at zero is considered. Such mixtures have been recently applied to stochastic search variable selection problems (see, for example, George & McCullogh (1993)). The mixture component with larger variance represents the significant coefficients, while the mixture component with smaller variance represents the negligible ones. The limiting case of a mixture of one normal distribution and a point mass at zero corresponds to the case where the smaller variance is actually set to zero.

An important distinction between the use of a scale mixture of two normal distributions (considered by Chipman, Kolaczyk & McCulloch (1997)) and a scale mixture of a normal distribution and a point mass at zero (considered by Clyde, Parmigiani & Vidakovic (1998) and Abramovich, Sapatinas & Silverman (1998)) in the Bayesian wavelet approach to nonparametric regression is the type of shrinkage obtained. In the former case, no wavelet coefficient estimate based on the posterior analysis will be exactly equal to zero. However, in the latter case, with a proper choice of a Bayes rule, it is possible to get wavelet coefficient estimates that are exactly zero, resulting in a *bonafide* thresholding rule. We consider here the mixture model of one normal distribution and a point mass at zero in detail. (Other choices of prior models will also be discussed in Section 4.3.)

Following the discussion given above, a hierarchical model that expresses the belief that some of the wavelet coefficients $\{d_{jk} : j = j_0, \ldots, J-1;\ k = 0, \ldots, 2^j - 1\}$ are zero is obtained by

$$d_{jk} \mid \gamma_{jk} \sim N(0, \gamma_{jk}\tau_j^2) \tag{13}$$

$$\gamma_{jk} \sim \text{Bernoulli}(\pi_j), \tag{14}$$

where $d_{jk} \mid \gamma_{jk}$ are mutually independent random variables. The binary random variables γ_{jk} determine whether the wavelet coefficients are nonzero ($\gamma_{jk} = 1$), arising from $N(0, \tau_j^2)$ distributions, or zero ($\gamma_{jk} = 0$), arising from point masses at zero. In the next stage of the hierarchy, it is assumed that the γ_{jk} have independent Bernoulli distributions with $P(\gamma_{jk} = 1) = 1 - P(\gamma_{jk} = 0) = \pi_j$ for some fixed hyperparameter $0 \leq \pi_j \leq 1$. The probability π_j gives the proportion of nonzero wavelet coefficients at resolution level j while the variance τ_j^2 is a measure of their magnitudes. The same prior parameters π_j and τ_j^2 for all coefficients at a given resolution level j are used, resulting in level-dependent wavelet threshold and shrinkage estimators.

Recall from Section 3.1 that is advisable to keep the coefficients on the lower coarse levels intact because they represent ‘low-frequency’ terms that usually contain important components of the function g. Thus, to complete the prior specification of g, the scaling coefficients $\{c_{j_0k} : k = 0, \ldots, 2^{j_0} - 1\}$ are assumed to be *mutual independent* random variables (and independent of the empirical scaling coefficients $\hat{c}_{j_0k}$) and vague priors are placed on them

$$c_{j_0k} \sim N(0, \epsilon), \quad \epsilon \to \infty. \tag{15}$$

Once the data are observed, the empirical wavelet coefficients $\hat{d}_{jk}$ and empirical scaling coefficients $\hat{c}_{j_0k}$ are determined, and we seek the posterior distributions on the wavelet coefficients d_{jk} and scaling coefficients c_{j_0k}. Using (12), (13) and (14), the d_{jk} are *a posteriori* conditionally independent

$$d_{jk} \mid \gamma_{jk}, \hat{d}_{jk}, \sigma^2 \sim N\left(\gamma_{jk}\, \frac{\tau_j^2}{\sigma^2 + \tau_j^2}\, \hat{d}_{jk}, \gamma_{jk}\, \frac{\sigma^2\tau_j^2}{\sigma^2 + \tau_j^2}\right). \tag{16}$$

In order to incorporate model uncertainty about which of the wavelet coefficients d_{jk} are zero, we now average over all possible γ_{jk}. Using (16), the marginal posterior distribution of d_{jk}

conditionally on σ^2, is then given by

$$\begin{aligned} d_{jk} \mid \hat{d}_{jk}, \sigma^2 \quad \sim \quad & p(\gamma_{jk} = 1 \mid \hat{d}_{jk}, \sigma^2)\, N\left(\frac{\tau_j^2}{\sigma^2 + \tau_j^2}\, \hat{d}_{jk}, \frac{\sigma^2 \tau_j^2}{\sigma^2 + \tau_j^2}\right) \\ & + (1 - p(\gamma_{jk} = 1 \mid \hat{d}_{jk}, \sigma^2))\, \delta(0), \end{aligned} \tag{17}$$

where $\delta(0)$ is a point mass at zero. It is not difficult to see the posterior probabilities that wavelet coefficients d_{jk} are nonzero can be expressed as

$$p(\gamma_{jk} = 1 \mid \hat{d}_{jk}, \sigma^2) = \frac{1}{1 + O_{jk}(\hat{d}_{jk}, \sigma^2)}, \tag{18}$$

where the posterior odds ratios $O_{jk}(\hat{d}_{jk}, \sigma^2)$ that $\gamma_{jk} = 0$ versus $\gamma_{jk} = 1$ are given by

$$O_{jk}(\hat{d}_{jk}, \sigma^2) = \frac{1 - \pi_j}{\pi_j} \frac{(\sigma^2 + \tau_j^2)^{1/2}}{\sigma} \exp\left(-\frac{\tau_j^2 \hat{d}_{jk}^2}{2\sigma^2(\sigma^2 + \tau_j^2)}\right). \tag{19}$$

Based on some Bayes rules (BR), as we shall discuss in Section 4.3, expressions (16) and (17) can be used to obtain wavelet threshold and shrinkage estimates $\mathrm{BR}(d_{jk} \mid \hat{d}_{jk}, \sigma^2)$, of the wavelet coefficients d_{jk}. Also, using (11) and (15), the c_{j_0k} are *a posteriori* conditionally independent

$$c_{j_0k} \mid \hat{c}_{j_0k}, \sigma^2 \sim N(\hat{c}_{j_0k}, \sigma^2) \tag{20}$$

and therefore, using (20), c_{j_0k} are estimated by $\hat{c}_{j_0k}$.

The posterior-based wavelet threshold and wavelet shrinkage coefficients are used to obtain a selective reconstruction of the response function. The resulting estimate can be written as

$$\hat{g}_{\mathrm{BR}}(t) = \sum_{k=0}^{2^{j_0}-1} \frac{\hat{c}_{j_0k}}{\sqrt{n}}\, \phi_{j_0k}(t) + \sum_{j=j_0}^{J-1} \sum_{k=0}^{2^j-1} \frac{\mathrm{BR}(d_{jk} \mid \hat{d}_{jk}, \sigma^2)}{\sqrt{n}}\, \psi_{jk}(t). \tag{21}$$

Finally, the vector $\hat{\mathbf{g}}_{\mathrm{BR}}$ of the corresponding estimates of the unknown response function g at the observed data-points can be derived by simply performing the IDWT of $\{\hat{c}_{j_0k},\ \mathrm{BR}(d_{jk} \mid \hat{d}_{jk}, \sigma^2)\}$ and the resulting four-step selective reconstruction estimation procedure can be summarized by the following diagram

$$\mathbf{y} \overset{\text{DWT}}{\longrightarrow} \left\{\hat{c}_{j_0k},\ \hat{d}_{jk}\right\} \overset{\text{Priors}}{\longrightarrow} \{c_{j_0k},\ d_{jk}\} \overset{\text{Posterior estimates}}{\longrightarrow} \left\{\hat{c}_{j_0k},\ \mathrm{BR}(d_{jk} \mid \hat{d}_{jk}, \sigma^2)\right\} \overset{\text{IDWT}}{\longrightarrow} \hat{\mathbf{g}}_{\mathrm{BR}}. \tag{22}$$

As mentioned in Section 2.2, for the equally spaced design and sample size $n = 2^J$ assumed throughout this section, both DWT and IDWT in (22) can be performed by a fast algorithm of order n, and so the whole process is computationally very efficient.

4 DESCRIPTION OF THE WAVELET ESTIMATORS

In this section, we give a brief description of the various classical and empirical Bayes wavelet shrinkage and thresholding estimators that we will be using in our simulation study given in Section 5.

4.1 CLASSICAL METHODS: TERM-BY-TERM THRESHOLDING

Given the basic framework of function estimation using wavelet thresholding as discussed in Section 3.1, there are a variety of methods to choose the threshold level λ in (9) or (10) for any given situation. These can be grouped into two categories: *global thresholds* and *level-dependent thresholds*. The former means that we choose a *single* value of λ to be applied globally to all empirical wavelet coefficients $\{\hat{d}_{jk} : j = j_0, \dots, J-1;\ k = 0, 1, \dots, 2^j - 1\}$ while the latter means that a *possibly different* threshold value λ_j is chosen for each resolution level $j = j_0, \dots, J-1$.

In what follows, we consider both global and level-dependent thresholds. These thresholds all require an estimate of the noise level σ. The usual standard deviation of the data values is clearly not a good estimator, unless the underlying response function g is reasonably flat. Donoho & Johnstone (1994) considered estimating σ in the wavelet domain and suggested a robust estimate that is based only on the empirical wavelet coefficients at the finest resolution level. The reason for considering only the finest level is that corresponding empirical wavelet coefficients tend to consist mostly of noise. Since there is some signal present even at this level, Donoho & Johnstone (1994) proposed a robust estimate of the noise level σ (based on the *median absolute deviation*) given by

$$\hat{\sigma} = \frac{\text{median}(\{|\hat{d}_{J-1,k}| : k = 0, 1, \dots, 2^{J-1} - 1\})}{0.6745}. \tag{23}$$

The estimator (23) has become very popular in practice and it is used in subsequent sections unless stated otherwise.

4.1.1 THE MINIMAX THRESHOLD

An optimal threshold, derived to minimize the constant term in an upper bound of the risk involved in estimating a function, was obtained by Donoho & Johnstone (1994). The proposed *minimax threshold*, that depends on the sample size n, is defined as

$$\lambda^{\mathrm{M}} = \hat{\sigma}\lambda_n^\star, \tag{24}$$

where $\lambda_n^\star$ is defined as the value of λ which achieves

$$\Lambda_n^\star := \inf_\lambda \sup_d \left\{ \frac{R_\lambda(d)}{n^{-1} + R_{\text{oracle}}(d)} \right\}, \tag{25}$$

where $R_\lambda(d) = E(\delta_\lambda(\hat{d}) - d)^2$ and $R_{\text{oracle}}(d)$ is the ideal risk achieved with the help of an *oracle*.

Two oracles were considered by Donoho & Johnstone (1994): *diagonal linear projection* (DLP), an oracle which tell us when to "keep" or "kill" each empirical wavelet coefficient, and *diagonal linear shrinker* (DLS), an oracle which tells you how much to shrink each wavelet coefficient. The ideal risks for these oracles are given by

$$R^{\text{DLP}}_{\text{oracle}}(d) := \min(d^2, 1) \quad \text{and} \quad R^{\text{DLS}}_{\text{oracle}}(d) := \frac{d^2}{d^2+1}.$$

Donoho and Johnstone (1994) computed the DLP minimax thresholds for the soft thresholding rule (5), while the DLP minimax thresholds for the hard thresholding rule (4) and the DLS minimax thresholds for both soft and hard thresholding rules were obtained by Bruce & Gao (1996). DLP minimax thresholds for the SCAD thresholding rule (8) were obtained by Antoniadis & Fan (2001). Numerical values for the sample sizes that we will be using in our simulative study in Section 5 are given in Table 1.

n	128	256	512	1024
HARD	2.913	3.117	3.312	3.497
SOFT	1.669	1.859	2.045	2.226
SCAD	1.691	1.881	2.061	2.241

Table 1: Diagonal linear projection minimax thresholds for HARD, SOFT and SCAD thresholding rules for various sample sizes.

Since the type of oracle used has little impact on the minimax thresholds, Table 1 only reports the DLP minimax thresholds and can be used as a look-up table in any software. These values were computed using a grid search over λ with increments $\Delta_\lambda = 0.0001$. At each point, the supremum over d in (24) was computed using a quasi-Newton optimisation with numerical derivatives (see, for example, Dennis & Mei, 1979).

Remark 4.1 *Although not considered in our simulation study in Section 5, we note that DLP minimax thresholds for the firm thresholding rule (6) and the nonnegative garrote thresholding rule (7) were considered by Gao & Bruce (1997) and Gao (1998) respectively. Numerical values for selected sample sizes can be found, for example, in Table 2 in Gao (1998).*

4.1.2 The Universal Threshold

As an alternative to the use of minimax thresholds, Donoho & Johnstone (1994) suggested thresholding of empirical wavelet coefficients $\{\hat{d}_{jk} : j = j_0, \ldots, J-1;\ k = 0, 1, \ldots, 2^j - 1\}$ by using the *universal threshold*

$$\lambda^{\mathrm{U}} = \hat{\sigma}\sqrt{2\log n}. \tag{26}$$

This threshold is easy to remember and its implementation in software requires no costly development of look-up tables. The universal threshold ensures, with high probability, that every sample in the wavelet transform in which the underlying function is exactly zero will be estimated as zero. This is so, because if $X_1, \ldots, X_n$ are independent and identically distributed $N(0,1)$ random variables, then

$$P\left\{\max_{1\leq i\leq n} |X_i| > \sqrt{2\log n}\right\} \sim \frac{1}{\sqrt{\pi\log n}}, \quad \text{as } n\to\infty.$$

The rate at which the probability above tends to one is however quite slow.

Remark 4.2 *In fact, it can be shown that*

$$P\left\{\max_{1\leq i\leq n} |X_i| > \sqrt{c\log n}\right\} \sim \frac{\sqrt{2}}{n^{c/2-1}\sqrt{c\pi\log n}}, \quad \text{as } n\to\infty. \tag{27}$$

Although not considered in our simulation study in Section 5, we mention that one can define other universal thresholds such that (27) converges to zero faster (ie, to suppress noise more thoroughly). For example, $\lambda_1^{\mathrm{U}} = \sqrt{4\log n}$ *makes the above convergence rate* $1/n\sqrt{2\pi\log n}$ *and* $\lambda_2^{\mathrm{U}} = \sqrt{6\log n}$ *makes the above convergence rate* $1/n^2\sqrt{3\pi\log n}$*. Such thresholds have been used for applications involving indirect noisy measurements (usually called* inverse problems*); in particular the latter universal threshold has been used by Abramovich & Silverman (1998) for the estimation of a function's derivative as a practical important example of an inverse problem.*

4.1.3 The Translation Invariant Threshold

It has been noted from scientists (and other users) that wavelet thresholding with either the minimax threshold (24) or the universal threshold (26) suffers from artifacts of various kinds. In other words, in the vicinity of discontinuities, these wavelet thresholding estimators can exhibit pseudo-Gibbs phenomena, alternating undershoot and overshoot of a specific target level. While these phenomena are less pronounced than in the case of Fourier-based estimates (in which Gibbs phenomena are global, rather than local, and of large amplitude), it seems reasonable to try to improve the resulting wavelet-based estimation procedures.

Coifman & Donoho (1995) proposed the use of the translation invariant wavelet thresholding scheme which helps to suppress these artifacts. The idea is to correct unfortunate mis-alignments between features in the function of interest and features in the basis. When the function of interest contains several discontinuities, the approach of Coifman & Donoho (1995) is to apply a *range* of shifts in the function, and *average* over the several results so obtained. Given data $\mathbf{y} = (y_1, \ldots, y_n)'$ from model (1), the *translation invariant* wavelet thresholding estimator is defined as

$$\hat{g}^{\mathrm{TI}} = \frac{1}{n}\sum_{k=i}^{n}(WS_k)'\delta_\lambda(WS_k\mathbf{y}), \tag{28}$$

where W is the order n orthogonal matrix associated with the DWT, S_k is the shift matrix

$$\begin{pmatrix} O_{k\times(n-k)} & I_{k\times k} \\ I_{(n-k)\times(n-k)} & O_{(n-k)\times k} \end{pmatrix},$$

δ_λ is either the hard thresholding rule (4) or the soft thresholding rule (5), $I_{k\times k}$ is the identity matrix with k rows, and $O_{r_1\times r_2}$ is an $r_1 \times r_2$ matrix of zeroes.

Alternatively, the translation invariant thresholding is via the nondecimated discrete wavelet transform (NDWT) (or stationary or maximal overlap transform) – see, for example, Nason & Silverman (1995) or Percival & Walden (2000). By denoting W^{TI} to be the matrix associated with the NDWT (which actually maps a vector of length $n = 2^J$ into a vector of length $(J+1)n$), then (28) is equivalent to

$$\hat{g}^{\mathrm{TI}} = W_{\mathrm{G}}^{\mathrm{TI}}\delta_\lambda(W^{\mathrm{TI}}\mathbf{y}), \tag{29}$$

where $W_{\mathrm{G}}^{\mathrm{TI}} = \left((W^{\mathrm{TI}})'W^{\mathrm{TI}}\right)^{-1}(W^{\mathrm{TI}})'$ is the generalised inverse of W^{TI}.

Remark 4.3 *Coifman & Donoho (1995) applied the above procedure by using the threshold* $\lambda = \hat{\sigma}\sqrt{2\log_e((n\log_2(n))}$. *Although lower threshold levels could be used in conjunction with translation invariant thresholding, Coifman & Donoho (1995) pointed out that such lower thresholds result in a very large number of noise spikes, apparently much larger than in the non-invariant case. Therefore, in our simulation study in Section 5, we only consider translation invariant thresholding in conjunction with the above threshold and with both hard thresholding (4) and soft thresholding (5).*

4.1.4 THRESHOLDING AS A MULTIPLE HYPOTHESES TESTING PROBLEM

The idea of wavelet thresholding can be viewed as a multiple hypotheses testing. For each wavelet coefficient $\hat{d}_{jk} \sim N(d_{jk}, \hat{\sigma}^2)$, test the following hypothesis

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

If H_0 is rejected, the coefficient $\hat{d}_{jk}$ is retained in the model; otherwise it is discarded.

Classical approaches to multiple hypotheses testing in this case face serious problems because of the large numbers of hypotheses being tested simultaneously. In other words, if the error is controlled at an *individual* level, the chance of keeping erroneously a coefficient is extremely high; if the *simultaneous* error is controlled, the chance of keeping a coefficient is extremely low. Abramovich & Benjamini (1995,1996) proposed a way to control such dissipation of power based on the *false discovery rate* (FDR) method of Benjamini & Hochberg (1995).

Let R be the number of empirical wavelet coefficients that are not dropped by the thresholding procedure for a given sample (thus, they are kept in the model). Of these R coefficients, S are correctly kept in the model and $V = R - S$ are erroneously kept in the model. The error in such a procedure is expressed in terms of the random variable $Q = V/R$ – the proportion of the empirical wavelet coefficients kept in the model that should have been dropped. (Naturally, it is defined that $Q = 0$ when $R = 0$ since no error of this type can be made when no coefficient is kept.) The FDR of empirical wavelet coefficients can now be defined as the *Expectation* of Q, reflecting the expected proportion of erroneously kept coefficients among the ones kept in the model. Following the method of Benjamini & Hochberg (1995), Abramovich & Benjamini (1995,1996) proposed maximizing the number of empirical wavelet coefficients kept in the model subject to condition $\mathbb{E}Q < \alpha$, for some prespecified level α, yielding the following procedure

1. Take $j_0 = 0$. For each of the $n-1$ empirical wavelet coefficients $\{\hat{d}_{jk} : j = 0, 1, \ldots, J-1;\ k = 0, 1, \ldots, 2^j - 1\}$ calculate the corresponding 2-sided p-value, p_{jk}, (testing $H_0 : d_{jk} = 0$),
$$p_{jk} = 2\left(1 - \Phi\left(\frac{|\hat{d}_{jk}|}{\hat{\sigma}}\right)\right),$$
where Φ is the cumulative distribution function of a standard normal random variable.

2. Order the p_{jk} according to their size, $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(n-1)}$ (i.e., each $p_{(i)}$ corresponds to some coefficient d_{jk}).

3. Find $k = \max\left(i : p_{(i)} < (i/m)\alpha\right)$. For this k, calculate
$$\lambda^{\text{FDR}} = \hat{\sigma}\Phi^{-1}\left(1 - \frac{p_{(k)}}{2}\right). \tag{30}$$

4. Apply the hard thresholding rule (4) or soft thresholding rule (5) to all empirical wavelet coefficients (ignoring the coefficients at the coarsest levels) $\{\hat{d}_{jk} : j = j_0, \ldots, J-1;\ k = 0, 1, \ldots, 2^j - 1\}$ with the threshold value (30) (using the traditional levels for significance testing, i.e., $\alpha = 0.01$ or $\alpha = 0.05$).

Remark 4.4 *We note that the universal threshold (26) can be viewed as a critical value of a similar test to the ones considered above. The level for this test is*

$$\alpha = P(|\hat{d}_{jk}| > \sigma\sqrt{2\log n} \mid H_0) \approx (n\sqrt{\pi \log n})^{-1},$$

which is also equal to its power against the alternative $H_1 : d_{jk} = d\ (\neq 0)$. *Thus, as mentioned by Abramovich & Benjamini (1995), the approach of Donoho & Johnstone (1994) based on the universal threshold (26) is equivalent to the 'panic' procedure of controlling the probability of even one erroneous inclusion of a wavelet coefficient at the level* $(n\sqrt{\pi \log n})^{-1}$, *but the level at which the error is controlled approaches zero as* n *tends to infinity.*

4.1.5 Thresholding Using Cross-Validation

One way to choose the threshold level λ is by minimising the mean integrated squared error (MISE) between a wavelet threshold estimator $\hat{g}_\lambda$ and the true function g. In symbols, the threshold λ should minimise

$$M(\lambda) = \mathbb{E} \int (\hat{g}_\lambda(x) - g(x))^2 \ dx. \tag{31}$$

In practice, the function g is unknown and so an estimate of M is required.

Cross-validation is widely used as an automatic procedure to choose the smoothing parameter in many statistical settings – see, for example, Green & Silverman (1994) or Eubank (1999). The classical cross-validation method is performed by systematically expelling a data point from the construction of an estimate, predicting what the removed value would be and, then, comparing the prediction with the value of the expelled point.

Cross-validation is usually numerically intensive unless there are some updating formulae that allow to calculate the 'leaving-out-one' predictions on the basis of the 'full' predictions only. In this respect very helpful is the 'leaving-out-one' Lemma 4.2.1 of Wahba (1990), which shows that such updating may be done when the so-called 'compatibility condition' holds. Although this condition is easy to derive for projection-type estimators, it fails to hold for nonlinear shrinkage or thresholding rules. One way to proceed is to pretend that this condition 'almost' holds. This approach to cross-validation in wavelet regression was adopted by Nason (1994, 1995, 1996). In order to directly apply the DWT, the latter author suggested breaking the original data set into 2 subsets of equal size: one containing only the even-indexed data, and the other, the odd-indexed data. The odd-indexed data will be used to 'predict' the even-indexed data, and vice-versa, leading to a 'leave-out-half' strategy.

To be more specific, given data $\mathbf{y} = (y_1, \ldots, y_n)'$ from model (1) with $n = 2^J$, remove all the odd-indexed y_i from the set. This leaves 2^{J-1} evenly indexed y_i which are reindexed from $j = 1, \ldots, 2^{J-1}$. These reindexed data are then used to construct a function estimate $\hat{g}_\lambda^{\mathrm{E}}$ by using a particular threshold parameter λ with either hard thresholding (4) or soft thresholding (5). To compare the function estimator with the left-out noisy data an interpolated version of

$\hat{g}_\lambda^{\mathrm{E}}$ is formed

$$\bar{g}_{\lambda,j}^{\mathrm{E}} = \frac{1}{2}(\hat{g}_{\lambda,j+1}^{\mathrm{E}} + \hat{g}_{\lambda,j}^{\mathrm{E}}), \quad j = 1, \ldots, n/2,$$

setting $\hat{g}_{\lambda,n/2+1}^{\mathrm{E}} = \hat{g}_{\lambda,1}^{\mathrm{E}}$ because g is assumed to be periodic. The $\bar{g}_\lambda^{\mathrm{O}}$ is computed for the odd-indexed points and the interpolant is, similarly, formed as

$$\bar{g}_{\lambda,j}^{\mathrm{O}} = \frac{1}{2}(\hat{g}_{\lambda,j+1}^{\mathrm{O}} + \hat{g}_{\lambda,j}^{\mathrm{O}}), \quad j = 1, \ldots, n/2.$$

The full estimate for the MISE given in (31) compares the interpolated wavelet estimators and the left out points

$$\hat{M}(\lambda) = \sum_{j=1}^{n/2} \left[(\bar{g}_{\lambda,j}^{\mathrm{E}} - y_{2j+1})^2 + (\bar{g}_{\lambda,j}^{\mathrm{O}} - y_{2j})^2 \right]. \tag{32}$$

It has been showed by Nason (1994) that one can almost always find a unique minimum of (32)

$$\lambda_{\min} = \arg\min_{\lambda \geq 0} \hat{M}(\lambda).$$

This minimum value depends on $n/2$ data points (since both estimates of g, $\hat{g}_\lambda^{\mathrm{E}}$ and $\bar{g}_\lambda^{\mathrm{O}}$ are based on $n/2$ data points) and, therefore, a correction for the sample size is needed. Nason (1994, 1995) considered the universal threshold λ^{U} given in (26) to supply a heuristic method for obtaining a cross-validated threshold for n data points. By using this adjustment, the *leave-out-half cross-validation* threshold is defined as

$$\lambda^{\mathrm{CV}} = \left(1 - \frac{\log 2}{\log n}\right)^{-1/2} \lambda_{\min}. \tag{33}$$

Remark 4.5 *Nason (1996) also developed a leave-one-out cross-validation method that works for* any *number of data points, removing the above algorithm's restriction of* $n = 2^J$ *data points. However, since we are only considered the case of* $n = 2^J$ *in this paper, this latter algorithm is not considered in our simulation study in Section 5.*

We also pinpoint that Weyrich & Warhola (1995a, 1995b) and Jansen, Malfait & Bultheel (1997), by mimicking the classical cross-validation, applied the method of generalized cross-validation to choose the threshold level λ*. The criterion they used is*

$$GCV(\lambda) = \frac{1}{n} \frac{\|\mathbf{w} - \mathbf{w}_\lambda\|^2}{\|\frac{n_0}{n}\|^2},$$

where $\mathbf{w}_\lambda$ *is the vector of thresholded normalised coefficients and* n_0/n *is the fraction of coefficients replaced by zero by this particular threshold value. They show that the minimizer of this function is an asymptotically optimal threshold in the mean-squared error sense. This alternative threshold, however, is not considered in our simulation study in Section 5.*

4.1.6 THE SURESHRINK THRESHOLD

Donoho & Johnstone (1995) introduced a scheme that uses the empirical wavelet coefficients at each resolution level j to choose a threshold value λ_j with which to threshold the empirical wavelet coefficients. The idea is to employ Stein's unbiased risk criterion (see Stein (1981)) to get an unbiased estimate of the l^2-risk.

Consider the following equivalent problem to (1) or, equivalently, to (2)–(3). Suppose $X_1, \ldots, X_s$ are independent $N(\mu_i, 1)$, $i = 1, \ldots, s$, random variables. The problem is to estimate the mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_s)'$ with minimum l^2-risk. A result of Stein (1981) states that the l^2-loss can be estimated unbiasedly for any estimator $\boldsymbol{\mu}$ that can be written as $\hat{\boldsymbol{\mu}}(\mathbf{X}) = \mathbf{X} + g(\mathbf{X})$, where the function $\mathbf{g} = (g_1, \ldots, g_s)' : \mathbb{R}^s \to \mathbb{R}^s$ is weakly differentiable. In other words, we have that

$$\mathbb{E}_{\boldsymbol{\mu}}||\hat{\boldsymbol{\mu}}(\mathbf{X}) - \boldsymbol{\mu}||^2 = s + \mathbb{E}_{\boldsymbol{\mu}}\{||\mathbf{g}(\mathbf{X})||^2 + 2 \bigtriangledown \cdot \mathbf{g}(\mathbf{X})\}, \tag{34}$$

where

$$\bigtriangledown \cdot \mathbf{g} \equiv \sum_{i=1}^{s} \frac{\partial g_i}{\partial x_i}.$$

Using the soft thresholding rule (5), we easily see that

$$||\mathbf{g}(\mathbf{X})||^2 = \sum_{i=1}^{s} [\min(|X_i|, \lambda)]^2 \quad \text{and} \quad \bigtriangledown \cdot \mathbf{g}(\mathbf{X}) = -\sum_{i=1}^{s} 1_{[-\lambda,\lambda]}(X_i),$$

where $1_A(x)$ is the usual indicator function for any set A. Then, the following quantity

$$\text{SURE}(\lambda; \mathbf{X}) = s - 2 \cdot \#\{i : |X_i| \leq \lambda\} + [\min(|X_i|, \lambda)]^2,$$

where $\#B$ denotes the cardinality of any set B, is an unbiased estimate of the l^2-risk, i.e.,

$$\mathbb{E}_{\boldsymbol{\mu}}||\hat{\boldsymbol{\mu}}_\lambda(\mathbf{X}) - \boldsymbol{\mu}||^2 = \mathbb{E}_{\boldsymbol{\mu}}\text{SURE}(\lambda; \mathbf{X}).$$

The threshold level λ is then set so as to minimise the estimate of the l^2-risk for a given data $X_1, \ldots, X_s$, i.e.,

$$\lambda = \arg \min_{0 \leq \lambda \leq \lambda^\star} \text{SURE}(\lambda; \mathbf{X}), \tag{35}$$

where $\lambda^\star = \sqrt{2 \log s}$. By considering $\hat{d}_{jk}/\hat{\sigma} = X_i$, $s = 2^j$ and applying (35) at any level $j = j_0, \ldots, J-1$, the *SureShrink threshold* is finally given by

$$\lambda_j^{\text{S}} = \arg \min_{0 \leq \lambda \leq \lambda^{\text{U}}} \text{SURE}\left(\lambda; \frac{\hat{d}_{jk}}{\hat{\sigma}}\right), \; j = j_0, ..., J-1; \; k = 0, ..., 2^j - 1, \tag{36}$$

where, in this case, λ^{U} is given in (26) with $n = 2^j$.

The SureShrink threshold (36) has a serious drawback in situations of extreme sparsity of the wavelet coefficients. In such cases, as noted by Donoho & Johnstone (1995), "... the noise contributed to the SURE profile by the many coordinates at which the signal is zero swamps the information contributed to the SURE profile by the few coordinates where the signal is nonzero". To avoid this drawback, Donoho & Johnstone (1995) considered a *hybrid* scheme of the SureShrink threshold by the following heuristic idea: if the set of empirical wavelet coefficients is judged to be sparsely represented, then the hybrid scheme defaults to the level-wise universal threshold $\lambda_j^{\mathrm{U}} = \hat{\sigma}\sqrt{2\log(2^j)}$; otherwise the SURE criterion is used to select a threshold value. In mathematical terms, the hybrid scheme of the SureShrink threshold is expressed, for $j = j_0, ..., J-1$, as

$$\lambda_j^{\mathrm{HS}} = \begin{cases} \lambda_j^{\mathrm{U}} & \text{if } \sum_{k=0}^{2^j-1} \hat{d}_{jk}^2 \leq \hat{\sigma}^2 2^{j/2}(2^{j/2} + j^{3/2}) \\ \lambda_j^{\mathrm{S}} & \text{otherwise.} \end{cases} \tag{37}$$

Remark 4.6 *The SureShrink threshold (36) and the hybrid SureShrink threshold (37) could also be obtained for the hard thresholding rule (4). However, the latter threshold is not continuous and does not have a bounded weak derivative, meaning that a more complicated SURE formula would be required to implement the idea on a particular data set. Also, it is possible to derive SureShrink-type thresholds for other thresholding rules, for example, as the ones given in (4)–(7). However, the simplicity of the SURE formula is again lost – see, for example, Gao (1998) for a SureShrink-type threshold for the nonnegative garrote threshold (7).*

4.1.7 THRESHOLDING AS A RECURSIVE HYPOTHESIS TESTING PROBLEM

The multiple hypotheses testing approach to thresholding discussed in Section 4.1.4 produces a global threshold λ. On the other hand, Ogden & Parzen (1996a) developed a hypothesis testing procedure that produces level-dependent thresholds λ_j, as does the SureShrink methods discussed in Section 4.1.6. Rather than seeking to include as many wavelet coefficients as possible (subject to constraint) as in Abramovich & Benjamini (1995, 1996), the procedure of Ogden & Parzen (1996a) includes a wavelet coefficient only when there is strong evidence that is needed in the reconstruction.

Consider the following equivalent problem to (1) or, equivalently, to (2)–(3). Let $X_1, \ldots, X_s$ be independent $N(\mu_s, 1)$, $i = 1, \ldots, s$, random variables that represent the empirical wavelet coefficients at any level $j = j_0, \ldots, J-1$ with $s = 2^j$. Let I_s represent a non-empty subset of indices $\{1, \ldots, s\}$. Then the multiple hypotheses testing problem could be expressed as

$$H_0 : \mu_i = 0,\ i \in I_s \quad \text{versus} \quad H_1 : \mu_i \neq 0 \text{ for all } i \in I_s; \quad \mu_i = 0 \text{ for all } i \notin I_s. \tag{38}$$

The approach of Ogden & Parzen (1996a) to test the set of hypotheses given in (38) is as follows. If the cardinality of the set I_s is not known, the standard likelihood ratio test for the above hypotheses would be based on the test statistic $\sum_{i=1}^{s} X_i^2 \sim \chi_s^2$ when H_0 is true. (Note that this is also the test statistic that would be used if it were known that $I_s = \{1, \ldots, s\}$.) However, this is not the most appropriate test statistics, since it is usually assumed that very few of the μ_i's are non-zero, resulting in poor power of detection when I_s contains only a few coefficients (since the noise of the zero wavelet coefficients will tend to overwhelm the signal of the nonzero wavelet coefficients).

If the cardinality of the set I_s is known, say equal to m, then the standard likelihood ratio test statistic would be the sum of squares of the m largest X_i's. However, in practice, m is unknown, so Ogden & Parzen (1996a) suggested a recursive testing procedure for I_s containing *only* one element each time. Hence, the appropriate test statistics is the largest of the squared X_i's. The α-critical point of this distribution is shown to be equal to

$$x_s^{\alpha} = \left\{ \Phi^{-1} \left[\frac{(1-\alpha)^{1/s} + 1}{2} \right] \right\}^2, \tag{39}$$

where Φ is the cumulative distribution function of a standard normal random variable. The recursive method then for choosing a threshold λ_j at each level $j = j_0, \ldots, J-1$ consists of the following steps

1. Compare the largest X_i^2 with the critical point x_s^{α} given in (39).
2. If the X_i^2 is larger, this indicates that there still significant signal among the wavelet coefficients. Remove the X_i with the largest absolute value from consideration, set s to $s-1$, and return to Step 1.
3. If $X_i^2 < x_s^{\alpha}$, then there is no strong evidence of strong signal among the (remaining) wavelet coefficients. The threshold λ_j for the current level j is set equal to the largest remaining X_i in absolute value.

By following the above algorithm, at each level j, we are throwing out 'large' wavelet coefficients from the data set until everything left (the set of 'small' wavelet coefficients) is not distinguishable from pure noise. By setting the threshold λ_j equal to the maximum absolute value of the 'small' wavelet coefficients, we are ensuring that they will all be shrunk to zero, and that each 'large' wavelet coefficient will be included in the reconstruction, but shrunk toward zero by the same amount. In other words, this procedure has a natural interpretation as a soft thresholding rule (5) with level-dependent thresholds λ_j.

In determining the thresholds λ_j, the user has some control over the amount of smoothness that is done via the choice of α involved in (39). In general, choosing a relatively small value of α will make it very difficult for a wavelet coefficient to be judged 'significant' resulting in a smoother estimate. On the other hand, choosing a relatively large value of α makes it easier for a wavelet coefficient to be included in the reconstruction, resulting in a less smooth estimate. The recommended value is $\alpha = 0.05$.

Remark 4.7 *Although it is not considered in our simulation study in Section 5, we mention that Ogden & Parzen (1996b) also developed a recursive hypothesis testing procedure to chose level-dependent thresholds λ_j that take into account not* only *the relative magnitudes of the wavelet coefficients (as in Abramovich & Benjamini (1995, 1996) and Ogden & Parzen (1996a)), but also the* relative *position of large wavelet coefficients. Their approach adapts standard change-point methods to test the set of hypotheses given in (38) based on omnibus tests that can be used to test the null hypothesis of equal means versus a very wide variety of possible alternatives. However, as mentioned by Ogden (1997), this approach suffers somewhat from a lack of power in detecting 'large' wavelet coefficients.*

4.2 CLASSICAL METHODS: BLOCK THRESHOLDING

The wavelet thresholding procedure described in Section 3.1 achieves adaptivity through term-by-term thresholding of the empirical wavelet coefficients. There, each individual empirical wavelet coefficient is compared with a predetermined threshold; a coefficient is retained if its magnitude in absolute value is above the threshold and is discarded otherwise. This approach achieves a degree of trade-off between variance and bias contribution to mean squared error. However, this trade-off is not optimal; it removes too many terms from the empirical wavelet expansion, with the result the estimator is too biased and has a sub-optimal L^2-risk convergence rate (and also in other metrics L^p, $1 \leq p \leq \infty$).

One way to increase estimation precision is by utilising information about neighboring empirical wavelet coefficients. In other words, empirical wavelet coefficients could be thresholded in *blocks* (or groups) rather than individually. As, a result, the amount of information available from the data for estimating the "average" empirical wavelet coefficient within a block, and making a decision about retaining or discarding it, would be an order of magnitude larger than the case of a term-by-term threshold rule. This would allow threshold decisions to be made more accurately and permit convergence rates to be improved. In what follows we consider both nonoverlapping and overlapping block thresholding estimators.

4.2.1 A NONOVERLAPPING BLOCK THRESHOLDING ESTIMATOR

A nonoveralpping block thresholding estimator was proposed by Cai (1999) via the approach of ideal adaptation with the help of an oracle.

At each resolution level $j = j_0, \ldots, J-1$, the empirical wavelet coefficients $\hat{d}_{jk}$ are grouped into nonoverlapping blocks of length L. In each case, the first few empirical wavelet coefficients might be re-used to fill the last block (which is called the *Augmented* case) or the last few remaining empirical wavelet coefficients might not be used in the inference (which is called the *Truncated* case), should L not divide 2^j exactly.

Let (jb) denote the set of indices of the coefficients in the bth block at level j, that is,

$$(jb) = \{(j,k) : (b-1)L + 1 \leq k \leq bL\},$$

and let $S^2_{(jb)}$ denote the L^2-energy of the noisy signal in the block (jb). Within each block (jb), estimate the wavelet coefficients d_{jk} simultaneously via a James-Stein thresholding rule

$$\tilde{d}^{(jb)}_{jk} = \max\left(0, \frac{S^2_{(jb)} - \lambda L \sigma^2}{S^2_{(jb)}}\right) \hat{d}_{jk}. \tag{40}$$

Then, an estimate of the unknown function g is obtained by applying the IDWT to the vector consisting of both empirical scaling coefficients $\hat{c}_{j_0 k}$ ($k = 0, 1, \ldots, 2^{j_0} - 1$) and thresholded empirical wavelet coefficients $\tilde{d}^{(jb)}_{jk}$ ($j = j_0, \ldots, J-1;\ k = 0, 1, \ldots, 2^j - 1$).

Cai (1999) suggested using $L = \log n$ and setting $\lambda = 4.50524$ which is the solution of the equation $\lambda - \log \lambda - 3 = 0$. This particular threshold was chosen so that the corresponding wavelet thresholding estimator is (near) optimal in function estimation problems. The resulting block thresholding estimator was called *BlockJS*.

Remark 4.8 *Hall, Penev, Kerkyacharian & Picard (1997) and Hall, Kerkyacharian & Picard (1998, 1999) considered wavelet block thresholding estimators by first obtaining a near unbiased estimate of the L^2-energy of the true coefficients within a block and then keeping or killing all the empirical wavelet coefficients within the block based on the magnitude of the estimate. Although it would be interesting to numerically compare their estimators, they require the selection of smoothing parameters – block length and threshold level – and it seems that no specific criterion is provided for choosing these parameters in finite sample situations.*

4.2.2 AN OVERLAPPING BLOCK THRESHOLDING ESTIMATOR

Cai & Silverman (2001) considered an overlapping block thresholding estimator by modifying the nonoverlapping block thresholding estimator of Cai (1999). The effect is that the treatment of empirical wavelet coefficients in the middle of each block depends on the data in the whole block.

At each resolution level $j = j_0, \dots, J-1$, group the empirical wavelet coefficients $\hat{d}_{jk}$ into nonoverlapping blocks (jb) of length L_0. Extend each block by an amount $L_1 = \max(1, [L_0/2])$ in each direction to form overlapping larger blocks (jB) of length $L = L_0 + 2L_1$.

Let $S^2_{(jB)}$ denote the L^2-energy of the noisy signal in the larger block (jB). Within each block (jb), estimate the wavelet coefficients simultaneously via the following James-Stein thresholding rule

$$\breve{d}^{(jb)}_{jk} = \max\left(0, \frac{S^2_{(jB)} - \lambda L \hat{\sigma}^2}{S^2_{(jB)}}\right) \hat{d}_{jk}. \tag{41}$$

Then, an estimate of the unknown function g is obtained by applying the IDWT to the vector consisting of both empirical scaling coefficients $\hat{c}_{j_0 k}$ $(k = 0, 1, \dots, 2^{j_0} - 1)$ and thresholded empirical wavelet coefficients $\breve{d}^{(jb)}_{jk}$ $(j = j_0, \dots, J-1;\ k = 0, 1, \dots, 2^j - 1)$.

Cai & Silverman (2001) suggested using either $L_0 = [\log n/2]$ and taking $\lambda = 4.50524$ (which results in the *NeighBlock* estimator) or $L_0 = L_1 = 1$ (i.e., $L = 3$) and taking $\lambda = \frac{2}{3}\log n$ (which results in the *NeighCoeff* estimator). NeighBlock uses neighbouring coefficients outside the block of current interest in fixing the threshold, whilst NeighCoeff chooses a threshold for each coefficient by reference not only to that coefficient but also to its neighbours.

Remark 4.9 *The above thresholding rule (41) is different to the one given in (40) since the empirical wavelet coefficients $\hat{d}_{jk}$ are thresholded with reference to the coefficients in the larger block (jB). One can envision (jB) as a sliding window which moves L_0 positions each time and, for each window, only half of the coefficients in the center of the window are estimated.*

4.3 BAYESIAN METHODS: TERM-BY-TERM SHRINKAGE AND THRESHOLDING

The basic Bayesian framework of function estimation as discussed in Section 3.2 can now be used to obtain wavelet shrinkage and threshold estimates. Obviously, using expressions (17), (18) and(19), different losses will lead to different Bayesian rules and, therefore, to different wavelet shrinkage and threshold estimates for the unknown function g.

The problem of eliciting the hyperparameters of the prior distributions is obviously dependent on the parametric forms chosen for the prior distributions. If the hyperparameters have meaningful interpretations or if a connection can be drawn between the hyperparameters and some quantities that are easier to specify, criteria for choosing the form of the prior distributions can be obtained. Alternatively, one can employ empirical Bayes methods that attempt to determine the hyperparameters of the prior distributions from the data being analysed. These latter methods will be discussed in subsequent sections.

4.3.1 Shrinkage estimates based on L^2-losses

Clyde & George (1999, 2000) obtained wavelet shrinkage estimates by considering level-dependent posterior mean estimates. Using (17), (18) and(19), it is easily seen that L^2-based Bayes rules $\mathrm{BR}(d_{jk} \mid \hat{d}_{jk}, \sigma^2)$ correspond to marginal posterior means of wavelet coefficients d_{jk} conditionally on σ^2, given by

$$\mathbb{E}(d_{jk} \mid \hat{d}_{jk}, \sigma^2) = \frac{1}{1 + O_{jk}(\hat{d}_{jk}, \sigma^2)} \frac{\tau_j^2}{\sigma^2 + \tau_j^2}\, \hat{d}_{jk}. \tag{42}$$

Expression (42) corresponds to a level-dependent shrinkage rule. It shrinks the empirical wavelet coefficients $\hat{d}_{jk}$ by a nonlinear factor of $(1 + \tau_j^2)/((1 + O_{jk}(\hat{d}_{jk}, \sigma^2))(\sigma^2 + \tau_j^2))$.

Estimates of the hyperparameters π_j, τ_j^2 and σ^2 can now be obtained by using empirical Bayes methods which are based on using marginal maximum likelihood estimates of the hyperparameters. Marginalizing over wavelet coefficients d_{jk} and model uncertainty γ_{jk}, and conditioning on π_j, τ_j^2 and σ^2, the empirical wavelet coefficients $\hat{d}_{jk}$ are independently distributed as a mixture of two normal distributions. Define $\hat{\mathbf{d}}_j = (\hat{d}_{jk} : \ k = 0, 1, \ldots, 2^j - 1)$ for $j = j_0, \ldots, J - 1$. At each level j, the marginal log-likelihood for π_j, τ_j^2 and σ^2 is therefore, up to a constant,

$$\mathcal{L}(\pi_j, \tau_j^2, \sigma^2 \mid \hat{\mathbf{d}}_j) = \sum_{k=0}^{2^j-1} \log \left\{ \pi_j (\sigma^2 + \tau_j^2)^{-1/2} \exp\left(-\frac{\hat{d}_{jk}^2}{2(\sigma^2 + \tau_j^2)} \right) + (1 - \pi_j)\sigma^{-1} \exp\left(-\frac{\hat{d}_{jk}^2}{2\sigma^2} \right) \right\}. \tag{43}$$

Because the empirical Bayes estimates of π_j, τ_j^2 and σ^2 based on (43) are generally not available in closed forms, approximations of the marginal maximum likelihood are used which could be maximized quickly.

Maximum likelihood estimation of π_j and τ_j^2 using the EM algorithm

By estimating the noise level σ with the robust estimate (23), we now discuss an approach to obtain the maximum likelihood estimates of π_j and τ_j^2 using the EM algorithm. This approach uses a complete data (or 'augmented') likelihood and applies the EM algorithm as developed in exponential family problems.

Rather than using the marginal log-likelihood (43) at each level j, consider now the log-likelihood given the latent or 'missing' vector $\boldsymbol{\gamma}_j = (\gamma_{jk} : \ k = 0, 1, \ldots, 2^j - 1)$. This complete data log-likelihood takes the form, up to an additive constant,

$$\begin{aligned} \mathcal{L}(\pi_j, \tau_j^2 \mid \hat{\mathbf{d}}_j, \boldsymbol{\gamma}_j) &= \left[\log\left(\frac{\pi_j}{1 - \pi_j} \right) - \frac{1}{2} \log(\hat{\sigma}^2 + \tau_j^2) \right] \sum_{k=0}^{2^j-1} \gamma_{jk} \\ &+ \ 2^j \log(1 - \pi_j) + \frac{\tau_j^2}{2(\hat{\sigma}^2 + \tau_j^2)} \sum_{k=0}^{2^j-1} \gamma_{jk} \frac{\hat{d}_{jk}^2}{\hat{\sigma}^2} \end{aligned} \tag{44}$$

which belongs to a regular exponential family of the form $[a_1(\boldsymbol{\zeta}_1)]^T b_1(\mathbf{X}) + c_1(\boldsymbol{\zeta}_1) + d_1(\mathbf{X})$, where $\boldsymbol{\zeta}_1 = (\pi_j, \tau_j^2)$, $a_1(\boldsymbol{\zeta}_1)$ is the vector of natural parameters, $\mathbf{X} = (\mathbf{d}_j, \boldsymbol{\gamma}_j)$ and $b_1(\mathbf{X}) = (\sum_k \gamma_{jk}, \sum_k \gamma_{jk}\hat{d}_{jk}^2/\hat{\sigma}^2)^T$ is the vector of sufficient statistics.

We can now apply the EM algorithm developed for exponential family problems (see Dempster, Laird & Rubin (1977)) that is particularly simple to implement. Hence, using (44), we have that

- *E-step:* It consists of computing the expectations of the sufficient statistics with respect to the distribution of $\boldsymbol{\gamma}_j$ given $\hat{\mathbf{d}}_j$, π_j and τ_j^2

$$\begin{aligned}\hat{b}_1^{(i)}(\mathbf{X}) &= \mathbb{E}\left(b_1(\mathbf{X}) \mid \hat{\mathbf{d}}_j, \pi_j^{(i)}, (\tau_j^2)^{(i)}\right) \\ &= \left(\sum_{k=0}^{2^j-1} \eta_{jk}^{(i)}(\hat{d}_{jk}, \hat{\sigma}^2), \sum_{k=0}^{2^j-1} \eta_{jk}^{(i)}(\hat{d}_{jk}, \hat{\sigma}^2)\frac{\hat{d}_{jk}^2}{\hat{\sigma}^2}\right)^T,\end{aligned}$$

  where

$$\eta_{jk}^{(i)}(\hat{d}_{jk}, \hat{\sigma}^2) = \frac{1}{1 + O_{jk}^{(i)}(\hat{d}_{jk}, \hat{\sigma}^2)}$$

  are the posterior expectations of γ_{jk}, and $O_{jk}^{(i)}(\hat{d}_{jk}, \sigma^2)$ are the posterior odds ratios that $\gamma_{jk} = 0$ versus $\gamma_{jk} = 1$ given by (19) evaluated using $\hat{\sigma}^2$ and the current estimates $\pi_j^{(i)}$ and $(\tau_j^2)^{(i)}$.

- *M-step:* It consists of maximizing $[a_1(\boldsymbol{\zeta}_1)]^T \hat{b}_1^{(i)}(\mathbf{X}) + c_1(\boldsymbol{\zeta}_1)$, resulting in the solution

$$\begin{aligned}\pi_j^{(i+1)} &= \frac{\sum_{k=0}^{2^j-1} \eta_{jk}^{(i)}(\hat{d}_{jk}, \hat{\sigma}^2)}{2^j} &(45)\\ (\tau_j^2)^{(i+1)} &= \max\left(0, \frac{\sum_{k=0}^{2^j-1} \eta^{(i)}(\hat{d}_{jk}, \hat{\sigma}^2)\, \hat{d}_{jk}^2}{\sum_{k=0}^{2^j-1} \eta^{(i)}(\hat{d}_{jk}, \hat{\sigma}^2)} - \hat{\sigma}^2\right). &(46)\end{aligned}$$

Because the complete data log-likelihood (44) belongs to a regular exponential family, if the parameter estimates are in the interior of the parameter space, standard exponential family theory ensures that the solutions for π_j and τ_j^2 are the unique global solutions, conditional on the values of the latent vector $\boldsymbol{\gamma}_j$. The E-step and M-step above are repeated until the estimates converge, and yield a stationary point of the marginal log-likelihood (43).

As in the case of direct maximization of the marginal log-likelihood (43) using the Gauss-Seidel algorithm (or other algorithms) applied to (43), the EM algorithm applied to the complete data log-likelihood (44) may converge to a local mode. Also, the direct maximization methods may result in faster convergence, because the convergence rate of the EM algorithm is linear (see Dempster, Laird & Rubin (1977)). However, the iterative solutions (45), (46) using the EM algorithm are in closed form and provide some insight into the problem and connections to the conditional maximum likelihood estimates that we will discuss below. Moreover, the M-step estimate (45) of π_j has a natural interpretation: is the posterior expected fraction of nonzero wavelet coefficients.

Maximum likelihood estimation of σ, π_j and τ_j^2 using the EM algorithm

Instead of using the robust estimate (23) of σ, Clyde & George (2000) observed that the complete data log-likelihood (44) at each level j can be combined to construct the complete data log-likelihood based on all levels for estimating σ^2 using the EM algorithm. By setting $\hat{\mathbf{d}} = (\hat{d}_{jk} : j = j_0, \ldots, J-1;\ k = 0, 1, \ldots, 2^j - 1)$, $\boldsymbol{\gamma} = (\gamma_{jk} :\ j = j_0, \ldots, J-1;\ k = 0, 1, \ldots, 2^j - 1)$, $\boldsymbol{\pi} = (\pi_j :\ j = j_0, \ldots, J-1)$ and $\boldsymbol{\tau}^2 = (\tau_j^2 :\ j = j_0, \ldots, J-1)$, it is not difficult to see that this complete data log-likelihood based on all levels j takes the form, up to a constant,

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\tau}^2, \sigma^2 \mid \hat{\mathbf{d}}, \boldsymbol{\gamma}) &= \frac{1}{2} \log\left(\frac{1}{\sigma^2}\right) \sum_{j=j_0}^{J-1} \sum_{k=0}^{2^j-1} (1 - \gamma_{jk}) - \frac{1}{2\sigma^2} \sum_{j=j_0}^{J-1} \sum_{k=0}^{2^j-1} (1 - \gamma_{jk}) \hat{d}_{jk}^2 \\
&+ \frac{1}{2} \sum_{j=j_0}^{J-1} \left[\log\left(\frac{1}{\sigma^2 + \tau_j^2}\right) \sum_{k=0}^{2^j-1} \gamma_{jk} - \frac{1}{\sigma^2 + \tau_j^2} \sum_{k=0}^{2^j-1} \gamma_{jk} \hat{d}_{jk}^2 \right] \\
&+ \sum_{j=j_0}^{J-1} \left[2^j \log(1 - \pi_j) + \log\left(\frac{\pi_j}{1 - \pi_j}\right) \sum_{k=0}^{2^j-1} \gamma_{jk} \right] \qquad (47)
\end{aligned}$$

which still belongs to a regular exponential family of the form $[a_2(\boldsymbol{\zeta}_2)]^T b_2(\mathbf{X}) + c_2(\boldsymbol{\zeta}_2) + d_2(\mathbf{X})$, where $\boldsymbol{\zeta}_2 = (\boldsymbol{\pi}, \boldsymbol{\tau}^2, \sigma^2)$, $a_2(\boldsymbol{\zeta}_2)$ is the vector of natural parameters, $\mathbf{X} = (\mathbf{d}, \boldsymbol{\gamma})$ and $b_2(\mathbf{X})$ is the $(2(J - j_0) + 1) \times 1$ vector of sufficient statistics with components $(\sum_{k=0}^{2^j-1} \gamma_{jk}, \sum_{k=0}^{2^j-1} \gamma_{jk} \hat{d}_{jk}^2)$ for $j = j_0, \ldots, J-1$ and $\sum_{j=j_0}^{J-1} \sum_{k=0}^{2^j-1} (1 - \gamma_{jk}) \hat{d}_{jk}^2$.

Applying again the EM algorithm developed for exponential family problems, we have

- *E-step:* It consists of computing the expectations of the sufficient statistics with respect to the distribution of $\boldsymbol{\gamma}$ given $\hat{\mathbf{d}}$, $\boldsymbol{\pi}$, $\boldsymbol{\tau}^2$ and σ^2

  $$\hat{b}_2^{(i)}(\mathbf{X}) = \mathbb{E}\left(b_2(\mathbf{X}) \mid \hat{\mathbf{d}}, \boldsymbol{\pi}^{(i)}, (\boldsymbol{\tau}^2)^{(i)}\right)$$

  which has components

  $$\left(\sum_{k=0}^{2^j-1} \eta_{jk}^{(i)}(\hat{d}_{jk}, \sigma^2), \sum_{k=0}^{2^j-1} \eta_{jk}^{(i)}(\hat{d}_{jk}, \sigma^2)\ \hat{d}_{jk}^2 \right)^T, \quad \text{for} \quad j = j_0, \ldots, J-1,$$

and

$$\sum_{j=j_0}^{J-1} \sum_{k=0}^{2^j-1} (1 - \eta_{jk}^{(i)}(\hat{d}_{jk}, \sigma^2))\, \hat{d}_{jk}^2.$$

The quantities

$$\eta_{jk}^{(i)}(\hat{d}_{jk}, \sigma^2) = \frac{1}{1 + O_{jk}^{(i)}(\hat{d}_{jk}, \sigma^2)}$$

are the posterior expectations of γ_{jk}, where $O_{jk}^{(i)}(\hat{d}_{jk}, \sigma^2)$ are the posterior odds ratios that $\gamma_{jk} = 0$ versus $\gamma_{jk} = 1$ given by (19) evaluated using the current estimates $\pi_j^{(i)}$, $(\tau_j^2)^{(i)}$ and $(\sigma^2)^{(i)}$.

- *M-step:* It consists of maximizing $[a_2(\boldsymbol{\zeta}_2)]^T \hat{b}_2^{(i)}(\mathbf{X}) + c_2(\boldsymbol{\zeta}_2)$,resulting in the solution

$$\pi_j^{(i+1)} = \frac{\sum_{k=0}^{2^j-1} \eta_{jk}^{(i)}(\hat{d}_{jk}, \sigma^2)}{2^j} \tag{48}$$

$$(\tau_j^2)^{(i+1)} = \max\left(0, \frac{\sum_{k=0}^{2^j-1} \eta_{jk}^{(i)}(\hat{d}_{jk}, \sigma^2)\, \hat{d}_{jk}^2}{\sum_{k=0}^{2^j-1} \eta_{jk}^{(i)}(\hat{d}_{jk}, \sigma^2)} - (\sigma^2)^{(i+1)}\right). \tag{49}$$

$$(\sigma^2)^{(i+1)} = \frac{\sum_{j=j_0}^{J-1} \sum_{k=0}^{2^j-1} (1 - \eta_{jk}^{(i)}(\hat{d}_{jk}, \sigma^2))\, \hat{d}_{jk}^2}{2^{(J-j_0)} - \sum_{j=j_0}^{J-1} \sum_{k=0}^{2^j-1} \eta_{jk}^{(i)}(\hat{d}_{jk}, \sigma^2)}. \tag{50}$$

As before, because the complete data log-likelihood (47) belongs to a regular exponential family, if the parameter estimates are in the interior of the parameter space, standard exponential family theory ensures that the solutions for π_j, τ_j^2 and σ^2 are the unique global solutions, conditional on the values of the latent vector $\boldsymbol{\gamma}$. The E-step and M-step above are repeated until the estimates converge, and yield a stationary point of the marginal log-likelihood (43). The M-step estimates (48) and (50) of π_j and σ^2, respectively, have natural interpretations: (48) is the posterior expected fraction of nonzero wavelet coefficients, while (50) is the ratio of the posterior expected error sum of squares to the posterior expected degrees of freedom.

Remark 4.10 *Rather than maximizing the marginal log-likelihood $\mathcal{L}$ given by (43) (with σ estimated by the robust estimate (23)) directly, Clyde & George (2000) also used the conditional maximum likelihood approach of George & Foster (2000) which can be maximized very quickly in practice. This method takes the 'augmented' log-likelihood (44) and evaluates it at the mode for γ_{jk}, rather than using the posterior mean, as in the EM algorithm discussed above.*

However, this latter algorithm is not considered in our simulation study in Section 5. This is so because the conditional maximum likelihood estimators have the same form as the EM maximum likelihood estimators (45) and (46), and are exactly the same when the posterior distribution of γ_{jk} is degenerate at one or zero. The difference between the EM maximum *likelihood and the conditional maximum likelihood estimates will be the most extreme when the posterior mean of* γ_{jk} *is 0.5 and when* τ_j^2 *is small. However, while the EM maximum likelihood estimates of* π_j *and* τ_j^2 *appear to be asymptotically consistent (as* $2^j \to \infty$*), this is not necessarily the case with the conditional maximum likelihood estimates (see, Johnstone & Silverman, 1998). On the other hand, because the conditional maximum likelihood estimators are very rapidly computable, they can be used as starting values for the EM algorithms for computing the maximum likelihood estimates of* π_j *and* τ_j^2*.*

4.3.2 THRESHOLDING ESTIMATES BASED ON L^1-LOSSES

Abramovich, Sapatinas & Silverman (1998) obtained wavelet thresholding estimates by considering level-dependent posterior median estimates. Using (17), (18) and(19), it is easily seen that L^1-based Bayes rules $\mathrm{BR}(d_{jk} \mid \hat{d}_{jk}, \sigma^2)$ correspond to marginal posterior medians of wavelet coefficients d_{jk} conditionally on σ^2, given by

$$\mathrm{Median}(d_{jk} \mid \hat{d}_{jk}, \sigma^2) = \mathrm{sign}(\hat{d}_{jk}) \max(0, \zeta_{jk}), \tag{51}$$

where

$$\zeta_{jk} = \frac{\tau_j^2}{\sigma^2 + \tau_j^2} |\hat{d}_{jk}| - \frac{\sigma \tau_j}{\sqrt{\sigma^2 + \tau_j^2}} \Phi^{-1} \left(\frac{1 + \min(O_{jk}(\hat{d}_{jk}, \sigma^2), 1)}{2} \right)$$

and Φ is the cumulative distribution function of a standard normal random variable. The quantity ζ_{jk} is negative for all $\hat{d}_{jk}$ in some implicitly defined interval $[-\lambda_j, \lambda_j]$, and hence

$$\mathrm{Median}(d_{jk} \mid \hat{d}_{jk}, \sigma^2) = 0$$

whenever $|\hat{d}_{jk}|$ falls below the threshold λ_j. Expression (51) corresponds to a *bonafide* thresholding rule (a level-dependent 'kill' or 'shrink' thresholding rule) with thresholds λ_j. Note that, unlike the soft thresholding rule (5), extent of shrinkage in (51) depends on $|\hat{d}_{jk}|$: large $|\hat{d}_{jk}|$ are shrunk less. For large $\hat{d}_{jk}$ the thresholding rule is asymptotic to linear shrinkage by a factor of $\tau_j^2/(\sigma^2 + \tau_j^2)$, since the value of Φ^{-1} above becomes negligible as $|\hat{d}_{jk}| \to \infty$.

One can now use the iterative EM solutions (48), (49) and (50) of Clyde & George (2000) to get maximum likelihood estimates of π_j, τ_j^2 and σ^2. Alternatively, by estimating the noise level σ with the robust estimate (23), one can use the iterative EM solutions (45), (46) of Clyde & George (1999) to get maximum likelihood estimates of π_j and τ_j^2.

As discussed in Section 4.3.1, the latter approach uses a complete data (or 'augmented') likelihood and applies the EM algorithm as developed in exponential family problems. Johnstone & Silverman (1998) suggested another approach that uses an EM algorithm based on a derivation that introduces an entropy function to create a modified likelihood, where the global maxima of this modified likelihood are the global maximum likelihood estimates of the marginal log-likelihood (43). We now present this alternative method to obtain the maximum likelihood estimates of π_j and τ_j^2.

Maximum likelihood estimation of π_j and τ_j^2 using the EM algorithm – an alternative derivation

Consider the following binary entropy function

$$H(\xi) = -\xi \log \xi - (1-\xi)\log(1-\xi) \tag{52}$$

which has conjugate

$$\log(1+e^{-x}) = \sup_{0\leq\xi\leq 1} (H(\xi) - \xi x), \tag{53}$$

the maximum being attained at $\xi = 1/(1+e^{x})$. At each level j, the marginal log-likelihood (43) can therefore be rewritten as, up to a constant,

$$\mathcal{L}(\pi_j, \tau_j^2 \mid \hat{\mathbf{d}}_j) = 2^j \log(1-\pi_j) + \sum_{k=0}^{2^j-1} \log[1 + \exp\{-h(\hat{d}_{jk}, \pi_j, \tau_j^2, \hat{\sigma}^2)\}], \tag{54}$$

where

$$h(\hat{d}_{jk}, \pi_j, \tau_j^2, \hat{\sigma}^2) = \log\left(\frac{1-\pi_j}{\pi_j}\right) + \log\left(\frac{(\hat{\sigma}^2+\tau_j^2)^{1/2}}{\hat{\sigma}}\right) - \frac{\tau_j^2}{2\hat{\sigma}^2(\hat{\sigma}^2+\tau_j^2)}\,\hat{d}_{jk}^2.$$

By defining

$$\mathcal{L}^{\star}(\pi_j, \tau_j^2;\ \xi_0, \xi_1, \ldots \xi_{2^j-1}) = \sum_{k=0}^{2^j-1} [H(\xi_k) - \xi_k h(\hat{d}_{jk}, \pi_j, \tau_j^2, \hat{\sigma}^2)] + 2^j \log(1-\pi_j), \tag{55}$$

it then follows from (53) that if we maximize (55) over all its arguments, we get the maximum of the original marginal log-likelihood (43).

For fixed π_j and τ_j^2, the function (55) is clearly a sum of concave functions of the individual ξ_k ($k = 0, 1, \ldots, 2^j - 1$). The unique maxima of each of these concave functions are given by solving, over $\xi_k \in [0,1]$, the equation

$$\frac{d}{d\xi_k} H(\xi_k) = h(\hat{d}_{jk}, \pi_j, \tau_j^2, \hat{\sigma}^2),$$

which after some manipulation lead to

$$\hat{\xi}_k = \frac{1}{1 + O_{jk}(\hat{d}_{jk}, \hat{\sigma}^2)}, \tag{56}$$

where $O_{jk}(\hat{d}_{jk}, \hat{\sigma}^2)$ are the posterior odds ratios that $\gamma_{jk} = 0$ versus $\gamma_{jk} = 1$ given by (19) with σ^2 replaced by $\hat{\sigma}^2$.

We can now find the global maximum of (55) over $\pi_j \in [0,1]$ and $\tau_j^2 \in [0,\infty)$ by fixing $\hat{\xi}_k$ $(k = 0, 1, \ldots, 2^j - 1)$. In this case, the function (55) can be rewritten as a sum of two functions Q_1 and Q_2. The first function, which does not involve π_j and τ_j^2, is given by

$$Q_1 = \log \hat{\sigma}^2 \sum_{k=0}^{2^j-1} \hat{\xi}_k - \sum_{k=0}^{2^j-1} \hat{\xi}_k \log \hat{\xi}_k - \sum_{k=0}^{2^j-1} [(1-\hat{\xi}_k)\log(1-\hat{\xi}_k)],$$

while the second function, which involves π_j and τ_j^2, is given by

$$\begin{aligned} Q_2 &= \frac{1}{2}\left(\frac{\tau_j^2}{\hat{\sigma}^2(\hat{\sigma}^2+\tau_j^2)} \sum_{k=0}^{2^j-1} \hat{\xi}_k \hat{d}_{jk}^2 - \log(\hat{\sigma}^2+\tau_j^2) \sum_{k=0}^{2^j-1} \hat{\xi}_k \right) \\ &\quad + 2^j \log(1-\pi_j) + \log\left(\frac{\pi_j}{1-\pi_j}\right) \sum_{k=0}^{2^j-1} \hat{\xi}_k. \end{aligned} \tag{57}$$

The last two terms of expression (57) correspond to a concave function of π_j with maximum given by

$$\hat{\pi}_j = \frac{\sum_{k=0}^{2^j-1} \hat{\xi}_k}{2^j}. \tag{58}$$

The first term of expression (57) is not a concave function though. However, we have

$$\frac{\partial}{\partial \tau_j^2} \mathcal{L}^\star(\pi_j, \tau_j^2;\ \hat{\xi}_0, \hat{\xi}_1, \ldots, \hat{\xi}_{2^j-1}) = \frac{1}{2(\hat{\sigma}^2+\tau_j^2)^2} \left(\sum_{k=0}^{2^j-1} \hat{\xi}_k \hat{d}_{jk}^2 - (\hat{\sigma}^2+\tau_j^2) \sum_{k=0}^{2^j-1} \hat{\xi}_k \right), \tag{59}$$

the product of a strictly positive quantity and a linearly decreasing function of τ_j^2. It follows then from (59) that the global maximum of (55) over $\tau_j^2 \in [0,\infty)$ is given by

$$\hat{\tau}_j^2 = \max\left(0, \frac{\sum_{k=0}^{2^j-1} \hat{\xi}_k\, \hat{d}_{jk}^2}{\sum_{k=0}^{2^j-1} \hat{\xi}_k} - \hat{\sigma}^2\right). \tag{60}$$

By optimizing alternately over the $\hat{\xi}_k$ and over (π_j, τ_j^2), we finally see the connection with the EM algorithm. The $\hat{\xi}_k$, as given by (56), are the posterior expected values of γ_{jk} given $\hat{d}_{jk}$ and the current values of π_j and τ_j^2. The function Q_2, as given by (57), is just the expected complete data log-likelihood for (π_j, τ_j^2) given $\hat{d}_{jk}$ and the previous iterate's estimates of (π_j, τ_j^2), as obtained in (56) (compare with (44)). Thus, the estimates (58) and (60) are just the ones obtained by (45) and (46), respectively.

Remark 4.11 *Although it is not considered in our simulation study in Section 5, we mention that Abramovich, Sapatinas & Silverman (1998) have studied a particular form for the hyperparameters* π_j *and* τ_j^2 *of the prior model (13), (14), and estimated* σ *by the robust estimate*

(23). These hyperparameters depend on additional hyperparameters through the following structure

$$\tau_j^2 = C_1 2^{-\alpha j} \quad \textit{and} \quad \pi_j = \min(1, C_2 2^{-\beta j}), \quad j = j_0, \ldots, J-1,$$

where α, β, C_1 and C_2 are non-negative constants.

Some interpretation of these constants were given by Abramovich, Sapatinas & Silverman (1998) to explain how they might be derived. Part of the novelty of the approached was the idea that by simulating observations from different Besov spaces, one can elicit the correct space from which to choose the prior. In particular, the parameters α and β and the Besov space parameters were connected, so that if a particular Besov space was chosen to represent prior beliefs, α and β could be numerically derived. A weakness of this approach is that while α and β has nice interpretations, the parameters C_1 and C_2 do not have good intrinsic interpretability, and so elicitation of these parameters would be very difficult. One recommendation was to choose the parameters C_1 and C_2 by arguments related to the method of moments and, for practical applications, to choose $\alpha = 1$ and $\beta = 0.5$ that found to work well for standard test cases.

4.3.3 THRESHOLDING ESTIMATES USING A BAYESIAN HYPOTHESIS TESTING APPROACH

Similar in spirit to the multiple hypotheses testing procedures discussed in Sections 4.1.4 and 4.1.7, a Bayesian method for obtaining a *bonafide* wavelet thresholding estimator was considered by Vidakovic (1998).

For each wavelet coefficient $\hat{d}_{jk} \mid d_{jk}, \sigma^2 \sim N(d_{jk}, \sigma^2)$, this method involves testing the following hypothesis

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

according to the Bayesian framework that requires a prior distribution that has a point mass component. Otherwise, the testing is impossible because any continuous prior density will give the prior (and hence the posterior) probability of zero to the precise hypothesis – see, for example, Berger (1985). If the hypothesis H_0 is rejected, then d_{jk} is estimated by $\hat{d}_{jk}$. In each level $j = j_0, \ldots, J-1$, the prior distribution could therefore be taken as

$$d_{jk} \sim \pi_j \xi(d_{jk}) + (1-\pi_j)\delta(0), \quad k = 0, 1, \ldots, 2^j - 1, \tag{61}$$

where, as before, $\delta(0)$ is a point mass at zero and ξ describes the behaviour of d_{jk} when d_{jk} is nonzero (i.e. when H_0 is false), which occurs with probability π_j.

Considering the above setting to the prior mixture model of one normal distribution and a point mass at zero (discussed in detail in Section 3.2), and applying the usual Bayes methods for hypothesis testing, Abramovich & Sapatinas (1999) obtained the following wavelet thresholding estimator (which is called the *Bayes factor* thresholding rule since the posterior odds ratio is obtained by multiplying the Bayes factor with the prior odds ratio)

$$\tilde{d}_{jk} = \hat{d}_{jk}\chi(\eta_{jk} < 1) \quad \text{with} \quad \eta_{jk} = \frac{P(H_0 \mid \hat{d}_{jk})}{P(H_1 \mid \hat{d}_{jk})}, \tag{62}$$

where χ is the usual indicator function and η_{jk} is the posterior odds ratio that is given by (19). It essentially mimics the hard thresholding rule (4), since a wavelet coefficient $\hat{d}_{jk}$ will be thresholded if the corresponding posterior odds ratio $\eta_{jk} > 1$ and will be kept as it is otherwise. The wavelet thresholding estimate (62) is then incorporated into expressions (21) or (22) in order to get an estimate of the unknown response function g.

To apply the wavelet thresholding rule (62), the parameters π_j, τ_j^2 and σ^2 should be chosen appropriately. One, as before, could use the robust estimate (23) of σ and the iterative EM solutions (45) and (46) of Clyde & George (1999) to get maximum likelihood estimates of π_j and τ_j^2 (or, equally, the estimates (58) and (60) obtained by Johnstone & Silverman, 1998). Alternatively, the iterative EM solutions (48), (49) and (50) of Clyde & George (2000) to get maximum likelihood estimates of π_j, τ_j^2 and σ^2 could be adopted.

Remark 4.12 *To compare the Bayesian thresholding rules (51) and (62), note that the latter rule is always a 'keep' or 'kill' thresholding, whilst the former rule is a 'shrink' or 'kill' thresholding, where extend of shrinkage depends on the absolute values of the wavelet coefficients. In addition, (62) thresholds* $\hat{d}_{jk}$ *if the corresponding* $\eta_{jk} \geq 1$. *One can verify that (51) will 'kill' those* $\hat{d}_{jk}$, *whose*

$$\eta_{jk} \geq 1 - 2\Phi\left(-\frac{\tau_j|\hat{d}_{jk}|}{\sigma\sqrt{\sigma^2+\tau_j^2}}\right)$$

and, hence, it will threshold more coefficients.

4.3.4 ALTERNATIVE SHRINKAGE ESTIMATES BASED ON L^2-LOSSES

Vidakovic & Ruggeri (2001) obtained wavelet shrinkage estimates by putting a distribution on σ^2 and considering a prior distribution for the wavelet coefficients d_{jk} that is similar in spirit to the prior mixture model (13), (14).

For each wavelet coefficient $\hat{d}_{jk} \mid d_{jk}, \sigma^2 \sim N(d_{jk}, \sigma^2)$, a standard way of integrating out σ^2 is to choose the prior distribution of σ^2 to be exponential, $\sigma^2 \sim \mathcal{E}(\mu)$, where $\mu > 0$. It is well known now that the exponential distribution is the entropy maximiser among all distributions supported on $(0, \infty)$ with a fixed first moment. Thus, given the moment, the exponential prior choice on σ^2 is most uninformative.

The marginal distribution is then the double exponential, $\hat{d}_{jk} \mid d_{jk} \sim \mathcal{DE}(d_{jk}, 1/\sqrt{2\mu})$ with probability density function given by

$$f(\hat{d}_{jk} \mid d_{jk}) = \frac{\sqrt{2\mu}}{2} \exp\left(-\sqrt{2\mu}\, |\hat{d}_{jk} - d_{jk}|\right)$$

which follows from the fact that the double exponential distribution is a scale mixture of normals. Vidakovic & Ruggeri (2001) observed that by using the prior distribution $d_{jk} \sim \mathcal{DE}(0, \nu)$, the marginal distribution (predictive distribution) of $\hat{d}_{jk}$ is given by

$$f(\hat{d}_{jk}) = \frac{\mu\nu \exp\left(-|\hat{d}_{jk}|/\nu\right) - \sqrt{\mu/2} \exp\left(-\sqrt{2\mu}\, |\hat{d}_{jk}|\right)}{2\mu\nu^2 - 1} \tag{63}$$

and the corresponding posterior means of wavelet coefficients d_{jk} are given by

$$\mathbb{E}(d_{jk} \mid \hat{d}_{jk}) = \frac{\nu(\nu^2 - 1/(2\mu))\hat{d}_{jk} \exp\left(-|\hat{d}_{jk}|/\nu\right) + \nu^2(\exp\left(-|\hat{d}_{jk}|\sqrt{2\mu}\right) - \exp\left(-|\hat{d}_{jk}|/\nu\right))/\mu}{(\nu^2 - 1/(2\mu))(\nu \exp\left(-|\hat{d}_{jk}|/\nu\right) - (1/\sqrt{2\mu}) \exp\left(-|\hat{d}_{jk}|\sqrt{2\mu}\right))}. \tag{64}$$

However, it can be seen that expression (64) is close to a linear shrinkage rule known to be under-performing in wavelet-based methods.

To obtain Bayesian shrinkage rules with a more desirable shape, Vidakovic & Ruggeri (2001) considered the ϵ-contaminated priors

$$d_{jk} \sim \epsilon_j \mathcal{DE}(0, \nu) + (1 - \epsilon_j)\delta(0) \tag{65}$$

which is similar in spirit to the prior mixture model (13), (14). Under the above prior mixture model, the marginal distribution (predictive distribution) of $\hat{d}_{jk}$ is given by

$$f^{(\epsilon)}(\hat{d}_{jk}) = \epsilon_j f(\hat{d}_{jk}) + (1 - \epsilon_j)\mathcal{DE}(0, 1/\sqrt{2\mu}) \tag{66}$$

and the corresponding L^2-based Bayes rules $\mathrm{BR}(d_{jk} \mid \hat{d}_{jk}, \sigma^2)$ correspond to posterior means of wavelet coefficients d_{jk} are given by

$$\mathbb{E}^{(\epsilon)}(d_{jk} \mid \hat{d}_{jk}) = \frac{\epsilon_j f(\hat{d}_{jk})\mathbb{E}(d_{jk} \mid \hat{d}_{jk})}{\epsilon_j f(\hat{d}_{jk}) + (1 - \epsilon_j)\mathcal{DE}(0, 1/\sqrt{2\mu})}, \tag{67}$$

where $f(\hat{d}_{jk})$ and $\mathbb{E}(d_{jk} \mid \hat{d}_{jk})$ are given by (63) and (64) respectively. The shrinkage rule (67) is now 'close' to a thresholding rule – it heavily shrinks small empirical wavelet coefficients while the large ones are shrunk slightly.

Expression (67) is called by Vidakovic & Ruggeri (2001) the *Bayesian adaptive multiresolution smoother* (BAMS). In order to apply it in practice, they proposed an empirical (moment matching) specification of the parameters μ, ν and ϵ_j that works well for standard test cases and emphasized that the nature of the data may call for different parameter values. They suggest the following choices:

μ: Since μ is the reciprocal of the mean for the prior on σ^2, σ is first estimated by a robust Tukey's $\hat{\sigma} = |Q_1 - Q_3|/C$, where Q_1 and Q_3 are the first and third quartile of the finest level $J-1$ of the wavelet decomposition, and $C \in [1.3, 1.5]$. Then, μ is estimated by $\hat{\mu} = 1/\hat{\sigma}$, which according to the Strong Law of large Numbers should be close to μ.

ϵ_j: Since $1-\epsilon_j$ is the weight of the point mass at zero in the prior distribution (65), it should be closed to one at the finest level of the wavelet decomposition and zero at the coarsest levels. A good estimator of ϵ_j is then given by a hyperbolic decay $\hat{\epsilon}_j = 1/(j - j_0 + 1)^{1.5}$ for all $j = j_0, \ldots, J-1$.

ν: Since ν is the scale of the 'spread part' (which is a double exponential having variance $2\nu^2$) in the prior distribution (65) and because of the independence between the noise and the signal parts, we have the $\sigma_s^2 = 2\epsilon_j^2\nu^2 + 1/\mu$, where σ_s^2 is the variance of the sample data. Taking $2\epsilon_j^2 \approx 1$ as a mid-point of range of ϵ_j, ν is estimated by $\hat{\nu} = \sqrt{\max(0, \sigma_s^2 - 1/\mu)}$.

Remark 4.13 *Although it is not considered in our simulation study in Section 5, we mention that Vidakovic (1998) proposed wavelet shrinkage estimates by considering a symmetric prior distribution on* d_{jk} *(i.e.,* $f(d_{jk}) = f(-d_{jk})$*). Although, the choice of normal distribution is not recommended for robustness reasons, Vidakovic (1998) suggested the prior distribution* $d_{jk} \sim t_n(0, \nu)$*, a t distribution with mean zero, scaling parameter* ν *and* n *degrees of freedom. Empirical specification of the parameters* μ *and* ν *and* n *that works well for standard test cases was also suggested. The corresponding* L^2*-based shrinkage rules also shrink small empirical wavelet coefficients heavily and large only slightly. However, close expressions are not available and Monte Carlo methods, for example, could be used to approximate the integrals involved.*

4.3.5 Shrinkage estimates based on Deterministic/Stochastic Decompositions

All the Bayesian approaches described previously to obtain wavelet shrinkage and wavelet thresholding estimates, assumed a prior for each wavelet coefficient d_{jk} with zero mean. Huang & Cressie (2000) proposed a Bayesian approach that does not put such a 'strong' assumption on the prior mean but rather they estimated it and plugged it into the wavelet shrinkage formulae. Moreover, they assumed that the underlying signal is composed of a piecewise-smooth *deterministic* part plus a zero mean *stochastic* part.

Consider the vector of empirical scaling coefficients $\hat{\mathbf{c}}_{j_0} = (\hat{c}_{j_0,0}, \hat{c}_{j_0,1}, \ldots, \hat{c}_{j_0,2^{j_0}-1})'$ and the vectors of empirical wavelet coefficients $\hat{\mathbf{d}}_j = (\hat{d}_{j,0}, \hat{d}_{j,1}, \ldots, \hat{d}_{j,2^j-1})'$ for levels $j = j_0, \ldots, J-1$. Similarly, let $\mathbf{c}_{j_0} = (c_{j_0,0}, c_{j_0,1}, \ldots, c_{j_0,2^{j_0}-1})'$ be the vector of scaling coefficients and let $\mathbf{d}_j = (d_{j,0}, d_{j,1}, \ldots, d_{j,2^j-1})'$ be the vectors of wavelet coefficients for levels $j = j_0, \ldots, J-1$.

Huang & Cressie (2000) considered the following Bayesian model

$$\boldsymbol{\omega} \mid \boldsymbol{\beta}, \sigma^2 \sim N(\boldsymbol{\beta}, \sigma^2\, I), \tag{68}$$

where $\boldsymbol{\omega} = (\hat{\mathbf{c}}'_{j0}, \hat{\mathbf{d}}'_{j0}, \ldots, \hat{\mathbf{d}}'_{J-1})'$ and the signal $\boldsymbol{\beta} = (\mathbf{c}'_{j0}, \mathbf{d}'_{j0}, \ldots, \mathbf{d}'_{J-1})'$ is assumed to have a prior distribution given as

$$\boldsymbol{\beta} \mid \boldsymbol{\mu}, \boldsymbol{\theta} \sim N(\boldsymbol{\mu}, \Sigma(\boldsymbol{\theta})),$$

where $\boldsymbol{\mu} = ((\mu^{\star}_{j0})', \mu'_{j0}, \ldots, \mu'_{J-1})'$ is the deterministic mean structure and $\Sigma(\boldsymbol{\theta})$ describes the variability and the correlation in the signal. The hyperparameter $\boldsymbol{\mu}$ represents the large-scale variation (low-frequency term) in $\boldsymbol{\beta}$. In other words, one could write

$$\boldsymbol{\beta} = \boldsymbol{\mu} + \boldsymbol{\eta}, \tag{69}$$

where $\boldsymbol{\eta} \sim N(\mathbf{0}, \Sigma(\boldsymbol{\theta}))$ is the stochastic component representing the small-scale variation (high-frequency term). Note that in the Bayesian approaches described earlier, $\boldsymbol{\mu} = \mathbf{0}$.

As pointed out by Huang & Cressie (2000), the presence of both deterministic mean structure and a stochastic structure help us to recover a wider variety of signals, including piecewise-smooth signals considered in nonparametric regression and nonsmooth signals often appearing in time series analysis and spatial statistics. The identification of the deterministic μ and stochastic $\boldsymbol{\eta}$ in (69) is however an ill-posed problem. Although, in finite samples, it is possible to separate them out asymptotically with some further assumptions on $\boldsymbol{\mu}$ and $\boldsymbol{\eta}$ (see, for example, Johnstone & Silverman (1997)), it is impossible to distinguish them.

It is easily seen that the corresponding L^2-based Bayes rule corresponds to the posterior mean of $\boldsymbol{\beta}$ conditionally on σ^2, given by

$$\mathbb{E}(\boldsymbol{\beta} \mid \boldsymbol{\omega}, \sigma^2) = \boldsymbol{\mu} + \Sigma(\boldsymbol{\theta})(\Sigma(\boldsymbol{\theta}) + \sigma^2\, I)^{-1}(\boldsymbol{\omega} - \boldsymbol{\mu}), \tag{70}$$

which is a shrinkage rule (called the *DecompShrink I* method). In order to apply it in practice, Huang & Cressie (2000) suggested the following empirical specification of the parameters σ^2, $\boldsymbol{\mu}$ and $\boldsymbol{\theta}$

σ: Rather than using the robust estimate (23), σ is estimated using a method based on the variogram of the original process $\mathbf{y} = (y_1, \ldots, y_n)'$ given in (1), resulting in a more reliable estimate when the signal $\boldsymbol{\beta}$ is either deterministic (i.e $\Sigma(\boldsymbol{\theta}) = \mathbf{0}$) or stochastic (i.e. $\Sigma(\boldsymbol{\theta}) \neq \mathbf{0}$). This is so, because when the underlying signal contain stochastic components, they will confound with the noise component in all the empirical wavelet coefficients and, therefore, (23) tends to overestimate the true value of σ. The estimator of σ^2 is $\hat{\sigma}^2$, where

$$\hat{\sigma} = \begin{cases} \left(\frac{k_2\hat{\gamma}(k_1)-k_1\hat{\gamma}(k_2)}{k_2-k_1}\right)^{1/2} & \text{if } \ k_2\hat{\gamma}(k_1) \geq k_1\hat{\gamma}(k_2) \geq k_1\hat{\gamma}(k_1) \\ \left(\frac{\hat{\gamma}(k_1)+\hat{\gamma}(k_2)}{2}\right)^{1/2} & \text{if } \ \hat{\gamma}(k_2) < \hat{\gamma}(k_1) \\ 0 & \text{otherwise} \end{cases} \tag{71}$$

for $0 < k_1 < k_2$ and $2\hat{\gamma}(k)$ is the robust estimator (based on the median absolute deviation) of the variogram $2\gamma(k) = \mathbb{V}(y_{t+k} - y_t)$ at lag k. The recommended values are $k_1 = 1$ and $k_2 = 2$.

$\boldsymbol{\mu}$: Because the scaling coefficients represent low-frequency features of the underlying function, $\mu^{\star}_{j_0}$ is estimated, as before, by $\hat{\mu}^{\star}_{j_0} = \hat{\mathbf{c}}_{j_0}$; it is declared that the empirical scaling coefficients $\hat{\mathbf{c}}_{j_0}$ are deterministic (i.e. its stochastic counterpart $\hat{\boldsymbol{\eta}}^{\star}_{j_0} = \mathbf{0}$). For the wavelet coefficients at each level j ($j = j_0, \ldots, J-1$), the deterministic mean $\boldsymbol{\mu}_j$ could be considered as coming from components that are potential outliers in the normal probability plot of $\hat{\mathbf{d}}_j$ because significant mean components usually stand out among the nonzero stochastic components, which are more evenly distributed at each scale. Therefore, quantities based on normal probability plots can be used (see, equations (8) and (9) in Huang & Cressie (2000)) to obtain estimates $\hat{\boldsymbol{\mu}}_j$ of $\boldsymbol{\mu}_j$ for $j = j_0, \ldots, J-1$.

$\boldsymbol{\theta}$: The vector of hyperparameters $\boldsymbol{\theta}$ is estimated by maximum likelihood based on the marginal distribution of the data $\boldsymbol{\omega}$, with the plug-in values $\hat{\sigma}^2$ and $\hat{\boldsymbol{\mu}}$ estimated above. That is,

$$\hat{\boldsymbol{\theta}} = \arg\inf_{\boldsymbol{\theta}}\{\log\left(|\Sigma(\boldsymbol{\theta}) + \hat{\sigma}^2 I|\right) + (\boldsymbol{\omega} - \hat{\boldsymbol{\mu}})'(\Sigma(\boldsymbol{\theta}) + \hat{\sigma}^2 I)^{-1}(\boldsymbol{\omega} - \hat{\boldsymbol{\mu}})\}. \tag{72}$$

The prior covariance matrix $\Sigma(\boldsymbol{\theta})$ is assumed to be a block diagonal matrix. Specifically, the stochastic scaling coefficients $\hat{\boldsymbol{\eta}}^{\star}_{j_0} = \mathbf{0}$ (which is in line with the earlier assumption that all the scaling coefficients $\hat{\mathbf{c}}_{j_0}$ are attributed to the deterministic mean component $\boldsymbol{\mu}^{\star}_{j_0}$). At each level $j = j_0, \ldots, J-1$, the stochastic wavelet coefficients $\boldsymbol{\eta}_j$ are assumed to be independent random variables with zero means (which allows one to model $\boldsymbol{\eta}_j$ at each level j separately) with $\mathbb{V}(\boldsymbol{\eta}_j) = \sigma^2_j I$. Therefore, from (72), the components of the pseudo maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta} = (\sigma^2_{j_0}, \ldots, \sigma^2_{J-1})$ are given by

$$\hat{\sigma}^2_j = \max\left(0, \frac{(\boldsymbol{\omega}_j - \hat{\boldsymbol{\mu}}_j)'(\boldsymbol{\omega}_j - \hat{\boldsymbol{\mu}}_j)}{2} - \hat{\sigma}^2\right), \quad j = j_0, \ldots, J-1. \tag{73}$$

Remark 4.14 *Although it is not considered in our simulation study in Section 5, we mention that Huang & Cressie (2000) also dealt with the case of a more general prior covariance matrix*

$\Sigma(\boldsymbol{\theta})$ which allows us for correlation between the stochastic components $\boldsymbol{\eta}_j$ for $j = j_0, \ldots, J-1$. Binary tree structures and an optimal-prediction Kalman-filter algorithm are used to obtain recursive estimates $\hat{\eta}_{jk}$ of η_{jk} for $j = j_0, \ldots, J-1$ and $k = 0, 1, \ldots, 2^j - 1$. The resulting wavelet shrinkage estimator is called DecompShrink II. However, as pointed out by Huang & Cressie (2000), the improvement of DecompShrink II in finite samples, if any, over DecompShrink I is small. Therefore, the latter method is preferable in terms of its good performance and ease of computation.

4.3.6 THRESHOLDING ESTIMATES BASED ON NONPARAMETRIC MIXED-EFFECTS MODELS

Similar in spirit to the deterministic/stochastic decomposition approach discussed in Section 4.3.5, Huang & Lu (2000) considered wavelet thresholding estimators based on nonparametric mixed-effect models.

By considering the wavelet expansion of the unknown response function g, a prior model for g is given by

$$g(t) = \sum_{k=0}^{2^{j_0}-1} \alpha_{j_0 k}\phi_{j_0 k}(t) + \delta Z(t) \quad \text{with} \quad Z(t) \sim \sum_{j=j_0}^{\infty} \sum_{k=0}^{2^j-1} \gamma_{jk}\psi_{jk}(t), \quad t \in [0,1], \tag{74}$$

where $\alpha_{j_0 k} = \langle g, \phi_{j_0 k}\rangle$ and $\gamma_{jk} = \langle g, \psi_{jk}\rangle$ ('$\sim$' means 'equal in distribution'). The coefficients $\alpha_{j_0 k}$ are now modelled as fixed effects (which usually reflect the main features of g), whilst the coefficients γ_{jk} are modelled as random effects (which usually reflect the fine features of g). These random coefficients are assumed uncorrelated with zero mean and $\mathbb{E}(\gamma_{jk}^2) = \lambda_j$. Huang & Lu (2000) used the prior model (74) to model the relation between the prior parameters and the Besov spaces (similar in spirit to the work of Abramovich, Sapatinas & Silverman (1998) – see Remark 4.11). It can be seen that the regularity of posterior estimators can be controlled via the prior parameter λ_j (in fact, the 'posterior' of g lies in space that is smoother than the 'prior' space).

Assuming that Z is a Gaussian process, the Bayes rule under the L^2-loss is the posterior mean of g given the observations $\mathbf{y} = (y_1, \ldots, y_n)'$ from model (1). It is easily seen that this can be calculated as

$$\mathbb{E}(g(t) \mid \mathbf{y}) = \mu(t) + \frac{\delta^2}{\sigma^2}[w(t)]' M^{-1}(\mathbf{y} - \boldsymbol{\mu}), \tag{75}$$

where $\mu(t) = \sum_{k=0}^{2^{j_0}-1} \alpha_{j_0 k}\phi_{j_0 k}(t)$, $w(t)$ is the $1 \times n$ matrix defined by $W(t, t_j)$ (for any $j = 1, \ldots, n$ and any $t \in [0,1]$), $M = w + (\delta^2/\sigma^2)W$, $\boldsymbol{\mu} = (\mu(1), \ldots, \mu(n))'$, w is the $n \times n$ matrix defined by $W(t_i, t_j)$ (for any $i = 1, \ldots, n$ and any $j = 1, \ldots, n$). Recall that W is the $n \times n$ orthogonal matrix associate with the DWT discussed in Section 2.2.

If the coefficients α_{j_0k} ($k = 0, 1, \ldots, 2^{j_0} - 1$) are unknown, ones needs to estimate them from the data $\mathbf{y}$. Huang & Lu (2000) proposed a generalized least squares estimate for α_{j_0k}. It turns out that the resulting shrinkage estimator is the *best linear unbiased predictor* (BLUP) and it is equivalent to a *method of regularization estimator* (MORE). Moreover, it is asymptotically equivalent to a diagonal shrinkage estimator. When the parameters values for σ, δ and λ_j are not available, adaptive estimators are necessary. The method of regularization (as discussed in Antoniadis (1996) and Amato & Vuza (1997)) could be used to obtain the resulting wavelet shrinkage estimator.

An alternative adaptive and computationally economical *thresholding* estimator for g was suggested by Huang & Lu (2000) given by

$$\hat{g}(t) = \sum_{k=0}^{2^{j_0}-1} \hat{\alpha}_{j_0k}\phi_{j_0k}(t) + \sum_{j=j_0}^{J-1}\sum_{k=0}^{2^j-1} \max\left(0, \frac{n\hat{\gamma}_{jk}^2 - \sigma^2}{n\hat{\gamma}_{jk}^2}\right)\hat{\gamma}_{jk}\psi_{jk}(t), \quad t \in [0,1], \tag{76}$$

where $\hat{\alpha}_{j_0k} = \frac{1}{n}\sum_{i=1}^n y_i\phi_{j_0k}(t_i)$ and $\hat{\gamma}_{jk} = \frac{1}{n}\sum_{i=1}^n y_i\psi_{jk}(t_i)$. (Note that the DWT can be applied to get the estimates $\hat{\alpha}_{j_0k}$ and $\hat{\gamma}_{jk}$.) When σ^2 is unknown, Huang & Lu (2000) suggested to select its value in (76) by generalized cross-validation (with low computational cost) and the resulting estimator is called the *GGV-BLUPWAVE*.

Remark 4.15 *We mention that similar ideas to the ones discussed above have been independently explored and developed by Angelini, De Canditiis & Leblanc (2003). We note that the methods discussed in these papers could also be applied to non-equispaced designs. A fast algorithm to obtain the resulting shrinkage estimator in such cases was developed by Angelini, De Canditiis & Leblanc (2003). However, since this paper only deals with the standard nonparametric regression model as defined in (1), this algorithm is not considered in our simulation study in Section 5.*

4.4 Bayesian Methods: Block Shrinkage and Thresholding

In Section 4.2, it was shown that one way to increase estimation precision in classical term-by-term thresholding estimators was by utilising information about neighboring empirical wavelet coefficients. In other words, empirical wavelet coefficients could be thresholded in *blocks* rather than individually using *global* thresholds. This idea has been recently discussed by Abramovich, Besbeas & Sapatinas (2002) in a Bayesian framework to obtain *level-dependent* nonoverlapping block shrinkage and nonoverlapping block thresholding estimates.

Consider the model given by (2) and (3). At each resolution level j ($j = j_0, \ldots, J-1$), the wavelet coefficients d_{jk} are grouped into nonoverlapping blocks b_{jK} of length $l_j = j$. In each case, the first few empirical wavelet coefficients might be re-used to fill the last block (which is called the *Augmented* case) or the last few remaining empirical wavelet coefficients might not be used in the posterior-based inference (which is called the *Truncated* case), should l_j not divide 2^j exactly.

Let m_j be the number of blocks (i.e. $K = 1, \dots, m_j$) and consider, at each level j, the following prior model on b_{jK}

$$b_{jK} \mid \gamma_{jK} \sim \gamma_{jK} N(0, V_j) + (1 - \gamma_{jK})\delta(0), \quad K = 1, \dots, m_j, \tag{77}$$

where $\delta(0)$ is a vector of l_j point masses at zero. The matrix V_j is an $l_j \times l_j$ nonsingular covariance matrix given by $V_j = \tau_j^2 P_j$ where P_j is the $l_j \times l_j$ matrix with elements $P_j[k,l] = \rho_j^{|k-l|}$ for $k, l = 1, \dots, l_j$ with $|\rho_j| < 1$ (otherwise P_j cannot be a positive definite matrix). It is also assumed that γ_{jK} has its own prior distribution given by

$$P(\gamma_{jK} = 1) = 1 - P(\gamma_{jK} = 0) = \pi_j, \quad \text{with} \quad 0 \leq \pi_j \leq 1$$

and that, at each level j, the blocks b_{jK} $(K = 1, \dots, m_j)$ are independent. The marginal prior distribution of b_{jK} is then of the form

$$b_{jK} \sim \pi_j N(0, V_j) + (1 - \pi_j)\delta(0), \quad K = 1, \dots, m_j. \tag{78}$$

Note that, at each level j, the same prior parameters are used for all blocks b_{jK} and that all the variances (i.e diagonal elements of V_j) are equal to τ_j^2. According to the prior model (78), a block b_{jK} is either zero with probability $1 - \pi_j$, or multivariate normally distributed with zero-mean and covariance V_j. This prior model supposes that if a wavelet coefficient is non-zero (zero), then its neighboring wavelet coefficients are likely to be non-zero (zero). As in the classical approach, to complete the Bayesian model, vague priors on the scaling coefficients $c_{j_0 k}, \ k = 0, \dots, 2^{j_0} - 1$ are placed which are therefore estimated by their empirical counterparts $\hat{c}_{j_0 k}, \ k = 0, \dots, 2^{j_0} - 1$. The above model is, obviously, an extension of the prior model of one normal distribution and a point mass at zero ($l_j = 1$) discussed in Section 3.2.

At each level j, consider the corresponding blocks $\tilde{b}_{jK}$ of empirical wavelet coefficients $\hat{d}_{jk} \mid d_{jk}, \sigma^2 \sim N(d_{jk}, \sigma^2)$. Then, by combining the prior model (78) with the normal likelihood $\tilde{b}_{jK} \mid b_{jK}, \sigma^2 \sim N(b_{jK}, \sigma^2 I)$, the posterior distribution of b_{jK} conditionally on σ^2 can be expressed as

$$b_{jK} \mid \tilde{b}_{jK}, \sigma^2 \sim \frac{1}{1 + O_{jK}(b_{jK}, \sigma^2)} N(A_j \tilde{b}_{jK}, \sigma^2 A_j) + \frac{O_{jK}(b_{jK}, \sigma^2)}{1 + O_{jK}(b_{jK}, \sigma^2)} \delta(0), \tag{79}$$

where $A_j = (\sigma^2 V_j^{-1} + I)^{-1}$ and the posterior odds ratio that $\gamma_{jK} = 0$ versus $\gamma_{jK} = 1$ is given by

$$O_{jK}(b_{jK}, \sigma^2) = \frac{1 - \pi_j}{\pi_j} \sqrt{\frac{\det(V_j)}{\sigma^{2 l_j} \det(A_j)}} \exp\left\{ -\frac{\tilde{b}_{jK}' A_j \tilde{b}_{jK}}{2\sigma^2} \right\}. \tag{80}$$

The posterior (79) can be used to generate block shrinkage and block thresholding estimators using Bayes rules under L^2 and L^1-losses. Define for the jK-th block the vector $\tilde{\hat{d}}_j = A_j \tilde{b}_{jK}$ and its elements $\tilde{\hat{d}}_{jk}$. Then, we have

- **posterior means**:
  It is immediately seen that the posterior mean of b_{jK} conditionally on σ^2 is given by
  $$E(b_{jK} \mid \tilde{b}_{jK}, \sigma^2) = \frac{1}{1 + O_{jK}(b_{jK}, \sigma^2)} \tilde{\hat{d}}_j. \tag{81}$$
  This is a *block* shrinkage estimator where each empirical wavelet coefficient within a block is shrunk by the same shrinkage factor depending on all coefficients within the block. It is called the *PostBlockMean* estimator.

- **marginal posterior medians**:
  It is easily seen that, for the posterior distribution given by (79), the marginal posterior distribution of d_{jk} conditionally on σ^2 is expressed as
  $$d_{jk} \mid \tilde{b}_{jK}, \sigma^2 \sim \frac{1}{1 + O_{jK}(b_{jK}, \sigma^2)} N(\tilde{d}_{jk}, \sigma^2 A_{jj}) + \frac{O_{jK}(b_{jK}, \sigma^2)}{1 + O_{jK}(b_{jK}, \sigma^2)} \delta(0),$$
  where A_{jj} is the diagonal entry of A_j (they are the same for all k). Hence, following the arguments of Abramovich, Sapatinas & Silverman (1988) discussed in Section 4.3.2, the posterior median of $d_{jk} \mid \tilde{b}_{jK}$ is of the following closed form
  $$\mathrm{Median}(d_{jk} \mid \tilde{b}_{jK}, \sigma^2) = \mathrm{sign}(\tilde{\hat{d}}_{jk}) \max(0, \zeta_{jK}), \tag{82}$$
  where
  $$\zeta_{jK} = |\tilde{\hat{d}}_{jk}| - \sigma \sqrt{A_{jj}} \Phi^{-1} \left(\frac{1 + \min(O_{jK}(b_{jK}, \sigma^2), 1)}{2} \right).$$
  This is an *individual* thresholding estimator where each empirical wavelet coefficient is thresholded utilising information about neighbouring coefficients within a block. It is called the *PostBlockMed* estimator.

The vector $\hat{\mathbf{g}}$ of the corresponding estimates of the unknown response function g at the observed data-points can be derived by simply performing the IDWT to the vector consisting of both the empirical scaling coefficients (obtained from applying any of the previous losses on the resulting posterior distributions using the vague priors on the scaling coefficients) and the shrunk or thresholded empirical wavelet coefficients (obtained from one of (81) or (82)).

As before, estimates of the hyperparameters π_j, τ_j^2, ρ_j and σ^2 can now be obtained by using empirical Bayes methods which are based on using marginal maximum likelihood estimates of the hyperparameters. At each level j, it is easily seen that the marginal distribution of the empirical blocks $\tilde{b}_{jK}$ is a mixture of two multivariate normal distributions. Therefore, defining $\tilde{\mathbf{b}}_{jK} = (\tilde{b}_{jK} : k = 1, \ldots, m_j)$, the marginal log-likelihood function is, up to a constant

$$\begin{aligned}\mathcal{L}(\pi, \tau_j^2, \rho_j, \sigma^2 \mid \tilde{\mathbf{b}}_{jK}) &= \sum_{K=1}^{m_j} \log \left\{ \pi_j (\det(B_j))^{-1/2} \exp\left(-\frac{1}{2} \tilde{b}_{jK}' B_j^{-1} \tilde{b}_{jK} \right) \right. \\ &\quad \left. + (1-\pi_j)\sigma^{-l_j} \exp\left(-\frac{1}{2\sigma^2} \tilde{b}_{jK}' \tilde{b}_{jK} \right) \right\}, \qquad (83)\end{aligned}$$

where $B_j = \sigma^2 I + \tau_j^2 P_j$.

Since expression (83) does not lead to closed form solutions for the maximum likelihood estimates of π_j, τ_j^2, ρ_j and σ^2, numerical minimisation of $-\mathcal{L}$ in (83) must be used in order to obtain the maximum likelihood estimates of these hyperparameters. Abramovich, Besbeas & Sapatinas (2002) suggested to use the robust estimate (23) for σ and to numerically minimize $-\mathcal{L}$. The log-likelihood function was reparametrised with

$$\pi_j = \frac{1}{1 + \exp(-\theta_{1j})}, \quad \tau_j = |\theta_{2j}| \quad \text{and} \quad \rho_j = \frac{2}{\pi} \arctan(\theta_{3j})$$

so that parameter estimates would lie in the ranges

$$0 \le \hat{\pi}_j \le 1, \quad \hat{\tau}_j \ge 0 \quad \text{and} \quad -1 < \hat{\rho}_j < 1,$$

respectively. The algorithm that they have used for the minimisation of $-\mathcal{L}$ is the Nelder-Mead simplex search method which does not require first derivatives of $-\mathcal{L}$.

Remark 4.16 *We have also considered hybrid schemes by applying, on the first few resolution levels, after a fixed level (which is a user-choice), the posterior mean and median estimators discussed in Sections 4.3.1 and 4.3.2, and using the block shrinkage and thresholding estimators (discussed in this section) on the remaining resolution levels.*

We mention that Abramovich, Besbeas & Sapatinas (2002) have also considered block thresholding estimators using a Bayesian hypothesis testing approach, similar in spirit to the one discussed in Section 4.3.3. Furthermore, at each resolution level j, they have considered blocks of size $O(j)$ and have studied the effect of various block sizes in the numerical performance of the resulting empirical Bayes block shrinkage and block thresholding estimators. Since there are 2^j empirical wavelet coefficients at each resolution level j, it is often more convenient to chose the block sizes l_j to be dyadic integers; this results in block sizes that evenly divide the empirical wavelet coefficients at each resolution level j into nonoverlapping blocks, and it has been also considered by Abramovich, Besbeas & Sapatinas (2002). However, for brevity, these alternative choices have not been considered in our simulation study in Section 5.

We conclude this section with Table 2 which reports, in a synthetic way, the main denoising procedures that were discussed and their corresponding properties.

Method	*Bayes*	*Global*	*Level*	*Shrink*	*Thresh*	*Block*	$\hat{\sigma} \neq MAD$
Minimax		●			●		
SCAD		●			●		
VisuShrink		●			●		
Translation Invariant		●			●		
Multiple Hypotheses Testing		●			●		
Cross-validation		●			●		
SureShrink			●		●		
Recursive Hypothesis Testing			●		●		
BlockJS			●		●	●	
NeighBlock			●		●	●	
Single Posterior Mean	●		●	●			
Single Posterior Mean 2	●		●	●			●
Single Posterior Median	●		●		●		
Single Posterior Median 2	●		●		●		●
Bayesian Hypothesis Testing	●		●		●		
BAMS	●		●	●			●
Decompsh	●		●	●			
Mixed	●		●	●			
Blocking Posterior Median	●		●		●	●	
Hybrid Posterior Median	●		●		●	●	
Blocking Posterior Mean	●		●	●		●	
Hybrid Posterior Mean	●		●	●		●	

Table 2: Main characteristic properties of the set of denoising procedures discussed in Section 4. The column *Bayes* groups all methods relying on a Bayesian procedure. The columns *Global* and *Level* refer to the method of choice for the threshold level in wavelet thresholding which are grouped into two categories: *global thresholds* and *level-dependent thresholds*. The columns *Shrink* and *Thresh* denote the type of thersholding used, while the column *Block* refers to methods for which the wavelet coefficients are thresholded in blocks rather than term-by-term. Finally, the column $\hat{\sigma} \neq MAD$ indicates whether or not the noise level σ is estimated with the robust estimate (23).

5 DESCRIPTION OF THE SIMULATION

In all cases, the data (x_i, y_i) were generated from a model of the form

$$y_i = f(x_i) + \epsilon_i, \quad \{\epsilon_i\} \text{ i.i.d. } N(0, \sigma^2),$$

where $\{x_i\}$ are equispaced in $[0, 1]$, $x_0 = 0$ and $x_n = 1$. The factors were

1. The sample sizes n.
2. The test functions $f(x)$.
3. The values of σ^2.

For each combination of these factor levels, a simulation run was repeated 100 times holding all factor levels constant, except the $\{\epsilon_i\}$ which were regenerated. In order to compare the behavior of the various estimation methods (see Table 3), each of them based on two different wavelet filters, we have used six different criteria: MSE, L1, RMSE, RMSB, MXDV and CPU. These criteria were computed as follows:

MSE: This is the average over 100 runs of

$$\frac{1}{n}\sum_{i=1}^{n}(f(x_i) - \hat{f}(x_i))^2.$$

L1: This is the average over the 100 runs of

$$\sum_{i=1}^{n}|f(x_i) - \hat{f}(x_i)|.$$

RMSE: The mean squared error was computed for each run and averaged over the 100 runs. Then its square root was taken.

RMSB: Let $\bar{f}(x_i)$ be the average of $\hat{f}(x_i)$ over the 100 runs. The RMSB is the square root of

$$\frac{1}{n}\sum_{i=1}^{n}(f(x_i) - \bar{f}(x_i))^2.$$

MXDV: This is the average over 100 runs of

$$\max_{1 \le i \le n} |f(x_i) - \hat{f}(x_i)|.$$

CPU: This is the average over 100 runs of the CPU time.

In these simulations we have concentrated on the Symmlet 8 wavelet basis (as described on page 198 of Daubechies (1992)), and on the Coiflet 3 basis (as described on page 258 of Daubechies (1992)).

The set of test functions, f, that we have used is similar to the one used by Marron, Adak, Johnstone, Neumann & Patil (1998). This set works both quite well and also quite poorly for a variety of wavelet estimators, and is shown in Figure 5.1. Explicit formulae of these curves are given in Appendix I. For the sake of completeness, we summarize here from Marron, Adak, Johnstone, Neumann & Patil (1998) the motivation behind each of these functions. A visual idea of the noise level that we have used in this paper is also given in Figure 5.2.

1. *Step:* This function should be very hard to estimate with linear methods, because of its jumps, but relatively easy for nonlinear wavelet estimators.

2. *Wave:* This is a sum of two periodic sinusoids. Since this signal is smooth, linear methods compare favorably with non linear ones.

3. *Blip:* This is essentially the sum of a linear function with a Gaussian density, and has been often used as a target function in nonparametric regression. To make the jump induced by the assumed periodicity visually clear, the function has been periodically rotated so the jump is at x = 0.8.

4. *Blocks:* This step function has many more jumps than the Step above, and has been used in several Donoho and Johnstone papers, for example Donoho & Johnstone (1994).

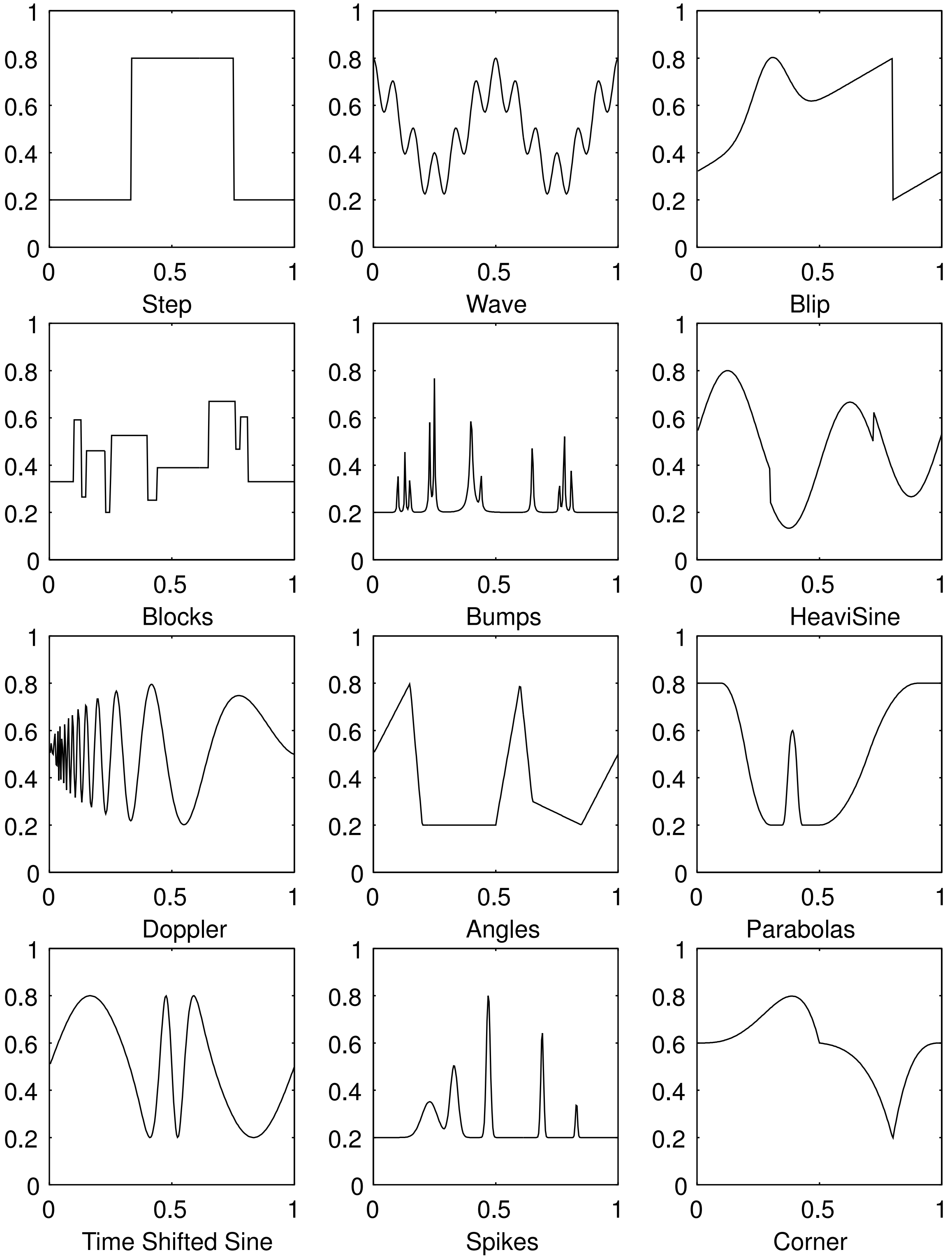
5. *Bumps:* This also comes from Donoho and Johnstone, and is very challenging for any smoother.

6. *HeaviSine:* Another Donoho and Johnstone example. This looks promising for linear smoothers, except for the two jumps.

7. *Doppler:* The final Donoho and Johnstone example. The time varying frequency makes this very hard for linear methods, with power spread all across the spectrum. It is more suitable for the wavelets, with their space time localization.

8. *Angles:* This function is piecewise linear, and continuous, but has big jumps in its first derivatives.

9. *Parabolas:* This function is piecewise parabolic. The function and its first derivative are continuous, but there are big jumps in its second derivative. It is ideal for the Symmlet 8



Step
Wave
Blip
Blocks
Bumps
HeaviSine
Doppler
Angles
Parabolas
Time Shifted Sine
Spikes
Corner

Figure 5.1: The twelve signals used in the simulation study in this paper, based on 256 design points.

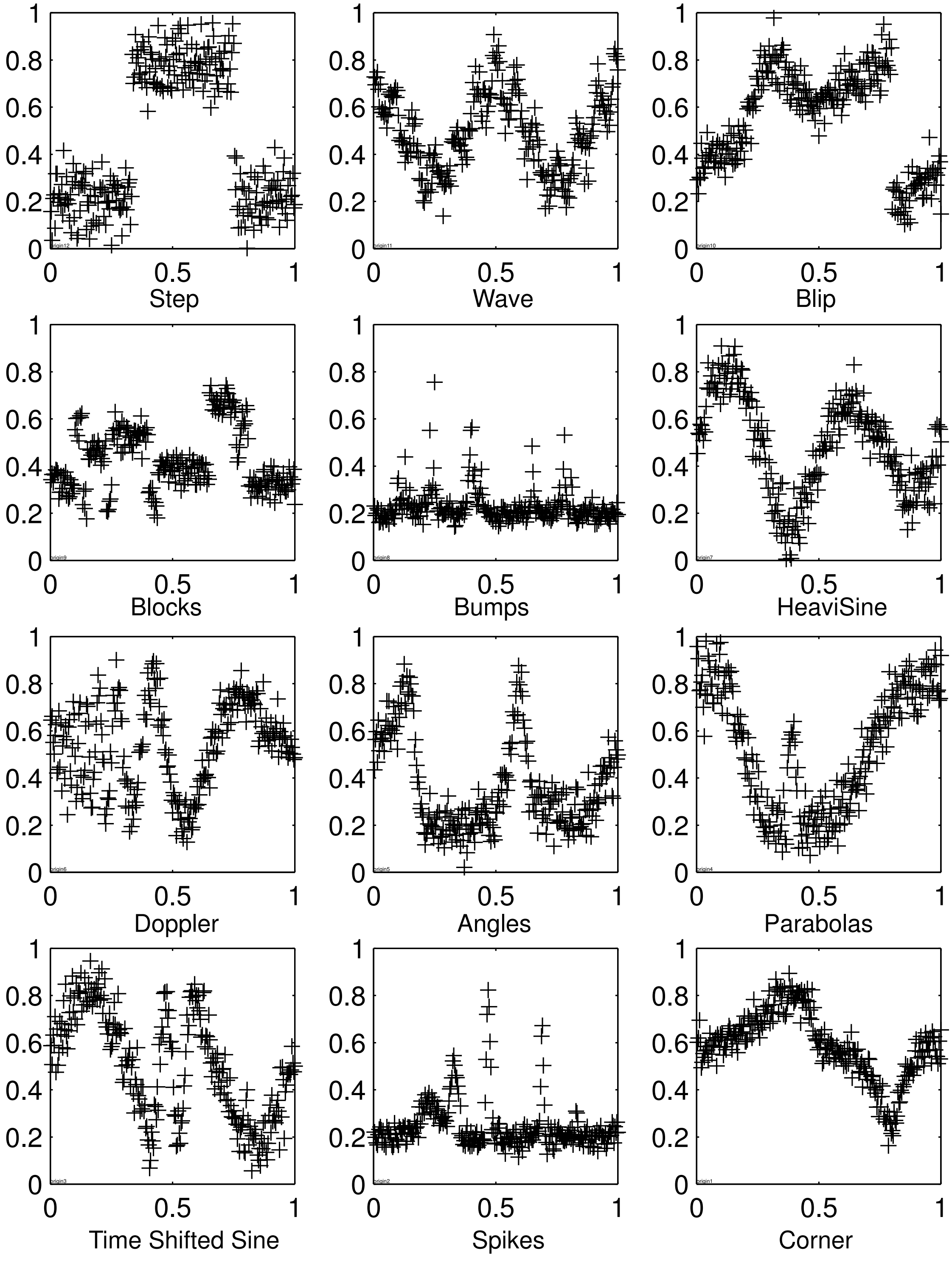


Figure 5.2: The noisy versions of the signals shown in Figure 5.1, giving a visual impression of our 'high noise' Gaussian errors setting.

bases. Estimation should be reasonable for linear methods because both the function and its first derivative are continuous.

10. *Time Shifted Sine:* This is a time shifted sine wave. It is intended to be a very smooth function, but rather far from a linear combination of sine waves. We view this as representing the type of curve that "traditional smoothers" would consider estimating.

For consistency of mnemonics, the various denoising procedures that we have used (acronyms, type of thresholding and reference to the appropriate sections of this paper) are summarized in Table 3. Insight into the performance of the various wavelet based denoising procedures described in Table 3 can be obtained from graphical outputs and numerical tables. However, the resulting graphical outputs and numerical tables across all criteria, sample sizes, test functions, noise levels, wavelet filters and denoising procedures are very extensive. Looking at our simulation results, it was apparent that the criteria L1 and MSE are closely correlated with RMSE. Also, most of the conclusions hold whatever wavelet filter is used. Hence, for reasons of space, we only report, in Appendix II, in detail the results for two samples sizes, $n = 128$ (a moderate sample size) and $n = 512$ (a large sample size), for four criteria (RMSE, RMSB, MXDV and CPU), a particular wavelet filer (Symmlet 8), a root-signal-to-noise ratio rsnr $= 3$ (a high noise level), for all test functions and all smoothing procedures. Even so, the graphical outputs summarizing the behaviour of the smoothers with respect to the various criteria are quite extensive. Graphical outputs for other combinations of sample sizes and noise levels, as well as numerical tables, can be derived using appropriate Matlab scripts. These scripts are based on Version 8 of the Wavelab toolbox for MATLAB (Buckheit, Chen, Donoho, Johnstone & Scargle (1995)) and are available at `http://www-lmc.imag.fr/SMS/software/GaussianWaveDen.html` or `http://www.ucy.ac.cy/∼fanis/links/software.html` . We refer at this point to Section 8 for more details on which files the user should download and how to start running them.

We remark below on some of the conclusions we drew after a careful examination on the simulations' output.

6 Summary of Results

As expected, for the same test function and denoising procedure, whatever wavelet filter is used, the RMSE is roughly proportional to the inverse of the square root of the sample size. In terms of RMSE, TI-H, BAMS and DECOMPSH are markedly better on the *Step* function, whatever the sample size is, and the results are quite similar with respect to the filter used. As for computational efficiency, for small sample sizes, TI-H and BAMS are much more efficient

1	VISU-H	VisuShrink	Hard	4.1.2
2	VISU-S	VisuShrink	Soft	4.1.2
3	SURE	SureShrink		4.1.6
4	HYBSURE	SureShrink	Hybrid	4.1.6
5	TI-H	Translation-Invariant	Hard	4.1.3
6	TI-S	Translation-Invariant	Soft	4.1.3
7	MINIMAX-H	Minimax	Hard	4.1.1
8	MINIMAX-S	Minimax	Soft	4.1.1
9	CV-H	Cross-Validation	Hard	4.1.5
10	CV-S	Cross-Validation	Soft	4.1.5
11	NEIGHBL	NeighBlock		4.2.2
12	BLOCKJS-A	Block Thresholding	Augment	4.2.1
13	BLOCKJS-T	Block Thresholding	Truncate	4.2.1
14	THRDA1	Hypothesis Testing	Soft	4.1.7
15	FDR-H	False Discovery Rate	Hard	4.1.4
16	FDR-S	False Discovery Rate	Soft	4.1.4
17	PENWAV	Linear Penalization		3.1
18	SCAD	Nonlinear Penalization	Hybrid	3.1 & 4.1.1
19	DECOMPSH	Deterministic/Stochastic		4.3.5
20	MIXED	Mixed-Effects		4.3.6
21	BLMED-A	Blocking Posterior Median	Augment	4.4
22	BLMED-T	Blocking Posterior Median	Truncate	4.4
23	HYBMED-A	Hybrid Blocking Median	Augment	4.4
24	HYBMED-T	Hybrid Blocking Median	Truncate	4.4
25	BLMEAN-A	Blocking Posterior Mean	Augment	4.4
26	BLMEAN-T	Blocking Posterior Mean	Truncate	4.4
27	HYBMEAN-A	Hybrid Blocking Mean	Augment	4.4
28	HYBMEAN-T	Hybrid Blocking Mean	Truncate	4.4
29	SINGLMED	Single Posterior Median		4.3.2
30	SINGLMED-2	Single Posterior Median		4.3.2
31	SINGLMEAN	Single Posterior Mean		4.3.1
32	SINGLMEAN-2	Single Posterior Median		4.3.1
33	SINGLHYP	Bayesian Hypothesis Testing		4.3.3
34	BAMS	Bayesian Adaptive Multiresolution		4.3.4

Table 3: Acronyms for the set of denoising procedures applied in the simulation study. The section column refers to the actual sections where these procedures have been defined.

than DECOMPSH, while for large sample sizes the three methods are equivalent. One would expect a similar conclusion with respect to the *Blocks* function. While this is true for BAMS and DECOMPSH, CV-S outperforms TI-H in this case in terms of RMSE. Note also that Symmlets are better suited than Coiflets in this case, which is mainly due to the presence of many jumps in the *Blocks* function.

For the smooth function *Wave*, most Bayesian methods do well. Among the non-Bayesian denoising procedures, PENWAV and TI-H are competitive, especially for large sample sizes. The fact that PENWAV is fine is not surprising given the fact that this is the type of functions where linear methods are generally equivalent to nonlinear ones. Similar conclusions hold also for the *Time Shifted Sine* function.

For the *Blip* function as well as for the *Parabolas* function, TI-H is markedly better than other methods, and the use of Coiflets improves the RMSE. At the other end, SURE, MIXED and SINGLHYP perform poorly for these test functions.

For the *Bumps*, *Angles* and *Spikes* functions almost all Bayesian procedures are equivalent (with the exception of SINGLHYP) and do markedly better than the other procedures in terms of RMSE. Among the non-Bayesian procedures, once again TI-H is fine. Both wavelet filters lead to similar results. For such functions it is therefore advisable to denoise using Bayesian procedures, if computational cost is not an issue.

Finally, for the *Heavisine* and *Doppler* functions almost all procedures give equivalent results, with the exception of SURE and SINGLHYP which perform poorly.

Criterion MXDV allows a finer comparison between Bayesian methods when these are in competition. Larger values of MXDV occur for functions with many spikes or discontinuities, but this is expected. The curious behavior of MXDV for some of the methods with the *Bumps* signal calls for some explanation. Recall that, throughout the simulations, the primary resolution level $j_0 = [\log_2(\log(n))] + 1$ was used for all methods. This value of j_0 affects whether or not the spikes in the *Bumps* signal are felt in the lowest level of wavelet coefficients. For $j_0 = 3$, the standard methods, especially PENWAV and DECOMPSH both smooth out the spike effect to a big extent.

Among all denoising procedures and almost all test functions, the minimum RMSB is achieved by the SURE procedure, which shows that the bias in the procedures is almost never a substantial contributor to RMSE, reflecting the capability of these automatic denoising procedures to fit a large variety of functions.

In terms of the mean squared error criterion, a conceivable competitor to SURE among the other methods is NEIGHBLOCK, especially when the underlying function is of significant spatial variability.

The strange behaviour of some of the methods with the Waves signal is probably due to the fact that for all methods the same primary resolution level $j_0 = [\log_2(\log(n))] + 1$ was used, and most methods smooth to some extent the high frequencies in the Waves signal.

Table 4 in the Appendix II shows the average of the CPU time involved in computing the estimates for the *Corner* function by each method and the two sample sizes. Our simulations show that non-Bayesian methods uniformly outperform Bayesian methods in terms of CPU time in all examples, and indeed the relative performance of Bayesian procedures is even worse for some other examples than the *Corner* presented in detail.

7 Overall Conclusions

As expected, no wavelet based denoising procedure uniformly dominates in all aspects. For larger sample sizes and when a function is expected to be mainly smooth, Coiflets lead to better results due to the fact that scaling functions associated to Coiflets have better approximation properties that other Daubechies filters. While Bayesian methods perform reasonably well at small sample sizes for relatively inhomogeneous functions, their computational cost may be a handicap, when compared with translation invariant thresholding procedures.

8 An illustrative example

This section explains in some detail which files the user should download and how to start running them. In order to provide some hints of how the functions should be used to analyse real data sets, a detailed practical step-by-step illustration of a wavelet denoising analysis on electrical consumption is provided.

Prerequisites

As already noted, our library makes an extensive use of the `Matlab` routines available in the `WaveLab` package developed by Buckheit, Chen, Donoho, Johnstone & Scargle (1995) at Stanford University. `WaveLab` has over 800 subroutines which are well documented, indexed and cross-referenced. The library is available, *free of charge*, over the Internet World Wide Web (WWW) access. Versions are provided for `Macintosh`, `Unix` and `Windows` platforms. The `WaveLab` package is made available as a compressed archive, in a format suitable for the machine in question: `.zip` (for MS-Windows), `.tar.Z` (for Unix) and `.sea.hqx` (for Macintosh). The archives may be accessed by WWW access to `http://www-stat.stanford.edu/~ wavelab`.

Once the appropriate compressed archive has been transferred to your machine, it should be decompressed, with the relevant tools, and installed. On a personal computer (`Macintosh` or `Windows`), the archives should be decompressed and installed as a subdirectory of the toolbox directory inside the `Matlab` folder. On a `Unix` workstation or server, the archives could either be installed in the systemwide `Matlab` directory, if you have permission to do this, or in your own personal `Matlab` directory, if you do not. Once the actual files are installed, you should have a number of subdirectories of `.m` files in the directory `WaveLab`. `Matlab` can automatically, at startup time, make all the `WaveLab` package available (read the installations instructions accompanying the archive). We will assume from now on that the `Wavelab` routines are available in your machine. We will also assume that you have already downloaded our `GaussianWaveDen` package from `http://www-lmc.imag.fr/SMS/software/GaussianWaveDen` or `http://www.ucy.ac.cy/∼fanis/links/software.html` and you have installed its subroutines by specifying its path in `Matlab`. Once this is done you can try running the demo that we will describe in this section. Each function in `GaussianWaveDen` has an accompanying `html` help documentation in the directory `html-help`.

An electrical consumption example

The example we present here involves a real-word signal – electrical consumption measured over the course of three days. This signal is particularly interesting because of noise introduced whenever a defect is present in the monitoring equipment. The data consist of measurement of a complex, highly-aggregated plant: the electrical load consumption, sampled minute by minute, over a 5-week period. The resulting time series of 50,400 points is partly plotted in the top panel of Figure 8.3. External information is given by electrical engineers, and additional indications can be found in Misiti, Misiti, Oppenheim & Poggi (1994). This information includes the following remarks:

- The load curve is the aggregation of hundreds of sensors measurements, thus generating measurement errors.
- The consumption is accounted for 50% by industry and for the other half by individual consumers. The component of the load curve produced by industry has a rather regular profile and exhibits low-frequency changes. On the other hand, the consumption of individual consumers may be highly irregular, leading to high-frequency components.
- There are more than 10 millions individual consumers.

- Daily consumption patterns also change according to rate changes at different times (e.g. relay-switched water heaters to benefit from special night rates).

- For the 3-day observations, indexed from 1 to 4096, the measurement errors are unusually high, due to sensors failures.

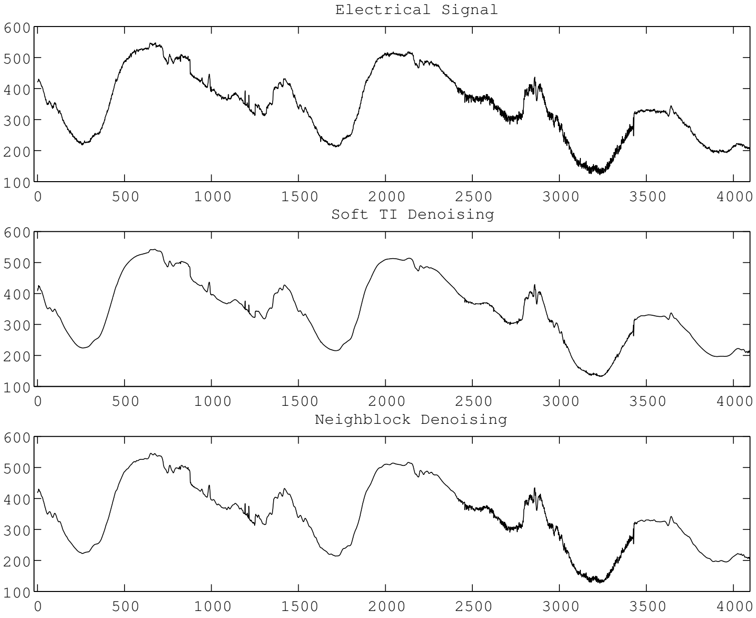


Figure 8.3: An electricity consumption signal and its denoised versions.

We shall not report here a complete analysis which is included in Misiti, Misiti, Oppenheim & Poggi (1994). We only want to illustrate some of the denoising procedures developed in this paper to the local description of this time series, which effectively remove the noise. We choose a portion of the sample signal corresponding to a midday period. Observe, however, that the midday period has a complicated structure because the intensity of the electricity consumers activity is high and it presents very large changes. An appropriate noise removal allows the identification of interesting features of the data.

Figure 8.3 has been generated using the following `Matlab` command lines.

```
% Load the original 1-D signal, choose a portion of it and plot it.
%
% load the signal in s.
s=file2var('eleccum.dat');
% fix the time axis.
```

```
x=(1:4096);
signal=s(x);
% Plot the original signal
plot(x,signal); title('Electrical Signal');
% Denoise the plotted portion of the signal using few of our procedures.
% Using the Translation Invariant procedure with soft thresholding.
f = recTI(signal,'S');
% Using the Neighblock procedure.
g=recneighblock(signal);
% Plot the results.
subplot(3,1,1);
plot(x,signal); axis([-20 4097 100 600]); title('Electrical Signal');
subplot(3,1,2);
plot(x,f); axis([-20 4097 100 600]); title('Soft TI Denoising')
subplot(3,1,3);
plot(x,g); axis([-20 4097 100 600]); title('Neighblock Denoising')
```

One may note on the denoised signal the abrupt changes due to automatic switches. Note also that the TI-S procedure produces a smooth fit, removing efficiently the massive and high frequency changes of personal electric appliances in the consumption, while the NEIGHBLOCK procedure undersmooths the high frequency portions of the observed signal.

To end this section, we would like to mention here a few types of signal processing problems where the wavelet methods discussed and compared in this paper have been used in practice. Wavelet denoising procedures have been found to be a particularly useful tool in machinery fault detection (see Staszewski & Tomlinson (1994) and Lin & McFadden (1997)). Typical examples of signals encountered in this field are vibration signals generated in defective bearings and gears rotating at constant speeds. When a machine or its parts change from one state into another, transients may be seen in the vibration signals. Transients usually have relatively high frequencies but relatively small time scales and contain rich information about machinery conditions. This explains why wavelet denoising procedures provide considerable improvement over certain traditional techniques in the fault detection of mechanical systems.

Wavelet denoising procedures, in conjunction with hypothesis testing, have also been used for detecting change points in several biomedical applications. Typical examples are the detection of life-threatening cardiac arrythmia (see Khadra, Al-Fahoum & Al-Nashash (1997)) in electrocardiographic signals (ECG) recorded during the monitoring of patients, or the detection of venous air embolism in doppler heart sound signals recorded during surgery when the incision wounds lie above the heart (see Chan, Chan, Lam, Lui & Poon (1997)). They also have been used in astronomical application for estimating periodicities in the light-curves of the variable star R Aquilae, after appropriate denoising (see Foster (1996)).

A number of interesting applications of wavelets may be also found in economic and financial applications. Ramsay, Usikov & Zaslavsky (1995) provides an account on research arising from earlier concerns in the analysis of the stock market.

In conclusion, it is apparent that wavelets are particularly well adapted to the statistical analysis of several types of data, and denoising tools, like the ones presented in this paper, will certainly be of great help in revealing features present in data.

Acknowledgements

This work was supported by the IMAG project AMOA whose financial support is acknowledged. The authors would like to thank Jean-Michel Poggi for his suggestions and comments on a preliminary version of this paper. Theofanis Sapatinas would like to thank Anestis Antoniadis for financial support and excellent hospitality while visiting Grenoble to carry out this work. Helpful comments made by the editor and an anonymous referee are gratefully acknowledged.

References

[1] Abramovich, F., Bailey, T.C. & Sapatinas, T. (2000). Wavelet analysis and its statistical applications. *The Statistician*, **49**, 1–29.

[2] Abramovich, F. & Benjamini, Y. (1995). Thresholding of wavelet coefficients as multiple hypotheses testing procedure. In *Wavelets and Statistics*, Antoniadis, A. & Oppenheim, G. (Eds.), Lecture Notes in Statistics **103**, pp. 5–14, New York: Springer-Verlag.

[3] Abramovich, F. & Benjamini, Y. (1996). Adaptive thresholding of wavelet coefficients. *Comput. Statist. Data Anal.*, **22**, 351–361.

[4] Abramovich, F., Besbeas, P. & Sapatinas, T. (2002). Empirical Bayes approach to block wavelet function estimation. *Comput. Statist. Data Anal.*, **39**, 435–451.

[5] Abramovich, F. & Sapatinas, T. (1999). Bayesian approach to wavelet decomposition and shrinkage. In *Bayesian Inference in Wavelet Based Models*, Müller, P. & Vidakovic, B. (Eds.), Lect. Notes Statist., **141**, pp. 33–50, New York: Springer-Verlag.

[6] Abramovich, F. & Silverman, B.W. (1998). The vaguelette-wavelet decomposition approaches to statistical inverse problems. *Biometrika*, **85**, 115–129.

[7] Abramovich, F., Sapatinas, T. & Silverman, B.W. (1998). Wavelet thresholding via a Bayesian approach. *J. R. Statist. Soc. B*, **60**, 725–749.

[8] Abramovich, F., Benjamini, Y., Donoho, D.L. & Johnstone, I.M. (2000). Adapting to unknown sparsity by controlling the false discovery rate. *Technical Report* **00-19**, Department of Statistics, Stanford University, USA.

[9] Amato, U. & Vuza, D.T. (1997). Wavelet approximation of a function from samples affected by noise. *Rev. Roumanie Math. Pure Appl.*, **42**, 481–493.

[10] Angelini, C., De Canditiis, D. & Leblanc, F. (2003). Wavelet regression estimation in nonparametric mixed effect models. *J. Multivariate Anal.*, **85**, 267–291.

[11] Antoniadis, A. (1996). Smoothing noisy data with tapered coiflets series. *Scand. J. Statist.*, **23**, 313–330.

[12] Antoniadis, A. (1997). Wavelets in statistics: a review (with discussion). *J. Ital. Statist. Soc.*, **6**, 97–144.

[13] Antoniadis, A. & Fan, J. (2001) Regularization of wavelets approximations (with discussion). *J. Am. Statist. Ass.*, **96**, 939–967.

[14] Antoniadis, A. & Oppenheim, G. (Eds.) (1995). *Wavelets and Statistics*, Lect. Notes Statist., **103**, New York: Springer-Verlag.

[15] Antoniadis, A. & Pham, D.T. (1998). Wavelet regression for random or irregular design. *Computat. Statist. Data Anal.*, **28**, 353–369.

[16] Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, **57**, 289–300.

[17] Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis.* New York: Springer-Verlag

[18] Breiman, L. & Peters, S. (1992). Comparing automatic smoothers (a public service enterprise). *Int. Statist. Rev.*, **60**, 271–290.

[19] Bruce, A.G. & Gao, H.-Y. (1996). Understanding WaveShrink: variance and bias estimation. *Biometrika*, **83**, 727–745.

[20] Buckheit, J.B. & Donoho, D.L. (1995). WaveLab and Reproducible Research. In *Wavelets and Statistics*, Antoniadis, A. & Oppenheim, G. (Eds.), Lect. Notes Statist., **103**, pp. 55–81, New York: Springer-Verlag.

[21] Buckheit, J.B., Chen, S., Donoho, D.L., Johnstone, I.M. & Scargle, J. (1995). About WaveLab. *Technical Report*, Department of Statistics, Stanford University, USA.

[22] Burrus, C.S., Gonipath, R.A. & Guo, H. (1998). *Introduction to Wavelets and Wavelet Transforms: A Primer.* Englewood Cliffs: Prentice Hall.

[23] Cai, T.T. (1999). Adaptive wavelet estimation: a block thresholding and oracle inequality approach. *Ann. Statist.*, **27**, 898–924.

[24] Cai, T.T. & Brown, L.D. (1998). Wavelet shrinkage for nonequispaced samples. *Ann. Statist.*, **26**, 1783-1799.

[25] Cai, T.T. & Brown, L.D. (1999). Wavelet estimation for samples with random uniform design. *Statist. Probab. Lett.*, **42**, 313–321.

[26] Cai, T.T. & Silverman, B.W. (2001). Incorporating information on neighbouring coefficients into wavelet estimation. *Sankhyā B*, **63**, 127–148.

[27] Chan, C.B., Chan, F.H., Lam, F.K., Lui, P.-W. & Poon, P.W. (1997). Fast detection of venous air embolism in doppler heart sound using the wavelet transform. *IEEE Trans. Biomed. Eng.*, **44**, 237–246.

[28] Chipman, H.A., Kolaczyk, E.D. & McCulloch, R.E. (1997). Adaptive Bayesian Wavelet Shrinkage. *J. Am. Statist. Ass.*, **92**, 1413–1421.

[29] Clyde, M. & George, E.I. (1999). Empirical Bayes estimation in wavelet nonparametric regression. In *Bayesian Inference in Wavelet Based Models*, Müller, P. & Vidakovic, B. (Eds.), Lect. Notes Statist., **141**, pp. 309–322, New York: Springer-Verlag.

[30] Clyde, M. & George, E.I. (2000). Flexible empirical Bayes estimation for wavelets. *J. R. Statist. Soc. B*, **62**, 681–698.

[31] Clyde, M., Parmigiani, G. & Vidakovic, B. (1998). Multiple shrinkage and subset selection in wavelets. *Biometrika*, **85**, 391–401.

[32] Coifman, R.R. & Donoho, D.L. (1995). Translation-invariant de-noising. In *Wavelets and Statistics*, Antoniadis, A. & Oppenheim, G. (Eds.), Lect. Notes Statist., **103**, pp. 125–150, New York: Springer-Verlag.

[33] Crouse, M.S., Nowak, R.D. & Baraniuk, R.G. (1998). Wavelet-based statistical signal processing using hidden Markov models. *IEEE Trans. Sig. Proc.*, **46**, 886–902.

[34] Daubechies, I. (1992). *Ten Lectures on Wavelets.* Philadelphia: SIAM.

[35] Delyon, B. & Juditsky, A. (1995). Estimating wavelet coefficients In *Wavelets and Statistics*, Antoniadis, A. & Oppenheim, G. (Eds.), Lect. Notes Statist., **103**, pp. 15–168, New York: Springer-Verlag.

[36] Delyon, B. & Juditsky, A. (1996). On minimax wavelet estimators. *Appl. Comput. Harm. Anal.*, **3**, 215–228.

[37] Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc. B*, **39**, 1–38.

[38] Dennis, J.E. & Mei, H.H.W. (1979). Two new unconstrained optimization algorithms which use function and gradient values. *J. Optimiz. Theory Appl.*, **28**, 453–483.

[39] Donoho, D.L. & Johnstone, I.M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**, 425–455.

[40] Donoho, D.L. & Johnstone, I.M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Am. Statist. Ass.*, **90**, 1200–1224.

[41] Donoho, D.L. & Johnstone, I.M. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.*, **26**, 879–921.

[42] Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. & Picard, D. (1995). Wavelet shrinkage: asymptopia? (with discussion). *J. R. Statist. Soc. B*, **57**, 301–337.

[43] Efromovich, S. (1999). Quasi-linear wavelet estimation. *J. Am. Statist. Ass.*, **94**, 189–204.

[44] Efromovich, S. (2000). Sharp linear and block shrinkage wavelet estimation. *Statist. Probab. Lett.*, **49**, 323–329.

[45] Eubank, R.L. (1999). *Nonparametric Regression and Spline Smoothing.* 2nd Edition, New York: Marcel Dekker.

[46] Fan, J. & Gijbels, I. (1996). *Local Polynomial Modelling and its Applications.* London: Chapman & Hall.

[47] Foster, G. (1996). Wavelets for period analysis of unevenly sampled time series. *Astron. J.*, **112**, 1709–1729.

[48] Gao, H.-Y. & Bruce, A.G. (1997). WaveShrink with firm shrinkage. *Statist. Sinica*, **7**, 855–874.

[49] Gao, H.-Y. (1998). Wavelet shrinkage denoising using the non-negative garrote. *J. Comp. Graph. Statist.*, **7**, 469–488.

[50] George, E.I. & Foster, D.P. (2000). Calibration and empirical Bayes variable selection. *Biometrika*, **87**, 731–747.

[51] George, E.I. & McCulloch, R. (1993). Variable selection via Gibbs sampling. *J. Am. Statist. Ass.*, **88**, 881–889.

[52] Green, P.J. & Silverman, B.W. (1994). *Nonparametric regression and generalised linear models.* London: Chapman & Hall.

[53] Hall, P., Kerkyacharian, G. & Picard, D. (1998). Block threshold rules for curve estimation using kernel and wavelet methods. *Ann. Statist.*, **26**, 922–942.

[54] Hall, P., Kerkyacharian, G. & Picard, D. (1999). On the minimax optimality of block thresholded wavelet estimators. rules for curve estimation using kernel and wavelet methods. *Statist. Sinica.*, **9**, 33–50.

[55] Hall, P., Penev, S., Kerkyacharian, G. & Picard, D. (1997). Numerical performance of block thresholded wavelet estimators. *Statist. Comput.*, **7**, 115–124.

[56] Hall, P. & Nason, G.P. (1997). On choosing a non-integer resolution level when using wavelet methods. *Statist. Probab. Lett.*, **34**, 5–11.

[57] Hall, P. & Patil, P. (1996a). Effect of threshold rules on performance of wavelet-based curve estimators. *Statist. Sinica*, **6**, 331–345.

[58] Hall, P. & Patil, P. (1996b). On the choice of smoothing parameter, threshold and truncation in nonparametric regression by non-linear wavelet methods. *J. R. Statist. Soc. B*, **58**, 361–377.

[59] Hall, P. & Turlach, B.A. (1997). Interpolation methods for nonlinear wavelet regression with irregularly spaced design. *Ann. Statist.*, **25**, 1912–1925.

[60] Härdle, W. (1990). *Applied Nonparametric Regression.* Cambridge: Cambridge University Press.

[61] Härdle, W., Kerkyacharian, G., Pikard, D. & Tsybakov, A. (1998). *Wavelets, Approximation, and Statistical Applications.* Lecture Notes in Statistics **129**, New York: Springer-Verlag.

[62] Huang, H.-C. & Cressie, N. (2000). Deterministic/stochastic wavelet decomposition for recovery of signal from noisy data. *Technometrics*, **42**, 262–276.

[63] Huang, S.Y. & Lu, H.H.-S. (2000). Bayesian wavelet shrinkage for nonparametric mixed-effects models. *Statist. Sinica*, **10**, 1021–1040.

[64] Jansen, M., Malfait, M. & Bultheel, A. (1997). Generalised cross-validation for wavelet thresholding. *Sig. Proc.*, **56**, 33–44.

[65] Johnstone, I.M. (1994). Minimax Bayes, asymptotic minimax and sparse wavelet priors. In *Statistical Decision Theory and Related Topics, V*, Gupta, S.S. and Berger, J.O. (Eds.), pp. 303–326, New York: Springer-Verlag.

[66] Johnstone, I.M. & Silverman, B.W. (1997). Wavelet threshold estimators for data with correlated noise. *J. R. Statist. Soc. B*, **59**, 319–351.

[67] Johnstone, I.M. & Silverman, B.W. (1998). Empirical Bayes approaches to mixture problems and wavelet regression. *Technical Report*, School of Mathematics, University of Bristol, UK.

[68] Khadra, L., Al-Fahoum, A.S. & Al-Nashash, H. (1997). Detection of life-threatening cardiac arrhythmias using the wavelet transformation. *Med. Biol. Eng. Comput.*, **35**, 626–632.

[69] Kovac, A. & Silverman, B.W. (2000). Extending the scope of wavelet regression methods by coefficient-dependent thresholding. *J. Am. Statist. Ass.*, **95**, 172–183.

[70] Lang, M., Guo, H., Odegard, J.E., Burrus, C.S. & Wells Jr, R.O. (1996). Noise reduction using an undecimated discrete wavelet transform. *IEEE Sig. Proc. Lett.*, **3**, 10–12.

[71] Lin, S.T. & McFadden, P.D. (1997). Gear vibration analysis by B-spline wavelet based linear wavelet transform. *Mech. Syst. Signal Proc.*, **11**, 603–609.

[72] Mallat, S.G. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Pattn Anal. Mach. Intell.*, **11**, 674–693.

[73] Mallat, S.G. (1999). *A Wavelet Tour of Signal Processing.* 2nd Edition, San Diego: Academic Press.

[74] Marron, J.S., Adak, S., Johnstone, I.M., Neumann, M.H. & Patil, P. (1998). Exact risk analysis of wavelet regression. *J. Comp. Graph. Statist.*, **7**, 278–309.

[75] Meyer, Y. (1992). *Wavelets and Operators.* Cambridge: Cambridge University Press.

[76] Misiti, M., Misiti, Y., Oppenheim, G. & Poggi, J.-M. (1994). Décomposition en ondelettes et méthodes comparatives : étude d'une courbe de charge éléctrique. *Revue de Statistique Appliquée*, **17**, 57–77.

[77] Müller, P. & Vidakovic, B. (Eds.) (1999). *Bayesian Inference in Wavelet Based Models.* Lect. Notes Statist., **141**, New York: Springer-Verlag.

[78] Nason, G.P. (1994). Wavelet regression by cross-validation. *Technical Report* **447**, Department of Statistics, Standord University, USA.

[79] Nason, G.P. (1995). Wavelet function estimation using cross-validation. In *Wavelets and Statistics*, Antoniadis, A. & Oppenheim, G. (Eds.), Lect. Notes Statist., **103**, pp. 261–280, New York: Springer-Verlag.

[80] Nason, G.P. (1996). Wavelet shrinkage using cross-validation. *J. R. Statist. Soc. B*, **58**, 463–479.

[81] Nason, G.P. (2002). Choice of wavelet smoothness, primary resolution and threshold in wavelet shrinkage. *Statist. Comput.*, **12**, 219–227.

[82] Nason, G.P. & Silverman, B.W. (1995). The stationary wavelet transform and some statistical applications. In *Wavelets and Statistics*, Antoniadis, A. & Oppenheim, G. (Eds.), Lect. Notes Statist., **103**, pp. 281–300, New York: Springer-Verlag.

[83] Neumann, M.H. & Spokoiny, V. (1995). On the efficiency of wavelet estimators under arbitrary error distributions. *Math. Meth. Stat.*, **4**, 137–166.

[84] Ogden, R.T. (1997). *Essential Wavelets for Statistical Applications and Data Analysis.* Boston: Birkhäuser.

[85] Ogden, R.T. & Parzen, E. (1996a). Change-point approach to data analytic wavelet thresholding. *Statist. Comput.*, **6**, 93–99.

[86] Ogden, R.T. & Parzen, E. (1996b). Data dependent wavelet thresholding in nonparametric regression with change-point applications. *Comput. Statist. Data Anal.*, **22**, 53–70.

[87] Percival, D.B. & Walden, A.T. (2000). *Wavelet Methods for Time Series Analysis.* Cambridge: Cambridge University Press.

[88] Ramsay, J.B., Usikov, D. & Zaslavsky, D. (1995). An analysis of U.S. stock price behavior using wavelets. *Fractals*, **3**, 377–389.

[89] Staszewski, W.J. & Tomlinson, G.R. (1994). Application of the wavelet transform to fault detection in a spur gear. *Mech. Syst. Signal Proc.*, **8**, 289–307.

[90] Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, **9**, 1135–1151.

[91] Strang, G. & Nguyen, T. (1996). *Wavelets and Filter Banks.* Wellesley: Wellesley-Cambridge Press.

[92] von Sachs, R. & MacGibbon, B. (2000). Non-parametric curve estimation by wavelet thresholding with locally stationary errors. *Scand. J. Statist.*, **27**, 475–499.

[93] Vannucci, M. & Corradi, F. (1999). Covariance structure of wavelet coefficients: theory and models in a Bayesian perspective. *J. R. Statist. Soc. B*, **61**, 971–986.

[94] Vidakovic, B. (1998a). Non-linear wavelet shrinkage with Bayes rules and Bayes factors. *J. Am. Statist. Ass.*, **93**, 173–179.

[95] Vidakovic, B. (1998b). Wavelet based nonparametric Bayes methods. In *Practical Nonparametric and Semiparametric Bayesian Statistics*, Dey, D.D., Müller, P. & Sinha, D. (Eds.), Lecture Notes in Statistics **133**, pp. 133-155, New York: Springer-Verlag.

[96] Vidakovic, B. (1999). *Statistical Modeling by Wavelets.* New York: John Wiley & Sons.

[97] Vidakovic, B. & Ruggeri, F. (2001). BAMS method: theory and simulations. *Sankhyā B*, **63**, 234–249.

[98] Wahba, G. (1990). *Spline Models for Observational Data.* Philadelphia: SIAM.

[99] Wang, Y. (1996). Function estimation via wavelet shrinkage for long-memory data. *Ann. Statist.*, **24**, 466–484.

[100] Wand, M.P. & Jones, M.C. (1995). *Kernel Smoothing.* London: Chapman & Hall.

[101] Weyrich, N. & Warhola, G.T. (1995a). De-noising using wavelets and cross-validation. *NATO Adv. Study Inst. C*, **454**, 523–532.

[102] Weyrich, N. & Warhola, G.T. (1995b). Wavelet shrinkage and generalized cross-validation for de-noising with applications to speech. In *Wavelets and Multilevel Approximation*, Chui, C.K. & Schumaker, L.L. (Eds.), Approximation Theory VIII **2**, pp. 407–414, Singapore: World Scientific.

[103] Wojtaszczyk, P. (1997). *A Mathematical Introduction to Wavelets.* Cambridge: Cambridge University Press.

Appendix I

Here are the analytical formulae of the test functions introduced in Section 5.

1. Step:
$$f_1(x) = 0.2 + 0.6\mathbf{I}_{(1/3,3/4)}(x).$$

2. Wave:
$$f_2(x) = 0.5 + 0.2\cos(4\pi x) + 0.1\cos(24\pi x).$$

3. Blip:
$$\begin{aligned} f_3(x) &= \left(0.32 + 0.6x + 0.3e^{-100(x-0.3)^2}\right)\mathbf{I}_{[0,0.8]}(x) + \\ &\quad \left(-0.28 + 0.6x + 0.3e^{-100(x-1.3)^2}\right)\mathbf{I}_{(0.8,1]}(x). \end{aligned}$$

4. Blocks: Donoho and Johnstone's (1994) Blocks function vertically rescaled to $[0.2, 0.8]$.

5. Bumps: Donoho and Johnstone's (1994) Bumps function vertically rescaled to $[0.2, 0.8]$.

6. Heavisine: Donoho and Johnstone's (1994) Heavisine function vertically rescaled to $[0.2, 0.8]$.

7. Doppler: Donoho and Johnstone's (1994) Doppler function vertically rescaled to $[0.2, 0.8]$.

8. Angles:
$$\begin{aligned} f_8(x) &= (2x + 0.5))\mathbf{I}_{[0,0.15]}(x) + (-12(x-0.15) + 0.8)\mathbf{I}_{(0.15,0.2]}(x) + \\ &\quad 0.2\mathbf{I}_{]0.2,0.5]}(x) + (6(x-0.5) + 0.2)\mathbf{I}_{(0.5,0.6]}(x) + \\ &\quad (-10(x-0.6) + 0.8)\mathbf{I}_{]0.6,0.65]}(x) + (-5(x-0.65) + 0.3)\mathbf{I}_{(0.65,0.85]}(x) + \\ &\quad (2(x-0.85) + 0.2)\mathbf{I}_{(0.85,1]}(x). \end{aligned}$$

9. Parabolas:
$$\begin{aligned} f_9(x) &= 0.8 - 30r(x,0.1) + 60r(x,0.2) - 30r(x,0.3) + \\ &\quad 500r(x,0.35) - 1000r(x,0.37) + 1000r(x,0.41) - 500r(x,0.43) + \\ &\quad 7.5r(x,0.5) - 15r(x,0.7) + 7.5r(x,0.9), \end{aligned}$$

   where $r(x,c) = (x-c)^2\mathbf{I}_{(c,1]}(x)$.

10. Time Shifted Sine:

$$f_{10}(x) = 0.3 \sin\{3\pi[g(g(g(g(x)))) + x]\} + 0.5,$$

where $g(x) = (1 - \cos(\pi x))/2$.

11. Spikes:

$$\begin{aligned} g(x) &= 15.6676 e^{-500(x-0.23)^2} + 2e^{-2000(x-0.33)^2} + \\ & \quad 4e^{-8000(x-0.47)^2} + 3e^{-16000(x-0.69)^2} + e^{-32000(x-0.83)^2}. \end{aligned}$$

$$f_{11}(x) = (0.6/\text{range}(g))g(x) + 0.2.$$

12. Corner:

$$\begin{aligned} g(x) &= 623.87x^3(1-4x)\mathbf{I}_{]0,0.5]}(x) + \\ & \quad 187.161(0.125 - x^3)x^4\mathbf{I}_{(0.5,0.8]}(x) + 3708.470441(x-1)^3\mathbf{I}_{(0.8,1]}(x), \end{aligned}$$

$$f_{12}(x) = (0.6/\text{range}(g))g(x) + 0.6.$$

Appendix II

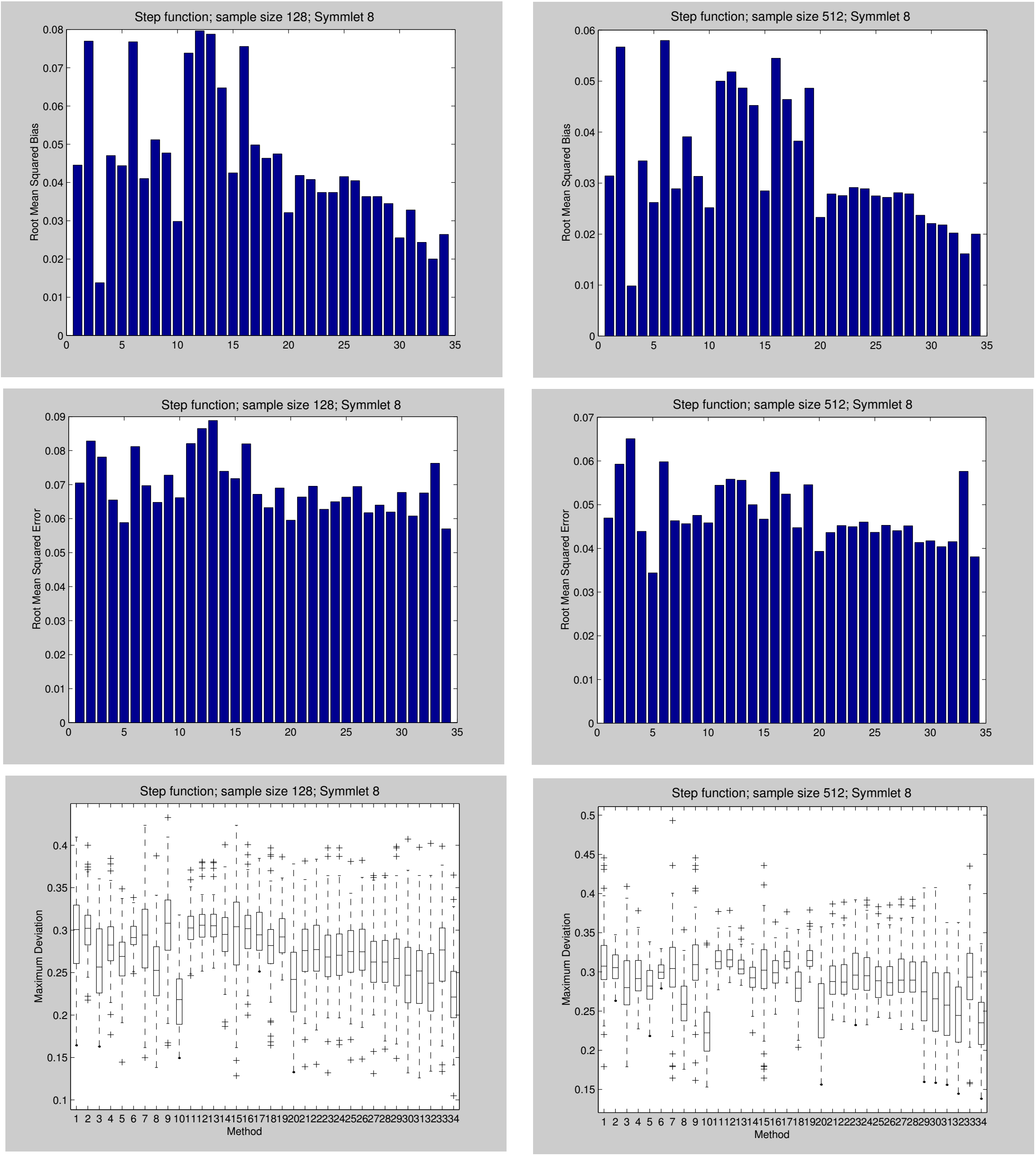


Figure 8.4: Performance of the estimators for the *Step* function over 100 simulations. The root signal-to-noise ratio is equal to 3 for sample sizes of 128 (left) and 512 (right) design points. The wavelet filter used is the Symmlet 8.

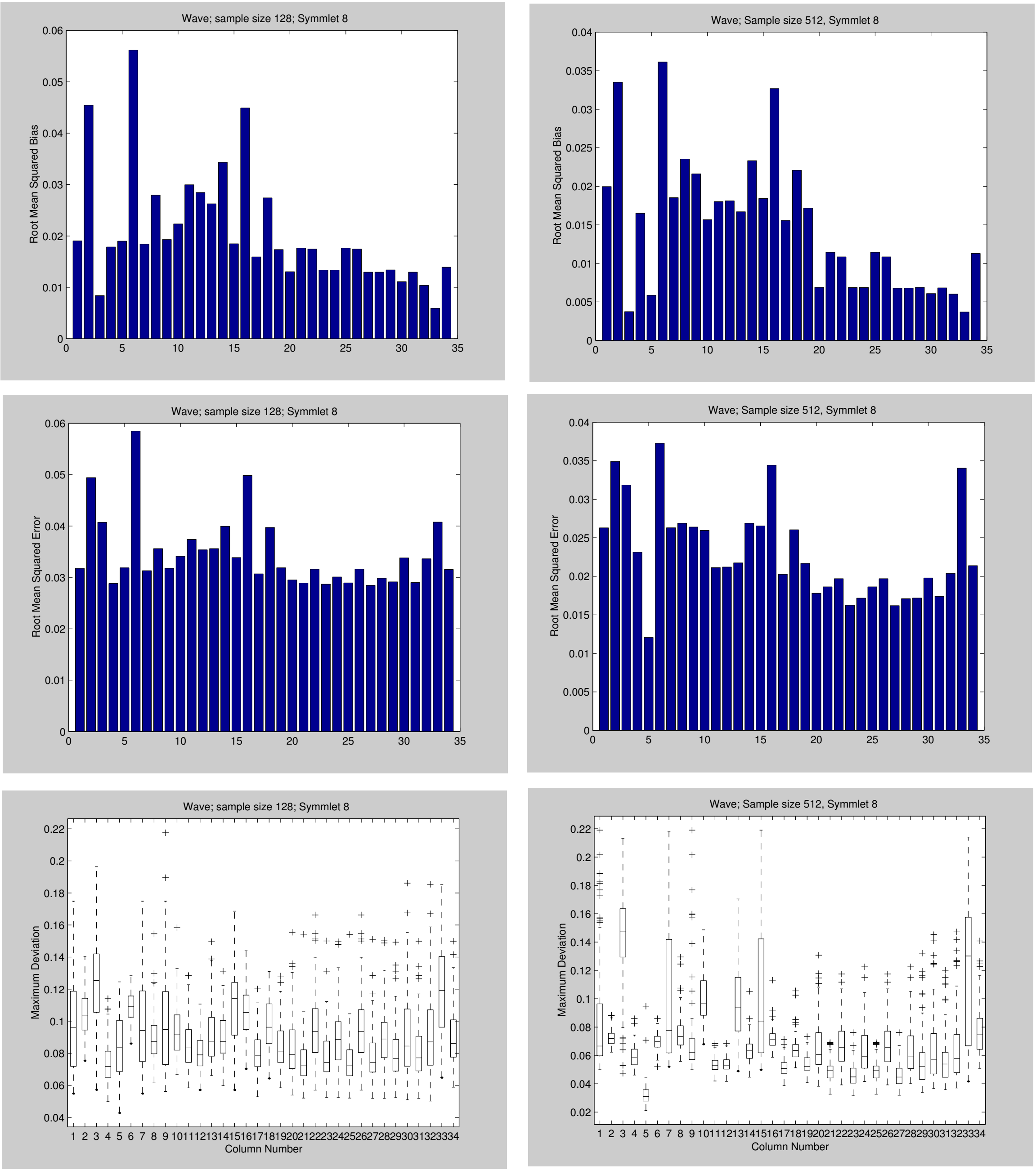


Figure 8.5: Performance of the estimators for the *Wave* function over 100 simulations. The root signal-to-noise ratio is equal to 3 for sample sizes of 128 (left) and 512 (right) design points. The wavelet filter used is the Symmlet 8.

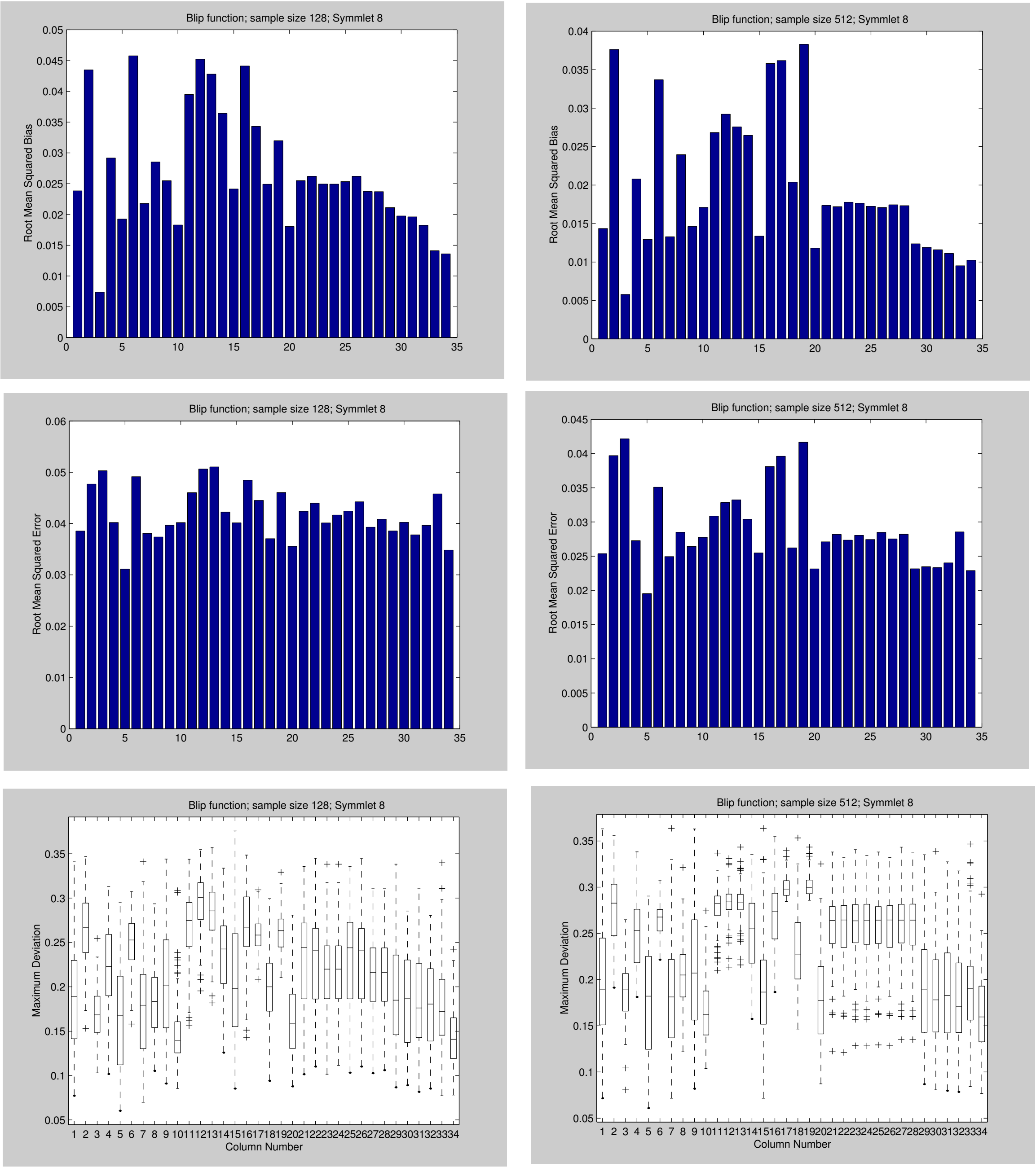


Figure 8.6: Performance of the estimators for the *Blip* function over 100 simulations. The root signal-to-noise ratio is equal to 3 for sample sizes of 128 (left) and 512 (right) design points. The wavelet filter used is the Symmlet 8.

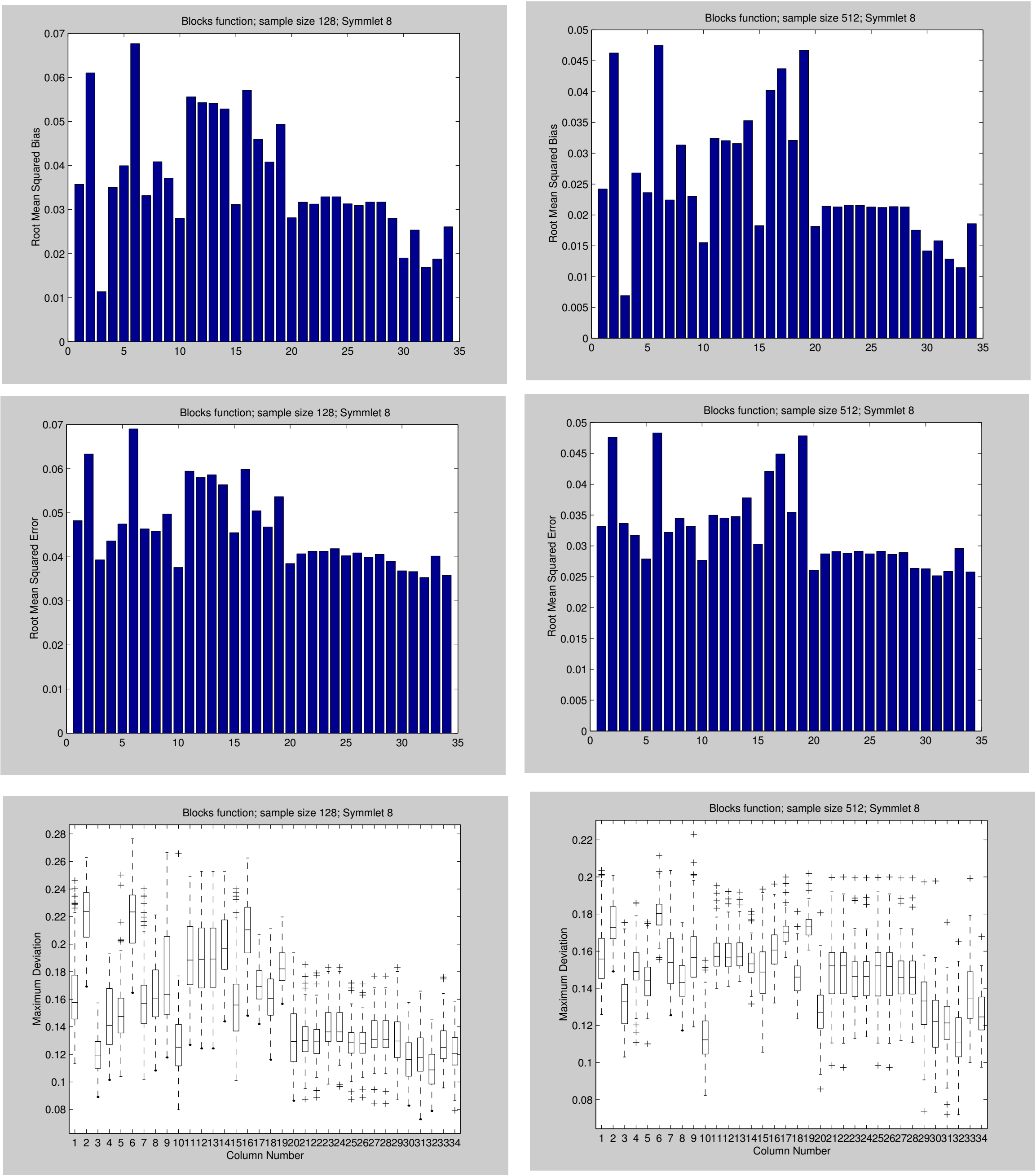


Figure 8.7: Performance of the estimators for the *Blocks* function over 100 simulations. The root signal-to-noise ratio is equal to 3 for sample sizes of 128 (left) and 512 (right) design points. The wavelet filter used is the Symmlet 8.

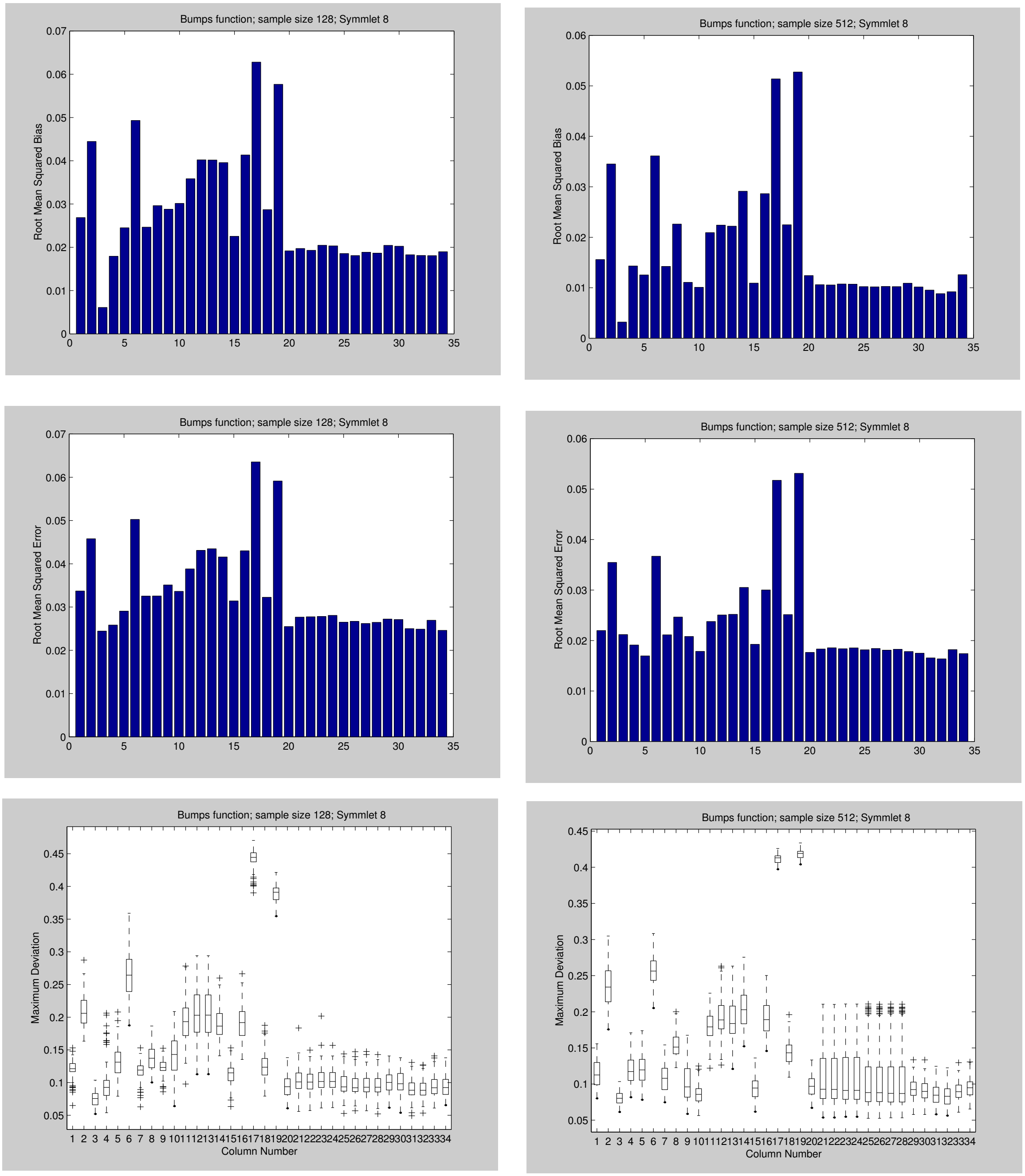


Figure 8.8: Performance of the estimators for the *Bumps* function over 100 simulations. The root signal-to-noise ratio is equal to 3 for sample sizes of 128 (left) and 512 (right) design points. The wavelet filter used is the Symmlet 8.

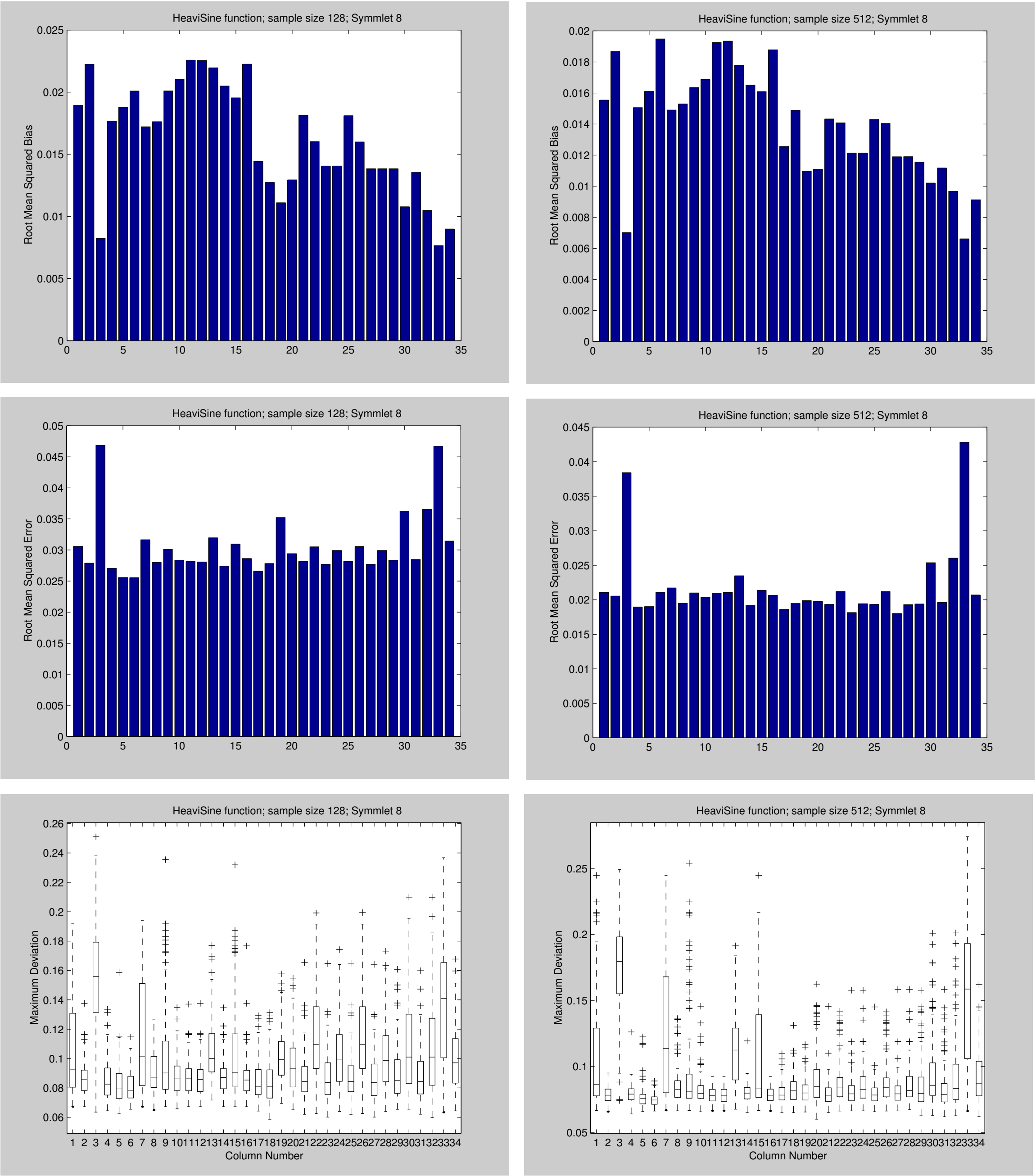


Figure 8.9: Performance of the estimators for the *HeaviSine* function over 100 simulations. The root signal-to-noise ratio is equal to 3 for sample sizes of 128 (left) and 512 (right) design points. The wavelet filter used is the Symmlet 8.

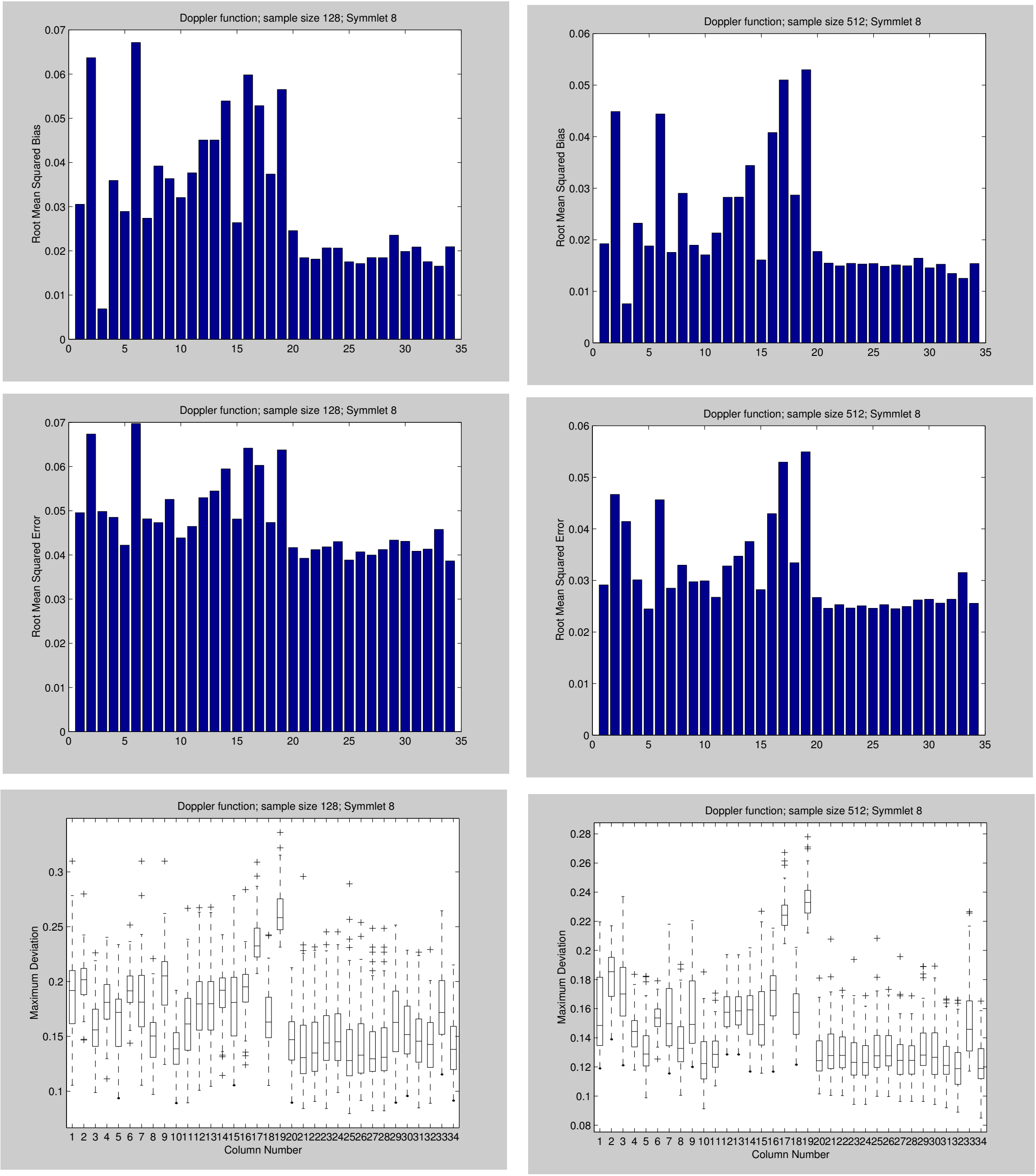


Figure 8.10: Performance of the estimators for the *Doppler* function over 100 simulations. The root signal-to-noise ratio is equal to 3 for sample sizes of 128 (left) and 512 (right) design points. The wavelet filter used is the Symmlet 8.

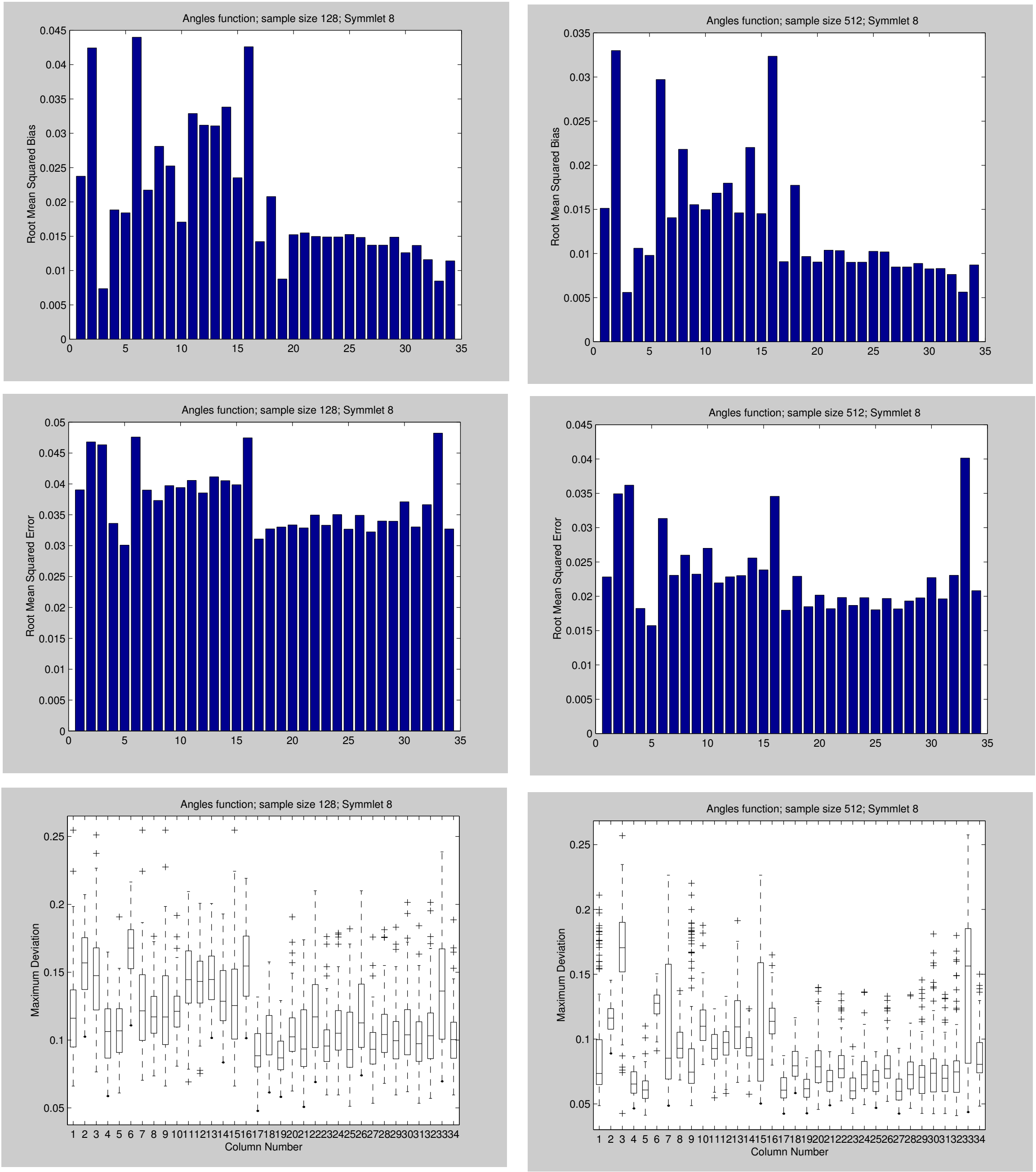


Figure 8.11: Performance of the estimators for the *Angles* function over 100 simulations. The root signal-to-noise ratio is equal to 3 for sample sizes of 128 (left) and 512 (right) design points. The wavelet filter used is the Symmlet 8.

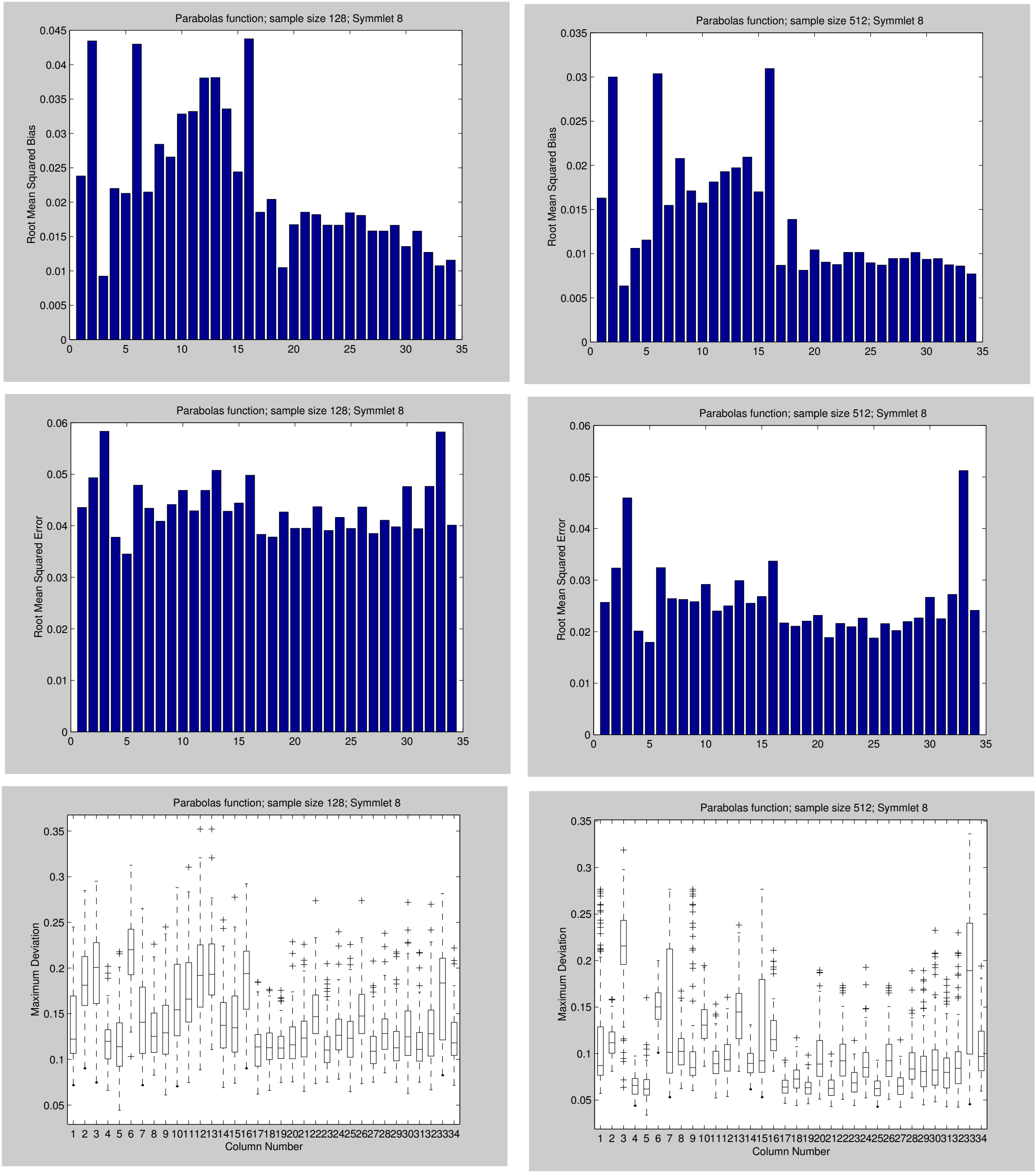


Figure 8.12: Performance of the estimators for the *Parabolas* function over 100 simulations. The root signal-to-noise ratio is equal to 3 for sample sizes of 128 (left) and 512 (right) design points. The wavelet filter used is the Symmlet 8.

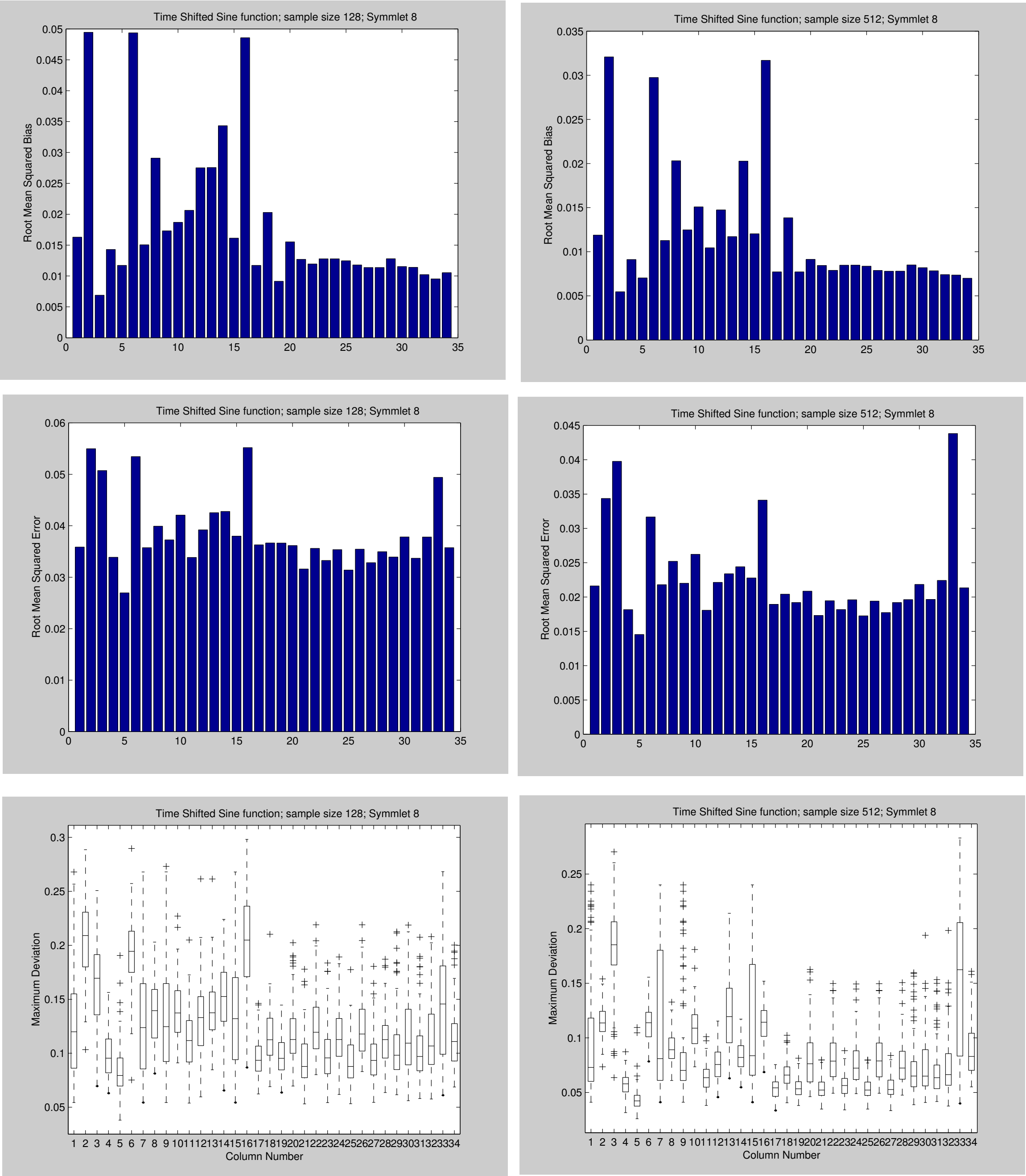


Figure 8.13: Performance of the estimators for the *Time Shifted Sine* function over 100 simulations. The root signal-to-noise ratio is equal to 3 for sample sizes of 128 (left) and 512 (right) design points. The wavelet filter used is the Symmlet 8.

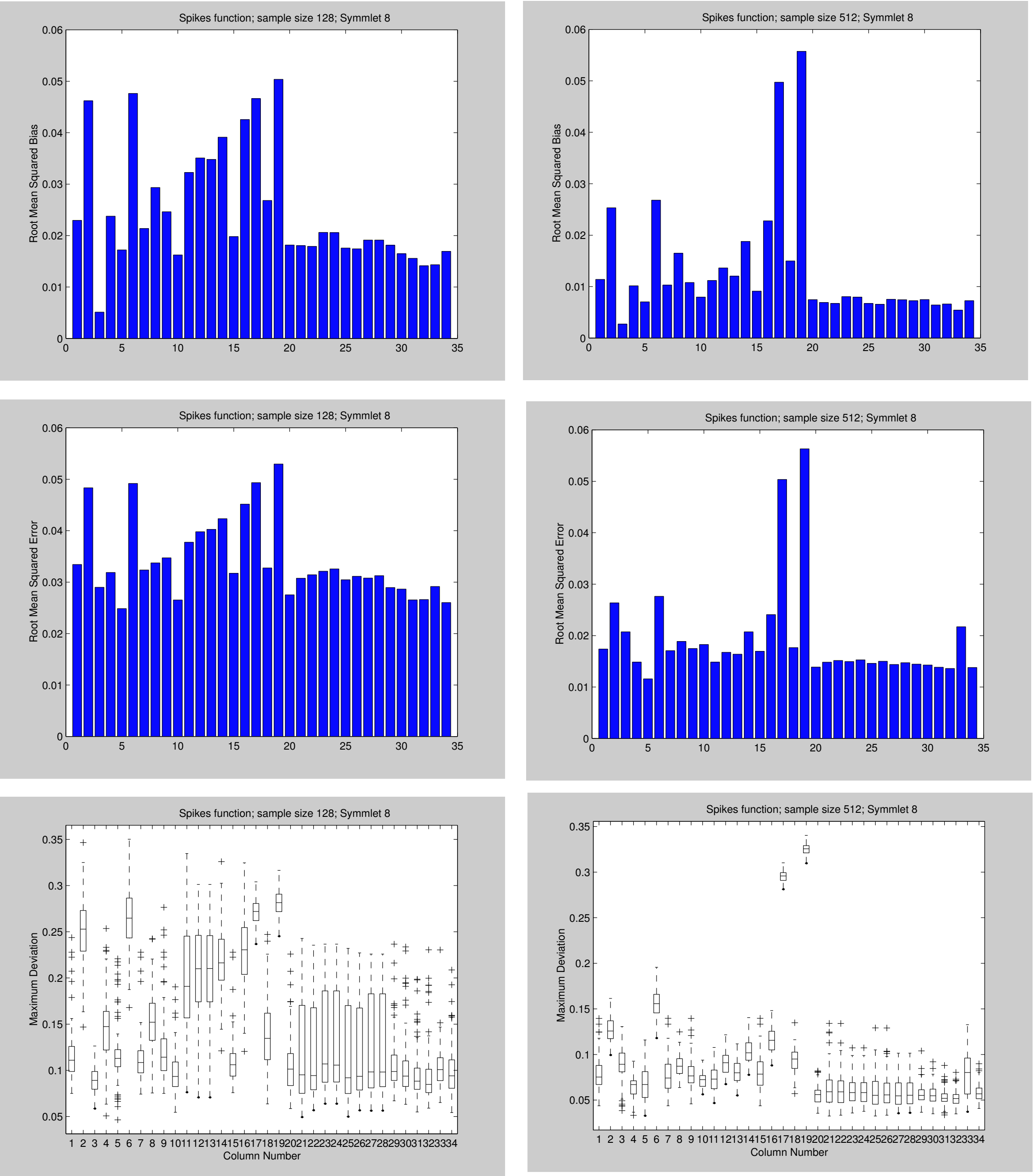


Figure 8.14: Performance of the estimators for the *Spikes* function over 100 simulations. The root signal-to-noise ratio is equal to 3 for sample sizes of 128 (left) and 512 (right) design points. The wavelet filter used is the Symmlet 8.

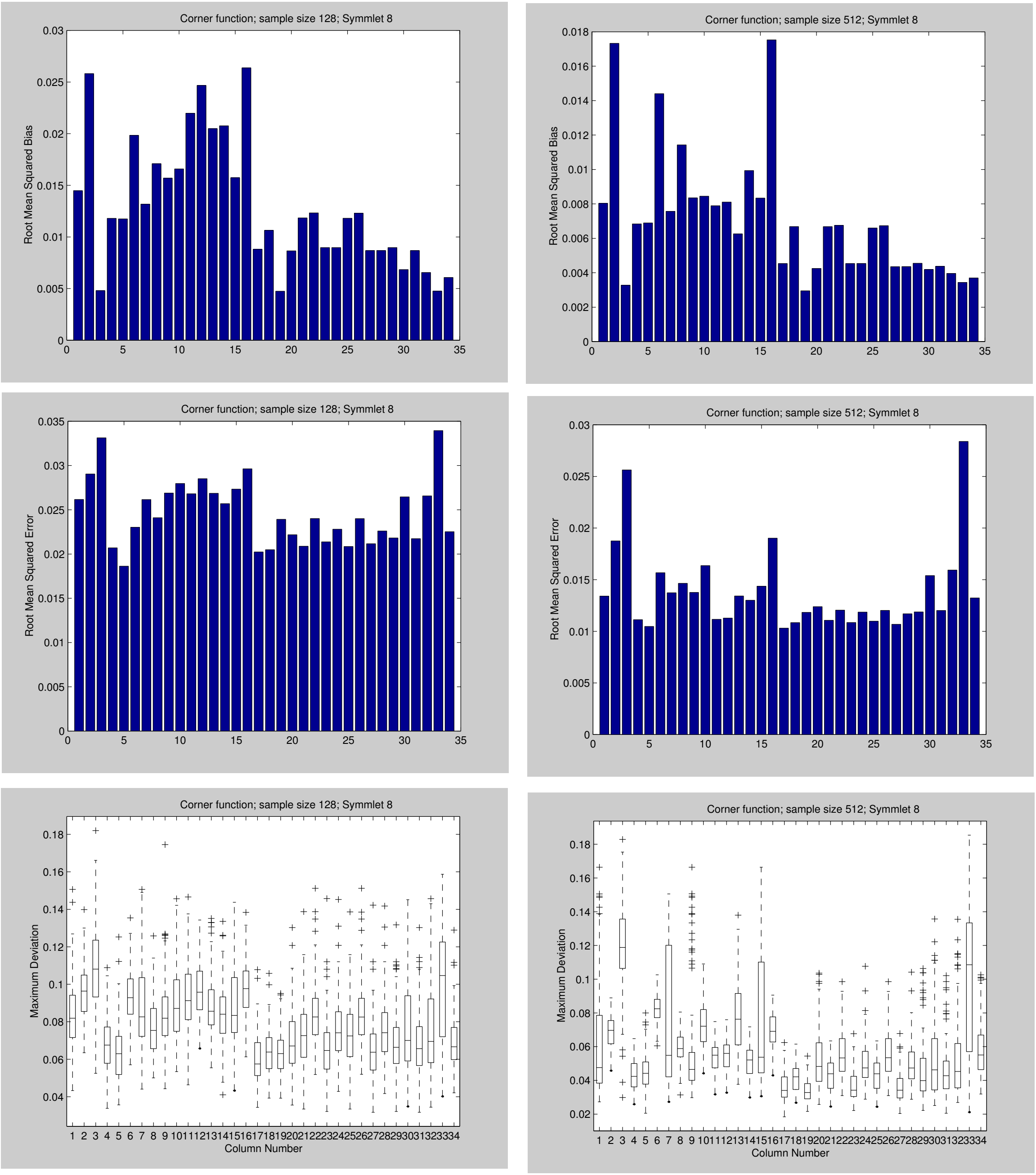


Figure 8.15: Performance of the estimators for the *Corner* function over 100 simulations. The root signal-to-noise ratio is equal to 3 for sample sizes of 128 (left) and 512 (right) design points. The wavelet filter used is the Symmlet 8.

Method	Average CPU Time (n=128)	St. Dev.	Average CPU Time (n=512)	St. Dev.
VISU-H	0.0047	0.0052	0.0057	0.005
VISU-S	0.0044	0.0052	0.0075	0.0044
SURE	0.0131	0.0053	0.0219	0.0044
HYBSURE	0.0108	0.0044	0.018	0.004
TI-H	0.0948	0.008	0.3842	0.0067
TI-S	0.0951	0.0085	0.3842	0.0075
MINIMAX-H	0.0071	0.0048	0.0089	0.0031
MINIMAX-S	0.007	0.0048	0.0094	0.0024
CV-H	0.788	0.0423	2.8711	0.1194
CV-S	0.8123	0.0803	2.9715	0.1487
NEIGHBL	0.0258	0.0065	0.0621	0.005
BLOCKJS-A	0.0131	0.0056	0.0295	0.0026
BLOCKJS-T	0.0124	0.0051	0.0282	0.0039
THRDA1	0.11	0.0236	0.1602	0.0155
FDR-H	0.0314	0.0053	0.0345	0.0052
FDR-S	0.0311	0.0053	0.0341	0.0057
PENWAV	0.0088	0.0041	0.0108	0.0031
SCAD	0.0059	0.0049	0.009	0.003
MIXED	0.0535	0.0083	0.0643	0.0056
DECOMPSH	0.1146	0.0073	0.2216	0.0072
BLMED-A	4.9719	0.7042	7.5799	0.7529
BLMED-T	4.8445	1.0492	8.4865	0.8633
HYBMED-A	8.1662	18.5313	10.4675	24.1012
HYBMED-T	8.1802	18.4423	10.4619	24.2981
BLMEAN-A	4.8831	0.7118	7.3968	0.7442
BLMEAN-T	4.776	1.0416	8.314	0.8564
HYBMEAN-A	8.1064	18.4009	10.3045	24.2128
HYBMEAN-T	8.0896	18.3561	10.3007	24.2864
SINGLMED	15.0696	36.0658	35.8483	60.1153
SINGLMED2	1.27	0.8122	2.0824	1.2284
SINGLMEAN	15.0723	36.1797	35.7626	60.0166
SINGLMEAN2	1.2227	0.8109	2.0086	1.2153
SINGLHYP	15.0615	36.1177	35.7836	60.0317
BAMS	0.0876	0.0077	0.3524	0.0062

Table 4: Average and its standard deviation (over 100 simulations) of the CPU time involved in computing the estimates for the *Corner* function by each method and the two sample sizes ($n = 128$, $n = 512$). Non-Bayesian methods uniformly outperform Bayesian methods in terms of CPU time in all examples.