Aust. N. Z. J. Stat. **44**(1), 2002, 23–39

TREE-BASED WAVELET REGRESSION FOR CORRELATED DATA USING THE MINIMUM DESCRIPTION LENGTH PRINCIPLE

THOMAS C.M. LEE[1]

Colorado State University

Summary

This paper considers the problem of non-parametric regression using wavelet techniques. Its main contribution is the proposal of a new wavelet estimation procedure for recovering functions corrupted by correlated noise, although a similar procedure for independent noise is also presented. Two special features of the proposed procedure are that it imposes a so-called 'tree constraint' on the wavelet coefficients and that it uses the minimum description length principle to define its 'best' estimate. The proposed procedure is empirically compared with some existing wavelet estimation procedures, for the cases of independent and correlated noise.

Key words: correlated noise; minimum description length principle; non-Gaussian noise; tree constraint; wavelet regression.

1. Introduction

In recent years wavelet techniques for non-parametric regression have attracted enormous attention from researchers across different fields. Two main reasons for this are that wavelet estimators enjoy excellent minimax properties and that they are capable of adapting to spatial and frequency inhomogeneities (see e.g. Donoho & Johnstone, 1994, 1995; Donoho *et al.*, 1995). Also, they are backed up by a fast algorithm (see e.g. Mallat, 1989).

Most existing wavelet-based non-parametric regression methods were designed for recovering data corrupted by independent Gaussian noise. The primary goal of this paper is to propose a wavelet function estimation procedure that is capable of handling correlated data. Two characteristics of this procedure are that it imposes a so-called 'tree constraint' on the class of all plausible function estimates and that the final function estimate is chosen by the minimum description length (MDL) principle (see Rissanen, 1989 Chapter 3, and references given therein). Of course, the proposed procedure also handles independent noise; furthermore, it can be made specialized to handle non-Gaussian noise (see Section 6.1).

This paper is arranged as follows. Section 2 provides background material for wavelet regression. Section 3 describes how the MDL principle can be applied to the problem of

Received April 1999; revised September 2000; accepted October 2000.

* Author to whom correspondence should be addressed.

[1] Dept of Statistics, Colorado State University, Fort Collins, CO 80523-1877, USA.
e-mail: tlee@stat.colostate.edu

Acknowledgments. The author thanks Victor Solo for many discussions, Michael Stein for many constructive comments and an insightful discussion which led to the idea of tree representation, Kenneth Wilder for discussions on classification trees, and Fanis Sapatinas and Bernard Silverman for providing their BUTHRESH routines. The author also thanks the referees and the associate editor for many useful comments and questions which have led to this substantially improved version of the paper. Numerical work was done using the `wavethresh` package of Nason & Silverman (1994). S-PLUS® codes for the SURE procedure were adopted from Luo & Wahba (1997).

wavelet regression. Section 4 presents an existing MDL criterion (due to Saito). In Section 5, we show that, to some extent, Saito's criterion can be simplified by imposing a tree constraint on the wavelet coefficients. Section 6 presents the proposed procedure. Section 7 reports simulation results and Section 8 draws conclusions.

While this paper was being revised, the author became aware of the closely related work by Moulin (1996), in which an MDL tree-based wavelet regression procedure is proposed. However, Moulin (1996) does not handle correlated noise or provide any simulation results regarding the practical performance of the procedure.

2. Background: wavelet regression

Suppose we observe n equidistant noisy observations:

$$y_i = f_i + e_i, \qquad f_i = f\Big(\frac{i}{n-1}\Big) \qquad (i = 0, \ldots, n-1)\,,$$

where f is an unknown function, and the e_i are noise. Our goal is to recover f using wavelet methods. For simplicity, we assume that $n = 2^{J+1}$ is an integer power of 2, and consider both independent and correlated noise.

Broadly speaking, wavelet methods for non-parametric regression involve two steps. The first step is to obtain the empirical wavelet coefficient vector $\boldsymbol{w}$ by applying a discrete wavelet transform (DWT) to the observations $\boldsymbol{y} = (y_0, \ldots, y_{n-1})$. If we denote the DWT matrix by $\boldsymbol{W}$ (see e.g. Donoho & Johnstone, 1994), $\boldsymbol{w}$ is given by $\boldsymbol{w} = \boldsymbol{W}\boldsymbol{y}$. Here $\boldsymbol{w}$ is an $n \times 1$ vector and we follow Donoho & Johnstone (1994) and use a double indexing scheme to label its elements:

$$\boldsymbol{w} = (\underbrace{w_{-1,0}}, \overbrace{w_{0,0}}, \underbrace{w_{1,0}, w_{1,1}}, \overbrace{w_{2,0}, w_{2,1}, w_{2,2}, w_{2,3}}, \ldots, w_{j,k}, \ldots, \underbrace{w_{J,0}, \ldots, w_{J,2^J-1}})\,.$$

Therefore, with the exception of the first element, the indexing scheme is: $w_{j,k}$, $j = 0, \ldots, J$; $k = 0, \ldots, 2^J - 1$.

The second step is to apply a filtering operation (e.g. thresholding) to $\boldsymbol{w}$ and obtain an *estimated* wavelet coefficient vector $\hat{\boldsymbol{w}}$. Then an estimate $\hat{\boldsymbol{f}}$ of $\boldsymbol{f} = (f_0, \ldots, f_{n-1})$ can be obtained by applying the inverse DWT to $\hat{\boldsymbol{w}}$: $\hat{\boldsymbol{f}} = \boldsymbol{W}^{\mathrm{T}}\hat{\boldsymbol{w}}$. Below we describe three types of filtering operation: thresholding, recursive partitioning and methods based on hidden Markov models.

2.1. Thresholding

Thesholding is perhaps the most widely studied filtering operation; it is widely known that the quality of $\hat{\boldsymbol{f}}$ is highly dependent on the threshold values. For the case of independent noise, various automatic methods have been proposed for choosing their values. These methods include the 'universal' thresholding scheme of Donoho & Johnstone (1994), the SURE thresholding scheme of Donoho & Johnstone (1995), the cross-validation scheme of Nason (1996) and the cross-validatory AIC scheme of Hurvich & Tsai (1998). Bayesian methods have also been proposed: see Chipman, Kolaczyk & McCulloch (1997), Abramovich, Sapatinas & Silverman (1998) and Vidakovic (1998). Of particular interest to the present paper is the MDL-based scheme proposed by Saito (1994). This MDL-based scheme was further studied by Antoniadis, Gijbels & Gregoire (1997) and it is discussed in detail in Section 4.

Correlated noise has also been considered (although to a much smaller extent). Wang (1996) provides some asymptotic minimax results, and modifies the independent noise 'universal' and cross-validation thresholding schemes so that they can be applied to correlated data. Johnstone & Silverman (1997) demonstrate that, even though the SURE thresholding scheme of Donoho & Johnstone (1995) was originally designed for independent noise, it is also applicable for correlated data. They also establish many theoretical minimax results. Another proposal for dealing with correlated noise is described in Solo (1998). His approach is to define the estimate $\hat{f}$ as the minimizer of an L_1-penalized weighted least squares criterion. Because the minimization of such a criterion is not trivial, a minimization algorithm, modified from the CLEAN algorithm of radiophysics, was developed.

2.2. Recursive partitioning

Engel (1994) provides another example of a filtering operation for handling independent data. His procedure is specialized to the Haar wavelet system, and is closely related to the CART method developed by Breiman *et al.* (1984). The idea behind it is to approximate the true function by a piecewise constant function and use a recursive partitioning scheme to find a 'best-fit' piecewise constant function. Note that in similar situations, a recursive partitioning scheme can usually be treated as some sort of tree-growing or pruning algorithm.

Donoho (1997) discusses a connection between the CART method for non-parametric regression and the BOB (best-ortho-basis) method for time-frequency analysis. He also addresses the denoising of noisy data using the BOB method. However, he does not discuss the issue of correlated data.

2.3. Hidden Markov models

Crouse, Nowak & Baraniuk (1998) introduced a novel wavelet regression procedure that uses hidden Markov models. Their procedure was designed for independent noise, and can be described as follows. Each wavelet coefficient is associated with a hidden binary state variable S; the value of S is unobservable. If $S = 0$, say, the corresponding wavelet coefficient is 'suspected' to be a 'noise coefficient', and is treated as a realization of a zero-mean Gaussian density with a small variance σ_S^2. On the other hand, if $S = 1$, say, the corresponding wavelet coefficient is 'suspected' to be a 'signal coefficient', and is treated as a realization of a zero-mean Gaussian density with a high variance σ_H^2. Both σ_S^2 and σ_H^2 are unknown.

Then the dependencies of the state variables are modelled by a probabilistic graph with a tree structure. Note that this is different from imposing a tree dependency structure directly on the wavelet coefficients (which is the approach taken by the present paper). The next step is to apply an EM algorithm to obtain the maximum likelihood estimates of the hidden state probabilities, σ_S^2 and σ_H^2. Once these estimates are computed, final wavelet coefficient estimates are obtained by an empirical Bayesian procedure.

3. MDL for wavelet regression

We motivate our discussion of the MDL principle by the following problem. Suppose a set of observed data z and a class of plausible models $\Theta = \{\theta_1, \ldots, \theta_m\}$ for z are given, and our goal is to select a 'best' model for z from Θ. It is allowed that different θ_i may have different numbers of parameters. One typical example is subset selection in the multiple regression context.

The MDL principle provides a powerful method for attacking such model selection problems. In short, the MDL principle defines the 'best' model as the one that enables the best encoding (or compression) of the data z, so that z can be transmitted in the most economical way. That is, the best fitted model is the one that produces the shortest codelength of z. In the present context, the codelength of z can be treated as the amount of memory space that is required to store z.

One general method for encoding z is to 'split' z into two components: a fitted model $\hat{\boldsymbol{\theta}}$ plus the corresponding residuals $\hat{\boldsymbol{r}}$. Rissanen (1989 pp. 54–58) called this 'two-part coding'. We follow Rissanen's notation and use $L(a)$ to denote the codelength for the object a. We have

$$L(z) = L(\hat{\boldsymbol{\theta}}) + L(\hat{\boldsymbol{r}} \mid \hat{\boldsymbol{\theta}}) .$$

The MDL principle defines the best $\hat{\boldsymbol{\theta}}$ as the one that gives the smallest $L(z)$. In the above expression we have stressed that $\hat{\boldsymbol{r}}$ is 'conditional' on $\hat{\boldsymbol{\theta}}$.

For the wavelet regression problem that we consider here, z corresponds to $\boldsymbol{y}$, $\hat{\boldsymbol{\theta}}$ corresponds to $\hat{\boldsymbol{f}}$ or $\hat{\boldsymbol{w}}$ (note that there is a one-to-one correspondence between $\hat{\boldsymbol{f}}$ and $\hat{\boldsymbol{w}}$), and $\hat{\boldsymbol{r}}$ corresponds to $\hat{\boldsymbol{e}} = \boldsymbol{y} - \hat{\boldsymbol{f}}$. In other words, the MDL principle suggests that $\hat{\boldsymbol{w}}$ should be chosen as the one that minimizes

$$L(\boldsymbol{y}) = L(\hat{\boldsymbol{w}}) + L(\hat{\boldsymbol{e}} \mid \hat{\boldsymbol{w}}) . \tag{1}$$

Thus to apply the MDL principle to the problem of wavelet regression, we must first derive an expression for $L(\boldsymbol{y})$, and then develop a procedure for minimizing the derived expression.

4. An existing MDL criterion for independent noise

When the noise is independent, Saito (1994) developed an MDL-based procedure for obtaining $\hat{\boldsymbol{w}}$ (and hence $\hat{\boldsymbol{f}}$). His procedure keeps the first $\hat{m}$ largest (in terms of absolute value) wavelet coefficients and deletes all the remaining ones, where $\hat{m}$ is chosen as the minimizer of an MDL criterion that he derived (i.e. an expression for $L(\boldsymbol{y})$). Thus his procedure is equivalent to a global thresholding scheme that has its threshold value chosen as any number between the $\hat{m}$th and $(\hat{m}+1)$th largest absolute values of the empirical wavelet coefficients. Antoniadis *et al.* (1997) provide supportive theoretical and simulation results for Saito's criterion.

To encode $\hat{\boldsymbol{w}}$ so that it can be transmitted, we need to encode (i) the indices, and (ii) the actual estimated values of those non-zero coefficients $\hat{w}_{jk}$ in $\hat{\boldsymbol{w}}$. Since the index of a non-zero coefficient $\hat{w}_{jk}$ is an integer between 1 and n, Saito asserts that the codelength for encoding such an index is $\log_2 n$. For the codelength of encoding the actual values of those estimated non-zero coefficients in $\hat{\boldsymbol{w}}$, we can apply the result of Rissanen (1989 pp. 55–56), which says the codelength for encoding one of these estimated values is $\frac{1}{2}\log_2 n$. One heuristic argument for this result is as follows. Each $\hat{w}_{jk}$ is estimated from n noisy observations, so there is no need to encode $\hat{w}_{jk}$ to a precision that is finer than its standard error. Now as the standard error is asymptotically of order $\sqrt{n}$, it suggests that $\hat{w}_{jk}$ can be effectively encoded with $\frac{1}{2}\log_2 n$ bits. Thus, if there are $\hat{m}$ non-zero coefficients in $\hat{\boldsymbol{w}}$, the total codelength for encoding $\hat{\boldsymbol{w}}$ is

$$\begin{aligned} L(\hat{\boldsymbol{w}}) &= L(\text{'indices'}) + L(\text{'actual estimated values'}) \\ &= \hat{m}\log_2 n + \tfrac{1}{2}\hat{m}\log_2 n = \tfrac{3}{2}\hat{m}\log_2 n . \end{aligned} \tag{2}$$

Now if an estimated $\hat{\boldsymbol{w}}$ is 'reasonable', the residual $\hat{\boldsymbol{e}}$ approximately satisfies the model assumption (independent Gaussian for the current situation). Rissanen (1989 pp. 54–55) shows that, by using this fact, the codelength $L(\hat{\boldsymbol{e}} \mid \hat{\boldsymbol{w}})$ for encoding $\hat{\boldsymbol{e}}$ is given by the negative of the conditional log-likelihood (base 2) of $\hat{\boldsymbol{e}}$ given $\hat{\boldsymbol{w}}$. When the noise is independent Gaussian and if we estimate σ^2 by $\hat{\sigma}^2 = \|\hat{\boldsymbol{e}}\|^2/n$, this negative conditional log-likelihood becomes ($\|\boldsymbol{x}\|^2$ is defined as $\boldsymbol{x}^{\mathsf{T}}\boldsymbol{x}$)

$$-\log_2 \left\{ \Big(\frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \Big)^n \exp\Big(-\frac{\|\hat{\boldsymbol{e}}\|^2}{\hat{\sigma}^2} \Big) \right\} = \tfrac{1}{2} n \log_2 2\pi + \tfrac{1}{2} n \log_2 \frac{\|\hat{\boldsymbol{e}}\|^2}{n} + n \,.$$

Thus, for minimization purposes, we can take

$$L(\hat{\boldsymbol{e}} \mid \hat{\boldsymbol{w}}) = \tfrac{1}{2} n \log_2 \frac{\|\hat{\boldsymbol{e}}\|^2}{n} \,. \tag{3}$$

Therefore, combining (1), (2) and (3), we have Saito's criterion

$$L_{\mathrm{S}}(\boldsymbol{y}) = L(\hat{\boldsymbol{w}}) + L(\hat{\boldsymbol{e}} \mid \hat{\boldsymbol{w}}) = \tfrac{3}{2} \hat{m} \log_2 n + \tfrac{1}{2} n \log_2 \frac{\|\hat{\boldsymbol{e}}\|^2}{n} \,. \tag{4}$$

However, in cases where the non-zero wavelet coefficients 'cluster' together, we claim that there is redundancy in Saito's criterion. By 'cluster' we mean that if $\hat{w}_{jk}$ is non-zero then $\hat{w}_{j-1,k}$, $\hat{w}_{j,k-1}$, $\hat{w}_{j,k+1}$ and $\hat{w}_{j+1,k}$ are also non-zero. Thus, when encoding such a $\hat{\boldsymbol{w}}$, we can make use of this fact and obtain a smaller value for $L(\hat{\boldsymbol{w}})$ (i.e. with a shorter codelength expression for $\hat{\boldsymbol{w}}$). This means that, in cases where 'clustering' is present (as defined previously), Saito's criterion over-penalizes the number of non-zero coefficients and hence has a tendency to underestimate the 'true' m.

5. Tree constraint

How can a $\hat{\boldsymbol{w}}$ be encoded so that the above 'clustering' property is captured? One possible approach is to use a tree to represent the indices of the non-zero elements of $\hat{\boldsymbol{w}}$. As we shall see later, another advantage of using a tree representation is that it agrees with the intuition that wavelet coefficients at a finer level (or resolution) should have a greater chance of being deleted than those coarser level wavelet coefficients at the same relative location. The first element $w_{-1,0}$ of a $\boldsymbol{w}$ carries the information of the mean of $\boldsymbol{y}$, so it should always be kept (unless $\boldsymbol{y}$ is a zero-mean white noise). For this reason, we ignore the index of $w_{-1,0}$ and focus on the indices of the remaining non-zero elements of a $\hat{\boldsymbol{w}}$.

Define $\hat{w}_{j+1,2k}$ and $\hat{w}_{j+1,2k+1}$ as the 'children' of $\hat{w}_{jk}$, $0 \le j < J$. We also call $\hat{w}_{jk}$ the 'parent' of $\hat{w}_{j+1,2k}$ and $\hat{w}_{j+1,2k+1}$. We impose the constraint that, if $\hat{w}_{jk}$ is deleted, none of its children can survive. That is, for $0 \le j < J$,

$$\hat{w}_{jk} = 0 \quad \Rightarrow \quad \hat{w}_{j+1,2k} = 0 \quad \text{and} \quad \hat{w}_{j+1,2k+1} = 0 \,.$$

This sort of constraint, sometimes known as tree constraint, is just a severe executioner of the intuition mentioned above that wavelet coefficients at finer levels should have higher chances of being deleted.

Once tree constraint is imposed, the idea of tree representation is straightforward. We illustrate it with an example. Suppose the only non-zero elements of $\hat{\boldsymbol{w}}$ are $\hat{w}_{00}$, $\hat{w}_{10}$, $\hat{w}_{11}$,

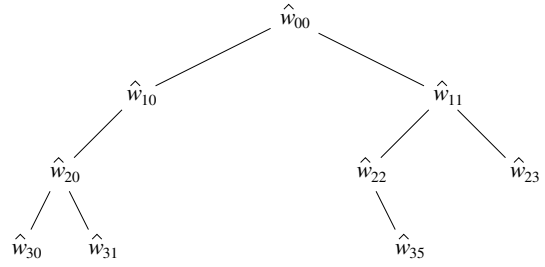

Figure 1. An example illustrating the tree representation of the indices of non-zero coefficients in $\hat{\boldsymbol{w}}$

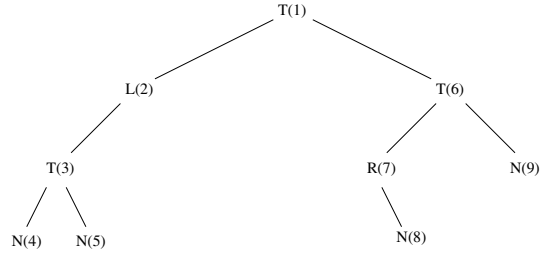

Figure 2. Node attributes and traversal order (in parenthesis) of the tree displayed in Figure 1

$\hat{w}_{20}$, $\hat{w}_{22}$, $\hat{w}_{23}$, $\hat{w}_{30}$, $\hat{w}_{31}$ and $\hat{w}_{35}$. The indices, or the relative positions, of these elements can be represented by the tree shown in Figure 1. Note that $\hat{w}_{10}$ and $\hat{w}_{22}$ have only one child.

By imposing tree-constraint on wavelet coefficients we can capture the 'clustering' property as defined previously. However, it is known that large coefficients do not always 'cluster' together, especially from one resolution level to another (i.e. from j to $j+1$). Indeed, it is true only along a sub-sequence of the resolution levels (see Daubechies, 1992 p. 300). Therefore, by imposing tree constraint we might delete large wavelet coefficients at high resolution level and induce numerical errors even without noise. Of course, there are cases where those numerical errors are smaller than the error due to noise and in such cases the proposed method is acceptable. Section 7.1 provides some examples.

The next task is to construct a method for encoding the tree structure. Various methods for encoding and optimizing classification trees are proposed in the literature; see e.g. Quinlan & Rivest (1989), Wallace & Patrick (1993) and Rissanen (1997). However, these methods are not suitable for our purposes because they all assume that all internal nodes of a classification tree have two children. Note also that the tree structure implicitly defined by the wavelet recursive partitioning scheme of Engel (1994) also falls into this category.

Here we suggest a tree-encoding method which allows for the possibility that an internal node possesses only one child. First notice that a node can only have one of the following four attributes: possesses no children (N), possesses a left child (L), possesses a right child (R) and possesses two children (T). Thus, if we agree to traverse a tree in a recursive, top-down, depth-first and left-first manner, we can encode the tree structure by just encoding the attributes of its nodes in the order that the nodes are visited.

To illustrate the idea, the tree in Figure 1 is re-drawn in Figure 2 with node attributes and traversal order displayed. Therefore, by using the above traversal scheme, the tree structure is represented by 'TLTNNTRNN'.

6. A tree-based MDL criterion

We present a new MDL criterion using a tree constraint on wavelet coefficients. First, we assume that the noise is independent and then we explain how to handle the case of correlated noise. We also briefly describe a tree-growing strategy for (approximately) minimizing this new criterion.

6.1. Independent noise

Recall that an MDL criterion, or codelength expression, for wavelet regression is of the form $L(\boldsymbol{y}) = L(\hat{\boldsymbol{w}}) + L(\hat{\boldsymbol{e}} \mid \hat{\boldsymbol{w}})$, and $L(\hat{\boldsymbol{w}})$ can be further decomposed into $L(\hat{\boldsymbol{w}}) = L(\text{'indices'}) + L(\text{'actual estimated values'})$; see (2). What is $L(\text{'indices'})$ when the indices are represented by a tree?

There are only four possible attributes for each node of a tree. Thus $\log_2 4 = 2$ bits are needed to encode the attribute of one node. If there are $\hat{m}$ non-zero wavelet coefficients (i.e. $\hat{m}$ nodes), $L(\text{'indices'}) = 2\hat{m}$ and hence $L(\hat{\boldsymbol{w}}) = 2\hat{m} + \frac{1}{2}\hat{m}\log_2 n$; see (2). Using the slightly vague argument that because n is usually large and $\hat{m}$ is small and thus the term $2\hat{m}$ is negligible, we approximate $L(\hat{\boldsymbol{w}})$ by $L(\hat{\boldsymbol{w}}) \approx \frac{1}{2}\hat{m}\log_2 n$.

If the noise is independent Gaussian, combining (1), $L(\hat{\boldsymbol{w}}) \approx \frac{1}{2}\hat{m}\log_2 n$ and (3), $L(\boldsymbol{y})$ can be approximated by

$$\text{MDL}_{\text{IND}}(\hat{\boldsymbol{f}}) = \tfrac{1}{2}\hat{m}\log_2 n + \tfrac{1}{2}n\log_2 \frac{\|\hat{\boldsymbol{e}}\|^2}{n}. \tag{5}$$

When the noise is known to be independent Gaussian, we propose that $\boldsymbol{f}$ can be estimated by the minimizer of $\text{MDL}_{\text{IND}}(\hat{\boldsymbol{f}})$, subject to the condition that the tree constraint is satisfied.

The criterion $\text{MDL}_{\text{IND}}(\hat{\boldsymbol{f}})$ can be modified for handling non-Gaussian noise if the density function of the noise, say g, is known up to a dispersion parameter ϕ. The second term of $\text{MDL}_{\text{IND}}(\hat{\boldsymbol{f}})$ is the codelength for encoding the residual $\hat{\boldsymbol{e}}$ and it is given by the negative of the conditional log-likelihood of $\hat{\boldsymbol{e}}$ given $\hat{\boldsymbol{w}}$. Thus a modified criterion for non-Gaussian noise can be obtained by replacing the second term in $\text{MDL}_{\text{IND}}(\hat{\boldsymbol{f}})$ by $-\log_2 g(\hat{\boldsymbol{e}} \mid \hat{\boldsymbol{w}})$. Previous work on non-Gaussian noise can be found, for example, in Moulin (1994), Neumann & von Sachs (1995) and Gao (1997).

6.2. Correlated noise

Suppose that the noise e_i is correlated, and can be adequately modelled by an autoregressive (AR) series of an unknown order p:

$$e_i = a_1 e_{i-1} + a_2 e_{i-2} + \cdots + a_p e_{i-p} + \tau_i,$$

where $\boldsymbol{a} = \{a_1, \ldots, a_p\}$ are unknown AR parameters and τ_i is a Gaussian innovation. This reduces to the independent noise case when $p = 0$. Extension to autoregressive moving-average (ARMA) noise or even autoregressive integrated moving-average (ARIMA) noise is conceptually simple, but for computational reasons we do not pursue it here.

Under the autoregressive dependence, the codelength expression $L(\hat{\boldsymbol{e}} \mid \hat{\boldsymbol{w}})$ can be decomposed into

$$L(\hat{\boldsymbol{e}} \mid \hat{\boldsymbol{w}}) = L(\hat{\boldsymbol{a}} \mid \hat{\boldsymbol{w}}) + L(\hat{\boldsymbol{\tau}} \mid \hat{\boldsymbol{a}}, \hat{\boldsymbol{w}}),$$

where $\hat{\boldsymbol{a}}$ is an estimate of $\boldsymbol{a}$, $\hat{\boldsymbol{\tau}} = (\hat{\tau}_1, \ldots, \hat{\tau}_n)$ with $\hat{\tau}_i = \hat{e}_i - \hat{a}_1 \hat{e}_{i-1} - \cdots - \hat{a}_{\hat{p}} \hat{e}_{i-\hat{p}}$, and $\hat{p}$ is an estimate of p. Now as each of $\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_{\hat{p}}$ is estimated from n data points, $L(\hat{\boldsymbol{a}} \mid \hat{\boldsymbol{w}}) \approx \frac{1}{2}\hat{p} \log_2 n$ (see the heuristic argument given in the second paragraph of Section 4). Also, if $\hat{\boldsymbol{w}}$ and $\hat{\boldsymbol{a}}$ are reasonable estimates, then, by model assumption, $\hat{\boldsymbol{\tau}}$ would be (approximately) independent normal, and hence $L(\hat{\boldsymbol{\tau}} \mid \hat{\boldsymbol{a}}, \hat{\boldsymbol{w}}) \approx \frac{1}{2} n \log_2(\|\hat{\boldsymbol{\tau}}\|^2/n)$ (similar to (3)). Therefore we have

$$L(\hat{\boldsymbol{e}} \mid \hat{\boldsymbol{w}}) = L(\hat{\boldsymbol{a}} \mid \hat{\boldsymbol{w}}) + L(\hat{\boldsymbol{\tau}} \mid \hat{\boldsymbol{a}}, \hat{\boldsymbol{w}}) \approx \tfrac{1}{2}\hat{p} \log_2 n + \tfrac{1}{2} n \log_2 \frac{\|\hat{\boldsymbol{\tau}}\|^2}{n}.$$

The $\hat{e}_i$ can be treated as an AR series, and the above codelength expression agrees with the classical one discussed for example in Hannan & Quinn (1979).

Using steps similar to those in Section 6.1, we obtain the following criterion, which can be taken as an approximation for $L(\boldsymbol{y})$ when the noise is autoregressively correlated:

$$\mathrm{MDL}_{\mathrm{COR}}(\hat{\boldsymbol{f}}) = \tfrac{1}{2}(\hat{m} + \hat{p}) \log_2 n + \tfrac{1}{2} n \log_2 \frac{\|\hat{\boldsymbol{\tau}}\|^2}{n}. \tag{6}$$

When the noise is correlated, we propose that $\boldsymbol{f}$ can be estimated by the minimizer of the criterion $\mathrm{MDL}_{\mathrm{COR}}(\hat{\boldsymbol{f}})$, subject to the tree constraint.

Observe that $\mathrm{MDL}_{\mathrm{COR}}(\hat{\boldsymbol{f}})$ is a generalization of $\mathrm{MDL}_{\mathrm{IND}}(\hat{\boldsymbol{f}})$, as $\mathrm{MDL}_{\mathrm{COR}}(\hat{\boldsymbol{f}})$ reduces to $\mathrm{MDL}_{\mathrm{IND}}(\hat{\boldsymbol{f}})$ when $\hat{p} = 0$. For this reason, if it is not clear whether the noise is independent or correlated, we recommend using $\mathrm{MDL}_{\mathrm{COR}}(\hat{\boldsymbol{f}})$ as the target. In fact, in the numerical experiments reported in Section 7, we always use $\mathrm{MDL}_{\mathrm{COR}}(\hat{\boldsymbol{f}})$ as our target even when the noise is independent.

6.3. Tree-growing

The imposition of the tree constraint suggests a natural tree-growing algorithm for approximating the minimizer of $\mathrm{MDL}_{\mathrm{IND}}(\hat{\boldsymbol{f}})$ (or $\mathrm{MDL}_{\mathrm{COR}}(\hat{\boldsymbol{f}})$). The idea is straightforward: the root node $(\hat{w}_{00})$ is the initial tree, and at each time-step the tree grows by the addition of a best-embryo node. An embryo node is a node which is directly linked but does not belong to the current tree, and a best-embryo node is defined as the embryo node that has the largest absolute value amongst all other embryo nodes. (A best-embryo node can also be defined in a different way. For example, a best embryo node can be defined as the embryo node such that, when it is added to the current tree, it gives the largest reduction in the value of $\mathrm{MDL}_{\mathrm{IND}}(\hat{\boldsymbol{f}})$ or $\mathrm{MDL}_{\mathrm{COR}}(\hat{\boldsymbol{f}})$.)

The tree-growing algorithm continues until the size, in terms of number of nodes, of the tree hits a pre-set limit S_{MAX} (a maximal tree is of size $n - 1$). Thus, when the algorithm finishes, a nested sequence of S_{MAX} trees is produced, and the tree that gives the smallest value of $\mathrm{MDL}_{\mathrm{IND}}(\hat{\boldsymbol{f}})$ or $\mathrm{MDL}_{\mathrm{COR}}(\hat{\boldsymbol{f}})$ is taken as the final estimate. When we are searching for the minimizer of $\mathrm{MDL}_{\mathrm{COR}}(\hat{\boldsymbol{f}})$, an additional step for searching for a 'best' AR order is required, and that a maximal AR searching order p_{MAX} has to be imposed. Also, we may want to impose a minimum size limit S_{MIN} for the final chosen tree, recognising that, in some existing wavelet thresholding schemes, low level wavelet coefficients are not thresholded. Throughout our numerical experiments we set $S_{\mathrm{MAX}} = \frac{1}{2}n$, $S_{\mathrm{MIN}} = 0$ and $p_{\mathrm{MAX}} = 4$.

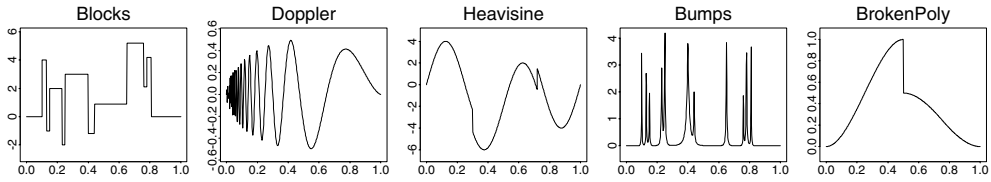

Figure 3. Plots of test functions

Saito's criterion (4) can, in principle, be extended to handle correlated noise, just as we extended criterion (5) to (6). However, the minimization of such a modified Saito's criterion may not be practically feasible, as the tree-growing algorithm can no longer be applied.

7. Numerical experiments

This section reports results of three numerical experiments. We used five test functions: the four test functions Blocks, Doppler, Heavisine and Bumps of Donoho & Johnstone (1994), and the broken polynomial test function (which we call BrokenPoly) of Nason & Silverman (1994). These five test functions are plotted in Figure 3. Also, throughout the three experiments, we used Daubechies's order 4 'least asymmetric' wavelet (Daubechies, 1992 pp. 198–199). For evaluating the quality of an estimated $\hat{\boldsymbol{f}}$, we used the numerical measure $\mathrm{MISE}(\hat{\boldsymbol{f}}) = \|\boldsymbol{f} - \hat{\boldsymbol{f}}\|^2$, which was taken as an approximation of the mean integrated squared error.

7.1. No noise

To investigate the possibility that imposing the tree constraint may prevent a true function from being completely recovered, we applied our tree-growing algorithm for minimizing $\mathrm{MDL}_{\mathrm{COR}}(\hat{\boldsymbol{f}})$ to recover the testing functions using the functions themselves as the observed data. That is, no noise was present. The recovered functions are virtually identical to the true functions, except for some numerical errors: the values of $\log \mathrm{MISE}(\hat{\boldsymbol{f}})$ for the five test functions are -46.82, -18.92, -23.06, -12.90 and -50.77, respectively. These values are much smaller than the corresponding values of $\log \mathrm{MISE}(\hat{\boldsymbol{f}})$ reported below.

7.2. Independent noise

In this subsection we investigate the relative practical performances of four wavelet regression methods when the noise is independent:

1. the SURE thresholding procedure of Donoho & Johnstone (1995);
2. the Bayesian thresholding procedure of Abramovich *et al.* (1998);
3. the MDL-based thresholding (MDL-Thresh) procedure of Saito (1994); and
4. the proposed tree-growing procedure (MDL-Tree) which aims to minimize $\mathrm{MDL}_{\mathrm{COR}}(\hat{\boldsymbol{f}})$ (rather than $\mathrm{MDL}_{\mathrm{IND}}(\hat{\boldsymbol{f}})$).

We defined the signal-to-noise ratio (SNR) as $\mathrm{SNR} = \|\boldsymbol{f}\|/\sigma$ (as in Donoho & Johnstone, 1994), and used three levels: high $\mathrm{SNR} = 9$, medium $\mathrm{SNR} = 7$ and low $\mathrm{SNR} = 5$. For each combination of test function and SNR, 50 sets of noisy observations were simulated, and the number of data points n for each dataset was 512.

For each simulated dataset, we applied the four wavelet methods listed above to estimate the test function. For each combination of test function and SNR, Figure 4 displays boxplots

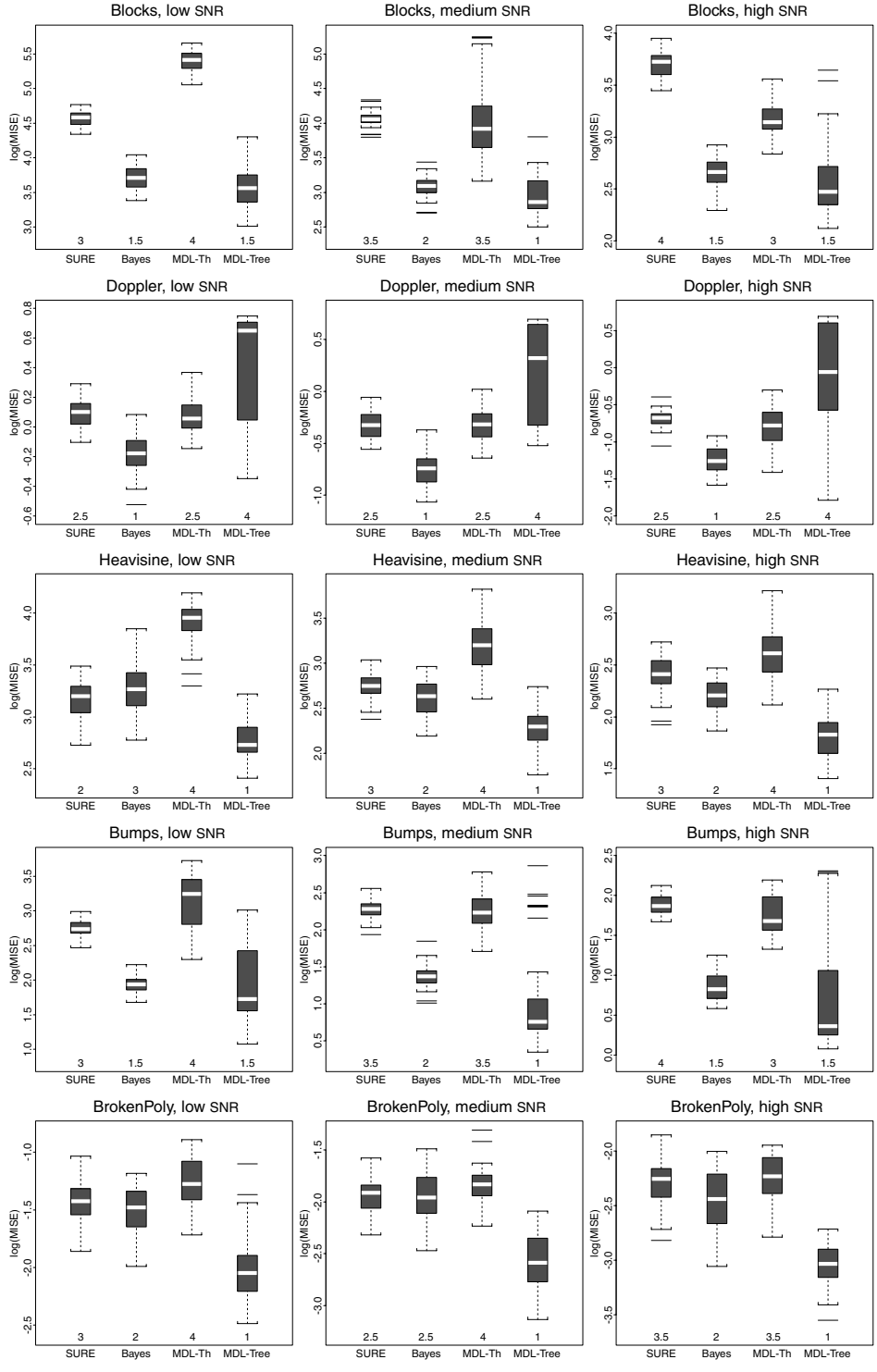

Figure 4. Boxplots of $\log(\mathrm{MISE}(\hat{f}))$ values for independent noise.
Numbers listed below the boxplots are relative paired Wilcoxon test rankings.

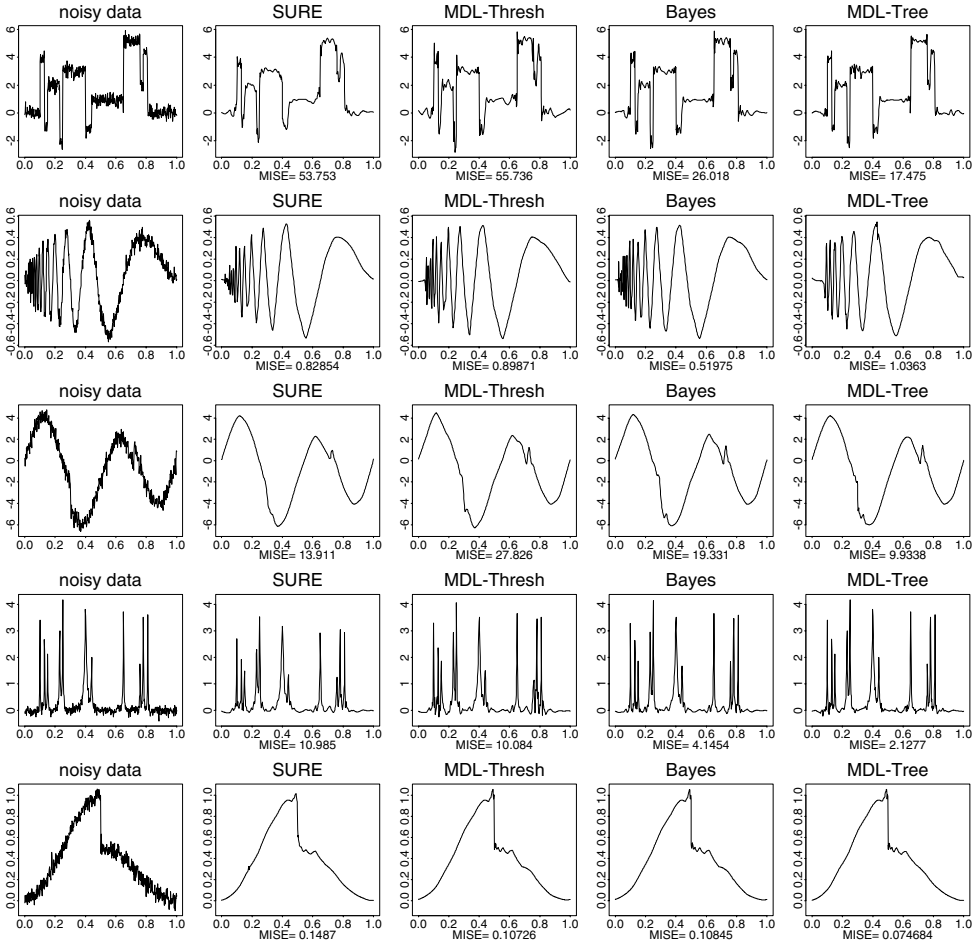

Figure 5. Estimates of the five test functions obtained from independent noisy observations, medium SNR

of the values of $\log \mathrm{MISE}(\hat{\boldsymbol{f}})$ (for all $\hat{\boldsymbol{f}}$). We also performed paired Wilcoxon tests to test if the difference between the median $\mathrm{MISE}(\hat{\boldsymbol{f}})$ values of two wavelet methods was significant or not. The significance level used was 1.25%, and the relative rankings, with 1 being the best, are also listed below the corresponding boxplots in Figure 4. Ranking the methods in this manner is not perfectly legitimate, but it provides an indicator of the relative merits of the methods (see Wand, 2000).

To visually evaluate and compare the performances of the four wavelet regression methods, we did the following. For the combination of the test function Blocks and medium SNR, we ranked the 50 $\hat{\boldsymbol{f}}$s obtained by the proposed tree-based method, according to their values of $\mathrm{MISE}(\hat{\boldsymbol{f}})$. The 25th best $\hat{\boldsymbol{f}}$, together with the corresponding simulated noisy data, is plotted in Figure 5. Estimates obtained by applying the other three wavelet regression methods to this same simulated noisy dataset are also plotted in Figure 5. The same procedure was repeated for the remaining four test functions, and the results are also displayed in Figure 5.

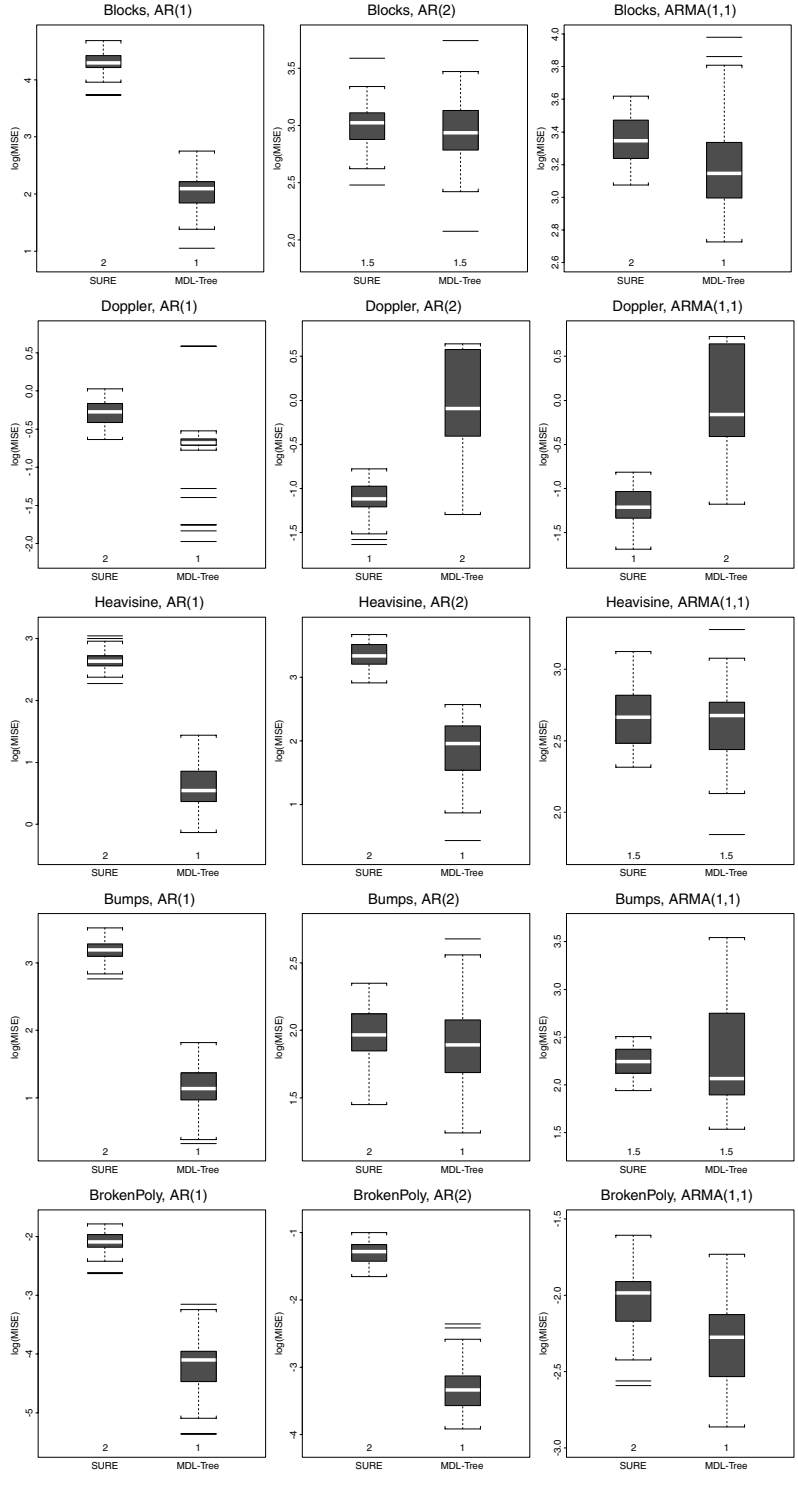

Figure 6. Boxplots of $\log(\mathrm{MISE}(\hat{f}))$ values for correlated noise.
Numbers listed below the boxplots are relative paired Wilcoxon test rankings.

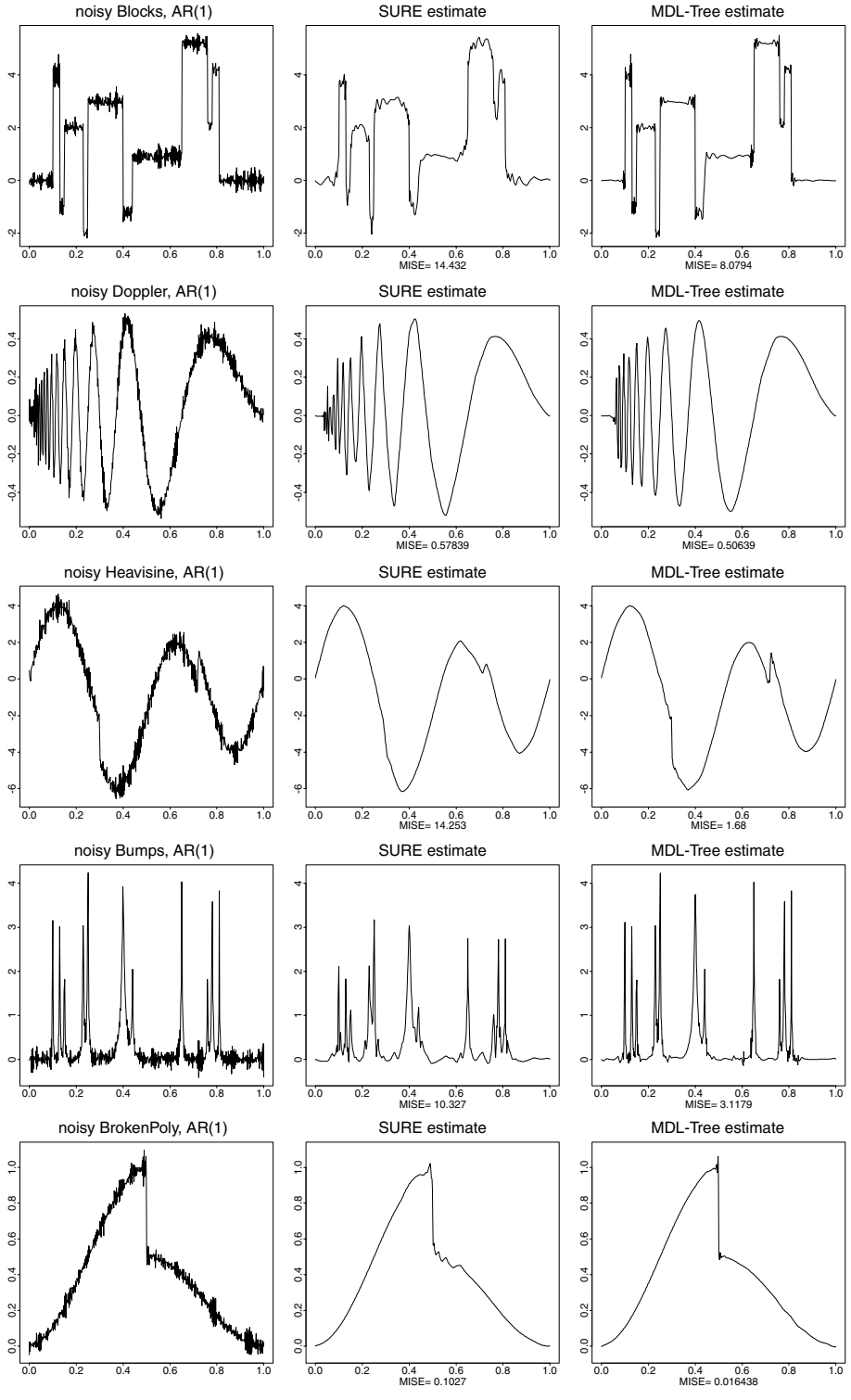

Figure 7. SURE and MDL-Tree estimates obtained from AR(1) correlated noisy observations

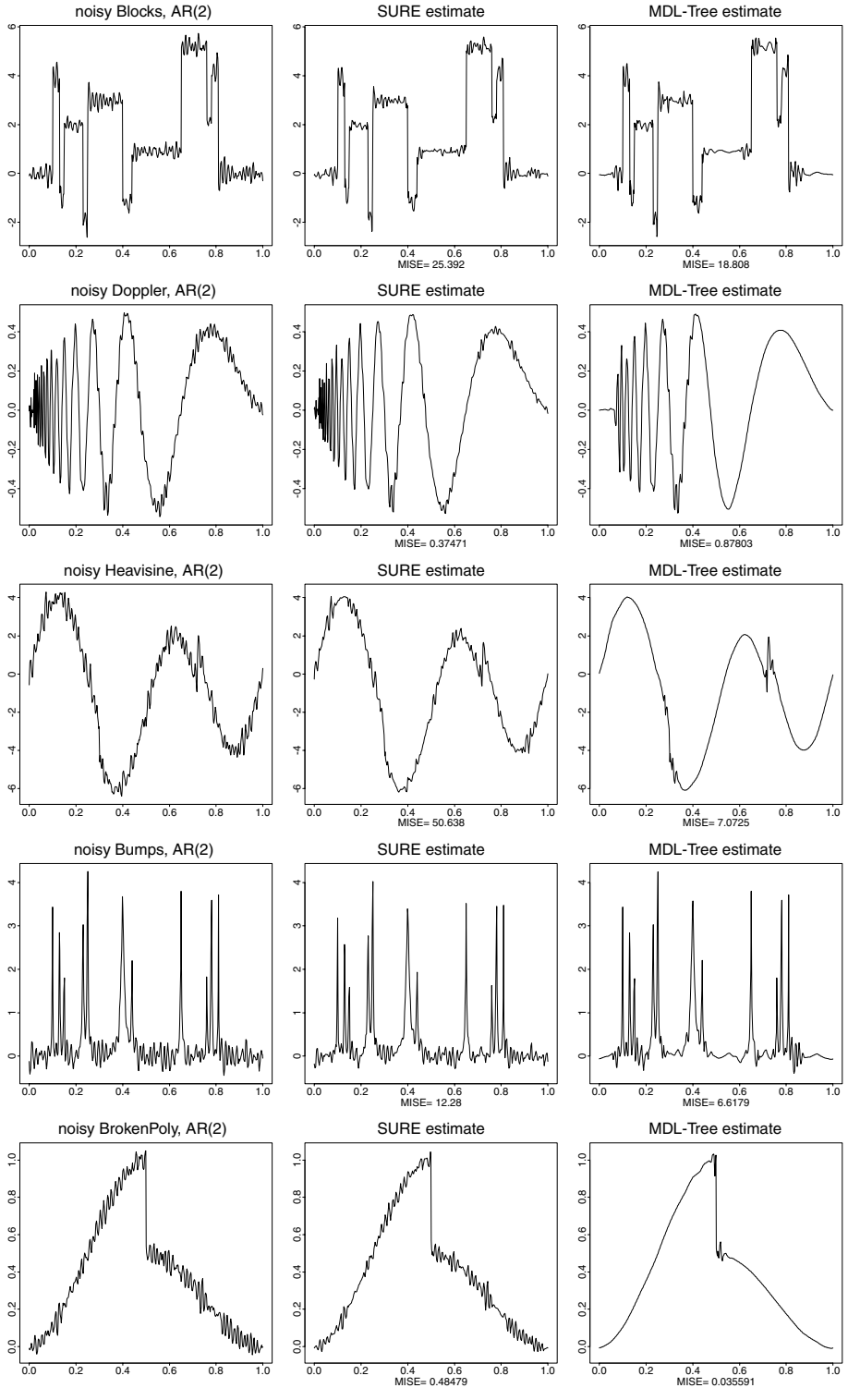

Figure 8. Similar to Figure 7 except for AR(2) noise

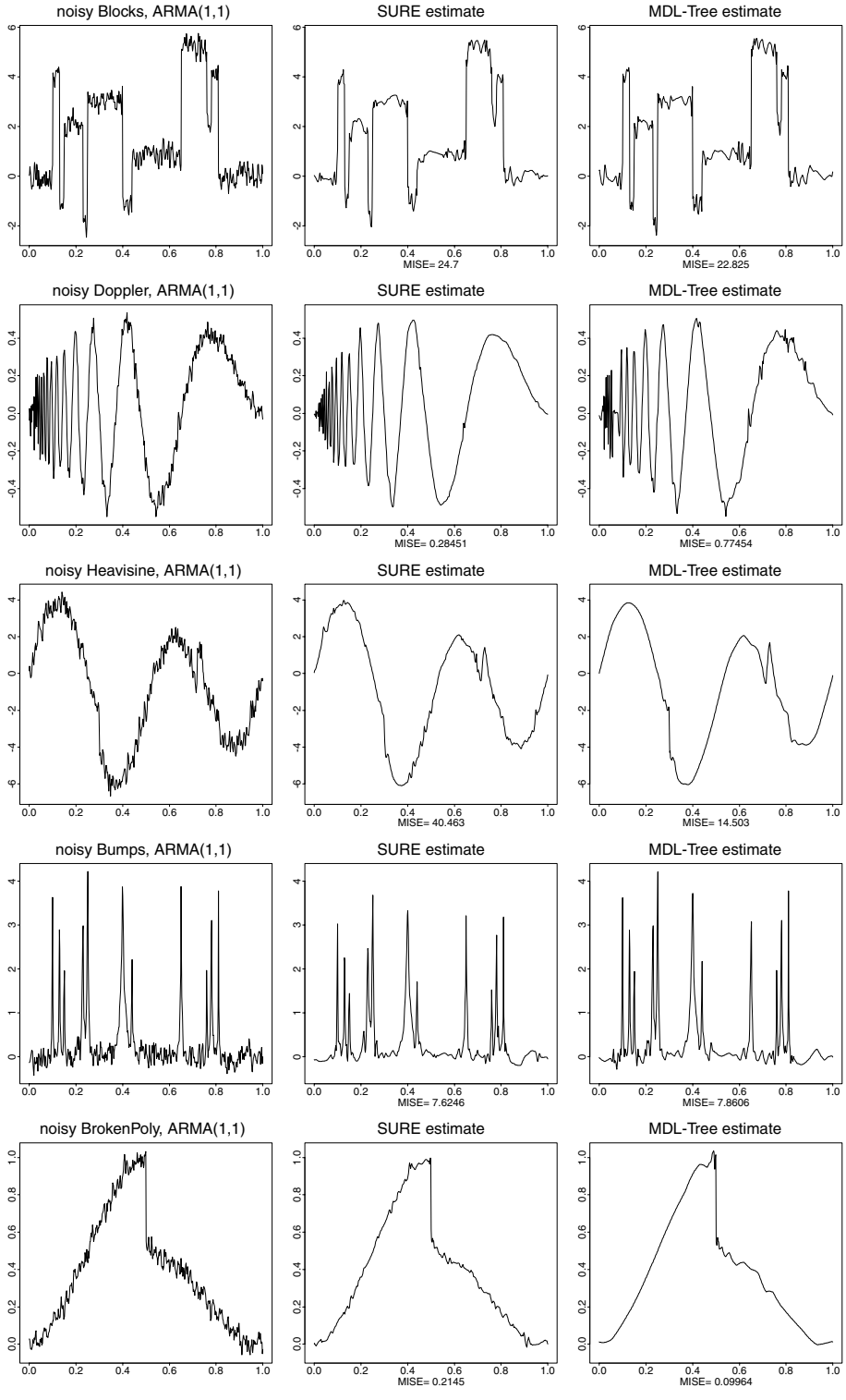

Figure 9. Similar to Figure 7 except for ARMA(1,1) noise

Except for those cases associated with Doppler, the MDL-Tree procedure compares favourably with the other three procedures, although it also exhibits some instability, as the length of some MDL-Tree boxplots is relatively large.

7.3. Correlated noise

Here we are interested in the performances of the SURE procedure and the proposed MDL-Tree procedure (which aims to minimize $\mathrm{MDL}_{\mathrm{COR}}(\hat{\boldsymbol{f}})$) when the noise is correlated. Johnstone & Silverman (1997) show that the SURE procedure is also applicable when the noise is correlated.

The setup for this correlated noise experiment was essentially the same as for the independent noise experiment, with the exception that the noise $\boldsymbol{e} = (e_1, \ldots, e_n)$ was generated from the ARMA(p, q) model

$$e_i = a_1 e_{i-1} + \cdots + a_p e_{i-p} + \tau_i + b_1 \tau_{i-1} + \cdots + b_q \tau_{i-q} \,,$$

where τ_i denotes a Gaussian innovation. Three different types of ARMA noise were considered: AR(1) with $a_1 = -0.8$, AR(2) with $a_1 = 4/3$ and $a_2 = -8/9$, and ARMA (1,1) with $a_1 = 0.2$ and $b_1 = -0.9$. Throughout the whole experiment, the noise was always linearly stretched so that $(\max(\boldsymbol{f}) - \min(\boldsymbol{f}))/(\max(\boldsymbol{e}) - \min(\boldsymbol{e})) = 5$. Note that Johnstone & Silverman (1997) also use the same AR(2) noise in their numerical examples, and that the ARMA(1,1) case does not satisfy the assumption made by $\mathrm{MDL}_{\mathrm{COR}}(\hat{\boldsymbol{f}})$.

Boxplots, together with the corresponding paired Wilcoxon test rankings, are displayed in Figure 6, and the '25th best estimates' are displayed in Figures 7–9. As before, except for Doppler, the MDL-Tree procedure compares favourably with the SURE procedure.

8. Conclusions

In this paper, a tree-based wavelet non-parametric regression procedure is proposed. This procedure is designed to handle autoregressively correlated noise, and applies the MDL principle to choose its best estimate. Results of numerical experiments demonstrate that, except for the cases of highly oscillating functions (e.g. Doppler), the proposed procedure provides satisfactory performances.

References

ABRAMOVICH, F., SAPATINAS, T. & SILVERMAN, B.W. (1998). Wavelet thresholding via a Bayesian approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **60**, 725–749.

ANTONIADIS, A., GIJBELS, I. & GREGOIRE, G. (1997). Model selection using wavelet decomposition and applications. *Biometrika* **84**, 751–763.

BREIMAN, L., FRIEDMAN, J.H., OLSHEN, R.A. & STONE, C.J. (1984). *Classification and Regression Trees.* Belmont, CA: Wadsworth.

CHIPMAN, H.A., KOLACZYK, E.D. & MCCULLOCH, R.E. (1997). Adaptive Bayesian wavelet shrinkage. *J. Amer. Statist. Assoc.* **92**, 1413–1421.

CROUSE, M.S., NOWAK, R.D. & BARANIUK, R.G. (1998). Wavelet-based statistical signal processing using hidden Markov models. *IEEE Trans. Signal Process.* **46**, 886–902.

DAUBECHIES, I. (1992). *Ten Lectures on Wavelets*, Vol. 1. Philadelphia: Society for Industrial and Applied Mathematics.

DONOHO, D.L. (1997). CART and best-ortho-basis: a connection. *Ann. Statist.* **25**, 1870–1911.

DONOHO, D.L. & JOHNSTONE, I.M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425–455.

DONOHO, D.L. & JOHNSTONE, I.M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.* **90**, 1200–1224.

DONOHO, D.L., JOHNSTONE, I.M., KERKYACHARIAN, G. & PICARD, D. (1995). Wavelet shrinkage: asymptopia? (With discussion). *J. R. Stat. Soc. Ser. B Stat. Methodol.* **57**, 301–369.

ENGEL, J. (1994). A simple wavelet approach to nonparametric regression from recursive partitioning schemes. *J. Multivariate Anal.* **49**, 242–254.

GAO, H-Y. (1997). Choice of thresholds for wavelet shrinkage estimate of the spectrum. *J. Time Ser. Anal.* **18**, 231–251.

HANNAN, E.J. & QUINN, B.G. (1979). The determination of the order of an autoregression. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **41**, 190–195.

HURVICH, C.M. & TSAI, C-L. (1998). A crossvalidatory AIC for hard wavelet thresholding in spatially adaptive function estimation. *Biometrika* **85**, 701–710.

JOHNSTONE, I.M. & SILVERMAN, B.W. (1997). Wavelet threshold estimators for data with correlated noise. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **59**, 319–351.

LUO, Z. & WAHBA, G. (1997). Hybrid adaptive splines. *J. Amer. Statist. Assoc.* **92**, 107–116.

MALLAT, S.G. (1989). A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans. Pattern Analysis and Machine Intelligence* **11**, 674–693.

MOULIN, P. (1994). Wavelet thresholding techniques for power spectrum estimation. *IEEE Trans. Signal Process.* **42**, 3126–3136.

MOULIN, P. (1996). Signal estimation using adapted tree-structured bases and the MDL principle. In *Proc. IEEE Signal Processing Symposium on Time-Frequency and Time-Scale Analysis*, pp. 141–143. Paris.

NASON, G.P. (1996). Wavelet shrinkage using cross-validation. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **58**, 463–479.

NASON, G.P. & SILVERMAN, B.W. (1994). The discrete wavelet transform. *J. Comput. Graph. Statist.* **3**, 163–191.

NEUMANN, M.H. & VON SACHS, R. (1995). Wavelet thresholding: beyond the Gaussian i.i.d. situation. In *Wavelets and Statistics*, eds A. Antoniadis & G. Oppenheim, pp. 301–330. New York: Springer-Verlag.

QUINLAN, J.R. & RIVEST, R.L. (1989). Inferring decision trees using the minimum description length principle. *Inform. and Comput.* **80**, 227–248.

RISSANEN, J. (1989). *Stochastic Complexity in Statistical Inquiry*. Singapore: World Scientific.

RISSANEN, J. (1997). Stochastic complexity in learning. *J. Comput. System Sci.* **55**, 89–95.

SAITO, N. (1994). Simultaneous noise suppression and signal compression using a library of orthonormal bases and the minimum description length criterion. In *Wavelets in Geophysics*, eds E. Foufoula-Georgiou & P. Kumar, pp. 299–324. New York: Academic Press.

SOLO, V. (1998). Wavelet signal estimation in coloured noise with extension to Transfer Function estimation. In *Proc. 37th IEEE CDC*, Tampa, Florida.

VIDAKOVIC, B. (1998). Nonlinear wavelet shrinkage with Bayes rules and Bayes factors. *J. Amer. Statist. Assoc.* **93**, 173–179.

WALLACE, C.S. & PATRICK, J.D. (1993). Coding decision trees. *Machine Learning* **11**, 7–22.

WAND, M.P. (2000). A comparison of regression spline smoothing procedures. *Comput. Statist.* **15**, 443–462.

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