Wavelet regression by cross-validation

G P Nason*†

Technical Report 447. Department of Statistics,
Stanford University, Stanford
14th April 1994

Abstract

This paper is about using wavelets for regression. The main aim of the paper is to introduce and develop a cross-validation method for selecting a wavelet regression threshold that produces good estimates with respect to L_2 error. The selected threshold determines which coefficients to keep in an orthogonal wavelet expansion of noisy data and acts in a similar way to a smoothing parameter in non-parametric regression.

The paper gives a very brief introduction to wavelets and how the discrete wavelet transform with a thresholding rule can be used to estimate functions from noisy data. The paper examines the integrated square error (ISE) of a soft-thresholded wavelet estimator as a function of the threshold and shows that the ISE is almost always convex and almost always has negative first derivative for a zero threshold. This implies that it is almost always possible to find a unique minimum for the ISE.

A cross-validation score is introduced to estimate the ISE of a soft-thresholded estimator. The score is based on splitting the data into two sets with one set consisting of odd-indexed observations and the other containing the even-indexed ones. Such a splitting is convenient because the discrete wavelet transform, which is involved in computing a regression, only operates on data sets containing a power of two number of data points. The theoretical behaviour of the cross-validation score as a function of the threshold is investigated and compared to the true ISE which it closely resembles.

The penultimate part of this paper compares the cross-validation procedure with other threshold-choice methods for a three functions and three different noise structures. The comparisons suggest that for sparse signals the cross-validation method successfully estimates the 'true' ISE and can outperform other methods. Comparisons using non-normal noise show

*Research partially supported by NSF grant DMS 9209130 at Stanford University
†Address for correspondence: Department of Mathematics, University of Bristol, University Walk, BRISTOL BS8 1TW, United Kingdom.

that the cross-validated choice is more resilient to outliers that other methods. However, all of the methods we used were unable to find exactly the "true" threshold when the data were correlated.

1 Introduction

This paper is about the regression modelling of data using wavelets. Wavelets are a recent unification and development of many ideas from the fields of signal processing, physics and applied mathematics (see Daubechies (1992) and Smith and Barnwell (1986) for example). Some of the potential uses of wavelets for statistical problems have been discussed and developed by Donoho and Johnstone (1993), Donoho and Johnstone (1994) and Fan, Hall, Martin and Patil (1993) for regression and Johnstone, Kerkyacharian and Picard (1992) for density estimation.

This paper relates mainly to the work on regression by Donoho and Johnstone (1993) and Donoho and Johnstone (1994). The former paper developed a theory for the minimax estimation of functions using wavelets and proved that orthonormal wavelet expansions with a simple thresholding rule provide near-optimal estimates of the true function (optimal in an asymptotic minimax sense and only with independent Gaussian noise). Further, the optimality of estimation extends to functions over a wide class of function spaces (not just L_2) because wavelets are a special kind of basis for certain function spaces called *unconditional* bases. The reader is referred to Donoho, Johnstone, Kerkyacharian and Picard (1994) for further details.

In this work we will consider the function space $L^2(\mathcal{R})$: the space of measurable, square-integrable one-dimensional functions. In principle the methods are also applicable to other more general function spaces. For regression wavelet-based estimators are expansions of orthogonal wavelets where the coefficients of the expansion are computed in some data-dependent fashion. They are used because of their ability to adapt to diverse features of a function such as discontinuities and varying frequency behaviour. Donoho and Johnstone (1993) suggest two types of thresholding rule (hard and soft) and two means of selecting an appropriate threshold (optimal and universal). In a subsequent paper they introduce another threshold selector called *SureShrink* that minimizes an unbiased estimate of the risk. Section 2.3 describes these methods. The main contribution of our work is the development of a another method: computation of a threshold using cross-validation. This cross-validation also attempts to minimize the risk (mean integrated square error) but makes no assumption about the type of noise structure.

A threshold can be thought of as a type of smoothing parameter for a wavelet estimator. In practice universal thresholding is used as it is simple to compute and it was the one of first means for choosing thresholds. In certain examples it has been found that universal thresholding tends to underfit (the threshold is too high and knocks out or shrinks too many wavelet coefficients). The practical consequence of underfitting depends on the underlying wavelet basis: if a rough wavelet is used then underfitting produces unnatural looking estimates that resemble the wavelet; if a smooth wavelet is used then the estimate just looks oversmoothed. Cross-validated (and *SureShrink*) estimates do not suffer from this oversmoothing and tend to match the optimal thresholds well. This tendency of universal thresholding to underfit was also noted by Fan, Hall, Martin and Patil (1993) and they suggest other means of correcting for it.

This paper is organized into five sections. In the next section we briefly describe wavelets and wavelet regression and outline the problems concerned with the regression. In Section 3 the integrated square error (ISE) of a soft-thresholded wavelet estimator as a function of the threshold is investigated and it is shown that the ISE is almost always convex and therefore has a unique minimum. Section 3 lays the foundation for Section 4 which introduces our cross-validation method as an estimator for the ISE. (The foundation consists of two parts: Section 3 shows that it is reasonable to try and minimise a cross-validated estimate of the ISE because it is almost always convex and secondly it provides formulae that aid interpretation of the cross-validation score itself). In Section 5 we compare a modified version of *SureShrink*, universal thresholding and the cross-validation method on three different functions using three different types of noise structure. Finally, in Section 6 an example is given of how wavelet-like methods may be of some use in signal extraction from noise.

2 Wavelet regression

Suppose data $g_1, \ldots, g_n$ can be modeled by

$$g_i = f(t_i) + \epsilon_i, \tag{1}$$

where the $\{\epsilon_i\}$ are iid $N(0, \sigma^2)$, $t_i = i/n$ and f is the function we would like to estimate. This model is exactly the same as the one considered by Donoho and Johnstone (1993) and like them we are interested in in examining the performance of an estimate $\hat{f}$ in terms of quadratic loss at the sample points. Following Donoho and Johnstone we let $\mathbf{f} = \{f(t_i)\}_{i=1}^n$ and $\hat{\mathbf{f}} = \left\{\hat{\mathbf{f}}(t_i)\right\}_{i=1}^n$ be the vectors of true and estimated sample values. Let $||\mathbf{v}_i^2||$ denote the usual squared ℓ_n^2 norm. Then performance is measured by the risk

$$R(\hat{f}, f) = n^{-1} E||\hat{\mathbf{f}} - \mathbf{f}||_{2,n}^2,$$

which we would like to make as small as possible. The remainder of this paper is about building $\hat{f}$ using wavelets. The next section gives a brief overview of wavelets and their role in regression.

2.1 Wavelets

A general introduction to the theory of wavelets can be found in Strang (1993). An introduction that considers the statistical uses of wavelets can be found in Nason and Silverman (1994). More technical and complete surveys include Daubechies (1992) and Chui (1992).

There are many families of wavelets. Each family can be derived from a single special function known as the *mother* wavelet and usually denoted $\psi(x)$. Within a family wavelets are derived from the mother wavelet by a process of scaling and translation in the following manner:

$$\psi_{j,k}(x) = 2^{\frac{j}{2}}\psi(2^j x - k), \quad j,k \in \mathcal{Z}.$$

Providing the mother wavelet is carefully chosen the family $\{\psi_{j,k}(x)\}$ furnishes the space L_2 with a particular orthonormal basis.

Properties of wavelets

One of the key properties of wavelets is their ability to spatially adapt to features of a function and to represent functions parsimoniously. For example, a discontinuity within a function could be represented efficiently by few wavelets whereas other bases, such as the Fourier, may require many more basis functions. This aspect of wavelets is part of a more general feature of wavelets: they exhibit excellent time-frequency localization properties.

There are several plausible wavelet families each generated by a single mother wavelet. It is possible to choose a particular family and tailor it for a specific application. The wavelets used in this paper are the 'extremal phase' wavelets created by Daubechies (1988). These wavelets have the advantage that they are compactly supported and there exists a hierarchy of possible mother wavelets of increasing regularity. The mothers are each indexed by N denoted by $\psi^N(x)$ and have regularity linear in N. The compact support aids time-frequency localization and interpretation of regression models in that it is very easy to see which wavelets contribute to a particular part of the regression. The choice of regularity is also useful when prior information is available as to the smoothness (or lack of) of the true function as it may be possible to tailor the choice of wavelet to the known degree of smoothness. For all the later examples we have arbitrarily chosen the $N = 6$ mother wavelet. This is depicted in Figure 1.

Suppose a function f from $L^2(\mathcal{R})$ is given. The *wavelet decomposition* of f with respect to the orthonormal wavelet basis $\{\psi_{j,k}(x)\}$ is written as:

$$f(x) = \sum_{jk} f_{jk}\psi_{jk}(x). \tag{2}$$

The *wavelet coefficients* are found in the usual way:

$$f_{jk} = \int f(x)\psi_{j,k}(x)\,dx. \tag{3}$$

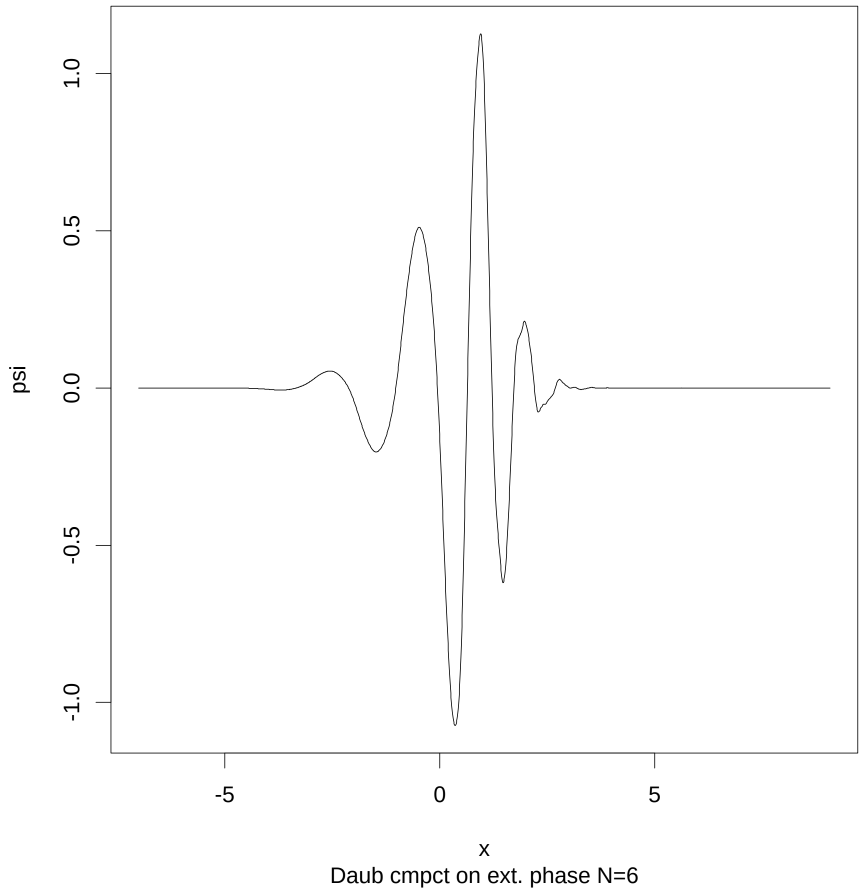


Figure 1: The mother wavelet– Daubechies'(1988) extremal phase wavelet $N = 6$.

The Discrete Wavelet Transform

Our ultimate objective is to estimate the function f. However, the noisy data $g_1, \ldots, g_n$ are discrete and consequently a discrete version of (3) must be performed to find wavelet coefficients. The discrete wavelet transform (DWT) is a discretized version of (3) and can be written as a linear transform involving an orthogonal matrix W derived from the mother wavelet. The DWT of data $g = (g_1, \ldots, g_n)$ is given by

$$w = Wg.$$

If this formula were implemented exactly as it is written then the transform would require $O(n^2)$ operations. However, a remarkable fast pyramidal algorithm exists when $n = 2^M$. This fast algorithm is due to Mallat (1989) and requires only $O(n)$ operations. We use the fast algorithm as implemented in SPlus by Nason (1993) as use of this package affords easy access to the rich statistical environment of S. The inverse DWT is easy because W is orthogonal. The recovery of g from w is achieved by

$$g = W^T w,$$

which also has a fast pyramidal version. From the reconstruction formula (inverse DWT) it can be seen that the rows of W correspond to discretized versions of the mother wavelet at various different scalings and translations. Indeed Donoho and Johnstone (1993) note the approximation

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

where $W_{j,k}(i)$ is the i^{th} element of the $(j,k)^{th}$ row of W. In this way the wavelet coefficients $\{w_{j,k}\}$ quantifies the contribution of the basis function $W_{j,k}$ which is localized to a spatial interval of size 2^{-j} and near frequency 2^j. In other words $w_{j,k}$ indicates the amount of signal around spatial location $2^{-j}k$ and near frequency proportional to 2^j. A popular method of plotting the wavelet coefficients can be seen in many figures in this paper. For example, Figure 7a shows a signal and Figure 7b shows a plot of its wavelet coefficients. Roughly speaking the horizontal axes in both plots correspond whilst the vertical axis in Figure 7b corresponds to frequency. A detailed interpretation of these plots appears in Nason and Silverman (1994).

2.2 Wavelet Regression

A wavelet regression estimator works as follows:

1. Find the DWT of $g = (g_1, \ldots, g_n)$ as follows:

$$w = Wg. \tag{4}$$

2. Modify the wavelet coefficients w to produce the set: w^*.

3. Reconstruct an estimate of f by

$$\hat{f} = W^T w^*. \tag{5}$$

The crucial part of the regression is step 2. The primary method of modification is *thresholding*. Donoho and Johnstone (1993) reasoned that thresholding was appropriate for the signal+noise model because

1. the wavelet transform of signal is concentrated into just a relatively small number of coefficients (parsimonity of transform);

2. the DWT of white noise is just white noise and is therefore spread amongst all coefficients (because W is orthogonal).

Therefore a suitable heuristic would be to find a robust estimate of the noise level and then threshold the coefficients using the estimated noise level. A robust estimate is necessary so that 'signal' coefficients do not unduly interfere with the noise level estimate. All that would remain would be the signal wavelet coefficients plus a small amount of noise. The next section describes Donoho and Johnstone's formalization of this heuristic.

2.3 Thresholding rules

Nearly all the relevant work on thresholding rules and optimal thresholds is contained within Donoho and Johnstone (1993), Donoho, Johnstone, Kerkyacharian and Picard (1994) and Donoho and Johnstone (1994). Given a wavelet coefficient w and threshold $t > 0$ the hard-thresholded value is given by

$$T_{\text{hard}}(w; t) = w\, I(|w| > t),$$

and the soft-thresholded value by

$$T_{\text{soft}}(w; t) = \text{sgn}(w)\,(|w| - t)\, I(|w| > t),$$

where I is the usual indicator function. Hard-thresholding is a 'keep-or-kill' strategy whereas soft-thresholding is 'shrink-or-kill'.

2.3.1 Choosing the threshold

Wavelet regression also requires a method for selecting a threshold t as well as having a thresholding rule. Donoho and Johnstone (1993) proposed two different means of selecting a threshold value and called them *optimal* and *universal*. Optimal thresholds are precomputed values that minimize a constant term in the upper bound for the minimax risk of estimating a function using a thresholded estimator. However, optimal thresholds may not work in all situations (considerations of risk) and have only been computed for soft-thresholding. The universal threshold is

$$t_{\mathrm{UV}} = \sqrt{2\log(n)}\hat{\sigma}. \tag{6}$$

The universal threshold combines an estimate of the noise level $\hat{\sigma}$ with the multiplier $\sqrt{2\log(n)}$. The multiplier ensures that the universal estimate is noise-free (with probability 1) because given a white noise sequence $\{z_i\}$ we have

$$P\left\{\max_i |z_i| > \sqrt{2\log(n)}\right\} \to 0, \qquad n \to \infty$$

This universal threshold is easy to remember, compute and is asymptotically optimal but as well shall see in later examples it tends to underfit the data.

Donoho and Johnstone (1994) devise another means of estimating a threshold value which they call *SureShrink*. This estimator is based on Stein's method for estimating the quadratic loss unbiasedly for a decision problem with independent Gaussian noise. The SureShrink method as described by Donoho and Johnstone (1994) works in a level by level fashion computing a threshold for each level in the wavelet decomposition. The threshold at each level is computed by the following method: decide whether there is enough signal to compute the SURE threshold and if

Yes find the minimum, t_{SURE} of the quantity $SURE(t; \mathbf{x})$ for that level;

No use the universal threshold on that level.

We performed a small simulation study (which is not reported here) that found that *VisuShrink* and *SureShrink* give very similar results. This comparison agrees with Donoho and Johnstone's simulation study in (1994, Section 5.1). In general, both methods can underfit (remove or shrink too many coefficients) especially at high resolution levels. However, it is interesting to note that using SureShrink simultaneously on all coefficients actually results in reasonable performance with respect to mean square error. We shall call this *GlobalSure* and compare it with the cross-validated method that is developed later. We emphasize most strongly that this paper does *not* use *SureShrink* itself but a modified version of it.

2.4 Problems with wavelet regression

At the present time there appear to be some practical obstacles to the appropriate use of wavelets in regression. The primary ones are:

- no method yet exists for when the design is non-uniform;
- no method yet exists for data sets that do not contain a power of two data elements;

When we write 'no method' above we really mean no fast method. In principle it is possible; just as it is possible to compute the ordinary discrete Fourier transform with any number of points. However, as with the discrete Fourier transform, the method is not necessarily fast.

The other problem of wavelet regression that this paper attempts to address is the choice of threshold. Donoho and Johnstone (1993) remark that the mean square error performance of the optimal minimax thresholds is noticeably better on samples of the size of tens or hundreds. The practical behaviour of the universal threshold seems to be that it is generally larger than the optimal minimax threshold. This means than more coefficients are either knocked out or shrunk by the universal thresholding and in general an underfitted estimate is produced. This underfitting was also noted by Fan et al. (1993) who suggested a way to correct for it by locally adapting the level of primary resolution. The *Sure* procedure of Donoho and Johnstone (1994) is one possible and promising method for obtaining a threshold. An alternative approach is adopted here: that of selecting a threshold by a cross-validation method.

3 ISE of soft-thresholded wavelet estimators

The objective for the regression is to minimize the mean integrated square error between the estimate $\hat{f}_t(x)$ and the true function $f(x)$. In symbols we wish to find a threshold t that minimizes

$$M(t) = E \int \left\{ \hat{f}_t(x) - f(x) \right\}^2 dx. \qquad (7)$$

In practice the function f is not known and so an estimate of M that can be minimized has to be devised. One popular way to do this in regression and many other statistical methods is to minimize this estimate of M and let this minimizer be our estimate of the threshold that minimizes M. A simple and well-known way of doing this for ordinary regressions is cross-validation: remove an observation from the pack; do regression on the remainder; measure the discrepancy between the deficient regression and the removed observation; repeat for each observation and an estimate obtained from all the discrepancies. Cross-validation in this form cannot be applied to the wavelet regression exactly as described above. The problem occurs because wavelet regression requires that you use 2^M data points. If you remove one point then you have $2^M - 1$ data points which is no longer a power of 2 (except for the trivial case of $M = 1$).

One way of adapting cross-validation is to remove half of the data points not just one. Then you are left with 2^{M-1} data points on which you can perform wavelet regression.

The cross-validation algorithm is described in Section 4. Before the algorithm is described the properties of the integrated square error (ISE) as a function of the threshold are investigated. We show that on the interval of interest the ISE is 'nearly-always' convex and if it is assumed that the wavelet transform of the true function is relatively sparse then the gradient of the ISE at $t = 0$ is 'nearly-always' negative. These two properties together show that the ISE can nearly always be minimized and that the cross-validated score estimates a well-behaved quantity.

3.1 Exploring the Integrated Square Error

Suppose data are given according to model (1). Then the discrete version of the integrated square error (7) is

$$m(t) = \sum_{j=1}^{M} \left(\hat{f}_{t,j} - f_j \right)^2 ,$$

where $\hat{f}_{t,j}$ is the discretized thresholded estimator of f_j. Let $\{v_{jk}\}$, $\{w_{jk}\}$ and $\{\epsilon_{jk}\}$ be the wavelet transforms of g_j, f_j and ϵ_j respectively. Note that

$$v_{jk} = w_{jk} + \epsilon_{jk},$$

and $\{\epsilon_{jk}\}$ is white noise because the wavelet transform is orthogonal. Then Parseval's relation may be used to express m in terms of wavelet coefficients:

$$m(t) = \sum_{j,k} \left\{ T_{\text{soft}}(v_{jk}; t) - w_{jk} \right\}^2 .$$

For notational convenience let $v^{(r)}$ and $w^{(r)}$ be the ordered versions of $|v_{jk}|$ and $|w_{jk}|$ (sometimes the ordering of w will be tied to v).

3.1.1 Differentiation of the Integrated Square Error

To obtain good estimates we will be interested in minimizing values of (7). Consequently, the minimizer of m is also of interest. The first derivative of m with respect to t is:

$$m'(t) = 2 \sum_{j,k} \left\{ T_{\text{soft}}(v_{jk}; t) - w_{jk} \right\} \frac{\partial T_{\text{soft}}(v_{jk}; t)}{\partial t}. \tag{8}$$

The derivative of the soft-threshold with respect to t is

$$\frac{\partial T_{\text{soft}}(v; t)}{\partial t} = -\text{sgn}(v)\, I(|v| > t),$$

for all $t \in \mathcal{R}$. Note that the derivative of T_{soft} is not continuous at $t = v$. The derivative of m from (8) becomes

$$m'(t) \quad = \quad -2 \sum_{j,k} \left\{ \text{sgn}(v_{jk})\,(|v_{jk}| - t)\, I(|v_{jk}| > t) - w_{jk} \right\} \text{sgn}(v_{jk}) I(|v_{jk}| > t)$$

$$= \quad 2 \sum_{j,k} I(|v_{jk}| > t) \left\{ w_{jk} \mathrm{sgn}(v_{jk}) - (|v_{jk}| - t) \right\} . \tag{9}$$

Now since $|v| = \mathrm{sgn}(v)\, v$ we have

$$m'(t) = 2 \sum_{j,k} I(|v_{jk}| > t) \left\{ \mathrm{sgn}(v_{jk})\, (w_{jk} - v_{jk}) + t \right\} , \tag{10}$$

for all $t \in \mathcal{R}$. The formula for the derivative of m shows that $m'(t)$ is linear increasing on each interval defined by increasing values of $|v_{jk}|$ and that the derivative of m has discontinuities at each $|v_{jk}|$. As t increases past $|v_{jk}|$ the term $\mathrm{sgn}(v_{jk})\,(w_{jk} - v_{jk}) + |v_{jk}|$ is dropped from (10). Therefore,

$$m'(|v_{jk}|^{+}) = m'(|v_{jk}|^{-}) - 2\mathrm{sgn}(v_{jk})\, w_{jk},$$

where $m'(\cdot^{-})$ and $m'(\cdot^{+})$ are the left and right derivatives of m at t.

3.1.2 The form of the first derivative

The derivative of m is linearly increasing on each interval defined by successive pairs of increasing $|v_{jk}|$. At each $|v_{jk}|$ there is a random jump of size

$$X_{jk} = -2\mathrm{sgn}(w_{jk} + \epsilon_{jk}) w_{jk}.$$

The ideal situation would be for all these jumps to be positive or zero. Then m would be convex. However, it is possible for X to be negative but at small values of $|v_{jk}|$ the quantity X_{jk} is almost always near zero (see Figure 2). This actually occurs because many of the wavelet coefficients of the signal w_{jk} will be nearly zero. This is a consequence of the fact that wavelets form unconditional bases (Donoho reference) and represent most signals sparsely. Also, in practice the cross-validation score is minimized on an interval $[0, t_{\mathrm{UV}}]$ where t_{UV} is the universal threshold of Donoho and Johnstone. It is usually the case that the large negative values of X_{jk} occur outside of the search interval ($|v_{jk}| > t_{\mathrm{UV}}$) and so the cross-validation score is almost-convex on the search interval.

The expectation of X_{jk} is:

$$EX = -w \left\{ 1 - 2\Phi\left(-\frac{w}{\sigma} \right) \right\} ,$$

and the variance is:

$$\mathrm{var}(X) = w^2 \left[1 - \left\{ 1 - 2\Phi\left(-\frac{w}{\sigma} \right) \right\}^2 \right] .$$

A plot of EX and $EX \pm 2\sqrt{\mathrm{var}(X)}$ appears in Figure 2. The picture shows that the for small w the expected value of X is zero.

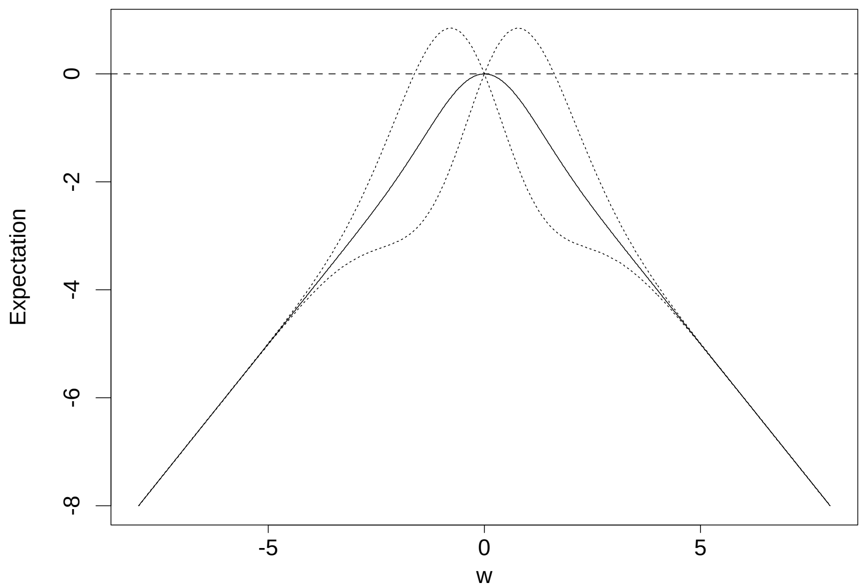


Figure 2: Mean and confidence bands for $X = -\text{sgn}(w + \epsilon)w$ as a function of w.

3.1.3 The first derivative at zero

The previous section showed heuristically that the first derivative is almost monotone increasing. In this section it is shown (again heuristically) that with overwhelming probability that the $Z = m'(0)$ is negative. This is shown by computing the mean and variance of Z and then showing it is approximately normal.

Definition: Z, Y_{jk}
Define Y_{jk} to be $\operatorname{sgn}(v_{jk})\epsilon_{jk}$. Let Z be the value of the derivative of m at $t = 0$. In symbols:

$$Z = m'(0) = -2 \sum_{j,k} Y_{jk}. \tag{11}$$

□

Lemma 3.1 *The mean and variance of Y_{jk} are:*

$$E(Y_{jk}) = 2\sigma\phi\left(\frac{w_{jk}}{\sigma}\right), \tag{12}$$

and

$$\operatorname{var}(Y_{jk}) = \sigma^2 \left\{1 - 4\phi^2\left(\frac{w_{jk}}{\sigma}\right)\right\}. \tag{13}$$

Proof: In the following proof the jk subscripts are dropped for clarity. The mean of Y_{jk} is given by

$$\begin{aligned} E(Y) &= E\left\{\operatorname{sgn}(w+\epsilon)\epsilon\right\} \\ &= \sigma\left\{\int_{-w/\sigma}^{\infty} u\phi(u)\,du - \int_{-\infty}^{-w/\sigma} u\phi(u)\,du\right\} \\ &= 2\sigma\int_{-w/\sigma}^{\infty} y\phi(y)\,dy \\ &= 2\sigma\left(\frac{w}{\sigma}\right). \end{aligned}$$

It is easy to see that $E(Y^2) = E(\epsilon^2) = \sigma^2$ and therefore

$$\operatorname{var}(Y) = \sigma^2 - 4\sigma^2\phi^2\left(\frac{w}{\sigma}\right) = \sigma^2\left\{1 - 4\phi^2\left(\frac{w}{\sigma}\right)\right\}.$$

□

Lemma 3.2 *The mean and variance of Z are given by:*

$$E(Z) = -4\sigma\sum_{jk}\phi\left(\frac{w_{jk}}{\sigma}\right), \tag{14}$$

and

$$\operatorname{var}(Z) = 4\sigma^2\sum_{jk}\left\{1 - 4\phi^2\left(\frac{w_{jk}}{\sigma}\right)\right\}. \tag{15}$$

Proof: These are obtained from (12) and (13) and the independence of the $\{\epsilon_{jk}\}$. $\square$

Assumption 1 *For a given $f \in L^2(\mathcal{R})$ assume that the wavelet transform w_{jk} is sparse. That is we assume that there exists N_{cutoff} not much less than N such that*

$$|w^{(r)}| \approx 0 \quad \textit{for all } r \leq N_{\text{cutoff}}.$$

We also assume that the remainder of the wavelet coefficients are at least an order of magnitude larger than the noise level.

Using the assumption the approximate normality of Z is now demonstrated.

Approximation 3.1 *Z, the first derivative of the ISE at $t = 0$, is approximately normally distributed with mean and variance given by (14) and (15).*

Proof: The next step is to show that Z is approximately normally distributed. To do this the formula for Z from (11) is broken into two pieces U and V at the cutoff point:

$$\begin{aligned} -Z/2 &= \sum_{jk} Y_{jk} = U + V \\ &= \sum_{r \leq N_{\text{cutoff}}} \operatorname{sgn}(w^{(r)} + \epsilon^{(r)})\epsilon^{(r)} + \sum_{r > N_{\text{cutoff}}} Y^{(r)}, \end{aligned} \tag{16}$$

where $\epsilon^{(r)}$ and $Y^{(r)}$ are ordered according to $w^{(r)}$.

For $r \leq N_{\text{cutoff}}$ we have $w^{(r)} \approx 0$ and therefore each component of U is like $\operatorname{sgn}(\epsilon)\epsilon$ and distributed with mean $(2\pi)^{-\frac{1}{2}} 2\sigma$ and variance $\sigma^2(1 - 2\pi^{-1})$. For large N_{cutoff} we have

$$U \sim N\{2(2\pi)^{-\frac{1}{2}}\sigma N_{\text{cutoff}},\, \sigma^2(1 - 2\pi^{-1})N_{\text{cutoff}}\},$$

approximately by the central limit theorem.

For $r > N_{\text{cutoff}}$ the value of w^r is large compared to the noise level. In this case $Y^{(r)}$ is distributed like $\operatorname{sgn}(w^{(r)})\epsilon^{(r)}$ which is distributed the same as $\epsilon^{(r)}$. Therefore

$$V \sim N\{0,\, \sigma^2(N - N_{\text{cutoff}})\}.$$

approximately. Therefore

$$Z \sim N\{-4(2\pi)^{-\frac{1}{2}}\sigma N_{\text{cutoff}},\, \sigma^2(N - 2\pi^{-1}N_{\text{cutoff}})\} \tag{17}$$

approximately. This shows that Z is approximately normally distributed. The mean and variance of Z were established in Lemma 3.2. $\square$

The next approximation shows how likely it is for Z to be negative.

Approximation 3.2 *The probability that Z is negative is given approximately by:*

$$P(Z<0)=\Phi\left(\frac{4(2\pi)^{-\frac{1}{2}}N_{\mathrm{cutoff}}}{\sqrt{N-2\pi^{-1}N_{\mathrm{cutoff}}}}\right).$$

A more accurate approximation may be obtained by also using the actual mean and variance of Z given by (14) and (15).

Proof: Uses the fact that Z is approximately normally distributed with the approximate mean given in Approximation 3.1. □

Approximation 3.2 shows that the probability of the first derivative at zero being negative is always greater than even. In fact, except for all but non-sparse signals, the probability is nearly one as Figure 3 shows. This shows that for 'most' signals the probability is usually 1. Finally, the main conclusion from this section is stated next.

Heuristic 1 *With large probability the integrated square error $m(t)$ is almost convex on $[0, t_{\mathrm{UV}}]$ with a negative first derivative at $t=0$. Therefore with large probability there exists a threshold t_{CV} that uniquely minimizes m.*

Proof: The discussion in Section 3.1.3 establishes the near-convexity of m. Approximation 3.2 shows that $m'(0)<0$ with large probability. □

4 Cross-Validation for Wavelet Regression

Suppose data $g_1,\ldots,g_n$ are given where $n=2^M$. The cross-validation algorithm then proceeds in two separate mirror-image stages. First all the odd indexed g_i are removed from the set. This leaves 2^{M-1} evenly indexed g_i which are re-indexed from $j=1,\ldots,2^{M-1}$. A regression estimate $\hat{f}_t^{\mathrm{EVEN}}$ is then constructed using a particular threshold t from the re-indexed g_j. Using the removed odd indexed data an interpolated version of the odd noisy data is formed:

$$\bar{g}_j^{\mathrm{ODD}}=\begin{cases}\frac{1}{2}\left(g_{2j-1}+g_{2j+1}\right) & j=1,\ldots,\frac{n}{2}-1\\ \frac{1}{2}\left(g_1+g_{n-1}\right) & j=\frac{n}{2}\end{cases}$$

The above is computed similarly for $\hat{f}_t^{\mathrm{ODD}}$ and $\bar{g}^{\mathrm{EVEN}}$. The full estimate for $M(t)$ is then

$$\hat{M}(t)=\sum_j\left(\hat{f}_{t,j}^{\mathrm{EVEN}}-\bar{g}_j^{\mathrm{ODD}}\right)^2+\left(\hat{f}_{t,j}^{\mathrm{ODD}}-\bar{g}_j^{\mathrm{EVEN}}\right)^2, \tag{18}$$

where $\hat{f}_{t,j}^{\mathrm{EVEN}}$ is the estimate of of $\hat{f}_t$ at the index point j. Notice that the estimate $\hat{M}$ relies on estimates of $\hat{f}_t$ based upon $\frac{n}{2}$ data points. From the work of Donoho and Johnstone (1993) it is known that the appropriate threshold

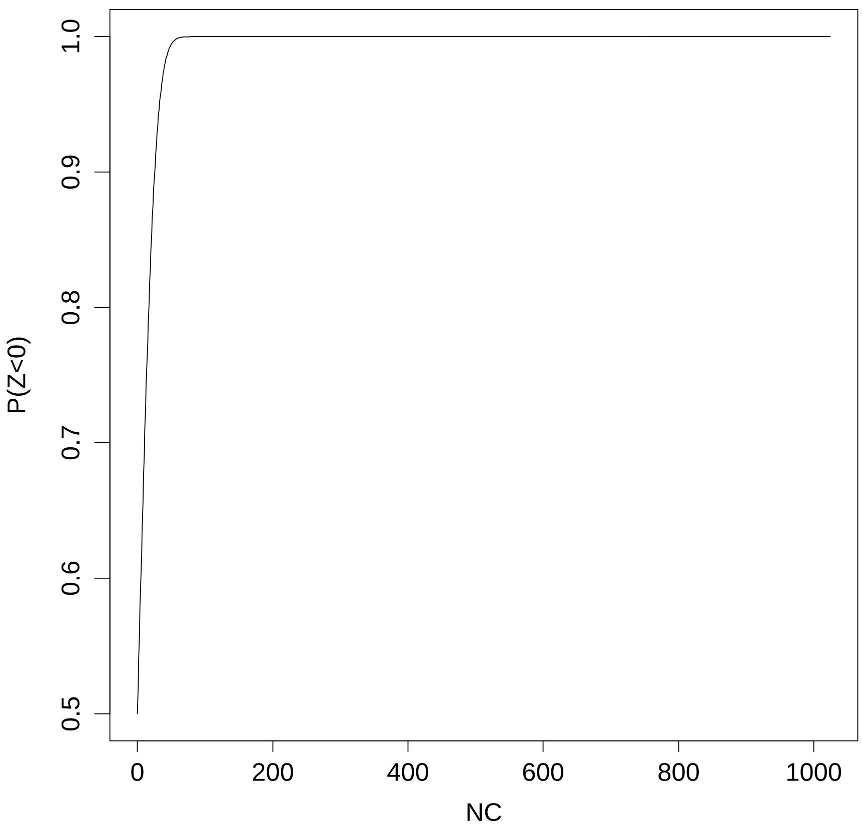


Figure 3: The probability that the first derivative at zero, Z, is negative as a function of sparsity of signal.. This example is based on a signal based on $N = 1024$ data points.

depends on n and asymptotically is $\sqrt{2\log(n)}\hat{\sigma}_n$. This supplies a heuristic method for obtaining a threshold for n data points. If the threshold for n points is $t_n = \sqrt{2\log(n)}\hat{\sigma}_n$ then the threshold for $\frac{n}{2}$ points will be $t_{\frac{n}{2}} = \sqrt{2\log(n/2)}\hat{\sigma}_{\frac{n}{2}}$ and therefore

$$t_n \approx \left(1 - \frac{\log 2}{\log n}\right)^{-\frac{1}{2}} t_{n/2}. \tag{19}$$

After the estimate $\hat{M}(t)$ has been minimized the correction (19) is applied to obtain the final cross-validated threshold.

4.1 Investigating the integrated square error

For a given data set write $\{e_{jk}\}$ and $\{o_{jk}\}$ for the wavelet transforms of the even and odd numbered indices of the noisy data g. Similarly write $\bar{o}_{jk}$ and $\bar{e}_{jk}$ for the wavelet transform of the interpolated noise $\bar{g}_j^{\text{ODD}}$ and $\bar{g}_j^{\text{EVEN}}$. Then by Parseval's relation:

$$\hat{M}(t) = \sum_{j,k} \{T_{\text{soft}}(e_{jk};t) - \bar{o}_{jk}\}^2 + \{T_{\text{soft}}(o_{jk};t) - \bar{e}_{jk}\}^2$$

Using formula (10) for the derivative of m the following formulae for $\hat{M}'(t)$ is obtained:

$$\begin{aligned} \hat{M}'(t) &= 2\sum_{j,k} I(|e_{jk}| > t)\{\text{sgn}(e_{jk})(\bar{o}_{jk} - e_{jk}) + t\} \\ &+ 2\sum_{j,k} I(|o_{jk}| > t)\{\text{sgn}(o_{jk})(\bar{e}_{jk} - o_{jk}) + t\}. \end{aligned} \tag{20}$$

The derivative will be continuous and linear increasing on intervals defined by increasing absolute values of $\{e_{jk}, o_{jk}\}$ and as t increases past $|e_{jk}|$ the term

$$\text{sgn}(e_{jk})(\bar{o}_{jk} - e_{jk}) + |e_{jk}| = \text{sgn}(e_{jk})\bar{o}_{jk}$$

is dropped from the derivative. (Similarly, as t increases past $|o_{jk}|$ the term $\text{sgn}(o_{jk})\bar{e}_{jk}$ is dropped from the derivative.) Typically, the (possibly negative) jumps in the interval of interest will be small and heuristically, on the average, the cross-validation $\hat{M}$ score will be convex. This time it is much less clear that the first derivative of $\hat{M}(t)$ is negative at $t = 0$. However, if $\bar{o}_{jk}$ estimates w_{jk} and e_{jk} estimates v_{jk} well then we can have some confidence that $\hat{M}'(t)$ will be negative with large probability.

To give a real example Figure 4 shows all values of the 'jumps' in $\hat{M}$ as a function of the absolute values at $\{e_{jk}, o_{jk}\}$. Figure 4 actually shows two things. First, that if the absolute value of a noisy coefficient is small then the corresponding jump is also small. Secondly, for relatively sparse signals most of the coefficients are small (in Figure 4 933 of the 1024 points are to the left of the

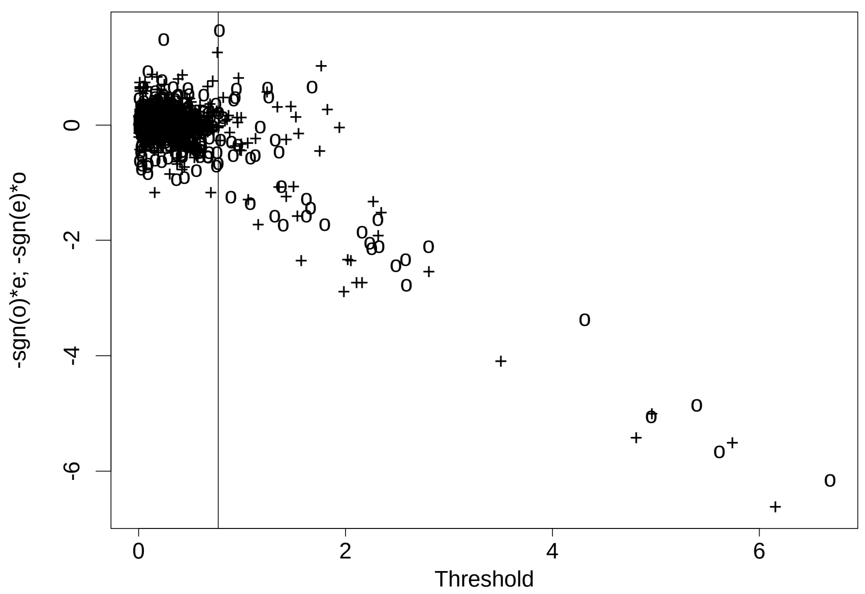


Figure 4: The jumps + $= (|e_{jk}|, -\mathrm{sgn}(e_{jk})\bar{o}_{jk})$ and o $= (|o_{jk}|, -\mathrm{sgn}(o_{jk})\bar{e}_{jk})$

line - i.e. most are small). Although Figure 4 is derived using a noisy chirp signal similar plots are obtained for other signals (see next section for a description of the signals that have been used). The preceding discussion provides a heuristic justification of the near-convexity for the cross-validation score with a negative first derivative at $t = 0$.

5 Comparing the Threshold-Choosing Methods

This section compares the cross-validation method with *GlobalSure* and universal thresholding. The following abbreviations are used for each method:

MAD: universal thresholding using a mean absolute deviation estimate of the noise level (corrected so that it estimates σ);

SURE: the *GlobalSure* procedure using a corrected mean absolute deviation estimate of the noise level;

CV: the cross-validation procedure.

Soft thresholding was used for every method. All simulation was carried out using the *WaveThresh* package of Nason (1993). All the wavelets that are used are Daubechies' 'extremal phase' wavelets (Daubechies (1988)). The same wavelet $N = 6$ as depicted in Figure 1 was used in every wavelet transform. For each signal both types of reconstruction were ran in two modes: considering all the wavelet coefficients and considering all coefficients not in the 'top' three levels (the lowest resolution ones). These two groups will be referred to as 'all' and 'subset'.

5.1 The data and added noise

We chose to ascertain the performance of the three methods on three functions. The three true functions appear as the top left plot in Figures 5, 6 and 7. The wavelet transforms for two of the original functions appears as the top right plot in Figure 6 and 7. For each combination of function and reconstruction method three different types of noise are used. They are abbreviated:

normal: independently distributed normal noise;

heavy: independently distributed Student's t noise on 3 d.f.;

corr: normal observations with an autocorrelation of $\frac{1}{2}$ at lag 1, otherwise uncorrelated.

All the noise is distributed with zero mean. The variance is also scaled appropriately to achieve a different noise level for each signal but constant over methods and noise structures.

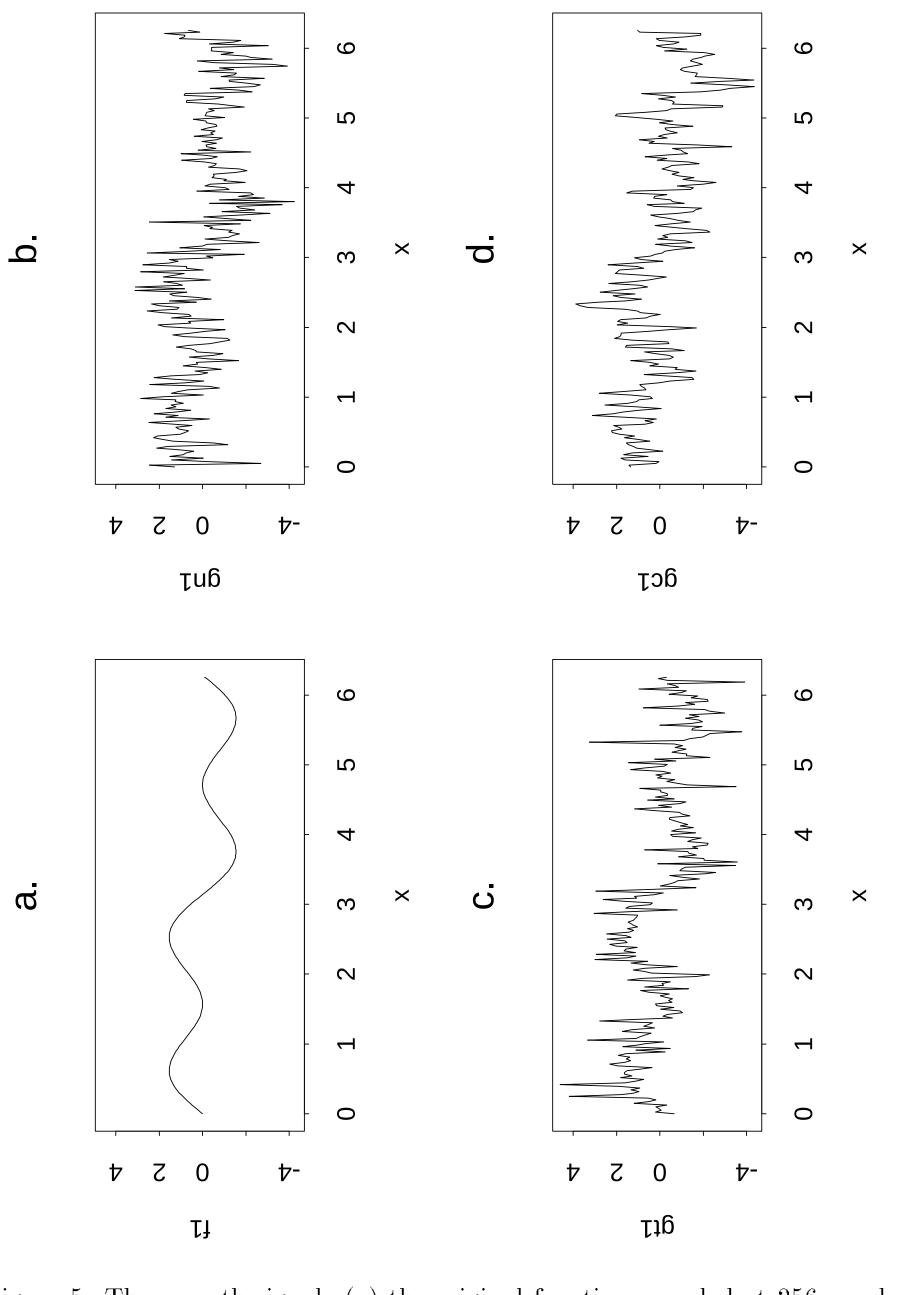


Figure 5: The smooth signal: (a) the original function sampled at 256 evenly spaced points; (b) (a) with independent pseudo-random normal noise; (c) (a) with independent pseudo-random Student's t noise; (d) (a) with pseduo-random autocorrelated normal noise. All noise was distributed with mean zero and unit variance.

The functional form of the smooth signal $f_1(x)$ is

$$f_1(x) = \sin(x) + \sin(3x)$$

and then this function is sampled 256 time evenly on the interval $[0, 2\pi]$. The normal, heavy and correlated noisy versions, $g_1(x)$, appear in Figure 5. The noise level for the smooth function is $\sigma = 1$.

The functional form of the piecewise polynomial with a discontinuity is the same function that appeared in Nason and Silverman (1994). It is defined on $[0, 1]$ to be:

$$f_2(x) = \begin{cases} 4x^2(3-4x) & x \in [0, \frac{1}{2}] \\ \frac{4}{3}x(4x^2 - 10 \times x + 7) - \frac{3}{2} & x \in (\frac{1}{2}, \frac{3}{4}] \\ \frac{16}{3}x(x-1)^2 & x \in (\frac{3}{4}, 1] \end{cases}$$

and then this function is sampled 256 times evenly on $[0, 1]$. The three noisy versions of this signal have the normal, heavy and correlated noise with noise level of 0.1. This time we choose to show the original signal, the autocorrelated signal and their wavelet transforms in Figure 6.

The functional form of the chirp signal is defined on the interval $[-1, 1]$ as:

$$f_3(x) = \sin\left(\frac{\pi}{x}\right).$$

This function is sampled 1024 times on [-1,1]. The noisy versions, $g_3(x)$, have independent Gaussian, Student's t and correlated normal pseudo-random noise with zero mean and a standard deviation of 0.3 added to $f_3(x)$. The noisy normal version, $g_3(x)$, and its wavelet transform appear in the bottom left and right of Figure 7.

5.2 Reconstruction results

To assess the different reconstruction procedures the following simulations were performed. For each of the true functions three types of noise were added and each of the three reconstruction methods selected a threshold. Each simulation was repeated 100 times to assess the variability of each reconstruction method. All the the reconstruction methods were applied twice in each simulation: once where all wavelet coefficients were subject to thresholding (ALL) and once where the three lowest frequency coefficient levels were left untouched (SUBSET) and the others subjected to the threshold. In almost all cases the SUBSET simulations were superior. This is because all of the reconstruction methods can (and usually do) kill at least one low-frequency components which then has a drastic affect on the reconstruction. So the results are reported below mostly for the SUBSET case.

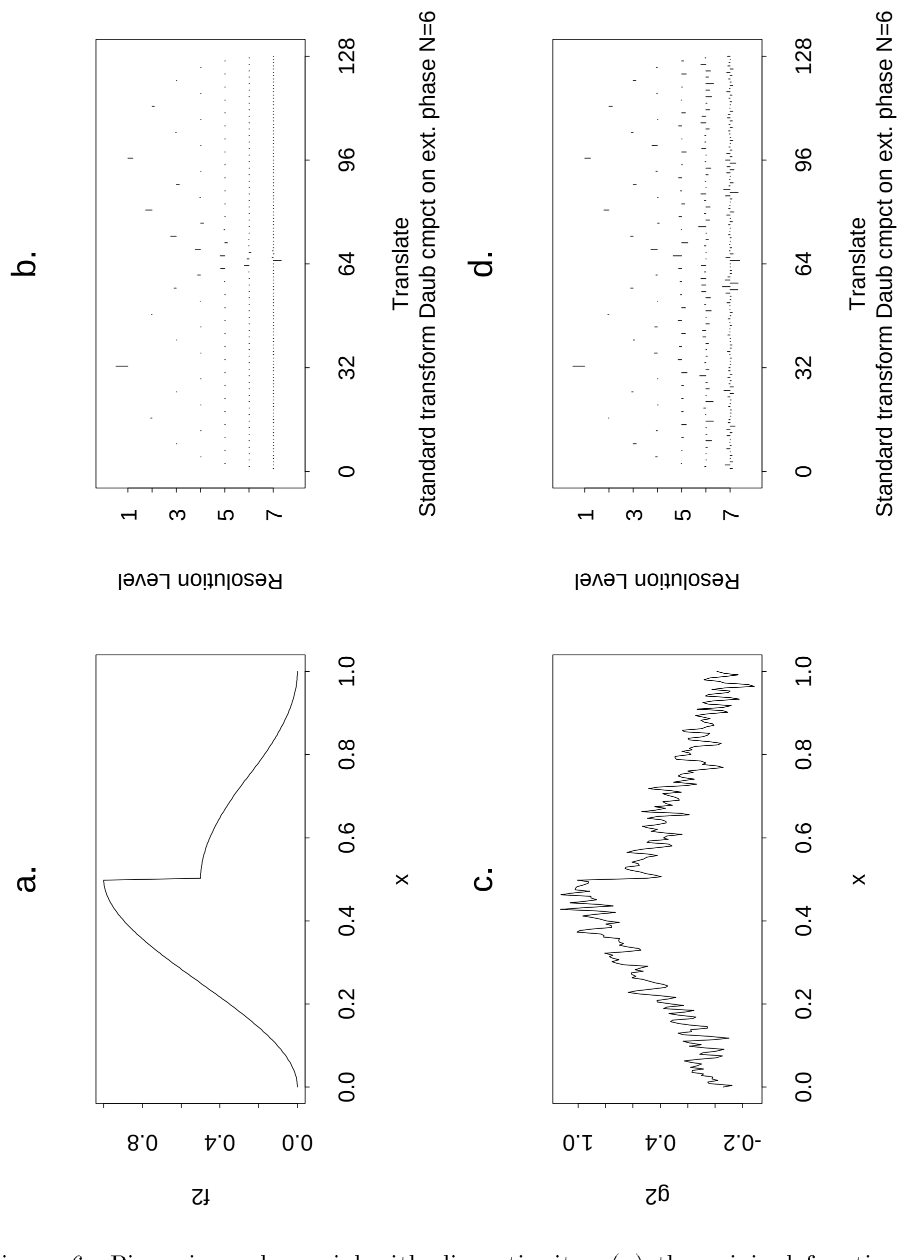


Figure 6: Piecewise polynomial with discontinuity: (a) the original function sampled at 256 evenly spaced points; (b) the discrete wavelet transform of (a); (c) (a) with added pseudo-random autocorrelated normal noise; (d) the discrete wavelet transform of (c). All added noise has zero mean and standard deviation of 0.1

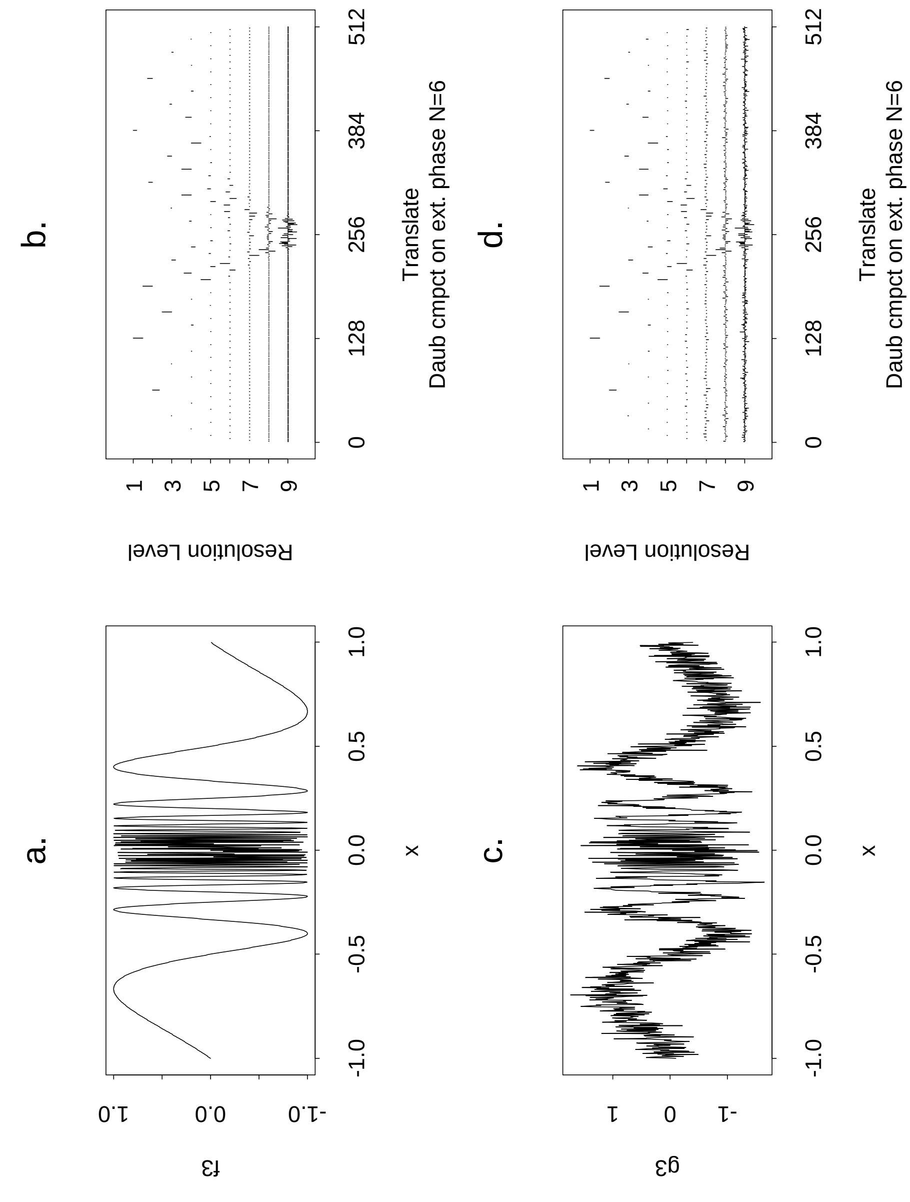


Figure 7: The chirp signal: (a) the original function sample at 1024 evenly spaced points; (b) the discrete wavelet transform of the original function using Daubechies wavelet $N = 6$; (c) the original function with independent pseudo-random normal noise with standard deviation of 0.3; (d) the discrete wavelet transform of (c) with Daubechies wavelet $N = 6$.

5.2.1 The best possible results.

The other quantity that needs to be evaluated is the 'true' optimal threshold. This is the threshold that minimizes the integrated square error between the soft-thresholded noisy function and the *true* original signal. This was computed for every combination of true signal, noise structure, and ALL and SUBSET level retention. Example figures for the independent normal noise appear in Figure 8.

5.2.2 How the simulation results are reported

For each combination of true signal and noise structure we provide a scatter-plot. On the x-axis the 'true' optimal thresholds are plotted (as computed by minimizing the integrated squared error between the original function and the thresholded noisy function). On the y-axis we plot the thresholds that were computed using each of the three reconstruction methods. The universal thresholds are marked by dots, *GlobalSure* by small circles and cross-validated thresholds by small crosses. The line $x = y$ is also plotted as this gives some idea of whether the reconstruction method underfits (above the line: too many wavelet coefficients were removed because the threshold was too high) or overfits (below the line: too many wavelet coefficients were retained because the chosen threshold was too low). Thresholds that are clustered on or near to the line $y = x$ can indicate a good reconstruction for that method on that signal/noise combination.

5.3 Reconstructions for normal independent noise

The reconstruction thresholds versus the 'true' optimal thresholds for normal noise and ALL level thresholding are plotted in Figure 9 and for the SUBSET levels thresholding in Figure 10. The same plots for piecewise polynomial appear in Figure 11 and for the chirp signal in 12 (SUBSET only, the ALL plots are very similar).

In using independently normally distributed data there seems to be no clear choice. The universal thresholding procedure seems to be the consistent worst performer of the three thresholding procedures. Six actual reconstructions for the piecewise polynomial are shown in Figure 13 and for the chirp signal in 14. Of course, the noisy data for the simulations changes each time and so the reconstructions are *not* what the reconstructions would have looked like. However, they do give a good idea of how the reconstructions vary for each reconstruction method. The cross-validation threshold choice does better than *GlobalSure* when estimating the smooth signal and worse when estimating the chirp signal. This may have something to do with the sparsity of the signals as *SureShrink* tends to do better with non-sparse signals. Figure 10 shows that no method does particularly well for the SUBSET choice. This is because for the simple

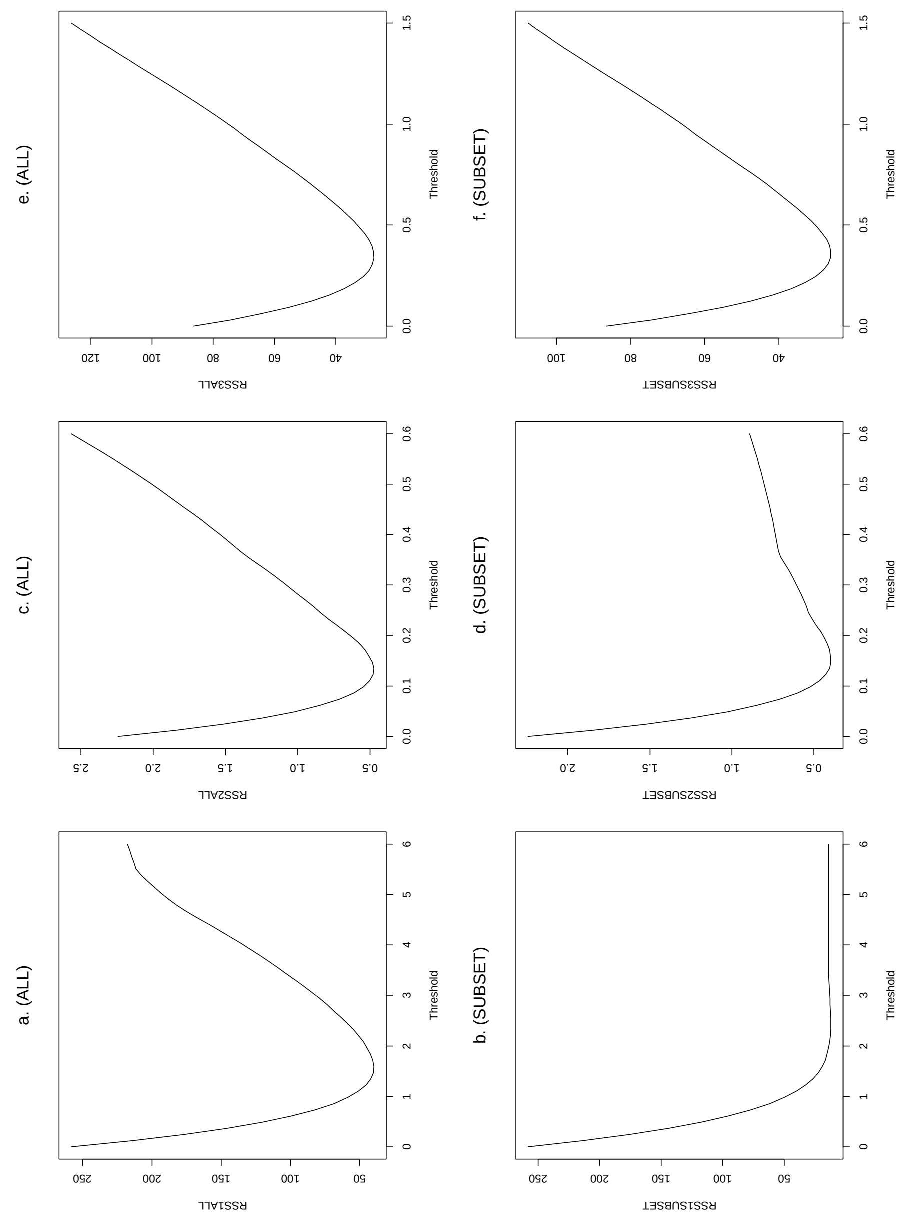


Figure 8: The residual sum of square between the thresholded estimates and the true functions for each of the three functions and ALL/SUBSET thresholding methods with normal added noise. The minima (with the value at the minima in parentheses) are: (a.) smooth function, ALL, 1.59 (30.8); (b.) smooth function, SUBSET, 2.45 (12.2); (c.) piecewise polynomial with discontinuity, ALL, 0.134 (0.474); (d.) piecewise polynomial with discontinuity, SUBSET, 0.146 (0.397); (e.) chirp function, ALL, 0.337 (27.5); (f.) chirp function, SUBSET, 0.367 (26.0).

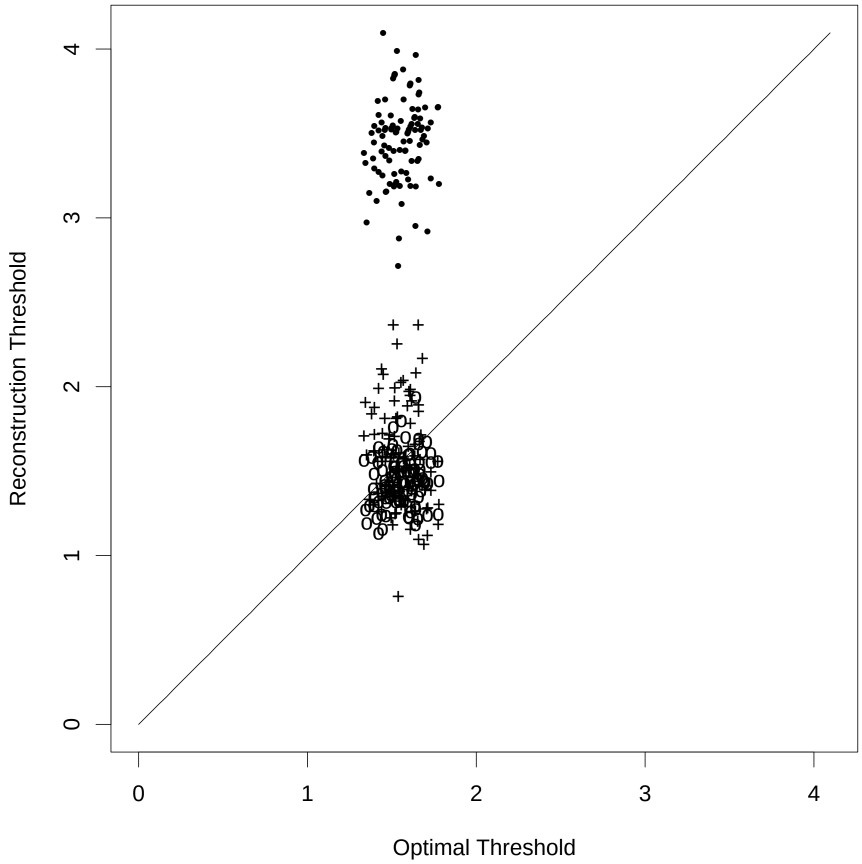


Figure 9: Reconstruction threshold plotted against 'true' optimal threshold for the smooth signal and normal independent noise. Thresholding ALL levels. ($\bullet$ = universal threshold; + = *GlobalSure* threshold; $\circ$ = cross-validated threshold).

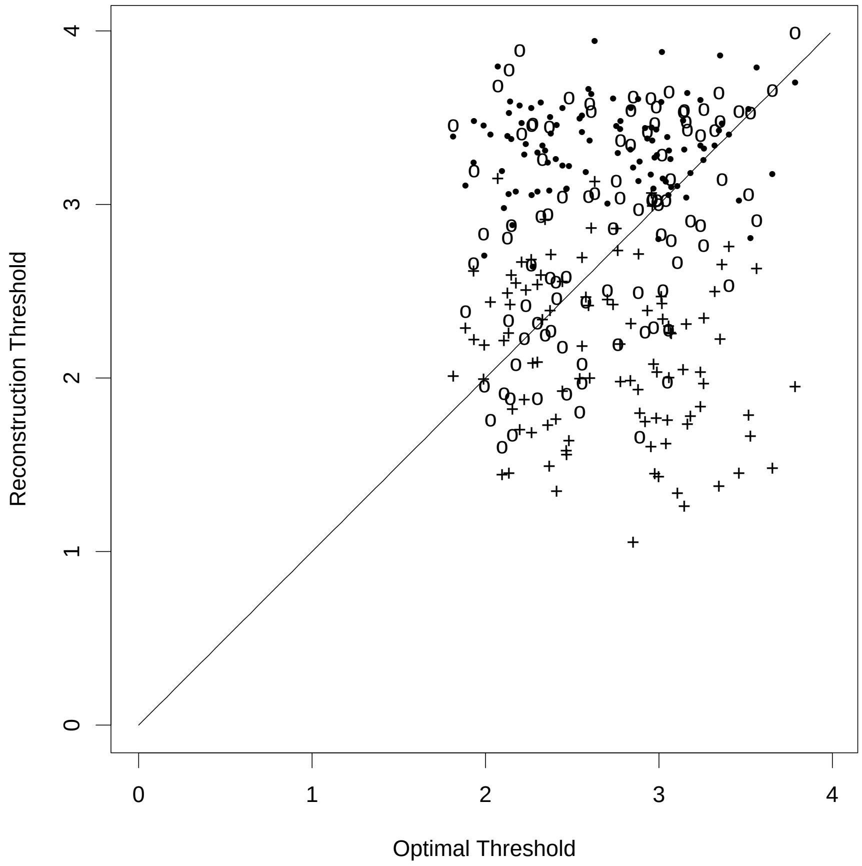


Figure 10: Reconstruction threshold plotted against 'true' optimal threshold for the smooth signal and normal independent noise. Thresholding SUBSET levels. (• = universal threshold; + = *GlobalSure* threshold; ∘ = cross-validated threshold).

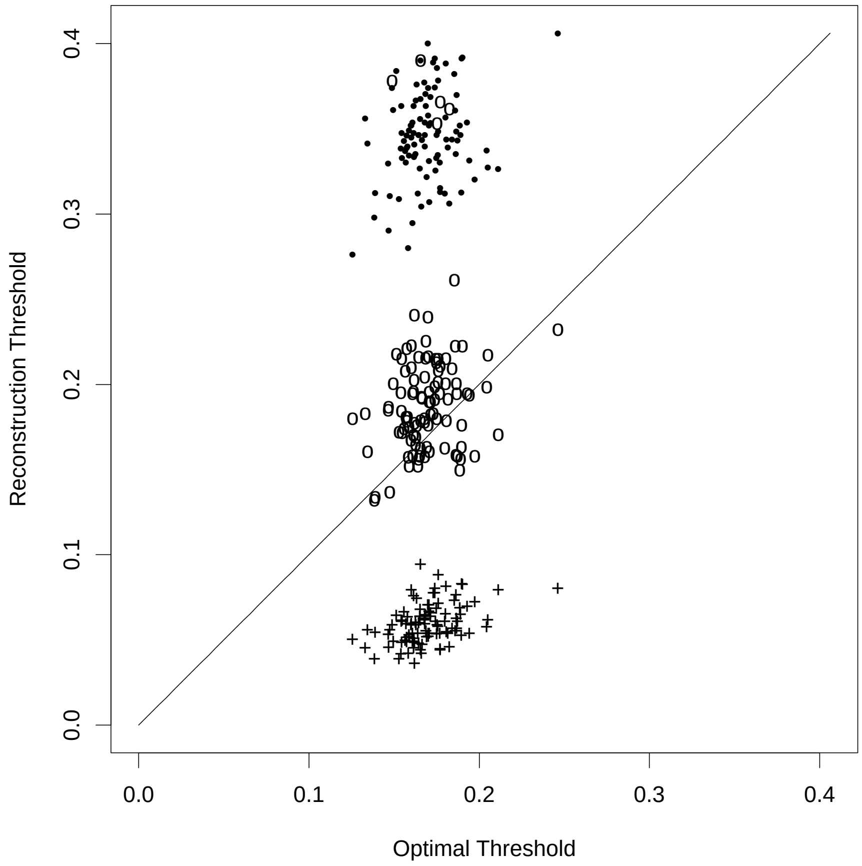


Figure 11: Reconstruction threshold plotted against 'true' optimal threshold for the piecewise polynomial and normal independent noise. Thresholding SUBSET levels. ($\bullet$ = universal threshold; + = *GlobalSure* threshold; $\circ$ = cross-validated threshold).

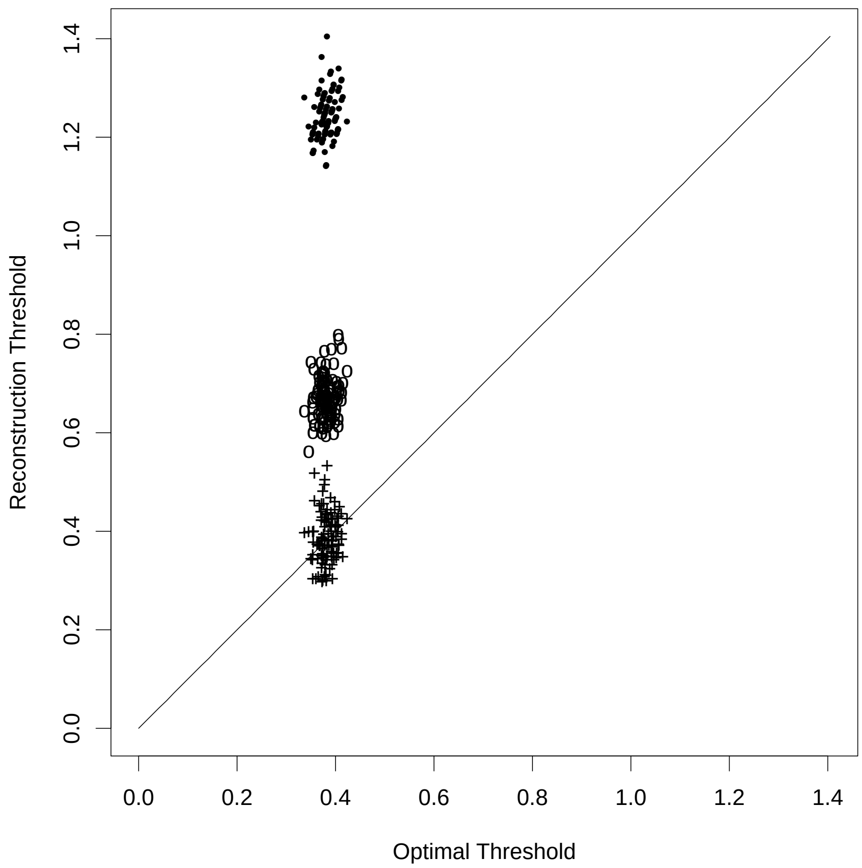


Figure 12: Reconstruction threshold plotted against 'true' optimal threshold for the chirp signal and normal independent noise. Thresholding SUBSET levels. ($\bullet$ = universal threshold; + = *GlobalSure* threshold; $\circ$ = cross-validated threshold).

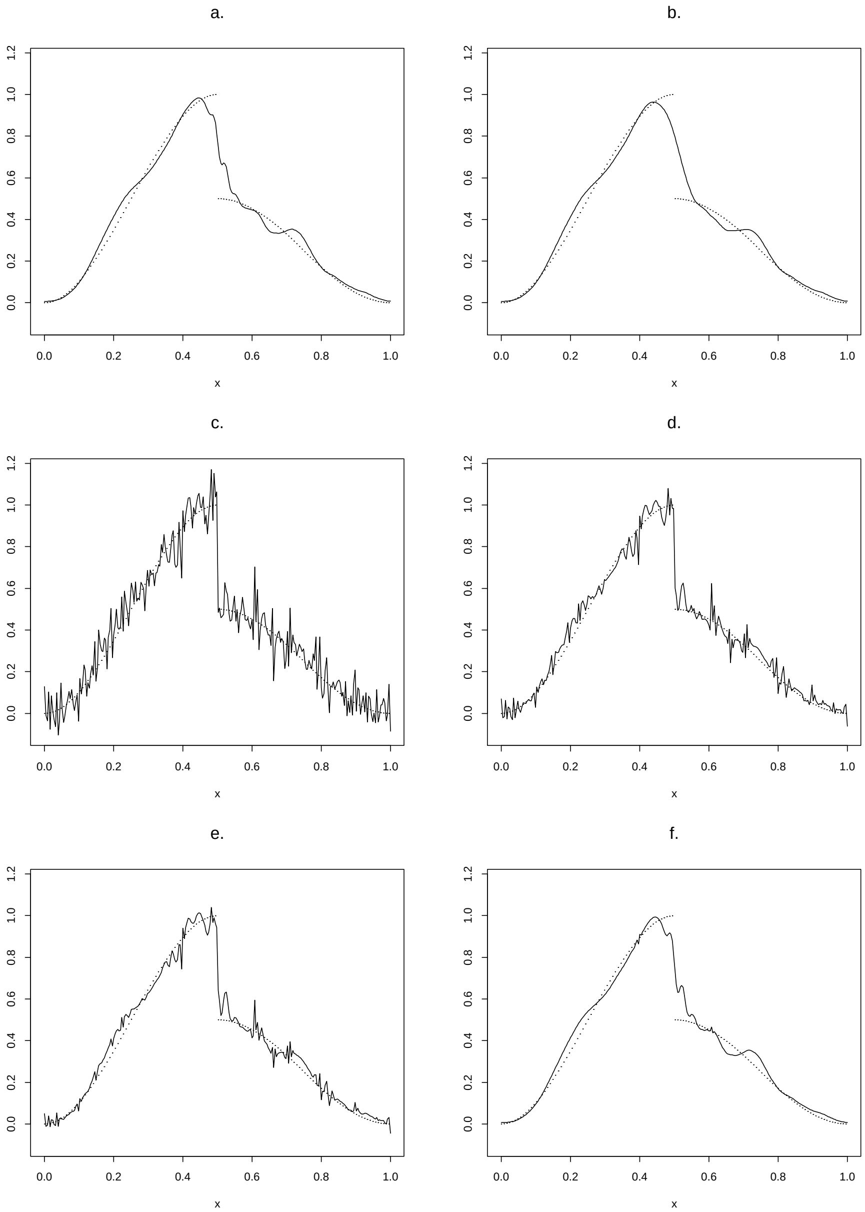


Figure 13: Reconstructions for the piecewise polynomial with normal noise and SUBSET thresholding. A high and low threshold, t, was chosen for each method from Figure 11. (ISE are in parentheses). (a.) Low universal: $t = 0.28$ (0.56); (b.) High universal: $t = 0.4$ (0.73); (c.) Low *GlobalSure*: $t = 0.03$ (1.39); (d.) High *GlobalSure*: $t = 0.1$ (0.51); (e.) Low CV: $t = 0.13$ (0.41); (f.) High CV: $t = 0.24$ (0.53). Figure (e.) has the lowest ISE.

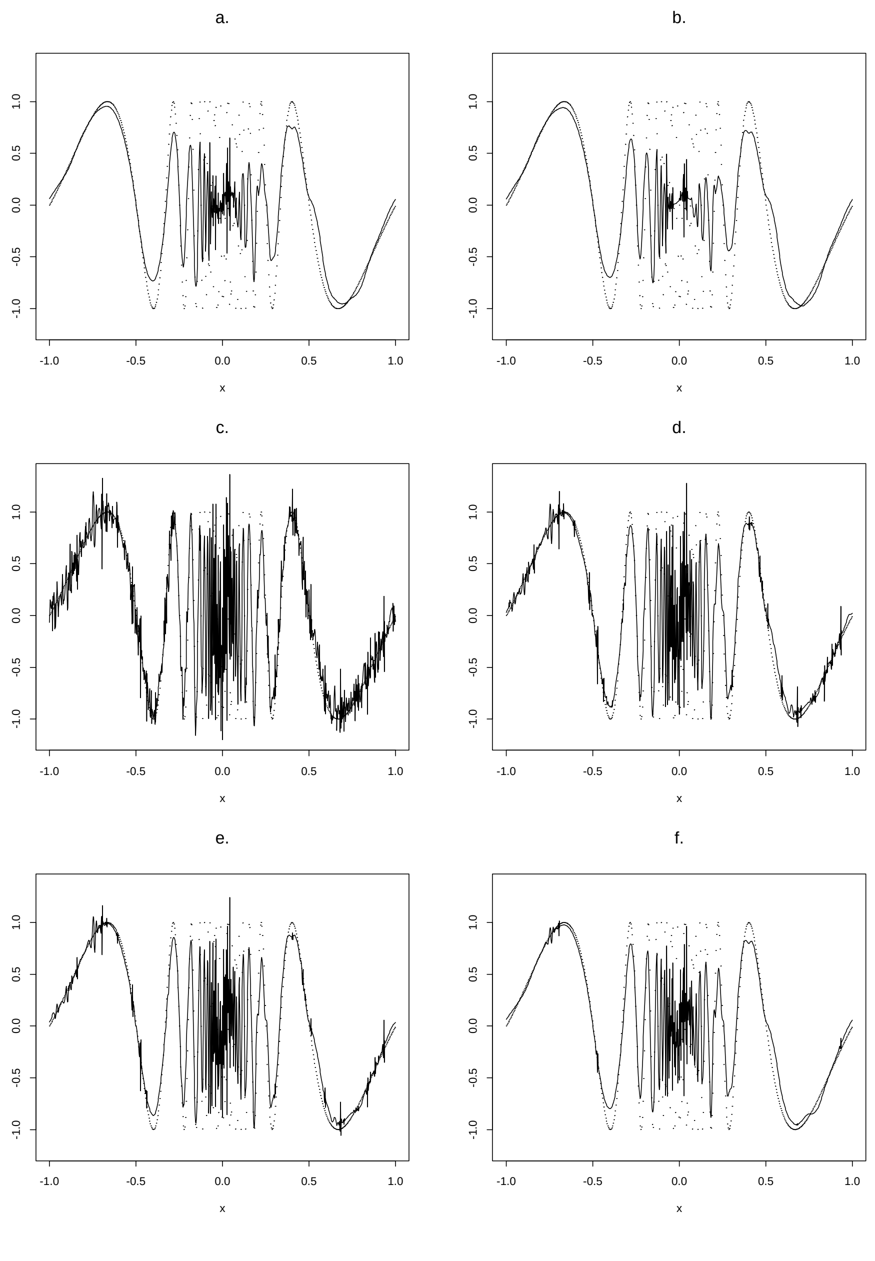


Figure 14: Reconstructions for the chirp signal with normal noise and SUBSET thresholding. A high and low threshold, t, was chosen for each method from Figure 12. (ISE are in parentheses). (a.) Low universal: $t = 1.15$ (78.4); (b.) High universal: $t = 1.4$ (100.0); (c.) Low *GlobalSure*: $t = 0.3$ (27.0); (d.) High *GlobalSure*: $t = 0.5$ (30.2); (e.) Low CV: $t = 0.55$ (33.0); (f.) High CV: $t = 0.8$ (50.8). Figure (c.) has the lowest ISE.

smooth signal all of the true coefficients lie mostly in the the top three levels and so none of the thresholding methods can affect them. Thus changing the threshold does not appreciably change the integrated square error.

5.4 Reconstructions for Student's t independent noise

The reconstruction thresholds versus the 'true' optimal thresholds for Student's t noise and for SUBSET levels thresholding appears in Figure 15. The same plots for the piecewise polynomial appear in Figure 16 and for the chirp signal in 17.

As for the normal noise there is no clear choice for a superior method. Both cross-validation and *GlobalSure* work well. One point to note for both the smooth and piecewise polynomial is that the 'true' thresholds on the x-axis have a large sample variance. This is because the Student's t noise occasionally contains very large outliers. A large threshold is then required to smooth this large outlier back to the original function. It is good to note that the cross-validation function gets close to this threshold 'optimal' threshold (the $x = y$ line). However, a level-by-level method may have been able to limit the damage. Figure 18 shows the reconstructions where the true threshold is large. It is true that the cross-validation method has smoothed the outlier better in that the response at the outlier is smaller when compared to *GlobalSure* and the universal (also the overall integrated square error is smaller than the others). However, because the threshold is large the discontinuity is completely smoothed out. The *GlobalSure* and universal procedures represent the discontinuity better but do not deal effectively with the outlier. Both of these effects are to be expected. Both *GlobalSure* and universal thresholding expect to deal with normal noise and cross-validation is minimizing a global measure of error. This example is rather extreme but does demonstrate the behaviour (or misbehaviour) of the methods rather well.

5.5 Reconstructions for correlated normal noise

The reconstruction thresholds versus the 'true' optimal thresholds for the correlated normal noise and for SUBSET levels thresholding appears in Figure 19. The same plots for the piecewise polynomial appear in Figure 20 and for the chirp signal in 21.

For the correlated noise there are few changes. Probably the most striking being the with the chirp signal the cross-validation method improves and the *GlobalSure* method gets a little worse. For the piecewise polynomial it is debatable which method is best—the answer depends on if you preferred your estimates under- or overfitted (in which case you would choose universal or cross-validated thresholding).

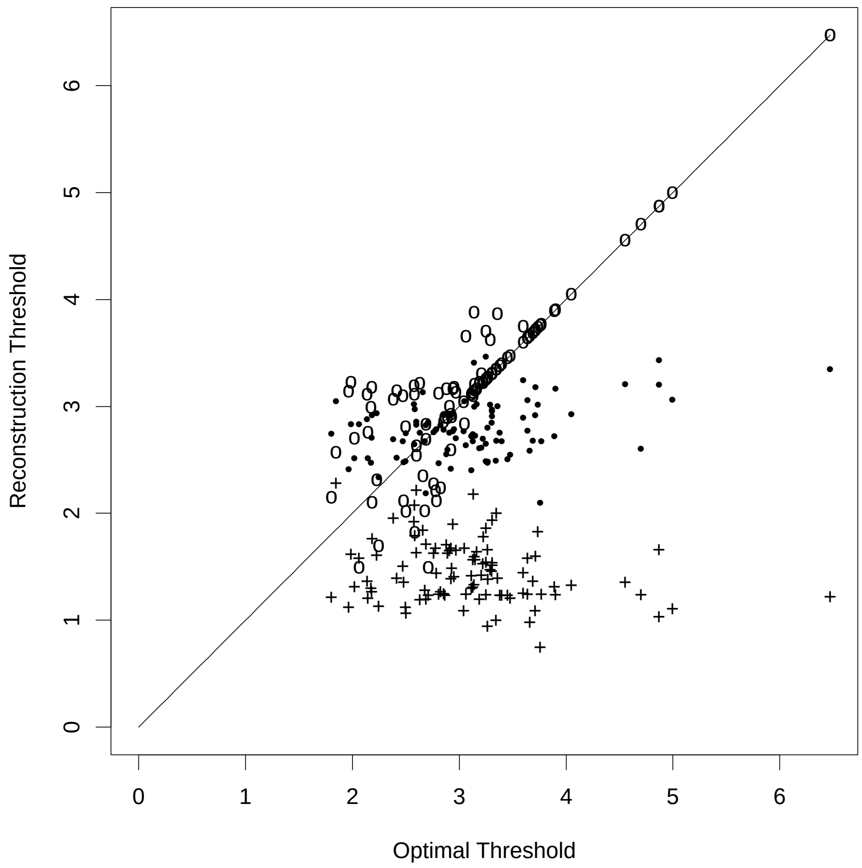


Figure 15: Reconstruction threshold plotted against 'true' optimal threshold for the smooth signal and Student's t independent noise. Thresholding SUBSET levels. ($\bullet$ = universal threshold; + = *GlobalSure* threshold; $\circ$ = cross-validated threshold).

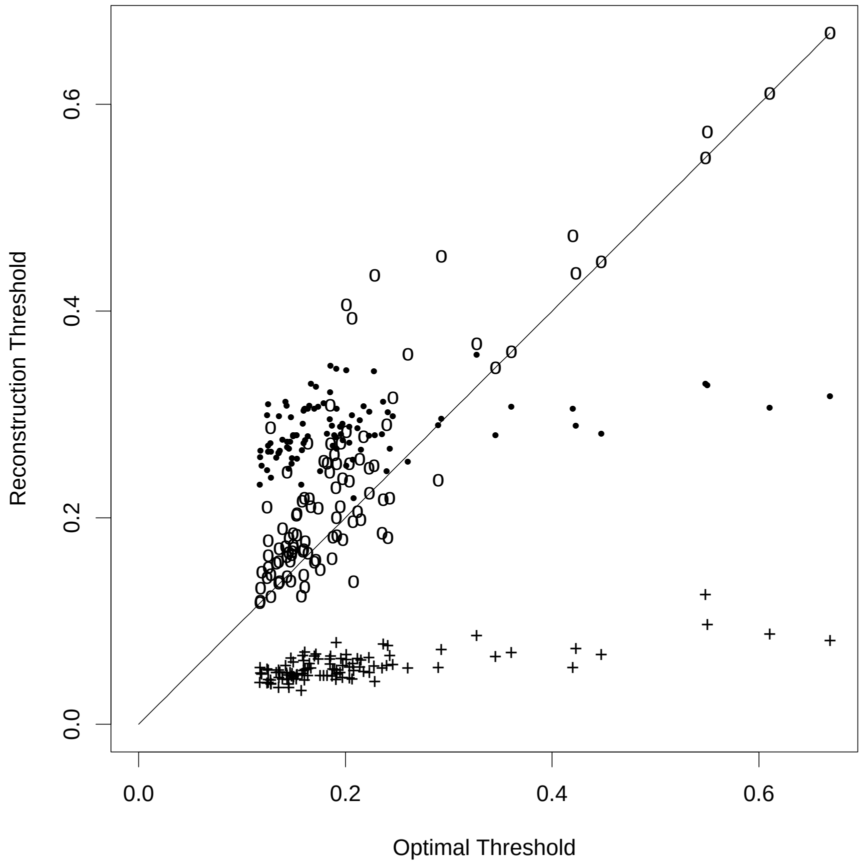


Figure 16: Reconstruction threshold plotted against 'true' optimal threshold for the piecewise polynomial and Student's t independent noise. Thresholding SUBSET levels. ($\bullet$ = universal threshold; + = *GlobalSure* threshold; $\circ$ = cross-validated threshold).

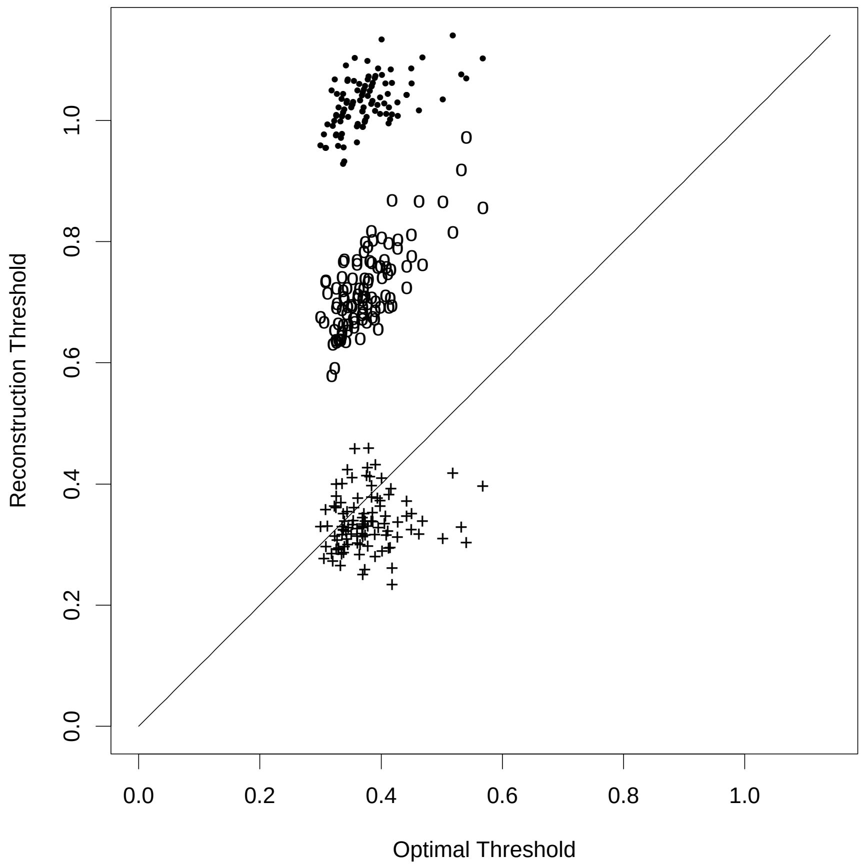


Figure 17: Reconstruction threshold plotted against 'true' optimal threshold for the chirp signal and Student's t independent noise. Thresholding SUBSET levels. ($\bullet$ = universal threshold; + = *GlobalSure* threshold; $\circ$ = cross-validated threshold).

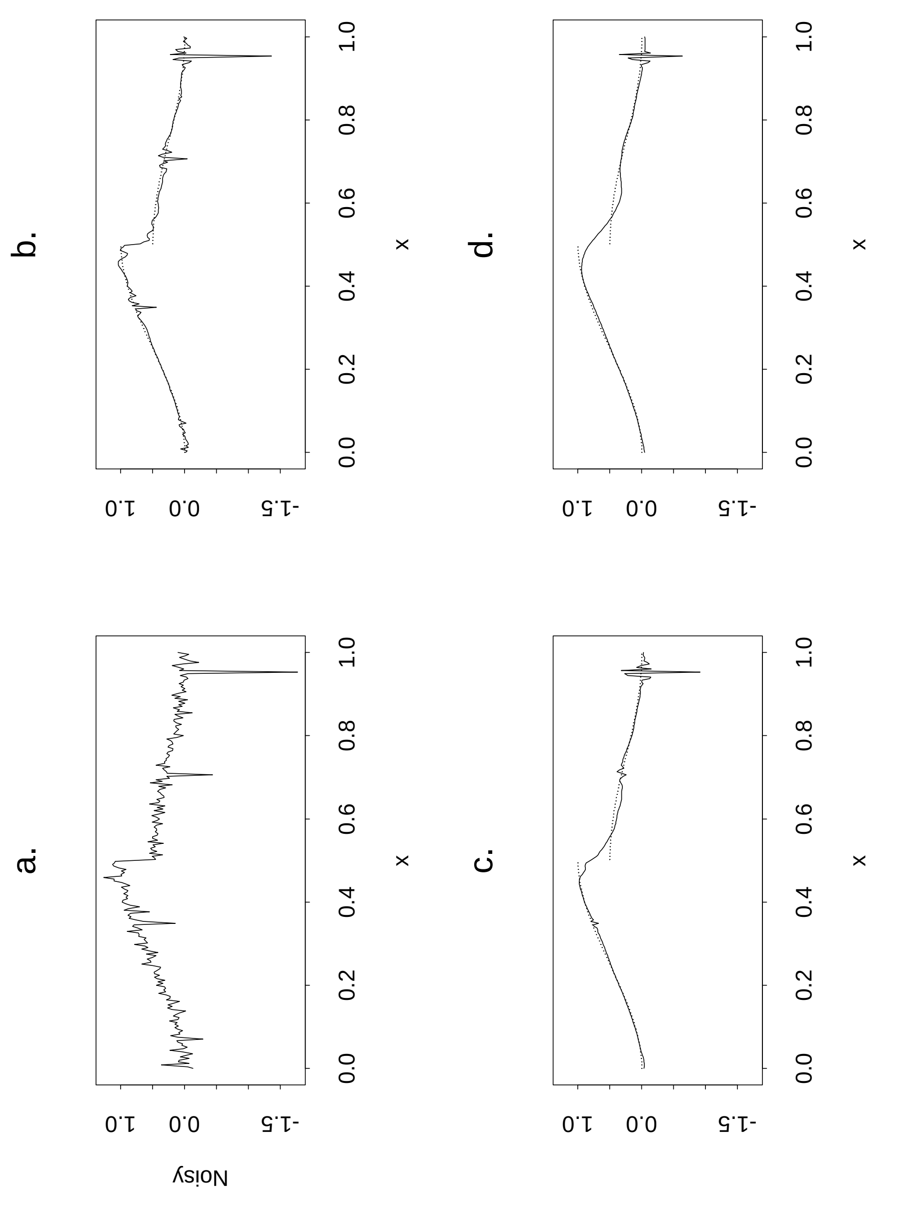


Figure 18: An example of reconstructions from Student's t noisy data with a large outlier. (Integrated square errors in parentheses) (a.) The noisy data; (b.) The *GlobalSure* reconstruction, threshold of 0.15 (2.63); (c.) The universal reconstruction, threshold of 0.35 (1.83); (d.) The cross-validated reconstruction, threshold of 0.58 (1.60).

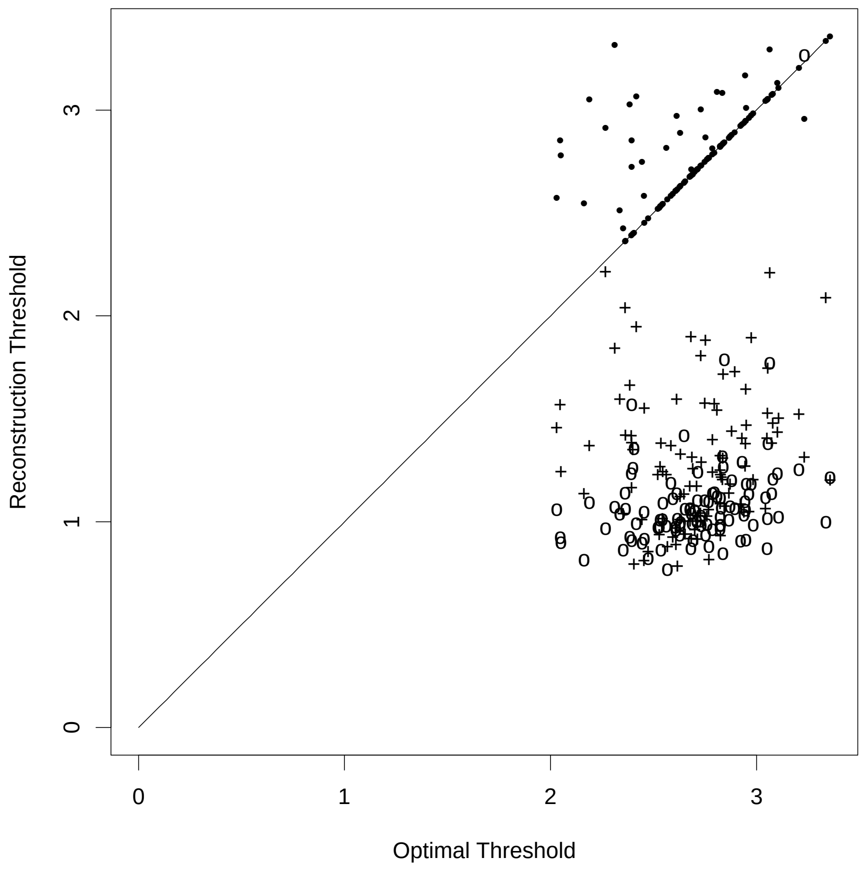


Figure 19: Reconstruction threshold plotted against 'true' optimal threshold for the smooth signal and correlated normal noise. Thresholding SUBSET levels. ($\bullet$ = universal threshold; + = *GlobalSure* threshold; $\circ$ = cross-validated threshold).

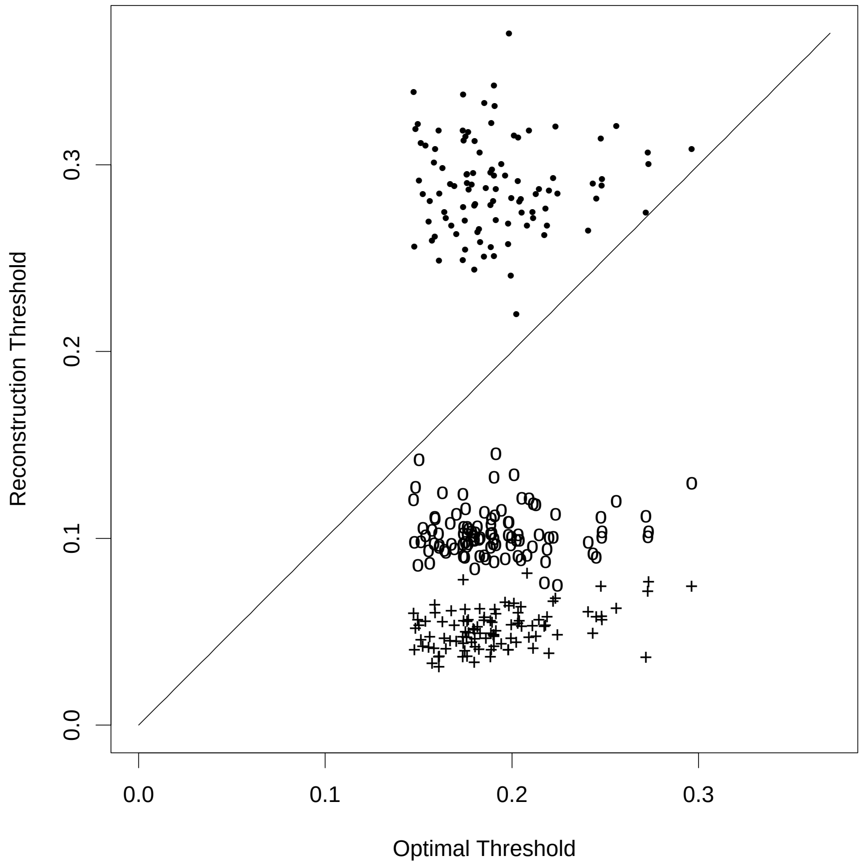


Figure 20: Reconstruction threshold plotted against 'true' optimal threshold for the piecewise polynomial and correlated normal noise. Thresholding SUBSET levels. ($\bullet$ = universal threshold; + = *GlobalSure* threshold; $\circ$ = cross-validated threshold).

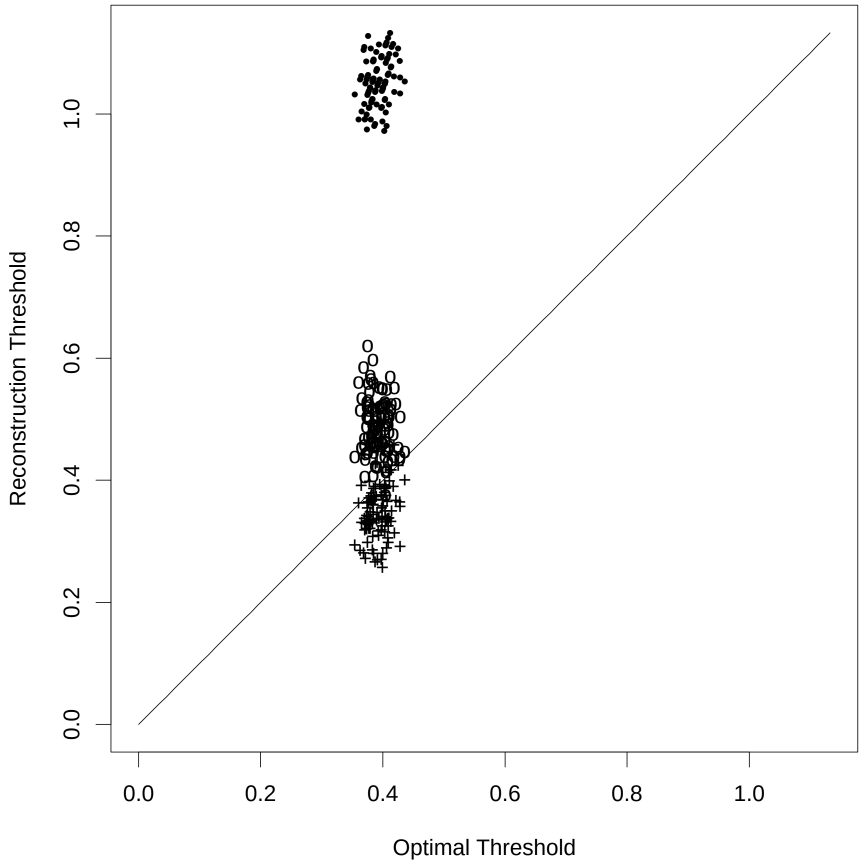


Figure 21: Reconstruction threshold plotted against 'true' optimal threshold for the chirp signal and correlated normal noise. Thresholding SUBSET levels. ($\bullet$ = universal threshold; + = *GlobalSure* threshold; $\circ$ = cross-validated threshold).

5.6 Some general remarks about the simulations

The original wavelet regression method was designed to deal with normal independent data. It is good to see that with the normal data the methods all produce reasonably tight threshold estimates. The cross-validation seems to do well for the sparse piecewise polynomial. The *GlobalSure* procedure does well for the chirp signal for all types of noise.

The reconstruction methods appear to be resilient against a heavy-tailed noise distribution in that for the chirp signal each reconstruction appears to do as well as for normal data and for the piecewise polynomial the cross-validated method tracks the change in the 'true' optimal threshold remarkably well.

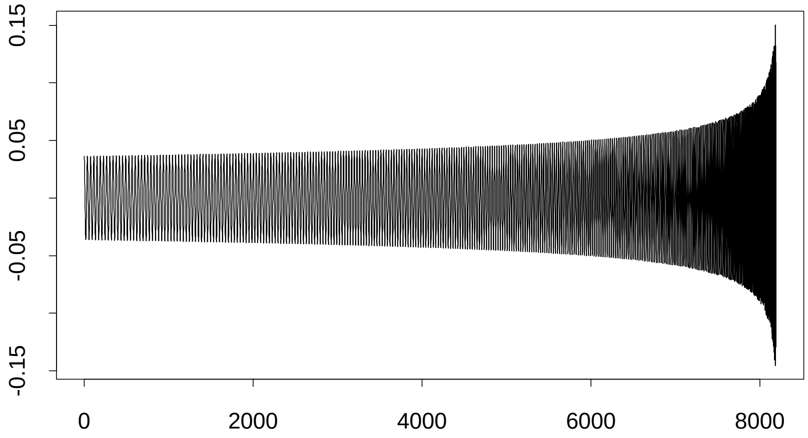
A correlated signal tends to derail each of the methods. Cross-validation does not do so well for the piecewise polynomial and *GlobalSure* not so well for the chirp signal.

It appears that correlation has a more drastic affect on the reconstruction results than divergences from error distribution. Johnstone and Silverman (1994) have investigated a procedure to adopt for correlated data that selects a different threshold for each level of wavelet coefficients. The real *SureShrink* of Donoho and Johnstone (1994) actually does this but the performance is similar to that of universal thresholding for the normal/independent case. The cross-validation method can be adapted to select different thresholds at different levels but the optimization becomes multiparameter with a large computational cost. We feel sure that different thresholds at different levels holds the key to the successful reconstruction of correlated data.

6 An application to chirpy data

One of the main areas for application for wavelets is to chirp type data such as that in Figure 7. The idea for using such data as a test example for wavelet regression arose because of a related signal processing problem from astrophysics. This involves simulated data that arises from a model of a gravity wave chirp and kindly supplied by David Nicholson of Cardiff University. The model of this particular kind of gravity wave is a chirp such as that depicted in Figure 22. The idea is to detect such a signal in the presence of white noise. There are two versions of the noisy signal. Figures 23a and b depict the low-noise version of the noisy chirp and its wavelet transform. It is possible to detect some semblance of the chirp in the wavelet transform when comparing it to Figure 22b. However, it is much more difficult to identify the chirp in the high-noise version in Figures 23c and d. The lack of any chirpyness is especially apparent in the higher levels (8-12) of Figure 23d where the coefficients appear to be just noise whereas in Figure 23b there is some sort of signal.

A procedure called *matched filtering* is one way of detecting a known signal from a noisy version. This involves correlating the noisy signal with a reversed a.



b.

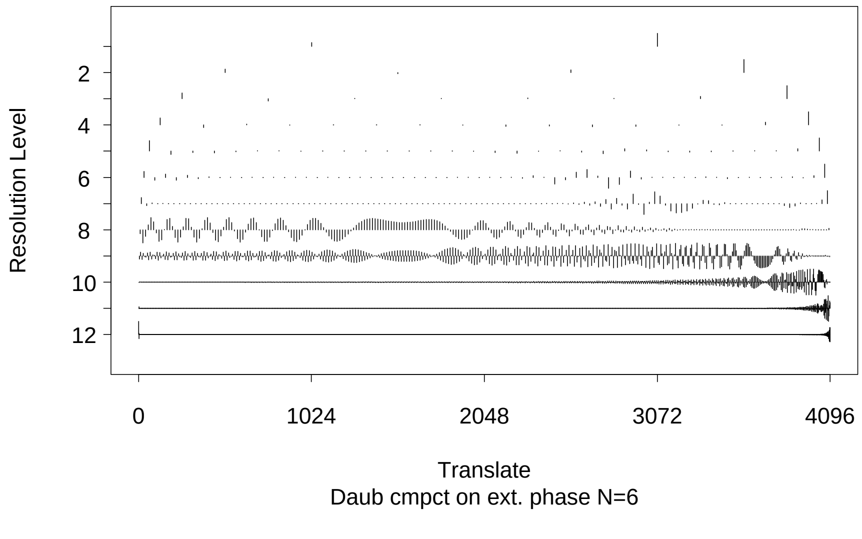


Figure 22: (a.) Simulated gravity wave sampled on 8192 points; (b.) Wavelet transform of a. using Daubechies N=6 wavelet.

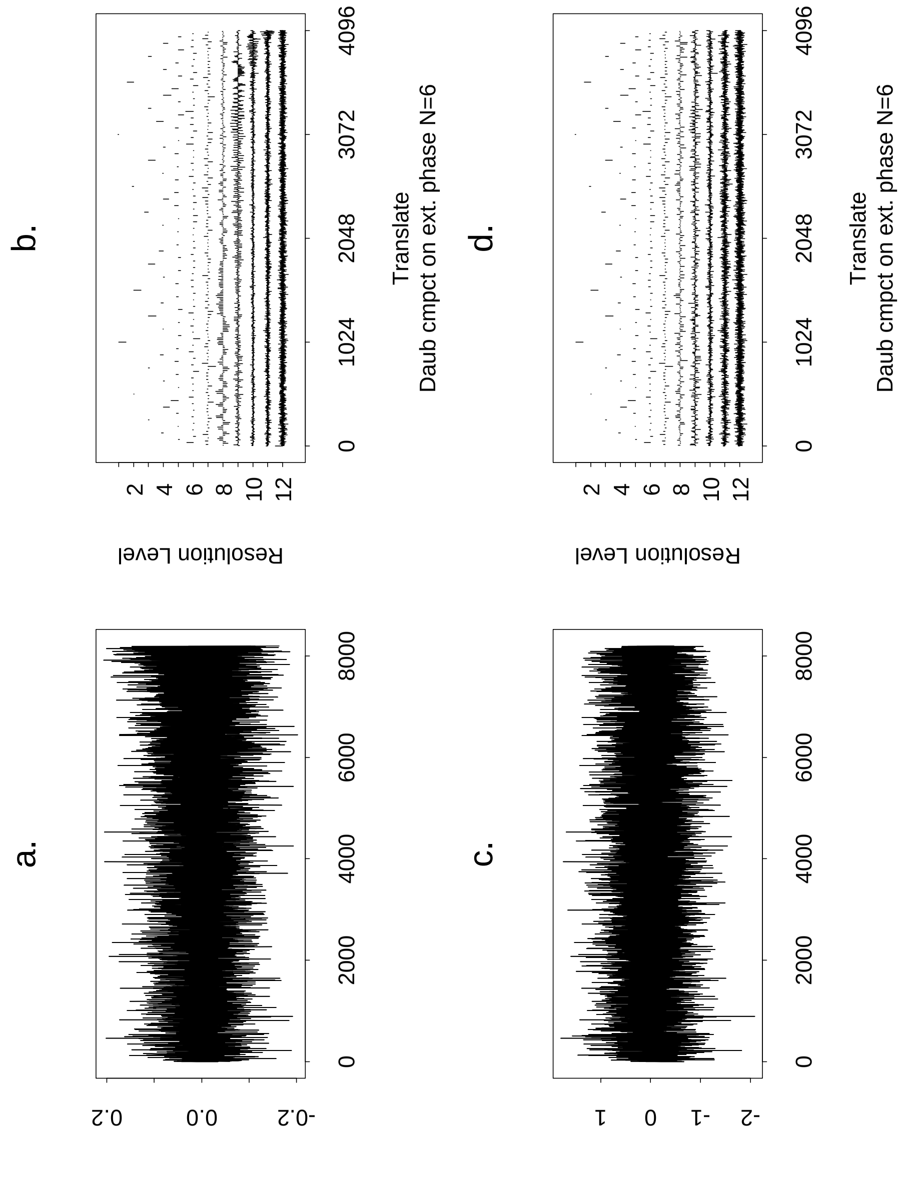


Figure 23: (a.) Low noise gravity wave chirp; (b.) Wavelet decomposition of a; (c.) High noise gravity wave chirp; (d.) Wavelet decomposition of c.

version of the known signal. The output consists of a series of spikes and the maxima of the output can correspond to successful matches. The obvious step is to denoise the signal before it goes into the matched filtering and then hopefully the detection should be improved. However, it is probable that there are methods that are better suited for this data. It is possible that the use of wavelet packets or a stationary wavelet transform would be able to be used in concert with matched filtering and produce better results. This is because these other wavelet-type transforms are even more flexible bases that can be used. Since the type of waveform being searched for is known it may be that one can build a dictionary of chirplets. Chirplets are basis functions that are similar in form to the chirp being searched for (see Mann and Haykin (1992) for example). This is a form of parametric modelling where you know the parametric form of the model behind the data but not necessarily the coefficients. It is likely that this sort of method will be much more successful in a matched filtering process.

Figure 24 shows the reconstructions of the low-noise chirp signal from Figure 23a. The drastic underfitting of the universal thresholding is apparent, but the cross-validated estimate appears very noisy. However, the error for the cross-validated version is smaller.

7 Conclusion

In this paper we introduced a method for computing a threshold for wavelet regression using cross-validation. We have investigated the ISE of a soft-thresholded wavelet estimator and investigated its behaviour as a function of the threshold. For some signals cross-validation is an appropriate means by which a threshold may be selected. For other signals a *SureShrink* based procedure may be more powerful. Further work remains to be done to develop the methods to work level-by-level as this is likely to enhance their performance for correlated data.

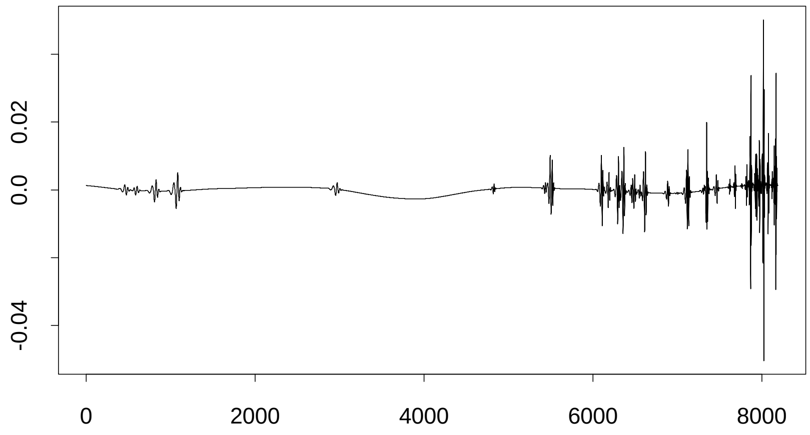
Acknowledgments

I would like to thank Bernard Silverman for helpful conversations and comments on earlier versions of this work. This work was performed with the support of a grant under the Complex Stochastic Systems Initiative of the UK Science and Engineering Research Council. I would also like to acknowledge and thank the Australian Graduate School of Management, The Australian National University and the Commonwealth Scientific and Industrial Research Organisation for support that enabled this work to be completed.

References

Chui, C.K. 1992. *An Introduction to Wavelets.* Boston: Academic Press.

a.



b.

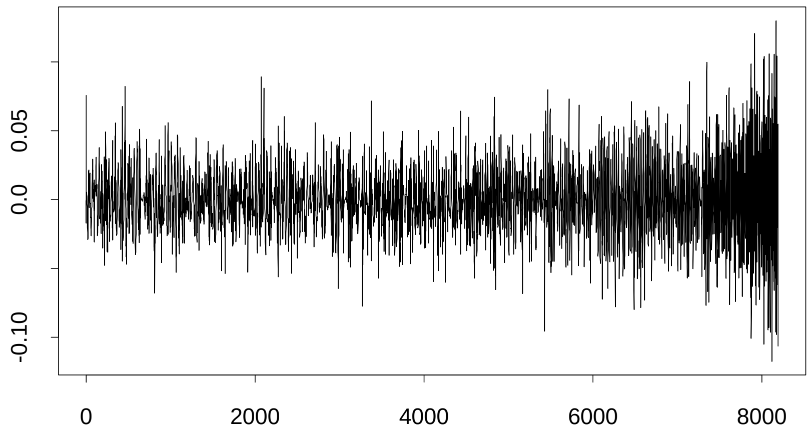


Figure 24: Reconstruction of chirp signal from low-noise using (a.) universally thresholded wavelet regression. Residual sum of squares between estimate and true signal was 9.94; (b.) cross-validatory thresholded wavelet regression. Residual sum of squares between estimate and true signal was 4.91.

Daubechies, I. 1988. Orthonormal bases of compactly supported wavelets. *Comm. on Pure and Applied Mathematics*, **41**, 909–996.

Daubechies, I. 1992. *Ten lectures on Wavelets.* Philadelphia: SIAM.

Donoho, D. L., Johnstone, I.M., Kerkyacharian, G., & Picard, D. 1994. Wavelet shrinkage: asymptopia? *J. R. Statist. Soc. B.* to appear.

Donoho, D.L., & Johnstone, I.M. 1993. Ideal spatial adaptation by wavelet shrinkage. *Biometrika.* (to appear).

Donoho, D.L., & Johnstone, I.M. 1994. Adapting to Unknown Smoothness via Wavelet Shrinkage. *J. Am. Stat. Ass.*, Accepted for publication.

Fan, J., Hall, P., Martin, M., & Patil, P. 1993. *Adaption to high spatial inhomogeneity based on wavelets and on local linear smoothing.* Tech. rept. CMA-SR18-93. Australian National University.

Johnstone, I.M., & Silverman, B.W. 1994. *Personal Communication.*

Johnstone, I.M., Kerkyacharian, G., & Picard, D. 1992. Estimation d'une densité de probabilité par méthode d'ondelettes. *Comptes Rendus Acad. Sciences Paris*, **315**, 211–216.

Mallat, S. G. 1989. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **11**, 674–693.

Mann, S., & Haykin, S. 1992. Adaptive "chirplet" transform: an adaptive generalization of the wavelet transform. *Optical Engineering*, **31**, 1243–1256.

Nason, G P. 1993. *The* `wavethresh` *package; wavelet transform and thresholding software for S.* Available from the StatLib archive.

Nason, G. P., & Silverman, B. W. 1994. The discrete wavelet transform in S. *Journal of Computational and Graphical Statistics.* To appear.

Smith, M.J.T., & Barnwell, T.P. 1986. Exact reconstruction techniques for tree-structured subband coders. *IEEE Transactions on Acoustics, Speech and Signal Processing*, **34**, 434–441.

Strang, G. 1993. Wavelet transforms versus Fourier transforms. *Bulletin (New Series) of the American Mathematical Society*, **28**, 288–305.