Choice of the threshold parameter in wavelet function estimation

G.P. Nason[1]

Department of Mathematics, University of Bristol, Bristol BS8 1TW, UK

Abstract

The procedures of Donoho, Johnstone, Kerkyacharian and Picard [DJKP] estimate functions by inverting thresholded wavelet transform coefficients of the data. The choice of threshold is crucial to the success of the method and is currently subject to an intense research effort. We describe how we have applied the statistical technique of cross-validation to choose a threshold and we present results that indicate that its performance for correlated data. Finally, to illustrate the techniques, we apply various wavelet-based estimation methods to some noisy one- and two-dimensional signals and display the results.

1 Introduction

This paper reviews various methods for selecting a threshold for wavelet function estimation. We concentrate our study on threshold selection using cross-validation.

Section 2 reviews the wavelet transforms that we use for the rest of the paper. Section 3 describes various methods of function estimation from noisy data using wavelet methods. Section 4 reviews cross-validation in this context. Section 5 provides an example where the two-fold cross-validation method performs badly because of the correlation structure within the data set (Nason [Na2, Na3] gives examples where the method works well!) Section 6 discusses extensions of the cross-validation method to more dimensions. Finally, Section 7 gives brief descriptions of work by Wang [Wa1] and Weyrich and Warhola [WW1] that improve on and extend the cross-validation methods described in this paper.

2 Wavelet overview

Wavelets are functions that are used to represent functions. In all that follows we will be interested solely in families of wavelets that act as orthonormal bases for various function spaces. For simplicity we restrict ourselves to the $L^2(\mathcal{R})$ function space in other words $f \in L^2(\mathcal{R})$ iff

$$\int_{-\infty}^{\infty} |f(x)|^2 \, dx < \infty.$$

It can be shown (Daubechies [Da2], Meyer [Me1]) that it is possible to construct a function $\psi(x)$, so that if $f \in L^2(\mathcal{R})$ then

$$f(x) = \sum_{k \in \mathcal{Z}} c_k \phi_{0k}(x) + \sum_{j<J, k \in \mathcal{Z}} d_{jk} \psi_{jk}(x), \tag{1}$$

where

$$c_k = \int_{\mathcal{R}} f(x) \phi_{0k}(x)\, dx,$$

and

$$d_{jk} = \int_{\mathcal{R}} f(x) \psi_{jk}(x)\, dx,$$

where J controls the maximum resolution. The functions $\psi_{jk}(x)$ are all derived from a single *mother* wavelet ψ by the relation

$$\psi_{jk}(x) = 2^{\frac{j}{2}} \psi(2^j x - k). \tag{2}$$

The derived functions $\psi_{jk}(x)$ are called wavelets and the mother wavelet is specially chosen so that the family $\{\psi_{jk}(x)\}$ forms an orthonormal basis for $L^2(\mathcal{R})$. The functions $\phi_{0k}(x)$ are all derived from a function $\phi(x)$, known as the *father* wavelet or *scaling* function, by using the dilation and translation formula given in (2). Typically *wavelets of class* m are specially constructed so that (Meyer [Me1]):

(orthonormal basis) the set $\{\psi_{jk}(x)\}$ forms an orthonormal basis for the space under consideration;

(regularity) if $m = 0$, $\psi(x)$ belongs to $L^\infty(\mathcal{R})$; if $m > 1$, $\psi(x)$ and all its derivatives up to order m belong to $L^\infty(\mathcal{R})$;

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

(oscillation) $\int_{\mathcal{R}} x^k \psi(x)\, dx = 0$ for $0 \le k \le m$.

The localization property mentioned above extends also to the frequency domain. Typically wavelets are well-localized in time and frequency. To achieve this wavelets are usually compactly supported in either the time or frequency domain (but not both so that the uncertainty principle is obeyed) and decay rapidly in both. The wavelets that we use in this paper are those created by Daubechies [Da1] who carefully constructed a series of mother wavelets (indexed by N) with each mother in the series having regularity proportional to N. Each of Daubechies' wavelets are compactly supported in the time domain. The expansion given in (1) has a discrete alternative. Given a data set $f_1, \ldots, f_n$ where $n = 2^J$ there exists an orthogonal matrix W such that the *discrete wavelet transform* θ_{jk} is given by

$$\theta = Wf,$$

where θ is the n-vector of discrete wavelet coefficients $\theta_{jk}, j = 0, \ldots, J-1$ and $k = 1, \ldots, 2^j$. In practice, the discrete wavelet transform is performed using an efficient algorithm that only requires $O(n)$ operations (see Mallat [Ma1]). The fast algorithm appeared earlier in the engineering literature as a two-channel subband coder (see Smith and Barnwell [SB1] for example). The inverse discrete wavelet transform is also easy to compute. The inverse transform may be represented by

$$f = W^T \theta.$$

There is a corresponding fast algorithm for this as well. Using the inverse transform formula it is possible to note that the rows of W correspond to discretized versions of the mother wavelets at various different scalings and translations. Donoho and Johnstone[DJ2] note the approximation:

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

$W_{j,k}(i)$ is the ith element of the (j,k)th row of W. It can be seen that the wavelet coefficient θ_{jk} quantifies the contribution of the basis functions $W_{j,k}$ which are localized to a spatial interval of size 2^{-j} and near frequency 2^j. In simpler terms θ_{jk} indicates the amount of signal around spatial location $2^{-j}k$ and near frequency 2^j.

We use the fast algorithm as implemented in SPlus by Nason [Na1]. Use of the SPlus package allows easy access to the comprehensive statistical facilities of S within an object-oriented environment.

3 Function estimation from noisy data

There are several good reasons why wavelets can be used for estimating functions. The main reasons are that wavelet shrinkage estimators are:

- nearly minimax for a wide range of loss functions and for general classes of functions;
- simple, practical and fast;
- adaptable to spatial and frequency inhomogeneities;
- readily extendable to high dimensions;
- applicable to various other problems such as density estimation and inverse problems.

A thorough review of these reasons and justification for them appears in Donoho *et al.* [DJKP].

Estimation by wavelet shrinkage is a simple procedure that computes the wavelet transform of data, modifies the transform coefficients and then inverts the modified coefficients to form the estimate.

3.1 Estimation and the discrete wavelet transform

Given observed data $g_1, \ldots g_n$ assume the model

$$g_i = f(t_i) + \epsilon_i, \tag{3}$$

where the $\{\epsilon_i\}$ is some noise process with variance σ^2, $t_i = i/n$ and f is the function that is to be estimated. Then

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

performs the wavelet transform on the noisy data. The wavelet coefficients are then modified by some procedure to form w^* and then the inverse transform is performed to obtain:

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

where $\hat{f}$ is the estimate of f at the points $\{t_i\}$. Donoho and Johnstone [DJ2] used wavelets designed for use on an interval devised by Cohen *et al.* [CDJV]. This paper uses Daubechies' [Da1] wavelets with periodic boundary correction. Further details of this particular transform can be found in Nason and Silverman [NS1].

The key question in wavelet function estimation is how should the wavelet coefficients, w, be modified to form w^*? Donoho *et al.* [DJKP] advise that shrinking wavelet coefficients produces estimates that possess the desirable properties in the list given above. To achieve shrinkage they propose thresholding the coefficients. Given a wavelet coefficient w and a 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. This paper considers soft-thresholding although hard-thresholding is a possible alternative. At the moment there is not much theory about which method is better (however, in cross-validation the smoothness of the soft-thresholder aids the optimization procedure). The question of how coefficients should be modified then reduces to: what should the threshold be? The choice is critical: if the threshold is too small/large then wavelet shrinkage estimators tend to over/underfit the data. The next section reviews some of the methods that have been suggested to choose the threshold value.

3.2 Threshold choosers

The following list describes some of the different methods that have been proposed to choose the threshold. It does not attempt to compare the methods *nor is it a complete list.*

(exact minimax): A policy that uses precomputed thresholds to minimize a constant term in the upper bound for the minimax risk of estimating a function using a thresholded estimator. See Donoho and Johnstone [DJ2].

(universal): Donoho and Johnstone [DJ2] proposed the *universal* threshold that is incorporated into their *VisuShrink* procedure. The universal threshold is

$$T_{\text{UV}} = \sqrt{(2 \log n)}\hat{\sigma}, \tag{6}$$

where n is the number of data points and $\hat{\sigma}$ is an estimate of the noise level σ (typically a scaled median absolute deviation of the empirical wavelet coefficients). One feature of *VisuShrink* is that it "guarantees" a noise-free reconstruction although by doing so it usually underfits the data (the underfitting was also noticed by Fan *et al.* [FHMP]).

(SURE): A threshold chooser based on Stein's [St1] unbiased risk estimation was proposed by Donoho and Johnstone [DJ1] and called *SureShrink*. The *SureShrink* chooser specifies a threshold value t_j for each resolution level j in a wavelet transform.

(cross-validation): The majority of the remainder of this paper concerns cross-validation. It is a technique that relies on attempting to minimize the prediction error generated by comparing a prediction based on a subset of the data and comparing it to the remainder of the data (see later in this paper or Nason [Na2, Na3], Neumann and Spokoiny [NS2], Weyrich and Warhola [WW1], Wang [Wa1] and Donoho and Johnstone [DJ1]).

(false-discovery rate): One can view the estimation of the true function's wavelet coefficients as a multiple hypothesis testing problem. Abramovich and Benjamini [AB1] have adapted Benjamini and Hochberg's[BH1] false discovery rate method for use with wavelets. Instead of choosing a threshold they keep (discard) a wavelet coefficient in the decomposition of the noisy data if a hypothesis test decides that the coefficient is non-zero (zero). A more sophisticated statistical argument is required for multiple hypothesis testing compared to when only one hypothesis test is involved. For example, suppose you wish to test the single hypothesis

$$H_0 : \theta = 0$$

versus

$$H_{\mathrm{A}} : \theta \neq 0$$

and test at the usual significance level of $P(\text{Reject } H_0 | H_0 \text{ is true}) = 0.05$. If, say, 1023 coefficients were to be independently tested at a significance level of 0.05 then approximately $1023 \times 0.05 \approx 50$ coefficients would not be zeroed (and for some functions this would be far too many). The false discovery rate idea can be formulated as follows. Abramovich and Benjamini [AB1] consider a situation where there are n hypotheses (coefficients) to be tested. Each hypothesis is of the form $H_{jk} : \theta_{jk} = 0$. Of these hypotheses n_1 are false, or equivalently the corresponding coefficients should be included in the reconstruction. The other $n_0 = n - n_1$ coefficients are indeed zero and ideally all the noisy versions should be set to zero. They define R to be the number of coefficients that are not zeroed by a given thresholding procedure (and thus will be included in any reconstruction). Of these R coefficients S are correctly kept in the model and V are kept in by mistake. Clearly $R = S + V$. The error is written as $Q = V/R$ and is the proportion of included coefficients that should have been zeroed. The False Discovery Rate of Coefficients (FDRC) is defined to be the expectation of Q. The procedure Abramovich and Benjamini adopt is to include as many coefficients as possible whilst controlling the FDRC to be at or below some level q. They provide an efficient algorithm to compute which coefficients to include and which to discard.

(Bayesian methods): Vidakovic [Vi1] adopts a Bayesian approach to the estimation of

$$\theta = Wf$$

and estimates the means θ_{jk} in $w_{jk} \sim N(\theta_{jk}, \sigma^2)$. Vidakovic obtains prior distributions for σ^2 and θ (that also involve hyperparameters) and with the assumed conditional distribution just mentioned have

$$w|\theta \sim f(w|\theta).$$

After observing w (the noisy wavelet coefficient) he tests the hypothesis $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$. If this hypothesis is "reject" than θ is estimated by w. He then develops some Bayes rules in the testing context that appear similar to the usual thresholding functions (and called *Bayes factor* thresholding). In particular, he shows that if

$$w|\theta \sim \mathcal{DE}\{\theta, (2\mu)^{-1/2}\},$$

where μ is a hyperparameter and $\mathcal{DE}$ is the double exponential and

$$\theta \sim \pi(\theta) = \pi_0\delta_0 + \pi_1\zeta(\theta)$$

is the prior distribution of θ (so π_0 is the probability that $\theta = 0$ and $\zeta(\theta)$ describes the spread of θ when it is not zero). Then w will be "thresholded" if

$$\frac{\pi_0 e^{-c|w|}}{\pi_0 e^{-c|w|} + \pi_1 \{\Pi_1(c) + \Pi_2(c)\}} \geq \frac{1}{2},$$

where $c = \sqrt{2\mu}$ and Π_1 and Π_2 are the Laplace transforms of $\zeta(\theta - w)$ and $\zeta(\theta + d)$.

(Ogden's methods): Amongst other ideas Ogden [Og1] develops two methods for thresholding a wavelet decomposition of noisy data which he calls *selection thresholding* and *data-analytic thresholding*. Selection thresholding is based on hypothesis testing of coefficients level-by-level. Given a set of coefficients at a particular level Ogden describes a test statistic that if large will prompt the user to include the largest (in absolute terms) coefficient into the reconstruction decomposition and then continue testing the remainder of the coefficients. If the test statistic is not large enough (when compared to some critical value) then the threshold is set to be the absolute value of the largest remaining coefficient. Data-analytic thresholding is based on looking at plots of cumulative sums of the squares of the coefficients at a particular level. Coefficients are removed from the level (and marked for inclusion in the reconstruction) if some test based on Brownian bridge sampling is significant and then the remaining coefficients are tested. The test tries to ascertain if the remaining coefficients are just white noise, by successively removing the larger coefficients until the test decides that the coefficients are indistinguishable from white noise. One important advantage of the data-analytic thresholding is that it does not separate coefficients that are close in time. For example, discontinuities can often cause two adjacent coefficients to be large (rather than just one) and this method identify these together, which would not necessarily be possible with other methods that separate coefficients and sort according to size.

4 Cross-validation methods

The aim of function estimation in this article is the minimization of the mean integrated square error (MISE) between the thresholded wavelet estimator $\hat{f}_t(x)$ and the true function $f(x)$. In symbols the threshold t should minimize

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

In practice the function f is not known and so an estimate of M has to be devised. It is often desirable that a loss function other than MISE be used and this can be easily achieved by replacing MISE by the appropriate loss in the estimate of M.

Cross-validation is widely used as an automatic procedure to choose a smoothing parameter in many statistical settings. The following sections describe how the cross-validation paradigm can be used for choosing the threshold for a wavelet shrinkage estimator.

The classic cross-validation method is performed by systematically expelling a data point from the construction of an estimate, predicting what the removed value would have been and comparing the prediction to the value of the expelled point. This simple leave-one-out procedure cannot be directly applied to wavelet shrinkage estimation because the discrete wavelet transform using Mallat's fast algorithm only operates on data sets that contain a power of 2 number of elements.

In artificial wavelet function estimation problems a data set of length 2^M is supplied and leave-one-out would imply performing the fast wavelet transform with $2^M - 1$ points which is not possible because $2^M - 1$ is not a power of two. In practice one is unlikely to receive a real data set with 2^M points and there are various ways around the power of two limitation:

1. truncate or extend the series in some way and pretend that you have 2^M points. Section 4.1 demonstrates a cross-validation method for handling these extended data sets.

2. devise some method of using wavelet shrinkage estimators for any number of points. The method is described in Nason [Na3] but be warned the method is computationally intensive.

3. use some other transform than the pyramidal algorithm. For example, use Kwong and Tang's [KT] W-matrices or Taswell and McGill's [TM1] algorithms for data sets of arbitrary length.

In principle a wavelet method should be able to cope with zero padding better than non-local schemes because most of the wavelet basis functions of the noisy signal are completely disconnected from those of the padding.

4.1 Two-fold cross-validation

This section describes a cross-validation procedure that can be used to automatically select a threshold for a wavelet shrinkage estimator based on 2^M points.

The procedure works by leaving out half of the data points. This leaves 2^{M-1} data points (a power of two) that are then used to form a wavelet shrinkage estimator using a particular threshold. The values of the expelled points can then be compared with the thresholded estimator to form an estimate of prediction error at a particular threshold. This quantity can be then numerically minimized over values of the threshold.

Two-fold cross-validation algorithm Given data $g_1, \ldots, g_n$ where $n = 2^M$. Remove all the odd-indexed g_i from the set. This leaves 2^{M-1} evenly indexed g_i which are re-indexed from $j = 1, \ldots, 2^{M-1}$. A function estimate $\hat{f}_t^{\mathrm{E}}$ is then constructed using a particular threshold t from the re-indexed g_j. To compare the function estimator with the left-out noisy data an interpolated version of $\hat{f}_t^{\mathrm{E}}$ is formed:

$$\bar{f}_{t,j}^{\mathrm{E}} = \begin{cases} \frac{1}{2}\left(\hat{f}_{t,j+1}^{\mathrm{E}} + \hat{f}_{t,j}^{\mathrm{E}}\right) & j = 1, \ldots, \frac{n}{2} - 1 \\ \frac{1}{2}\left(\hat{f}_{t,1}^{\mathrm{E}} + \hat{f}_{t,n-1}^{\mathrm{E}}\right) & j = \frac{n}{2}. \end{cases} \tag{8}$$

The estimate $\bar{f}_{t,\frac{n}{2}}^{\mathrm{E}}$ is formed from the first and last $\hat{f}_t$ values because f is assumed to be periodic. The $\hat{f}_t^{\mathrm{O}}$ is computed for the odd indexed points and the interpolant $\bar{f}_t^{\mathrm{O}}$ computed as above. The full estimate for $M(t)$ compares the interpolated wavelet estimators and the left-out points:

$$\hat{M}(t) = \sum_{j=1}^{\frac{n}{2}} \left\{ \left(\bar{f}_{t,j}^{\mathrm{E}} - g_{2j+1}\right)^2 + \left(\bar{f}_{t,j}^{\mathrm{O}} - g_{2j}\right)^2 \right\}. \tag{9}$$

Note that the estimate $\hat{M}$ relies on two estimates of $\hat{f}_t$ based upon $n/2$ data points. From the work of Donoho and Johnstone [DJ2] it is known that the appropriate threshold depends on n and asymptotically behaves like $T_{\mathrm{UV}}(n) = \sqrt{(2\log n)}\hat{\sigma}_n$. This quantity supplies a heuristic method for obtaining a cross-validated threshold for n data points. If the threshold for n points is $T_{\mathrm{UV}}(n)$ then the threshold for $n/2$ points will be $T_{\mathrm{UV}}(n/2)$ and therefore

$$T_{\mathrm{UV}}(n) \approx \left(1 - \frac{\log 2}{\log n}\right)^{-1/2} T_{\mathrm{UV}}(n/2). \tag{10}$$

After the estimate $\hat{M}(t)$ has been minimized the correction (10) is applied to obtain the final cross-validated threshold.

4.1.1 A note about two-fold cross-validation Burman [Bu1] provides a comparative study of ordinary cross-validation, ν-fold cross-validation and repeated learning-testing methods. The two-fold cross-validation algorithm above is a special case of ν-fold cross-validation where $\nu = 2$. It is not quite 2-fold validation as it is an example of *uncontrollable* cross-validation in the sense of Stone [St1] because ideally 2-fold validation requires a random split into two groups. The 2-fold wavelet algorithm above is forced to split the data into equal halves because of the 2^M restriction imposed by the fast discrete wavelet transform.

The correction term introduced above for the two-fold algorithm is rather heuristic. Burman [Bu1] introduces proper terms necessary to correct for bias caused by performing ν-fold cross-validation rather than ordinary cross-validation. It would be desirable to

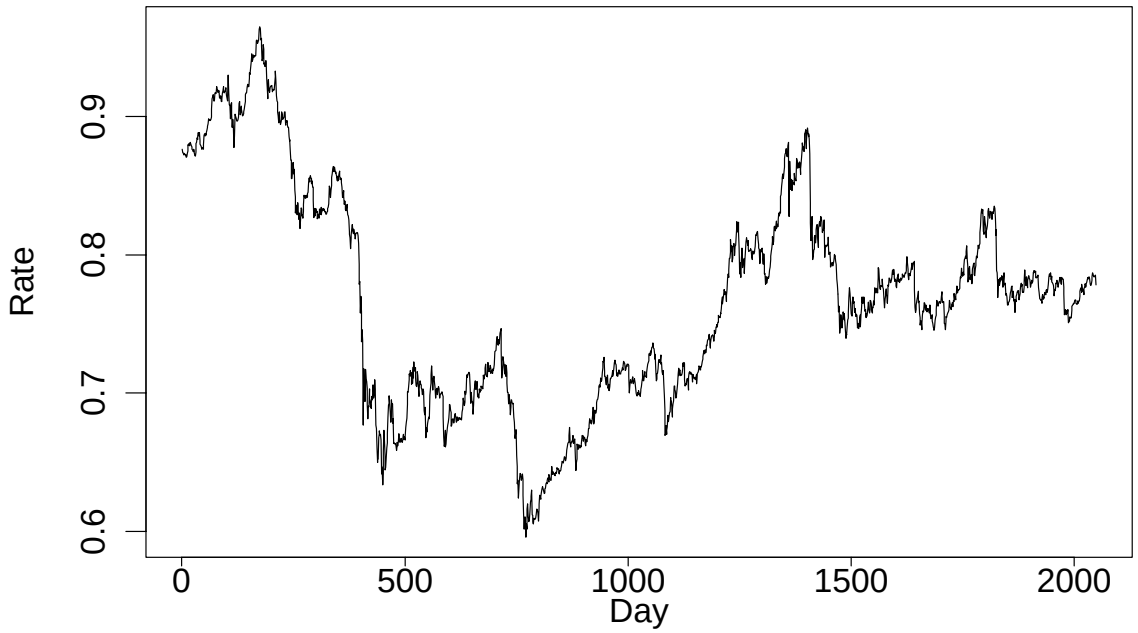

Figure 1: Plot of Australian versus US dollar exchange rate data.

repeat Burman's analyses for the wavelet based 2-fold cross-validation. However, such analyses require a von Mises expansion of

$$T(z, F_n) = \left\{ f(x) - \hat{f}_t(x) \right\}^2$$

about $T(z, F)$ where $z = \{x, f(x)\}$. For the expansion to be possible the functional $T(z, \cdot)$ has to be von Mises differentiable (Serfling [Se1]). It is not obvious whether this is the case for soft-thresholded estimators and seems unlikely for hard-thresholded ones.

5 An example

Nason [Na2, Na3] gives examples where the simple two-fold cross-validation algorithm works well. This is mainly in the case of normal independent noise. If heavy-tailed or correlated noise is used then the cross-validation methods do not do so well (although see the description of Wang's work in Section 7.1, also Johnstone and Silverman [JS]).

This section concentrates on exchange rate data of the Australian dollar against the US dollar. This example is chosen to illustrate two things that can go *wrong* with the simple cross-validation scheme proposed in the previous section (although it is hoped that developments of this and other methods can overcome such problems). The version of the real data set is plotted in Figure 1. A possible model for this data set is exponential growth with the occasional negative shock (supposedly some piece of bad news hitting the market). Finally, the whole data set is subject to normal independent noise. To model this idea we create a simulated data set that consists of a randomly selected number of shocks (according to a Poisson distribution with mean 3), and at each shock there is a negative jump the size of which is the absolute value of the Student's t-distribution on 3 degrees of

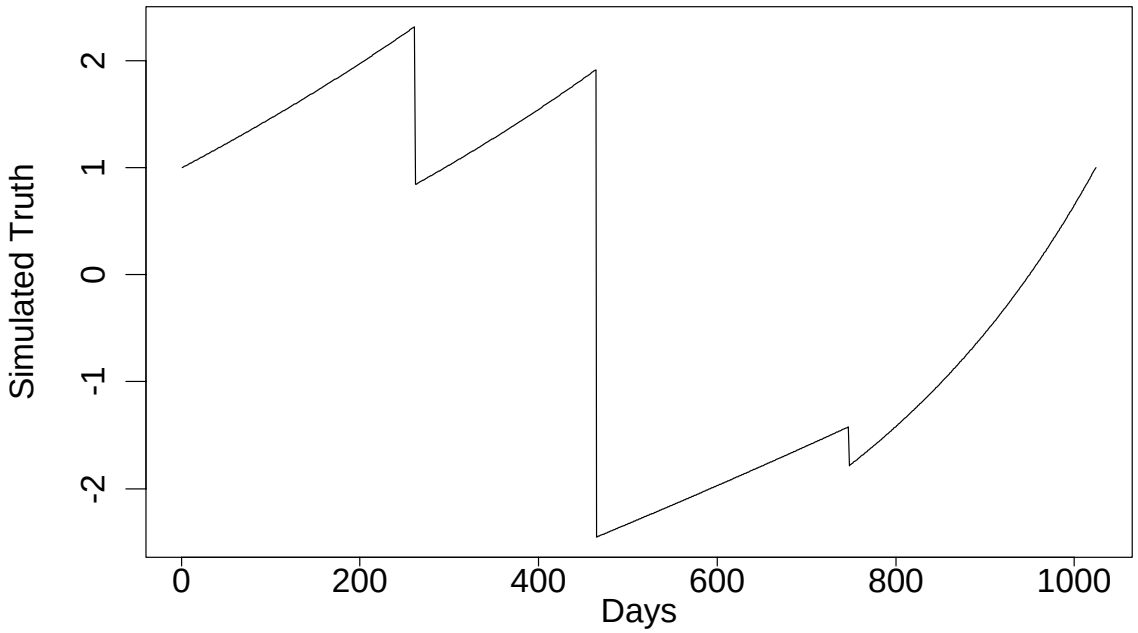

Figure 2: Simulation using exchange rate model

freedom. Then normal independent noise was added to the system. Figure 2 and Figure 3 show both the "truth" and a noisy version. The variances of the jumps, normal noise and exponential growth can all be varied to attempt to mimic the real data set (but of course, the point of this example is to demonstrate the shortcomings of the cross-validation method and only provide a barely reasonable model for the data.) Figure 4 shows a simulation using the described model and three reconstructions: universal, *GlobalSure* and cross-validation (*GlobalSure* is the *SureShrink* procedure of Donoho and Johnstone but altered so that one threshold is chosen for all levels). All the reconstructions in Figure 4 have their advantages and drawbacks. The universal reconstruction is certainly noise-free but tends to under-fit the data. The noise-free character could be extremely useful in some situations. In this case the universal reconstruction has only detected the large discontinuity and has a larger estimated l_2 norm when compared to the *GlobalSure* and cross-validation reconstructions (but remember that the purpose of universal thresholding is to minimize minimax risk, not l_2 error). The other two reconstructions have found the largest two discontinuities and it is arguable whether the cross-validation procedure has indicated the smallest one. Comparing the cross-validation and *GlobalSure* procedures the latter produces a visually preferable estimate and indeed the l_2 error is slightly smaller. Compare this example to the simulations performed by Nason [Na2] where the cross-validation procedure appeared to be superior in performance to *GlobalSure* for normal independent noise. Clearly, for more complex noise structures the ordinary cross-validation is not very good and Wang's modification given by [Wa1] is likely to perform better.

The result of applying the cross-validation method to the original data from Figure 1 appears in Figure 5. Figure 6 shows the universal reconstruction which is much smoother. The original data has large short-term autocorrelations and this destroys the performance

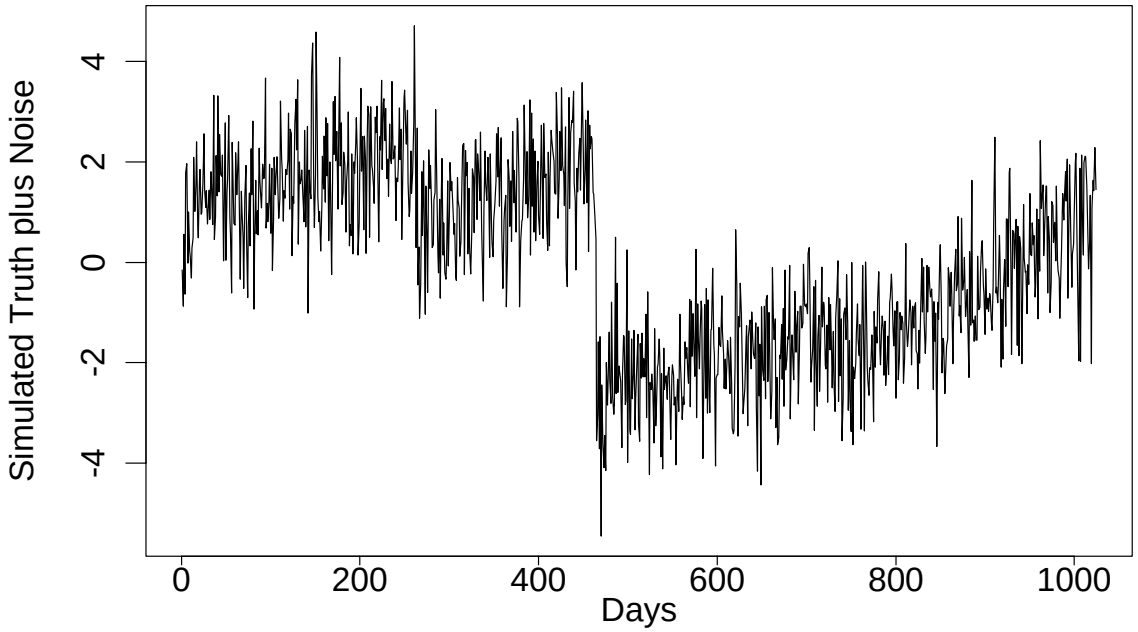

Figure 3: Simulation using exchange rate model plus noise

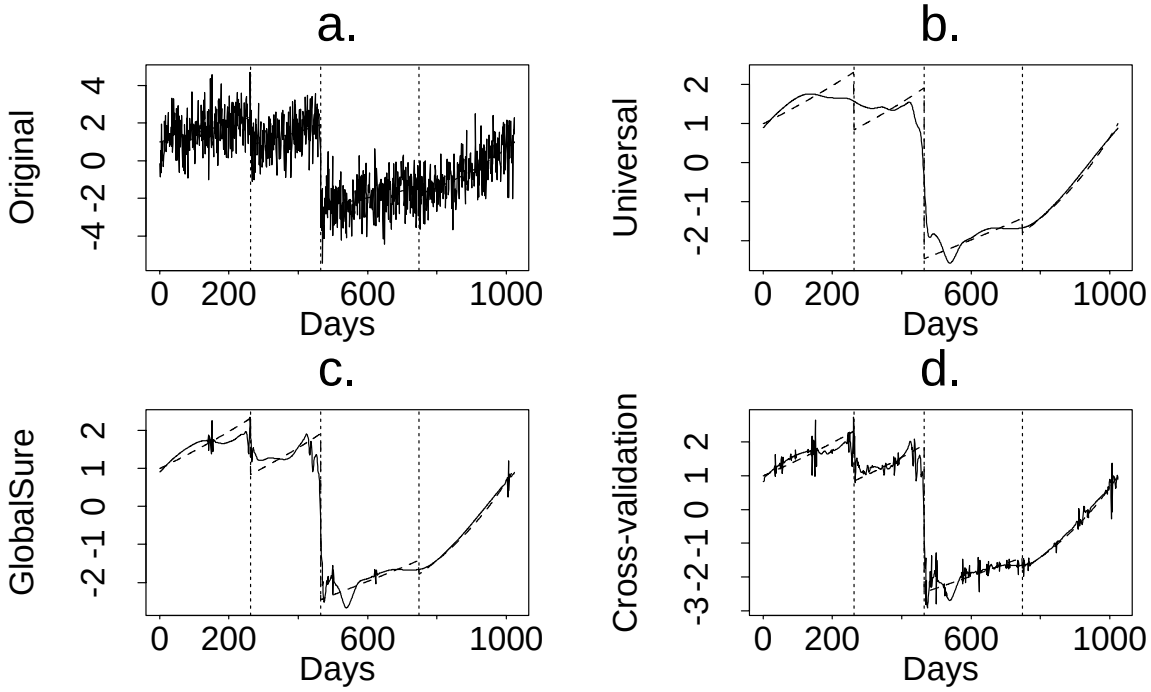

Figure 4: A noisy signal (a) with three jumps and denoised versions: (b) universal thresholding (T_{UV}= 4.15, l_2 norm=10.7); (c) *GlobalSure* thresholding (t_{SURE}= 2.31, l_2 norm=8.38); (d) cross-validation (t_{CV}= 1.72, l_2 norm=8.41). The vertical dotted lines indicate the jumps and the dashed lines indicate the true signal. The norm refers to the estimated l_2 norm between the true signal and the reconstruction.

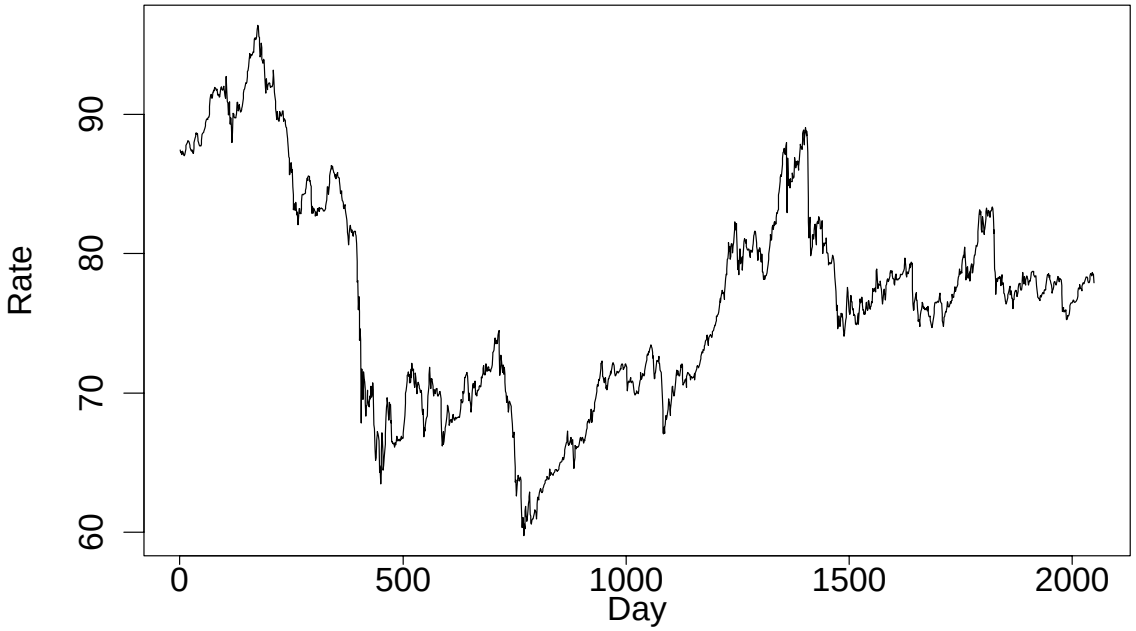

Figure 5: Cross-validation reconstruction from exchange rate data from Figure 1 (threshold value was 0.112)

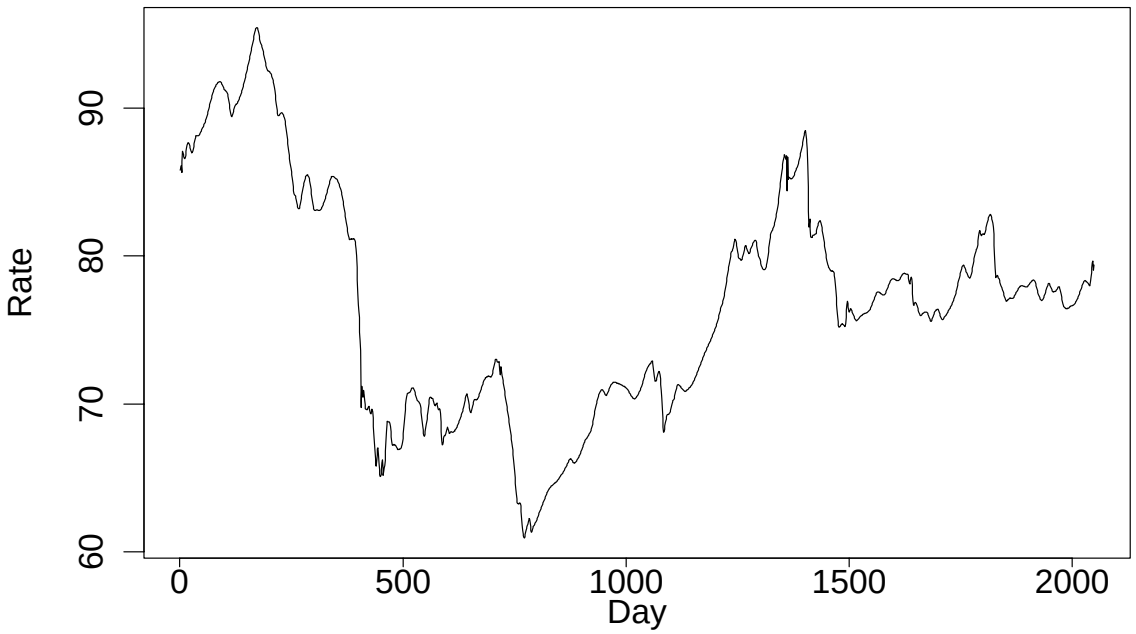

Figure 6: Universal reconstruction from exchange rate data from Figure 1 (threshold value was 1.364)

of the cross-validation procedure. By comparison the universal procedure is more robust.

6 Cross-validation in more dimensions

The extension of the two-fold cross-validation of Section 4.1 to k dimensions is achieved by using the multidimensional DWT of Mallat [Ma1]. As in Section 4 the cross-validation algorithm minimizes an estimate of the k-dimensional MISE (the x in equation (7) is now a vector in k-dimensional space). The next section develops an estimate of the k-dimensional MISE.

6.1 2^k-fold cross-validation algorithm

Assume now that the k-dimensional data may be denoted by $g_{i_1,\ldots,i_k}$ with $i_j \in \{1,\ldots,2^M\}$ for $j = 1,\ldots,k$. Suppose the data are arranged on a fixed equally spaced k-dimensional hypergrid H. For each of the k subscripts of g it is possible to select either the odd or evenly subscripted observations. Denote the selection of an even subscript by 0 and an odd subscript by 1. The selection 0 or 1 for each subscript provides a subset of H that is 2^{-k} times the size of H and equally spaced on a subgrid of H. For example, the selection g_{101} would select all the odd-indexed observations on subscripts 1 and 3 and the even-indexed observations on subscript 2.

Let S be the set containing the 2^k possible binary strings of length k and denote the ith largest binary string by b_i and the subgrid defined by b_i by H_i. Denote the k-dimensional wavelet shrinkage estimator with threshold t based on data g_{b_i} by $\hat{f}_{t,b_i}$. Denote the quantity $\bar{f}^j_{t,b_i}$ to be the interpolant of $\hat{f}_{t,b_i}$ to the grid defined by b_j by multiple repeats of the univariate interpolation scheme (8). This interpolation scheme is invariant with respect to the order in which each univariate interpolant is applied. Then the k-dimensional cross-validation score is given by:

$$\hat{M}(t) = \sum_{i\in S} \sum_{j\in S\setminus i} \sum_{m\in H_j} \left\{ \bar{f}^j_{t,b_i}(m) - g_{b_j}(m) \right\}^2 ,$$

where the final sum is over all indices m in the subgrid H_j.

6.2 Cross-validation for images

Images are two-dimensional objects and therefore 2^2-fold cross-validation can be used. This section illustrates the above algorithm using a 8×8 image on the pixel grid in Figure 7a. The grid H_4 is illustrated in Figure 7b. Mallat's two-dimensional DWT will be applied to the data in H_4 and a thresholded wavelet estimator constructed from it at threshold t. The estimator is then interpolated:

right to match the H_3 grid;

down to match the H_2 grid.

To match H_1 it is possible to either interpolate the H_3 grid downwards or the H_2 grid to the right — this demonstrates the invariance with respect to the ordering of the univariate interpolating procedure. Each of the interpolates ($\bar{f}^1_{t,b_4}$, $\bar{f}^2_{t,b_4}$, and $\bar{f}^3_{t,b_4}$) is compared to g_{00}, g_{01} and g_{10} using quadratic loss and each component summed to form the part of the

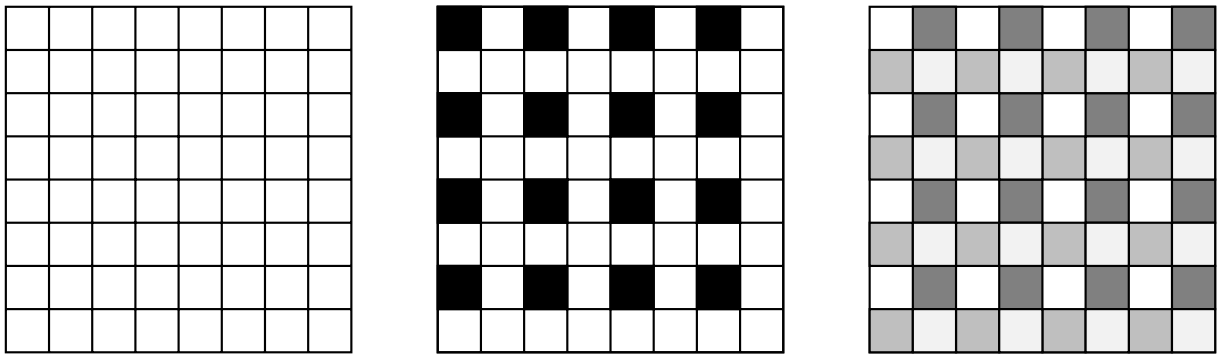

Figure 7: Organization for 2^2-fold cross-validation. a. The 8×8 pixel grid H. b. The data g_{11} forming the first subgrid H_4 associated with binary string $b_4 = 11$. c. The other three subgrids ■= H_3, ▒=H_2 and □=H_1 containing data g_{10}, g_{01} and g_{00} respectively.

estimate of $\hat{M}$ using H_4 as a starting point for constructing a wavelet estimator. This procedure is then repeated using H_1, H_2 and H_3 as starting points and the contributions are summed to form the final $\hat{M}$.

6.3 An example

Figure 8 shows the original image that we use for our example. Figure 9 shows the original image plus normal independent noise. Figures 10 and 11 show the universal and cross-validated reconstructions.

6.4 Computational effort and optimization

The 2-fold algorithm requires $O(n)$ operations. The 2^k-fold requires $O\{(2n)^k\}$ operations where n is the length of each side of the hypercube H.

The optimization algorithm that is used in all cases is the simple golden section search as mentioned in Press *et al.* [PTVF]. The algorithm works extremely well in practice. This is mainly because the function $\hat{M}$ is very nearly convex (to the eye on a large scale it looks convincingly convex). Detailed investigation of $\hat{M}$ by Nason [Na2] shows that the first derivative of $\hat{M}(t)$ is continuous and linear increasing on intervals defined by increasing $\{|w_{jk}|\}$ where $\{w_{jk}\}$ are the noisy wavelet coefficients formed from the transform of $g_1, \ldots, g_n$. At the points $t = |w_{jk}|$ the derivative may experience a discontinuity. Nason [Na2] provides heuristics that indicate that although these jumps may be negative they are usually small (only negative jumps cause non-convexity of $\hat{M}$) and therefore the zero-derivative point of $\hat{M}$ is usually well-determined.

Figure 8: Original Lennon image.

Figure 9: Noisy Lennon image (signal to noise ratio is $\frac{1}{2}$).

Figure 10: Universal threshold reconstruction using noisy image in Figure 9. The reconstruction threshold was 534 and the l_2 norm was 6543.

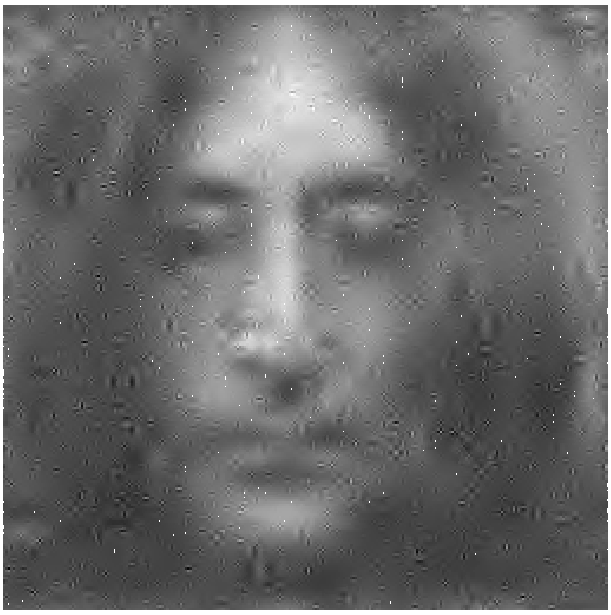

Figure 11: Cross-validated threshold using noisy image in Figure 9. The reconstruction threshold was 222 and the l_2 norm was 5660.

7 Improved Wavelet Cross-Validation Methods

7.1 Wang's method.

Wang [Wa1] is concerned with data that exhibit long-range dependence. That is data that obey model (3) but with

$$\mathrm{Cor}(\epsilon_i, \epsilon_j) \approx |i-j|^{-\alpha}$$

for $0 < \alpha < 1$ and dependent ϵ_i. By using a wavelet-vaguelette decomposition (WVD, see Donoho [Do1]) Wang nearly decorrelates fractional Gaussian noise which approximates long-range dependence processes. The correct procedure for cross-validation (or indeed any thresholding scheme in this case) is to apply a different threshold t_j to each level of the transform. Wang suggests two methods: one based on a level dependent universal threshold and one based on a generalization of Nason [Na2]. Wang's method serves as a general method for cross-validation for correlated data (see also Johnstone and Silverman [JS]). Suppose data $(g_1, \ldots, g_n)$ are given with $n = 2^M$, Wang's cross-validation algorithm is as follows:

1. Select an integer l_n and divide the data into 2^{M-l_n} groups as follows: the first 2^{l_n} observations $g_1, \ldots, g_{2^{l_n}}$ comprise the first group, the next 2^{l_n} observations $g_{2^{l_n}+1} \ldots, g_{2^{l_n+1}}$ form the second group and so on;

2. Remove the first observation from each group and form a data set of size 2^{M-l_n} from the removed observations. Then construct a wavelet estimate $\hat{f}_1^t$ from these points using a threshold $t_j = t2^{(H-1/2)(M-j)}$ where H is the parameter of the fractional Brownian motion behind the model that Wang assumes. Using the rest of the values interpolate to obtain $\bar{g}_1$ corresponding to the $\hat{f}_1^t$ values;

3. Repeat the previous step with the second, third, ..., up to the 2^{l_n}th observation from each group and obtain $(\hat{f}_2^t, \bar{g}_2), \ldots, (\hat{f}_{2^{l_n}}^t, \bar{g}_{2^{l_n}})$.

4. Define

$$M(t) = \sum_{l}^{2^{l_n}} \sum_{j} \left[\hat{f}_l^t(j) - \bar{g}_l(j)\right]^2.$$

   Let t^{CV} be the value of t that minimizes M. Then the cross-validation threshold is defined to be

$$t_j^{\mathrm{CV}} = \left(1 - \frac{l_n \log 2}{\log n}\right)^{-\frac{1}{2}} 2^{(H-1/2)(M-j)} t^{\mathrm{CV}}$$

   The term multiplying t^{CV} is a bias correction term equivalent to that in (10), except that there are 2^{l_n} observations in each group.

Notice that Nason's [Na2] cross-validation method is recovered for $l_n = 1$ and there are 2^{M-1} groups (pairs) consisting of two observations each. This method reduces the effect of inter-observation dependence and hopefully the selected threshold will be closer to the optimal threshold.

7.2 Weyrich and Warhola's method

Weyrich and Warhola [WW1] also discuss ordinary cross-validation and introduce a method of generalized cross-validation for wavelet regression. We only give the briefest of details here and refer the reader to [WW1] for a fuller discussion. They define a *generalized cross-validation* (GCV) criterion by

$$GCV(\delta) = \frac{\frac{1}{n}||(I_n - A)\mathbf{g}||_2^2}{\left\{\frac{1}{n}\mathrm{Tr}(I_n - \mathbf{A})\right\} 2},$$

where A is the operator

$$A = W^{-1} D_\delta W,$$

where W is the wavelet transform and D_δ is the thresholding operator. The trace of A is very simple to compute for the thresholding case and they define the GCV estimator of the ideal threshold to be that δ that minimizes the GCV criterion. They also give details of two extensions to the basic GCV algorithm:

1. The first extension introduces a threshold for each level in the transform. They then minimize the GCV by altering each level parameter separately and holding the others constant.

2. They also briefly mention using wavelet packets.

As for Wang, their simulation results appear extremely encouraging.

8 Conclusion

We hope that we have convinced the reader that cross-validation methods can make an important contribution to the estimation of functions using wavelet-based methods. They are simple to implement and understand, it is possible to utilize measures of fit that are appropriate to the application and they readily adapt to more than one dimension. It is likely that more powerful forms of transform such as those based on wavelet basis libraries will provide an even more adaptive and flexible class of representors and it is probable that cross-validation methods will still be of use.

9 Acknowledgments

We would like to thank the organizers of the "XV$^{\mathrm{e}}$ Rencontres Franco-Belges de Statisticiens" for financial support and for organizing a most enjoyable and intellectually stimulating meeting. We would also like to thank Geoff Eagleson, Australian Graduate School of Management, University of New South Wales, Australia for the exchange rate data. We would also like to thank the referee for many helpful comments.

References

[AB1] Abramovich, F., Benjamini, Y.: Adaptive thresholding of wavelet coefficients. (submitted for publication), (1994).

[BH1] Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Statist. Soc. B. **57** (1995), 289–300.

[Bu1] Burman, P.: A comparative study of ordinary cross-validation, ν-fold cross-validation and the repeated learning-testing methods. Biometrika. **76** (1989), 503–514.

[CDJV] Cohen, A., Daubechies, I., Jawerth, B., Vial, P.: Multiresolution analysis, wavelets, and fast algorithms on an interval. Compt. Rend. Acad. Sci. Paris A. **316** (1993), 417–421.

[Da1] Daubechies, I.: Orthonormal bases of compactly supported wavelets. Comms Pure Appl. Math.. **41** (1988), 909–996.

[Da2] Daubechies, I.: *Ten lectures on Wavelets.* SIAM, (1992).

[DJ1] Donoho, D.L., Johnstone, I.M.: Adapting to unknown smoothness via wavelet shrinkage. J. Am. Statist. Ass., (to appear).

[DJ2] Donoho, D.L., Johnstone, I.M.: Ideal spatial adaptation by wavelet shrinkage. Biometrika, (to appear).

[DJKP] Donoho, D.L., Johnstone, I.M., Kerkyacharian, G., and Picard, D.: Wavelet shrinkage: asymptopia? (with discussion). J. R. Statist. Soc. B. **57** (to appear).

[Do1] Donoho, D.L.: Nonlinear solution of linear-inverse problems by wavelet-vaguelette decomposition. Technical Report 403, Department of Statistics, Stanford University, Stanford, (1992).

[FHMP] Fan, J., Hall, P., Martin, M., Patil, P.: Adaption to high spatial inhomogeneity based on wavelets and on local linear smoothing. Technical Report CMA-SR18-93, Centre for Mathematics and Its Applications, Australian National University, Canberra, (1993).

[JS] Johnstone, I.M., Silverman, B.W.: Wavelet threshold estimators for data with correlated noise. (in preparation).

[KT] Kwong, M.K., Tang, P.T.P.: W-matrices, nonorthogonal multiresolution analysis, and finite signals of arbitrary length. Technical Report MCS-P449-0794, Argonne National Laboratory, (1994).

[Ma1] Mallat, S.G.: A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans. Pattn Anal. Mach. Intell. **11** (1989), 674–693.

[Me1] Meyer, Y.: *Wavelets and Operators.* Cambridge University Press, Cambridge, (1992).

[Na1] Nason, G.P.: The `WaveThresh` package; wavelet transform and thresholding software for S. Available from the StatLib archive, (1993).

[Na2] Nason, G.P.: Wavelet function estimation using cross-validation. (submitted for publication), (1994).

[Na3] Nason, G.P.: Wavelet regression by cross-validation. Technical Report 447, Department of Statistics, Stanford University, Stanford, (1994).

[NS1] Nason, G.P., Silverman, B.W.: The discrete wavelet transform in S. Journal of Computational and Graphical Statistics, **3** (1994), 163–191.

[NS2] Neumann, M.H., Spokoiny, V.G.: On the efficiency of wavelet estimators under arbitrary error distributions. The IMS Bulletin, **23** (1994) 218.

[Og1] Ogden, R.T.: Wavelet thresholding in nonparametric regression with change point applications. PhD thesis, Texas A&M University, (1994).

[PTVF] Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: *Numerical Recipes in C, the Art of Scientific Computing.* Cambridge University Press, Cambridge, (1992).

[SB1] Smith, M.J.T., Barnwell, T.P.: Exact reconstruction techniques for tree-structured subband coders. IEEE Transactions on Acoustics, Speech and Signal Processing. **34** (1986), 434–441.

[Se1] Serfling, R.J.: *Approximation Theorems of Mathematical Statistics.* Wiley, New York, (1980).

[St1] Stein, C.: Estimation of the mean of a multivariate normal distribution. Ann. Statist., **9** (1981), 1135–1151.

[St1] Stone, M.: Cross-validatory choice and assessment of statistical predictions (with discussion). J. R. Statist. Soc. B, **36** (1974), 111–147.

[TM1] Taswell, C., McGill, K.C.: Wavelet transform algorithms for finite duration discrete-time signals. Technical Report Numerical Analysis Project Manuscript NA-91-07, Department of Computer Science, Stanford University, (1991).

[Vi1] Vidakovic, B. Nonlinear wavelet shrinkage with Bayes rules and Bayes factors. (submitted for publication), (1994).

[Wa1] Wang, Y.: Function estimation via wavelets for data with long-range dependence. Technical Report, Univeristy of Missouri, Columbia, (1994).

[WW1] Weyrich, N., Warhola, G.T.: De-noising using wavelets and cross-validation. Technical Report AFIT/EN/TR/94-01, Department of Mathematics and Statistics, Air Force Institute of Technology, AFIT/ENC, 2950 P ST, Wright-Patterson Air Force Base, Ohio, 45433-7765, (1994).