Generalized Cross Validation for wavelet thresholding *

Maarten Jansen, Maurits Malfait, Adhemar Bultheel

January 1996
Revised July 26, 1996

keywords: Noise reduction, wavelets, Cross Validation, thresholding

Abstract

Noisy data are often fitted using a smoothing parameter, controling the importance of two objectives that are opposite to a certain extent. One of these two is smoothness and the other is closeness to the input data.

The optimal value of this paramater minimizes the error of the result (as compared to the unknown, exact data), usually expressed in the L_2 norm. This optimum cannot be found exactly, simply because the exact data are unknown. In spline theory, the Generalized Cross Validation (GCV) technique has proven to be an effective (though rather slow) statistical way for estimating this optimum.

On the other hand, wavelet theory is well suited for signal and image processing. This paper investigates the possibility of using GCV in a noise reduction algorithm, based on wavelet-thresholding, where the threshold can be seen as a kind of smoothing parameter. The GCV method thus allows choosing the (nearly) optimal threshold, without knowing the noise variance.

Both an original theoretical argument and practical experiments are used to show this successful combination.

*Appeared in *Signal Processing* **56(1)**, pp. 33 – 44, January 1997

Contents

1 Introduction

During the last decade wavelets have become a popular tool in signal and image processing [2]. Wavelet theory supports the idea of multiresolution [14] in a natural way. A multiresolution analysis of a signal allows to look at the signal at different scales. From the behaviour of a signal at successive scales, it is possible to derive its characteristics [13] and to detect singularities [12]. Unlike the classic Fourier analysis, a wavelet transform also maintains space or time information. A discontinuity in a function at a certain place only has influence on those coefficients that correspond to basis functions at that place. Since sine and cosine functions have an infinite support, a Fourier analysis is not able to localise edges. Wavelet basis functions are thus local in time and frequency. This permits to represent a regular signal with a small number of coefficients. Moreover, a wavelet transform is fast.

One of the applications of this theory is the problem of noise reduction. Many algorithms use some of the properties of a wavelet decomposition, to distinguish the regular part of a signal from a random perturbation. One, important class of algorithms defines a criterion to divide the wavelet coefficients into two groups. The first group contains the coefficients dominated by noise, while the other coefficients are sufficiently clean. The most simple procedures [5, 4, 3] are based on the observation that a limited number of coefficients are sufficient to reconstruct a regular signal. These algorithms eliminate all wavelet coefficients below a certain threshold because those coefficients are dominated by noise. Donoho and Johnstone [6] showed that this method has statistical optimality properties.
The approach of Xu, Weaver and collaborators [19] classifies the coefficients by their correlation between the successive scales.
Mallat and his collaborators [12] use the wavelet coefficients to compute local regularity parameters of the signal. Practically, their method only examines the regularity in the extrema of the wavelet coefficients.
All these methods are based on an individual, binary classification of wavelet coefficients. To incorporate spatial coherence between clean coefficients, Bayesian procedures [11, 10] introduce an a priori model for configurations of important coefficients.

This paper concentrates on the simple threshold algorithm. In this procedure, the threshold value is a parameter that the user can choose. The optimal threshold makes the result as close as possible to the noise-free signal. However, since this original signal is unknown, we cannot compute the error of the result and hence not minimize it. Donoho and Johnstone [4] propose a threshold proportional to the noise level.

But in many practical cases the actual amount of noise is not known. Instead of estimating the noise level, we try to find a good threshold directly, only using the input data. Weyrich and Warhola [18] and Nason [15] applied the idea of Cross Validation [7, 17, 1], and obtained excellent results. This Cross Validation is a function of the threshold value only based on the input data. Its minimum is a good approximation for the optimal threshold. Wahba [17] uses the same idea to find an optimal smoothing parameter for a spline fitting procedure. Unfortunately, computations are expensive in this case.

In this text we prove that the minimum of the "Generalized Cross Validation" is an asymptotically optimal threshold. In this way we justify the method of Weyrich and Warhola [18] and explain its success. Unlike the spline fitting case, removing small wavelet coefficients is a non-linear operation. The main problems in our proof arise from this non-linear character. Beside this theoretical evidence, we also illustrate the method with an example in image denoising. We emphasize that computations are faster than for spline smoothing.

This paper is organised as follows. In the second section we introduce some notation and formulate the problem. This leads to the definition of an error function which has to be minimized. In Section 3 we examine this error function. The following section tries to find out what we can do if we compare the input and output data for a given threshold. Section 5 introduces the idea of Cross Validation in an informal way. In Section 6 we prove that the method is the asymptotically optimal. In Section 7 we illustrate that computations are fast. The minimisation procedure is no bottleneck in the noise reduction algorithm. In Section 8 we indicate how we can apply this method to images and suggest a way to speed up the calculations for such large data sets.

2 Notations and problem

We consider the following model of a discrete noisy signal:

$$y_i = f_i + \varepsilon_i, \quad i = 1, \ldots, N,$$

or, in vector notation:

$$\boldsymbol{y} = \boldsymbol{f} + \boldsymbol{\varepsilon}.$$

The vector $\boldsymbol{y}$ represents input signal and $\boldsymbol{f}$ is an unknown, deterministic signal. We suppose that the noise $\boldsymbol{\varepsilon}$ is a stationary stochastic signal, i.e.

all its values are identically distributed with zero mean and variance σ^2. Thus $\mathrm{E}\varepsilon_i = 0$ and $\mathrm{E}\varepsilon_i^2 = \sigma^2, \forall i = 1, \ldots, N$. By $h(\varepsilon)$ we denote the density function of the noise. For the purpose of this text we restrict ourselves to uncorrelated (white) noise. This means that $\mathrm{E}\varepsilon_i\varepsilon_j = \delta_{ij}\sigma^2$.

To reconstruct the original data, we use a wavelet representation. We do not explain details of wavelet theory here. Basic theory can be found in many papers. We mention Daubechies [2] or Mallat [14]. For the purpose of this text, we use simple non-redundant, orthogonal, discrete wavelet transforms. This operation can be represented by an orthogonal matrix W. We consider the following transforms:

$$\begin{aligned} \boldsymbol{v} &= W\,\boldsymbol{f}, \\ \boldsymbol{\omega} &= W\,\boldsymbol{\varepsilon}, \\ \boldsymbol{w} &= W\,\boldsymbol{y} = \boldsymbol{v} + \boldsymbol{\omega}. \end{aligned}$$

This transform localizes the most important spatial and frequential characteristics of a regular signal in a limited number of wavelet coefficients. On the other hand, it is easy to prove that an orthogonal transform of stationary, white noise results in stationary white noise. This means that the expected noise energy is the same in all coefficients. If this energy is not too large, noise has a relatively small influence on the important large signal coefficients. These observations suggest to replace the small coefficients by zero, because they are dominated by noise and carry only a small amount of information. We will use Donoho's "soft-thresholding" or "shrinking" function, shown in Figure 1: a wavelet coefficient w between $-\delta$ and δ is set to zero, while the others are shrunk in absolute value.

The shrinking (soft-thresholding) operations can be represented as

$$\boldsymbol{w_\delta} = D_\delta \cdot \boldsymbol{w}, \tag{1}$$

where

$$D_\delta = \mathrm{diag}[d_{ii}], \tag{2}$$

with

$$d_{ii} = \begin{cases} 0 & \text{if } |w_i| < \delta, \\ 1 - \frac{\delta}{|w_i|} & \text{otherwise.} \end{cases} \tag{3}$$

Note that the notation $\boldsymbol{w_\delta} = D_\delta \cdot \boldsymbol{w}$ may be deceiving because the mapping $\boldsymbol{w} \mapsto \boldsymbol{w_\delta}$ is nonlinear. The elements of the "matrix" D_δ depend on the signal $\boldsymbol{w}$. In the same way we have:

$$\boldsymbol{v_\delta} = D_\delta \cdot \boldsymbol{v}.$$

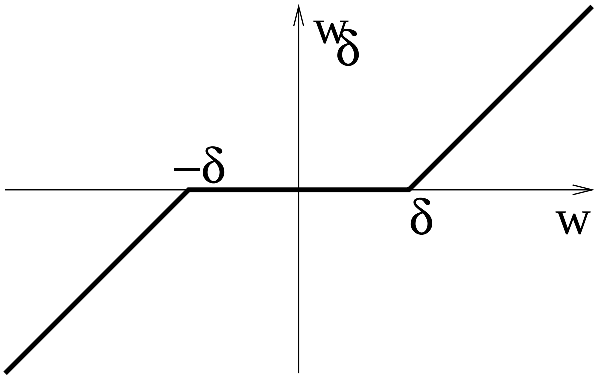

Figure 1: Soft-threshold function: a wavelet coefficient w with an absolute value below the threshold δ is replaced by 0. Coefficients with higher absolute values are shrunk.

Inverse transforms give the result:

$$\boldsymbol{y_\delta} = W^{-1} \cdot \boldsymbol{w_\delta}. \tag{4}$$

The overall operation can then be represented by:

$$\boldsymbol{y_\delta} = A_\delta \cdot \boldsymbol{y}, \tag{5}$$

where

$$A_\delta = W^{-1} \cdot D_\delta \cdot W. \tag{6}$$

We call A_δ the *influence matrix.* Through D_δ, it depends on the threshold value but also on the input signal $\boldsymbol{y}$.

A natural question arising from this procedure is how to choose the threshold δ. This choice should be optimal in a certain way. If $\boldsymbol{y_\delta}$ is the result of applying the threshold procedure to the wavelet coefficients of a signal $\boldsymbol{y}$, and $\boldsymbol{\varepsilon_\delta} = \boldsymbol{y_\delta} - \boldsymbol{f}$ is the noise of this result, then an often used criterion to measure the quality of this result is its signal-to-noise ratio ($SNR(\delta)$):

$$SNR(\delta) = 10 \cdot \log_{10} \frac{\sum_i f_i^2}{\sum_i \varepsilon_{\delta i}^2},$$

An optimal choice of δ should maximize $SNR(\delta)$. This is equivalent to minimizing the mean square error R:

$$R(\delta) := \frac{\sum_{i=1}^N (y_{\delta i} - f_i)^2}{N} = \frac{1}{N} \|\boldsymbol{y_\delta} - \boldsymbol{f}\|^2 = \frac{1}{N} \|\boldsymbol{\varepsilon_\delta}\|^2, \tag{7}$$

where we used the classical Euclidian vector norm based on the inner product $\langle p, q\rangle = \sum_i p_i q_i$. Because of the orthogonality of W, we can also compute R from the wavelet coefficients as:

$$R(\delta) = \frac{1}{N}\|\boldsymbol{\omega_\delta}\|^2, \tag{8}$$

where $\boldsymbol{\omega_\delta} = \boldsymbol{w_\delta} - \boldsymbol{v} = W\,\boldsymbol{\varepsilon_\delta}$ is the noise after the operation in the wavelet domain.

However, because $\boldsymbol{f}$ is unknown the function $R(\delta)$ is not computable and hence it cannot be used to find an optimal δ. The optimal threshold has to be estimated. Donoho and Johnstone [4] propose to use the "universal threshold" estimation:

$$\delta = \sqrt{2\log(N)}\;\sigma. \tag{9}$$

This formula and other, more complicated estimators require knowledge of the noise variance σ^2, which may not be readily available in practical applications. Weyrich and Warhola [18] therefore suggest to adapt Wahba's Generalized Cross Validation (GCV) [17, 1] for automatic spline smoothing. Applied to our wavelet procedure, this GCV should be a function of the threshold value, using only known data and having approximately the same minimum as the residual function R.

3 The mean-square error function $R(\delta)$

Since we are looking for an approximation of $R(\delta)$, it is useful to examine this function first. In the wavelet domain we have:

$$R(\delta) = \frac{1}{N}\|\boldsymbol{\omega_\delta}\|^2. \tag{10}$$

The expectation of this function can then be written as:

$$\begin{aligned} \mathrm{E}R(\delta) \;&=\; \frac{1}{N}\|\boldsymbol{v_\delta} - \boldsymbol{v}\|^2 + \frac{1}{N}\mathrm{E}\|\boldsymbol{w_\delta} - \boldsymbol{v_\delta}\|^2 \\ &\quad + \frac{2}{N}\langle(\boldsymbol{v_\delta} - \boldsymbol{v}), \mathrm{E}(\boldsymbol{w_\delta} - \boldsymbol{v_\delta})\rangle. \end{aligned} \tag{11}$$

Because exact coefficients are also transformed by the thresholding operations, the value of $\boldsymbol{\omega_\delta}$ contains a bias. This is reflected in the first term of equation (11). Therefore we also define:

$$\boldsymbol{\eta_\delta} := \boldsymbol{w_\delta} - \boldsymbol{v_\delta}.$$

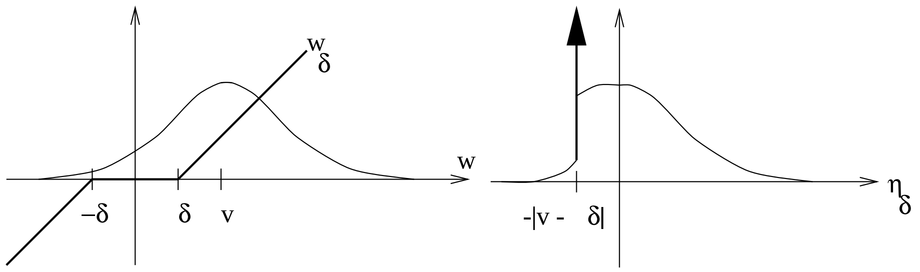

Figure 2: Left: Distribution of w, noisy wavelet coefficient, and shrinking function. Right: Distribution of $\eta_\delta := w_\delta - v_\delta$. This distribution contains a Dirac impulse at $-|v-\delta|$.

For linear operations, such as spline smoothing, the third term in (11) would be zero. Since shrinking is a non-linear operation, we cannot use this argument. But in the case where $|v_i| \gg \delta$, we have with a high probability that also $|w_i| \gg \delta$ and thus $\eta_{\delta i} := w_{\delta i} - v_{\delta i} = w_i - v_i = \omega_i$, and so $\mathrm{E}(w_{\delta i} - v_{\delta i}) = 0$. Problems occur only when $v_i \approx \delta$. Figure 2 shows how we can deduce the distribution of $\eta_{\delta i}$ in that case. The first part contains the distribution of w_i around its mean value v_i. If $-\delta \leq w_i \leq \delta$, then we have $\eta_{\delta i} = -|v_i - \delta|$. Thus this value of $\eta_{\delta i}$ has a finite probability, which can be seen on the second figure as a Dirac impulse in the distribution function of $\eta_{\delta i}$. Clearly this distribution does not have a zero mean. Taking into account other coefficients v_i, we can assume that this effect will be largely compensated, at least if the clean wavelet signal is sufficiently symmetric. We can conclude that the mean square error is approximately a sum of two terms. The first is a bias, and the second is due to the noise.

We now return to equation (11) and we consider the sum of the first and last term of the right-hand side. For the purpose of our proof, we need this sum to be positive. This holds true if for each index i:

$$(v_{\delta i} - v_i)^2 + 2(v_{\delta i} - v_i)\mathrm{E}\eta_{\delta i} \geq 0$$

For a Gaussian distribution of the noise, numerical computations show that all possible values of v_i satisfy this condition if

$$\delta > 0.735\,\sigma.$$

Since the interesting threshold values are mostly larger than σ, no problems arise from this restriction.

By this argument we can write for $\delta > 0.735\,\sigma$:

$$\mathrm{E}R(\delta) = b^2(\delta) + \mu_2(\delta) \cdot \sigma^2, \tag{12}$$

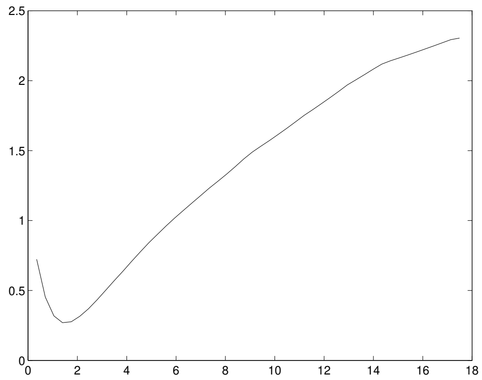

Figure 3: Typical form of $R(\delta)$. For small values of the threshold δ, the "input" noise has still an important contribution to the mean square error of the result. If the threshold is large, the shrinking operation deforms the original signal and causes a bias.

with:

$$b^2(\delta) = \frac{1}{N}\|\boldsymbol{v_\delta} - \boldsymbol{v}\|^2 + \frac{2}{N}\langle(\boldsymbol{v_\delta} - \boldsymbol{v}), \mathrm{E}(\boldsymbol{w_\delta} - \boldsymbol{v_\delta})\rangle, \tag{13}$$

and:

$$\mu_2(\delta) = \frac{\|\mathrm{E}(\boldsymbol{w_\delta} - \boldsymbol{v_\delta})\|^2}{N\sigma^2}. \tag{14}$$

Figure 3 shows a typical form of the function $R(\delta)$. For more information about this function, we refer to Nason [15].

4 A first estimator for $R(\delta)$

4.1 The effect of the threshold operation

We are looking for an estimator for $R(\delta)$, which is based on known variables. Therefore we first investigate the effect of the threshold operation on the input data.

Define

$$T(\delta) = \frac{\sum_{i=1}^N (w_{\delta i} - w_i)^2}{N} = \frac{1}{N}\|\boldsymbol{w_\delta} - \boldsymbol{w}\|^2. \tag{15}$$

Since neither $\boldsymbol{v}$ nor $\boldsymbol{v_\delta}$ are stochastic, we have $\mathrm{E}\langle\boldsymbol{\omega}, \boldsymbol{\omega_\delta}\rangle = \mathrm{E}\langle\boldsymbol{\omega}, \boldsymbol{\eta_\delta}\rangle$, and so we may write:

$$\mathrm{E}T(\delta) \quad = \quad \mathrm{E}\frac{1}{N}\left(\|\boldsymbol{w} - \boldsymbol{v}\|^2 + \|\boldsymbol{v} - \boldsymbol{w_\delta}\|^2 + 2\cdot\langle(\boldsymbol{w} - \boldsymbol{v}), (\boldsymbol{v} - \boldsymbol{w_\delta})\rangle\right)$$

$$
\begin{aligned}
&= \sigma^2 + \mathrm{E}R(\delta) - \frac{2}{N}\mathrm{E}\langle \boldsymbol{\omega}, \boldsymbol{\omega_\delta} \rangle \\
&= \sigma^2 + \mathrm{E}R(\delta) - \frac{2}{N}\mathrm{E}\langle \omega, \eta_\delta \rangle. \qquad (16)
\end{aligned}
$$

We investigate the third term in detail because this leads to some essential equations. Therefore we define:

$$\mu_1(\delta) := \frac{\sum_{i=1}^N \mathrm{E}[\omega_i \eta_{\delta i}]}{N\sigma^2}. \qquad (17)$$

4.2 Another expression for μ_1

We now prove :

Lemma 1 *If the density $h(\omega_i)$ is Gaussian, then*

$$\mathrm{E}[\omega_i \eta_{\delta i}] = \sigma^2 \mathrm{P}(|w_i| > \delta). \qquad (18)$$

Proof:

Since $h(\omega_i) = K \exp(-\omega_i^2/2\sigma^2)$, we have

$$\omega_i h(\omega_i) = -\sigma^2 h'(\omega_i), \qquad (19)$$

and so:

$$
\begin{aligned}
\mathrm{E}[\omega_i \eta_{\delta i}] &= \int_{-\infty}^{\infty} \eta_{\delta i}\, \omega_i\, h(\omega_i)\, \mathrm{d}\omega_i \\
&= -\sigma^2 \int_{-\infty}^{\infty} \eta_{\delta i}\, h'(\omega_i)\, \mathrm{d}\omega_i \\
&= -\sigma^2 \eta_{\delta i}\, h(\omega_i)\Big|_{-\infty}^{\infty} + \sigma^2 \int_{-\infty}^{\infty} \frac{\partial \eta_{\delta i}}{\partial \omega_i}\, h(\omega_i)\, \mathrm{d}\omega_i.
\end{aligned}
$$

Integration by parts is allowed since $\eta_{\delta i}(\omega_i)$ is a continuous function.
It is easy to see that:

$$\frac{\partial \eta_{\delta i}}{\partial \omega_i} = \begin{cases} 0 & \text{if } |w_i| < \delta, \\ 1 & \text{otherwise,} \end{cases}$$

from which (18) follows. $\square$

This lemma is in fact a special case of more general results by Hudson [8] and Stein [16].

We may conclude:

$$\mu_1(\delta) = \frac{1}{N} \sum_{i=1}^N \mathrm{P}(|w_i| > \delta). \qquad (20)$$

4.3 The derivative influence matrix

We now introduce a new matrix:

$$D'_{ij} = \frac{\partial w_{\delta i}}{\partial w_j}. \tag{21}$$

Note that if $i \neq j$, then $D'_{ij} = 0$.
For $i = j$ we have

$$D'_{ii} = \begin{cases} 0 & \text{if } |w_i| < \delta, \\ 1 & \text{otherwise.} \end{cases}$$

Thus, if $\mathrm{Tr}(D')$ is the trace of D', then

$$\mathrm{Tr}(D') = \#\{i \mid w_{\delta i} \neq 0\}.$$

Furthermore we consider the Jacobian matrix A' with entries

$$A'_{ij} = \frac{\partial y_{\delta i}}{\partial y_j}. \tag{22}$$

Then it is easy to see that

$$A' = W^{-1} \cdot D' \cdot W, \tag{23}$$

and, since W is non-singular,

$$\mathrm{Tr}(A') = \mathrm{Tr}(D').$$

A' is called the *derivative influence matrix.*

With these notations, and since for a Bernoulli variable

$$\mathrm{E}D'_{ii} = \mathrm{P}(D'_{ii} = 1), \tag{24}$$

we can rewrite μ_1 as:

$$\begin{aligned} \mu_1(\delta) &= \frac{1}{N}\sum_{i=1}^{N} \mathrm{P}(D'_{ii} = 1) \\ &= \frac{1}{N}\sum_{i=1}^{N} \mathrm{E}D'_{ii} \\ &= \frac{\mathrm{Tr}(\mathrm{E}A')}{N}. \end{aligned} \tag{25}$$

Starting from $\langle \boldsymbol{\omega}, \boldsymbol{\eta_\delta} \rangle$, which is not computable in practice, we end up with $\sigma^2 \mathrm{Tr}(A')$, which is easy to find while both have the same expectation. Thus, from (16), (17), and (25) we can construct

$$\mathrm{SURE}(\delta) := T(\delta) - \sigma^2 + 2\sigma^2 \cdot \frac{\mathrm{Tr}(A')}{N} \tag{26}$$

as an approximation for $R(\delta)$. However this function still requires a value for σ^2. Application of Stein's Unbiased Risk Estimator [16] leads to the same result [5].

5 Ordinary Cross Validation

This section introduces the idea of Cross Validation in an informal way. Our aim is to minimize the error function based on an unknown exact signal. We therefore try to find a good compromise between goodness of fit and smoothness. We assume that the original signal is *regular* to some extent, which means that the value f_i can be approximated by an linear combination of its neighbours. So, by considering $\tilde{y}_i$, a combination of y_j, not depending on y_i itself, we can eliminate the noise in this particular component. Since we replace it by a weighted average of its neighbours, noise in these components is smoothed, and so we end up with a relatively clean, noise-independent value. This value can be used in the computation of an approximation for $R(\delta)$.

To investigate the closeness of fit, we compute the result of the threshold operation for the modified signal $\tilde{\boldsymbol{y}}$, in which the i–th component y_i was replaced by $\tilde{y}_i$, i.e.,

$$\tilde{\boldsymbol{y}} = A \cdot (y_1, \ldots, y_{i-1}, \tilde{y}_i, y_{i+1}, \ldots, y_N)^T.$$

We then consider the ability of $\tilde{y}_{\delta i}$ to "predict" the value y_i as a measure for the optimality of the choice of the threshold [1].

For (too) small values of δ the difference $y_i - \tilde{y}_{\delta i}$ is dominated by noise, while for large values of δ the signal itself is too much deformed. We repeat the same procedure for all components and compute

$$OCV := \frac{1}{N} \sum_{i=1}^{N} (y_i - \tilde{y}_{\delta i})^2 \tag{27}$$

to express the compromise. This function is called "Ordinary Cross Validation". This name indicates that we use the values of the *other* components in the calculation for one point.

Many combination formulas are possible for $\tilde{y}_i$. Most obvious is to take $\tilde{y}_i = \frac{1}{2} \cdot (y_{i-1} + y_{i+1})$. But taking $\tilde{y}_i$ so that $\tilde{y}_{\delta i} = \tilde{y}_i$ will turn out to be an interesting choice. This value always exists, since the threshold algorithm has a levelling effect. Indeed, taking $\tilde{y}_{\delta i} = \max_i y_i$, we obtain $\tilde{y}_{\delta i} \leq \tilde{y}_i$, while the opposite is true for $\tilde{y}_{\delta i} = \min_i y_i$. So, by continuity arguments, one can expect such a value to exist.

For this last choice of $\tilde{y}_i$ we can write:

$$y_i - \tilde{y}_{\delta i} = \frac{y_i - y_{\delta i}}{1 - a_i^*},$$

with:

$$a_i^* = \frac{y_{\delta i} - \tilde{y}_{\delta i}}{y_i - \tilde{y}_{\delta i}} = \frac{y_{\delta i} - \tilde{y}_{\delta i}}{y_i - \tilde{y}_i} \approx \frac{\partial y_{\delta i}}{\partial y_i} = A'_{ii}.$$

So we have:

$$OCV \approx \frac{1}{N} \sum_{i=1}^{N} (y_i - y_{\delta i})^2 . w_i^2(\delta),$$

with:

$$w_i(\delta) = \frac{1}{(1 - A'_{ii})}.$$

However this cannot be used in practical computations, since A'_{ii} is 0 or 1. Therefore we take some kind of mean value for $w_i(\delta)$:

$$w_i(\delta) = w(\delta) = \frac{1}{\frac{1}{N} \cdot \sum_{i=1}^{N} (1 - A'_{ii})}.$$

This gives us the formula of the so called “Generalized Cross Validation”.

6 Generalized Cross Validation

6.1 Definition

So we have as a definition of the Generalized Cross Validation:

$$GCV(\delta) = \frac{\frac{1}{N} \cdot \|\boldsymbol{y} - \boldsymbol{y_\delta}\|^2}{[\frac{\mathrm{Tr}(I - A')}{N}]^2} = \frac{T(\delta)}{S(\delta)}, \tag{28}$$

with $T(\delta)$ as in Section 4 and

$$S(\delta) = \left[\frac{\mathrm{Tr}(I - A')}{N} \right]^2. \tag{29}$$

If the wavelet transform is orthogonal, the same formula can be used, mutatis mutandis, in the wavelet domain.

6.2 Asymptotic behaviour

In this paragraph we shall prove that if $\delta^* = arg \min R(\delta)$ and $\tilde{\delta} = arg \min GCV(\delta)$, then for $N \to \infty$, both minimizers yield a result of the same quality:

$$\frac{\mathrm{E}R(\tilde{\delta})}{\mathrm{E}R(\delta^*)} \downarrow 1, \tag{30}$$

The first difficulty is due to the fact that, unlike in the spline case, $GCV(\delta)$ is a quotient of two variables both depending on the input signal. Next, we compare the result obtained by the minimal GCV-threshold $\tilde{\delta}$ with the result for the optimal threshold δ^*. We give an upper bound for the ratio $\frac{R(\tilde{\delta})}{R(\delta^*)}$. Finally we show that this upper bound tends to one.

6.2.1 A quotient of two random variables

$GCV(\delta)$ is a quotient of two random, mutually dependent, variables. We therefore use asymptotic arguments to obtain that for $N \to \infty$ (recall (25)):

$$\mathrm{E}GCV(\delta) \to \frac{\mathrm{E}T(\delta)}{[1-\mu_1(\delta)]^2}.$$

6.2.2 Quality of the GCV estimate of the optimal threshold

For $N \to \infty$ we can write:

$$\begin{aligned}
\frac{\mathrm{E}R(\delta) - (\mathrm{E}GCV(\delta) - \sigma^2)}{\mathrm{E}R(\delta)} &= 1 - \frac{\mathrm{E}GCV(\delta)}{\mathrm{E}R(\delta)} + \frac{\sigma^2}{\mathrm{E}R(\delta)} \\
&\approx 1 - \frac{\mathrm{E}R + \sigma^2 - 2\sigma^2\mu_1}{(1-\mu_1)^2 \cdot \mathrm{E}R} + \frac{\sigma^2}{\mathrm{E}R} \\
&= (1 - \frac{1}{(1-\mu_1)^2}) + \frac{\sigma^2}{\mathrm{E}R} \cdot [\frac{-1+2\mu_1}{(1-\mu_1)^2} + 1] \\
&= \frac{-\mu_1(2-\mu_1)}{(1-\mu_1)^2} + \frac{\sigma^2}{b^2 + \sigma^2 \cdot \mu_2} \cdot \frac{\mu_1^2}{(1-\mu_1)^2}.
\end{aligned}$$

Because $\mu_1 \leq 1$, we have $2\mu_1 \geq \mu_1^2$, and so:

$$\begin{aligned}
\left|\frac{\mathrm{E}R(\delta) - (\mathrm{E}GCV(\delta) - \sigma^2)}{\mathrm{E}R(\delta)}\right| &\leq \frac{1}{(1-\mu_1)^2} \cdot (|-2\mu_1 + \mu_1^2| + \left|\frac{\sigma^2\mu_1^2}{b^2 + \sigma^2\mu_2}\right|) \\
&\leq \frac{1}{(1-\mu_1)^2} \cdot (2\mu_1 - \mu_1^2 + \frac{\mu_1^2}{\mu_2}) \\
&\leq \frac{1}{(1-\mu_1)^2} \cdot (2\mu_1 + \frac{\mu_1^2}{\mu_2}) =: h(\delta).
\end{aligned}$$

If $\mathrm{E}GCV(\tilde{\delta}) = \min_\delta \mathrm{E}GCV(\delta)$ and $\mathrm{E}R(\delta^*) = \min_\delta \mathrm{E}R(\delta)$, then:

$$[1 - h(\tilde{\delta})]\mathrm{E}R(\tilde{\delta}) \leq \mathrm{E}GCV(\tilde{\delta}) - \sigma^2 \leq \mathrm{E}GCV(\delta^*) - \sigma^2 \leq [1 + h(\delta^*)]\mathrm{E}R(\delta^*),$$

or:

$$1 \leq \frac{\mathrm{E}R(\tilde{\delta})}{\mathrm{E}R(\delta^*)} \leq \frac{1 + h(\delta^*)}{1 - h(\tilde{\delta})}. \tag{31}$$

6.2.3 Limit behaviour of the upper bound

If $h(\delta) \to 0$, then $\mathrm{E}R(\tilde{\delta}) \to \mathrm{E}R(\delta^*)$. To this end, it is sufficient that $\mu_1(\delta) \to 0$ and $\frac{\mu_1^2(\delta)}{\mu_2(\delta)} \to 0$. We explain that these two conditions are fulfilled for $N \to \infty$ if we accept the following assumption:

Assumption 1 *The original signal $\boldsymbol{f}$ is* smooth, *i.e. it can be represented by a limited number of coefficients in the wavelet basis.*
More precisely, this means the following:
Denote by V_0 the set of clean wavelet coefficients that are approximately equal to zero: $V_0 = \{i \mid |v_i| < r\delta\}$, and by V_1 the other coefficients: $V_1 = \{i \mid |v_i| > r\delta\}$, where r is some small value. If $N = N_0 + N_1$ with $N_0 = \#V_0$, and $N_1 = \#V_1$, then we assume that for N sufficiently large N_1 remains constant.

We can write:

$$\begin{aligned}\mu_1(\delta) &\leq \frac{\sum_{i\in V_1} 1 + \sum_{i \in V_0} \mathrm{P}(|w_i| > \delta)}{N} \\ &\leq \frac{N_1}{N} + \frac{N_0}{N}\mathrm{P}(|\omega_i| > (1-r)\,\delta).\end{aligned}$$

If we suppose that $\delta_{\mathrm{opt}} \approx \sqrt{2\log(N)}\sigma$, and we use the asymptotic expression

$$\int_x^\infty \mathrm{e}^{\frac{-u^2}{2\sigma^2}}\,\mathrm{d}u \sim \frac{\sigma^2 \mathrm{e}^{\frac{-x^2}{2\sigma^2}}}{x}, \tag{32}$$

then we have:

$$\mu_1(\delta) \sim \frac{N_1}{N} + \frac{1}{\sqrt{\pi}N^{(1-r)^2}\sqrt{\log N}}. \tag{33}$$

To show that $\frac{\mu_1^2}{\mu_2} \to 0$, we use the fact that for positive a, b, c, d:

$$\frac{a}{b} < \frac{c}{d} \Rightarrow \frac{a}{b} < \frac{a+c}{b+d} < \frac{c}{d}. \tag{34}$$

We have:

$$\frac{\mu_1}{\mu_2} = \frac{\sum_{i=1}^N \mathrm{E}(\omega_i \eta_{\delta i})}{\sum_{i=1}^N \mathrm{E}(\eta_{\delta i}^2)} \leq \max_{i=1\ldots N} \frac{\mathrm{E}(\omega_i \eta_{\delta i})}{\mathrm{E}(\eta_{\delta i}^2)}. \tag{35}$$

To find this maximum, we distinguish three cases:

1. $v_i > \delta$

   For $w_i > \delta$, we have $\eta_{\delta i} = \omega_i$, while for $|w_i| < \delta$, this becomes $\eta_{\delta i} = \delta - v_i$. If $w_i < \delta$, then $\eta_{\delta i} < \delta - v_i$ and consequently, $\omega_i \eta_{\delta i} > (\delta - v_i)^2$. So we have:

   $$\mathrm{E}(\eta_{\delta i}^2) \geq \int_{\delta - v_i}^{\infty} \omega_i^2\, h(\omega_i)\, \mathrm{d}\omega_i + (\delta - v_i)^2 \int_{-\infty}^{\delta - v_i} h(\omega_i)\, \mathrm{d}\omega_i.$$

   The right-hand side only depends on the difference $b := \delta - v_i$, not on δ itself. Moreover it is strictly positive with a minimum $C > 0$. This leads for this case to:

   $$\frac{\mathrm{E}(\omega_i \eta_{\delta i})}{\mathrm{E}(\eta_{\delta i}^2)} < \frac{\sigma^2}{C}.$$

2. The case $v_i < -\delta$ is completely similar.

3. $|v_i| < \delta$

   Set $a = -\delta - v_i$, and $b = \delta - v_i$, then we have:

   $$\begin{aligned}\frac{\mathrm{E}(\omega_i \eta_{\delta i})}{\mathrm{E}(\eta_{\delta i}^2)} &= \frac{\int_{-\infty}^a \omega_i(\omega_i - a)\, h(\omega_i)\, \mathrm{d}\omega_i + \int_b^\infty \omega_i(\omega_i - b)\, h(\omega_i)\, \mathrm{d}\omega_i}{\int_{-\infty}^a (\omega_i - a)^2\, h(\omega_i)\, \mathrm{d}\omega_i + \int_b^\infty (\omega_i - b)^2\, h(\omega_i)\, \mathrm{d}\omega_i} \\ &\leq \max\left(\frac{\int_{-\infty}^a \omega_i(\omega_i - a)\, h(\omega_i)\, \mathrm{d}\omega_i}{\int_{-\infty}^a (\omega_i - a)^2\, h(\omega_i)\, \mathrm{d}\omega_i}, \frac{\int_b^\infty \omega_i(\omega_i - b)\, h(\omega_i)\, \mathrm{d}\omega_i}{\int_b^\infty (\omega_i - b)^2\, h(\omega_i)\, \mathrm{d}\omega_i} \right) \\ &\leq \max_b \frac{\int_b^\infty \omega_i(\omega_i - b)\, h(\omega_i)\, \mathrm{d}\omega_i}{\int_b^\infty (\omega_i - b)^2\, h(\omega_i)\, \mathrm{d}\omega_i}.\end{aligned}$$

   It is easy to check (numerically) that this maximum is reached for $b = 2\delta$. We call

   $$\mu_q(\delta) = \frac{\int_{2\delta}^\infty \omega_i(\omega_i - 2\delta)\, h(\omega_i)\, \mathrm{d}\omega_i}{\int_{2\delta}^\infty (\omega_i - 2\delta)^2\, h(\omega_i)\, \mathrm{d}\omega_i}. \tag{36}$$

   Using (19) we have:

   $$\begin{aligned}\int_b^\infty \omega_i(\omega_i - b)\, h(\omega_i)\, \mathrm{d}\omega_i &= -\sigma^2 \int_b^\infty (\omega_i - b) h'(\omega_i)\, \mathrm{d}\omega_i \\ &= -\sigma^2 (\omega_i - b)\, h(\omega_i) \Big|_b^\infty + \sigma^2 \int_b^\infty h(\omega_i)\, \mathrm{d}\omega_i \\ &= \sigma^2 \int_b^\infty h(\omega_i)\, \mathrm{d}\omega_i. \end{aligned} \tag{37}$$

Now (36) becomes:

$$\mu_q(\delta) = \frac{\sigma^2 \int_{2\delta}^{\infty} h(\omega_i)\,\mathrm{d}\omega_i}{(\sigma^2 + (2\delta)^2) \int_{2\delta}^{\infty} h(\omega_i)\,\mathrm{d}\omega_i - 2\delta \int_{2\delta}^{\infty} \omega_i\, h(\omega_i)\,\mathrm{d}\omega_i}. \tag{38}$$

A long but trivial calculation shows that

$$\mu_q(\delta) \sim \frac{2\delta^2}{\sigma^2}. \tag{39}$$

If $\delta_{\mathrm{opt}} \approx \sqrt{2\log(N)}\sigma$, we have

$$\frac{\mu_1^2(\delta)}{\mu_2(\delta)} = \mu_1(\delta)\frac{\mu_1(\delta)}{\mu_2(\delta)} \leq \mu_1(\delta) \cdot \max\left(\mu_q(\delta)\,,\,\frac{\sigma^2}{C}\right) \sim \frac{\log N}{N} \to 0. \tag{40}$$

6.3 Conclusion

We have proven:

Theorem 1 *for $N \to \infty$:*

$$\frac{\mathrm{E}R(\tilde{\delta})}{\mathrm{E}R(\delta^*)} \downarrow 1, \tag{41}$$

and in the neighbourhood of δ^*:

$$\mathrm{E}GCV(\delta) \approx \mathrm{E}R(\delta) + \sigma^2. \tag{42}$$

Figure 4 compares both functions $R(\delta)$ and $V(\delta)$ for a typical case. The noise variance was 1.1925.

7 Computational aspects

The procedure to be executed can be described as follows:

1. Compute $\boldsymbol{w} = W \cdot \boldsymbol{y}$, using the fast wavelet transform algorithm.
2. Choose a starting threshold value.
3. Minimise $GCV(\delta)$.
4. Compute $\boldsymbol{w_\delta} = D_\delta \cdot \boldsymbol{w}$.
5. Inverse transform: $\boldsymbol{y_\delta} = W^{-1} \cdot \boldsymbol{w_\delta}$.

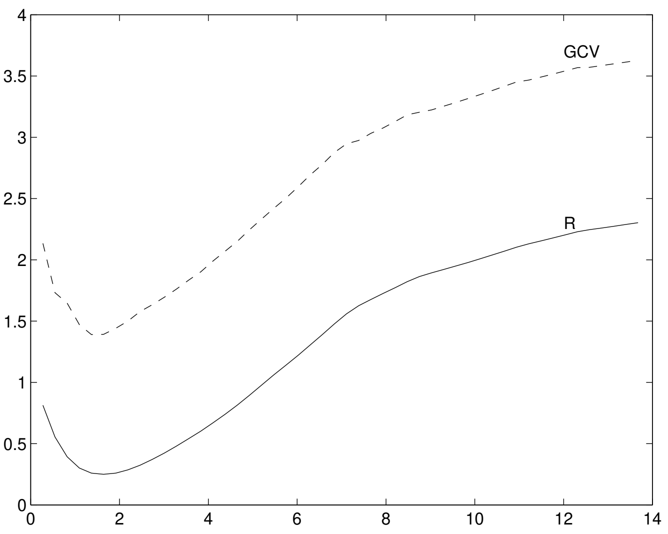

Figure 4: GCV and mean square error of the result in function of the threshold δ .

Because $GCV(\delta)$ is an approximation itself, it is not useful to compute its minimum very precisely. Moreover, in most cases this is not necessary either, due to the smooth curve of $R(\delta)$ in the neighbourhood of its minimum.

A relative accuracy of 10^{-4} will do. Using a classic minimisation procedure (such as Fibonacci) this requires approximately 20 function evaluations.

Computation of $GCV(\delta)$ can be performed completely in the wavelet domain. Only at the beginning of the minimisation procedure a wavelet transform is needed. As we said before, the denominator

$$\mathrm{Tr}(I - A'_\delta) = \mathrm{Tr}(I) - \mathrm{Tr}(A'_\delta) = N - \mathrm{Tr}(D'_\delta)$$

counts the number of coefficients that are set to zero. This does not require any floating point operation. Computation of the numerator can be done with $2N$ floating point operations. So 20 function evaluations lead to some $40N$ floating point operations.

For a fast wavelet transform we need $2F2N$ flops, where F is the number of filter coefficients. For $F = 4$, we have $16N$ flops. To reconstruct the signal after the operation with optimal δ, we need an inverse transform too. This makes the minimisation procedure not too expensive, as compared with the wavelet transform.

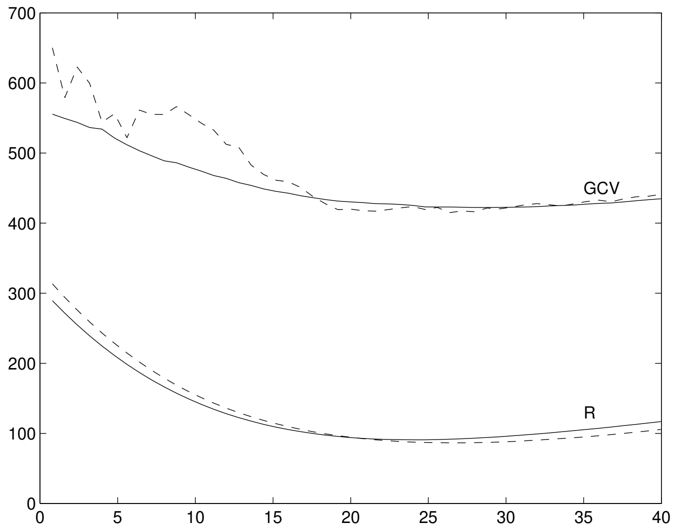

Figure 5: $GCV(\delta)$ based on all (256×256) pixels in full line, based on 1000 pixels in dashed line. Idem for $R(\delta)$.

8 Application to images

An image can be seen as a function of two independent variables, say x and y. If we have a two-dimensional wavelet transform, GCV theory can easily be adapted to the coefficients computed by this formula.

A trivial way to construct a 2D-wavelet transform, starting from its 1D version, is to use tensor products. The resulting "square wavelet transform" is described in [9]. Obviously, this is not optimal, since the axes are mostly arbitrarily chosen lines which do not have any meaning for the image. This method thus introduces artifacts and more sophisticated methods therefore produce better results. However, since this text illustrates the possibilities of GCV theory for image denoising, this simple choice is sufficient here.

Though computation of $GCV(\delta)$ is quick, for large images time may become crucial. To deal with this problem, one can base the computation of $GCV(\delta)$ not on all pixels, but on a well selected, representative part of it. Of course $GCV(\delta)$ cannot be computed exactly in this way. For a 256×256 pixel image we used a very simple equidistant sampler of 1000 pixels and obtained the results given in figure 5. This figure shows a detail of the $GCV(\delta)$ curve, together with its approximation and also the corresponding curves for the real square error function.

In this case the minimum of the approximate GCV is accidentally a

Figure 6: Result of the threshold operation with the minimizer of $GCV(\delta)$, based on 1000 pixels. On the left hand we have the original image, in the middle the noisy image (SNR = 10.0dB), on the right the result (SNR = 15.5dB).

better approximation for the optimal threshold than the minimum of the true GCV. Figure 6 shows the eventual result.

9 Discussion

We introduced Generalized Cross Validation for wavelet shrinking, based on the absolute value of the wavelet coefficients. This combination has been proved to be successful. Without knowledge of the noise level, one can find a nearly optimal threshold. However wavelet thresholding is a very simple, straightforward denoising algorithm. It would be interesting to know whether a similar combination is also possible with more sofisticated denoising procedures. In Section 4, we explicitly used the shrinking formula to compute $\mathrm{E}[\omega_i \eta_{\delta i}]$. At this point our arguments do not hold for other algorithms.

A possible modification is to compute a separate threshold for each resolution level. Although this is more adaptive, problems might occur from the fact that at low levels the Generalized Cross Validation function is based on too little coefficients (recall that this procedure is only asymptotically optimal). One can therefore group some levels together.

Acknowledgement

This research was financed by a grant from the Flemish Institute for the Promotion of Scientific and Technological Research in the Industry (IWT). We also thank Wim Sweldens for carefully proof reading this text. Discussion with him not only provided useful feedback but also led to interesting questions for possible future research.

References

[1] P. Craven and G. Wahba. Smoothing noisy data with spline functions. *Numerische Mathematik*, 31:377–403, 1979.

[2] I. Daubechies. *Ten Lectures on Wavelets.* CBMS-NSF Regional Conf. Series in Appl. Math., Vol. 61. Society for Industrial and Applied Mathematics, Philadelphia, PA, 1992.

[3] D. Donoho. De-noising via soft thresholding. Technical Report 409, Department of Statistics, Stanford University, 1992.

[4] D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, 81:425–455, 1994.

[5] D. L. Donoho and I. M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.*, to appear, 1995.

[6] D. L. Donoho and I. M. Johnstone. Wavelet shrinkage: Asymptopia? *Journal of the Royal Statistical Society, Series B*, 57(2):301–369, 1995.

[7] P. Hall and I. Koch. On the feasibility of cross-validation in image analysis. *SIAM J. Appl. Math.*, 52(1):292–313, 1992.

[8] H. M. Hudson. A natural identity for exponential families with applications in multiparameter estimation. *Annals of Statistics*, 6(3):473–484, 1978.

[9] B. Jawerth and W. Sweldens. An overview of wavelet based multiresolution analyses. *SIAM Review*, 36(3):377–412, 1994.

[10] M. Malfait. *Stochastic Sampling and Wavelets for Bayesian Image Analysis.* PhD thesis, Department of Computer Science, K.U.Leuven, Belgium, 1995.

[11] M. Malfait. Using wavelets to suppress noise in biomedical images. In A. Aldroubi and M. Unser, editors, *Wavelets in Medicine and Biology*, pages 191–208. CRC Press, Boca Raton, FL, April 1996.

[12] S. Mallat and W. L. Hwang. Singularity detection and processing with wavelets. *IEEE Transactions on Information Theory*, 38(2):617–643, 1992.

[13] S. Mallat and S. Zhong. Characterization of signals from multiscale edges. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14:710–732, 1992.

[14] S. G. Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, 1989.

[15] G. P. Nason. Wavelet regression by cross validation. Preprint, Department of Mathematics, University of Bristol, UK, 1994.

[16] C. Stein. Estimation of the mean of a multivariate normal distribution. *Annals of Statistics*, 9(6):1135–1151, 1981.

[17] G. Wahba. *Spline Models for Observational Data*, chapter 4, pages 45–65. CBMS-NSF Regional Conf. Series in Appl. Math. Society for Industrial and Applied Mathematics, Philadelphia, PA, 1990.

[18] N. Weyrich and G. T. Warhola. De-noising using wavelets and cross validation. In S.P. Singh, editor, *Approximation Theory, Wavelets and Applications*, volume 454 of *NATO ASI Series C: Mathematics and Physical Sciences*, pages 523–532. Kluwer, 1995.

[19] Y. Xu, J. B. Weaver, D. M. Healy, and J. Lu. Wavelet transform domain filters: a spatially selective noise filtration technique. *IEEE Transactions on Image Processing*, 3(6):747–758, 1994.

List of Figures

Correspondence to

Maarten Jansen
Department of Computer Science
Katholieke Universiteit Leuven
Celestijnenlaan 200A
B-3001 Heverlee
BELGIUM
E-mail: maarten.jansen@cs.kuleuven.ac.be