Generalized Cross Validation for wavelet thresholding

Maarten Jansen
Maurits Malfait
Adhemar Bultheel

Report TW 236, January 1996



Katholieke Universiteit Leuven
Department of Computer Science
Celestijnenlaan 200A – B-3001 Heverlee (Belgium)

Generalized Cross Validation for wavelet thresholding

Maarten Jansen
Maurits Malfait
Adhemar Bultheel

Report TW 236, January 1996

Department of Computer Science, K.U.Leuven

Abstract

Noisy data are often fitted using a smoothing parameter, controling the importance of two objectives that are opposite to a certain extend. 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 in 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. Moreover, computations turn out to be very fast.

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

Keywords : wavelets, soft-thresholding, generalized cross validation
AMS(MOS) Classification : 41A30, 94A12, 93E14, 60G35, 65D10, 68U10

Contents

1 Introduction **2**

2 Notations for wavelet transforms and threshold operations **3**

3 The mean-square error function $R(\delta)$ **4**

4 A first estimator for $R(\delta)$ **5**
- 4.1 The effect of the threshold operation 5
- 4.2 Further calculations 6
- 4.3 The derivative influence matrix 7

5 Ordinary Cross Validation **8**

6 Generalized Cross Validation **9**
- 6.1 Definition 9
- 6.2 Asymptotic behavoiur 10
- 6.3 Conclusion 11

7 Computational aspects **11**

8 Application to images **12**

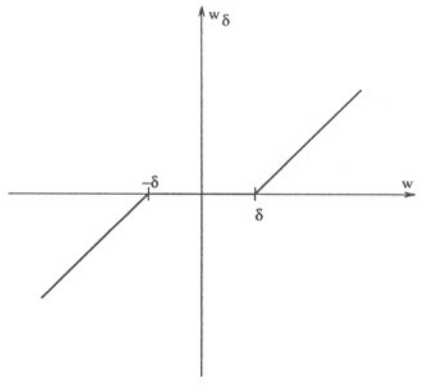


Figure 1: soft-threshold function.

1 Introduction

We consider the following model of a discrete noisy signal:

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

or, for short, in vector notation:

$$y = f + \epsilon.$$

y_i are input data and f is an unknown, deterministic signal. We suppose the noise ϵ is a stationary stochastic signal, i.e. all its values are identically distributed with zero mean and variance σ. Thus $\mathrm{E}\epsilon_i = 0$ and $\mathrm{E}\epsilon_i^2 = \sigma^2, \forall i = 1, \ldots, N$. By $h(\epsilon)$ we shall denote the density function of the noise. For the purpose of this text we restrict ourselves to uncorrelated (white) noise. This means that $\mathrm{E}\epsilon_i\epsilon_j = \delta_{ij}\sigma^2$.

To reconstruct the original data, we use a wavelet transform. This operation 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 white noise is completely spread out over all the wavelet coefficients. If noise is not too big, these arguments suggest to throw away all the wavelet coefficients, which are below a certain threshold 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 lying between $-\delta$ and δ is set to zero, while the others are diminished in absolute value.

A natural question arising from this procedure is of course how to choose the threshold δ. This choice should be optimal in a certain way. If y_δ is the result of applying the threshold procedure to signal y, then the most commonly used criterion to measure the quality of this result is its signal-to-noise ratio ($SNR(\delta)$):

$$SNR(\delta) = 10 \cdot \log_{10} \frac{\sum_i [y_{\delta i} - \overline{y}_\delta]^2}{\sum_i [\epsilon_i - \overline{\epsilon}]^2}.$$

where for an arbitrary signal f, we set $\overline{f} = \frac{1}{N}\sum_i f_i$. An optimal choice of δ should maximize $SNR(\delta)$. This is equivalent to minimizing the mean square error R:

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

where we used the classical Euclidian vector norm based on the inner product $\langle p, q \rangle = \sum_i p_i q_i$.

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

$$\delta = \sqrt{2\log(N)}\,\sigma. \tag{1}$$

This formula and other, more complicated estimators require knowledge of the noise variance σ, which may not be readily available in practical applications. Weyrich and Warhola [8] therefore suggest adapting Wahba's [7], [2] Generalized Cross Validation (GCV) 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. Because we can only rely on the given data, it seems logical that this function should express a compromise between goodness of fit and smoothness of the result.

For spline smoothing, Wahba has found a function, that provides an asymptotically optimal method. Unfortunately computations are very expensive. We shall adapt her proof to wavelet tresholding and show that the method leads to a fast algorithm. By this we try to explain the excellent results of this method, obtained by Weyrich and Warhola [8]. In a last section 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 for wavelet transforms and threshold operations

We shall not explain details of wavelet theory here. Basic theory can be found in a lot of papers. We mention Daubechies [3] B. Jawerth and W. Sweldens [5], or A. Bultheel [1]. 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:

$$v = W \cdot f,$$

$$w = W \cdot y.$$

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

$$w_\delta = D_\delta \cdot w, \tag{2}$$

where

$$D_\delta = \mathrm{diag}[d_{ii}], \tag{3}$$

with

$$d_{ii} = \begin{cases} 0 & \text{if } |w_i| < \delta, \\ 1 - \frac{\delta}{|w_i|} & \text{otherwise.} \end{cases} \tag{4}$$

δ is the choosen value of the threshold. Note that the notation $w_\delta = D_\delta \cdot w$ is deceiving because the mapping $w \mapsto w_\delta$ is nonlinear. The operator D_δ is not a matrix in the usual

sense, because its elements depend on the signal w, to which it is applied. In the same way we have:

$$v_\delta = D_\delta \cdot v.$$

Back transforms give the result:

$$y_\delta = W^{-1} \cdot w_\delta. \tag{5}$$

The entire operation can then be represented by:

$$y_\delta = A_\delta \cdot y, \tag{6}$$

where

$$A_\delta = W^{-1} \cdot D_\delta \cdot W. \tag{7}$$

We call A_δ the *influence matrix*. By D_δ, it depends on the threshold value but also on the input signal y.

For the rest of this text, we will mostly work in the wavelet domain. Because white, stationary noise is transformed to white, stationary noise, we will use the same notation ϵ_i in the wavelet domain.

Furthermore we consider the noise of the result in the wavelet domain:

$$\epsilon_\delta := w_\delta - v.$$

Because exact coefficients are also transformed by the thresholding operations, this value contains a bias. Therefore we also define:

$$\eta_\delta := w_\delta - v_\delta.$$

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. Because of the orthogonality of W, we can compute R from the wavelet coefficients as:

$$R(\delta) = \frac{1}{N}\|w_\delta - v\|^2. \tag{8}$$

The expectation of this function can then be written as:

$$\begin{aligned} \mathrm{E}R(\delta) &= \frac{1}{N}\|v_\delta - v\|^2 + \frac{1}{N}\|\mathrm{E}(w_\delta - v_\delta)\|^2 \\ &\quad + \frac{2}{N}\langle (v_\delta - v), \mathrm{E}(w_\delta - v_\delta)\rangle. \end{aligned} \tag{9}$$

The first term is due to the bias.

For linear operations, such as spline smoothing, the third term would be zero. Since shrinking is a non-linear operation, we can't use this argument. But in the case where $|v_i| \gg \delta$, it is almost sure that also $|w_i| \gg \delta$ and thus $\eta_{\delta i} := w_{\delta i} - v_{\delta i} = w_i - v_i = \epsilon_i$, and so $\mathrm{E}(w_{\delta i} - v_{\delta i}) = 0$. Only when $v_i \approx \delta$, problems occur. 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 $\eta_{\delta i} = -|v_i - \delta|$. This value of $\eta_{\delta i}$ thus has a finite probability, which can be seen on the second figure as a pulse in the

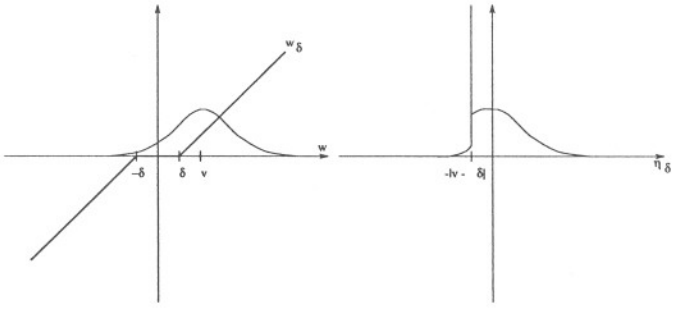


Figure 2: Left: Distribution of w, noisy wavelet coefficient and shrinking function. Right: Distribution of $\eta_\delta := w_\delta - v_\delta$

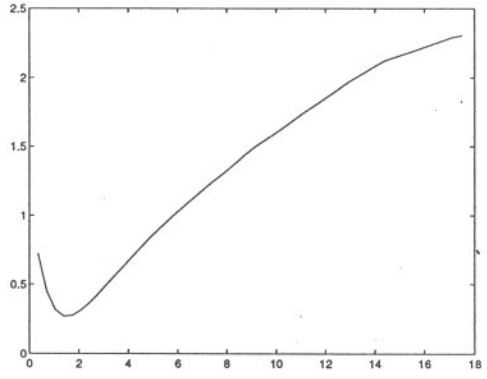


Figure 3: Typical form of $R(\delta)$.

distribution function of $\eta_{\delta i}$. This distribution clearly does not have a zero mean. Taking into account other coefficients v_i, we can assume this effect will be largely compensated at least if the wavelet signal is sufficiently symmetric.

By this we can write:

$$\mathrm{E}R(\delta) = b^2(\delta) + \mu_2(\delta) \cdot \sigma^2, \tag{10}$$

with:

$$b^2(\delta) = \frac{1}{N}\|v_\delta - v\|^2, \tag{11}$$

and:

$$\mu_2(\delta) = \frac{\|\mathrm{E}(w_\delta - v_\delta)\|^2}{N\sigma^2} \tag{12}$$

Figure 3 shows a typical form of the function $R(\delta)$. More about this function can be found in Nason's text [6].

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

4.1 The effect of the threshold operation

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

Define

$$T(\delta) = \frac{\sum_{i=1}^{N}(w_{\delta i} - w_i)^2}{N} = \frac{1}{N}\|w_\delta - w\|^2, \tag{13}$$

Since neither v or v_δ are stochastic we have $\mathrm{E}\langle \epsilon, \epsilon_\delta \rangle = \mathrm{E}\langle \epsilon, \eta_\delta \rangle$, and so we can write:

$$\begin{aligned} \mathrm{E}T(\delta) &= \mathrm{E}\frac{1}{N}\left(\|w - v\|^2 + \|v - w_\delta\|^2 + 2 \cdot \langle (w - v), (v - w_\delta) \rangle\right) \\ &= \sigma^2 + \mathrm{E}R(\delta) - \frac{2}{N}\mathrm{E}\langle \epsilon, \epsilon_\delta \rangle \\ &= \sigma^2 + \mathrm{E}R(\delta) - \frac{2}{N}\mathrm{E}\langle \epsilon, \eta_\delta \rangle. \end{aligned} \tag{14}$$

We shall investigate the third term in detail because this will lead to some essential equations. Let us therefore define:

$$\mu_1(\delta) = \frac{\sum_{i=1}^{N} \mathrm{E}[\epsilon_i \eta_{\delta i}]}{N\sigma^2}. \tag{15}$$

4.2 Further calculations

We shall prove now that:

Theorem 1 *If the density $h(\epsilon)$ is Gaussian, then*

$$\mathrm{E}[\epsilon_i \eta_{\delta i}] = \mathrm{P}(|w_i| > \delta). \tag{16}$$

First we prove a lemma:

Lemma 1 *If $h(x) = K\mathrm{e}^{\frac{-x^2}{2\sigma^2}}$ then:*

$$\forall b \in \mathbb{R} : \int_b^\infty x^2\, h(x)\, dx = b \int_b^\infty x\, h(x)\, dx + \sigma^2 \int_b^\infty h(x)\, dx.$$

Proof:

By a substitution $u = \frac{x^2}{2\sigma^2}$, one can verify that

$$\int x\, K\mathrm{e}^{\frac{-x^2}{2\sigma^2}}\, dx = -K\sigma^2 \mathrm{e}^{\frac{-x^2}{2\sigma^2}} + C.$$

And so, we have:

$$\begin{aligned} \int_b^\infty x^2\, K\mathrm{e}^{\frac{-x^2}{2\sigma^2}}\, dx &= -K\sigma^2 \mathrm{e}^{\frac{-x^2}{2\sigma^2}}\, x\Big|_b^\infty + \sigma^2 \int_b^\infty K\mathrm{e}^{\frac{-x^2}{2\sigma^2}}\, dx \\ &= b \int_b^\infty x\, K\mathrm{e}^{\frac{-x^2}{2\sigma^2}}\, dx + \sigma^2 \int_b^\infty K\mathrm{e}^{\frac{-x^2}{2\sigma^2}}\, dx, \end{aligned}$$

which proves the lemma.

For the theorem, we distinguish three cases:

1. $v_i > \delta$

   If $w_i > \delta$, then of course $\eta_{\delta i} = \epsilon_i$. For $|w_i| < \delta$, we have $\eta_{\delta i} = \delta - v_i$. We shall ignore the case $w_i < -\delta$, because this is very unlikely.

   Here we suppose that the density $h(\epsilon)$ is Gaussian and so, by our lemma:

$$\forall b \in \mathbb{R} : \int_b^\infty \epsilon^2 \, h(\epsilon) \, d\epsilon = b \int_b^\infty \epsilon \, h(\epsilon) \, d\epsilon + \sigma^2 \int_b^\infty h(\epsilon) \, d\epsilon. \tag{17}$$

   We now take $b = \delta - v_i$ to find:

$$\begin{aligned} \mathrm{E}[\epsilon_i \eta_{\delta i}] &= \int_{-\infty}^b \epsilon_i b \, h(\epsilon_i) \, d\epsilon_i + \int_b^\infty \epsilon_i^2 \, h(\epsilon_i) \, d\epsilon_i \\ &= \sigma^2 \int_b^\infty h(\epsilon_i) \, d\epsilon_i + b \int_{-\infty}^\infty \epsilon_i \, h(\epsilon_i) \, d\epsilon_i \\ &= \sigma^2 \mathrm{P}(|w_i| > \delta) + b \cdot 0 \end{aligned} \tag{18}$$

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

3. $|v_i| < \delta$

   We now call $b = \delta - v_i$ as before and $a = -\delta - v_i$ and then we have:

$$\mathrm{E}[\epsilon_i \eta_{\delta i}] = \int_{-\infty}^a \epsilon_i (\epsilon_i - a) \, h(\epsilon_i) \, d\epsilon_i + \int_a^b \epsilon_i \cdot 0 \, h(\epsilon_i) \, d\epsilon_i + \int_b^\infty \epsilon_i (\epsilon_i - b) \, h(\epsilon_i) \, d\epsilon_i.$$

   Again using (17) we obtain

$$\begin{aligned} \mathrm{E}[\epsilon_i \eta_{\delta i}] &= \sigma^2 \left(\int_{-\infty}^a h(\epsilon_i) \, d\epsilon_i + \int_b^\infty h(\epsilon_i) \, d\epsilon_i \right) \\ &= \sigma^2 \mathrm{P}(|w_i| > \delta). \end{aligned} \tag{19}$$

4.3 The derivative influence matrix

For all cases we may conclude:

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

We now introduce a new matrix:

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

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

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

Furthermore we also 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:

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

We started from $\langle \epsilon, \eta_\delta \rangle$, which, in practice, is not computable. We ended up with $\sigma^2 \mathrm{Tr}(A')$, which is very easy to find and both have the same expectation. Thus, from (14), (15) and (25) we can construct

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

as an approximation for $R(\delta)$. Unfortunately, this function requires a value for σ^2.

5 Ordinary Cross Validation

This section will introduce 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 the original signal is *regular* to some extend, which means that some 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 noise. To investigate the closeness of fit, we compute the result of the threshold operation for the modified signal $\tilde{y}$, where the ith component was replaced. So

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

For (too) small values of δ the difference $y_i - \tilde{y}_{\delta i}$ will be 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 $\tilde{y}_i$ such that $\tilde{y}_{\delta i} = \tilde{y}_i$ will turn out to be a very interesting choice. This value can always be found, since the threshold algorithm has some leveling effect. Taking $\tilde{y}_{\delta i} = \max_i y_i$, we have $\tilde{y}_{\delta i} \leq \tilde{y}_i$, while the opposite will be 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:

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

Unfortunately, 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:

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

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

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

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

6.2 Asymptotic behavoiur

In this paragraph we shall prove that for $N \to \infty$:

$$arg\min GCV(\delta) = arg\min R(\delta).$$

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

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

So, for $N \to \infty$ we can write:

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

Before going on, we look for a relation between μ_1 and μ_2. We have:

$$\begin{aligned}
\mu_2(\delta) &= \frac{\sum_{i=1}^N \mathrm{E}\eta_{\delta i}^2}{N\sigma^2} = \frac{\sum_{i=1}^N \mathrm{E}(w_{\delta i} - v_{\delta i})^2}{N\sigma^2} \\
&\geq \frac{\sum_{i=1}^N \mathrm{E}\left(\eta_{\delta i}^2 | D'_{ii} = 1\right) \cdot \mathrm{P}(D'_{ii} = 1)}{N\sigma^2}.
\end{aligned}$$

For a given δ and σ, it can be seen that

$$E\left(\eta_{\delta i}^2 | D'_{ii} = 1\right) > 0,$$

and so we set

$$\frac{\sigma^2}{C} := \min_{v_i} \mathrm{E}\left((w_{\delta i} - v_{\delta i})^2 | w_i \geq \delta\right). \tag{30}$$

Again using (24) for the Bernoulli variable D_{ii}, we then have:

$$C \cdot \mu_2(\delta) \geq \frac{\sum_{i=1}^N \mathrm{E}D'_{ii}}{N} = \frac{\mathrm{Tr}(\mathrm{E}A')}{N}.$$

By (25) this becomes:

$$\frac{\mu_1}{\mu_2} \leq C. \tag{31}$$

For common values of δ, experiments show that $\frac{\mu_1}{\mu_2}$ lies between 1 and 2.

Because $\mu_1 \leq 1$, one sees that $2\mu_1 \geq \mu_1^2$, and so:

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

with C definied in equation (30).

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:

$$1\leq\frac{\mathrm{E}R(\tilde{\delta})}{\mathrm{E}R(\delta^*)}\leq\frac{1+h(\delta^*)}{1-h(\tilde{\delta})}. \tag{32}$$

If $h(\delta)\to 0$, then $\mathrm{E}R(\tilde{\delta})\to\mathrm{E}R(\delta^*)$. It's intuitively clear that for large N indeed $h(\delta)\to 0$. To this end, it is sufficient that $\mu_1(\delta)\to 0$. But $\mu_1(\delta)$ counts the relative number of wavelet-coefficients not put to zero. In practice signals (and images) have limited bandwidths, and, at a certain N, more frequent sampling doesn't introduce much more frequencies. This causes the new coefficients to grow smaller and smaller.

6.3 Conclusion

We can conclude for $N\to\infty$:

$$\mathrm{E}R(\tilde{\delta})\approx\mathrm{E}R(\delta^*), \tag{33}$$

and in the neighbourhood of δ^*:

$$\mathrm{E}GCV(\delta)\approx\mathrm{E}R(\delta)+\sigma^2. \tag{34}$$

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 $w=W\cdot y$, using the fast wavelet transform algorithm.
2. Choose a starting threshold value.
3. Minimise $GCV(\delta)$.

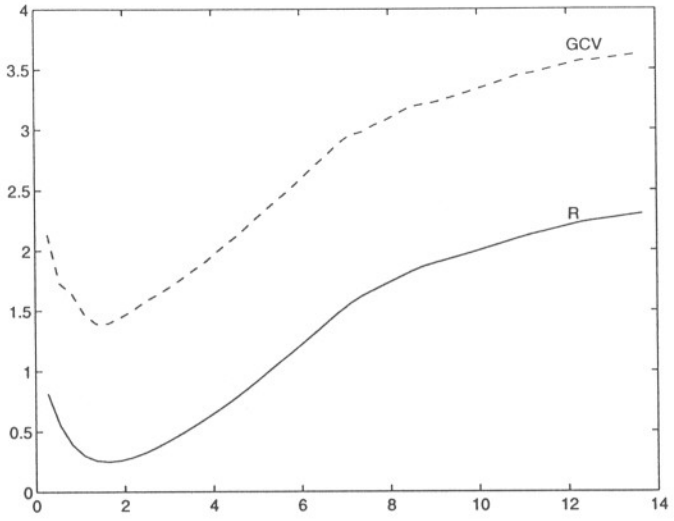


Figure 4: GCV and mean square error of the result in function of the threshold δ .

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.

We know that $0 \leq \delta^* \leq \max_i w_i$. A relative accurancy of 10^{-4} will do. Using a classic minimisation procedure (such as Fibonacci) this requires some 20 function evaluations.

Computation of $GCV(\delta)$ can be completely performed 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 put to zero. This does not require any flop. Computation of the numerator can be done with $2N$ flops. So 20 function evaluations will lead to some $40N$ flops.

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 a back transform too. This makes the minimisation procedure not too expensive, as compared with the wavelet transform.

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, is to repeat the one dimensional transform for two directions (x and y). Of course, this is not optimal, since the axes are mostly arbitrarily choosen lines which do not have any meaning for the image. This method will introduce artefacts.

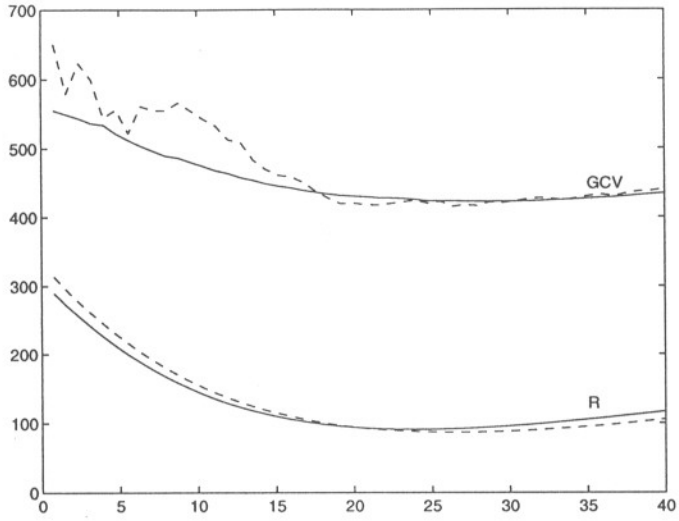


Figure 5: $GCV(\delta)$ based on all (256×256) pixels in full line, based on 1000 pixels in broken line. Idem for $R(\delta)$.

Though computation of $GCV(\delta)$ is quick, for large images time may become a problem. 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 pictured 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 fast computed GCV is a better approximation for the optimal threshold than the minimum of the 'real' GCV, which is of course a coincidence. Figure 6 shows the eventual result.

References

[1] A. Bultheel. Learning to swim in a sea of wavelets. *Bull. Belgian Mathematical Society Simon Stevin*, 2(1):1–45, 1995.

[2] P. Craven and G. Wahba. Smoothing noisy data with spline functions. *Numerische Mathematik*, 1979.

[3] 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.

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

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



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 noise image ($SNR = 10.0$dB, on the right the result ($SNR = 15.5$dB).

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

[7] 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.

[8] 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*, pages 523–532, 1995.