Wavelet thresholding using generalized cross validation

Maarten Jansen, Adhemar Bultheel

October 1997

Abstract

De-noising algorithms based on wavelet thresholding replace small wavelet coefficients by zero and keep or shrink the coefficients with absolute value above the threshold. The optimal threshold minimizes the error of the result as compared to the unknown, exact data. To estimate this optimal threshold, we use Generalized Cross Validation. This procedure is fast and does not require an estimation for the noise energy. Moreover, the method is shown to be asymptotically optimal.

In its original form,this method assumes uncorrelated noise and orthogonal wavelet transforms. It turns out that extension to biorthogonal transforms and correlated noise is possible by applying level-dependent thresholds. We also present experiments with integer wavelet transforms, and illustrate the method with some results in image de-noising.

keywords: noise reduction, wavelets, threshold, generalized cross validation

1 Introduction

Wavelet thresholding is a straightforward though effective method in non-parametric regression problems. It is based on the assumption that wavelet coefficients are a parsimonious representation of "regular" data: a limited number of large wavelet coefficients captures all important signal features, such as discontinuities, whereas the input noise is spread out evenly over all coefficients.

The extensive literature on wavelet thresholding mainly focusses on two questions. The first is which threshold rule to use. Most popular are hard- and soft-thresholding. The second, and perhaps most important, issue is the selection of an appropriate threshold. The "universal threshold" by Donoho

and Johnstone [6]: is widely known and essentially takes a threshold value proportional to the amount of noise. The "SURE"-threshold [7] minimizes an unbiased estimation of risk. The approach of Ogden and Parzen [15] is based on standard statistical testing procedures: it chooses a threshold such that no significant signal coefficients can be detected in the set of coefficients with absolute value below the threshold. Nason [13], and Weyrich and Warhola [22] describe cross validation procedures. Bayesian procedures [1, 17, 20] start from a statistical model for noisy *and* noise-free coefficients.

This paper discusses the possibilities of a generalized cross validation [21, 5] method. Weyrich and Warhola introduce it in a wavelet threshold procedure and obtain nice results. Asymptotical optimality properties [9] explain these results. For the greater part, the results of this paper summerize findings of previous papers [9, 8].

Section 2 introduces the method of wavelet thresholding and formulates the problem of threshold estimation. In Section 3, we propose the method of GCV for the case of uncorrelated, stationary, additive Gaussian noise. Section 4 investigates some modifications for more general correlated noise models. We end with some illustrations in image de-noising and conclusions.

2 Model, problem, and objective

Consider the following additive model for discrete data with noise:

$$\boldsymbol{y} = \boldsymbol{f} + \boldsymbol{\varepsilon}. \tag{1}$$

$\boldsymbol{f}$ represents an unknown, "clean" data vector and $\boldsymbol{y}$ is the input for our algorithm. The noise $\boldsymbol{\varepsilon}$ is *stationary*. In principle, this means that all components of the vector should have the same distribution. In practice, we want the mean $\mathrm{E}\varepsilon_i$ and the variance $\sigma^2 = \mathrm{E}\varepsilon_i^2$ to be constants, i.e. independent of i. This is *second order stationarity*. In the first instance, we also assume that the noise is *white* or *uncorrelated*: $\mathrm{E}\varepsilon_i\varepsilon_j = \delta_{ij}\sigma^2$.

The algorithm starts with a wavelet transform. By the linearity of this operation, the model in terms of wavelet coefficients remains unchanged:

$$\boldsymbol{w} = \boldsymbol{v} + \boldsymbol{\omega}.$$

The vector $\boldsymbol{v} = W\boldsymbol{f}$ contains the wavelet coefficients of the original data, $\boldsymbol{\omega} = W\boldsymbol{\varepsilon}$ are the noise coefficients and $\boldsymbol{w} = W\boldsymbol{y}$. Only for orthogonal W the wavelet decomposition of the noise $\boldsymbol{\omega}$ is stationary and white.

After applying a soft threshold δ, we get the modified wavelet coefficients $\boldsymbol{w_\delta}$. The inverse wavelet transform then yields the restored data $\boldsymbol{y_\delta}$. The

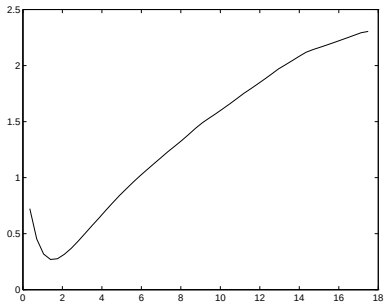


Figure 1: Wavelet thresholding for noisy images: mean square error of the result as a function of the threshold. If the threshold is too small, too much noise remains in the result. A large threshold deforms too much the original image (bias). The optimal threshold minimizes the mean square error.

optimal threshold minimizes the mean square error function $R(\delta)$ of the result, where

$$R(\delta) = \frac{1}{N}\|\boldsymbol{y_\delta} - \boldsymbol{f}\|^2. \tag{2}$$

Orthogonal wavelet transforms allow to express this function in terms of wavelet coefficients. In the remainder of this paper, we always minimize in "wavelet-domain", whether the transform is orthogonal or not. This error function has a typical form, as depicted in Figure 1. A threshold equal to zero leaves the input unchanged. These input coefficients are unbiased estimates for the underlying, unknown, true coefficients, but they are very noisy! A higher threshold reduces the variance at the expense of an increasing bias. At a certain moment this increase of bias exceeds the reduction of variance, and the result gets over-smoothed. The expectation $\mathrm{E}R(\delta)$ is called *Risk* function.

However, the noise-free data are unknown, and so the optimal threshold cannot be computed exactly.

3 The principle of Generalized Cross Validation

To estimate the risk, we apply the Generalized Cross Validation (GCV) function [21, 5]. This is a formalized generalization of a leaving-out-one operation: to validate the performance of a smoothing parameter at a given data point y_i, we leave out this point and apply the smoothing procedure on the remaining data. We interpolate this intermediate result in the missing point and call this interpolation $\tilde{y_i}$. The quality of $\tilde{y_i}$ as a prediction for

y_i is then considered as a measure for the optimality of the smoothing parameter. For more technical details, we refer to [22, 9]. In our terminology, smoothing corresponds to wavelet thresholding, the smoothing parameter is the threshold, and the input data are wavelet coefficients.

Formalizing and generalizing this procedure leads to the GCV-function, which for wavelet thresholding reduces to:

$$GCV(\delta) = \frac{\frac{1}{N}\|\boldsymbol{y} - \boldsymbol{y_\delta}\|^2}{[\frac{N_0}{N}]^2}, \tag{3}$$

where N is the total number of wavelet coefficients and N_0 the number of these coefficients that were replaced by zero.

This GCV-function has following properties:

1. As illustrated in Figure 2 for a typical case, $GCV(\delta)$ and $R(\delta)$ have approximately the same form. This is due to the *asymptotic* properties of GCV (see below). In practice, a thousand coefficients are necessary to guarantee a successful application of GCV.

2. Unlike for some other applications, the use of GCV for wavelet thresholding leads to a simple and fast algorithm: minimization of $GCV(\delta)$ is of linear computational complexity.

3. No explicite estimate for the noise variance is necessary.

In [9] we prove that the minimizer of $GCV(\delta)$ is an asymptotically optimal threshold choice, i.e.:

Theorem 1 *If* $\delta^* = arg\min R(\delta)$ *and* $\tilde{\delta} = arg\min GCV(\delta)$, *then for* $N \to \infty$*:*

$$\frac{\mathrm{E}R(\tilde{\delta})}{\mathrm{E}R(\delta^*)} \downarrow 1, \tag{4}$$

and in the neighbourhood of δ^**:*

$$\mathrm{E}\,GCV(\delta) \approx \mathrm{E}R(\delta) + \sigma^2. \tag{5}$$

Unlike spline smoothing, wavelet thresholding is a non-linear operation. This has an influence on the proof, since it makes computation of expectations more difficult. The proof makes use of Stein's unbiased risk estimation [19, 7].

Furthermore, the proof imposes some restrictions on data and noise:

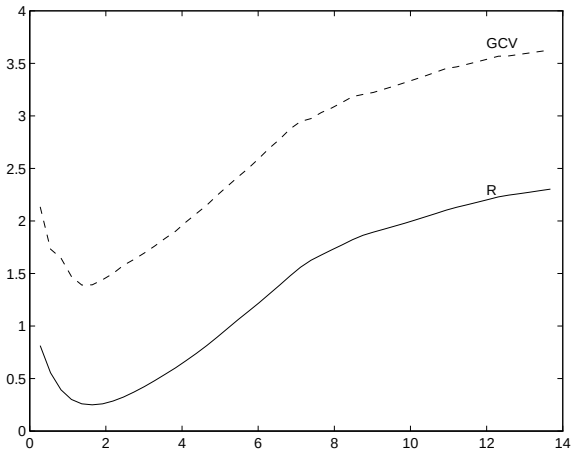


Figure 2: GCV and mean square error of the result in function of the threshold δ. The input was an image with artificial, white, and stationary noise.

1. The original data $\boldsymbol{f}$ are *smooth* in the sense that they can be represented parsimoniously by taking a wavelet transform. In fact, this assumption justifies the use of wavelets, since the localizing properties of these basis functions guarantee such a compact representation for most inputs. Without this assumption, the wavelet transform would not be necessary.

2. The noise in wavelet domain should be *stationary*. This is necessary not only to find an estimate for the optimal threshold with GCV but also to reduce noise decently with wavelet thresholding anyway: one threshold cannot perform well on a coefficient with a large noise variance *and* on another coefficient with little noise at the same time. Stationarity in wavelet coefficients requires:

   (a) stationary input noise,

   (b) white input noise,

   (c) and an orthogonal wavelet transform.

   An orthognal transform of uncorrelated, stationary noise is uncorrelated and stationary. The fact that the coefficient noise is uncorrelated is not necessary. However, if one of the above requirements is violated, the coefficient noise cannnot be guaranteed to be stationary.

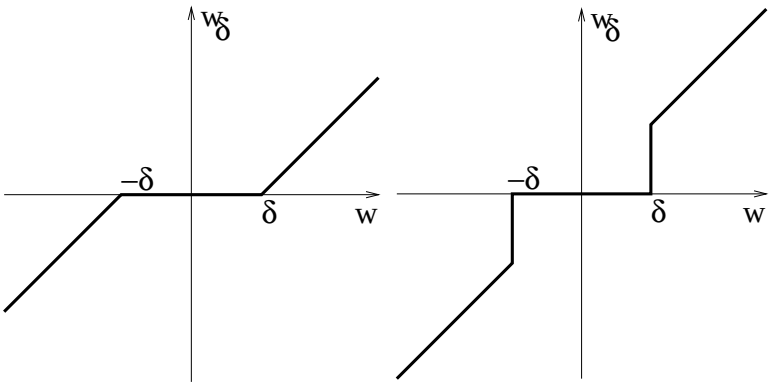


Figure 3: Noise reduction by wavelet thresholding. On the left: Soft-thresholding: a wavelet coefficient w with an absolute value below the threshold δ is replaced by 0. Coefficients with higher absolute values are shrunk. On the right: Hard-thresholding: Coefficients with an absolute value above the threshold are kept. This operation is not *continuous*.

3. The noise should be Gaussian with zero mean. Experiments showed that, in practice, the GCV-method performs well for other zero mean stationary distributions of the noise.

4. We use soft-thresholding. The method fails in the hard-thresholding case, because this operation is not continuous, as illustrated in Figure 3.

4 Extensions

4.1 Level-dependent thresholds

It is clear that one threshold cannot perform well for images with heteroscedastic (non-stationary) noise. Indeed, the wavelet transform of such noise is not stationary either and a "good" threshold always depends — explicitly or implicitly — on the amount of noise. If this amount differs for all coefficients, one common threshold is not able to remove the noise decently.

For images with correlated noise, the correlation itself is no problem, but the wavelet transform of such a noise is not stationary either. However, it is easy to prove that a wavelet transform of stationary, correlated noise is stationary *within each resolution level* [10, 8]. Note that, more generally for one-dimensional data, the correlation matrix of the wavelet transform equals

the two-dimensional, rectangular wavelet transform of the correlation matrix of the input data. Silverman [18] describes a fast computation for banded correlation matrices.

Using this observation, we can construct a GCV-estimate for a threshold at each resolution level. As mentioned before, this procedure requires no value or estimate for the noise standard deviation. In this case of correlated noise, where we need a different threshold at each level, this advantage becomes more pronounced. Indeed, for data with white and stationary noise, an orthogonal wavelet transform yields stationary noise and typically, the noise variance can be estimated from the wavelet coefficients at the highest (finest) resolution. This level has most wavelet coefficients and only a few of them contain important data: at this level, most of the coefficients are "almost" purely noise.

4.2 Non-decimated wavelet transforms

All conditions for a successful application of GCV remain satisfied if we use a non-decimated wavelet transform and level-dependent thresholds to remove stationary, correlated noise. The non-decimated or stationary [4, 16, 11, 14, 12, 8] wavelet transform is translation-invariant. It has the same number of coefficients at all resolution levels. This number is equal to the size of the original input data. In this way, all levels have sufficient coefficients to justify the use of asymptotic procedure, such as GCV. A non-decimated wavelet transform is a redundant representation of the input. Several independent reconstructions are possible. A combination of these reconstruction schemes, after thresholding, causes additional smoothing in the result.

4.3 Integer wavelet transforms

Integer wavelet transforms are invertible transforms mapping integer input values onto integer wavelet coefficients [2]. This transform is non-linear, due to rounding in each resolution step. Therefore no assumption can be made about the noise on the wavelet coefficients: it is neither stationary, nor additive. In spite of these theoretic problems, application of GCV for integer wavelet thresholding seems to work in practice (see below for an example). All calculations can be done with integers only, which may be usefull, for instance, if one aims a DSP-implementation.



Figure 4: Left: a clean test-image. Right: the same image with artificial, additive and correlated noise. The noise is the result of a convolution of white noise with a FIR highpass filter. Signal-to-noise ratio is 4.97 dB.

5 Examples in image de-noising

We illustrate the method with some experiments on noisy images. Figure 4 shows the original image and the image with noise. The signal-to-noise (SNR) ration is 4.97 dB, where we define SNR as:

$$SNR = 20 \log_{10} \frac{\|\boldsymbol{f}\|}{\|\boldsymbol{\varepsilon}\|}. \tag{6}$$

Figure 5 shows the result of the level-dependent threshold procedure applied on the fast wavelet transform of the noisy image. Signal-to-noise ratio is now 15.87 dB, but the result shows artifacts. The same figure also shows the result for the redundant wavelet transform. It illustrates the smoothing effect of this alternative: most of the artifacts have disappeared and the SNR-value increases to 19.01 dB. For this example, we used biorthognal wavelets from the Cohen-Daubechies-Feauveau family [3] with two vanishing moments for primal and dual basis. A fast integer algorithm, based on the same filters, yields the result in Figure 6.

6 Conclusion

This paper has presented an automatic selection of level-dependent thresholds for wavelet coefficients, to remove stationary, correlated noise. Most existing algorithms use an estimation of the noise energy at all scales to



Figure 5: Result after applying the level-dependent threshold procedure on the wavelet transform of the noisy image. Left: the fast wavelet transform. Signal-to-noise ratio is 15.87 dB. Right: redundant wavelet transform. Signal-to-noise ratio is 19.01 dB.



Figure 6: Result after applying the level-dependent threshold procedure on a fast integer wavelet transform of the noisy image. Signal-to-noise ratio is 15.52 dB.

choose these thresholds. Since we need one threshold for each scale, the benefits of an automatic selection of these thresholds are still more pronounced than for the choice of one threshold. However, our selection is based on an asymptotical method, which fails at coarse scales, since only a few wavelet coefficients correspond to these levels. Therefore, we introduce the redundant wavelet transform. This alternative provides, at the expense of more computation time and higher memory requirements, sufficiently coefficients at all scales.

References

[1] F. Abramovich, F. Sapatinas, and B. W. Silverman. Wavelet thresholding via a bayesian approach. *Journal of the Royal Statistical Society, Series B*, 58, 1997.

[2] R. Calderbank, I. Daubechies, W. Sweldens, and B.-L. Yeo. Wavelet transforms that map integers to integers. Technical report, Department of Mathematics, Princeton University, 1996.

[3] A. Cohen, I. Daubechies, and J. Feauveau. Bi-orthogonal bases of compactly supported wavelets. *Comm. Pure Appl. Math.*, 45:485–560, 1992.

[4] R. R. Coifman and D. L. Donoho. Translation-invariant de-noising. In A. Antoniadis and G. Oppenheim, editors, *Wavelets and Statistics*, Lecture Notes in Statistics, pages 125–150, 1995.

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

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

[7] D. L. Donoho and I. M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.*, 90:1200–1224, 1995.

[8] M. Jansen and A. Bultheel. Multiple wavelet threshold estimation by generalized cross validation for data with correlated noise. TW Report 250, K.U.Leuven, Department of Computer Science, Leuven, Belgium, January 1997.

[9] M. Jansen, M. Malfait, and A. Bultheel. Generalized cross validation for wavelet thresholding. *Signal Processing*, 56(1):33–44, January 1997.

[10] I. M. Johnstone and B. W. Silverman. Wavelet threshold estimators for data with correlated noise. *Journal of the Royal Statistical Society, Series B*, 59:319–351, 1997.

[11] G. Lang, H. Guo, C. S. Burrus, and R. O. Wells. Noise reduction using an undecimated discrete wavelet transform. *IEEE Signal Processing Letters*, 3(1):10–12, 1996.

[12] M. Malfait and D. Roose. Wavelet based image denoising using a markov random field a priori model. *IEEE Transactions on Image Processing*, 6(4):549–565, 1997.

[13] G. P. Nason. Wavelet shrinkage using cross validation. *Journal of the Royal Statistical Society, Series B*, 58:463–479, 1996.

[14] G. P. Nason and B. W. Silverman. The stationary wavelet transform and some statistical applications. In A. Antoniadis and G. Oppenheim, editors, *Wavelets and Statistics*, Lecture Notes in Statistics, pages 281–299, 1995.

[15] T. Ogden and E. Parzen. Data dependent wavelet thresholding in non-parametric regression with change-point applications. *Computational Statistics and Data Analysis*, 22:53–70, 1996.

[16] J.-C. Pesquet, H. Krim, and H. Carfantan. Time invariant orthonormal wavelet representations. *IEEE Transactions on Signal Processing*, 44(8):1964–1970, 1996.

[17] F. Ruggeri and B. Vidakovic. A bayesian decision theoretic approach to wavelet thresholding. Preprint 95-35, Duke University, Durham, NC, 1995.

[18] B. Silverman. Wavelet regression: Relaxing the assumptions. Presented at workshop on wavelets and statistics, Duke University, October 1997.

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

[20] B. Vidakovic. Nonlinear wavelet shrinkage with Bayes rules and Bayes factors. Preprint 94-24, Duke University, Durham, NC, 1994.

[21] 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.

[22] 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.