A SOFT THRESHOLDING APPROACH FOR MDL DENOISING

Janne Ojanen and Jukka Heikkonen

Laboratory of Computational Engineering, Helsinki University of Technology (TKK)
P.O.Box 9203, FI-02015 TKK, FINLAND
phone: +358 9 451 4837, fax: +358 9 451 4833, email: janne.ojanen@tkk.fi
web: http://www.lce.hut.fi/

ABSTRACT

The existing MDL method for wavelet denoising is extended with a soft thresholding approach. We assume that the wavelet coefficients are comprised of an informative part and a noise part. We propose a soft thresholding method based on the earlier MDL hard thresholding approach equivalent to fitting two Gaussian density functions to the wavelet coefficients, one for the informative part in the data and the other for noise. Our approach is data-dependent and since it is completely characterized by the properties of the MDL hard thresholding solution, it does not require any additional parameters to be estimated. We show that our method improves the results of the existing MDL denoising method for both artificial and natural test signals.

1. INTRODUCTION

Wavelets are widely utilized in several statistical signal processing research areas [1]. During the past decade wavelets have gained steady popularity in denoising, the task of removing uninformative noise from signals. The popularity of wavelets in denoising is largely due to the sparsity of the wavelet representation of data and the computationally efficient algorithms. By sparsity it is meant that the distribution of the wavelet coefficients is highly concentrated: the large majority have very small values while only a small subset of coefficients have large values [2].

Probably the best known wavelet-based denoising methods are those proposed by Donoho and Johnstone [3, 4]. These methods are based on hard thresholding, in which the coefficients with magnitudes below a certain threshold are set to zero, while the remaining coefficients are retained unmodified. The methods proposed by Donoho and Johnstone aim at minimizing the worst-case risk, and they have been shown to be minimax optimal over a large class of functions. The Bayesian approach in wavelet denoising is based on minimizing the expected risk, with the expectation taken over a postulated prior distribution supposedly governing the underlying true signal [5, 6]. An example of a prior distribution with a very good experimental performance is the generalized Gaussian density [7].

A very different approach to wavelet denoising is based on the Minimum Description Length (MDL) principle [8, 9, 10]. The MDL principle can be employed in denoising problems by defining noise to be that part in the data that cannot be compressed with the given model class. In other words, noise is defined to be the part in the data in which the given model class cannot find any regular features. Ideally, this definition of noise does not include any ad hoc assumptions of the noise distribution, even though a Gaussian noise model is usually assumed. Although several different MDL denoising methods have been proposed [11, 12, 13], this paper concentrates on the approach originally suggested by Rissanen [14] and further developed by Roos et al. [15, 16]. This method has been shown to perform well, and it is considered to be the most theoretically rigorous MDL denoising approach.

Denoising problem can be stated formally as follows. The observed 1-D signal is represented as a real-valued column vector

\[ x = (x_1, \ldots, x_n)^T \]

of length \( n \). Besides the obvious 1-D data, this signal model can also be extended, for example, to 2-D image data. By defining an \( n \times m \) regressor matrix \( W \), whose columns are the basis vectors, the signal \( x \) can be written as a linear combination of the basis vectors weighted with a coefficient vector \( \beta = (\beta_1, \ldots, \beta_m)^T \) and Gaussian i.i.d. noise,

\[ x = W \beta^0 + e, \]

where the elements of \( e \) are i.i.d. Gaussians with a common variance \( \sigma_e^2 \), \( e_i \sim N(0, \sigma_e^2) \). The common convention, which we also adopt, is to restrict the regressor matrix \( W \) in such a manner that the basis vectors form a complete orthonormal basis. This implies that the basis vectors are orthogonal unit vectors and the number of the basis vectors is equal to the size of the data, \( m = n \). Therefore the regressor matrix becomes \( n \times n \), and its inverse is given by the transpose, \( W^{-1} = W^T \). This restriction is satisfied with a number of generally used wavelet transforms, such as the Daubechies transforms. The transform \( W^T x \) is referred to as the wavelet transform and \( W \beta^0 \) as the inverse wavelet transform. The orthonormal transform preserves the Euclidean norm, so that the Parseval’s equality \( ||\beta^0|| = ||W\beta^0|| \) holds. Moreover, the orthonormality of the transform implies that spherically symmetric densities, such as Gaussians, are invariant under the transform.

The optimal solution to the denoising problem is to obtain a coefficient vector \( \hat{\beta} \) giving the signal estimate \( \hat{x} = W \hat{\beta} \) consisting of the informative part in the data, while the difference \( x - \hat{x} \) is considered as noise. The conventional maximum likelihood result for the optimal coefficient vector \( \hat{\beta} = W^T x \) is not applicable, since it gives a reconstruction equal to the original signal \( x \). In most hard thresholding methods a threshold optimal in some sense is calculated, and all the coefficients with larger magnitudes than the threshold are retained while the rest are set to zero. Typically the set of retained coefficients is relatively small. The soft thresholding methods remove the coefficients with magnitudes smaller than the threshold, but also shrink the retained coefficients in order to decrease the effect of noise. The rationale in soft thresholding is that a single wavelet coefficient is not solely representing either noise or information, but is a combination of a relatively small magnitude and some information. The soft thresholding method removes those coefficients.
of both.

The MDL solution proposed by Rissanen [14] is to consider each possible subset of basis vector separately and choose the one giving the smallest description length defined by the normalized maximum likelihood (NML) code length. Here we propose a new denoising method with improved performance by extending the original hard thresholding MDL method with a soft thresholding like modification.

The paper is structured as follows. The MDL approach in wavelet denoising is discussed in Sec. 2. In Sec. 3 we propose a soft thresholding like extension to the MDL denoising framework. In Sec. 4 the MDL soft thresholding denoising method is shown to improve the performance in denoising artificial test signals and natural images. The conclusions are presented in Sec. 5.

2. MDL DENOISING

The derivation of the NML model and the corresponding code length in [14] requires an evaluation of an integral undefined unless the range of integration is restricted. The restriction of range introduces hyperparameters, which are removed by a renormalization procedure resulting in a renormalized maximum likelihood model. Recently Roos et al. [15, 16] gave a different interpretation of the method, based on fitting a Gaussian density function to the informative coefficients and another to the non-informative coefficients. This interpretation gives equivalent denoising criterion while avoiding the cumbersome renormalization procedure. The following description of this derivation closely follows [16].

Begin by considering a fixed subset of the coefficient indices, $\gamma \subseteq \{1, \ldots, n\}$. The coefficients $\beta_i$, $i \in \gamma$, are modeled as independent outcomes from a Gaussian distribution with variance $\gamma$. In the hard thresholding approach all coefficients $\beta_i$, $i \notin \gamma$, are set to zero. This results in the model

$$x^n = W^T \beta^n + \epsilon^n, \quad \{ \begin{array}{l} e_i \sim N(0, \sigma_i^2), \\
\beta_i \sim N(0, \gamma) \end{array}, \quad \text{if } i \in \gamma, \beta_0 = 0, \quad \text{otherwise.}$$

The wavelet representation of the observed signal $x^n$ is

$$x^n = W^T e^n = W^T \beta^n + W^T \epsilon^n,$$

where $e^n$ are the observed wavelet coefficients and $W^T \epsilon^n$ gives the noise in wavelet domain. For this model the optimal parameters $\hat{\beta}_n$ are the maximum likelihood parameters

$$\hat{\beta}_i = \{ \begin{array}{l} c_i, \quad \text{if } i \in \gamma \\
0, \quad \text{otherwise.} \end{array}$$

Because the wavelet transform is orthonormal $W^T \epsilon^n$, the transform of the i.i.d. Gaussian distribution for the noise $e^n$, is also a Gaussian with the same variance $\gamma$. Each observed wavelet coefficient $c_i$ has a distribution defined by the sum of two Gaussian random variables. In the case $i \notin \gamma$ the density is simply the noise density, whereas for $i \in \gamma$ it is a Gaussian with variance $\sigma_i^2 = \sigma + \gamma$. This results in the model for the observed wavelet coefficients

$$x^n = W^T e^n, \quad c_i \sim \{ \begin{array}{l} N(0, \sigma_i^2), \quad \text{if } i \in \gamma \\
N(0, \sigma_i^2), \quad \text{otherwise.} \end{array}$$

Denoising can now be viewed as a task of choosing the subset $\gamma$: each wavelet coefficient belong either to the set of informative coefficients with variance $\sigma_i^2$ or to the set of non-informative coefficients with variance $\sigma_i^2$. The MDL principle states that the best subset is the one minimizing the description length for the observed data. The NML density can now be defined as

$$f_{\text{NML}}(x^n; \gamma) = \frac{f(x^n; \hat{\sigma}_1^2, \hat{\sigma}_2^2)}{C_\gamma},$$

where $\hat{\sigma}_1^2 = \frac{1}{n} \sum_{i=1}^n c_i^2$ and $\hat{\sigma}_2^2 = \frac{1}{n-k} \left( \sum_{i=1}^n c_i^2 - \sum_{i=1}^n \hat{\sigma}_1^2 \right)$ are the maximum likelihood estimates for the variances, $k$ is the number of coefficients for which $i \in \gamma$ (the non-zero elements in $e^n$) and $C_\gamma = \int_{\mathbb{R}^n} f(z^n; \hat{\sigma}_1^2, \hat{\sigma}_2^2) dz^n$ is the normalizing constant, also known as the parametric complexity of the model class defined by the structure index $\gamma$. Roos et al. [16] show that the code length corresponding to the negative logarithm of the NML density function in Eq. (3) is approximated by

$$\frac{n-k}{2} \ln \frac{\sum_{i=1}^n c_i^2 - \sum_{i=1}^n \hat{\sigma}_1^2}{n-k} + \frac{k}{2} \ln \frac{\sum_{i=1}^n \hat{\sigma}_1^2}{k} + \frac{1}{2} \ln k(n-k).$$

Incidentally, this is exactly the same form as proposed earlier by Rissanen [14].

Roos et al. [16] have also shown that in regression problems where the number of different possible models is very large, such as $2^n$ in denoising, the approximation in Eq. (4) may not be sufficient and the model index must also be encoded to correct the criterion. Their encoding scheme for the locations of the $k$ retained informative coefficients results in an additional code length term $\ln \binom{n}{k}$ giving the denoising criterion

$$\frac{n-k}{2} \ln \frac{\sum_{i=1}^n c_i^2 - \sum_{i=1}^n \hat{\sigma}_1^2}{n-k} + \frac{k}{2} \ln \frac{\sum_{i=1}^n \hat{\sigma}_1^2}{k} + \frac{1}{2} \ln k(n-k) + \ln \binom{n}{k}.$$

Minimizing the criterion in Eq. 3 gives the optimal set of wavelet coefficients $e^n$. In this paper this method is referred to as the MDL hard thresholding method. Fig. 1 shows an example of the empirical distribution of the wavelet coefficients and the two fitted Gaussian densities with variances resulting from minimizing the MDL criterion.

3. SOFT THRESHOLDING FOR MDL DENOISING

Although the MDL hard thresholding method described in Sec. 2 has been shown to work well, it suffers from the same drawbacks as the other hard thresholding approaches. The hard thresholding methods typically retain a very small number of coefficients, and the results are often overly smoothed. While this might seem to give good results with respect to a certain error measure, the visual quality of the denoised signals is often not so good. The view that the observed wavelet coefficients correspond to either informative signal or noise taken in the hard thresholding methods is very strict. The soft thresholding methods are based on idea that the coefficients have contributions from both the informative signal and noise, which is also assumed in Eq. 1, so that shrinking the retained coefficients attempts to attenuate the effects.
of noise. Typically in image data the smaller wavelet coefficients consist not only of noise, but also of important image details such as edges. By retaining a slightly larger amount of coefficients and shrinking them the soft thresholding methods usually give better results than the hard thresholding methods.

However, it has proven to be difficult to combine soft thresholding with the selection of model class according to shortest description length in a theoretically rigorous ways. An example of one such attempt can be found in [16]. Since incorporating the soft thresholding approach into the MDL denoising framework would be very beneficial especially when denoising natural images, we describe an useful approach which can be easily employed in connection with the MDL hard thresholding denoising method. Although our method is not theoretically pure MDL, it is an intuitive extension following from the interpretation involving the separate Gaussian densities for the non-informative and the informative coefficients.

The MDL hard thresholding method gives an optimal division of the coefficients into informative and non-informative sets and also gives the variances of their distributions. Our aim is to find out the proportion of the informative signal density function the weight tends to zero, while for the large coefficients the curve approaches the diagonal.

With the weight vector \( \mathbf{w}^n = (w_1, \ldots, w_n)^T \) given in Eq. (5) we can now compute the modified coefficients

\[
\tilde{c}_i = w_i c_i, \quad i = 1, \ldots, n
\]

whether they were considered by the MDL hard thresholding method to be informative signal or noise. In Fig. 2 the typical behavior of the weighted coefficients is shown: when the noise density function has significantly larger values than the informative signal density function the weight tends to zero, and in the opposite case the weight tends to one.

4. EXPERIMENTAL RESULTS

The effect of the proposed soft thresholding extension to the original MDL hard thresholding method was studied with a set of artificial 1-D signals scaled for the range of 200 and 8-bit grayscale natural images with a range of 255. The signals were corrupted with Gaussian random noise with known variance and the denoised signals were compared with the originals. The error was measured with the peak-signal-to-noise ratio (PSNR) defined as

\[
\text{PSNR} = 10 \log_{10} \left( \frac{\max(x^n) - \min(x^n))^2}{\text{MSE}} \right),
\]

where the squared range is calculated from the signal \( x^n \) and MSE is the mean squared error. While the PSNR error measure is useful in comparing the effects of different methods, the visual quality is also very important, especially for the images. The Daubechies db6 wavelet basis was used in all experiments. For comparison we also give results for two generally used soft thresholding denoising methods, BayesShrink [7] and VisuShrink [3].

Fig. 4 illustrates the denoising results for the 'Lena' image with noise standard deviation \( \sigma = 30 \). The best method with respect to PSNR is BayesShrink. MDL hard thresholding and VisuShrink give visually poor results. Although MDL soft thresholding has slightly lower PSNR than

![Two Gaussian densities fitted to noisy 'Lena' image](image1)

Figure 1: Two Gaussian densities fitted to noisy 'Lena' image (noise standard deviation \( \sigma = 10 \)). The empirical histogram of wavelet coefficients is drawn in gray. The Gaussian density for the noise coefficients (\( \hat{c}_n = 12.0 \)) is drawn in solid line, and the Gaussian density for the informative coefficients (\( \hat{c}_f = 398.6 \)) in dotted line.

![Behavior of the proposed soft thresholding approach](image2)

Figure 2: The behavior of the proposed soft thresholding approach on the noisy 'Lena' whose observed coefficient histogram and the fitted Gaussian densities were presented in Fig. 1. The solid line shows \( \tilde{c}_i \) plotted against original \( c_i \).
BayesShrink, the visual quality of the denoising result is almost as good. Moreover, the MDL soft thresholding approach gives clearly better results than the MDL hard thresholding method.

For the 1-D test signals the results are similar in the PSNR sense: BayesShrink consistently performs well, although the MDL soft threshold gives better results in some cases with low noise levels. However, the quality of the MDL soft threshold results decreases as the noise level increases due to an increasing amount of noise being retained in the denoising results. The denoising results for the 'Blocks' test signal are presented in Fig. 4.

These observations hold with other signals and images as well, as seen in Table 1. The MDL soft thresholding method improves the denoising results of the original MDL hard thresholding method especially in the low noise variance region. However, when the noise variance increases, increasing amounts of noise is left in the denoised signals, resulting in a similar PSNR levels. The good performance of BayesShrink is no surprise, since it is considered to be one of the most efficient wavelet denoising approaches [18]. The results in Table 1 show that with low noise levels the MDL soft thresholding method performs as well as BayesShrink.

5. CONCLUSIONS

We have extended an existing MDL hard thresholding wavelet denoising method with a soft thresholding like modification. The MDL soft thresholding method was based on an interpretation of the original MDL hard thresholding denoising method in which two Gaussian densities, one corresponding to the informative part and the other to noise, are fitted to the wavelet coefficients. The MDL soft thresholding modification was shown to improve the denoising results of the original MDL denoising method with artificial test signals and natural images, and in some cases the results were as good as those obtained with BayesShrink.

The performance of the proposed method was seen to decline with high noise variance. A possible approach in enhancing the denoising performance with high noise levels is to consider a subband-wise approach taken in many denoising methods such as BayesShrink. By extending the MDL soft thresholding method into a subband-adaptive method the denoising performance would most probably improve even in the high noise region, although this requires extending the
Table 1: Numerical simulation results. The peak-signal-to-noise ratio for various images and 1-D signals, denoising methods and noise standard deviations. First column shows the noise standard deviation, and the rest of the columns give the average PSNR for different methods. Average is taken over 15 repetitions. 'MDL hard' refers to the MDL hard thresholding method, 'MDL soft' to the MDL soft thresholding method, 'Bayes' to BayesShrink and 'Visu' to VisuShrink.

<table>
<thead>
<tr>
<th></th>
<th>Lena</th>
<th>Mandrill</th>
<th>Blocks</th>
<th>Bumps</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MDL hard</td>
<td>MDL soft</td>
<td>Bayes</td>
<td>Visu</td>
</tr>
<tr>
<td>σ = 10</td>
<td>31.1</td>
<td>33.1</td>
<td>33.5</td>
<td>27.8</td>
</tr>
<tr>
<td>σ = 20</td>
<td>28.1</td>
<td>29.4</td>
<td>30.3</td>
<td>25.1</td>
</tr>
<tr>
<td>σ = 30</td>
<td>26.4</td>
<td>26.7</td>
<td>28.6</td>
<td>23.7</td>
</tr>
<tr>
<td>σ = 40</td>
<td>25.3</td>
<td>24.5</td>
<td>27.5</td>
<td>22.7</td>
</tr>
<tr>
<td>σ = 10</td>
<td>22.8</td>
<td>26.4</td>
<td>28.7</td>
<td>21.3</td>
</tr>
<tr>
<td>σ = 20</td>
<td>21.2</td>
<td>24.3</td>
<td>25.0</td>
<td>20.1</td>
</tr>
<tr>
<td>σ = 30</td>
<td>20.4</td>
<td>22.7</td>
<td>23.1</td>
<td>19.5</td>
</tr>
<tr>
<td>σ = 40</td>
<td>19.9</td>
<td>21.4</td>
<td>22.0</td>
<td>19.2</td>
</tr>
<tr>
<td>σ = 5</td>
<td>37.1</td>
<td>38.6</td>
<td>37.0</td>
<td>32.0</td>
</tr>
<tr>
<td>σ = 10</td>
<td>32.0</td>
<td>32.6</td>
<td>32.2</td>
<td>28.1</td>
</tr>
<tr>
<td>σ = 20</td>
<td>26.4</td>
<td>27.1</td>
<td>28.4</td>
<td>25.2</td>
</tr>
<tr>
<td>σ = 30</td>
<td>23.7</td>
<td>23.6</td>
<td>26.0</td>
<td>23.7</td>
</tr>
<tr>
<td>σ = 5</td>
<td>38.9</td>
<td>40.5</td>
<td>40.4</td>
<td>34.5</td>
</tr>
<tr>
<td>σ = 10</td>
<td>33.4</td>
<td>34.6</td>
<td>35.0</td>
<td>29.9</td>
</tr>
<tr>
<td>σ = 20</td>
<td>28.5</td>
<td>28.8</td>
<td>30.5</td>
<td>26.3</td>
</tr>
<tr>
<td>σ = 30</td>
<td>25.7</td>
<td>25.1</td>
<td>27.5</td>
<td>24.0</td>
</tr>
</tbody>
</table>

model in Eq. (1) to take into account the subbands of the wavelet transform.

REFERENCES


