

---

# Extensions to MDL denoising

---

**Janne Ojanen**

Helsinki University of Technology TKK  
Department of Biomedical Engineering and Computational Science  
P.O.Box 9203 (Tekniikantie 14), FI-02015 TKK, Finland  
jiojanen@lce.hut.fi

**Jukka Heikkonen**

European Commission - Joint Research Centre  
Institute for the Protection and Security of the Citizen (IPSC)  
G04 Maritime Affairs (Fishreg Sector) TP 051, Via Fermi 1, I-21020 Ispra (VA), Italy  
jukka.heikkonen@jrc.it

## Abstract

The minimum description length principle in wavelet denoising can be extended from the standard linear-quadratic setting in several ways. We describe briefly three extensions: soft thresholding, histogram modeling and a multicomponent approach. The MDL hard thresholding approach based on the normalized maximum likelihood universal modeling can be extended to include soft thresholding shrinkage, which can be considered to give better results in some applications. In MDL histogram denoising approach the assumptions of the parametric density models for the data can be relaxed. The informative and noise components of the data are modeled with equal bin width histograms. The method can cope with different noise distributions. In multicomponent approach more than one non-noise components are included in the model, because it is possible that in addition to the random noise there may be other disturbing signal elements, or that the informative signal is comprised of several different components which we may want to observe, separate or remove. In these cases adding informative components in the model may result in better performance than in the NML denoising approach.

## 1 Introduction

The observed data is thought to be corrupted by additive noise,  $y^n = x^n + \epsilon^n$ , where the noise term  $\epsilon^n$  is often assumed to be comprised of i.i.d. Gaussians. Given the orthonormal regression matrix  $\mathbf{W}$  the discrete wavelet transform (DWT) of the noisy data is defined as  $c^n = \mathbf{W}^T y^n$ . The aim of wavelet denoising is to obtain modified coefficients  $\tilde{c}^n$  representing the informative part in the data. In MDL setting wavelet denoising is seen as a model selection task. The linear regression model can be rewritten as a density function  $f(y^n | c_\gamma^n, \sigma^2, \gamma) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2} \|y^n - \mathbf{W}c_\gamma^n\|^2\right\}$ , where the structure index  $\gamma$  defines which rows of the regressor matrix are included in the model, or equivalently, which elements of  $c_\gamma^n$  are non-zero. We may now define the NML density function, compute the well-known maximum likelihood estimates for parameters and calculate the normalizing factor by a renormalization scheme

discussed in Rissanen [1], and the result is a denoising criterion

$$\begin{aligned} & \frac{k}{2} \ln \left( \frac{1}{k} \sum_{i \in \gamma} c_i^2 \right) + \frac{n-k}{2} \ln \left( \frac{1}{n-k} \sum_{j \notin \gamma} c_j^2 \right) \\ & + \frac{1}{2} \ln k(n-k) + L(\gamma) \end{aligned} \quad (1)$$

approximating the stochastic complexity. The selection of  $\gamma$  and the resulting coefficient vector  $c_\gamma^n$  are obtained by minimizing the criterion. Furthermore, the criterion is shown to be minimized by the  $k$  coefficients with largest magnitudes. For the code length for the model class a code length function  $L(\gamma) = \ln \binom{n}{k}$  is recommended in [2].

## 2 MDL soft thresholding

An extension to wavelet denoising is to include soft thresholding shrinkage [3]. In essence, soft thresholding the observed wavelet coefficients with a threshold parameter  $\lambda$  gives two subsets indexed by  $\gamma_1$  and  $\gamma_2$ :  $\hat{c}_{\gamma_1} = (\hat{c}_{\gamma_1(1)}, \dots, \hat{c}_{\gamma_1(k)})$  corresponding to the shrunk  $k$  'informative' coefficients  $\hat{c}_i = \text{sign}(c_i)(|c_i| - \lambda)$ , and  $\tilde{c}_{\gamma_2} = (\tilde{c}_{\gamma_2(1)}, \dots, \tilde{c}_{\gamma_2(n-k)})$  containing the  $n-k$  unmodified 'noise' coefficients for which  $\text{sign}(c_i)(|c_i| - \lambda) < 0$ . A useful analogy is to think the process as data transmission over a channel. The sender must transmit enough information over a channel to the receiver so that the receiver is capable of reconstructing the original data from the transmitted signal. In this case we transmit, with as short a code length as possible,  $k$  soft thresholded coefficients  $\hat{c}_{\gamma_1}$  and  $n-k$  noise coefficients  $\tilde{c}_{\gamma_2}$ , so that when  $\lambda$  (which also must be transmitted) is known the receiver is able to reconstruct the original data.

The code length for the wavelet coefficients is obtained by encoding the subsets  $\hat{c}_{\gamma_1}$  and  $\tilde{c}_{\gamma_2}$  with separate NML codes  $L_{\text{NML}}(\hat{c}_{\gamma_1}|\gamma_1)$  and  $L_{\text{NML}}(\tilde{c}_{\gamma_2}|\gamma_2)$ , respectively. For computing the NML code length for any sequence, see [4]. The code length of the model class,  $L(\gamma_1, \gamma_2, \lambda)$ , is also required for describing the parameter of the shrinkage function as well as the index sets  $\gamma_1$  and  $\gamma_2$ . The code length may be further divided into  $L(\gamma_1, \gamma_2, \lambda) = L(\gamma_1, \gamma_2|\lambda) + L(\lambda)$ , where  $L(\gamma_1, \gamma_2|\lambda) = \ln \binom{n}{k}$  gives the code length for choosing the  $k$  coefficients into  $\gamma_1$  out of a total of  $n$  coefficients when  $\lambda$  is fixed.  $L(\lambda)$  is required to describe the threshold parameter value. However,  $L(\lambda)$  may be considered to be a constant that can be ignored in the final criterion. Finally, the encoding is performed by a two-part encoding where the total code length is given by the sum  $L_{\text{NML}}(\hat{c}_{\gamma_1}|\gamma_1) + L_{\text{NML}}(\tilde{c}_{\gamma_2}|\gamma_2) + L(\gamma_1, \gamma_2, \lambda)$ . Applying the Stirling's approximation to Gamma functions and ignoring all terms constant with respect to  $k$  gives the criterion for choosing the optimal parameter  $\lambda$ ,

$$\begin{aligned} \min_{\lambda} \left[ \frac{k}{2} \ln \left( \frac{1}{k} \sum_{i=1}^k \hat{c}_{\gamma_1(i)}^2 \right) \right. \\ \left. + \frac{n-k}{2} \ln \left( \frac{1}{n-k} \sum_{i=1}^{n-k} \tilde{c}_{\gamma_2(i)}^2 \right) \right. \\ \left. + \frac{1}{2} \ln k(n-k) + L(\gamma) \right]. \quad (2) \end{aligned}$$

The criterion (2) is almost identical to the original MDL denoising criterion (1): the difference is in the first term, where in the soft thresholding criterion there are shrunk wavelet coefficient values instead of the originals.

### 3 MDL histogram denoising

The NML approach is restricted to the quadratic-linear case in which noise is assumed to follow Gaussian distribution. We obtain another denoising criterion by employing histogram models. The main idea is to model the wavelet coefficients representing the denoised signal,  $\hat{c}^n$ , by an equal bin width histogram at each resolution level of the wavelet transform, and the coefficients representing the noise,  $\tilde{c}^n$  by a single equal bin width histogram. Minimization of the total code length yields the optimal way of dividing the coefficients into ones representing informative signal and noise. For information on how to compute the stochastic complexity for the data string given the number of bins in the fixed bin width histogram, see [5, 6].

The key parts of the MDL-histo denoising algorithm discussed in more detail in [7] are summarized as follows:

1. Obtain the set of wavelet coefficients  $c^n = c_1^{n_1}, \dots, c_r^{n_r}$  through the  $r$ -level wavelet transform.
2. Recursively on resolution levels  $i = 1, \dots, r$  fit a an  $m$ -bin histogram  $H_i$  to the coefficients  $c_i^{n_i}$  and select a tentative collection of bins  $S_i$ , with the number of chosen bins  $m_i = |S_i|$ . Denote by  $n_{i,(j)}$  the number of points falling in the bin of  $H_i$  having index  $(j)$ . The bins in  $S_i$  contain  $k_i$  retained coefficients. The retained and residual coefficients at level  $i$  are written as two strings  $\hat{c}_i^{n_i}$  and  $\tilde{c}_i^{n_i}$ , respectively.
3. Fit a histogram with  $M$  bins to the residual coefficients  $\tilde{c}^n = \tilde{c}_1^{n_1}, \dots, \tilde{c}_i^{n_i}, c_{i+1}^{n_{i+1}}, \dots, c_r^{n_r}$  where the first  $i$  residual strings are obtained by setting the already retained coefficients to zero.
4. Find the optimal  $S_i$  by minimizing the criterion

$$\begin{aligned} \min_{S_i, M} \left\{ \log_2 \binom{n_i}{n_{i,(1)}, \dots, n_{i,(m_i)}, (n_i - k_i)} + \right. \\ \log_2 \binom{n_i + m_i + 1}{n_i} + \log_2 \binom{n - \sum_{j=1}^{i-1} \hat{k}_j - k_i}{\nu_1, \dots, \nu_M} \\ + \log_2 \binom{n - \sum_{j=1}^{i-1} \hat{k}_j - k_i + M}{M} + k_i \log_2 \left( \frac{R}{m} \right) \\ + k_i \log_2 \left( \frac{M}{R_i} \right) + \left( \sum_{j=1}^{i-1} \hat{k}_j \right) \log_2 \left( \frac{M}{R_i} \right) \\ \left. - (n - 1) \log_2 M + 2 \log_2 \log_2 M \right. \\ \left. + n \log_2 R_i + \log_2 R_i + 2 \log_2 \log_2 R_i \right\}, \quad (3) \end{aligned}$$

where  $\nu_j$  is the number of coefficients falling into the  $j$ th bin of the  $M$ -bin histogram fitted to the residual string  $\tilde{c}^n$ ,  $R$  is the range of wavelet coefficients,  $R_i$  are the levelwise ranges of the coefficients and  $\sum_{j=1}^{i-1} \hat{k}_j$  denotes the number of retained coefficients in the so far optimized sets  $S_j, j < i$ . For the first level  $i = 1$  this sum is zero.

The denoised signal results from the inverse transform of the sequence of retained coefficients  $\hat{c} = \hat{c}^n = \hat{c}_1^{n_1}, \dots, \hat{c}_r^{n_r}$ .

### 4 Multicomponent denoising

It is possible that in addition to the random noise there may be other disturbing signal elements, or that the informative signal is comprised of several different components which we may want to observe, separate

or remove. With more than one informative component in the noisy measured data a multicomponent approach may result in better performance than the original MDL denoising method [8].

Roos et al. [9, 10, 2] have shown that a criterion similar to the renormalization result can be obtained by a different derivation, details of which can be found in [10, 2]. In short, they define a model for the wavelet coefficients, in which each coefficient is distributed according to a zero-mean Gaussian density with variance  $\sigma_I^2$  if it belongs to the set of informative coefficients indexed by  $\gamma$ , or according to a zero-mean Gaussian density with variance  $\sigma_N^2$  if it represents noise, with the restriction  $\sigma_I^2 \geq \sigma_N^2$ . Again, the optimal denoising result is given by the  $\gamma$  minimizing the normalized maximum likelihood code length of the data given the model class defined by  $\gamma$ .

This approach may be extended by using  $m$  Gaussian components and specifying the restriction for their variance parameters,  $\sigma_1^2 \geq \dots \geq \sigma_m^2$ . The NML code length for this model can be calculated in a manner following the derivation in [10] for the two-component denoising criterion. The derivation turns out to be straightforward since the normalizing integral factors into  $m$  parts, each depending only on the coefficients determined by the respective index set  $\gamma_i$ . We obtain a criterion

$$\sum_{i=1}^m \left( \frac{k_i}{2} \ln \frac{1}{k_i} \sum_{j \in \gamma_i} c_j^2 + \frac{1}{2} \ln k_i \right) + L(\gamma_1, \dots, \gamma_m) + m \log \log \frac{\sigma_{\max}^2}{\sigma_{\min}^2} + \text{const}, \quad (4)$$

where const refers to terms constant with respect to the index sets and  $m$ , and  $\sigma_{\max}^2$  and  $\sigma_{\min}^2$  are hyperparameters for the maximum and minimum variance, respectively. The last two terms can be ignored if we wish to find the optimal  $m$ -component result. However, if we want to compare the results for two approaches with different number of components, for example  $m_1 = 3$  and  $m_2 = 4$ , we cannot remove the term involving the hyperparameters as it affects the code length.

## References

- [1] J. Rissanen. MDL denoising. *IEEE Transactions on Information Theory*, 46(7):2537–2543, 2000.
- [2] T. Roos. *Statistical and Information-Theoretic Methods for Data-Analysis*. PhD thesis, University of Helsinki, 2007.
- [3] J. Ojanen and J. Heikkonen. MDL and wavelet denoising with soft thresholding. *Submitted to 2008 Workshop on Information Theoretic Methods in Science and Engineering*, 2008.
- [4] J. Rissanen, editor. *Information and Complexity in Statistical Modeling*. Springer, 2007.
- [5] P. Hall and E.J. Hannan. On stochastic complexity and nonparametric density estimation. *Biometrika*, 75(4):705–714, December 1988.
- [6] J. Rissanen, T.P. Speed, and B. Yu. Density estimation by stochastic complexity. *IEEE Transactions on Information Theory*, 38(2):315–323, March 1992.
- [7] V. Kumar, J. Heikkonen, J. Rissanen, and K. Kaski. Minimum description length denoising with histogram models. *IEEE Transactions on Signal Processing*, 54(8):2922–2928, August 2006.
- [8] J. Ojanen, J. Heikkonen, and K. Kaski. Towards the multicomponent MDL denoising. In *Festschrift in Honor of Jorma Rissanen on the Occasion of his 75th Birthday*. Tampere University Press, 2008.
- [9] T. Roos, P. Myllymäki, and H. Tirri. On the behavior of MDL denoising. In Robert G. Cowell and Zoubin Ghahramani, editors, *AISTATS05*, pages 309–316. Society for Artificial Intelligence and Statistics, 2005.
- [10] T. Roos, P. Myllymäki, and J. Rissanen. MDL denoising revisited. *Submitted for publication*, 2006. Preprint available at: <http://www.arxiv.org/abs/cs.IT/0609138>.