MDL AND WAVELET DENOISING WITH SOFT THRESHOLDING

Janne Ojanen[1] *and Jukka Heikkonen*[2]

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

ABSTRACT

We propose a soft thresholding approach to the minimum description length wavelet denoising. Our method is based on combining two-part coding with normalized maximum likelihood universal models to give a soft thresholding denoising criterion. Experiments with the proposed MDL soft thresholding method indicate that our denoising criterion leads to fairly similar performance as with the well-known VisuShrink method.

1. INTRODUCTION

Denoising, the task of removing or suppressing uninformative noise from signals, is an important part of many signal or image processing applications. Wavelets are common tools in the field of signal processing [1, 2]. The popularity of wavelets in denoising is largely due to the computationally efficient algorithms as well as to the sparsity of the wavelet representation of data. By sparsity we mean that majority of the wavelet coefficients have very small magnitudes whereas only a small subset of coefficients have large magnitudes [3]. We may informally state that this small subset contains the interesting informative part of the signal, whereas the rest of the coefficients describe noise and can be discarded to give a noise-free reconstruction.

The best known wavelet denoising methods are thresholding approaches, see e.g. [4, 5]. In hard thresholding all the coefficients with greater magnitudes than the threshold are retained unmodified as they are thought to comprise the informative part of data, while the rest of the coefficients are considered to represent noise and set to zero. However, it is reasonable to assume that coefficients are not purely either noise or informative but mixtures of those. To cope with this soft thresholding approaches have been proposed. In soft thresholding the coefficients with magnitudes smaller than the threshold are set to zero, but the retained coefficients are also shrunk towards zero by the amount of the threshold value in order to decrease the effect of noise assumed to corrupt all the wavelet coefficients.

Probably the most popular wavelet-based denoising methods are thresholding approaches proposed by Donoho and Johnstone aiming at minimizing the worst-case risk, and they have been shown to be minimax optimal over a large class of functions [4, 5, 6]. Another group of popular methods in wavelet denoising are Bayesian approaches often based on minimizing the expected risk, with the expectation taken over a postulated prior distribution supposedly governing the underlying true signal [7, 8, 9].

A different approach to wavelet denoising is based on the minimum description length (MDL) principle [10, 11, 12, 13]. 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 assumptions of the noise distribution, even though a Gaussian noise model is usually assumed. Although several different MDL denoising methods have been proposed [14, 15, 16], this paper concentrates on the normalized maximum likelihood (NML) approach originally suggested by Rissanen [17] and further developed by Roos et al. [18, 19]. The NML denoising method may be considered to be the most theoretically rigorous MDL denoising approach.

The MDL denoising methods proposed this far have been based on selecting a subset of wavelet coefficients to represent the informative signal, which is equivalent to hard thresholding. However, a theoretically sound MDL soft thresholding method would be useful, because soft thresholding has been found in some cases superior to hard thresholding. Some soft thresholding ideas in MDL denoising have been proposed in [19, 20]. In this paper we propose a soft thresholding MDL method based on NML and two-part coding generalizing the hard thresholding approach. We also demonstrate that our soft thresholding approach gives results fairly similar to the VisuShrink method of Donoho and Johnstone [4].

2. MDL PRINCIPLE

The general ideas of the MDL principle in model class selection are introduced before describing MDL wavelet denoising. Let $x^n = (x_1, \ldots, x_n)^T$ be a data sequence, equivalently viewed as a column vector when necessary. A model class

$$\mathcal{M}_\gamma = \{f(x^n|\theta,\gamma) : \theta \in \Theta_\gamma \subset R^k, \gamma \in \Gamma\} \quad (1)$$

is defined as the set of density functions $f(x^n|\theta,\gamma)$ where the structure index γ defines the dimensionality of the real-valued parameter vector $\theta = (\theta_{\gamma(i)}, \ldots, \theta_{\gamma(k)})^T$. For example, in linear regression the structure index γ defines which k input variables are included in the model while θ defines the values of the regression coefficients. In model class selection a model class indexed by γ is selected among the set $\mathcal{M} = \bigcup_\gamma \mathcal{M}_\gamma$.

Each model class may be represented by a single universal model. The normalized maximum likelihood (NML) model is a universal model with minimax optimality properties important enough to consider the code length associated with the NML model, $-\ln f_{\text{NML}}$, to be the stochastic complexity, the shortest achievable description of the data given the model class. The NML density function f_{NML} is defined for the data x^n given the model class $\mathcal{M}_\gamma$ as

$$f_{\text{NML}}(x^n|\gamma) = \frac{f(x^n|\hat{\theta}(x^n),\gamma)}{C_{n,\gamma}} , \quad (2)$$

where the normalizing constant is given by

$$C_{n,\gamma} = \int_{z^n} f(z^n|\hat{\theta}(z^n),\gamma)\mathrm{d}z^n . \quad (3)$$

The normalizing integral (3), also known as the parametric complexity, is problematic as it is unbounded for many useful and realistic models such as the density functions of the exponential family. In order to keep the parametric complexity bounded typically either the range of data or parameters is restricted.

The MDL principle tells us to select the model class minimizing the total code length

$$\min_\gamma \{-\ln f_{\text{NML}}(x^n|\gamma) + L(\gamma)\} \quad (4)$$

defined as a two part code length composed of the stochastic complexity for the data given the model class and the code length $L(\gamma)$ for encoding the model class. In applications where the number of compared model classes is very small related to the number of data points the code length for the model class may be ignored and the model class selection can be done according to the stochastic complexities of the model classes. However, in applications such as denoising where the number of parameters may be close to the number of data points the model class code length may have significant impact on the results.

3. WAVELET DENOISING AND MDL

The denoising problem can be described formally in a linear regression setting. The observed data is represented as a real-valued column vector x^n. This signal model can be easily extended, for example, into the two-dimensional image data. We define an $n \times n$ wavelet regressor matrix $\mathbf{W}$, whose columns are basis vectors forming a complete orthonormal basis. Due to orthonormality the inverse of the regressor matrix is given by its transpose, $\mathbf{W}^{-1} = \mathbf{W}^T$. The data x^n can be written as a linear combination of the basis vectors weighted with a coefficient vector $\beta = (\beta_1, \ldots, \beta_n)^T$ and Gaussian noise,

$$x^n = \mathbf{W}\beta + \epsilon^n , \quad (5)$$

where the elements of ϵ^n are i.i.d. Gaussians, $\epsilon_i \sim N(0,\sigma^2)$, with a common variance σ^2. Given the regression matrix $\mathbf{W}$ the discrete wavelet transform (DWT) of the noisy data is given by

$$c^n = \mathbf{W}^T x^n = \beta + \mathbf{W}^T\epsilon^n , \quad (6)$$

where the noise in wavelet domain $\mathbf{W}^T\epsilon^n$ is also a Gaussian due to the orthonormality of the wavelet transform. The aim in denoising is to obtain estimates for the noise-free wavelet coefficients $\hat{\beta}$ and to produce a denoised signal $\hat{x}^n$ with the inverse discrete wavelet transform (IDWT): $\hat{x}^n = \mathbf{W}\hat{\beta}$. The conventional maximum likelihood method fails unless the number of parameters is somehow restricted; using the ML solution $\hat{\beta} = \mathbf{W}^T x^n$ of the full model in the IDWT gives the observed noisy data $\hat{x}^n = \mathbf{W}\mathbf{W}^T x^n = x^n$.

The most common wavelet denoising methods are based on hard or soft thresholding. In hard thresholding a subset of coefficients with magnitudes larger than the threshold are retained,

$$\hat{c}_i = c_i \mathrm{I}_{\{|c_i|>\lambda\}}, \quad i = 1, \ldots, n , \quad (7)$$

where λ denotes the threshold and I is an indicator function. In soft thresholding the retained coefficients are also shrunk towards zero

$$\hat{c}_i = \mathrm{sign}(c_i)(|c_i| - \lambda)\mathrm{I}_{\{|c_i|>\lambda\}}, \quad i = 1, \ldots, n . \quad (8)$$

A common choice for the threshold is the so-called universal threshold, $\lambda = \hat{\sigma}\sqrt{2\ln n}$, where the noise standard deviation estimate $\hat{\sigma}$ is usually obtained as the median absolute deviation (MAD) estimate from the wavelet coefficients of the finest detail level [4]. Soft thresholding with the universal threshold is known as the VisuShrink method.

In MDL setting wavelet denoising is seen as a model class selection task. We can rewrite the linear regression model in (5) as a density function

$$f(x^n|\beta_\gamma,\sigma^2,\gamma) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}||x^n - \mathbf{W}\beta_\gamma||^2\right\}, \quad (9)$$

where the structure index γ defines which columns of the regressor matrix are included in the model, or equivalently, which elements of β_γ are non-zero. We may now define the NML density function (2), in which the maximum likelihood estimates for the parameters are the well-known $\hat{\beta}_i = c_i$ for $i \in \gamma$ and $\hat{\sigma}^2 = \frac{1}{n}||x^n - \mathbf{W}\hat{\beta}||^2$. Calculating the NML density function requires restricting the range of data to keep the parametric complexity bounded. Rissanen [17] solves this by introducing hyperparameters restricting the ML parameter estimates. Because the hyperparameters affect the resulting code length, they must be removed by a second normalization over the hyperparameters, resulting in a criterion

$$\frac{k}{2}\ln\left(\frac{1}{k}\sum_{i\in\gamma}c_i^2\right) + \frac{n-k}{2}\ln\left(\frac{1}{n-k}\sum_{j\notin\gamma}c_j^2\right) + \frac{1}{2}\ln k(n-k) + L(\gamma) \quad (10)$$

approximating the stochastic complexity, which is shown to be minimized by the k coefficients with largest magnitudes [17]. Therefore, instead of optimizing over γ it is sufficient to find optimal $\hat{k}$ minimizing (10). Terms constant with respect to k or γ (for example, terms containing the hyperparameters) have been discarded as they do not affect the model class selection task. In fact, (10) determines a hard thresholding rule where the threshold value is defined implicitly by minimizing the code length. For the code length for the model class a code length function $L(\gamma) = \ln\binom{n}{k}$ is recommended in [19].

4. MDL SOFT THRESHOLDING

Consider the observed wavelet coefficients c^n and a fixed threshold λ. Soft thresholding results in two coefficient sequences,

$$\hat{c}_i = \begin{cases} \text{sign}(c_i)(|c_i| - \lambda) & |c_i| > \lambda \\ 0 & |c_i| \le \lambda \end{cases} \quad (11)$$

defining the coefficients $\hat{c}^n$ corresponding to the informative signal and

$$\tilde{c}_i = \begin{cases} \text{sign}(c_i)\lambda & |c_i| > \lambda \\ c_i & |c_i| \le \lambda \end{cases} \quad (12)$$

defining $\tilde{c}^n$ describing noise. A reconstruction of the noise-free signal is obtained through the IDWT, $\hat{x}^n = \mathbf{W}\hat{c}^n$.

MDL may be used to determine the optimal coefficient vector $\hat{c}^n$. 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, enough coefficients from both $\hat{c}^n$ and $\tilde{c}^n$ so that when λ (which also must be transmitted) is known the receiver is able to reconstruct the original data. In fact, we have to encode the k non-zero coefficients from $\hat{c}^n$, because we cannot replicate the original data from the respective k elements of $\tilde{c}^n$. Vice versa, the $n-k$ remaining coefficients must be taken from $\tilde{c}^n$ because the zeros in $\hat{c}^n$ cannot be inverted to give the original coefficients. In other words, fixing the threshold λ also explicitly gives the division into two subsets indexed by γ_1 and γ_2, $\hat{c}_{\gamma_1} = (\hat{c}_{\gamma_1(1)}, \ldots, \hat{c}_{\gamma_1(k)})$ and $\tilde{c}_{\gamma_2} = (\tilde{c}_{\gamma_2(1)}, \ldots, \tilde{c}_{\gamma_2(n-k)})$.

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. 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 γ_1 and γ_2. Finally, the encoding is performed by a two-part encoding where the total code length L is given by

$$L = L_{\text{NML}}(\hat{c}_{\gamma_1}|\gamma_1) + L_{\text{NML}}(\tilde{c}_{\gamma_2}|\gamma_2) + L(\gamma_1, \gamma_2, \lambda) \; . \quad (13)$$

A real valued sequence may be encoded using a normalized maximum likelihood coding [13]. The required NML code lengths for the coefficient sets are given by

$$L_{\text{NML}}(\hat{c}_{\gamma_1}|\gamma_1) = \frac{k}{2}\ln\left(k\pi\hat{\tau}(\hat{c}_{\gamma_1})\right) - \ln\Gamma\left(\frac{k}{2}\right) + \ln\ln\frac{\hat{\tau}_{\max}}{\hat{\tau}_{\min}} \quad (14)$$

and

$$L_{\text{NML}}(\tilde{c}_{\gamma_2}|\gamma_2) = \frac{n-k}{2}\ln\left((n-k)\pi\hat{\tau}(\tilde{c}_{\gamma_2})\right) - \ln\Gamma\left(\frac{n-k}{2}\right) + \ln\ln\frac{\hat{\tau}_{\max}}{\hat{\tau}_{\min}} \; , \quad (15)$$

where the maximum likelihood variance estimates are given by $\hat{\tau}(\hat{c}_{\gamma_1}) = \frac{1}{k}\sum_{j=1}^{k}\hat{c}^2_{\gamma_1(j)}$ and $\hat{\tau}(\tilde{c}_{\gamma_2}) = \frac{1}{n-k}\sum_{j=1}^{n-k}\tilde{c}^2_{\gamma_2(j)}$, respectively. Hyperparameters $\hat{\tau}_{\min}$ and $\hat{\tau}_{\max}$ define the minimum and maximum of the ML variance estimates. These hyperparameters must be introduced to make the parametric complexity (3) bounded. However, while the hyperparameters clearly affect the code length, they are later seen to have no effect on the model class selection.

The code length $L(\gamma_1, \gamma_2, \lambda)$ is the cost of thresholding the DWT coefficients and assigning them into two subsets. The code length may be further divided into

$$L(\gamma_1, \gamma_2, \lambda) = L(\gamma_1, \gamma_2|\lambda) + L(\lambda) = \ln\binom{n}{k} + L(\lambda) \; , \quad (16)$$

where $L(\gamma_1, \gamma_2|\lambda) = \ln\binom{n}{k}$ gives the code length for choosing the k coefficients into γ_1 out of a total of n coefficients when λ 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.

We combine the code lengths (14), (15) and (16), apply the Stirling's approximation to Gamma functions and ignore all terms constant with respect to k (for example, the terms containing the hyperparameters in (14) and (15) are seen to have no effect on the criterion). The criterion for choosing the optimal parameter λ is given by

$$\min_{\lambda}\Bigg[\frac{k}{2}\ln\left(\frac{1}{k}\sum_{i=1}^{k}\hat{c}^2_{\gamma_1(i)}\right) + \frac{n-k}{2}\ln\left(\frac{1}{n-k}\sum_{i=1}^{n-k}\tilde{c}^2_{\gamma_2(i)}\right) + \frac{1}{2}\ln k(n-k) - k\ln k - (n-k)\ln(n-k)\Bigg], \quad (17)$$

where the two last terms come from the Stirling's approximation to the model class code length $\ln\binom{n}{k}$. The criterion (17) is almost identical to the original MDL denoising criterion (10): the difference is in the first term, where in the soft thresholding criterion there are shrunk wavelet coefficients instead of the originals. Furthermore, taking the hard thresholding function and going through the equations leading to (17) gives exactly (10).

5. EXPERIMENTS

The performance of the proposed MDL soft thresholding method was studied with a set of artificial 1-D signals [4] scaled for the range of 200 and 8-bit grayscale natural images[1] 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

$$\text{PSNR} = 10\log_{10}\left(\frac{[\max(x^n) - \min(x^n)]^2}{\text{MSE}}\right), \quad (18)$$

where the squared range is calculated from the signal x^n and MSE is the mean squared error. Especially for the images visual quality of the results is also important. The Daubechies 'db5' wavelet basis was used in all experiments with $N = 5$ decomposition levels in the multiresolution wavelet transform. In practice, the approximation coefficients are often retained without shrinking. We adopted this custom to keep the results comparable.

Compared to the original MDL denoising soft thresholding seems to have only little effect on the number of retained coefficients. Typically only a small number of extra coefficients are retained in soft thresholding approach, so that the main difference in the denoising results comes from the shrinkage effect. An example denoising result for 'Lena' image is shown in Figure 1, where the denoising result for VisuShrink is also presented. Soft thresholding results are typically somewhat oversmoothed compared to hard thresholding, which is also seen in Figure 1.

Further comparisons with VisuShrink revealed that the MDL soft thresholding approach gives similar results in terms of error measures and visual quality. However, there are differences in threshold values and therefore in the number of retained coefficients. With 1-D signals there is some dispersion in the threshold values, but on the average the thresholds and therefore other properties of the MDL soft thresholding results are fairly similar to VisuShrink, as can be see in Figure 2. While the denoising results with 2-D images also are similar, MDL method consistently retains more coefficients than VisuShrink especially at higher noise levels. This is reflected in lower thresholds, as can be seen in Figure 3. Despite these clear differences, the effect on denoising results seems to be almost negligible.

[1] USC-SIPI image database, http://sipi.usc.edu/database/

Figure 1. Example denoising results. (a) Noisy 'Lena' image, 512×512, noise standard deviation $\sigma = 15$, PSNR $= 24.6$; (b) MDL hard thresholding, PSNR $= 29.0$; (c) MDL soft thresholding, PSNR $= 26.6$; (d) VisuShrink, PSNR $= 26.4$.

6. CONCLUSIONS

We have proposed a method for employing soft thresholding shrinkage in MDL wavelet denoising based on combining two-part and NML coding. We have shown that when our method is adapted to hard thresholding, we obtain the criterion suggested in the original MDL denoising approach [17]. We have also demonstrated that our MDL soft thresholding approach has fairly similar denoising performance as the well-known VisuShrink method.

The method described in this article can be extended in several ways. Here we have discussed only global denoising methods. However, many existing denoising methods are level-dependent: they use the multiresolution properties of the wavelet transform by applying different thresholds at each DWT decomposition level in order to obtain better separation of noise and underlying informative signal. It is possible to extend our method to be level-dependent by dividing the coefficients into subsets

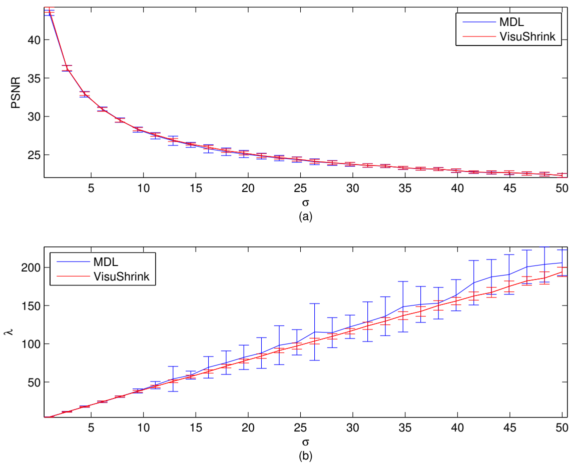

Figure 2. A comparison between MDL soft threshold and VisuShrink. (a) Results (PSNR) for denoising 'Blocks' [4] signal corrupted with Gaussian noise of varying standard deviation. The results at each noise standard deviation are mean values of 10 independent simulations, with the error bars drawn to show two times the standard deviation of the results. (b) The average threshold values of the same simulations.

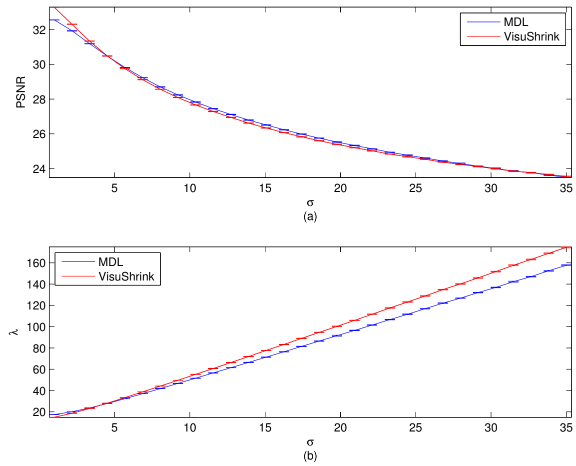

Figure 3. A comparison between MDL soft threshold and VisuShrink. (a) Results (PSNR) for denoising 'Lena' corrupted with Gaussian noise of varying standard deviation. The results at each noise standard deviation are mean values of 10 independent simulations, with the error bars drawn to show two times the standard deviation of the results. (b) The average threshold values of the same simulations.

according to their decomposition levels and encoding the retained coefficients with NML codes. Also, a similar idea could be used to include more than one informative component in the data. In addition, other shrinkage functions could be used instead of soft thresholding. An interesting extension would be to consider other universal coding systems than NML, which is known to have problems with the unbounded parametric complexity. For example, conditional normalized maximum likelihood [21] coding could be used to compute the code length for the Laplacian distribution which then could be used as a model for the wavelet coefficients.

7. ACKNOWLEDGMENTS

This work was supported in part by the Graduate School in Computational Methods of Information Technology at Helsinki University of Technology, the Finnish Technology Agency under project KUKOT, Institute for the Protection and Security of the Citizen (IPSC) of the Joint Research Centre - European Commission, and the Center of Excellence Program of the Academy of Finland.

8. REFERENCES

[1] I. Daubechies, *Ten Lectures on Wavelets*, vol. 61 of *CBMS-NSF Regional Conference Series in Applied Mathematics*, Capital City Press, 1992.

[2] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, 1998.

[3] S. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674–693, July 1989.

[4] D.L. Donoho and I.M. Johnstone, "Ideal spatial adaption by wavelet shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425–455, August 1994.

[5] D.L. Donoho, "De-noising by soft-thresholding," *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 613–627, May 1995.

[6] D.L. Donoho and I.M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *Journal of the American Statistical Association*, vol. 90, no. 432, pp. 1200–1224, December 1994.

[7] B. Vidakovic, "Nonlinear wavelet shrinkage with Bayes rules and Bayes factors," *Journal of the American Statistical Association*, vol. 93, no. 441, pp. 173–179, 1998.

[8] F. Ruggeri and B. Vidakovic, "A Bayesian decision theoretic approach to the choice of thresholding parameter," *Statistica Sinica*, vol. 9, pp. 183–197, 1999.

[9] S.G. Chang, B. Yu, and M. Vetterli, "Adaptive wavelet thresholding for image denoising and compression," *IEEE Transactions on Image Processing*, vol. 9, no. 9, pp. 1532–1546, September 2000.

[10] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[11] A.R. Barron, J. Rissanen, and B. Yu, "The minimum description length principle in coding and modeling," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2743–2760, 1998.

[12] P. Grünwald, I.J. Myung, and M.A. Pitt, Eds., *Advances in Minimum Description Length: Theory and Applications*, MIT Press, 2005.

[13] J. Rissanen, Ed., *Information and Complexity in Statistical Modeling*, Springer, 2007.

[14] N. Saito, "Simultaneous noise suppression and signal compression using a library of orthonormal bases and the minimum description length criterion," in *Wavelets in Geophysics*, E. Foufoula-Georgiu and P. Kumar, Eds., pp. 299–324. New York: Academic, 1994.

[15] M. Hansen and B. Yu, "Wavelet thresholding via MDL for natural images," *IEEE Transactions on Information Theory*, vol. 46, no. 5, pp. 1778–1788, August 2000.

[16] V. Kumar, J. Heikkonen, J. Rissanen, and K. Kaski, "Minimum description length denoising with histogram models," *IEEE Transactions on Signal Processing*, vol. 54, no. 8, pp. 2922–2928, August 2006.

[17] J. Rissanen, "MDL denoising," *IEEE Transactions on Information Theory*, vol. 46, no. 7, pp. 2537–2543, 2000.

[18] T. Roos, P. Myllymäki, and H. Tirri, "On the behavior of MDL denoising," in *AISTATS05*, Robert G. Cowell and Zoubin Ghahramani, Eds. 2005, pp. 309–316, Society for Artificial Intelligence and Statistics.

[19] T. Roos, *Statistical and Information-Theoretic Methods for Data-Analysis*, Ph.D. thesis, University of Helsinki, 2007.

[20] J. Ojanen and J. Heikkonen, "A soft thresholding approach for MDL denoising," in *Proceedings of the 15th European Signal Processing Conference (EUSIPCO 2007)*, 2007, pp. 1083–1087.

[21] J. Rissanen and T. Roos, "Conditional NML universal models," *Information Theory and Applications Workshop, 2007*, pp. 337–341, Jan. 29 2007-Feb. 2 2007.