On the Behavior of MDL Denoising

Teemu Roos **Petri Myllymäki**
Helsinki Institute for Information Technology
Univ. of Helsinki & Helsinki Univ. of Technology
P.O. Box 9800 FIN-02015 TKK, Finland

Henry Tirri
Nokia Research Center
P.O. Box 407 Nokia Group
FIN-00045, Finland

Abstract

We consider wavelet denoising based on minimum description length (MDL) principle. The derivation of an MDL denoising criterion proposed by Rissanen involves a renormalization whose effect on the resulting method has not been well understood so far. By inspecting the behavior of the method we obtain a characterization of its domain of applicability: good performance in the low variance regime but over-fitting in the high variance regime. We also describe unexpected behavior in the theoretical situation where the observed signal is pure noise. An interpretation for the renormalization is given which explains both the empirical and theoretical findings. For practitioners we point out two technical pitfalls and ways to avoid them. Further, we give guidelines for constructing improved MDL denoising methods.

1 INTRODUCTION

Most natural signals such as audio and images are typically redundant in that the neighboring time-slots or pixels are highly correlated. Wavelet representations of such signals are very sparse, meaning that most of the wavelet coefficients are very small and the information content is concentrated on only a small fraction of the coefficients (Mallat, 1989). This can be exploited in data compression, pattern recognition, and denoising, i.e., separating the informative part of a signal from noise. In statistics the denoising problem has been analyzed in terms of statistical risk, i.e., the expected distortion under an assumed model where typically distortion is defined as squared error and the model consists of deterministic signal plus additive Gaussian noise. Donoho & Johnstone (1994) prove that certain thresholding methods are nearly minimax optimal for a large class of signals. In the Bayesian approach a prior distribution is postulated for the signal and the expected (Bayes) risk is minimized (Ruggeri & Vidakovic, 1999). Both approaches require that parameters such as noise variance are known beforehand or determined as a part of the process.

The minimum description length (MDL) philosophy offers an alternative view where the noise is *defined* as the incompressible part of the signal (Rissanen, 2000). We analyze Rissanen's MDL denoising method and characterize its domain of applicability. We show that the method performs well in the low variance regime but fails in the high variance regime when compared to a thresholding method proposed by Donoho and Johnstone. In particular, in the theoretical situation where the noise completely dominates the signal, the MDL denoising method retains a majority of the wavelet coefficients even though in this case discarding all coefficients is the optimal solution in terms of both statistical risk and what we intuitively understand as separating information from noise.

We explain the behavior of the MDL method by showing that it results not from the MDL principle itself but from a renormalization technique used in deriving the method. We also point out two technical pitfalls in the implementation of MDL denoising that practitioners should keep in mind. Further, we give guidelines for constructing MDL denoising methods that have a wider domain of applicability than the current one and list objectives for future research in this direction.

2 MDL PRINCIPLE

We start by introducing some notation and briefly reviewing some of the relevant parts of MDL theory. A recent introduction to MDL is given by Grünwald (2005), see also Barron *et al.* (1998).

2.1 STOCHASTIC COMPLEXITY

Let y^n be a sequence of observations. We define a model class as a set of densities $\{f(y^n\,;\theta)\colon \theta\}$ indexed by a finite-dimensional parameter vector θ. The maximum likelihood estimator of the parameter vector is denoted by $\hat{\theta}(y^n)$. The normalized maximum likelihood (NML) density for a model class parameterized by parameter vector θ is defined by

$$\bar{f}(y^n) = \frac{f(y^n\,;\hat{\theta}(y^n))}{C^n}, \tag{1}$$

where C^n is a normalizing constant:

$$C^n = \int_Y f(y^n\,;\hat{\theta}(y^n))\; dy^n. \tag{2}$$

Implicit in the notation is the range of integration Y within which the data y^n is restricted. A range other than the full domain of y^n is necessary in cases where the integral is otherwise unbounded.

The difference between the ideal code-length (negative logarithm) of the NML density and the unachievable maximum likelihood code-length is given by the *regret* which is easily seen to be constant for all data sequences y^n:

$$-\ln \bar{f}(y^n) - [-\ln f(y^n\,;\hat{\theta}(y^n))] = \ln C^n.$$

The NML density is the unique minimizer in Shtarkov's minimax problem (Shtarkov, 1987):

$$\min_q \max_{y^n} -\ln q(y^n) - [-\ln f(y^n\,;\hat{\theta}(y^n))] = \ln C^n,$$

and the following more general problem:

$$\min_q \max_p E_p - \ln q(y^n) - [-\ln f(y^n\,;\hat{\theta}(y^n))] = \ln C^n,$$

where the expectation over y^n is taken with respect to the worst-case data generating density p. For any density q other than the NML density, the maximum (expected) regret is greater than $\ln C^n$. Further, the NML is also the *least favorable* distribution in that is the unique maximizer of the maximin problem with the order of the min and max operators in the latter problem above exchanged. For these reasons the NML code is said to be universal in that it gives the shortest description of the data achievable with a given model class, deserving to be defined as the *stochastic complexity* of the data for the model class. The MDL principle advocates the choice of the model class for which stochastic complexity is minimized.

2.2 PARAMETRIC COMPLEXITY

It is instructive to view NML as seeking a balance between fit versus complexity. The numerator measures how well the best model in the model class can represent the observed data while the denominator 'penalizes' too complex model classes. The logarithm of the denominator, $\ln C^n$, is termed *parametric complexity* of the model class. Currently one of the most active areas of research within the MDL framework is the problem of unbounded parametric complexity which makes it impossible to define the NML density for models such as geometric, Poisson, and Gaussian families, see (Grünwald, 2005).

For model classes with unbounded parametric complexity, Rissanen (1996) proposes to use a two-part scheme where the range of the data is first encoded using a code based on an universal code for integers after which the data is encoded using NML taking advantage of the restricted range. Foster & Stine (2001, 2005) analyze similar schemes where the range of the *parameters* is restricted instead that of the the data. A weakness in such solutions is that they typically result in two-part codes that are not complete, i.e., the corresponding density integrates to less than one.

Rissanen (2000) describes an elegant renormalization scheme where the hyperparameters defining the range of the data are optimized and a second normalization is performed such that the resulting code is complete. This 'renormalized' NML can be used for model selection in linear regression and denoising. We discuss the renormalization and the resulting MDL denoising criterion more thoroughly in Sec. 4.

3 WAVELET DENOISING

Wavelet denoising can be seen as a special case of linear regression with regressor selection. For a good textbook on wavelets, see (Daubechies, 1992). An ex-

tensive review of statistical uses of wavelets is given by Abramovich *et al.* (2000).

3.1 WAVELET REGRESSION

This section closely follows Rissanen (2000). Let X be an $n \times k$ matrix of *regressor variables* (independent variables), and y^n be a vector of n *regression variables* (dependent variables). In a linear regression model the regression variables are dependent on the regressor variables and a $k \times 1$ parameter vector β through the equation $y^n = X\beta + \epsilon^n$, where ϵ^n is a vector of n noise terms that are modeled as independent Gaussian with zero mean and variance σ^2. This is equivalent to the equation

$$f(y^n; \beta, \sigma) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{\|y^n - X\beta\|^2}{2\sigma^2}\right), \quad (3)$$

where $\|\cdot\|^2$ denotes the squared Euclidean norm. The regressor matrix X is considered fixed and given in all of the following and therefore omitted in the notation. We define the matrices $Z = X'X$ and $\Sigma = n^{-1}Z$ which are assumed to be positive definite in order to guarantee uniqueness of maximum likelihood estimates. The maximum likelihood estimators of β and σ^2 are independent and given by

$$\hat{\beta}(y^n) = Z^{-1}X'y^n, \quad (4)$$

$$\hat{\sigma}^2(y^n) = \frac{1}{n}\|y^n - X\hat{\beta}'(y^n)\|^2. \quad (5)$$

Now, assume the vector y^n can be considered a series, i.e., the data points are ordered in a meaningful way. We can then obtain a regressor matrix X by various transformations of the index i of the y_i variables. Thus, we define for each $j \leq k$, $X_{i,j} = f_j(i)$, where f_j are arbitrary basis functions. One both theoretically and practically appealing way to define the functions f_j is to use a *wavelet basis*, see e.g., Daubechies (1992). By letting the regressor matrix be square, i.e., $k = n$, and taking as the basis functions $f_j(i)$ an appropriate wavelet basis, we get an *orthogonal* regressor matrix X, i.e., X has as its inverse the transpose X' and we have $Z = X^{-1}X = I$, where I is the identity matrix.

Instead of using all the basis vectors, we may also choose a subset γ of them. This gives the reconstructed version $\hat{y}^n_\gamma = X\hat{\beta}_\gamma(y^n)$, and the difference to the original signal is left to be modeled as noise. Since the basis is orthogonal, the maximum likelihood values of any subset of all the parameters are equal to the corresponding maximum likelihood parameters in the full model and one gets the parameter vector

$$\hat{\beta}_\gamma(y^n) = (\delta_i(\gamma)\hat{\beta}_i(y^n))',$$

where $\delta_i(\gamma)$ is equal to one if the index i is in the index set γ of retained coefficients and zero otherwise. The maximum likelihood estimator of the noise variance becomes

$$\begin{aligned}\hat{\sigma}^2_\gamma(y^n) &= \frac{1}{n}\|X\hat{\beta}'(y^n) - X\hat{\beta}'_\gamma(y^n)\|^2 \\ &= \frac{1}{n}\|\hat{\beta}(y^n) - \hat{\beta}_\gamma(y^n)\|^2,\end{aligned}$$

which is seen to be the sum of the discarded coefficients divided by n. We denote for convenience the squared norm of the maximum likelihood coefficient vector corresponding to γ by S_γ:

$$S_\gamma = \|\hat{\beta}_\gamma(y^n)\|^2 = \sum_{i\in\gamma}\beta_i^2.$$

The squared norm of the coefficient in the full model with $k = n$ is denoted simply by S. From orthogonality it follows that S is equal to the squared norm of the data $\|y^n\|^2$.

3.2 THE DENOISING PROBLEM

The denoising problem is now to choose a subset γ such that the retained coefficients would give a good reconstruction of the informative part of the signal while the discarded coefficients would contain as much of the noise in the signal as possible, The sparseness of wavelet representations, i.e., the fact that a large fraction of the coefficients are essentially zero in the 'noise-free' or informative part of the signal (see (Mallat, 1989)) makes it plausible to recover the informative part by identifying and discarding the coefficients that are likely to contain pure noise.

The idea of *wavelet thresholding* was proposed soon after Mallat's paper independently by Donoho & Johnstone (1991) and Weaver *et al.* (1991). In wavelet thresholding a threshold value is first determined and the coefficients whose absolute value is less than the threshold are discarded. Using the maximum likelihood estimates as the values of the retained coefficients is called *hard thresholding* while in *soft thresholding* the retained coefficients are also shrunk towards zero in order to reduce the noise distorting the informative coefficients.

In statistical wavelet denoising the denoising problem is often formalized using the concept of *statistical risk*, i.e., the expected distortion (usually squared error) of the reconstructed signal when compared to a true signal. This requires an assumed model typically involving i.i.d. noise added to a true signal. In the statistical approach the signal is considered deterministic and the worst-case risk over a class of signals is minimized while in the Bayesian approach (see, e.g., (Ruggeri & Vidakovic, 1999; Chang *et al.* , 2000)) a prior distribution on the true signal is postulated and the expected (Bayes) risk is minimized. Donoho & Johnstone (1994) have derived a set of wavelet denoising methods including the following hard threshold:

$$t_{DJ} = \sigma\sqrt{2\log n}, \tag{6}$$

where σ is the standard deviation of noise.

In order to apply the method in practice, one usually needs to estimate σ. Donoho & Johnstone suggest using as an estimator the median of the coefficients on the finest level divided by .6745 which usually works well as long as the signal is contained mainly in the low frequency coefficients. There are also several other, more refined denoising methods suggested by the mentioned authors and others but due to space limitations and the fact that our real focus is in understanding the behavior of MDL based denoising, these methods are not discussed in the current paper. Fodor & Kamath (2003) present an empirical comparison of different wavelet denoising methods; see also Ojanen *et al.* (2004) for a comparison of the Donoho-Johnstone method and MDL denoising.

4 MDL DENOISING

The MDL principle offers a different approach to denoising where the objective is to separate information and noise in the observed signal. Unlike in the statistical approach, information and noise are *defined* as the compressible and the incompressible part of the signal respectively, thus depending on the model class used for describing the signal.

4.1 MDL APPROACH TO DENOISING

One of the most characteristic features of the MDL approach to statistical modeling is that there is no need to assume a hypothetical generating model whose existence would be very hard to verify. Any background information regarding the phenomenon under study is incorporated in the choice of the model class. The only assumption is that at least one of the model classes under consideration allows compression of the data which is clearly much easier to accept than the assumption that the assumed model is indeed an exact replica of the true generating mechanism.

In denoising, MDL model selection is performed by considering each subset of the coefficients as a model class and minimizing the stochastic complexity of the data given the model class. Unfortunately, for wavelet based models and more generally, for linear regression models, the normalizer in the NML density is unbounded and NML is not defined unless the range of the data is restricted. The problem can be solved by resorting to universal models other than NML, such as two-part or mixture models in defining the stochastic complexity. Hansen & Yu (2000) propose a combination of two-part and mixture codes for wavelet denoising. Their method also includes an estimation step similar to the one used by Donoho & Johnstone, and is thus not completely faithful to the MDL philosophy.

4.2 RENORMALIZED NML

Rissanen (2000) solves the problem of unbounded parametric complexity by two-fold normalization. The data range is first restricted such that the squared (Euclidean) norm of the maximum likelihood values of the wavelet coefficients $\|\hat{\beta}_\gamma(y^n)\|^2$ is always less than some maximal value R and the maximum likelihood variance $\hat{\sigma}^2_\gamma(y^n)$ is greater than some minimal value σ_0^2. We then obtain an NML density with limited support for each pair (R, σ_0^2). It is now possible to construct a 'renormalized' or 'meta' NML density by taking the obtained NML densities as a new model class[1].

After the application of Stirling's approximation to gamma functions and ignoring constant terms it can be shown that the code-length to be minimized becomes[2]

$$\frac{(n-k)}{2}\ln\frac{S-S_\gamma}{n-k} + \frac{k}{2}\ln\frac{S_\gamma}{k} + \frac{1}{2}\ln(k(n-k)). \tag{7}$$

[1] In fact even the renormalization requires the data range to be restricted but it turns out that the final range doesn't affect the resulting criterion.

[2] Multiplying the code-length formula by two gives an equivalent minimization problem. Note the last term that was incorrect in some of the earlier publications.

It can be shown that the criterion is always maximized by choosing γ such that either the k largest or the k smallest coefficients are retained for some k. We consider this an artefact of the renormalization performed and assume in the what follows that the k largest coefficients are retained. We return to the issue in Sec. 5.3.

4.3 PRACTICAL ISSUES

We point out two issues of a rather technical nature that nevertheless deserve to be noted by practitioners since we have found them to result in very poor performance in more than one case. First, in all wavelet thresholding methods, it should be made sure that the wavelet transform used is such that the coefficients are scaled properly, in other words, that the corresponding basis is orthogonal. This is essential for all wavelet thresholding methods. It is easy to check that the sum of squares of the original data and the transformed coefficients are always equal.

Secondly, since the criterion is derived for continuous data and involves densities, problems may occur when it is applied to low-precision or discrete, say integer, data. If the data can be represented exactly by some number k_0 of coefficients, the criterion becomes minus infinity for all $k \geq k_0$ because the first term includes a logarithm of zero. Also, for k almost as large as k_0 the criterion takes a very small value and such a value of k is often selected as the optimal one potentially resulting in severe over-fitting. This problem may either be solved by using a lower bound for $(S - S_\gamma)/(n - k)$ corresponding to a lower bound on the variance. Alternatively, once a sudden drop to minus infinity in the criterion is recognized it is possible to reject all values of k that are near the point where the drop occurs.

5 BEHAVIOR OF MDL DENOISING

By inspecting the behavior of the MDL denoising criterion as a function of noise variance, we are able to give a rough characterization of its domain of applicability. This makes way towards a more important goal, the understanding of renormalized NML, and potential ways of generalizing and improving it.

5.1 EMPIRICAL OBSERVATIONS

Fig. 1 illustrates the behavior of the MDL denoising method and the method by Donoho & Johnstone described in Sec. 3 with Daubechies N=4 wavelet basis. The original image is distorted by Gaussian noise to get a noisy signal. When there is little noise, the difference is small, MDL method performing better in terms of standard error. However, when there is much noise the methods produce very different results. The Donoho-Johnstone method retains only 0.3 percent of the coefficients while the MDL method retains 46.9 percent of them, the former giving a better result in terms of standard error.

Figure 1: Lena Denoised. *Top left:* Noisy image ($\sigma = 10.0$); *middle left:* Donoho-Johnstone (2.2 % retained, std. error 8.1); *bottom left:* MDL (7.6 % retained, std. error 6.8); *top right:* Noisy image ($\sigma = 47.5$); *middle right:* Donoho-Johnstone (0.3 % retained, std. error 17.3); *bottom right:* MDL (46.9 % retained, std.error 44.9).

The effect of the standard deviation of noise on the behavior of the two methods can be clearly seen in Fig. 2. It can be seen that the MDL method outperforms the Donoho-Johnstone method when the noise standard deviation is less than 15. However, outside this range the performance of the MDL method degrades linearly

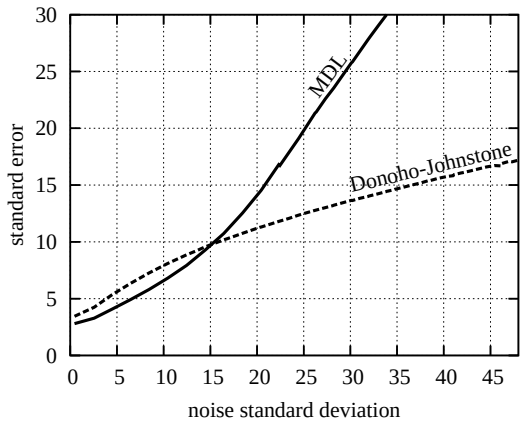

Figure 2: Effect of noise.

due to retaining too many coefficients. The standard error of the noise should be compared to the standard deviation of the original signal which in this case was 46.6. Experiments with other natural images indicate that the standard deviation of the signal determines the scale but does not affect the shape of the curves. As a rough characterization of the domain of applicability of the MDL method it can be said that the noise standard deviation should be at most half of the standard deviation of the signal.

5.2 THEORETICAL ANALYSIS

The degradation of performance of the MDL denoising criterion is underlined when the noise variance is very large. This can be demonstrated theoretically by considering what happens when the noise variance grows without bound so that in the limit the signal is pure Gaussian noise. Since the criterion is scale invariant we may without loss of generality assume unit variance. Essentially, we need to evaluate the asymptotics of S_k, the squared sum of the k largest coefficients in absolute value. Let $\beta_{i_1}^2 \leq \beta_{i_2}^2 \leq ... \leq \beta_{i_n}^2$ be the squared coefficients ordered in ascending order. We have

$$S_k = \sum_{j=n-k+1}^{n} \beta_{i_j}^2 = \sum_{\beta_i^2 \geq t_k^2} \beta_i^2,$$

where we assumed that the first retained coefficient $t_k := \beta_{i_{n-k+1}}$ is unique. If we consider t_k a fixed parameter instead of a random variable, the terms in the above sum are independent with expectation given by:

$$E[\beta_i^2 \mid \beta_i \geq t_k] = \frac{1}{1-\Phi(t_k)} \int_{t_k}^{+\infty} \frac{x^2 e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} dx,$$

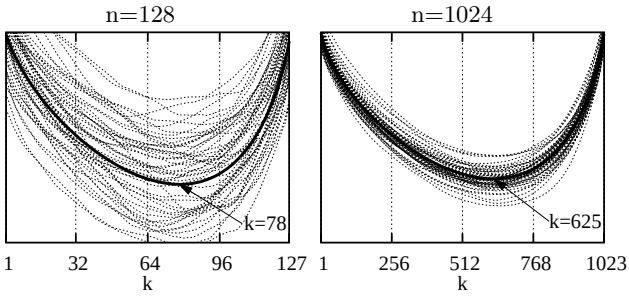

Figure 3: The renormalized NML denoising criterion with pure Gaussian noise.

where the expectation is taken with respect to the standard normal distribution whose distribution function is denoted by Φ. The integral is given by

$$\int \frac{x^2 e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} dx = \frac{-x e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} + \Phi(x),$$

and the expectation becomes

$$E[\beta_i^2 \mid \beta_i \geq t_k] = \frac{t_k e^{-\frac{t_k^2}{2}}}{\sqrt{2\pi}(1-\Phi(t_k))} + 1. \qquad (8)$$

Now in order to contain a k/n fraction of Gaussian random variates as n goes to infinity, the limiting value of the cut-point t_k must be

$$\lim_{n\to\infty} t_k = \Phi^{-1}\left(1 - \frac{k}{2n}\right).$$

(Division of k by two comes from the fact that also negative coefficients with large absolute value are included.) Plugging this into Eq. (8) in place of t_k gives the asymptotic behavior of the average S_k/k. Since the expectation of all coefficients under the unit variance Gaussian noise model is equal to one, the expectation of $(S_n - S_k)/(n-k)$, i.e., the expectation of the $n-k$ *smallest* squared coefficients can be easily obtained once the expectation of the k largest coefficients is known.

Fig. 3 shows the values of the renormalized NML denoising criterion with sample sizes $n = 128$ (on the left), and $n = 1024$ (on the right), with 50 repetitions in each case. Data is pure Gaussian noise with unit variance. The theoretical minima for the two samples sizes are $k = 78$ and $k = 625$ respectively. The asymptotic curve is plotted with a solid line. By evaluating the criterion for large n it can be seen that the MDL method tends to keep about $625/1024 \approx 61$ % of the coefficients. This is suboptimal in terms of both statistical risk and the natural meaning of information

and noise in data. If all data is indeed pure noise the method should indicate that there is no information in the data at all.

5.3 INTERPRETATION

Let us now consider the interpretation of the renormalized NML denoising criterion in order to understand the above described behavior. The code-length function (7) is the negative logarithm of a corresponding density of the following form (ignoring normalization constants):

$$(S - S_\gamma)^{-(n-k)/2} S_\gamma^{-k/2} = \|\hat{\beta}_{\gamma^c}\|^{-(n-k)} \|\hat{\beta}_\gamma\|^{-k}, \quad (9)$$

where γ^c denotes the complement of γ, i.e, the set of $n-k$ discarded coefficients.

Incidentally, the form in Eq. (9) is equivalent to using a zero-mean Gaussian density with optimized variance for both the retained and the discarded coefficients. This can be seen as follows. Given a vector x of k random variates, the maximal density achievable with a zero-mean Gaussian density assuming the entries in the vector are independent is given by

$$\max_\sigma (2\pi\sigma^2)^{-k/2} \exp\left(-\frac{\|x\|^2}{2\sigma^2}\right) = \left(2\pi e k^{-1} \|x\|^2\right)^{-k/2} \quad (10)$$

which is seen to be proportional to $\|x\|^{-k}$. Thus the two factors in Eq. (9) correspond to maximized Gaussian densities of the kind in (10). Fig. 4 gives an illustration verifying that the threshold is at the intersection points of two Gaussian densities fitted to the discarded and the retained coefficients respectively. The latter density has very high variance because the empirical distribution of the coefficients has heavy tails. The fact that both retained and discarded coefficients are encoded with a Gaussian density explains many aspects of the behavior reported above.

It is quite easy to derive rough conditions on when the criterion performs well. From orthogonality of the wavelet transform it follows that each of the informative coefficients is a sum of an information term and a noise term. Assuming independent noise, the density of the sum is given by the convolution of the densities of the summands. For instance, if the original signal has Gaussian density, the convolution is Gaussian as well with variance equal to the sum of the signal variance σ_S^2 and the noise variance σ_N^2. As long as the signal variance is large compared to the noise variance, the variance of the informative coefficients, $\sigma_S^2 + \sigma_N^2$, is significantly larger than that of the noise coefficients. Consequently, the criterion based on Gaussian densities with different variances is able to separate the informative and non-informative coefficients as long as the noise variance is not too high.[3] It is also easy to understand that fitting two Gaussian densities to a single one gives nonsensical results which explains the behavior in the pure noise scenario of Sec. 5.2.

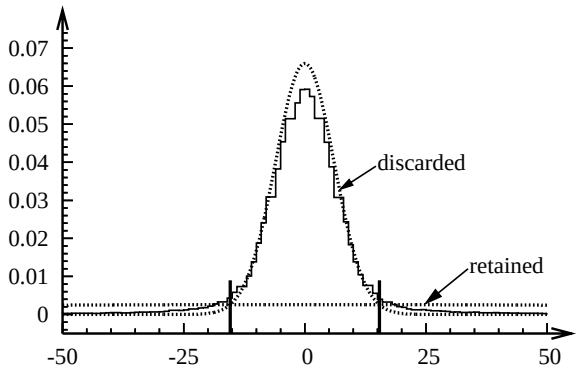

Figure 4: Gaussian densities fitted to noisy Lena ($\sigma = 10.0$). The empirical histogram is plotted with solid line. Gaussian densities with variance adjusted for the discarded ($\hat{\sigma} = 6.0$) and the retained ($\hat{\sigma} = 153.7$) coefficients are shown with dotted curves. Threshold is at ± 15.4.

It has been observed that wavelet coefficients in natural images tend to be well modeled by generalized Gaussian densities of the form $K \exp(-(|x|/\alpha)^\beta)$ where K is a normalization constant (Mallat, 1989). The typical values of β are near one which corresponds to the Laplacian (double exponential) density. This suggests that the density of the observed coefficients can be modeled by a convolution of the Laplace and Gaussian densities. Ruggeri & Vidakovic (1999) consider Bayes optimal hard thresholding in this model when the scale parameters of both densities are known. Chang *et al.* (2000) estimate the scale parameters from the observed signal. The construction of an NML model based on Laplacian and generalized Gaussian models with a proper treatment of the scale parameters is an interesting future research topic.

[3] Similar reasoning also shows that while the criterion is symmetric in the two sets of coefficients, one should always retain the k largest coefficients instead of the k *smallest* coefficients.

6 CONCLUSIONS

In its general form, the MDL principle essentially aims at separating meaningful information from noise, and thus provides a very natural approach to denoising as an alternative to the statistical and Bayesian approaches. There are, however, some intricate issues in applying MDL to the denoising problem related to unbounded parametric complexity of Gaussian families. We discussed a solution by Rissanen involving a renormalization whose effect has been unclear so far and is of considerable interest not only in denoising applications but in the MDL framework in general.

The reported empirical and theoretical findings suggested a characterization of the domain of applicability for Rissanen's denoising method. It was seen that over-fitting is likely in the high noise regime. For practitioners, we pointed out two technical pitfalls and ways to avoid them. We gave an interpretation of the renormalization by showing that it results in a code based on two Gaussian densities, one for the retained wavelet coefficients and one for the discarded ones. Based on the interpretation we were able to explain both the empirical and the theoretical findings. The interpretation also facilitates understanding of the problem of unbounded parametric complexity in general and suggests generalizations of the renormalization procedure, potentially leading to improved MDL methods for denoising as well as other applications.

Acknowledgments

We thank Jorma Rissanen, Peter Grünwald, and Ursula Gather for useful discussions. This work was supported in part by the Academy of Finland under projects Minos and Cepler, the Finnish Technology Agency under project PMMA, and the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views.

References

Abramovich, F., Bailey, T., & Sapatinas, T. 2000. Wavelet analysis and its statistical applications. *Journal of the Royal Statistical Society, Series D*, **49**(1), 1–29.

Barron, A., Rissanen, J., & Yu, B. 1998. The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, **44**(6), 2743–2760.

Chang, G., Yu, B., & Vetterli, M. 2000. Adaptive wavelet thresholding for image denoising and compression. *IEEE Transactions on Image Processing*, **9**(9), 1532–1546.

Daubechies, I. 1992. *Ten Lectures on Wavelets*. Society for Industrial & Applied Mathematics (SIAM), Philadelphia, PA.

Donoho, D., & Johnstone, I. 1991. *Minimax estimation via wavelet shrinkage*. Tech. rept., Stanford University.

Donoho, D., & Johnstone, I. 1994. Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, **81**(3), 425–455.

Fodor, I.K., & Kamath, C. 2003. Denoising through wavelet shrinkage: an empirical study. *Journal of Electronic Imaging*, **12**(1), 151–160.

Foster, D.P., & Stine, R.A. 2001. The competitive complexity ratio. *Proceedings of the 2001 Conference on Information Sciences and Systems*, 1–6.

Foster, D.P., & Stine, R.A. 2005. The contribution of parameters to stochastic complexity. *To appear in:* Grünwald, P., Myung, I.J., & Pitt, M. (eds), *Advances in Minimum Description Length: Theory and Applications*, MIT Press, Cambridge, MA.

Grünwald, P. 2005. A Tutorial introduction to the minimum description length principle. *To appear in:* Grünwald, P., Myung, I.J., & Pitt, M. (eds), *Advances in Minimum Description Length: Theory and Applications*, MIT Press, Cambridge, MA.

Hansen, M., & Yu, B. 2000. Wavelet thresholding via MDL for natural images. *IEEE Transactions on Information Theory*, **46**(7), 1778–1788.

Mallat, S. 1989. A Theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11**(7), 674–693.

Ojanen, J., Miettinen, T., Heikkonen, J., & Rissanen, J. 2004. Robust denoising of electrophoresis and mass spectrometry signals with minimum description length principle. *FEBS Letters*, **570**(1–3), 107–113.

Rissanen, J. 1996. Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, **42**(1), 40–47.

Rissanen, J. 2000. MDL denoising. *IEEE Transactions on Information Theory*, **46**(7), 2537–2543.

Ruggeri, F., & Vidakovic, B. 1999. A Bayesian decision theoretic approach to the choice of thresholding parameter. *Statistica Sinica*, **9**(1), 183–197.

Shtarkov, Yu M. 1987. Universal sequential coding of single messages. *Problems of Information Transmission*, **23**(3), 3–17.

Weaver, J.B., Yansun, X., Healy, D.M. Jr., & Cromwell, L.D. 1991. Filtering noise from images with wavelet transforms. *Magnetic Resonance in Medicine*, **21**(2), 288–295.