MDL Denoising

J. Rissanen
IBM Research Division
Almaden Research Center, DPE-B2/802
San Jose, CA 95120-6099, rissanen@almaden.ibm.com

1/12/1999

Abstract

The so-called denoising problem, relative to normal models for noise, is formalized such that 'noise' is defined as the incompressible part in the data while the compressible part defines the meaningful information bearing signal. Such a decomposition is effected by minimization of the ideal code length, called for by the *Minimum Description Length* (*MDL*) principle, and obtained by an application of the *normalized maximum likelihood* technique to the primary parameters, their range, and their number. For any orthonormal regression matrix, such as defined by wavelet transforms, the minimization can be done with a threshold for the squared coefficients resulting from the expansion of the data sequence in the basis vectors defined by the matrix.

keywords: linear regression, wavelet transforms, threshold, stochastic complexity, Kolmogorov sufficient statistics

1 Introduction

Intuitively speaking the so-called 'denoising' problem is to separate an observed data sequence $x_1, x_2, \ldots, x_n$ into a 'meaningful' signal $\hat{x}_i$ and the remaining 'noise' e_i thus $x_i = \hat{x}_i + e_i$. Taking the traditional approach Donoho and Johnstone in the pioneering paper [4] on application of wavelets to statistics posed this as the problem of estimating a function $f(t_i)$ from its noisy values

$$x_i = f(t_i) + \epsilon_i, \tag{1}$$

where the noise ϵ_i is a normal, independent, identically distributed (iid) sequence of zero mean and known variance $\tau = \sigma^2$. With the notations $f(t_i) = \bar{x}_i$ and $\hat{f}(t_i) = \hat{x}_i$ for the values of the function and its estimates, respectively, the traditional criterion to assess the goodness of the various estimators is the risk

$$R(\hat{f}, f) = R(\hat{x}, \bar{x}) = n^{-1} E \|\hat{x} - \bar{x}\|^2,$$

where $\hat{x} = \hat{x}_1, \ldots, \hat{x}_n$ and similarly $\bar{x}$ denote the sequences of the estimated and modeled numbers, respectively. Taking advantage of the special property of orthonormal wavelet bases that the coefficients with small absolute value tend to be attributed to the noise, the authors converted the

problem into one of finding a threshold such that all coefficients whose square exceeds the threshold are retained and others set to zero. By quite ingenious arguments they showed that the very simple threshold $2\tau \ln n$ will have the minmax property that for the worst case signal $\bar{x}_i$ the asymptotically minimum risk $2\tau \ln n$ is attained; see also [5]. Moreover, this was turned into a universal threshold by a procedure to estimate the variance from the portion of the $n/2$ coefficients of the basis vectors corresponding to the finest resolution in the wavelet transform.

Despite its simplicity the authors' threshold works well for data where the assumed mean $\bar{x}$ consists of piecewise smooth segments. Even when the break points are unknown, the risk is not appreciably larger than if an 'Oracle' would tell the true break points. However, if there is no 'true' signal of the kind assumed, the signal recovered by the procedure tends to be too smooth. This happens for instance in speech data, where the information bearing signal has rapidly changing large swings as well as even more rapidly changing components mixed with noise. We also find the traditional approach with its risk untenable in principle, because it needs two 'smooth' signals, only one of which, $\hat{x}$, depends on the data. Hence, there are really two ideas of noise and its estimated variance involved: the relevant one $e_i = x_i - \hat{x}_i$ and another imagined one $\epsilon_i = x_i - \bar{x}_i$. The two cannot be equated, because the risk would vanish and we end up in circular reasoning: the noise gets defined by the estimated signal $\hat{x}$, which, in turn, depends on the estimated variance of the noise. Clearly, there cannot be any unique way to imagine or model a signal $\bar{x}_i$, which means that any estimate of its variance must be arbitrary.

In [7] a different approach to the denoising problem was studied based on an early version of the formula for the shortest code length for the observed sequence required in the *MDL* principle, [10]. In the notations above it is the same as that for the deviations $x_i - \hat{x}_i = e_i$, which is determined by the way they are modeled. Two types of distributions for them were considered, the first being normal with zero mean and known variance and the second Huber's $\epsilon-$contaminated normal distribution, also of zero mean and known variance, which was shown to satisfy a maximum entropy property. As in [4] the required noise variance cannot be estimated from the squared difference between the data and the recovered signal. Interestingly enough, the minimizing number of the coeffients was shown to define a threshold, which for normally modeled noise reduces asymptotically to the minmax threshold of Donoho and Johnstone. The authors also suggest the use of the criterion to find the best wavelet basis within a parametric family.

Inspired by the two papers mentioned we describe a different basis for the denoising problem: noise is *defined* to be that part of the data that cannot be compressed with the models considered, while the rest defines the meaningful information bearing signal, which need not be smooth. In fact, rapidly changing data may well be compressible, in which case they should not be eliminated as noise. This happens for instance in speech data as illustrated below. In the linear regression with normally distributed noise the decomposition required can be found by minimization of a criterion,

defined by the negative logarithm of a *Normalized Maximum Likelihood* (*NML*) density function, for which an exact formula exists even nonasymptotically. Such an *NML* criterion is derived by a two-fold extension of the normalization procedure done by Dom, [3], which, in turn, sharpens the asymptotic formula in [9] that is applicable to more general parametric model classes. The resulting decomposition is similar to Kolmogorov's sufficient statistics in the algorithmic theory of information, [2],[1], [10], [11], and it will also be seen to extend the usual sufficient statistics decomposition of parametric likelihood functions of exponential type.

Because the *NML* criterion involves the sum of the squares of both the residuals, which define the ML estimate of the noise variance, and the constructed signal, it is no longer obvious that the optimal decomposition can be found by any threshold. However, we prove that one still exists so that the procedure for finding the information bearing signal remains simple. Under suitable assumptions the threshold turns out to behave asymptotically like one half of the threshold in [4] and [7], so that the minmax bound for the risk in the former reference will not be reached. This is, however, of no consequence, since the minmax bound is reached only by pathological signals which are unbounded. The same criterion can also be applied to find the best data driven wavelet basis among a parametric family.

Finally, we add, in order to avoid misunderstandings, that the purpose of this paper is to derive exact nonasymptotic formulas for Gaussian models arising in linear-quadratic regression problems. Unlike in earlier attempts these formulas involve no arbitrarily selected hyperparameters, and they split the data in a natural manner into the noise and the signal that has all the useful information that can be extracted with the model class considered. The question, however, whether such models are even adequate for all denoising problems, let alone the best, is completely beyond the scope of this paper.

2 The *NML* Criterion for Linear Regression

The denoising problem is a special instance of the familiar linear regression problem, where each observed number x_i for $i = 1, \ldots, n$ is modeled as

$$x_i = \sum_{j=1}^{k} \beta_j z_{ji} + \epsilon_i, \tag{2}$$

and ϵ_i is taken as having a normal distribution with zero mean and variance $\tau = \sigma^2$. The entries z_{ji} define a $k \times n$ regression matrix Z'. Usually this matrix is a submatrix of a bigger $n \times n$ regressor matrix W, defined by the rows with indices in a set $\gamma = \{i_1, \ldots, i_k\}$.

Extending such a model by independence to sequences $\mathbf{x}' = (x_1, \ldots, x_n)$, viewed as a row vector,

we get the density function

$$f(\mathbf{x};\gamma,\beta,\tau) = \frac{1}{(2\pi\tau)^{n/2}} e^{-\frac{1}{2\tau}\sum_t (x_t - \beta' \mathbf{z}_t)^2}, \tag{3}$$

where $\beta' = \beta_1, \ldots, \beta_k$ and $\mathbf{z}_t$ is the t'th column of Z'. Write $\Sigma = Z'Z$, which is taken to be positive definite.

The maximum likelihood solution of the parameters is given by

$$\begin{aligned} \hat{\beta}(\mathbf{x}) &= \Sigma^{-1} Z' \mathbf{x} \\ \hat{\tau}(\mathbf{x}) &= \frac{1}{n} \sum_t (x_t - \hat{\beta}'(\mathbf{x}) \underline{z}_t)^2. \end{aligned} \tag{4}$$

In the following three subsections we describe three universal *NML* density functions of increasing degree of universality, the last giving the final criterion. We also use the same letter 'f' and the notation of type $f(\mathbf{x};\cdot)$ where the dot is replaced by parameters of various kind. Clearly, then density functions so denoted are different depending on the nature and the number of the parameters involved. This should not cause confusion, for the nature of the parameters will indicate which density function is being discussed.

2.1 First Level Universal

We begin by the derivation of the *NML* density function, for which an asymptotic formula for general parametric model classes was given in [9]. It was recognized by Dom that for the normal linear regression models one could derive an exact formula, [3], which, however, includes certain hyperparameters to be chosen properly. Our derivation, which is simpler, was also done in [1], but we repeat the steps needed for further development. The *NML* density function is defined as

$$\hat{f}(x;\gamma) = \frac{f(\mathbf{x};\gamma,\hat{\beta}(\mathbf{x}),\hat{\tau}(\mathbf{x}))}{\int_{Y(\tau_0,R)} f(\mathbf{y};\gamma,\hat{\beta}(\mathbf{y}),\hat{\tau}(\mathbf{y}))d\mathbf{y}}, \tag{5}$$

where $\mathbf{y}$ is restricted to the set

$$Y(\tau_0, R) = \{\mathbf{y} | \hat{\tau}(\mathbf{y}) \geq \tau_0, \hat{\beta}'(\mathbf{y})\Sigma\hat{\beta}(\mathbf{y}) \leq nR\}. \tag{6}$$

In this the parameters τ_0 and R are such that the ML estimates fall within $Y(\tau_0, R)$.

It was shown in [13] that a normalization process like in (5) gives $\hat{f}(\mathbf{x};\gamma)$ as the solution to the minmax problem

$$\min_q \max_{\mathbf{x}} \ln \frac{f(\mathbf{x};\gamma,\hat{\beta}(\mathbf{x}),\hat{\tau}(\mathbf{x}))}{q(\mathbf{x})}.$$

The same density function was recently shown to solve even the following minmax problem

$$\min_q \max_g E_g \log \frac{f(X^n;\hat{\theta}(X^n),\gamma)}{q(X^n)}, \tag{7}$$

where q and g range over any distributions, [12]. Hence, in partiular g is not restricted to be of the normal type (3). In words: $\hat{f}(x^n;\gamma)$ is the unique density function which minimizes over all density functions $q(x^n)$ the mean ideal code length difference between $-\log q(x^n)$ and the shortest $-\log f(x^n;\hat{\theta}(x^n),\gamma)$ we can ask for, the mean taken with respect to the worst case data generating distribution. We call the negative logarithms of densities 'ideal' code lengths, because they can be converted into integer-valued code lengths as upper bounds for negative logarithms of probabilities induced on the data quantized to a finite precision.

The numerator in (5) is given by

$$f(\mathbf{x};\gamma,\hat{\beta}(\mathbf{x}),\hat{\tau}(\mathbf{x})) = 1/(2\pi e\,\hat{\tau}(\mathbf{x}))^{n/2}, \tag{8}$$

and we need to evaluate the denominator. When the data are generated by (3) the maximum likelihood estimates (4) are sufficient statistics, and we have the factorization

$$f(\mathbf{x};\gamma,\beta,\tau) = f(\mathbf{x}|\hat{\beta}(\mathbf{x}),\hat{\tau}(\mathbf{x}))g_1(\hat{\beta}(\mathbf{x});\beta,\tau)g_2(\hat{\tau};\tau), \tag{9}$$

where the conditional density function in the first factor does not depend on the parameters. The second factor is the normal density function with mean β and covariance matrix $\tau\Sigma^{-1}$, and the third is induced by the χ^2- density function for $n\hat{\tau}(\mathbf{x})/\tau$ with $n-k$ degrees of freedom.

We can evaluate the denominator in (5) by integrating the conditional density function over its range, while keeping $\hat{\beta}(\mathbf{y}) = \hat{\beta}$ and $\hat{\tau}(\mathbf{y}) = \hat{\tau}$ at fixed values, which gives unity, and then integrating the product $g_1(\hat{\beta};\hat{\beta},\hat{\tau})g_2(\hat{\tau};\hat{\tau})$ over the range of $\hat{\beta}$ and $\hat{\tau}$. The product turns out to be a function of $\hat{\tau}$ only, and it is given by

$$g(\hat{\tau}) = A_{n,k}\hat{\tau}^{-k/2-1}, \tag{10}$$

where

$$A_{n,k} = \frac{|\Sigma|^{1/2}}{(\pi n)^{k/2}\Gamma(\frac{n-k}{2})}(\frac{n}{2e})^{n/2}. \tag{11}$$

The integration gives then

$$\begin{aligned} C(\tau_0,R) &= \int_{\tau_0}^{\infty} d\hat{\tau} \int_{\hat{\beta}'\Sigma\hat{\beta}\le nR} g(\hat{\tau})d\hat{\beta} && (12)\\ &= A_{n,k}V_k\frac{2}{k}\left(\frac{R}{\tau_0}\right)^{k/2}, && (13) \end{aligned}$$

where

$$V_k R^{k/2} = |\Sigma|^{-1/2}\frac{2\pi^{k/2}(nR)^{k/2}}{k\Gamma(k/2)} \tag{14}$$

is the volume of the ellipsoid defined by Σ and R. Finally, the logarithm of the *NML* density function for k such that $0 < k < n$ is given by

$$-\log\hat{f}(\mathbf{x};\gamma,\tau_0,R) = \frac{n}{2}\ln\hat{\tau} + \frac{k}{2}\ln\frac{R}{\tau_0} - \ln\Gamma(\frac{n-k}{2}) - \ln\Gamma(\frac{k}{2}) + \ln\frac{4}{k^2} + \frac{n}{2}\ln(n\pi). \tag{15}$$

We conclude this subsection by extending Equation (15) for $k = 0$. In this case write the *NML* density function as $\hat{f}(\mathbf{x}; 0)$, and the same technique gives

$$-\ln \hat{f}(\mathbf{x}; 0) = \frac{n}{2} \ln(n\pi\hat{\tau}) - \ln \Gamma(\frac{n-1}{2}) - \frac{1}{2} \ln((n-1)\pi) - \ln \mathcal{S}_{n-1}(t) + \ln \ln \frac{\tau_1}{\tau_0}, \tag{16}$$

where $\mathcal{S}_{n-1}(t)$ for $t = \sqrt{n-1}\bar{x}/\sqrt{\hat{\tau}}$ is Student's *t-distribution* of $n-1$ degrees of freedom and τ_1 is another parameter greater than τ_0; $\bar{x}$ is the arithmetic mean of the data. The values of the parameters will not be needed.

When looking for the optimal value for k among $k > 0$ the norm of the parameter vectors $\hat{\beta}(\mathbf{x})$ affects the value of the criterion but otherwise is of no particular interest. However, if by chance the optimal value is 1, and we need to check if the value $k = 0$ is even better, we should examine the situation for small values of the single parameter against its value being 0. This means that we should not compare $\hat{f}(\mathbf{x}; \gamma, \tau_0, R)$, as obtained from (15) for $k = 1$, with (16), but we should recompute it for $k = 1$ for a small value of R, which will increase it and make it more competitive against (16). Much as in hypothesis testing with the *NML* criterion, [11], an appropriate value for the range is obtained with $R = c^2\hat{\tau}/(n-1)$, where c is a positive constant to be determined presently. Then writing the *NML* density function for $k = 1$ as $\hat{f}(\mathbf{x}; 1)$, we get with the same technique as above the simple result

$$\frac{\hat{f}(\mathbf{x}; 0)}{\hat{f}(\mathbf{x}; 1)} = 2c\mathcal{S}_{n-1}(t). \tag{17}$$

The density $\hat{f}(\mathbf{x}; 1)$ is maximized for the positive square root of $\hat{c}^2 = (n-1)\hat{\beta}^2\Sigma/(n\hat{\tau})$. We then pick the optimal number $\hat{k} = 0$, if the ratio is greater than or equal to unity. Otherwise, the winner is $\hat{k} = 1$.

2.2 Second Level Universal

We wish to get rid of the two parameters R and τ_0, which clearly affect the criterion (15) in an essential manner, or rather we replace them with other parameters which do not influence the relevant criterion. In [11] and [6] this was done simply by setting the two parameters to the values that minimize (15): $R = \hat{R}$, and $\tau_0 = \hat{\tau}$, where $\hat{R} = n^{-1}\hat{\beta}'(\mathbf{x})\Sigma\hat{\beta}(\mathbf{x})$. However, the resulting $\hat{f}(\mathbf{x}; \gamma, \hat{\tau}(\mathbf{x}), \hat{R}(\mathbf{x}))$ is not a density function. We can of course correct this by multiplying it by a prior $w(\hat{\tau}(\mathbf{x}), \hat{R}(\mathbf{x}))$, but the result will be a density function on the triplet $x, \hat{\tau}(\mathbf{x}), \hat{R}(\mathbf{x})$, which is not quite right. We proceed instead by the same normalization process as above:

$$\hat{f}(\mathbf{x}; \gamma) = \frac{\hat{f}(\mathbf{x}; \gamma, \hat{\tau}(\mathbf{x}), \hat{R}(\mathbf{x}))}{\int_Y \hat{f}(\mathbf{y}; \gamma, \hat{\tau}(\mathbf{y}), \hat{R}(\mathbf{y}))d\mathbf{y}}, \tag{18}$$

where the range Y will be defined presently. By (9) and the subsequent equations we also have the factorization

$$\hat{f}(\mathbf{x};\gamma,\tau_0,R) = f(\mathbf{x}|\gamma,\hat{\beta},\hat{\tau})g(\hat{\tau})/C(\tau_0,R) = f(\mathbf{x}|\gamma,\hat{\beta},\hat{\tau})\frac{k}{2}\hat{\tau}^{-k/2-1}V_k^{-1}\left(\frac{\tau_0}{R}\right)^{k/2}. \tag{19}$$

As above we can now integrate the conditional while keeping $\hat{\beta}$ and $\hat{\tau}$ constant, which gives unity. Then by setting $\tau_0 = \hat{\tau}$ and $R = \hat{R}$ we integrate the resulting function $1/\hat{\tau}$ of $\hat{\tau}$ over a range $[\tau_1,\tau_2]$. Finally, noticing that the remaining function of $\hat{\beta}$ is constant, namely $n\hat{R}$, on the surface of the ellipsoid $\hat{\beta}'\Sigma\hat{\beta} = n\hat{R}$, its integral amounts to the integration of the surface area multiplied by $\hat{R}^{-k/2}$ with respect to $\hat{R}$ over a range $[R_1,R_2]$. All told we get

$$\int_Y \hat{f}(\mathbf{y};\gamma,\hat{\tau}(\mathbf{y}),\hat{R}(\mathbf{y}))d\mathbf{y} = \frac{k}{2}\ln\frac{\tau_2 R_2}{\tau_1 R_1}, \tag{20}$$

where the parameters are such that the ranges they define include $\hat{\tau}(\mathbf{x})$ and $\hat{R}(\mathbf{x})$, respectively. The negative logarithm $-\ln\hat{f}(\mathbf{x};\gamma)$ is then given by

$$-\ln\hat{f}(\mathbf{x};\gamma) = \frac{n-k}{2}\ln\hat{\tau} + \frac{k}{2}\ln\hat{R} - \ln\Gamma(\frac{n-k}{2}) - \ln\Gamma(\frac{k}{2}) + \ln\frac{2}{k} + \frac{n}{2}\ln(n\pi) + \ln\ln\frac{\tau_2 R_2}{\tau_1 R_1}. \tag{21}$$

Because the last term does not depend on γ nor k, we do not indicate the dependence of $\hat{f}(\mathbf{x};\gamma)$ on the new parameters.

2.3 Third Level Universal

If we intended to use the criterion (21) to find the optimal index set γ and their number k, we would have to add a code length of the optimal index set γ, and the result would not be the shortest code length for the data alone. Therefore, we extend the density function $\hat{f}(\mathbf{x};\gamma)$ to the larger class of models, defined as the union over all submatrices Z of the fixed W, and to obtain a criterion for finding the optimal index set γ and the associated optimal model. There are two basic cases to consider, one in which a prior $w(\gamma)$ for the index γ in some set Ω exists to be taken advantage of, and the other where only the range Ω is selected. In the former case, the optimal index $\bar{\gamma}$ is the one that maximizes the product $\hat{f}(\mathbf{x};\gamma)w(\gamma)$, and the resulting parameter-free density function for the model class, which now includes w, is given by

$$\bar{f}(\mathbf{x};\Omega) = \frac{\hat{f}(\mathbf{x};\bar{\gamma}(\mathbf{x}))}{\bar{P}_n(\bar{\gamma}(\mathbf{x}))}w(\bar{\gamma}(\mathbf{x})), \tag{22}$$

where

$$\bar{P}_n(\gamma) = \int_{\{\mathbf{y}:\bar{\gamma}(\mathbf{y})=\gamma\}} \hat{f}(\mathbf{y};\bar{\gamma}(\mathbf{y}))d\mathbf{y} \tag{23}$$

is a normalization needed to obtain a density function.

In the second case, which is far more common, no meaningful prior exists, and we construct a parameter-free *NML* density function for the data by imitating the technique above. We begin with the *MDL* estimator $\hat{\gamma}(\cdot)$, obtained by minimizing the ideal code length for the data $-\ln \hat{f}(\mathbf{x};\gamma)$ with respect to γ. Although the result $\hat{f}(\mathbf{x};\hat{\gamma}(\mathbf{x}))$ is not a density function we get one by the normalization process

$$\hat{f}(\mathbf{x};\Omega) = \frac{\hat{f}(\mathbf{x};\hat{\gamma}(\mathbf{x}))}{\int_{\hat{\gamma}(\mathbf{y})\in\Omega} \hat{f}(\mathbf{y};\hat{\gamma}(\mathbf{y}))d\mathbf{y}}, \tag{24}$$

where Ω is a set of indices such that it includes $\hat{\gamma}(\mathbf{x})$. The denominator in (24), call it C, is given by

$$C = \sum_{\gamma\in\Omega} \hat{P}_n(\gamma), \tag{25}$$

where

$$\hat{P}_n(\gamma) = \int_{\{\mathbf{y}:\hat{\gamma}(\mathbf{y})=\gamma\}} \hat{f}(\mathbf{y};\hat{\gamma}(\mathbf{y}))d\mathbf{y}. \tag{26}$$

This defines the *canonical* prior

$$\hat{\pi}_n(\gamma) = \frac{\hat{P}_n(\gamma)}{\sum_{\alpha\in\Omega} \hat{P}_n(\alpha)}, \tag{27}$$

and we get the factorization

$$\hat{f}(\mathbf{x};\Omega) = \frac{\hat{f}(\mathbf{x};\hat{\gamma}(\mathbf{x}))}{\hat{P}_n(\hat{\gamma}(\mathbf{x}))}\hat{\pi}_n(\hat{\gamma}(\mathbf{x})). \tag{28}$$

In analogy with $\hat{f}(\mathbf{x};\gamma)$ we call $\hat{f}(\mathbf{x};\Omega)$ the *NML* density function for the model class with the index sets in Ω. It then gives the final decomposition

$$-\ln \hat{f}(\mathbf{x};\Omega) = \frac{n-\hat{k}}{2}\ln\hat{\tau} + \frac{\hat{k}}{2}\ln\hat{R} - \ln\Gamma(\frac{n-\hat{k}}{2}) - \ln\Gamma(\frac{\hat{k}}{2}) + \ln\frac{1}{\hat{k}} + \ Const, \tag{29}$$

where we include in $Const$ all the terms that do not depend on the optimal index set $\hat{\gamma}$ of size $\hat{k}$. The terms other than the first define the length of a code from which the optimal normal model, defined by the ML parameters, in the subclass specified by the term $Const$ can be decoded, while the first term represents the code length of the part of the data that adds no further information about the optimal model. It may be viewed as noise. Hence this decomposition is similar to Kolmogorov's sufficient statistics decomposition in the algorithmic theory of information, and it is also seen to extend the ordinary sufficient statistics, as defined for certain parametric families, to something that could be called parameter free *universal* sufficient statistics.

By applying Stirling's approximation to the $\Gamma-$functions we get the *NML* criterion for $0 < k < n$

$$\min_{\gamma\in\Omega}\{(n-k)\ln\hat{\tau} + k\ln(n\hat{R}) + (n-k-1)\ln\frac{n}{n-k} - (k+1)\ln k\}, \tag{30}$$

where k denotes the number of elements in γ. This differs from the criterion in [11] and that in [6] only by an added term $2\ln(2/k)$. We emphasize that the optimal index set γ with which the shortest ideal code length for the data $-\log \hat{f}(\mathbf{x};\Omega)$ in the minmax sense results, is indeed $\hat{\gamma}(\mathbf{x})$, obtained by minimizing the criterion (21) without any additional code length for γ.

It may be of interest to calculate the probability $\hat{P}_n(\hat{\gamma}(\mathbf{x}))$ and the prior $\hat{\pi}_n(\gamma)$, which also give the normalizing constant (25). We can do it asymptotically by use of the theory of the *MDL* estimates $\hat{\gamma}(\mathbf{x})$, which are consistent provided the number of parameters to be estimated does not grow with n; for the general proof pattern we refer to [1]. This motivates the assumptions in the theorem

Theorem 1 *Let the data be generated by some model $f(\mathbf{y}^n;\gamma,\beta,\tau)$ in the class specified by γ in a finite set Ω, and let $Prob(\hat{\gamma}(y^n) \neq \gamma) \to 0$. Then for each $\gamma \in \Omega$*

$$\hat{P}_n(\gamma) \rightarrow 1 \tag{31}$$

$$\hat{\pi}_n(\gamma) \rightarrow 1/|\Omega|. \tag{32}$$

Proof: We have

$$1 = \hat{P}_n(\gamma) + \int_{\{y^n:\hat{\gamma}(y^n)\neq\gamma\}} \hat{f}(y^n;\hat{\gamma}(y^n))dy^n \tag{33}$$

where the integral converges to zero by assumption. The claim follows.

The theorem has the counterintuitive implication that if we consider the unnormalized 2-part code length $-\ln \hat{f}(\mathbf{x};\gamma) - \ln \hat{\pi}_n(\gamma)$ then all index sets γ, regardless of the number of indices in them, should be encoded with the same code length. Moreover, the need for the normalization by the probability $\hat{P}_n(\hat{\gamma})$ disappears.

3 Regression with Orthonormal Bases

Let W be an orthonormal $n \times n-$ matrix so that $W^{-1} = W'$. Its columns define an orthonormal basis, and we have the transforms

$$\begin{aligned} \mathbf{x} &= W\mathbf{c} \\ \mathbf{c} &= W'\mathbf{x}, \end{aligned} \tag{34}$$

where $\mathbf{x}$ and $\mathbf{c}$ denote the column vectors of the strings of the data $\mathbf{x}' = x_1,\ldots,x_n$ and the coefficients $\mathbf{c}' = c_1,\ldots,c_n$, respectively. Because of orthonormality Parseval's equality $\mathbf{c}'\mathbf{c} = \sum_t c_t^2 = \mathbf{x}'\mathbf{x} = \sum_t x_t^2$ holds.

Let $\gamma = \{i_1,\ldots,i_k\}$ denote a set of indices, $0 < k < n$, and let $\hat{\mathbf{c}}'$ denote the row vector of components $\{\delta_i(\gamma)c_i\}$, where $\delta_i(\gamma) = 1$ for $i \in \gamma$, and zero otherwise. In words, the non-zero

components in $\hat{\mathbf{c}}'$ with indices in γ are the corresponding components of $\mathbf{c}'$. These define the ML estimates $\hat{\beta}$ in the notations of the previous section. The ML estimate $\hat{\tau}$ is now given by

$$n\hat{\tau} = \sum_{t=1}^{n}(x_t - \hat{x}_t)^2 = \mathbf{c}'\mathbf{c} - \hat{\mathbf{c}}'\hat{\mathbf{c}}. \tag{35}$$

The criterion (30) for finding the best subset γ, including the number k of its elements, is then equivalent with

$$\min_{\gamma} C_\gamma(\mathbf{x}) = \min_{\gamma}\{(n-k)\ln\frac{\mathbf{c}'\mathbf{c} - \hat{S}_k}{n-k} + k\ln\frac{\hat{S}_k}{k} - \ln\frac{k}{n-k}\}, \tag{36}$$

where

$$\hat{S}_k = \hat{\mathbf{x}}'\hat{\mathbf{x}} = \hat{\mathbf{c}}'\hat{\mathbf{c}}. \tag{37}$$

Because of Parseval's equality the sum of the squared deviations $\hat{\tau}$ is minimized by the k largest coefficients in absolute value, which makes the thresholding schemes in [4] possible. Because $\hat{S}$ in the second term of the criterion $C_\gamma(\mathbf{x})$ is maximized by the k largest squared coefficients, it is not clear at all that any threshold exists, and that the search for the minimizing index set $\hat{\gamma}$ through all the 2^n possible index sets can be avoided. However, we have the theorem

Theorem 2 *For orthonormal regression matrices the index set $\hat{\gamma}$ that minimizes the criterion (36) is given either by the indices $\hat{\gamma} = \{(1),\ldots,(k)\}$ of the k largest coefficients in absolute value or the indices $\hat{\gamma} = \{(n-k+1),\ldots,(n)\}$ of the k smallest ones for some $k = \hat{k}$. If $C_{(k)}(\mathbf{x})$ and $\bar{C}_{(k)}(\mathbf{x})$ denote the corresponding values of the criteria, respectively, then*

$$\bar{C}_{(k)}(\mathbf{x}) = C_{(n-k)}(\mathbf{x}) + 2\ln\frac{n-k}{k}, \tag{38}$$

and $\hat{\gamma}$ can be found by no more than n evaluations of the criterion.

Proof: Let γ be an arbitrary collection of a fixed number of indices k, and let $\hat{S}_k$ be the corresponding sum of the squared coefficients. Let $u_i = c_i^2$ be a term in $\hat{S}_k$. The derivative of $C_\gamma(\mathbf{x})$ with respect to u_i is then

$$\frac{dC_\gamma(\mathbf{x})}{du_i} = \frac{k}{\hat{S}_k} - \frac{n-k}{\mathbf{c}'\mathbf{c} - \hat{S}_k}, \tag{39}$$

which is nonpositive when $\hat{S}_k/k \geq \hat{T}_k/(n-k)$, where $\hat{T}_k = n\hat{\tau}_k = \mathbf{c}'\mathbf{c} - \hat{S}_k$, and positive otherwise. The second derivative is always negative, which means that $C_\gamma(\mathbf{x})$ as a function of u_i is concave.

If for some γ, $\hat{S}_k/k > \hat{T}_k/(n-k)$, we can reduce $C_\gamma(\mathbf{x})$ by replacing, say, the smallest square c_i^2 in γ by a larger square outside of γ, and get another γ for which $\hat{S}_k$ is larger and $\hat{T}_k$ smaller. This process is possible until $\gamma = \{(1),\ldots,(k)\}$ consists of the indices of the k largest squared coefficients. Similarly, if for some γ, $\hat{S}_k/k < \hat{T}_k/(n-k)$, we can reduce $C_\gamma(\mathbf{x})$ until γ consists of

the indices of the k smallest squared coefficients. Finally, if for some γ, $\hat{S}_k/k = \hat{T}_k/(n-k)$, then all the squared coefficients must be equal, and the claim holds trivially. The proof of (38) follows by a direct verification.

It may seem weird that the threshold defined by $\hat{k}$ could require setting coefficients that exceed the threshold to zero. However, we have made no assumptions about the data to which the criterion is to be applied, and it can happen that the signal $\hat{x} = \hat{x}^n$ recovered is defined by a model which is more complex than the noise $x^n - \hat{x}^n$, relative to the two classes of distributions considered, the normal one for the noise and the uniform for the models. Hence, in such a case it may pay to reverse the roles of the information bearing signal and the noise.

It seems that in most denoising problems the data are such that the information bearing part is simpler than the noise in the sense the index set $\hat{\gamma}$ minimizing the criterion (36) has fewer than $n/2$ elements. For such data Equation (38) implies that $\bar{C}_{(k)}(\mathbf{x})$ cannot be the minimum. In fact, $C_{\hat{\gamma}}(\mathbf{x}) \leq C_{(n-i)}(\mathbf{x}) < \bar{C}_{(i)}(\mathbf{x})$ for all $i < n/2$, which in view of Theorem 2 implies $C_{\hat{\gamma}}(\mathbf{x}) = C_{(\hat{k})}(\mathbf{x})$. For denoising, then, we should simply optimize the criterion (36) over the k largest coefficients in absolute value thus

$$\min_k C_{(k)}(\mathbf{x}) = \min_k \{(n-k)\ln\frac{c'c - \hat{S}_{(k)}}{n-k} + k\ln\frac{\hat{S}_{(k)}}{k} - \ln\frac{k}{n-k}. \tag{40}$$

With $\hat{c}^n$ denoting the column vector defined by the coefficients $\hat{c}_1, \ldots, \hat{c}_n$, where $\hat{c}_i = c_i$ for $i \in \{(1), \ldots, (\hat{k})\}$ and zero, otherwise, the signal recovered is given by $\hat{x}^n = W\hat{c}^n$.

We conclude this section by studying the behavior of the threshold $\lambda = c^2_{(\hat{k})}$ for large n. For small n it depends in an intricate manner on the data sequence, but for large n we can get some information if we assume a certain type of asymptotic behavior of the data sequence. For wavelet transforms a data sequence can increase either by sampling a function, defined on an interval of the real line, with finer and finer resolutions or by keeping the resolution fixed while increasing the number of samples. In either case a reasonable assumption for many data sequences might be that for $\hat{k} = o(n)$ both $\hat{\mathbf{x}}'\hat{\mathbf{x}}/n$ and $\hat{\tau}_{\hat{k}}$ would stay within an interval $[a_1, a_2]$ for $0 < a_1$.

We have for $1 < k < n$

$$C_{(k)}(\mathbf{x}) - C_{(k-1)}(\mathbf{x}) = -(n-k+1)\ln(1+\frac{\lambda}{n\hat{\tau}(k)}) + k\ln(1+\frac{\lambda}{\hat{S}_{(k-1)}}) + \ln\frac{\hat{S}_{(k-1)}}{n\hat{\tau}(k)} + \ln n + F(k,n), \tag{41}$$

where

$$F(k,n) = \ln(1-k/n) - k\ln(1+\frac{1}{k-1}) - \ln k. \tag{42}$$

If we use the approximation $\ln(1+O(1/n)) = O(1/n)$ we get

$$\lambda = \hat{\tau}(\hat{k})\ln n + o(\ln n). \tag{43}$$

This is one half of the threshold in [4] as well as that in [7], provided that $\hat{\tau}(k)$ as a function of the increasing data sequence converges to some τ. That this threshold does not achieve the minmax bound $2\tau \ln n$ for the risk in [4] is of little consequence since for all but pathological signals the risk will be much smaller. In fact, we can see at once that if the mean signal $\bar{\mathbf{x}}$ is such that $\bar{\mathbf{x}}_i^2 < M$ for some constant M, which certainly is satisfied by many signals of interest, the risk will be uniformly bounded. Indeed, $(1/n)E(\bar{\mathbf{x}} - \hat{\mathbf{x}})'(\bar{\mathbf{x}} - \hat{\mathbf{x}}) = (1/n)E[(\mathbf{x} - \bar{\mathbf{x}}) + (\hat{\mathbf{x}} - \mathbf{x})]'[(\mathbf{x} - \bar{\mathbf{x}}) + (\hat{\mathbf{x}} - \mathbf{x})] < 2(\tau + M)$ no matter what the threshold is.

4 Examples

We calculate two examples using wavelets defined by Daubechies' N=6 scaling function. The first example, taken from [8], is a case where the *MDL* threshold is close to the ones in [4] as well as to the threshold in [8], obtained with a rather complicated cross-validation technique. It is clear from the criterion (40) that the *MDL* threshold is a function of the data. In [4] it becomes a function of data by replacing the assumed 'true' variance by its estimate, and that is the case in [8], too, by the nature of the cross-validation procedure. The second example is about real speech data, where the information bearing signal, recovered by the *NML* criterion (40), differs a lot from the result obtained with the universal threshold in [4].

In the first example the mean signal $\bar{\mathbf{x}}_i$ consists of 512 equally spaced samples of the following function defined by three piecewise polynomials

$$x(t) = \begin{cases} 4t^2(3-4t) & \text{for } t \in [0, .5] \\ \frac{4}{3}t(4t^2 - 10t + 7) - \frac{3}{2} & \text{for } t \in [.5, .75] \\ \frac{16}{3}t(t-1)^2 & \text{for } t \in [.75, 1] \end{cases}$$

We added pseudorandom normal 0-mean noise with standard deviation of 0.1 to the data points $\bar{x}_i$ are , which defined the data sequence x_i.

The threshold obtained with the *NML* criterion is $\lambda = 0.246$. This is between the two thresholds called VisuShrink $\lambda = 0.35$ and GlobalSure $\lambda = 0.14$, both of the type in [4], as reported in [8]. It is also close to the threshold $\lambda = 0.20$, obtained with the much more complex cross-validation procedure in [8]. In all cases the recovered signals looked alike.

In the second example the data sequence consists of 128 samples from a voiced portion of speech. The *NML* criterion retains 42 coefficients exceeding the threshold $\lambda = 7.3$ in absolute value. It gives the value $\hat{\tau} = 5.74$ for the noise variance. The estimation procedure in [4] gives $\hat{\tau} = 10.89$ and the threshold $\lambda = \sqrt{2\hat{\tau} \ln 128} = 10.3$. Figure 1 shows the original signal together with the information bearing signal extracted by the *NML* criterion and the much smoother signal, marked 'DJ signal', obtained with the threshold in [4]. We see that the latter is far too smooth and fails to capture the important large pulses in the original signal. We also see that the *NML* criterion has not removed the sharp peaks in the large pulses despite the fact that they have locally high frequency content. They

simply can be compressed by use of the retained coefficients, and by the general principle behind the criterion they are not regarded as noise.

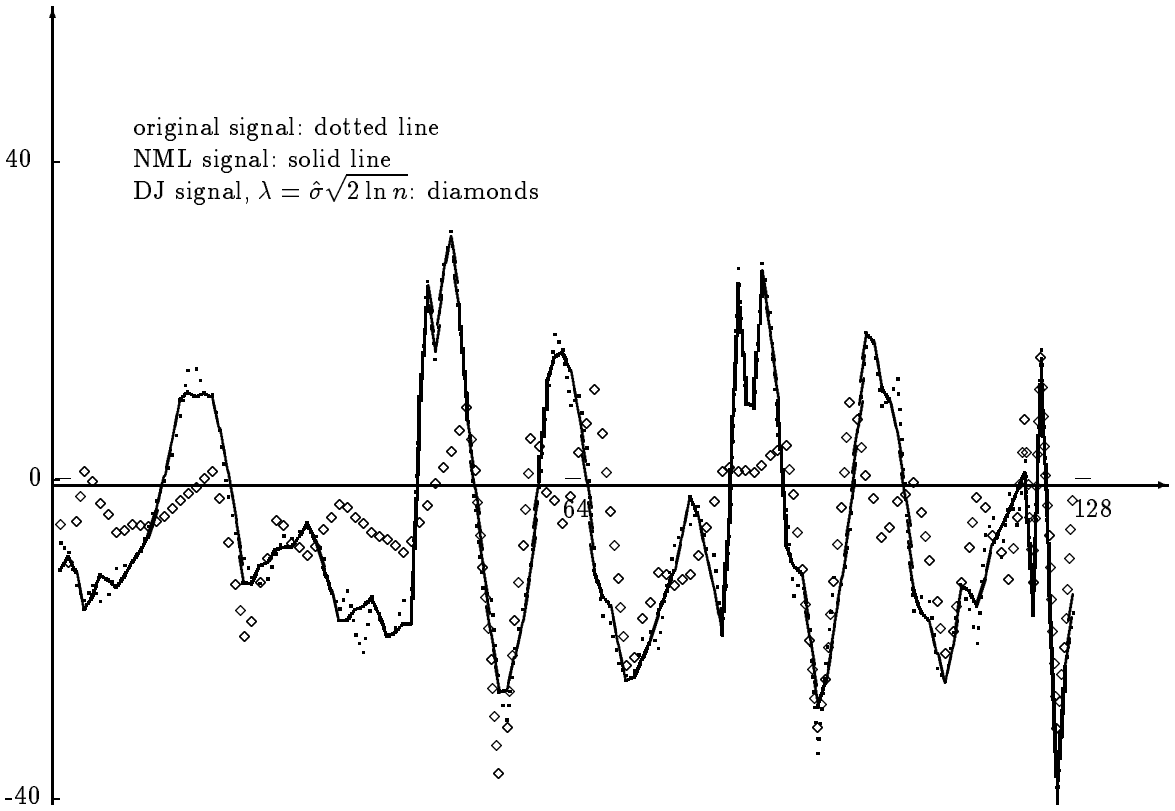

Figure 1. Speech signal smoothed with Daubechies' N=6 wavelet

References

[1] Barron, A.R., Rissanen, J., and Yu, B. (1998) 'The MDL Principle in Modeling and Coding', special issue of *IEEE Trans. Information Theory* to commemorate 50 years of information theory, Vol. **IT-44**, No. 6, October 1998, pp 2743-2760

[2] Cover, T. M. and Thomas, J. A. (1991), *Elements of Information Theory*, Wiley.

[3] Dom, B. (1996), 'MDL Estimation for Small Sample Sizes and Its Application to Linear Regression', *IBM Research Report* **RJ 10030**, June 13, 1996.

[4] Donoho, D.L. and Johnstone, I.M. (1994), 'Ideal Spatial Adaptation by Wavelet Shrinkage', *Biometrika*, **81**, 425-455.

[5] Foster, D.P. and George, E.I. (1994), 'The Risk Inflation Criterion for Multiple Regression', *Annals of Statistics*, Vol. **22**, No. 4, pp 1947-75

[6] Hansen, M.H. and Yu, B. (1998), 'Model Selection and the Principle of Minimum Description Length', (to appear in *JASA*), can be obtained from the web at http://cm.bell-labs.com/stat/doc/mdl.ps.Z

[7] Krim, H. and Schick, I.C. (1999), 'Minimax Description Length for Signal Denoising and Optimized Representation', *IEEE Trans. Information Theory*, Vol. **IT-45**, No. 3, pp 898-908

[8] Nason, G.P. (1996), 'Wavelet Shrinkage using Cross-validation', *J. R. Statist. Soc. B* 58, No. 2, pp 463-479

[9] Rissanen, J. (1996), 'Fisher Information and Stochastic Complexity', *IEEE Trans. Information Theory*, Vol. **IT-42**, No. 1, pp 40-47

[10] Rissanen, J. (1989), Stochastic Complexity in Statistical Inquiry, World Scientific, New Jersey, second edition 1998, 177 pages

[11] Rissanen, J. (1998), 'Hypothesis Selection and Testing by the MDL Principle', invited paper to *The Computer Journal*, Vol. **42**, Nr 4, pp 260-269

[12] Rissanen, J. (2000), 'A Generalized MinMax Bound for Universal Coding', *Proceedings of IEEE International Symposium on Information Theory*, ISIT2000, Sorrento, Italy, June 25-30, 2000.

[13] Shtarkov, Yu. M. "Universal Sequential Coding of Single Messages," *Problems of Information Transmission*, Vol. 23, No. 3, 3-17, July-September 1987.