MDL Regression and Denoising

Teemu Roos

November 30, 2004

Abstract

Rissanen's proofs for linear regression and wavelet denoising reproduced. Generalized NML density introduced.

1 Linear regression

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 τ. The linear regression model is equivalent to

$$\begin{aligned} f(y^n\,;\beta,\tau) &= (2\pi\tau)^{-n/2} \exp\left(-\frac{1}{2\tau}\|y^n - X\beta\|^2\right) \\ &= (2\pi\tau)^{-n/2} \exp\left(-\frac{1}{2\tau}\sum_t (y_t - \vec{x}_t\beta)^2\right), \qquad (1) \end{aligned}$$

where $\|\cdot\|^2$ denotes the square norm, and $\vec{x}_t$ denotes the tth row of X. We define the matrices $Z = X'X$ and $\Sigma = n^{-1}Z$ which are assumed to be positive definite[1]. The maximum likelihood estimators of β and τ are independent and given by

$$\hat{\beta}(y^n) = Z^{-1}X'y^n, \qquad (2)$$

$$\hat{\tau}(y^n) = \frac{1}{n}\|y^n - X\hat{\beta}'(y^n)\|^2 = \frac{1}{n}\sum_t \left(y_t - \hat{\beta}'(y^n)\vec{x}_t\right)^2. \qquad (3)$$

[1] Positive definiteness guarantees that there are unique maximum likelihood parameters.

2 Normalized maximum likelihood

The normalized maximum likelihood (NML) density for a model class parameterized by parameter vector θ is defined by

$$\bar{f}(y^n) = \frac{f(y^n\,; \hat{\theta}(y^n))}{C}, \tag{4}$$

where

$$C = \int_Y f(y^n\,; \hat{\theta}(y^n))\; dy^n \tag{5}$$

is a normalizing constant, and Y is the range of integration 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 infinite.

Assume that the maximum likelihood estimator $\hat{\theta}(y^n)$ is sufficient for θ. Then the conditional density of y^n given $\hat{\theta}(y^n)$ is independent of the generating density so that we can write

$$f(y^n \mid \hat{\theta}(y^n)\,; \hat{\theta}(y^n)) = f(y^n \mid \hat{\theta}(y^n)).$$

Since the maximum likelihood estimator $\hat{\theta}(y^n)$ is a function of the data y^n, we have the factorization

$$\bar{f}(y^n) = f(y^n \mid \hat{\theta}(y^n))\; g(\hat{\theta}(y^n)\,; \hat{\theta}(y^n))\; C^{-1}, \tag{6}$$

where $g(\hat{\theta}(y^n)\,; \theta)$ is the density of the maximum likelihood estimator. When evaluated at the maximum likelihood point $\theta = \hat{\theta}(y^n)$ as above, the density gives the so called *canonical prior.*

The difference between the 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:

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

The NML density is the unique solution to Shtarkov's minmax problem:

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

and the following more general problem:

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

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 $\ln C$. Thus, NML can be said to give the shortest description of the data achievable with a given model class.

Codes based on the NML density are sometimes characterized as a computable probabilistic analogues of the Kolmogorov sufficient statistic decomposition in algorithmic information theory. However, this interpretation is not very straightforward. The main difficulty is that NML does not define a two-part code where information and noise would be easily separated. In particular, while the definition of NML (4) as the ratio of the maximum likelihood density and the normalizing constant might suggest a code where first the logarithm of the normalizing constant gives the code-length for the maximum likelihood parameters containing all the interesting information, and secondly, noise is encoded using the density indexed by the parameter values, the interpretation fails because such a code is redundant: code-words are reserved for encoding data with parameters other than the maximum likelihood ones. Being a density, NML corresponds to a complete code, i.e., the Kraft inequality is satistied as an equality and such redundancy is not possible.

The factorization (6) in fact allows a two-part eccoding scheme where the maximum likelihood parameters are encoded with optimal finite precision which doesn't affect the resulting density and is not explicit in the formula. However, given the (truncated) maximum likelihood parameters, the noise is encoded not using one of the densities in the model class but benefiting in addition from the restriction that the maximum likelihood parameters for the data agree with the truncated versions in the first part of the code. This contradicts the idea of the sufficient statistic decomposition where data is always encoded using one of the models in the model class. Indeed, it seems that the sufficient statistic decomposition is by its nature not optimal for encoding data since it implies redundancy of the kind described above and therefore, the two objectives of coding and modeling may at the end of the day lead to somewhat different solutions.

3 NML for linear regression

Consider the NML density in the case of linear regression:

$$\bar{f}(y^n\,; \tau_0, R) = \frac{f(y^n\,; \hat{\beta}(y^n), \hat{\tau}(y^n))}{C_{\tau_0,R}}, \tag{7}$$

with the numerator given by Eq. (1) and the normalizing constant given by

$$C_{\tau_0,R} = \int_{Y(\tau_0,R)} f(y^n\,; \hat{\beta}(y^n), \hat{\tau}(y^n))\; dy^n. \tag{8}$$

where $Y(\tau_0, R)$ is the range of integration. The density is defined only for data sequences y^n within the range. A range is necessary since over the n-dimensional real space $\mathbb{R}^n$ the integral would be infinite. Given hyperparameters τ_0 and R we define the range $Y(\tau_0, R)$ as

$$Y(\tau_0, R) = \{y^n : \hat{\tau}(y^n) \geq \tau_0, \hat{\beta}'(y^n)\Sigma\hat{\beta}(y^n) \leq R\}.$$

By plugging the maximum likelihood estimators (2) and (3) in the density (1) we obtain the numerator of the NML density:

$$f(y^n\,; \hat{\beta}(y^n), \hat{\tau}(y^n)) = (2\pi e \hat{\tau}(y^n))^{-n/2}. \tag{9}$$

In order to do the integral in the denominator, we need to work the same quantity in a different form. Since $\hat{\beta}(y^n)$ and $\hat{\tau}(y^n)$ are sufficient statistics and independent of each other, we have the factorization

$$f(y^n\,; \beta, \tau) = f(y^n \mid \hat{\beta}(y^n), \hat{\tau}(y^n)) g_1(\hat{\beta}(y^n)\,; \beta, \tau) g_2(\hat{\tau}(y^n)\,; \beta, \tau), \tag{10}$$

The density of the maximum likelihood estimator $\hat{\beta}(y^n)$ is Gaussian with mean β and covariance $\tau/n\Sigma^{-1}$:

$$\begin{aligned} g_1(\hat{\beta}(y^n)\,; \beta, \tau) &= \mathcal{N}(\hat{\beta}(y^n)\,; \beta, \tau) \\ &= \frac{1}{(2\pi)^{k/2}|\tau/n\Sigma^{-1}|^{1/2}} \exp\left(\frac{1}{2}(\hat{\beta}(y^n) - \beta)'(\tau/n\Sigma^{-1})^{-1}(\hat{\beta}(y^n) - \beta)\right) \\ &= \frac{n^{k/2}|\Sigma|^{1/2}}{(2\pi\tau)^{k/2}} \exp\left(\frac{n}{2\tau}(\hat{\beta}(y^n) - \beta)'\Sigma(\hat{\beta}(y^n) - \beta)\right), \end{aligned}$$

where we used the equations $|a\Sigma| = a^k|\Sigma|$, $|\Sigma^{-1}| = |\Sigma|^{-1}$, and $(a\Sigma)^{-1} = 1/a\Sigma^{-1}$, for all $a \neq 0$ and Σ invertible and of size $k \times k$. When evaluated at $\beta = \hat{\beta}(y^n)$ the exponential term disappears and we get

$$g_1(\hat{\beta}(y^n)\,; \hat{\beta}(y^n), \hat{\tau}(y^n)) = \frac{n^{k/2}|\Sigma|^{1/2}}{(2\pi\hat{\tau}(y^n))^{k/2}}. \tag{11}$$

Consider next the the density of $n\hat{\tau}(y^n)/\tau$ which is χ^2 (Chi-squared) with $n-k$ degrees of freedom. Recall that the density of a χ^2 distributed random variable with $n-k$ degrees of freedom is given by

$$\chi^2_{n-k}(\xi) = \frac{\xi^{(n-k)/2-1}\exp\left(-\frac{\xi}{2}\right)}{\Gamma\left(\frac{n-k}{2}\right)2^{(n-k)/2}}.$$

By change of variables the density of $\hat{\tau}(y^n)$ is obtained by plugging $n\hat{\tau}(y^n)/\tau$ in the above density and multiplying by n/τ. Thus, the density of $\hat{\tau}(y^n)$ is in fact independent of β and we can write

$$g_2(\hat{\tau}(y^n)\,; \tau) = \frac{(n\hat{\tau}(y^n)/\tau)^{(n-k)/2-1}\exp\left(-\frac{n\hat{\tau}(y^n)}{2\tau}\right)}{\Gamma\left(\frac{n-k}{2}\right)2^{(n-k)/2}}\frac{n}{\tau}.$$

Evaluating the above at the maximum likelihood parameter $\tau = \hat{\tau}(y^n)$ yields

$$g_2(\hat{\tau}(y^n)\,;\hat{\tau}(y^n)) = \frac{n^{(n-k)/2-1}\exp\left(-\frac{n}{2}\right)}{\Gamma\left(\frac{n-k}{2}\right)2^{(n-k)/2}}\frac{n}{\hat{\tau}(y^n)} = \frac{n^{(n-k)/2}\exp\left(-\frac{n}{2}\right)}{\Gamma\left(\frac{n-k}{2}\right)2^{(n-k)/2}\hat{\tau}(y^n)}. \tag{12}$$

By independence, the joint density of $(\hat{\beta}(y^n), \hat{\tau}(y^n))$ is obtained by combining (11) and (12) which yields

$$\begin{aligned} g(\hat{\beta}(y^n), \hat{\tau}(y^n)\,;\hat{\beta}(y^n), \hat{\tau}(y^n)) &= g_1(\hat{\beta}(y^n)\,;\hat{\beta}(y^n), \hat{\tau}(y^n))\; g_2(\hat{\tau}(y^n)\,;\hat{\tau}(y^n)) \\ &= \hat{\tau}(y^n)^{-k/2-1}\left(\frac{n}{2e}\right)^{n/2}\frac{|\Sigma|^{1/2}}{\pi^{k/2}\Gamma\left(\frac{n-k}{2}\right)}. \end{aligned} \tag{13}$$

By the factorization (10), the integral (8) can be performed by subsequent integration over the maximum likelihood parameters and the corresponding data vectors:

$$\begin{aligned} C_{\tau_0,R} &= \int_{Y(\tau_0,R)} f(y^n\,;\hat{\beta}(y^n), \hat{\tau}(y^n))\; dy^n \\ &= \int_{\tau_0}^{\infty}\int_{B_R} g(\hat{\beta}, \hat{\tau}\,;\hat{\beta}, \hat{\tau})\int_{\{y\,:\,\hat{\beta}(y^n)=\hat{\beta}, \hat{\tau}(y^n)=\hat{\tau}\}} f(y^n \mid \hat{\beta}(y^n), \hat{\tau}(y^n))\; dy^n d\hat{\beta} d\hat{\tau}, \end{aligned}$$

where B_R is the ellipsoid

$$B_R = \{\beta \in \mathbb{R}^k : \beta'\Sigma\beta \leq R\}, \tag{14}$$

and the inner integral is over data vectors y^n such that the maximum likelihood parameters are $\hat{\beta}$ and $\hat{\tau}$. By definition, the value of the inner integral is unity, and by Eq. (13) we get

$$C_{\tau_0,R} = \left(\frac{n}{2e}\right)^{n/2}\frac{|\Sigma|^{1/2}}{\pi^{k/2}\Gamma\left(\frac{n-k}{2}\right)}\int_{B_R} d\hat{\beta}\int_{\tau_0}^{\infty}\hat{\tau}^{-k/2-1}d\hat{\tau}. \tag{15}$$

Next we evaluate the volume of the ellipsoid B_R. Consider first the k-dimensional sphere of radius $R^{1/2}$:

$$S_R = \{b \in \mathbb{R}^k : b'b \leq R\}. \tag{16}$$

Its volume is given by

$$Vol(S_R) = \int_{S_R} db = \frac{2\pi^{k/2}R^{k/2}}{k\Gamma(k/2)}.$$

Since by assumption, Σ is positive definite, it has a positive definite square root $\Sigma^{1/2}$ which further has the inverse $\Sigma^{-1/2}$. By the symmetry of Σ, both $\Sigma^{1/2}$ and $\Sigma^{-1/2}$ are symmetric. Matrix $\Sigma^{-1/2}$ defines a linear map $\mathbb{R}^k \to \mathbb{R}^k : \beta = \Sigma^{-1/2}b$

which is one-to-one, and we have $b = \Sigma^{1/2}\Sigma^{-1/2}b$. The image of sphere S_R under map $\Sigma^{-1/2}$ is given by

$$\begin{aligned}\Sigma^{-1/2}(S_R) &= \{\beta : b'b \le R\} \\ &= \{\beta : (\Sigma^{1/2}\Sigma^{-1/2}b)'(\Sigma^{1/2}\Sigma^{-1/2}b) \le R\} \\ &= \{\beta : (\Sigma^{1/2}\beta)'(\Sigma^{1/2}\beta) \le R\} \\ &= \{\beta : \beta'\Sigma\beta \le R\} = B_R.\end{aligned}$$

Thus, the image of the sphere S_R under map $\Sigma^{1/2}$ is the ellipsoid B_R whose volume is thereby

$$Vol(B_R) = \int_{B_R} d\hat{\beta} = |\Sigma^{-1/2}|\, Vol(S_R) = |\Sigma|^{-1/2} \frac{2\pi^{k/2}R^{k/2}}{k\Gamma(k/2)}. \tag{17}$$

The value of the latter integral in Eq. (15) is given by

$$\int_{\tau_0}^{\infty} \hat{\tau}^{-k/2-1} d\hat{\tau} = \frac{2}{k\tau_0^{k/2}}. \tag{18}$$

Combining Eqs. (15), (17) and (18) gives

$$C_{\tau_0,R} = \frac{4n^{n/2}R^{k/2}}{(2e)^{n/2}k^2\Gamma\left(\frac{n-k}{2}\right)\Gamma\left(\frac{k}{2}\right)\tau_0^{k/2}}. \tag{19}$$

The NML density (7) is then given by

$$\bar{f}(y^n\,; \tau_0, R) = \frac{k^2\Gamma\left(\frac{n-k}{2}\right)\Gamma\left(\frac{k}{2}\right)\tau_0^{k/2}}{(n\pi\hat{\tau}(y^n))^{n/2}4R^{k/2}}, \tag{20}$$

and the negative logarithm of this is

$$\begin{aligned}&-\ln \bar{f}(y^n\,; \tau_0, R) \\ &\quad = \frac{n}{2}\ln\hat{\tau}(y^n) - \ln\Gamma\left(\frac{n-k}{2}\right) - \ln\Gamma\left(\frac{k}{2}\right) + \ln\frac{4}{k^2} + \frac{k}{2}\ln\frac{R}{\tau_0} + \frac{n}{2}\ln(n\pi).\end{aligned}$$

4 Second level NML

In order to get rid of the last term that depends on k and the choice of the hyperparameters R and τ_0, and thus, affects the criterion, we do a second normalization. Let $\hat{R}(y^n)$ and $\hat{\tau}_0(y^n)$ denote the maximum likelihood values of R and τ_0. Their values are given by

$$\hat{R}(y^n) = \hat{\beta}'(y^n)\Sigma\hat{\beta}(y^n), \tag{21}$$

$$\hat{\tau}_0(y^n) = \hat{\tau}(y^n). \tag{22}$$

The second level NML density is given by

$$\bar{f}(y^n) = \frac{\bar{f}(y^n\,;\hat{\tau}_0(y^n), \hat{R}(y^n))}{C_{\tau_1,\tau_2,R_1,R_2}}, \tag{23}$$

where

$$C_{\tau_1,\tau_2,R_1,R_2} = \int_{Y(\tau_1,\tau_2,R_1,R_2)} \bar{f}(y^n\,;\hat{\tau}_0(y^n), \hat{R}(y^n))\; dy^n, \tag{24}$$

and the range of integration is defined by

$$Y_{\tau_1,\tau_2,R_1,R_2} = \{y^n : \tau_1 \leq \hat{\tau}_0(y^n) \leq \tau_2, R_1 \leq \hat{R}(y^n) \leq R_2\}.$$

By Eqs. (7), (10) and (13) the integrand allows the factorization

$$\begin{aligned}&\bar{f}(y^n\,;\hat{\tau}_0(y^n), \hat{R}(y^n))\\ &\quad = f(y^n \mid \hat{\beta}(y^n), \hat{\tau}(y^n)) g(\hat{\beta}(y^n), \hat{\tau}(y^n)\,;\hat{\beta}(y^n), \hat{\tau}(y^n))/C_{\hat{\tau}_0(y^n),\hat{R}(y^n)}.\end{aligned}$$

Again, the integral (24) can be performed by subsequent integration over the maximum likelihood parameters and the corresponding data vectors:

$$\begin{aligned}C_{\tau_1,\tau_2,R_1,R_2} &= \int_{Y(\tau_1,\tau_2,R_1,R_2)} \bar{f}(y^n\,;\hat{\tau}_0(y^n), \hat{R}(y^n))\; dy^n\\ &= \int_{B_{R_1,R_2}} \int_{\tau_1}^{\tau_2} g(\hat{\beta}, \hat{\tau}\,;\hat{\beta}, \hat{\tau})/C_{\hat{\tau},\hat{R}}\\ &\qquad \int_{\{y\,:\,\hat{\beta}(y^n)=\hat{\beta}, \hat{\tau}(y^n)=\hat{\tau}\}} f(y^n \mid \hat{\beta}(y^n), \hat{\tau}(y^n))\; dy^n d\hat{\tau} d\hat{\beta},\end{aligned}$$

where B_{R_1,R_2} contains parameters $\hat{\beta}$ inside ellipsoid B_{R_2} but outside ellipsoid B_{R_1} as defined in Eq. (14). Again, the inner integral equals unity. By Eqs. (13) and (19) the fraction equals

$$g(\hat{\beta}, \hat{\tau}\,;\hat{\beta}, \hat{\tau})/C_{\hat{\tau},\hat{R}} = \frac{k^2|\Sigma|^{1/2}\Gamma\,(k/2)}{4\pi^{k/2}} \hat{\tau}^{-1} \hat{R}^{-k/2}.$$

Thus the integral becomes

$$\begin{aligned}C_{\tau_1,\tau_2,R_1,R_2} &= \frac{k^2|\Sigma|^{1/2}\Gamma\,(k/2)}{4\pi^{k/2}} \int_{\tau_1}^{\tau_2} \hat{\tau}^{-1} d\hat{\tau} \int_{B_{R_1,R_2}} \hat{R}^{-k/2} d\hat{\beta}\\ &= \frac{k^2|\Sigma|^{1/2}\Gamma\,(k/2)}{4\pi^{k/2}} \ln \frac{\tau_2}{\tau_1} \int_{B_{R_1,R_2}} \hat{R}^{-k/2} d\hat{\beta}.\end{aligned} \tag{25}$$

The set B_{R_1,R_2} is the union of surface areas of ellipsoids $B_{\hat{R}}$ with $\hat{R}$ in $[R_1, R_2]$. The surface area of ellipsoid $B_{\hat{R}}$ is given by

$$Area(B_{\hat{R}}) = \frac{\partial\, Vol(B_{\hat{R}})}{\partial \hat{R}} = |\Sigma|^{-1/2} \frac{2\pi^{k/2}}{k\Gamma(k/2)} \frac{\partial \hat{R}^{k/2}}{\partial \hat{R}} = |\Sigma|^{-1/2} \frac{\pi^{k/2} \hat{R}^{k/2-1}}{\Gamma(k/2)}.$$

On the surface of $B_{\hat{R}}$ the integrand in Eq. (25) is constant and integration can be done in terms of the surface areas by varying $\hat{R}$ instead of $\hat{\beta}$:

$$\int_{B_{R_1,R_2}} \hat{R}^{-k/2} d\hat{\beta} = \int_{R_1}^{R_2} \hat{R}^{-k/2} |\Sigma|^{-1/2} \frac{\pi^{k/2} \hat{R}^{k/2-1}}{\Gamma(k/2)} d\hat{R} = |\Sigma|^{-1/2} \frac{\pi^{k/2}}{\Gamma(k/2)} \ln \frac{R_2}{R_1}. \tag{26}$$

Plugging this into Eq. (25) and cancelling like terms gives

$$C_{\tau_1,\tau_2,R_1,R_2} = \frac{k^2}{4} \ln \frac{\tau_2}{\tau_1} \ln \frac{R_2}{R_1}. \tag{27}$$

Putting together Eqs. (23), (7), (9), (19) and (27) yields the following formula for the second level NML density:

$$\begin{aligned} \bar{f}(y^n) &= \frac{\bar{f}(y^n\,; \hat{\tau}_0(y^n), \hat{R}(y^n))}{C_{\tau_1,\tau_2,R_1,R_2}} = \frac{f(y^n\,; \hat{\beta}(y^n), \hat{\tau}(y^n))}{C_{\hat{\tau}_0(y^n),\hat{R}(y^n)} C_{\tau_1,\tau_2,R_1,R_2}} \\ &= \frac{k^2 \Gamma\left(\frac{n-k}{2}\right) \Gamma\left(\frac{k}{2}\right) \hat{\tau}_0(y^n)^{k/2} 4}{(n\pi\hat{\tau}(y^n))^{n/2} 4 \hat{R}(y^n)^{k/2} \ln \frac{\tau_2}{\tau_1} k^2 \ln \frac{R_2}{R_1}} \\ &= \frac{\Gamma\left(\frac{n-k}{2}\right) \Gamma\left(\frac{k}{2}\right)}{\hat{\tau}(y^n)^{(n-k)/2} (\pi n)^{n/2} \hat{R}(y^n)^{k/2} \ln \frac{\tau_2}{\tau_1} \ln \frac{R_2}{R_1}}, \end{aligned} \tag{28}$$

where we noted that by Eq. (22) the maximum likelihood estimates $\hat{\tau}$ and $\hat{\tau}_0$ are equal. The negative logarithm of this is

$$\begin{aligned} -\ln \bar{f}(y^n) = \frac{n-k}{2} \ln \hat{\tau}(y^n) + \frac{k}{2} \ln \hat{R}(y^n) - \ln \Gamma\left(\frac{n-k}{2}\right) - \ln \Gamma\left(\frac{k}{2}\right) \\ + \frac{n}{2} \ln(\pi n) + \ln\left(\ln \frac{\tau_2}{\tau_1} \ln \frac{R_2}{R_1}\right). \end{aligned} \tag{29}$$

Unlike in the case of first level NML, the hyperparameters now affect only terms that are constant for all models.

5 Stirling approximation

The Gamma functions can be approximated by applying the Stirling approximation

$$\begin{aligned} \ln \Gamma(n+1) = \ln n! \approx \left(n + \frac{1}{2}\right) \ln n - n + \frac{1}{2} \ln(2\pi) \\ \ln \Gamma(n) = \ln \Gamma(n+1) - \ln n \approx \left(n - \frac{1}{2}\right) \ln n - n + \frac{1}{2} \ln(2\pi). \end{aligned}$$

Applying the approximation to Eq. (29) yields

$$\begin{aligned} -\ln \bar{f}(y^n) \approx \frac{n-k}{2} \ln \hat{\tau}(y^n) + \frac{k}{2} \ln \hat{R}(y^n) - \left(\frac{n-k-1}{2}\right) \ln \left(\frac{n-k}{2}\right) + \frac{n-k}{2} \\ - \frac{1}{2} \ln(2\pi) - \left(\frac{k-1}{2}\right) \ln \frac{k}{2} + \frac{k}{2} - \frac{1}{2} \ln(2\pi) + \frac{n}{2} \ln(\pi n) + \ln\left(\ln \frac{\tau_2}{\tau_1} \ln \frac{R_2}{R_1}\right). \end{aligned}$$

When interested only in minimizing the code-length formula with respect to k we can simplify the criterion further by multiplying it by two and dropping terms that do not have k in them[2]:

$$\begin{aligned}
&-2\ln \bar{f}(y^n) \\
&\qquad \overset{\pm}{=} (n-k)\ln \hat{\tau}(y^n) + k \ln \hat{R}(y^n) - (n-k-1)\ln\left(\frac{n-k}{2}\right) - (k-1)\ln\frac{k}{2} \\
&\qquad = (n-k)\ln \hat{\tau}(y^n) + k \ln \hat{R}(y^n) - (n-k-1)\ln(n-k) - (k-1)\ln k \\
&\qquad = (n-k)\ln \frac{\hat{\tau}(y^n)}{n-k} + k \ln \frac{\hat{R}(y^n)}{k} + \ln(k(n-k)), \qquad (30)
\end{aligned}$$

The approximation to the Gamma functions is accurate except for very small k or $n-k$, and the error can be bounded.

6 Wavelet denoising

Assume the vector y^n can be considered a series, i.e., the data points are ordered in a meaningful way. We can obtain the 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 of defining the functions f_j is to use a *wavelet basis.* By letting the regressor matrix be square and taking as the n basis functions $f_j(i)$ a wavelet basis, we get an *orthogonal* regressor matrix X, i.e., X has as its inverse the transpose X':

$$Z = X'X = X^{-1}X = I,$$

where I is the identity matrix. Further, the maximum likelihood parameters are given by

$$\hat{\beta}(y^n) = Z^{-1}X'y^n = X'y^n \qquad (31)$$

$$\hat{\tau}(y^n) = \frac{1}{n}\|y^n - X\hat{\beta}'(y^n)\|^2 = \frac{1}{n}\|y^n - y^n\|^2 = 0, \qquad (32)$$

i.e., the reconstructed version $\hat{y}^n = X\hat{\beta}(y^n) = y^n$ is identical to the original signal and nothing remains to be modeled as noise, thus ending up with $\hat{\tau}(y^n) = 0$.

Instead of using all the basis vectors, it may be useful to choose a subset γ of them. This gives the reconstructed version

$$\hat{y}^n_\gamma = X\hat{\beta}_\gamma(y^n)$$

[2] This can be seen to be the same criterion as the version Rissanen gives in Lecture Notes: $(n-k)\ln\hat{\tau} + k\ln(n\hat{R}) + (n-k-1)\ln\frac{n}{n-k} - (k-1)\ln k$.

not equal the original signal but an approximated version with the difference to the original signal modeled as noise. With a wavelet basis, the noise variance is easily obtained. First, 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 difference between the reconstructed version and the original signal is then

$$\|y^n - \hat{y}^n_\gamma\|^2 = \|X\hat{\beta}(y^n) - X\hat{\beta}_\gamma(y^n)\|^2 = \|X\left(\hat{\beta}(y^n) - \hat{\beta}_\gamma(y^n)\right)\|^2.$$

Since X is an orthogonal matrix it preserves the norm, i.e.,

$$\|X\left(\hat{\beta}(y^n) - \hat{\beta}_\gamma(y^n)\right)\|^2 = \|\hat{\beta}(y^n) - \hat{\beta}_\gamma(y^n)\|^2.$$

Thus, the noise variance is simply the mean of the squared coefficients that are set to zero:

$$\hat{\tau}_\gamma(y^n) = \frac{1}{n}\|\hat{\beta}(y^n) - \hat{\beta}_\gamma(y^n)\|^2 = \frac{1}{n}\sum_{i=1}^{n}(1-\delta_i(\gamma))(\hat{\beta}_i(y^n))^2.$$

Thus, the sum of the retained coefficients and the sum of squared errors between the original and reconstructed signals is always equal to the sum of squares of the original signal:

$$\begin{aligned}\|y^n\|^2 = \|X'y^n\|^2 = \|\hat{\beta}(y^n)\|^2 &= \|\hat{\beta}(y^n) - \hat{\beta}_\gamma(y^n)\|^2 + \|\hat{\beta}_\gamma(y^n)\|^2 \\ &= \|y^n - \hat{y}^n_\gamma\|^2 + \|\hat{\beta}_\gamma(y^n)\|^2.\end{aligned}$$

Define

$$S(y^n) = \|y^n\|^2, \quad \text{and} \quad S_\gamma(y^n) = \|\hat{\beta}_\gamma(y^n)\|^2.$$

We have then

$$\hat{\tau}_\gamma(y^n) = \frac{S(y^n) - S_\gamma(y^n)}{n}. \tag{33}$$

The maximum likelihood value for the hyperparameter R is by Eq. (21) simply

$$\hat{R}_\gamma(y^n) = \hat{\beta}'_\gamma(y^n)\Sigma\hat{\beta}_\gamma(y^n) = \frac{1}{n}\hat{\beta}'_\gamma(y^n)X'X\hat{\beta}_\gamma(y^n) = \frac{1}{n}\|\hat{\beta}_\gamma\|^2 = \frac{S_\gamma(y^n)}{n}. \tag{34}$$

Thus, the criterion (30) becomes

$$\begin{aligned}(n-k)\ln\frac{\hat{\tau}_\gamma(y^n)}{n-k} &+ k\ln\frac{\hat{R}_\gamma(y^n)}{k} + \ln(k(n-k)) \\ &= (n-k)\ln\frac{S(y^n) - S_\gamma(y^n)}{n(n-k)} + k\ln\frac{S_\gamma(y^n)}{nk} + \ln(k(n-k)) \\ &= (n-k)\ln\frac{S(y^n) - S_\gamma(y^n)}{n-k} + k\ln\frac{S_\gamma(y^n)}{k} + \ln(k(n-k)),\end{aligned}$$

where k is the number of retained coefficients determined by γ. It is remarkable that the criterion is symmetric in the two sets of coefficients; the ones that are set to zero and the retained ones. 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.

7 Generalized NML

In some cases, such as the linear regression case, unless the range of possible values of the data sequence y^n is restricted, the integral in the denominator of the NML density is not be bounded and thus, NML is not defined. In the above, the problem was solved by bounding the range of integration in a set defined by some hyperparameters (τ_0 and R), fixing the hyperparameters to their maximum likelihood values, and doing a second normalization. Another solution to the problem of unbounded integrals is to use the following *generalized NML (gNML) density* for a model family parameterized by parameter vector θ:

$$\bar{f}_w(y^n) = \frac{f(y^n\,; \hat{\theta}(y^n))\, w(\hat{\theta}(y^n))}{C^w} \tag{35}$$

where w is a non-negative function of the parameter vector called the *slack function*, and the normalizer C^w is given by

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

The range of integration Y may or may not be bounded. Using a constant function as the slack function w yields the standard first level NML density (4).

The name 'slack function' comes from the fact that the gNML density falls short of achieving the minmax regret of the standard NML density, i.e., the regret for sequence y^n is given by

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

which is seen not to be constant unless w is constant. In other words, for some sequences y^n the gNML density allows more *slack* in terms of regret than the standard NML density (and *vice versa* for some other sequences) if the latter is defined.

For gNML, the factorization (10) of the numerator becomes

$$f(y^n\,; \hat{\theta}(y^n))\; w(\hat{\theta}(y^n)) = f(y^n \mid \hat{\theta}(y^n))\; g(\hat{\theta}(y^n)\,; \hat{\theta}(y^n))\; w(\hat{\theta}(y^n)), \tag{36}$$

which shows that the slack function can also be seen as using instead of the canonical prior an arbitrary prior for the maximum likelihood estimator.

8 Generalized NML for linear regression

In the linear regression case, the denominator of the gNML density is given by

$$C^w = \int_Y f(y^n\,; \hat\beta(y^n), \hat\tau(y^n))\, w(\hat\beta(y^n), \hat\tau(y^n))\; dy^n.$$

Let the range Y be defined as the set of data sequences for which the maximum likelihood estimates $\hat\beta(y^n)$ and $\hat\tau(y^n)$ are such that the slack function w takes non-zero values. The integral can be done in three parts as before:

$$C^w = \int\int\int_{\{y:\hat\beta(y^n)=\hat\beta, \hat\tau(y^n)=\hat\tau\}} f(y^n\,; \hat\beta(y^n), \hat\tau(y^n))\, w(\hat\beta(y^n), \hat\tau(y^n))\; dy^n d\hat\beta d\hat\tau,$$

Using the factorization (36) we get

$$\begin{aligned} C^w = \int\int g(\hat\beta, \hat\tau\,; \hat\beta, \hat\tau)\; w(\hat\beta, \hat\tau) \\ \int_{\{y:\hat\beta(y^n)=\hat\beta, \hat\tau(y^n)=\hat\tau\}} f(y^n \mid \hat\beta(y^n), \hat\tau(y^n))\; dy^n d\hat\beta d\hat\tau. \end{aligned} \tag{37}$$

Once again, the innermost integral equals unity. Plugging in the canonical prior (13) yields

$$C^w = \left(\frac{n}{2e}\right)^{n/2} \frac{|\Sigma|^{1/2}}{\pi^{k/2}\Gamma\left(\frac{n-k}{2}\right)} \int \hat\tau^{-k/2-1} \int w(\hat\beta, \hat\tau)\; d\hat\beta d\hat\tau. \tag{38}$$

It now remains to be decided how to choose the slack function w so that the integral (38) remains bounded and the gNML density is defined.

Recall the second level NML density (23):

$$\bar f(y^n) = \frac{\bar f(y^n\,; \hat\tau_0(y^n), \hat R(y^n))}{C_{\tau_1,\tau_2,R_1,R_2}}.$$

The numerator is given by Eq. (20):

$$\bar f(y^n\,; \hat\tau_0(y^n), \hat R(y^n)) = \frac{k^2\Gamma\left(\frac{n-k}{2}\right)\Gamma\left(\frac{k}{2}\right)\hat\tau(y^n)^{k/2}}{(n\pi\hat\tau(y^n))^{n/2} 4\hat R(y^n)^{k/2}},$$

which differs from the numerator of the first level NML density (9) given by

$$f(y^n\,; \hat\beta(y^n), \hat\tau(y^n)) = (2\pi e\hat\tau(y^n))^{-n/2}$$

by a factor dependent on $\hat\tau(y^n)$ and $\hat\beta(y^n)$ through $\hat R(y^n)$. Let the function w be given by

$$w(\beta, \tau) = \tau^{k/2}\,(\beta'\Sigma\beta)^{-k/2}, \quad \text{if } \tau_1 \le \tau \le \tau_2,\ R_1 \le \beta'\Sigma\beta \le R_2, \tag{39}$$

and zero otherwise.

For data such that the slack function takes non-zero value at the maximum likelihood estimates, the numerator of the gNML density becomes

$$\begin{aligned} f(y^n\,;\hat{\beta}(y^n),\hat{\tau}(y^n))\,w(\hat{\beta}(y^n),\hat{\tau}(y^n)) &= (2\pi e\hat{\tau})^{-n/2}\hat{\tau}(y^n)^{k/2}\,(\hat{\beta}'(y^n)\Sigma\hat{\beta}(y^n))^{-k/2} \\ &= (2\pi e)^{-n/2}\hat{\tau}(y^n)^{-(n-k)/2}\,\hat{R}(y^n)^{-k/2}, \end{aligned}$$

which is seen to be equivalent to the numerator of the second level NML density (where we have $\hat{\tau}_0(y^n) = \hat{\tau}(y^n)$) except for a constant c' independent on the parameter values:

$$f(y^n\,;\hat{\beta}(y^n),\hat{\tau}(y^n))\,w(\hat{\beta}(y^n),\hat{\tau}(y^n)) = c'\bar{f}(y^n\,;\hat{\tau}_0(y^n),\hat{R}(y^n)).$$

Since the range of integration (support) is the same for both the gNML density and the second level (standard) NML density, the two densities must be identical, which shows that the two-fold normalization procedure is equivalent to using the generalized NML density with the slack function given by Eq. (39). Whether the slack function induced by the two-fold normalization procedure is a reasonable choice is of course an interesting question.

Example 1. Fig. 1 compares the standard NML density and the second level NML (or equivalently, the gNML density with the slack function given by Eq. (39)) in terms of the likelihood ratio of each of the two densities versus the maximized likelihood (9). The second level NML density gives non-zero likelihood to data with the maximum likelihood estimate $\hat{R}(y^n)$ within the range $[R_1, R_2]$, whereas the standard NML density gives non-zero likelihood to data with $\hat{R}(y^n) \leq R$. In case both densities give non-zero density values, the second level NML density gives higher density for data with $\hat{R}(y^n)$ small than the standard NML density.

9 Generalized NML for wavelet denoising

For any wavelet basis, the shape of the slack function (39) is very simple. First, by Eqs. (33) and (34) we have

$$w(\hat{\beta}(y^n),\hat{\tau}(y^n)) = n^{-k}\left(\sqrt{S(y^n)-S_\gamma(y^n)}\right)^k\left(\sqrt{S_\gamma(y^n)}^{-1}\right)^k,$$

which is a product of three factors; a constant together with a factor that depends on the coefficients that are set to zero, and another factor that depends on the retained ones. Further, the two factors are both of the form $(\cdot)^k$ where in the latter case the argument is the inverse of the Euclidean norm of the vector of retained coefficients. Especially this latter factor suggests an interpretation

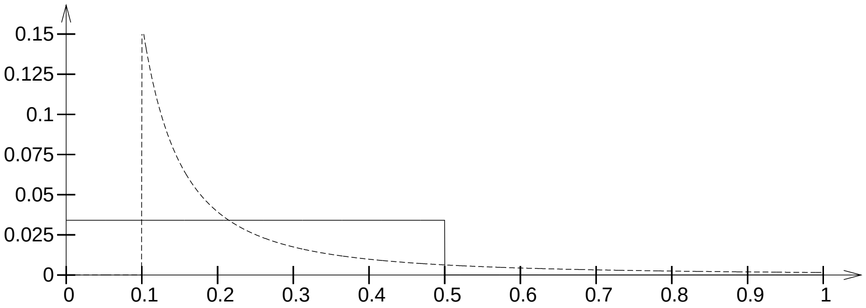

Figure 1: Ratios of the standard NML density (solid line) and the second level NML density (dashed line) to the maximum likelihood density as a function of $\hat{R}(y^n)$. The setting is defined by $n = 8, k = 4, \hat{\tau}(y^n) = 0.5$; the hyperparameters of the standard NML density are given by $\tau_0 = 0.1, R = 0.5$, and for the second level NML density by $\tau_1 = 0.1, \tau_2 = 1, R_1 = 0.1, R_2 = 1$.

of the slack function as a spherical prior density for the retained coefficients proportional to the inverse of the Euclidean norm. Such a prior density is not proper for the whole real space but this is not a problem since the range of integration is bounded by the hyperparameters R_1 and R_2.

It has been suggested that wavelet coefficients in ‘noiseless’ natural images tend to be well modeled by a Laplacian density. It might therefore be reasonable to use a slack function that has the shape of a Laplacian density in terms of the components of β.