Complexity and Information in Modeling

J. Rissanen[*]
Helsinki Institute for Information Technology,
Tampere and Helsinki Universities of Technology, Finland, and
University of London, England

1 Introduction

There are three fundamental notions in model selection, *information, complexity*, and *noise.* And yet none of them have been defined in a precise manner in traditional statistics. Intuitively, 'information' has to do with the regular features and restrictions constraining the data in a statistical manner, and 'complexity' is associated with the number of parameters in the models, while 'noise' is often viewed as the high frequency part in the data. In this paper we define these three notions in a formal manner so that they can be measured, and we regard the purpose of model selection to be to extract either all the information from the data that can be extracted with a model class available or a desired amount by restricting the model complexity appropriately. All the information can be extracted with the *MDL* (Minimum Description Length) principle, [9] and [10], which then gives yet another interpretation for this principle.

It has been claimed that the *MDL* principle is equivalent with posterior maximization in Bayesian approaches, or its role has been restricted to a rederivation of the model selection criterion BIC, [4]. In fact, one may regard the posterior maximization to produce only an asymptotically justified approximation of the shortest code length called for by the *MDL* principle, which is what also BIC is, [12]. Unlike BIC, which has not changed at all over the years and fails to account for the structure in which parameters appear as well as the varying influence of the parameters to the likelihood function, the *MDL* principle has evolved and is still evolving. It has grown into a totally different theory of modeling and indeed of all statistical inference with new powerful techniques for solutions that cannot be arrived at by Bayesian nor orthodox statistical procedures. And, moreover, unlike the traditional approaches, the new one is logically sound without any metaphysical assumptions. It spells out the limitations in all model selection, above all the fact that to find the best model is noncomputable, which should put an end to the futile and misleading endeavors to find one. For a simple tutorial on the *MDL* principle we refer to [5], and a more advanced tutorial can be found in [2].

In all the traditional approaches model selection is done by assuming that the data resulted

[*] This author's work was supported by the Academy of Finland under contract 53884. Chapter IV in the book *Computability, Complexity and Constructivity in Economic Analysis*, (Editor: K Vela Velupillai ISBN 1405130784), Blackwell Publishings, Oxford, Publications Dates: US: May, 2005; Rest of the World - March, 2005; Pages: 324

from sampling a 'true' probability distribution, which is to be estimated from the data by use of various principles such as the maximum likelihood principle or by minimization of some mean loss, the mean taken with respect to the imagined 'true' distribution or the posterior in the Bayesian approaches. Since frequently the sought for regular features are bafflingly complex the 'true' imagined distribution must also be assumed to be complex, which creates the fundamental difficulty of selecting the appropriate complexity of the models to be fitted to the data. Experience shows that fitting too complex a model will not capture the regular features well, which is reflected in poor performance of the found model in new data generated by the same physical machinery. In the traditional approaches the fundamental issues of 'information' and 'complexity' are dealt with in ad hoc manner combined with sound judgement such as by addition of various 'complexity' penalizing terms to the model selection criterion.

Although there is a fundamental flaw in statistical procedures based on the assumption of the 'true' data generating distribution, it is deep rooted. It is often assumed in a disguised form as an ideal unreachable model to be estimated by simpler real models, which then leads to the awkward situation that we are modeling a model. It is reflected even in the often quoted and well meaning phrase that "all models are wrong but some are useful", which strictly speaking is meaningless without the existence of something that is 'right' or 'true'. This is particularly clear if by a model of a data string is meant a finite set that includes the string, as was done by Kolmogorov in the algorithmic theory of complexity, which we discuss below. It would be absurd to regard one such model 'true' and the others 'false'. I'm convinced that a proper and in the end useful theory of model selection cannot be done without abandoning such a preposterous idea of model and without addressing in a formal way the fundamental issues of 'complexity', 'information' and 'noise'.

2 Kolmogorov's structure function

The basic ideas of 'noise', 'information', and 'complexity' are easiest to explain in the algorithmic theory of information, which can be done with very few technicalities.

The *Kolmogorov complexity* of a binary string $x = x^n = x_1, \ldots, x_n$, relative to a universal computer U, is defined as

$$K(x) = \min_{p(x)} |p(x)|,$$

where $|p(x)|$ is the length of a self-delimiting program in the language U that generates the string, [16], [6]. Such a program, also a binary string, is a codeword of the string x, which can be decoded by running the program. The requirement 'self-delimiting' means that no program is a prefix of another, which can be obtained in any universal language by adding a few simple instructions. If we organize the countable set of all self-delimiting programs in a binary tree, each program is a leaf, and the important Kraft-inequality holds

$$\sum_x 2^{-K(x)} \leq 1, \tag{1}$$

where the summation is over all finite binary strings. This, in turn, means that by normalization we can define a *universal* probability distribution for all binary strings. Since by a suitable coding all data strings can be represented as binary strings we have a universal distribution for all data strings. Such a universal distribution has the remarkable property that the ratio of the probability it assigns to any binary string and any algorithmically defined probability for the same string is bounded away from zero by a constant, not dependent on the length of the string; for more details we refer to [7].

Originally the complexity of a string was defined to be its information, which terminology was in keeping with Shannon's probabilistic information. This term was a bad choice and led to criticism of the kind that the majority of strings obtained by flipping a fair coin have the maximum amount of information, because their shortest programs were about as long as the length of the strings. After all, one can easily write a program that simply tells the computer to copy the string symbol for symbol, and no shorter program exists. And yet intuitively we would regard such a string to give us no information. In light of such criticism the name 'information' was changed to 'complexity'. And yet there is a place even for 'information', which roughly speaking means the amount of regular features a string has. Since the regular features in a string define its 'model', the 'information' in a string will be the shortest description or code length of the regular features, which forms the foundation for the *MDL* principle in statistical modeling even before Kolmogorov's structure function formalization.

Kolmogorov defined a *model* of string x to be a finite set S that includes the string. This corresponds to intuition in that all strings in the set share the same properties, namely the properties that define the set. These, in turn, can be defined by a program, and the length of a shortest one may be written as $K(S)$. It is clear that if a set S that includes the string $x = x^n$ is given we can encode x with no more than $\log |S|$ bits, where the logarithm base is two and $|S|$ denotes the number of elements in S. Indeed, we can order the elements of S lexically, and encode each as its ordinal the largest of which is $|S|$. We now have the ingredients in Kolmogorov's structure function of a parameter α, [17],

$$h_x(\alpha) = \min_{S \ni x} \{\log |S| : K(S) \leq \alpha\}. \tag{2}$$

A small set captures more properties than a large one. For instance the set X^n of all strings of length n clearly captures no properties of $x = x^n$ other than the length n, and the singleton set $\{x\}$ captures all conceivable properties of x shared by no other string. Bearing this in mind we see that the minimizing set S_α extracts all properties from x on the level α; that is, with 'model cost' (= code length needed to describe S) not exceeding α.

If $\alpha > \alpha'$, $h_x(\alpha) < h_x(\alpha')$, because the minimization is done over a larger collection of sets. Hence $h_x(\alpha)$ decreases from its maximum, about $n = \log |X^n|$ at $\alpha = 0$ to zero at $\alpha = K(x)$, the complexity of the singleton set $\{x\}$. Actually in the algorithmic theory of information equalities and inequalities between code lengths; ie, program lengths, are taken to within constants, and we write them as $\doteq$ and $\dot{\geq}$. The code length $h_x(\alpha) + \alpha$ for the pair x, S cannot be smaller than the code length $K(x)$ for x. Hence,

$$h_x(\alpha)\dot{\geq}K(x)-\alpha,$$

the right hand side defining the *sufficiency line*. There is a special value $\bar{\alpha}$ defined as

$$\bar{\alpha}=\min\{\alpha : h_x(\alpha)+\alpha\doteq K(x)\}. \tag{3}$$

The two-part code length

$$h_x(\bar{\alpha})+\bar{\alpha}$$

represents the Kolmogorov *minimal sufficient statistics decomposition*, in which $S_{\bar{\alpha}}$ represents all learnable properties of x that can be captured by finite sets leaving $h_x(\bar{\alpha})$ as the code length for noninformative 'noise'.

3 Probability Model Classes

The Kolmogorov complexity is noncomputable, and so is the sufficient statistics decomposition, which means that the constructs described in the algorithmic theory cannot be applied directly. The *MDL* theory was patterned after the algorithmic theory but with a far less powerful language in which to represent the regular features in data, namely, a class of probability models. Because the models must be capable of being fitted to data they must be finitely describable and hence in the end parametric. For instance, histograms both with variable and constant bin width, will be considered parametric. This differs from the traditional usage where a model can be even nonparametric which cannot be fitted to data, and the likelihood function is regarded as a model. In our case the likelihood function corresponds to a class of parametric density or probability functions as models

$$\mathcal{M}_\gamma=\{f(x^n;\theta,\gamma):\theta\in\Omega_\gamma\subseteq R^k\},\;\; \mathcal{M}=\bigcup_\gamma \mathcal{M}_\gamma,$$

where γ is a structure index such as the indices of some of the rows of a regressor matrix and $\theta=\theta_1,\ldots,\theta_k$, k depending on γ. For much of the discussion the structure index will be constant, and in order to simplify notations we write $f(x^n;\theta)$ for $f(x^n;\theta,\gamma)$ and $\mathcal{M}_k$ for $\mathcal{M}_\gamma$.

We make no assumption about the data that they be samples from any distribution. Also the data may consist of pairs $(y^n,x^n)=(y_1,x_1),(y_2,x_2),\ldots,(y_n,x_n)$ of symbols of any kind, in which case the models are written as $f(y^n|x^n;\theta,\gamma)$. This generalization does not introduce any relevant changes to the discussion, and we consider the previous case. We also take the data as real numbers. The model class is selected on prior knowledge about the data to be modeled. No meaningful theory can exist for its selection. However, we can compare the goodness of several suggestions.

In order to define the structure function for the probability models we need to replace $K(x)$ by the *stochastic complexity* as the negative logarithm of the normalized maximum likelihood density function, [12], the model cost $K(S)$ by the shortest code length $L(\theta^d,k)$ for quantized parameters θ^d and their number k, and $\log|S|$ by the worst case code length in the set of 'typical' strings of $f(\cdot;\theta^d)$, all of which will be discussed in the following subsections.

3.1 Stochastic Complexity

Consider the normalized maximum likelihood *NML* density function

$$\hat{f}(x^n;\mathcal{M}_k) = \frac{f(x^n;\hat{\theta}(x^n))}{C_{n,k}} \tag{4}$$

$$\begin{aligned} C_{n,k} &= \int_{\hat{\theta}(y^n)\in\Omega^\circ} f(y^n;\hat{\theta}(y^n))dy^n \qquad (5)\\ &= \int_{\hat{\theta}\in\Omega^\circ} g(\hat{\theta};\hat{\theta})d\hat{\theta}, \end{aligned}$$

where Ω° is the interior of Ω, assumed to be compact, and $g(\hat{\theta};\theta)$ is the density function on the statistic $\hat{\theta}(y^n)$ induced by $f(y^n;\theta)$. The notations dy^n and $d\hat{\theta}$ refer to differential volumes.

We take the expression

$$-\log\hat{f}(x^n;\mathcal{M}_k) = -\log f(x^n;\hat{\theta}(x^n)) + \log C_{n,k} \tag{6}$$

as the 'shortest code length' for the data x^n that can be obtained with the model class $\mathcal{M}_k$ and call it the **Stochastic Complexity** of x^n, given $\mathcal{M}_k$, [12]. This term needs clarification. First, if the models are density functions $f(x)$ they induce probabilities $p(x) \cong f(x)\delta$ to the necessarily rational valued data, written to some precision δ. Hence if we add the number $-n\log\delta$ to the stochastic complexity we get a code length that differs from the stochastic complexity by an irrelevant number. Secondly, the stochastic complexity cannot represent literally the shortest code length for every data sequence. Rather, it will be that in a probabilistic sense, which however is strong enough to mean the shortest for all intents and purposes unless n is small.

If the CLT holds for $\hat{\theta}(y^n)$ we have the convergence,

$$C_{n,k}\left(\frac{2\pi}{n}\right)^{k/2} \to \int_{\Omega} |J(\theta)|^{1/2}d\theta, \tag{7}$$

where

$$J(\theta) = \lim_{n\to\infty} -n^{-1}\{E_\theta \frac{\partial^2 \log f(X^n;\theta)}{\partial\theta_i\partial\theta_j}\} \tag{8}$$

is a generalization of Fisher's information matrix. We assume its elements to be continuous functions of θ.

The first justification of the term 'stochastic complexity' is the following maxmin problem

$$\max_g \min_q E_g \log\frac{f(X^n;\hat{\theta}(X^n))}{q(X^n)} = \max_g \min_q [D(g\|q) - D(g\|\hat{f}(x^n;\mathcal{M}_k) + \log C_{n,k}] = \log C_{n,k},$$

solved by $\hat{q} = \hat{g} = \hat{f}(x^n;\mathcal{M}_k)$, where $D(g\|q)$ is the Kullback-Leibler distance, see Appendix. We see that $\log 1/f(x^n;\hat{\theta}(x^n))$ is an unreachable target for any code length obtainable with the members in the model class, so that the *NML* density function gets closest to this in the mean, the mean taken with respect to the worst case data generating distribution. To see this notice that the minimizing q for any g is $\hat{q}(g) = g$, and the unique maximizing g is $\hat{f}(x^n;\mathcal{M}_k)$.

Also $\hat{q} = \hat{f}(x^n; \mathcal{M}_k)$ and any g solve the minmax problem

$$\min_q \max_g [D(g\|q) - D(g\|\hat{f}(x^n; \mathcal{M}_k) + \log C_{n,k}] = \log C_{n,k}.$$

In fact, the minmax value is lower bounded by the maxmin value, and since $\hat{g}(q) = g$, any g, for $q = \hat{f}(x^n; \mathcal{M}_k)$ is a maximizing g reaching the value $\log C_{n,k}$, we are done. The minmax problem is also seen to be equivalent with Shtarkov's minmax problem, [15]:

$$\min_q \max_{x^n} \log \frac{f(x^n; \hat{\theta}(x^n))}{q(x^n)} = \log C_{n,k}.$$

The second justification is the theorem, [11], stating that if the model class satisfies mild conditions and if we assume that the data have been generated by some distribution $f(x^n; \theta)$ in the model class $\mathcal{M}_k$, then for all ϵ, all θ, and all sufficiently large n,

$$E_{g \in \mathcal{M}_k} \frac{1}{n} \log 1/q(X^n) - H(g) \geq \frac{k - \epsilon}{2n} \log n,$$

except for θ in a set whose volume goes to zero as n grows. In view of the formula for $C_{n,k}$ in (7) we see that the mean of $-\log \hat{f}(x^n; \mathcal{M}_k)$, the mean taken with respect to a distribution $\mathcal{M}_k$, cannot be beaten to within a constant with any code except for θ in a vanishing set.

3.2 Code Length for Models

In order to get the code length for models we need the existence of a special partition of the compact parameter space Ω. The construct is somewhat intricate, and we just outline it; for more details we refer to [14]. We want the partitioning to consist of curvilinear hyper rectangles such that the Kullback-Leibler distance $D(f_i\|f_j)$ between the models $f_i = f(y^n; \theta^i)$ and $f_j = f(y^n; \theta^j)$, defined by the centers of two adjacent rectangles $\theta^i = \theta(i)$ and $\theta^j = \theta(j)$, is the same for any pair. To achieve this apply Taylor's expansion to the two adjacent models, which gives

$$D(f_i\|f_j) = \frac{n}{2}(\theta^j - \theta^i)' J(\tilde{\theta})(\theta^j - \theta^i),$$

where $\tilde{\theta}$ is a point between θ^i and θ^j.

Consider a hyper ellipsoid centered at θ^i

$$\delta' J(\theta^i)\delta = d/n, \tag{9}$$

where $J(\theta^i)$ is the Fisher information matrix, (8), and $\delta = \theta - \theta^i$. It encloses a rectangle $B_{i,n}(d)$ of maximum volume

$$V = \left(\frac{4d}{nk}\right)^{k/2} |J(\theta^i)|^{-1/2}. \tag{10}$$

Actually since $J(\theta^i)$ is not constant over the parameter space the rectangles have to be curvilinear defined by differential equations in order for them to form a partition. However, we are interested in the case where n is large and the rectangles small so that the volume difference between the straight line edge and the curvilinear rectangles is ignorable.

Consider the *canonical* 'prior' density function for $\hat{\theta}$

$$w(\hat{\theta}) = \frac{g(\hat{\theta};\hat{\theta})}{\int_\Omega g(\theta;\theta)d\theta} \cong \frac{|J(\hat{\theta})|^{1/2}}{\int_\Omega |J(\theta)|^{1/2}d\theta},$$

the approximation being Jeffreys' prior. This defines a probability distribution for the centers θ^i, which tends to *uniform* as n grows:

$$\begin{aligned} \pi_d(\theta^i) &= \int_{B_{i,n}(d)(\theta^i)} w(\theta)d\theta && (11)\\ \frac{\pi_d(\theta^i)}{w(\theta^i)|B_{i,n}(d)|} &\to 1 && (12)\\ \frac{\pi_d(\theta^i)}{\left(\frac{2d}{\pi k}\right)^{k/2} C_{n,k}} &\to 1. && (13) \end{aligned}$$

Here $|B_{i,n}(d)|$ denotes the volume of $B_{i,n}(d)$. With this approximation we get the code length for the model, defined by the center θ^i,

$$L_d(\theta^i) \cong \frac{k}{2}\log\frac{\pi k}{2d} + \log C_{n,k}. \tag{14}$$

This also gives the number of rectangles partitioning Ω:

$$C_{n,k}\left(\frac{k\pi}{2d}\right)^{k/2}.$$

We mention that in [1] $C_{n,k}$ was given the interpretation of the number of optimally *distinguishable* models from data x^n in a somewhat intricate sense.

4 Structure Function

We consider the set $X_{i,n}(d) = \{y^n : \hat{\theta}(y^n) \in B_{i,n}(d)\}$ as the set of typical strings of the model defined by θ^i. Just as in the algorithmic theory $\log|S|$ is the code length of the worst case sequence in S, we need the code length of the worst case sequence y^n in $X_{i,n}(d)$. If $B_{i,n}(d)$ is small $f(x^n;\hat{\theta}(x^n))$ does not vary much for $x^n \in X_{i,n}(d)$, and the worst case sequence y^n is one for which

$$n(\hat{\theta}(y^n) - \theta^i)'J(\theta^i)(\hat{\theta}(y^n) - \theta^i) = d.$$

It will be convenient to take the logarithm base as natural. We define the structure function for the model class $\mathcal{M}_k$ as follows

$$h_{x^n}(\alpha) = \min_d\{-\ln f(x^n;\hat{\theta}(x^n)) + \frac{d}{2} : L_d(\theta^i) \leq \alpha\}. \tag{15}$$

For the minimizing d the inequality will have to be satisfied with equality,

$$\alpha = \frac{k}{2}\ln\frac{\pi k}{2d} + \ln C_{n,k},$$

and with the asymptotic approximation (14) we get

$$d_\alpha = \frac{\pi k}{2}C_{n,k}^{2/k}e^{-2\alpha/k}, \tag{16}$$

and

$$h_{x^n}(\alpha) = -\ln f(x^n; \hat{\theta}(x^n)) + d_\alpha/2. \tag{17}$$

We may ask for the values of α for which the structure function is closest to the sufficiency line defined by

$$-\ln \hat{f}(x^n; \mathcal{M}_k) - \alpha,$$

which amounts to the minimization of the 2-part code length

$$\min_\alpha \{h_{x^n}(\alpha) + \alpha\}. \tag{18}$$

With (17) and (16) we get the minimizing α as

$$\bar{\alpha} = \frac{k}{2} \ln \frac{\pi}{2} + \ln C_{n,k}, \tag{19}$$

and $d_{\bar{\alpha}} = k$. We then get the *universal sufficient statistics decomposition* of the model class $\mathcal{M}_k$,

$$h_{x^n}(\bar{\alpha}) + \bar{\alpha} = -\ln f(x^n; \hat{\theta}(x^n)) + \frac{k}{2} + \frac{k}{2} \ln \frac{\pi}{2} + \ln C_{n,k}. \tag{20}$$

The last two terms as the code length of the optimal model represent the optimal amount of information one can extract from the string with the model class $\mathcal{M}_k$, leaving the first two terms, $h_{x^n}(\bar{\alpha})$, as the code length of whatever remains in the data, the 'noise'. The models for which the values of α are larger than $\bar{\alpha}$ also extract all the information from the data, but in so doing they try to explain some of the noise. The interesting models correspond to the range $\alpha \leq \bar{\alpha}$, for they incorporate a portion of the learnable properties for a smaller 'model cost' or the code length for the optimal model on that level, and they leave a greater amount as unexplained noise.

We next study the universal sufficient statistics decomposition for the model class $\mathcal{M}$. We need a probability function $q(k)$ for the number of parameters, which can be taken as uniform, say $1/n$. The code length for the models is then increased from the previously discussed case by adding $\ln n$ to it. The structure function is now

$$h_{x^n}(\alpha) = \min_{d,k} \{-\ln f(x^n; \hat{\theta}(x^n)) + \frac{1}{2} d : -\ln \pi_d(\theta^i) + \ln n \leq \alpha\}. \tag{21}$$

For each k the minimizing value for d is

$$d_{\alpha,k} = \frac{\pi k}{2} (n C_{n,k})^{2/k} e^{-2\alpha/k},$$

and it is reached when the code length for the optimal model is α.

The minimum of the 2-part code length $h_{x^n}(\alpha) + \alpha$ is

$$\min_{d,k} [-\ln f(x^n; \hat{\theta}(x^n)) + \frac{1}{2} d - \ln \pi_d(\theta^i) + \ln n]. \tag{22}$$

For each k the minimizing value for d is $\hat{d} = k$, as before, and we are left with the minimization

$$\min_k \{-\ln \hat{f}(x^n; \mathcal{M}_k) + \ln n + \frac{k}{2} \ln \frac{\pi e}{2}\}.$$

Letting $\hat{k}$ denote the minimizing number of parameters, we get

$$h_{x^n}(\hat{\alpha}) = -\ln f(y^n; \theta^i, \hat{k}) = -\ln f(x^n; \hat{\theta}(x^n)) + \hat{k}/2, \tag{23}$$

where

$$\hat{\alpha} = \frac{\hat{k}}{2} \ln \frac{\pi}{2} + \ln(C_{n,\hat{k}} n). \tag{24}$$

This gives

$$h_{x^n}(\hat{\alpha}) + \hat{\alpha} = -\ln \hat{f}(x^n; \hat{\gamma}) + \ln n + \frac{\hat{k}}{2} \ln \frac{\pi e}{2}. \tag{25}$$

The volume of the $\hat{k}-$dimensional cells $B_{i,n}(\hat{k})$ is given by

$$(\frac{4}{n})^{\hat{k}/2} |J(\theta^i)|^{-1/2}. \tag{26}$$

This means that the number of the cells corresponding to $d = \hat{k}$ is $(\pi/2)^{n\hat{k}/2} C_{n,k}$. As above, $h_{\mathbf{x}^n}(\alpha)$ stays above the sufficiency line $L(\alpha) = -\ln \hat{f}(x^n; \mathcal{M}_{\hat{k}}) + d_{\alpha,\hat{k}}/2$, except at the point $\hat{\alpha}$, where they are equal. We also see that the optimal 2-part code length exceeds the stochastic complexity only by the constant $d_{\alpha,\hat{k}}/2$.

We conclude this section with an example.

Example: Consider the class of Bernoulli models with one parameter $\theta = Prob(X = 0)$. The parameter space is the unit interval. The width of the equivalence class for the parameter d is

$$|B_{i,n}(d)| = (\frac{4d}{n})^{1/2} ((i/n)(1 - i/n))^{1/2}. \tag{27}$$

The canonical probability distribution for the centers of the equivalence classes is from (14)

$$\pi_d(i/n) = \frac{\sqrt{2d/\pi}}{C_{n,1}},$$

where $C_{n,1}$ is given by

$$\log C_{n,1} = \frac{1}{2} \log \frac{n\pi}{2} + o(1).$$

The structure function is given by

$$h_{x^n}\alpha = \min_d \{nh(n_0/n) + \frac{d}{2} : -\log \pi_d(i/n) \leq \alpha\}, \tag{28}$$

where n_0 denotes the number of 1's in the string x^n and $h(n_0/n)$ is the binary entropy function evaluated at the point n_0/n. Further, i/n is the center of the equivalence class where n_0/n falls. The minimizing value for d is given by

$$d_\alpha = \frac{n\pi^2}{4} 2^{-2\alpha}.$$

5 Linear Regression

The normal density functions arising in the linear quadratic regression problem pose a special problem in calculating the universal sufficient statistics decomposition due to the fact that the normalizing coefficient $C_{n,\gamma}$ does not exist unless we restrict the range of integration. This can be done but it requires hyperparameters, which affect the important problem of selecting the optimal collection of the explanatory or regressor variables. However, with a double normalization we obtain hyperparameters which do not affect the optimal selection of the regressor variables, [13]. Instead of calculating the universal sufficient statistics strictly as explained above we ignore the fact that the real valued parameters should be quantized optimally as defined by the parameter $d_{\bar{\alpha}}$ and do the decomposition in terms of the stochastic complexity. The difference in finding the optimal structure index γ will be ignorable.

Consider the basic linear regression problem, where we have data of type $(y_t, x_{1t}, x_{2t}, \ldots, x_{Kt})$ for $t = 1, 2, \ldots, n$. We fit a linear model of type

$$y_t = \beta' \underline{x}_t + \epsilon_t = \sum_{i \in \gamma} \beta_i x_{it} + \epsilon_t, \tag{29}$$

where $\gamma = \{i_1, \ldots, i_k\}$ denotes a subset of the indices of the regressor variables; the prime denotes transposition, and for the computation of the required code lengths the deviations ϵ_t are modeled as samples from an iid Gaussian process of zero mean and variance $\tau = \sigma^2$, also as a parameter. In such a model the response data $y^n = y_1, \ldots, y_n$ are also normally distributed with the density function

$$f(y^n | X_\gamma; \beta, \tau) = \frac{1}{(2\pi\tau)^{n/2}} e^{-\frac{1}{2\tau} \sum_t (y_t - \beta' \underline{x}_t)^2}, \tag{30}$$

where $X'_\gamma = \{x_{it} : i \in \gamma, t = 1, \ldots, n\}$ is the $k \times n$ matrix defined by the values of the regressor variables with indices in γ; we also write $f(y^n; \gamma, \beta, \tau)$ for $f(y^n | X_\gamma; \beta, \tau)$. Write $Z_\gamma = X'_\gamma X_\gamma = n\Sigma_\gamma$, which is taken to be positive definite. The development for a while will be for a fixed γ, and we drop the subindex γ in the matrices above. The maximum likelihood solution of the parameters is given by

$$\begin{aligned} \hat{\beta}(y^n) &= Z^{-1} X' y^n & (31)\\ \hat{\tau}(y^n) &= \frac{1}{n} \sum_t (y_t - \hat{\beta}'(y^n) \underline{x}_t)^2, & (32) \end{aligned}$$

where $\underline{x}_t$ denotes the column vector of the data x_{it} for $i \in \gamma$. The density function (30) admits the sufficient statistics factorization in the usual sense rather than in the sense discussed above

$$\begin{aligned} f(y^n; \gamma, \beta, \tau) &= f(y^n | \hat{\beta}, \hat{\tau}) p_1(\hat{\beta}; \beta, \tau) p_2(n\hat{\tau}/\tau; \tau) \frac{n}{\tau} & (33)\\ f(y^n | \gamma, \hat{\beta}, \hat{\tau}) &= (2\pi)^{\frac{k-n}{2}} n^{n/2} |\Sigma|^{-1/2} \Gamma(\frac{n-k}{2}) 2^{\frac{n-k}{2}} \hat{\tau}^{1 - \frac{n-k}{2}} & (34)\\ p_1(\hat{\beta}; \beta, \tau) &= \frac{n^{k/2} |\Sigma|^{1/2}}{(2\pi\tau)^{k/2}} e^{\frac{n}{2\tau} (\hat{\beta} - \beta)' \Sigma (\hat{\beta} - \beta)} & (35)\\ p_2(n\hat{\tau}/\tau; \tau) &= (n\hat{\tau}/\tau)^{\frac{n-k}{2} - 1} e^{-n\frac{\hat{\tau}}{2\tau}} 2^{-\frac{n-k}{2}} \Gamma^{-1}(\frac{n-k}{2}). & (36) \end{aligned}$$

We calculate next the *NML* density function

$$\hat{f}(y^n; \gamma) = \frac{f(y^n; \gamma, \hat{\beta}(y^n), \hat{\tau}(y^n))}{\int_{Y(\tau_0, R)} f(z^n; \gamma, \hat{\beta}(z^n), \hat{\tau}(z^n)) dz^n}, \tag{37}$$

where

$$Y(\tau_0, R) = \{z^n : \hat{\tau}(z^n) \geq \tau_0,\, \hat{\beta}'(y^n)\Sigma\hat{\beta}(y^n) \leq R\}. \tag{38}$$

The numerator has a very simple form

$$f(y^n; \gamma, \hat{\beta}(y^n), \hat{\tau}(y^n)) = 1/(2\pi e\hat{\tau}(y^n))^{n/2}, \tag{39}$$

and the problem is to evaluate the integral in the denominator.

Putting $\theta = \beta, \tau$ and integrating the conditional $f(y^n|\hat{\theta}(y^n); \gamma, \theta) = h(y^n)$ over y^n such that $\hat{\theta}(y^n)$ equals any fixed value $\hat{\theta}$ yields unity. Therefore with $p(\hat{\theta}; \gamma, \hat{\theta}) \equiv g(\hat{\tau})$ we get from the expression for the χ^2 density function in (33),

$$\begin{aligned} C(\tau_0, R) &= \int_{Y(\tau_0,R)} f(z^n; \gamma, \hat{\theta}(z^n))dz^n & (40)\\ &= \int_{\tau_0}^{\infty} \hat{\tau}^{-\frac{k+2}{2}} d\hat{\tau} \int_{B_R} d\beta & (41)\\ &= A_{n,k} V_k \frac{2}{k} \left(\frac{R}{\tau_0}\right)^{k/2}, & (42)\end{aligned}$$

where $B_R = \{\beta : \beta'\Sigma\beta \leq R\}$ is an ellipsoid,

$$V_k R^{k/2} = |\Sigma|^{-1/2} \frac{2\pi^{k/2} R^{k/2}}{k\Gamma(k/2)} \tag{43}$$

its volume, and

$$A_{n,k} = \frac{|\Sigma|^{1/2}}{\pi^{k/2}} \frac{(\frac{n}{2e})^{\frac{n}{2}}}{\Gamma(\frac{n-k}{2})}. \tag{44}$$

We then have the *NML* density function itself for $0 < k < n$ and $y = y^n$

$$-\ln \hat{f}(y; \gamma, \tau_0, R) = \frac{n}{2} \ln \hat{\tau} + \frac{k}{2} \ln \frac{R}{\tau_0} - \ln\Gamma(\frac{n-k}{2}) - \ln\Gamma(\frac{k}{2}) + \ln \frac{4}{k^2} + \frac{n}{2} \ln(n\pi). \tag{45}$$

We wish to get rid of the two parameters R and τ_0, which clearly affect the criterion in an essential manner, or rather we replace them with other parameters which do not influence the relevant criterion. This is done by renormalization. To do it we need to set the two parameters to the values that minimize (45): $R = \hat{R}$, and $\tau_0 = \hat{\tau}$, where $\hat{R} = \hat{\beta}'(y)\Sigma\hat{\beta}(y)$. Then the new NML density function is given by

$$\hat{\hat{f}}(y; \gamma) = \frac{\hat{f}(y; \gamma, \hat{\tau}(y), \hat{R}(y))}{\int_Y \hat{f}(z; \gamma, \hat{\tau}(z), \hat{R}(z))dz}, \tag{46}$$

where the range Y will be defined presently. By (33) and the subsequent equations we also have the factorization

$$\hat{f}(y; \gamma, \tau_0, R) = f(y|\gamma, \hat{\beta}, \hat{\tau})g(\hat{\tau})/C(\tau_0, R) = f(y|\gamma, \hat{\beta}, \hat{\tau})\frac{k}{2}\hat{\tau}^{-k/2-1}V_k^{-1}\left(\frac{\tau_0}{R}\right)^{k/2}. \tag{47}$$

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 of $\hat{\tau}$ over a range $[\tau_1, \tau_2]$ and $\hat{R}(\beta)$ over the volume between the hyper ellipsoids bounded by $[R_1 \leq \hat{R} \leq R_2]$. All told we get

$$\begin{aligned} \int_Y \hat{f}(z; \gamma, \hat{\tau}(z), \hat{R}(z))dz &= \frac{k}{2} V_k^{-1} \int_{\tau_1}^{\tau_2} \int_{R_1}^{R_2} \tau^{-1} R^{-k/2} d\tau dV \\ &= (\frac{k}{2})^2 \ln \frac{\tau_2}{\tau_1} \ln \frac{R_2}{R_1}, \end{aligned}$$

where we expressed the volume element as

$$dV = \frac{k}{2} V_k R^{\frac{k}{2}-1} dR.$$

(This corrects a small error in [13].)

The negative logarithm of $\hat{f}(y;\gamma)$ is then given by

$$-\ln \hat{f}(y;\gamma) = \frac{n-k}{2} \ln \hat{\tau} + \frac{k}{2} \ln \hat{R} - \ln \Gamma(\frac{n-k}{2}) - \ln \Gamma(\frac{k}{2}) + \frac{n}{2} \ln(n\pi) + \ln[\ln \frac{\tau_2}{\tau_1} \ln \frac{R_2}{R_1}]. \qquad (48)$$

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

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

$$\min_{\gamma} \{(n-k) \ln \hat{\tau} + k \ln \hat{R} + (n-k-1) \ln \frac{1}{n-k} - (k-1) \ln k\}, \qquad (49)$$

where k denotes the number of elements in γ. This has no other parameters than the set of indices of the rows of the regressor matrix, and it can be applied to testing various hypotheses and to find the best index set.

5.1 Denoising

An important special case of regression is the denoising problem, in which the regressor matrix, written now as W, is $n \times n$, and the regression data as x^n. The problem is to remove noise from the data sequence $x' = x^n = x_1, \ldots, x_n$, taken as a row vector, so that the remaining 'smooth' signal $\hat{x}' = \hat{x}^n = \hat{x}_1, \ldots, \hat{x}_n$ represents the information bearing data:

$$x_t = \hat{x}_t + \epsilon_t, \;\; t = 1, \ldots, n. \qquad (50)$$

As in the regression problem $\hat{x}^n$ is written as a linear combination of a collection $\gamma = \{i_1, \ldots, i_k\}$ of the rows $\underline{w}'_i = w_{i1}, \ldots, w_{in}$ of W with indices in $\gamma = \{i_1, \ldots, i_k\}$ thus

$$\hat{x}_t = \sum_{j \in \gamma}^{k} c_j w_{jt}.$$

A great simplification results if W is orthonormal, which can be achieved with wavelets, and it is not even necessary to write down the matrix W. Its inverse is then the transpose W', and we have the 1-1 transformation

$$\begin{aligned} x &= Wc \\ c &= W'x, \end{aligned} \qquad (51)$$

where x and c denote the column vectors of the strings of the data $x' = x_1, \ldots, x_n$ and the coefficients $c' = c_1, \ldots, c_n$, respectively. Hence, the least squares coefficients indexed by any collection γ are exactly the corresponding coefficients of the transform c'. Because of orthonormality Parseval's equality $c'c = \sum_t c_t^2 = x'x = \sum_t x_t^2$ holds.

The criterion (49) for finding the best subset γ, including the number k of its elements, is then equivalent with

$$\min_{\gamma}\{(n-k)\ln\frac{\mathbf{c}'\mathbf{c}-\hat{S}_k}{n-k}+k\ln\frac{\hat{S}_k}{k}+\ln(k(n-k))\}, \tag{52}$$

where

$$\hat{S}_k=\hat{\mathbf{x}}'\hat{\mathbf{x}}=\hat{\mathbf{c}}'\hat{\mathbf{c}}. \tag{53}$$

This is equivalent with the universal sufficient statistics decomposition, and it provides a natural definition of noise as the part of the data that cannot be compressed with the selected normal model class, while the rest, defined by the optimal model, gives the desired information bearing signal. Moreover, a theorem was proved in [13] stating that the optimal index set consists either of k largest or k smallest coefficients in absolute value for some k.

In [3] the denoising problem was viewed as one of estimating the 'true' signal $\bar{x}_t$, to which independent normally distributed noise of 0-mean and variance τ is added. By rather intricate minmax arguments a threshold of the kind $\lambda=\sqrt{2\tau\ln n}$ was derived for the selection of the index set such that any coefficient with absolute value less than the threshold is set to zero. The trouble with this is the estimation of the noise variance τ, because any estimate involves circular reasoning: the noise gets defined by the threshold which is determined by the variance estimate of the noise. In the paper the issue was settled by estimation of the noise variance from the high frequency part of the data. This works well for data where the noise indeed is in the high frequencies but not for others. At any rate the method is certainly not 'data driven' as claimed.

In [8] the denoising criterion (52) was applied to a number of biological data sequences along with three other denoising algorithms, including that in [3]. The best in every case could be judged to be (52) as also was the case with synthetic data, where the added noise could be compared with its estimate.

Appendix

In coding we want to transmit or store sequences of elements of a finite set $A=\{a_1,\ldots,a_m\}$ in terms of binary sequences of 0 and 1. The set A is called the *alphabet* and its elements are called *symbols*, which can be of any kinds, often numerals. The sequences are called *messages*, or often just data when the symbols are numerals. A code, then, is a one-to-one function $C:A\rightarrow B^*$ taking each symbol in the alphabet into a finite binary string, called the *codeword.* It is extended to sequences $x=x_1,\ldots,x_n$

$$C:A^*\rightarrow B^*$$

by the operation of *concatenation*: $C(xx_t)=C(x)C(x_t)$, where xy denotes the string obtained when symbol y is appended to the end of the string x. We want the extended code, also written as C, to be not only invertible but also such that the codewords $C(a_i)$ can be separated and recognized in the code string $C(x)$ without a comma. This implies an important restriction on the codes, the so-called prefix property, which states that *no codeword is a prefix of another.* This requirement, making the code a *prefix code*, implies the important Kraft inequality

$$\sum_{a\in A} 2^{-n_a} \le 1, \tag{54}$$

where $n_a = |C(a)|$ denotes the length of the codeword $C(a)$. It is easily extended even to countable alphabets, where the sum becomes the limit of strictly increasing finite sums bounded from above by unity.

The codeword lengths of a prefix code define a probability distribution by $P(a) = 2^{-n_a}/K$, where K denotes the sum on the left hand side of the Kraft inequality. Even the converse is true in essence: Given a set of natural numbers such that the Kraft inequality holds a prefix code can be defined such that the codeword lengths coincide with the given natural numbers. The code is not unique. If we have a probability mass function $P(a)$ defined on A, the natural numbers $\lceil \log 1/P(a) \rceil$ clearly satisfy the Kraft inequality and hence define a prefix code. If the alphabet is large, as it is if we take it as the set of binary strings of length n, say $n > 10$, we may ignore the requirement that the codeword lengths are integers, and we regard $\log 1/P(a)$ as an *ideal* code length. This means that we may equate prefix codes with probability mass functions. Suppose further that we have a density function $f(x^n)$ defined on strings of real numbers. If we write each number x_i to a precision δ, say $\bar{x}_i$, then we obtain probabilities $P(\bar{x}_i)$ given roughly as $f(\bar{x}_i)\delta$, and $\log 1/f(x^n)$ differs from an ideal code length $\log 1/P(\bar{x}^n)$ by the constant $n \log 1/\delta$, which does not affect anything we need to do with code lengths. Hence, we regard even $\log 1/f(x^n)$ as an ideal code length.

We conclude this appendix with the fundamental result of Shannon's: For any two probability mass functions q and g

$$\min_q E_g \log \frac{1}{q} \ge E_g \log 1/g = H(g),$$

the equality holding if and only if $q = g$. The function $H(g)$ is the Shannon entropy. This also holds for density functions with the minor and irrelevant change in the equality statement: $q = g$ almost everywhere. This has the important corollary that $D(g\|q) = E_g \log \frac{g}{q}$ may be taken as a distance measure between the two density functions, the Kullback-Leibler distance.

References

[1] Balasubramanian, V. (1996),'Statistical Inference, Occam's Razor and Statistical Mechanics on the Space of Probability Distributions', *Neural Computation*, **9**, No. 2, 349-268, 1997

[2] Grünwald, P. (2004), 'Tutorial on Minimum Description Length', Chapter in *Advances in Minimum Description Length: Theory and Applications, MIT Press*, (P. Grünwald, I.J. Myung and M.A. Pitt, eds).

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

[4] Hastie, T., Tibshiriani, R., and Friedman, J. (2001), The elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer Verlag, 536 pages

[5] Hansen, A.J. and B. Yu, (2001), 'Model Selection and the Principle of Minimum Description Length', *J. of the American Statistical Association* 96(454), 746-774

[6] Kolmogorov, A.N. (1965), 'Three Approaches to the Quantitative Definition of Information', Problems of Information Transmission **1**, 4-7

[7] Li, M. and Vitanyi, P.M.B. (1997), *An Introduction to Kolmogorov Complexity and Its Applications, Springer-Verlag, 1997,* second edition

[8] Ojanen J., Miettinen T., Heikkonen J., Rissanen J. (2004), 'Robust denoising of electrophoresis and mass spectrometry signals with Minimum Description Length principle', FEBS Letters, Vol. 570, No. 1-3, pp 107-113 (invited paper)

[9] Rissanen, J. (1978), 'Modeling By Shortest Data Description', *Automatica*, Vol. **14**, pp 465-471

[10] Rissanen, J. (1983), 'A Universal Prior for Integers and Estimation by Minimum Description Length', *Annals of Statistics*, Vol **11**, No. 2, 416-431

[11] Rissanen, J. (1986), 'Stochastic Complexity and Modeling', *Annals of Statistics*, Vol **14**, 1080-1100

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

[13] Rissanen, J. (2000), 'MDL Denoising', *IEEE Trans. on Information Theory*, Vol. **IT-46**, Nr. 7, November 2000.

[14] Rissanen, J. and Tabus, I (2004), 'Kolmogorov Structure Function in MDL Theory and Lossy Data Compression', Chapter in *Advances in Minimum Description Length: Theory and Applications, MIT Press*, (P. Grünwald, I.J. Myung and M.A. Pitt, eds).

[15] Shtarkov, Yu. M. (1987), ‘Universal Sequential Coding of Single Messages’, Translated from Problems of Information Transmission, Vol. 23, No. 3, 3-17, July-September 1987.

[16] Solomonoff, R.J. (1964), ‘A Formal Theory of Inductive Inference’, Part I, Information and Control **7**, 1-22; Part II, Information and Control **7**, 224-254.

[17] Vereshchagin, N. and Vitanyi, P. (2002), ‘Kolmogorov’s Structure Functions with an Application to the Foundation of Model Selection’, *Proc. 47’th IEEE Symp. Found. Compput. Sci*, pages 751-760, 2002