Stochastic Complexity
Author(s): Jorma Rissanen
Source: *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 49, No. 3 (1987), pp. 223-239
Published by: Wiley for the Royal Statistical Society
Stable URL: http://www.jstor.org/stable/2985991
Accessed: 21-10-2017 06:13 UTC

J. R. Statist. Soc. B (1987)
49, *No.* 3, *pp.* 223–239 and 252–265

Stochastic Complexity

By JORMA RISSANEN

IBM Research

[*Read before the* Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday, March 4th, 1987,* Professor A. F. M. Smith *in the Chair*]

SUMMARY

It is argued that all the useful information in observed data that can be extracted with a selected class of modeled distributions, will be obtained if we calculate the stochastic complexity, defined to be the shortest description length of the data. The same quantity also determines the greatest lower bound for prediction errors when the data are sequentially predicted. An abstract definition of stochastic complexity is given along with two fundamental theorems which justify the notion. Further, three explicit model selection criteria to approximate the stochastic complexity are described and the associated optimal models are interpreted to define asymptotically sufficient statistics for the data.

Keywords: MINIMUM CODE LENGTH; MODEL SELECTION CRITERIA; PRIOR KNOWLEDGE

1. INTRODUCTION

"What do the data have to tell?" is an often repeated phrase in statistical discourse, which rather aptly defines its central problem. While there should be little disagreement about the worthiness of trying to learn the objective information in observations, serious difficulties arise already in attempts to define precisely the nature of the purported message. To this we may add the difficulty of avoiding the advertently or inadvertently introduced subjective biases and other arbitrary choices which tend to distort the content. In this paper, we discuss a new approach to these problems based on the recently introduced concept of *stochastic complexity* to measure the amount of uncertainty in the data. We offer this, together with associated computation procedures, as a rational basis for model selection and statistical inference in general.

The basic concepts in this theory are 1) a parametric class of probabilistic models, in which the number of parameters may range over all natural numbers, and 2) a distinguished "utility function", whose minimized value, broadly speaking, is the stochastic complexity. As in Bayesian theory the class of models is not intended to include any "true" distribution for the data, but rather it is only regarded as a language in which the properties of the data are to be expressed. This is a minimum requirement for any kind of learning, for how can we find regular features in the data unless we can describe them! Unlike in the Bayesian theory, however, we do not admit just any kinds of utility functions, but rather we adopt one and one function only; namely, the number of binary digits with which the observations can be described by taking advantage of the selected class of models.

One may wonder why this particular loss function is chosen rather than something else. We give three independent justifications. First, the main interpretation of the stochastic complexity as the greatest lower bound for the description length may be taken as a formal measure of the amount of randomness in the data, defined relative to the selected class of models. Importantly, this randomness stems from both the sampling uncertainty, which occupies such a central place in traditional statistics, and the uncertainty due to the distribution itself. Hence, if we could determine a model with which the stochastic complexity is reached, we would have learned all the useful information in the data that on the whole can be extracted with the

† *Address for correspondence*: IBM Research, K54/802, San Jose, CA95120, USA

0035–9246/87/49223 $2.00

chosen class, and the only way anyone could teach us more is to offer a better model class. For this reason such a model may be regarded to represent the sufficient statistic for the data in the spirit of the related algorithmic notion due to Kolmogorov, see Cover (1985).

Secondly, if we let $I(x)$ denote the infinum of the code lengths for the data x, relative to a class of models, then $P(x) = 2^{-I(x)}$ defines a probability, which is not only a proper distribution but also gives the largest possible value of the "global" or unconditional likelihood as a representation of the intuitive idea of the "most likely explanation" of the data that can be obtained with the same class. We say "global" to distinguish it from the likelihoods $P(x \mid \theta)$, which provide probabilities conditioned on the parameters, and for which the maximized likelihood $P(x \mid \hat{\theta}(x))$ no longer defines a proper distribution. Because simply by conditioning on more parameters the maximized probabilities in general increase, the number of parameters itself, of course, cannot be found by the otherwise so useful traditional maximum likelihood technique, which we regard as its main flaw. Conversely, any procedure that achieves a maximized probability $P(x)$ for the data will necessarily produce the infinum of the code length—log $P(x)$. Hence, no other loss function admits such a "total maximum likelihood" interpretation.

We get a third interpretation for the stochastic complexity whenever it is desirable to measure models' performance by prediction errors. Indeed, if we instead of encoding the observed data sequence consider predicting it, and the prediction error is measured by some error function, then the accumulated prediction errors can be interpreted to define a code length for the data, and its minimized value gives the stochastic complexity. In this, the predictions involved must be "honest", which simply means that the prediction of each data item x_{i+1} must be a function of the previously predicted data points only. By contrast, "predictions" made, for example, in the various cross-validation procedures, Stone (1974), Lachenbruch (1975), and Geisser and Eddy (1979), are not "honest" in this sense, and, hence, the associated accumulated errors do not define a proper density function. The same is true about many other well-known model-selection criteria such as the AIC (Akaike, 1974), where the objective is to estimate either a mean prediction error or the Kullback distance, both of which involve the expectation operation relative to an imagined and non existing "true" distribution. (Still, we are indebted to Akaike's pioneering and innovative work for inspiration in our own efforts.)

The distinct advantage of our approach to statistical inference is that, unlike the usual procedures, both Bayesian and non-Bayesian alike, it is virtually free of arbitrary choices to be made from case to case. To be sure, the selection of the model class still has to be made, but, we maintain, no successful statistical theory of inference is possible without such a choice. Moreover, in principle at least, any choice of the model class should be regarded only as a provisional one to be held until another with a smaller stochastic complexity is found. Regarding the ultimate model, no algorithmic procedure to find it can exist, as shown in the theory of the algorithmic complexity, Solomonoff (1964), Kolmogorov (1965), Chaitin (1973), which also is the spiritual father of our main notion. Hence, as it seems to us, statistical inquiry is fundamentally a search of a steadily improving sequence of models, not unlike the theories in physics.

Whether or not subjective choices do have a place in statistical inquiry, a legitimate concern about our approach is whether it is general enough to be useful. After all, one may suspect that the adopted unique utility function may simply not be flexible enough to satisfy the various demands in applications, especially since it seems to deprive the statistician of the ability to influence the model search towards some desired goal. To answer this we note that the choice of the model class leaves us with a considerable power to influence the search, since many of the usual utility functions may be incorporated into such models. Among these are the various prediction error measures, each of which gives rise to a specific density function. The most common of them, the quadratic measure for example, is represented by a gaussian family of densities. In conclusion, although our view implies a rather different interpretation of the purpose and meaning of some statistical problems (such as hypothesis testing) we cannot think of any application that could not be dealt with in such a manner.

In this paper we describe the basic ideas in the stochastic complexity approach to statistical reasoning. As already mentioned, we regard the algorithmic notion of complexity as the origin of ours, although substantial changes are required to get workable results. The two main difficulties are to define the complexity in such a manner that it corresponds to the shortest code length and that it can be computed, at least approximately. The code length as a criterion for model selection has been proposed independently by several people. Apparently the earliest paper is due to Wallace and Boulton (1968), where a two-part code length, resulting from a particular coding process, was used as a criterion for classification problems without any attempt, however, to show that the resulting code length is even approximately the shortest among all available codings. Recently, Rissanen (1986a), we proved certain code lengths to be asymptotically shortest, and we took them to define stochastic complexities of their respective types. These are now seen to be computed approximations of the new conceptually more satisfying notion defined below, which, in turn, was inspired by the Bayesian studies, Aitchison (1975), and above all the "prequential" probability in Dawid (1984), to which stochastic complexity gives a new wider meaning. The current definition has the further advantage that it can be shown to reach the asymptotic lower bound for the mean code length under much weaker conditions than before, which is done in Secton 4. As a demonstration of the power of the new definition, we derive a closed formula for the complexity in the linear selection-of-variables problem with gaussian models, of which an approximation gives a new and simple model selection criterion. As to the analyzable properties of the *MDL* estimates, as we call the estimators associated with the search for the stochastic complexity, the works of Hannan (1980), Hannan and Rissanen (1982), and Rissanen (1986b), imply that the estimates of the number of parameters are consistent, while the usual properties of the maximum likelihood estimates of the real valued parameters are retained.

2. STOCHASTIC COMPLEXITY

Our intent is to define the stochastic complexity such that it should represent the very shortest code length for any set of data that can be obtained when the codes are somehow designed with help of a selected class of probabilistic models. In this we really mean the code length of the particular set of observations at hand, rather than a mean length relative to some distribution, perhaps constructed out of the model class. The reason for this is that the model class itself is not unique, which means that we can get as small a mean length as we like by just picking the class so that it assigns a high probability to a simple sequence, say the one consisting of all 0s. As a result, we would not get the universal yardstick for models' performance which is central in our approach. The problem that we face in defining our complexity, however, is to make precise the way the model class is supposed to be used, which places us in a quandary. On one hand, if we permit too much, we get back the algorithmic complexity, which we do not want. On the other hand, we do not wish to spell out too narrowly the way the models are supposed to be used, because then we would defeat the purpose with stochastic complexity. From a practical stand-point, the task is clear enough: use the models any way you wish to obtain as short a code length as possible, which leaves the general philosophy intact without producing a workable criterion, however. Hence, for example, it is perfectly admissible to consider the model which assigns all the probability to the given sequence, but since we must encode this distribution itself nothing is gained. A way out of this predicament is to compromise. We shall propose an explicitly defined candidate for the stochastic complexity, which at least for long strings has the desired property. Although for short strings the possibility remains open that a shorter code length might result from some coding, the same formula still provides a target, which appears to be hard to beat, except for very short strings, of course, for which little useful statistics can be done anyway. In some cases, such as in Example 1 below, the given complexity will be near optimal for most strings even when the length is only moderate, as can be shown by enumerative arguments.

Let $\{f(x \mid k, \theta) \mid \theta = (\theta_1, \ldots, \theta_k), k = 0, 1, \ldots\}$ denote a parametric class of distributions, represented by densities, such that for each $f(\cdot \mid k, \theta)$ the compatibility (marginality) conditions required for a random process are satisfied. Here, $x = x_1, \ldots, x_n$ stands for a finite string of observed data, which we often write also as x^n to indicate its length. Further, for each k let $\pi(\theta \mid k)$ be a strictly positive distribution in the k-dimensional parameter space. These "prior" distributions need not be interpreted in any particular manner as representing definite prior knowledge about the parameters. Rather, in this approach, prior knowledge can be applied to the selection of the model class, which includes both of the two types of densities, and its role is suggestive, only. The final choice of the models is done in the light of the data in order to achieve the shortest total code length and hence to extract the maximum amount of information from the observed sample.

We define the *stochastic complexity* of the data x of length n, relative to a class of distributions, to be

$$I(x) = -\log \sum_{k=0}^{R} Q(k) \int f(x \mid k, \theta) d\pi(\theta \mid k), \tag{2.1}$$

where $Q(k) = 1/(R+1)$, and R is the range, not greater than n, of the number of parameters. Indeed, it seems pointless to fit more parameters than the number of data points. The integration is over the k-dimensional space of the parameters. In the degenerate case where there are no free parameters this definition reduces to the Shannon information $I(x) = -\log f(x)$, or complexity as we would like to call it.

The stochastic complexity is defined only relative to a class of models consisting of $f(x \mid k, \theta)$ and $\pi(\theta \mid k)$. By Kolmogorov's theorem (Kolmogorov, 1965) there cannot be an algorithmic procedure for the selection of an optimal model class which would guarantee that the very shortest code length can be found. Therefore, the class will have to be selected by us with use of intuition and prior knowledge. However, a choice need by no means be final, and, in principle at least, we can try to find better classes with smaller values for the stochastic complexity. So long as we have only a handful of candidates, we need not consider the code length required to describe the model class or, equivalently, a prior for it. If, however, the model classes themselves are indexed, for example, by the natural numbers or by the reals so that their specification requires a substantial amount of bits, then, of course, the description of the class itself must be included in the total description. In particular, the "priors" $\pi(\theta \mid k)$ ought to be so found that the complexity gets minimized. We can see in (2.1) that, generally speaking, a good prior ought to have the bulk of its probability at or near the maximum likelihood estimate. Hence, we might say, the purpose of prior knowledge is to specify this estimate. We have several possibilities. For instance, we can do the specification before the data, which means that we only take advantage of the broad constraints about the parameters, such as their range, or we wait until the maximum likelihood estimate is calculated and construct a prior accordingly. In this latter case, however, we must specify the estimate itself, which increases the total code length and the complexity. Hence, a good prior ought to be both easy to describe and it ought to place a considerable amount of the probability mass at or near the maximum likelihood estimate. Sometimes, a useful class of priors can be found of the form $\pi(\theta \mid k, \alpha)$, where α is a parameter about which only its range is specified. Then, we define

$$\pi(\theta \mid k) = \int \pi(\theta \mid k, \alpha) d\pi(\alpha), \tag{2.2}$$

where $\pi(\alpha)$, in turn, is chosen to represent the prior knowledge resulting from the range information. We discuss this further below.

All this represents a conceptual deviation from the way prior knowledge is dealt with traditionally, although there appears to be a certain similarity with the idea of the so-called reference prior, Bernardo (1979). This is the prior that maximizes Shannon's mutual information between the parameters and the data and, hence, reaches the "channel capacity", to use coding theoretic terms. Although this prior need not coincide with our optimizing prior, which, to be sure, is a rather diffuse notion unless the set of allowable models is defined precisely, it nevertheless does provide a candidate worth checking: see Example 2 below.

It is reasonable to expect that if there is no particular prior knowledge about the parameters, an attempt to optimize the prior will produce only a marginal reduction in the code length, and we might just as well try to pick it to reflect the lack of prior knowledge. If the range of the parameters is bounded, the familiar uniform distribution is not only an intuitively reasonable choice but it can also be justified on formal grounds, such as having maximum entropy, Jaynes (1968), as well as representing a natural minimax mean code length, Rissanen (1983). Somewhat surprisingly, perhaps, the same minimax approach works even when the range is unbounded. As discussed in this reference, encoding a natural number k by a so-called prefix code requires about $L^*(k) = \log^*k + \log C$ binary digits, where $\log^*k = \log k + \log\log k + \ldots$, the sum including all the positive iterates of the base-two logarithm, and C is the constant, about 2.865, that makes $\Sigma_{n=1}^{\infty} 2^{-L^*(n)} = 1$. (All logarithms in this paper are either to the base two, written "log", or to the natural base, written "ln".) The code length $L^*(k)$ has an asymptotic optimality property such that the ratio of its mean to the entropy, both relative to any distribution in a large class, approaches unity, which means that even if we were given a distribution for the integers to design a code with, we would not do asymptotically better than we do with the "universal" code of length $L^*(k)$. In fact, it is very difficult to describe even moderately large integers with fewer bits, which the reader may wish to try to do.

In equivalent probabilistic terms, the properties of $L^*(k)$ mean that it is hard to find a distribution where the tail would decrease less fast than

$$Q^*(k) = \frac{1}{C}\, 2^{-\log^*(k)}, \tag{2.3}$$

and we take this to represent a universal prior for the positive integers, reflecting the state of affairs that nothing special is known about them on prior grounds. At any rate, for our purpose where the aim is to find the shortest code length, no distribution exists which would improve the end product by a significant amount. The distribution (2.3) can be extended to all real numbers, Rissanen (1986a), if we define

$$q^*(y) = \frac{1}{2C}\, 2 - \log^*(\bar{y}), \tag{2.4}$$

where $\bar{y}$ is the smallest positive integer greater than or equal to $|y|$. This density, when multiplied by the precision, say $\delta = 2^{-r}$, to which a real number is truncated, gives a probability whose negative logarithm represents its complexity. Hence, the complexity grows with the number of fractional digits in the real number regardless of its size, as it clearly should. Moreover, this again has the asymptotic optimality property that makes it virtually impossible to write down a number with fewer binary digits than given by (2.4), unless the number is very simple, which is precisely the property which we regard as the foremost requirement in any prior that can claim universality. Notice that (2.3) modifies Jeffreys' improper prior $1/k$ just enough to get a proper distribution. However, the same is not true about (2.4), which differs substantially from $1/|y|$, especially for $|y| < 1$. We have used coding theory to express in a formal way the usual intuitive idea of no prior information. It seems to us that the several known attempts to do the same are not quite satisfactory, if only for the reason that they often end up in an improper prior. The universal prior (2.4) has the drawback that the integral required in (2.1) cannot be calculated in a closed form. However, in practice we wish to approximate the complexity anyway, and there are general ways to do it as discussed in the next section.

The complexity $I(x)$ is seen to be the negative logarithm of a density $f(x)$, which relates it to a code length. In fact, it is a standard result in coding that whenever we have a probability function $P(x)$ for discrete data, then a code can be constructed with length as the least integral upper bound to $-\log P(x)$. Hence, if δ is the precision to which the data items are truncated, $f(x_t)\delta$ is a probability, and the stochastic complexity, calculated per observation, differs from a code length only by about $-\log \delta$, which is independent of the data as well as of the model class involved. For this reason we call $I(x)$ itself a code length. As to its minimality, it is clear already from Shannon's complexity that it cannot be the shortest for every sample x. Rather, it will have to be in a suitably defined mean sense. We give theorems in Section 4 which spell this out exactly. In essence, it is true that unless the sample is very short or very special, it will be difficult to design a code using the given class of models such that its length would be below $I(x)$, which gives substance to our general principle to search for models with small complexity or, equivalently, distributions that assign a large probability to the data.

Example 1. Let x be a string in the symbols $1, \ldots, d$, the number d regarded as fixed and known, and let the class of models be defined by the symbol probabilities $\theta = (p_1, \ldots, p_{d-1})$, extended to strings by independence. The priors are given by the uniform density function in the $d-1$ free parameters p_i. The integral in (2.1) turns out to be given in terms of the multinomial and binomial as

$$P(x) = \binom{n}{n_i}^{-1} \binom{n+d-1}{n}^{-1},$$

where n_i denotes the number of times i occurs in x, and $n = \Sigma_i n_i$. The complexity (2.1) is given by $I(x) = -\log P(x)$, which for large n is, with Stirling's formula, well approximated by

$$I(x) \simeq nH(\{n_i/n\}) + \frac{1}{2} \log \frac{n^{2d-1}}{\prod_{i=1}^{d} n_i},$$

where, $H(\{n_i/n\}) = \Sigma_i (n_i/n) \log (n/n_i)$. These computations extend in a straightforward manner to give the stochastic complexity of a string relative to the class of all Markov models, Rissanen (1986c).

Example 2. We consider a problem of Jeffreys, discussed in Bernardo (1985), where a sample of n individuals out of a population of size N is observed, and all were found to possess a certain property. Let t, $t = 0, 1, \ldots, n$, denote the variable number of individuals in a sample of size n that could have the said property for all we know about the population, and let the parameter θ denote the number of individuals in the total population having the same property. Consider a model class defined by

$$P(t \mid \theta, n, N) = \binom{\theta}{t} \binom{N-\theta}{n-t} \Big/ \binom{N}{n}$$

together with a prior $\pi(\theta)$, which we pick in two different ways. The first choice is the uniform distribution in the range determined after the sample is taken, namely, $\pi_1(\theta) = 1/(N-n+1)$, for all parameters in the range. The resulting stochastic complexity is

$$I_1(n) = -\log P(n \mid \theta, n, N) = -\log \sum_{i=n}^{N} \pi_1(i) \binom{i}{n} \Big/ \binom{N}{n} = \log \frac{(N-n+1)(n+1)}{N+1}.$$

For increasing N and fixed n this is seen to approach the value $\log(n+1)$, which makes us suspect that the chosen prior is not too good. Indeed, to encode a binary string of known length n, where all the symbols are the same, should require only about one bit, namely, the symbol itself, or none, if the symbol is known, too.

For the second choice of priors we consider $\pi_2(\theta \mid \alpha)$, defined as $\pi_2(N \mid \alpha) = \alpha$ and $\pi_2(i \mid \alpha) = (1-\alpha)/N$, for $i = 0, 1, \ldots, N-1$. Accordingly, the event that all the N individuals have the given property is singled out and assigned the probability α, while the remaining probability $1-\alpha$ is shared by the other N possibilities. In Bernardo (1985) the value $\alpha = 1/2$ was arrived at by the reference prior arguments, but here we get the same value from (2.2) by assigning the uniform distribution to the range of α. The stochastic complexity with this prior is

$$I_2(n) = \log \frac{2N(n+1)}{2N + (N-1)n} = 1 + \log \frac{n+1}{n+2-n/N}.$$

The second choice of the prior is clearly better by our general principle, and, in fact, can hardly be improved upon. In order to arrive at the same conclusion by more traditional arguments, one needs to compute the posterior probability of the hypothesis $\theta = N$, given that $t = n$, with the two priors. The result with the former is $(n+1)/(N+1)$ and $(n+1)/(n+2-n/N)$ with the latter. Because the latter approaches unity when both n and N grow in accordance with intuition, while the former approaches zero for instance with $n = \log N$, the latter prior is judged as more satisfactory. We conclude that while it may require intelligence and cleverness to spot a bad prior and propose a good one, no new principle nor intuition is needed in our approach to do the same.

We conclude this section with a second important class of models for which the stochastic complexity can be calculated in a closed form, namely, the class arising in linear regression with gaussian distributions and the familiar conjugate priors for the parameters, Cox and Hinkley (1974). The result, which we give without the rather lengthy derivation, turns out to generalize the predictive distributions in Aitchison and Dunsmore (1975) in quite nontrivial manner. In Section 3 we then derive a new model selection criterion as an approximation to this complexity.

Example 3. The model class is defined by $f(x_i \mid k, \theta, \tau)$, which is taken as normal with variance τ and mean $\mu_t = \Sigma_{i=1}^k \theta_i u_{it}$, where the u_{ij} denote the observed regressor variables defining the $k \times t$ matrices $U_{k,t} = \{u_{ij}\}$. This density is extended to strings of data by independence. For simplicity, we assume $C_{k,t} = U_{k,t} U_{k,t}^{\mathrm{T}}$, where $^{\mathrm{T}}$ denotes transposition, to be nonsingular for all k and $t > k$. For each k-dimensional parameter vector θ we pick the prior $\pi(\theta \mid k, c)$ as normal with mean zero and variance $c\tau I$, c being a positive parameter and I the identity matrix. For τ, again, we pick the prior as a gamma distribution for its inverse, or

$$f(\tau \mid a, b) = \frac{(a/2)^{\mathrm{b}}}{\Gamma(b)} \tau^{-b-1} e^{-\alpha/2\chi},$$

where the two parameters a and b are positive. We can now derive with some calculations a formula for the joint density $f(x^n, \theta, \tau \mid k, a, b, c)$ and the marginal density $f(x^n \mid k, a, b, c)$ for $n > 2$, which, in turn, gives the conditional densities $f(x_{t+1} \mid x^t, k, a, b, c)$, valid for $t > \max\{k, 2\}$. These can be extended to $b = 0$ and $c = \infty$ to represent the case with little prior knowledge, and with extensive calculations they can be written as

$$f(x_t \mid x^{t-1}, k, a) = \frac{\Gamma\left(\frac{t}{2}\right)}{\Gamma\left(\frac{t-1}{2}\right)(\pi S_{k,t-1})^{1/2}\left[1 + \frac{(x_t - \hat{x}_t)^2}{S_{k,t-1}}\right]^{t/2}}, \tag{2.5}$$

which turn out to define valid densities for $t > \max\{1, k\}$. In this formula $\hat{x}_t = \hat{\theta}^{\mathrm{T}}(t-1)\bar{u}_t = \Sigma_{i=1}^k \hat{\theta}_i(t-1)u_{i.t}$ denotes the predicted value of x_t based upon the least squares estimates, obtained from the data up to time $t-1$, where we wrote $\bar{u}_t$ as the vector of the components in the sum. Further

$$S_{k,t} = (a + R_{k,t})(1 + \rho_t)$$

$$R_{k,t} = \sum_{i=1}^{t} x_i^2 - \hat{\theta}^{\mathrm{T}}(t) C_{k,t} \hat{\theta}(t) \tag{2.6}$$

$$\rho_t = \bar{u}'_{t+1} C_{k,t}^{-1} \bar{u}_{t+1}.$$

We take the so-far undefined density, corresponding to $k=0$, as follows

$$f(x_1 \mid a) = \frac{\Gamma(1)}{\pi \sqrt{a} \left(1 + \frac{x_1^2}{a}\right)}. \tag{2.7}$$

When the conditional densities $f(x_t \mid x^{t-1}, \kappa(k, t-1), a)$, where $\kappa(k, t) = \min\{t, k\}$, are multiplied we get a new density $f(x \mid k, a)$ and the associated stochastic complexity

$$I(x \mid a) = \log n - \log \sum_{k=0}^{n-1} f(x \mid k, a). \tag{2.8}$$

In this we cannot set the parameter $a=0$. In principle, we should integrate the product $f(x \mid k, a) q^*(a)$ over the positive reals, but since we cannot get a closed form solution we determine the value $\hat{a}(x)$ such that it minimizes (2.8). This causes the complication that $2^{-I(x \mid \hat{a}(x))}$ no longer represents a proper density. The problem is overcome by setting

$$I(x) = \log n - \log \sum_{k=0}^{n-1} f(x \mid k, \hat{a}(x)) - \log q^*(\hat{a}(x)). \tag{2.9}$$

3. THREE MODEL SELECTION CRITERIA

We shall consider various conceptual coding procedures for the data, each of which gives a code length as an approximation to the stochastic complexity and which acts as a model selection criterion. Such approximations are needed, because the complexity (2.1) is an abstract quantity involving no specific model, while in practice it is often the model that we want, rather than the complexity itself. An exception, however, is hypothesis testing, where the optimal models are not required. Each approximation turns out to be an upper bound, at least for long strings, as might be expected in the light of the intended interpretation of the complexity. It is, of course, not really necessary to consider coding procedures to get the approximations, but they suffice, and they have the advantage that the parameters can be treated on the same level as the data, namely, as objects that have to be encoded. Moreover, we only need to make sure that the (imagined) coding is done in a decodable manner, which automatically will avoid the trap of ending up with a criterion without a natural data dependent interpretation. In this regard, all criteria that seek to approximate an expected quantity, such as the prediction error or the Kullback distance, lack a credible interpretation, because there cannot be any unique "inherent" parent distribution behind any set of data relative to which the expectation could be defined. Similarly, any cost function in the Bayesian theory that does not admit a code length interpretation, represents a subjective and arbitrary choice and should be accepted with suspicion, quite regardless of how well it might work in a specific application.

We begin with the first basic type of upper bound, which can be interpreted as the code length when the coding is done by a special predictive procedure. Such a code length was in Rissanen (1986a) defined to be a predictive version of stochastic complexity. The result is also related to the prequential probability in Dawid (1984). With the notation

$$f(x\,|\,k)=\int f(x\,|\,k,\,\theta)d\pi(\theta\,|\,k) \tag{3.1}$$

we get an upper bound for the stochastic complexity by including only the largest term in the sum (2.1). Hence,

$$I(x)\leqslant \min_k\,\{-\log\, f(x\,|\,k)+\log\,(n+1)\Big\}, \tag{3.2}$$

Consider the conditional density for each observation x_{i+1} given the past,

$$f(x_{t+1}\,|\,x^t,\,k)=\frac{f(x^{t+1}\,|\,k)}{f(x^t\,|\,k)}.$$

Then the inequality (3.2) can also be written as

$$I(x)\leqslant \min_k \left\{-\sum_{t=0}^{n-1}\log\, f(x_{t+1}\,|\,x^t,\,k)+\log(n+1)\right\}. \tag{3.3}$$

By Shannon's coding theorem, $-\log f(x_{t+1}\,|\,x^t,\,k)-\log\,\delta$ represents the shortest code length (in the mean, rather than literally) with which x_{t+1}, written to some precision δ, can be encoded using the given distribution. Hence, the right hand side of (3.3) represents to within a constant term the code length for the entire sequence, resulting from such a predictive coding procedure. Let the minimizing *PMDL* (Predictive Minimum Description Length) estimate be $\hat{k}$. With this estimate we can further compute the maximum likelihood estimate $\hat{\theta}=\hat{\theta}(x)$ having $\hat{k}$ components to give a complete *PMDL* model.

The criterion in the right hand side of (3.3) can be further approximated as follows:

$$\min_k \left\{-\sum_{t=0}^{n}\log\, f(x_{t+1}\,|\,x^t,\,k,\,\hat{\theta}(t))\right\}, \tag{3.4}$$

where $\hat{\theta}(x^t)=\hat{\theta}(t)$ is the estimate with k components which would have minimized the past code length $-\log f(xt\,|\,k,\,\theta)$ if we had had the foresight to use it. But because, obviously, this could not be done until x_t is seen, we somewhat belatedly apply it to encode the observation x_{t+1}. We are here applying a sound principle of inductive inference, namely, always to encode the "next" value with a model that would have worked best in the past, in the hope that the future is like the past. The initial density function $f(x_1\,|\,k,\,\hat{\theta}(0))$ will have to be picked somehow the reflect the prior knowledge. But notice that this can be done without estimating the initial parameter. For the early values of t a reasonable rule is not to estimate more parameters than what can be uniquely solved. This time (3.4) is not necessarily and upper bound for the stochastic complexity.

We see that this type of coding is just like the process of prediction; see also Dawid (1984). In fact, as discussed in Rissanen (1986a), prediction with any error criterion may be interpreted as coding with a suitably selected family of distribution, and the corresponding stochastic complexity represents the prediction errors that result when the data are predicted in the best possible manner, Rissanen (1984). An important special case arises from the selection of gaussian models with a fixed variance, for then the criterion (3.4) becomes a *predictive* least squares criterion, with strictly extends the applicability of the classical least squares criterion in allowing the estimation of both the number of the parameters and their values, see Section 5.2.

In the case where the data are modeled as independent, the predictive criteria have the disadvantage that they must impose an ordering for the data, which makes them order dependent, and, furthermore, the estimates $\hat{\theta}(t)$ for each k must be computed n times. In the important linear regression case we, however, obtain a far simpler criterion by substituting

$f(x \mid k, a)$ from Example 3 into (3.2). A further simplification results if we ignore the initial conditional densities in the product $f(x \mid k, 0)$ for $t \leqslant k$, which gives the criterion

$$\min_k \left\{ \frac{n}{2} \log R_{k,n} + \frac{1}{2} \log \mid C_{k,n} \mid \right\}, \tag{3.5}$$

provided, of course, that we do not have exceptional samples for which $R_{k,n} = 0$. *This criterion also extends to ARMA* estimation if we interpret $C_{k,n}$ as the matrix, defined by the double derivatives of the sum of the squared deviations, evaluated at the least squares estimates.

The preceding predictive coding process in (3.4) uses the maximum likelihood estimates $\hat{\theta}(t)$, which are determined from the past data only. But we can do coding even in a non-predictive way when the parameters are estimated from all the data, and this results in another upper bound for the stochastic complexity, which at the same time will provide a simple proof of the second main theorem in Section 4. We need to make two fairly reasonable assumptions. First, we assume that $f(x \mid k, \theta)$ has continuous double derivatives with respect to the parameters. Let $\lambda_{\max}(x^n)$ denote the largest eigenvalue of the positive definite matrix

$$\frac{-1}{n} \left\{ \frac{\partial^2 \log f(x^n \mid \theta)}{\partial \theta_j \partial \theta_j} \right\}, \tag{3.6}$$

evaluated at the maximum likelihood estimate $\hat{\theta} = \hat{\theta}(x^n)$. We assume that there is a positive constant α such that

$$\lambda_{\max}(x^n) < \alpha \tag{3.7}$$

for all samples. This condition is certainly satisfied in the important gaussian family.

By expanding $f(x \mid k, \theta)$ in Taylor's series in a fixed neighbourhood S of $\hat{\theta}$, we get the inequality

$$f(x \mid k, \theta) \geqslant f(x \mid k, \hat{\theta}) \left[1 - \frac{\beta n}{2} \sum_{i=1}^{k} (\theta_i - \hat{\theta}_i)^2 \right],$$

where β is a positive constant not dependent on x. Such a constant exists because of (3.7) and the assumption that the second partials are continuous, which imply that the maximum eigenvalue of the matrix (3.6) for θ in S is close to $\lambda_{\max}(x^n)$, uniformly in x^n. We then deduce the following inequalities, the first being obvious, and the second follows by carrying out the integration of the Taylor expansion and taking $\pi(\theta \mid k)$ as the k-fold product of the universal density (2.2),

$$\int f(x \mid k, \theta) d\pi(\theta \mid k) \geqslant \int_{\Delta} f(x \mid k, \theta) d\pi(\theta \mid k) \geqslant 2^k \delta^k f(x \mid k, \hat{\theta}) \prod_{i=1}^{k} q^*(\bar{\theta}_i)(1 - (\beta k n/6)\delta^2),$$

where Δ denotes a k-dimensional rectangular neighbourhood of $\hat{\theta}$ with edge length δ, picked small enough so that Δ remains inside of S. Further, $\bar{\theta}$ is the point where the product of the $q^*(\theta_i)$ is at minimum in S. We now find the value of δ which maximizes the right-most expression so as to give the best possible lower bound. The result is

$$\int f(x \mid k, \theta) d\pi(\theta \mid k) \geqslant 2 \left(\frac{24}{\beta} \right)^{k/2} (k+2)^{-(1+k/2)} \times f(x \mid k, \hat{\theta}) \prod_{i=1}^{k} q^*(\bar{\theta}_i) n^{-/2}. \tag{3.8}$$

The optimal edge length is proportional to $1/\sqrt{\mathrm{n}}$, so that Δ stays inside S for large enough n, and our arguments remain valid. From (3.8) we obtain the desired upper bound approximation for the stochastic complexity for all sufficiently large n,

$$I(x) \leqslant \min_{k,\theta} \left\{ -\log f(x \mid k, \theta) + \frac{k}{2} \log n + \left(\frac{k}{2} + 1 \right) \log(k+2) + O(\log \log n) \right\}, \tag{3.9}$$

provided that the maximum likelihood estimates $\hat{\theta}$ remain uniformly bounded and hence also the term $\Sigma_{i=1}^{k} \log |\hat{\theta}_i|$, which otherwise is to be added to (3.9). We recognize (3.9) as essentially the model selection criterion derived in Schwarz (1978) for special Bayesian priors and in Rissanen (1978), (1983), where it was derived as the length of a non-predictive code. Indeed, we may regard $-\log f(x \mid k, \hat{\theta})$ as the ideal code length for the data, given the estimate $\hat{\theta} = \hat{\theta}(x)$. Because this depends on all the data and hence cannot be computed by the decoder at the time it is needed, its value must be given as a preamble in the code string. When each component is truncated to its optimal precision and encoded with $q^*(.)$, the resulting code length is $(k/2) \log n + O(\log \log n)$. The present derivation of the criterion is far simpler, however, and it establishes an exact upper bound for the complexity for all n larger than some number. For later use we call the pair of minimizing parameters $(\hat{k}, \hat{\theta})$ the *MDL* (Minimum Description Length) estimates. We note in passing that (3.5) may be regarded as a refinement of (3.9) in the quadratic case and one, which does not need an asymptotic justification. Indeed, frequently the elements of $C_{k,n}$ for large values of n are approximately proportional to n, so that $|C_{k,n}|$ behaves like n^k, and the second term of (3.5) is like the second term of (3.9).

4. MAIN THEOREMS

We reproduce from Rissanen (1986a) a theorem, which together with another theorem, which is new and which we prove, is fundamental in our entire theory of stochastic complexity.

Theorem 4.1. In the model class $\{f(x \mid k, \theta): k \geqslant 0\}$ let for each positive integer k the parameters θ range over a compact subset Ω^k with nonempty interior of the k-dimensional Euclidean space. Assume that there exist estimates $\hat{\theta}(x^n)$ such that the tail probabilities are uniformly 'summable as follows

$$P_\theta\{\sqrt{n} \,\| \hat{\theta}(x^n) - \theta \| \geqslant \log n\} \leqslant \delta(n), \text{ for all } \theta, \text{ and } \sum_n \delta(n) < \infty, \tag{4.1}$$

where $\| \theta \|$ denotes a norm. If g is any density defined on the observations, satisfying the compatibility conditions for a random process, then for all k and all $\theta \in \Omega^k$, except in a set of Lebesgue measure zero,

$$\liminf_n \frac{E_{k,\theta} \log [f(x^n \mid k, \theta)/g(x^n)]}{\frac{k}{2} \log n} \geqslant 1. \tag{4.2}$$

The mean is taken relative to the distribution defined by $f(x \mid k, \theta)$.

The particular density $g(x) = 2^{-I(x)}$ satisfies the stated compatibility condition, because each member $f(x \mid k, \theta)$ does it by assumption. Hence, the mean stochastic complexity, defined relative to a class satisfying the conditions in the theorem, also satisfies the inequality (4.2). The condition (4.1) is not directly seen to hold in general, nor do we know whether it is necessary. However, it does not seem to be overly restrictive, because it can be shown to hold in such important classes as the Markov chains, Rissanen (1986c), and, of course, in the gaussian families occurring in regression. It has also been verified for the *ARMA* processes in Gerencse'r (1985). If we keep n fixed and set $k = 0$, meaning that there are no free parameters, the statement of the theorem degenerates to Shannon's famous coding inequality.

The second theorem states that the stochastic complexity reaches the lower bound under only slightly more stringent conditions than in Theorem 4.1. In Rissanen (1986a) the independence assumption and more was needed to prove a similar theorem. We refer to a result of the same nature, which holds in almost sure sense, to Dawid (1984).

Theorem 4.2. Let in addition to the conditions in Theorem 4.1 $f(x \mid k, \theta)$ satisfy (3.7). Then for all k and all $\theta \in \Omega^k$, except in a set of Lebesgue measure zero,

$$E_{k,\theta} \frac{I(x^n) + \log f(x^n \mid k, \theta)}{\frac{k}{2} \log n} \to 1. \tag{4.3}$$

Proof. The proof follows from Theorem 4.1 and the inequality (3.9). Indeed,

$$\begin{aligned} E_{k,\theta} I(x) &\leqslant E_{k,\theta} \min_{m,\alpha} \left\{ -\log f(x \mid m, \alpha) + \frac{m}{2} \log n + \left(\frac{m}{2} + 1\right) \log(m+2) + O(\log \log n) \right\} \\ &\leqslant -E_{k,\theta} \log f(x \mid k, \theta) + \frac{k}{2} \log n + O(\log \log n), \end{aligned} \tag{4.4}$$

which with Theorem 4.1 implies the claim. In this formula α denotes a parameter vector with m components.

Remarks. The first application of these theorems is that they define precisely the sense in which the complexity serves as a lower bound for "regular" code lengths. Indeed, we define a code length $L(x)$ to be *regular*, if it satisfies an inequality $L(x) \geqslant -\log g(x)$, where $g(x)$ is some density such that the marginality conditions for a random process hold. Such a regularity condition is a natural sharpening of the well-known Kraft inequality, required for Shannon's coding inequality to hold, namely, $\Sigma_x 2^{-L(x)} \leqslant 1$, where x runs through all strings over the finite set of symbols with the same length. In fact, all codes known to us are regular. By (4.2) and (4.3) we deduce that for all positive numbers ε, all k and all $\theta \in \Omega^k$, except in a set of Lebesgue measure zero,

$$\frac{1}{n} E_{k,\theta}(L(x^n) - I(x^n)) \geqslant -\varepsilon k \frac{\log n}{n} \tag{4.5}$$

for all large enough n.

A second and the main application of these results is that they provide a rational basis for, what we regard as the most fundamental of all the problems in statistics, namely, the comparison of models and, hence, for the assessment of the goodness of the estimates of the number of parameters. In our opinion, only quite unsatisfactory attempts, such as estimating probabilities of "correct" estimates, to do the same have been made in the past. Indeed, given any estimator $(\tilde{k}, \tilde{\theta})$, we determine the regular code length $-\Sigma_t \log f(x_{t+1} \mid x^t, \tilde{k}, \tilde{\theta})$ and compare the mean with the entropy according to (4.2), which will settle the issue of how good the proposed estimators are. Unless the samples are very short we do not need to compute the mean, but rather we simply compare the length with the stochastic complexity, which is an advantage over the test for efficient estimators by Cramér-Rao inequality, where, besides, the number of parameters must remain fixed. Efficient estimators play no rôle in the present theory, for it is not the small variance in estimation errors that counts, which is meaningful only with the true distribution assumption and which also requires the artificial restriction of unbiasedness, but the goodness of the resulting model, for which we offer the code length as the universal yardstick. Although the covariances of the estimation errors do not appear in our theory, they still could be used to provide information about the curvature of the likelihood function. Whether or not such information can be regarded to represent the sampling uncertainty in a meaningful manner depends entirely on how well the optimal model represents the constraints in the machinery providing the data. To assess this in an absolute sense, however, can be very difficult, except in gambling and related problems where the machineries are carefully designed by humans. In some cases, such as in the selection-of-variables problem, discussed in Section 5.2, the preferred performance criterion is the prediction error. Such a variant, however, can be derived from Theorems 4.1 and 4.2, Rissanen (1984, 1986b).

Finally, the predictive approximation in (3.3), too, is seen to achieve the lower bound in Theorem 4.1. This is because by (3.1) and (3.8) the right hand side of (3.9) is an upper bound for the right hand side of (3.3).

To conclude this section we define the estimate $(\hat{k}, \hat{\theta})$ and the associated model to represent asymptotically the *minimal sufficient statistics* of the *data*, relative to the considered class of models, in the spirit of Kolmogorov's related algorithmic notion. Moreover, their complexity is given asymptotically by $\frac{1}{2}\hat{k}$ log n. The justification for this is that we may think of encoding the data relative to this model. Specifically, we may define the "noise" sequence $e_{t+1} = x_{t+1} - \hat{x}_{t+1}$ where $\hat{x}_{t+1}$ denotes that value of z for which $f(z \mid x^t) = f(x^t, z)/f(x^t)$ reaches its maximum, assuming a unique maximizing point for simplicity. The noise sequence is not necessarily an innovation sequence, but its code length, about $-\Sigma_t \log f(x_{t+1} \mid x^t)$, cannot be further reduced within the given model class. We may say that the noise sequence contains essentially no useful information about the string x. Instead, all the useful information is in the optimal model.

5. APPLICATIONS

The *MDL* and the *PMDL* principles together with the stochastic complexity they seek to compute, extend strictly the maximum likelihood principle to estimation of models with different numbers of parameters or different structures. Hence, the potential applications cover nearly all aspects of statistical problems. Most typical of them are the time series modeling problems, but also pattern recognition and classification problems involving the design of an optimal size decision tree can rather elegantly be solved by this approach without the usually needed fudge factors and arbitrary performance measures. Here, we discuss only two applications, hypothesis testing, which we do in a rudimentary way to illustrate the rather different thinking involved, and the selection-of-variables problem, where more definite analytic results have been obtained, Rissanen (1986b).

5.1. *Hypothesis Testing*

We view the purpose of hypothesis testing to be to find out which of two classes of models better fits or explains the observed data. Once the result is found, we regard it as a separate issue to decide whether the winning hypothesis is better to a sufficient degree to justify its adoption over the competing hypotheses for whatever practical purpose. Such a decision, which normally is embedded within the hypothesis testing procedure, depends on factors outside the observed data, and in our opinion it can be and should be made separately. Usually, the decision is reached with help of thresholds, which, however, really are introduced from necessity to overcome the inability to compare models with different numbers of parameters rather than by desire.

Let the null hypothesis H_0 be represented by a class of models with k_0 free parameters and the alternative hypothesis H_1 by a class with k_1 free parmeters. Then in the light of the observed data x we prefer the null hypothesis, if $I_0(x) \leqslant I_1(x)$, where the stochastic complexities are defined relative to the two model classes corresponding to the two competing hypotheses. Hence, we may sat that $T(x) = I_0(x) - I_1(x)$ acts as a universal test statistic, but unlike in ordinary tests the decision can be computed without the distribution of the test statistic nor arbitrarily set thresholds. Instead, the decision is reached by balancing the ability of the two hypotheses to render the observed sequence random with the length of the sufficient statistics in each case as the "cost".

To test the meaningfulness of this procedure, which at first sight may appear as too simple and, perhaps, unreliable, we consider two examples.

Example 4. Consider a binary string of data $x = x_1, \ldots, x_n$. The null hypothesis states that the probability of the occurrence of symbol 1 is $p = 1/2$, while the alternative composite hypothesis states the opposite, namely $p \neq 1/2$. We pick the class of Bernoulli models. By

Example 1 in Section 2,

$$T(x) = n - \log \frac{(n+1)!}{n_0!\,n_1!} = n(1 - H(n_1/n)) - \frac{1}{2} \log \frac{n_0 n_1}{n} + O(1/n). \tag{5.1}$$

Since the first term can only be non-negative, we see that in order for hypothesis H_1 to win, the ratio n_1/n must differ sufficiently from 1/2 to overcome the "cost" of one free parameter; i.e., the complexity of the sufficient statistic under H_1, the same under H_0 being zero. This translates to the error of the first kind, or the confidence level, which turns out to be a few percent for sample size up to a few hundred, after which it declines gradually to zero, as it should. This agrees well with the value normally selected by "sound judgment" except, perhaps, for long strings, for which "sound judgment" is more difficult to apply; for a different attempt to assign the size of a test in a universal, data-independent way, we refer to Martin-Löv (1974).

As another example we consider a 2-way contingency table from Cramér, (1961, Table 30.6.1). The same approach applies to all contingency tables, which we study in another paper.

Example 5. Consider the table of two rows and 12 columns, where the first row, 3743, 3550, 4017. 4173, 4117, 3944, 3964, 3797, 3712, 3512, 3392, 3761, consists of the number of boys born in Sweden in the consecutive months of the year 1935, and the second row, 3537, 3407, 3866, 3711, 3775, 3665, 3621, 3596, 3491, 3391, 3160, 3371, consists of the number of girls born in the same months. The objective is to test whether the percentage of boy births from all the births does not depend on the month of the year, which is taken as the null hypothesis. We denote the elements of the 2×12 – matrix by n_{ij}, $i = 1, 2, j = 1, \ldots, 12$.

Under the null hypothesis the table represents counts of a binary string in the boy and the girl births, and from Example 1 of Section 2 we have under the homogeneous model, or the null hypothesis,

$$I_0(x) = \ln \frac{(n+1)!}{n_1!\,n_2!}, \tag{5.2}$$

where n_i, $i = 1, 2$, denotes the sum of the counts in the two rows, or the boy and the girls births, respectively, and n is the total count. For the non-homogeneous model the table is taken to represent counts of 12 distinct binary strings, and

$$I_1(x) = \sum_{j=1}^{12} \ln \frac{(n_j + 1)!}{n_{1j}!\,n_{2j}!}, \tag{5.3}$$

where $n_{.j}$ denote the column counts. By Stirling's formula we get the double test statistic

$$2T(x) \simeq 2 \sum_{ij} n_{ij} \ln \frac{n_{ij}}{n_{i.} n_{.j}/n} + \ln \frac{n^3}{n_{1.} n_{2.}} - \sum_j \ln \frac{n_{.j}^3}{n_{1j} n_{2j}}. \tag{5.4}$$

The sum term is exactly the so-called Kullback G^2 measure. It has asymptotically a χ^2 distribution with 11 degrees of freedom. Therefore, for large values of n our test is like the ordinary test except that the arbitrarily selected confidence dependent threshold is replaced by the second and the the third terms in (5.4) expressing the difference between the complexities of the two sufficient statistics involved.

With the given numerical values, $I_0(x) = 61301$ and $I_1(x) = 61341$, which favors the homogeneity hypothesis. The sum in (5.4) has the value 21.6 while the second term is -111.04, giving $T(x)$ the value -44.7; this is reasonably close to the difference between (5.2) and (5.3), which is almost exactly -40. Hence, the probability for the test statistic to be positive (to reject the homogeneity hypothesis) is the probability by which a χ^2-distributed variable with 11 degrees of freedom exceeds 111.04, or negligible.

5.2. *The Selection-of-Variables Problem*

Consider the linear regression problem

$$y_t = \theta_1 x_{1,t} + \ldots + \theta_k x_{k,t} + \varepsilon_t,\ t = 1, \ldots, n, \tag{5.5}$$

where ε_t represents the amount by which the linear model fails to explain y_t. If we pick the family $f(y_{t+1} \mid k, \theta)$ as normal with mean $\hat{y}_{t+1} = \theta_1 x_{1,t+1} + \ldots + \theta_k x_{k,t+1}$ and variance 1/2, the criterion (3.4) turns into the following predictive least squares criterion

$$I(y \mid k) = \sum_{t=0}^{n-1} e_{t+1}^2, \tag{5.6}$$

to be minimized over k. Here, $e_{t+1} = y_{t+1} - \hat{y}_{t+1}$, where $\hat{y}_{t+1}$ is the predictor which results when θ is replaced by the k-component (ordinary) least least squares estimate $\hat{\theta}(t)$, determined from the past data up to time t. Alternatively, if we also let the variance parameter be free, as in Example 3 of section 2, the simpler to compute criterion (3.5) is more appropriate in cases where the prediction errors should be normalized by dividing them by the estimated variance. Both criteria may be minimized not only over the number of the regressor variables but also over collections of them, as often required.

In the basic case where the observations result from a gaussian process of type (5.5) with ε_t having independent normal distribution, $N(0, \sigma)$, and the conditional mean being a linear combination of the first m regressor variables, one can prove under weak regularity conditions that the *PMDL* estimates $\hat{k}(n)$ are consistent. Curiously enough, the consistency property would be lost if the squared terms were weighted so as to cause their variance to be the same, as was done for a different purpose in Brown, Durbin, and Evans (1975), who also considered what we call "honest" prediction errors. Further, under the same conditions one can also prove that the mean of the sum of the squared prediction errors (5.6) is asymptotically minimum among all predictors. Hence, for example, in this rather relevant sense, there is no better estimator for the number m than $\hat{k}(n)$. In particular, Akaike's AIC estimates would not be as good, for they would not reach the lower bound. Neither are the more traditional estimates such as in Helms (1974) nor the various cross-validation estimates. These results complement the optimality properties of the ordinary least suares estimates and lend support to the contention that the search for the stochastic complexity, indeed does provide a rational basis for model selection.

We illustrate the use of the criterion (3.5) with a numerical example, in which not only the number of the regressor variables is to be found but also the best subset of them. Such problems are often tackled by a combination of painstaking analysis of the variances and subjective judgment.

Example 6. The first five columns in Table 1 consist of 16 measurements of five properties of asphalt, marked X1, X2, X3, X4, X5, which affect in some more or less random way the wear and tear of roads made of that asphalt, quantitatively measured as the rut depth in the sixth column, marked Y. The data are taken from Figure 6B.1 of Daniel and Wood (1971). The objective is to predict the rut depth by a linear function of the five regressor variables and a constant term, and the problem is to find the best subset.

The best linear combination involves the constant term and the variables X1, X2, X3, and X5, and the value of the criterion (3.5) is 59.88. The second best combination is obtained with the constant term and all the five variables, which give the criterion the value 61.14, while the third best combination results by dropping the last variable X5 from the winning combination, which increases the criterion to the value 61.22.

In conclusion, one may wonder why a greater reliance should be placed on the optimal predictor found this way than on one resulting from the conventional analysis of variance, which involves subjective judgments. The answer is that we have tried all the combinations of the available regressor variables and found the one with which all the available data were

TABLE 1
Data of 16 *asphalt roads*

$X1$	$X2$	$X3$	$X4$	$X5$	Y
2.80	4.68	4.87	8.4	4.916	6.75
1.40	5.19	4.50	6.5	4.563	13.00
1.40	4.82	4.73	7.9	5.321	14.75
3.30	4.85	4.76	8.3	4.865	12.60
1.70	4.86	4.95	8.4	3.776	8.25
2.90	5.16	4.45	7.4	4.397	10.67
3.70	4.82	5.05	6.8	4.867	7.28
1.70	4.86	4.70	8.6	4.828	12.67
0.92	4.78	4.84	6.7	4.865	12.58
0.68	5.16	4.76	7.7	4.034	20.60
6.00	4.57	4.82	7.4	5.450	3.58
4.30	4.61	4.65	6.7	4.853	7.00
0.60	5.07	5.10	7.5	4.257	26.20
1.80	4.66	5.09	8.2	5.144	11.67
6.00	5.42	4.41	5.8	3.718	7.67
4.40	5.01	4.74	7.1	4.715	12.25

predicted best. Hence, we really are compelled to pick the so-found best predictor, unless we have external information or insist in acting irrationally.

REFERENCES

Aitchison, J. (1975) Goodness of prediction fit. *Biomerika*, **62**, 547–554.

Aitchison, J. and Dunsmore, I. R. (1975), *Statistical Prediction Analysis*. Cambridge: University Press.

Akaike, H. (1977) On Entropy Maximization Principle, in *Applications of Statistics* (P. R. Krishnaiah, ed.), pp. 27–41. Amsterdam: North-Holland.

Bernardo, J. M. (1979) Reference Posterior Distributions for Bayesian Inference. *J. R. Statist. Soc.* B, **41**, 113–147.

Bernardo, J. M. (1985) On a Famous Problem of Induction. *Trabajos de Estadistica y de Investigacion Operativa*, **36**, 14–30.

Brown, R. L., Durbin, J., Evans, J. M. (1975) Techniques for Testing the Constancy of Regression Relationships over Time. *J. R. Statist. Soc.* B, 149–163.

Chaitin, G. J. (1975) A Theory of Program Size Formally Identical to Information Theory. *J. ASS. Comp. Mach.*, **22**, 329–340.

Cover, T. (1985) Kolmogorov Complexity, Data Compression, and Inference. In *The Impact of Processing Techniques on Communications*, (J. K. Skwirzynski, ed.), pp. 23–33, Nijhoff.

Cox, D. R. and Hinkley, D. V. (1974) *Theoretical Statistics*. London: Chapman and Hall.

Cramér, H. (1961), *Mathematical Methods of Statistics*. Princeton, N.J.: Princeton University Press.

Daniel, C. and Wood, F. S. (1971) *Fitting Equations to Data*. New York: Wiley.

Dawid, A. P. (1984) Present Position and Potential Developments: Some Personal Views, Statistical Theory, The Prequential Approach. *J. R. Statist. Soc.* A, **147**, 278–292.

Geisser, S. and Eddy, W. (1979), "A Predictive Approach to Model Selection", *J. Amer. Statist. Ass.*, **74**, 153–160.

Gerencse'r, L. (1985) On the Normal Approximation of the Maximum Likelihood Estimator of ARMA Parameters (working paper IV/41, Computer and Automation Institute, Hungarian Academy of Sciences, November 22, 1985).

Hannan, E. J. (1980) The Estimation of the Order of an ARMA Process. *Ann. Statist.* **8**, 1071–1081.

Hannan, E. J. and Rissanen, J. (1982) Recursive Estimation of ARMA Order. *Biometrika*, **69**, 81–94.

Helms, R. W. (1974) The Average Estimated Variance Criterion for the Selection-of-Variables Problem in General Linear Regression. *Technometrics*, **16**, 261–273.

Jaynes, E. T. (1968) Prior Probabilities. *IEEE Trans. Systems, Science, and Cybernetics*, SCC-**4**, 227–291.

Kolmogorov, A. N. (1965) Three Approaches to the Quantitative Definition of Information, *Problems of Information Transmission*, **1**, 4–7. (English trans. in *Selected Translations Math. Stat.* and *Prob.*, **7**, 00–00, IMS and AMS, 1968.)

Lachenbruch, P. A. (1975) *Discriminant Analysis*. Haffner Press.

Martin-Löv, P. (1974) The Notion of Redundancy and Its Use as a Quantitative Measure of the Discrepancy between a Statistical Hypothesis and a set of Observational Data. *Scand. J. Statist.*, **1**, 3–18.

Rissanen, J. (1978) Modelling by shortest data description. *Automatica*, **14**, 465–471.

Rissanen, J. (1983) A Universal Prior for Integers and Estimation by Minimum Description Length. *Ann. Statist.*, **11**, 416–431.

Rissanen, J. (1984) Universal Coding, Information, Prediction, and Estimation. *IEEE Trans. Inf. Thy,* IT-**30**, 629–636.

Rissanen, J. (1985) Minimum Description Length Principle, in *Encyclopedia of Statistical Sciences,* **5**, (S. Kotz & N. L. Johnson, eds.), pp. 523–527. New York: Wiley.

Rissanen, J. (1986a) Stochastic Complexity and Modeling. *Ann. Statist.*, **14**, 1080–1100.

Rissanen, J. (1986b) A Predictive Least Squares Principle *J. of Math. Control and Information*, (special issue: Parametrization Problems) **3**, 211–222.

Rissanen, J. (1986c) Complexity of Strings in the Class of Markov Sources. *IEEE Inf. Thy*, **29**, IT-32, No. 4, 526–532.

Schwarz, G. (1978) Estimating the Dimension of a Model. *Ann. Statist.*, **6**, 461–464.

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.

Stone, M. (1974) Cross-Validatory Choice and Assessment of Statistical Predictions. *J. Roy. Statist. Soc.*, B, **36**, 111–147.

Wallace, C. S. and Boulton, D. M. (1968) An Information Measure for Classification. *Computer J.*, **11**, 185–194.