Stochastic Complexity and Modeling

Jorma Rissanen

Annals of Statistics, Volume 14, Issue 3 (Sep., 1986), 1080-1100.

Stable URL:
http://links.jstor.org/sici?sici=0090-5364%28198609%2914%3A3%3C1080%3ASCAM%3E2.0.CO%3B2-O

The Annals of Statistics
1986, Vol. 14, No. 3, 1080–1100

STOCHASTIC COMPLEXITY AND MODELING

BY JORMA RISSANEN

IBM Almaden Research Center

As a modification of the notion of algorithmic complexity, the stochastic complexity of a string of data, relative to a class of probabilistic models, is defined to be the fewest number of binary digits with which the data can be encoded by taking advantage of the selected models. The computation of the stochastic complexity produces a model, which may be taken to incorporate all the statistical information in the data that can be extracted with the chosen model class. This model, for example, allows for optimal prediction, and its parameters are optimized both in their values and their number. A fundamental theorem is proved which gives a lower bound for the code length and, therefore, for prediction errors as well. Finally, the notions of "prior information" and the "useful information" in the data are defined in a new way, and a related construct gives a universal test statistic for hypothesis testing.

1. Introduction. The purpose of statistical model fitting is to "understand" the observed data. If by understanding we mean the ability to remove redundancies in the data and hence to discover regular statistical features, then the ultimate measure of the success of such attempts must be the length with which the data can be described, say, in terms of binary digits. Indeed, if such a shortest description of the data, to be called *stochastic complexity*, is found in terms of the models of a selected class, there is nothing further anyone can teach us about the data; we know all there is to know. This is the rationale behind the MDL (minimum description length) criterion, which we, inspired by the algorithmic notion of information, Solomonoff (1964), Kolmogorov (1965), and Chaitin (1975), introduced in Rissanen (1978) and (1983) in a particular nonpredictive form. The criterion also reduces to the maximum likelihood criterion in the special cases where the use of the latter is appropriate.

We may regard our work as a continuation of the program that Fisher began with his information, which is defined in terms of the covariance of the estimated parameters about a "true" parameter. Wishing to remove the untenable assumption of data generating systems and "true" parameters, we instead regard the class of models to provide a language in which to express the regular features in the data. Because then the models cannot be restricted to have a fixed number of parameters Fisher's information does not apply, and we must consider the complexity and information directly in the observed data. That the resulting complexity still is a meaningful concept is evidenced by the fact that even prediction errors can be so expressed. The fact that no "true" parameters are needed in our notion implies that the associated optimal model is not an

Received July 1984; revised October 1985.

AMS 1980 *subject classifications.* 62A99, 62M10, 62F03, 60F99.

Key words and phrases. Inference, number of parameters, model selection criteria, prediction, coding.

approximation of anything at all. Rather, it acquires its own data dependent meaning from the two interpretations of the complexity, the first as the shortest code length and the second as the smallest prediction errors. As a further consequence, any two models, even with different numbers of parameters, can be fairly compared, which, for example, in the important selection-of-variables problem pretty much settles a major question about the least-squares estimates left open by Gauss, namely, how to estimate the number of the regression parameters.

The first contribution in this paper is to define the notion of stochastic complexity, especially when the coding is done in a predictive manner. This will provide a criterion, which, unlike the earlier nonpredictive MDL criterion, penalizes the number of the parameters in the fitted models without any explicitly added term. The associated *predictive* MDL modeling principle turns out to be closely related to the "prequential" principle, discovered independently by Dawid (1984) for the case where the data have a natural order. Still another related idea, called "forward validation", appears in Hjorth (1982), where it, however, was used in a traditional manner to provide unbiased estimates or estimates with reduced bias.

The main contribution in this paper is a fundamental theorem, which sets a tight lower bound for the code length with which long strings of data can be encoded with help of a class of models. Because prediction is just another form of coding, the theorem also gives a universal lower bound for the mean prediction errors of any predictors. The theorem may further be used to assess the goodness of estimators, which, unlike in the Cramér–Rao bound, may include the estimates of the number of the parameters as well. This theorem with a companion theorem, stating that the complexity achieves the lower bound, may be taken to provide a rational basis for model comparison, regardless of whether the models have the same number of parameters or not.

As the third contribution of this paper we formalize concepts such as "useful information," and "prior information" in the data. We also define a universal test statistic for hypothesis testing, which appears to have a number of advantages. We illustrate the idea by a test of two-way contingency tables.

2. Predictive MDL principle. The probabilistic models we consider consist of indexed densities $f_a(x|u)$, or ultimately probabilities $P_a(x|u)$, where $x = x_1, \ldots, x_n$, also written as x^n, denotes a sample of length n as a response or "output" to another "input" sample $u = u^n$ of the same length. Because the input sample adds nothing new in principle, we drop it to simplify the notations; we illustrate its use in Example 2 below. We denote both random variables and their values with lower case letters, letting the context tell which is meant. The data items are often numerical, but, of course, not always. When numerical, each number in the binary notation, say, has only some number r of fractional digits. Hence, when the model is a density it assigns a probability to x, which is obtained by integrating the density over the n-dimensional cube of edge length 2^{-r} with x as the center. We denote this induced probability function by $P_a(x)$ without indicating the implicitly understood precision r, which we otherwise do

not need. The index α may be taken sufficiently general to allow comparison of nested and nonnested models alike. However, it is the number of parameters that turns out to be the interesting quantity, and we for simplicity take the index to be of the form $\alpha = (k, \theta)$, where k denotes the number of components in the parameter vector $\theta = (\theta_1, \ldots, \theta_k)$, and $k = 0, 1, \ldots$. The value $k = 0$ corresponds to the empty parameter λ.

We are interested in predicting the sequence x as well as coding it. The former may be viewed as a special predictive form of coding, and we gain generality by proceeding with the coding interpretation. Often, we wish to model the data such that the individual observations are independent. Then, instead of coding a sequence the relevant problem is to encode the n-element unordered "list" $\{x_i\}$, where repeated occurrences of a value are preserved. The required modification for such a case will be discussed below. Predictive coding means that we model the conditional density for the possible values of the "next" observation x_{t+1} thus

$$f_{k,\hat{\theta}(t)}(x_{t+1}|x^t), \tag{2.1}$$

where $\hat{\theta}(t) = \hat{\theta}(x^t)$ is obtained with an estimation algorithm for the parameter θ with k components. Such a density allows us to encode the observation x_{t+1} to the precision r with the "ideal" code length $-\log P_{k,\hat{\theta}(t)}(x_{t+1}|x^t)$, which, as just explained, is represented by $-\log f_{k,\hat{\theta}(t)}(x_{t+1}|x^t)$. The word "ideal" means that if the possible values of the next observation indeed are distributed as modeled, then no prefix code exists with a shorter mean length. Whenever we wish to express the code length as the number of binary digits in the coded string, the logarithm is to be taken to the base 2; otherwise, its base does not matter. By adding all these ideal code lengths, we get the total code length

$$L(x|k) = -\sum_{t=0}^{n-1} \log f_{k,\hat{\theta}(t)}(x_{t+1}|x^t). \tag{2.2}$$

This may be minimized with respect to k to give the estimator $\hat{k}(n) = \hat{k}(x^n)$, which with the last data point defines the final estimate $\hat{\theta}(n)$ having $\hat{k}(n)$ components.

How should we select the estimate $\hat{\theta}(t)$ for each k? On first thought one might think of picking it so as to minimize the ideal code length $-\log f_{k,\theta}(x_{t+1}|x^t)$. But, clearly, this cannot be done, because such a minimization would make $\hat{\theta}(t)$ a function of x_{t+1}, which, in turn, would make decoding impossible. Indeed, decoding of x_{t+1} requires the knowledge of $\hat{\theta}(t)$, which therefore must not depend on the value x_{t+1} to be decoded. We are faced with the central issue in inductive inference, and we reason as follows: In the light of past observations the best single value of the parameter for encoding the "next" observations, x_{i+1}, $i = 0, 1, \ldots, t-1$, is the value that minimizes the sum $-\sum_{i=0}^{t-1} \log f_{k,\theta}(x_{i+1}|x^i)$. This is the maximum likelihood estimate $\hat{\theta}(t)$, except that we add the restriction that the predicted density (2.1) is positive for every possible value of x_{t+1}, which is required to make (2.2) meaningful for all data sequences. We might then say that this choice for the estimator $\hat{\theta}(t)$ is based

upon the hope that the predicted distribution (2.1) for the new observation x_{t+1} is like it was in the past, which to us seems to be as sound a principle for statistical inference as any. After all, by its nature inductive inference is based precisely upon such a faith; the same reasoning was also applied in the "prequential" procedure, Dawid (1984).

The minimization of (2.2) requires the initial estimate $\hat{\theta}(0)$ for each number of components k. The traditional way to calculate such is to select more or less arbitrarily a prior density function for the parameters and then take one of the maximizing values as the estimate $\hat{\theta}(0)$. The predictive approach, however, offers a different way, and one which avoids the both conceptually and technically difficult problem of specifying the prior densities. Indeed, what (2.2) really requires is the specification of a density function $f(x_1)$ for the first observation such that it reflects our prior knowledge about its value. Technically, we may take this density function to be in the parametric family and specified by the empty parameter λ. Such a distribution is often much easier to pick than a prior for the parameters. For example, if the prior knowledge consists of the fact that the set of possible values of x_1 is finite, M, put $-\log f(x_1) = \log M$. The procedure to compute (2.2) for each selected number of parameters k is then as follows: The first observation x_1 is encoded with the ideal code length $-\log f(x_1)$, where the density is selected to represent our knowledge, often ignorance, about the value x_1. We continue encoding the next observations with this same density until one parameter can be uniquely fitted, and we increase the number of fitted parameters in this manner one by one until the set value k, needed in the evaluation of (2.2), is reached.

The minimized code length (2.2) does not quite represent the complexity of the sequence x, because it is conditioned on the optimizing number of parameters, which clearly is required in the decoding process. This value can be given in a coded form as a preamble in the entire code string. Because the decoder will have to be able to separate the binary codeword representing $\hat{k}(n)$ from the subsequent code of the data without a separating comma, the preamble must be a so-called prefix code. As discussed in Rissanen (1983), encoding the natural number k by a prefix code requires

$$L^*(k) = \log^* k + \log c \tag{2.3}$$

binary digits, where $\log^* k = \log k + \log\log k + \cdots$, the sum including all the positive iterates, and c is the constant, about 2.865, that makes $\sum_{n=1}^{\infty} 2^{-L^*(n)} = 1$. Therefore, we may define the (semi) *predictive* (stochastic) *complexity* of the sequence x, relative to the selected class of models, as

$$I_{SP}(x) = \min_k \{L(x|k) + \log^* k + c\}. \tag{2.4}$$

The word "semi" suggests that the optimizing number of parameters, which we still write as $\hat{k}(n)$, is not determined the predictive way. To avoid misunderstandings we emphasize that the main effect for penalizing the number of parameters in (2.4) is by no means due to the second term, $\log^* k$. In fact, in most if not all the cases of interest the minimizations of (2.2), where no such term appears, and (2.4) produce exactly the same number of parameters, which is why we may safely use the same symbol to denote both.

We can apply the above discussed inductive reasoning to obtain a purely predictive complexity. Indeed, let $\hat{k}(t)$ denote the minimizing number of parameters in (2.2), where n is replaced by t. Then we may regard the pair $(\hat{k}(t), \hat{\theta}(t))$ to represent our best estimate of the conditional density for the possible values of the "next" observation x_{t+1} available at time t. Adding the resulting ideal code lengths we get the purely predictive (stochastic) complexity as follows

$$l_{\mathrm{P}}(x) = -\sum_{t=0}^{n-1} \log f_{\hat{k}(t),\hat{\theta}(t)}(x_{t+1}|x^t), \tag{2.5}$$

where $\hat{k}(0) = 0$ and $\hat{\theta}(0) = \lambda$, representing the empty set of parameters. In other words, the initial density $f(x_1)$ is determined as described above.

In the case where the observed data do not form a natural sequence, and they are modeled as independent, we should modify (2.2) by minimizing it over all permutations. In other words, we should find that order which allows for the shortest code for the unordered list. Except for very small sample sizes such a search is far too complex, and we construct a symmetric function of the data by a local optimization procedure as follows:

$$L(x|k) = \sum_{t=0}^{n-1} \min_{j \notin \{i(1),\ldots,i(t)\}} \left\{ -\log f_{k,\hat{\theta}(x_{i(1)},\ldots,x_{i(t)})}(x_j) \right\}, \tag{2.6}$$

where the minimizing index j defines $i(t+1)$. The associated predictive and semi-predictive complexities for the unordered list of observed data are then defined analogously with (2.5) and (2.4). We illustrate this procedure by Example 2 below.

DISCUSSION. In Rissanen (1978) and (1983) we, in effect, defined a third, purely nonpredictive notion of complexity, which to within terms of order $\log n$ is given by

$$l_{\mathrm{NP}}(x) = \min_{k,\theta} \left\{ -\log f_{k,\theta}(x) + \frac{k}{2} \log n \right\}. \tag{2.7}$$

This formula results from a particular way of coding the data, where the second term represents the number of digits required to encode k parameters to an optimal precision. Clearly, this nonpredictive stochastic complexity cannot be meaningfully interpreted in terms of prediction errors. The criterion (2.7) is seen to be identical in form but not in scope nor in content with Schwarz's criterion, Schwarz (1978).

We now have three different versions of complexity for a sequence. This abundance is a reflection of the difficulty in defining the notion of complexity in an objective way for short data sequences. The trouble arises just as soon as we try to make precise the way one is allowed to use the models in the selected class to do the coding. For example, one cannot permit the estimators $\hat{\theta}(t)$ to be completely arbitrary, because there may exist the estimator that assigns the probability unity to the actually occurring value x_{t+1} for each t, which would give a perfect prediction. To be sure, such an estimator would require the

knowledge of this value, but for a given sequence such a "predictor" exists as a mathematical function. One could bar out such estimators by requiring their description to have a uniformly bounded length, independent of the sample size n, but then we would be back in the algorithmic notion of complexity, and we would not be able to use our model classes in any meaningful way as the "language" in which to look for the regular features in the data.

For these reasons, in the three notions of complexity the estimators $\hat{\theta}(t)$ to be used are specified one way or another. Although the chosen estimators, of course, are meaningful and natural, their selection was done on subjective grounds just the same, which is what we ideally would have liked to avoid. (Such qualms might be considered as being of no practical significance, but our aim in this paper is to seek a foundation for statistical reasoning which is as free from arbitrary choices as we can make it.) The situation improves rapidly with the growing sample size n, because then all the three notions of complexities tend to be equal, and, moreover, all of them, indeed, represent asymptotically the shortest code length, calculated per observation, available with any ways of doing the coding. This, of course, is the reason why it at all is meaningful to talk about complexities. We prove such results in the next section.

EXAMPLE 1. Consider the class of Bernoulli models. Hence, $k = 1$ and $\theta = p$, the probability of the occurrence of symbol 1. In lack of prior knowledge the first symbol is taken to occur with probability $\frac{1}{2}$. Hence no prior density in the parameter space nor the Bayesian formalism is needed. Having observed m occurrences of the symbol 1 in the past t symbols, we form the estimate $\hat{p}(t) = (m + 1)/(t + 2)$, which is seen to modify the maximum likelihood estimate such that the values 0 and 1 are avoided. By an easy induction the entire string of length n with n_i occurrences of symbol i, $i = 0, 1$, gets the probability $P(x) = n_0!n_1!/(n + 1)!$. Notice that the sum of this over all strings of length n is unity. The predictive and the semi-predictive complexities, taken either for strings or for unordered lists, agree, and they all are given by $-\log P(x)$, which by Stirling's formula also can be written as

$$l(x) = nH\left(\frac{n_0}{n}\right) + \tfrac{1}{2}\log n + O(1/n), \tag{2.8}$$

where $H(p) = -p \log p - (1 - p)\log(1 - p)$.

We next study to what extent prediction error measures can be interpreted as code lengths, which at the same time illustrates how the large classes of models as studied here are typically generated. Let $\hat{x}_{t+1} = g_\theta(x^t)$ denote a parametric predictor of x_{t+1}, where the parameter is to be determined from the past data. Usually, the predictors are defined by recurrence equations such as of the ARMA type. One may view this process as a means of accounting for the dependencies in the data, which when done well causes the prediction errors to be nearly independent. Next, let $\delta_t(x_{t+1}, \hat{x}_{t+1})$ denote a measure of the prediction error.

Now define

$$f_{t,\theta}(x_{t+1}|x^t) = K(x^t, \theta)2^{-\delta_t(x_{t+1}, \hat{x}_{t+1})}, \tag{2.9}$$

where $K(x^t, \theta)$ denotes that number for which $\int f_{t,\theta}(y|x^t)\,dy = 1$. We then see that

$$-\log f_{t,\theta}(x_{t+1}|x^t) = \delta_t(x_{t+1}, \hat{x}_{t+1}) - \log K(x^t, \theta)$$

represents an ideal code length for the observation x_{t+1} given the past data. With a suitable estimator $\hat{\theta}(t) = \hat{\theta}(x^t)$ the total ideal code length takes the form

$$L(x|k) = \sum_{t=0}^{n-1} \delta_t(x_{t+1}, \hat{x}_{t+1}) - \sum_{t=0}^{n-1} \log K(x^t, \hat{\theta}(t)). \tag{2.10}$$

We see that this predictive MDL criterion differs from the first sum, involving the prediction errors, only to the extent the second term depends on k. Most of the usual prediction error measures actually depend only on the difference $e_{t+1} = x_{t+1} - \hat{x}_{t+1}$, and, moreover, often the possible values of x_{t+1} range from $-\infty$ to $+\infty$. Then we see that $K(x^t, \hat{\theta}(t)) = K(x^t)$, and the difference between the two criteria amounts to a constant.

With the quadratic error function, giving rise to gaussian models, the predictive MDL principle reduces to a predictive least-squares principle, which we illustrate in the important selection-of-variables regression problem; for ARMA estimation, see Rissanen (1986a).

EXAMPLE 2. The observations consist of n tuples, $(x(i), u_1(i), \ldots, u_m(i))$, $i = 1, \ldots, n$, $m + 1$ elements in each, where m, the number of the regressor variables or "inputs" in our terminology, may be large. The basic problem is to find out which subcollection of the regressor variables gives the best prediction of the variable x. In order to simplify the description, we only look for subcollections consisting of the first k variables, and ask for an optimum value for k. The general case is similar except numerically more complex. We consider a linear predictor of the usual type

$$\hat{x}(i) = \sum_{j=0}^{k-1} a_j u_j(i), \tag{2.11}$$

where $u_0(i) = 1$. We measure the prediction errors by the sum of the squares, $\delta(x_{t+1} - \hat{x}_{t+1}) = \frac{1}{2}e_{t+1}^2$, where $e_{t+1} = x_{t+1} - \hat{x}_{t+1}$. Then $f_\theta(x_{t+1}|x^t)$, defined by (2.9), is seen to be normal with mean $\hat{x}_{t+1}$ and variance 1.

Ignoring initial knowledge we look for the smallest observation $x(i_1)$ according to (2.6), which we predict as 0. For $k = 0$, which means that we ignore all the regressor functions, the best fit for a_0 from the past observations at times $i_1, \ldots, i_t$ is the average $\hat{a}_0(t) = 1/t\Sigma_{j=1}^t x(i_j)$, which is taken as the prediction of that observation $x(i_{t+1})$, among those not yet predicted, for which the prediction error is smallest. Adding such prediction errors $(x(i_t) - \hat{a}_0(t))^2$ over all the data gives $L(0)$. For $k = 1$ we still predict $x(i_1)$ as 0. Recursively, suppose we have calculated $i_1, \ldots, i_t$ (which need not coincide with the indices found for $k = 0$), at

which points the corresponding prediction errors $e^2(i_j)$ also have been determined. We find $\hat{a}_{0,1}(t)$ and $\hat{a}_{1,1}(t)$ by minimizing $\sum_{r=1}^{t}(x(i_r) - a - bu_1(i_r))^2$ with respect to a and b. This gives the prediction error $e(i_{t+1}) = x(i_{t+1}) - \hat{x}(i_{t+1})$, where $\hat{x}(j) = \hat{a}_{0,1}(t) - \hat{a}_{1,1}(t)u_1(j)$, and i_{t+1} denotes the index of the variables not yet predicted for which the prediction error is smallest. Adding again the squared prediction errors over all the data we get the sum $L(1)$. We continue this way calculating $L(k)$ for each k, and we find the minimizing number $\hat{k} = \hat{k}(n)$, which, in turn, defines the least squares estimates for the real valued parameters from all the data.

This routine for fitting polynomials to a scattered points, marked by hand on the screen of a computer, was programmed. The displayed optimum degree polynomials agreed in an uncanny way with the best polynomial judged by the human eye. Also, it was found to be essential to compute the predictive complexity for an unordered list, rather than for the data ordered in a random way, to avoid initialization problems. We proved in Rissanen (1984b) that under reasonable conditions on the regressor variables the estimates $\hat{k}(n)$ are consistent, and that the estimates of all the parameters are asymptotically optimal in minimizing the mean per observation prediction errors $E(1/n)S(\hat{k}(n))$ in the case where such a proof makes sense, namely, where a "true" set of parameters exists. These results, in effect, settle the issue of how one should estimate the number of regressor variables, because prediction is the very reason we want them.

3. Main bound. The main result to be stated requires certain smoothness conditions on the parametric densities, which determine the way certain estimates of the parameters converge. These conditions need a verification in each individual class of models. By the theory of large deviations they can be shown to be satisfied for the class of Markov chains, as shown by H. Künsch in Rissanen (1986b). By a rather different (and difficult) analysis they can also be verified for the gaussian ARMA models. For the gaussian regression case, Example 2 above, the required conditions are trivially satisfied.

THEOREM 1. *Let for each k the parameters θ range over a compact subset Ω^k with nonempty interior of the k-dimensional Euclidean space. We assume that there exist estimates $\hat{\theta}(x^n)$ satisfying the central limit theorem such that the tail probabilities are uniformly summable as follows*

$$(3.1)\quad P_\theta\{\sqrt{n}\,\|\hat{\theta}(x^n) - \theta\| \geq \log n\} \leq \delta(n), \quad \text{for all } \theta, \text{ and } \sum_n \delta(n) < \infty,$$

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,

$$(3.2)\qquad \liminf \frac{E_{k,\theta}\log[f_{k,\theta}(x^n)/g(x^n)]}{(k/2)\log n} \geq 1.$$

The mean is taken relative to the distribution defined by $f_{k,\theta}$.

The proof is given in Appendix A.

DISCUSSION. The claim can also be stated thus: For all k, all positive numbers ε, and for all points $\theta \in \Omega^k$, except in a null set,

$$E_{k,\theta}\log\frac{f_{k,\theta}(x^n)}{g(x^n)} \geq \left(\tfrac{1}{2} - \varepsilon\right)k \log n, \tag{3.3}$$

for all but finitely many values of n.

This theorem has many uses, the most important of which is that it justifies the notions of stochastic complexity thereby providing a rational basis for model assessment regardless of the number of parameters in them. To see this we first demonstrate how the theorem may be regarded as a generalization of Shannon's famous coding theorem. Let $L(x)$ be any real-valued function, interpreted as a code length, which satisfies the inequality

$$L(x) \geq -\log g(x), \quad \text{all } x, \tag{3.4}$$

where $g(x)$ is a density defining a random process. Because it integrates to unity over the sequences of the same length, we see that (3.4) requires the length to satisfy a generalized Kraft inequality. But g also satisfies the compatibility conditions for a random process, which is reflected in a similar but weaker requirement for the code length. These conditions are natural enough, and all the usual codes satisfy them. We call such a code length *regular* for brevity. By applying the theorem to the density $g(x)$ the inequality (3.4) converts (3.3) to the desired result,

$$E_{k,\theta}L(x^n) \geq H_{k,\theta}(n) + \left(\tfrac{1}{2} - \varepsilon\right)k \log n, \tag{3.5}$$

where $H_{k,\theta}(n)$ denotes the entropy of the data sequences of length n. Shannon's inequality results from $k = 0$, which represents the case of a model class having only one member, and fixing n.

It is readily seen that (3.4) holds for the semi-predictive complexity $L(x) = l_{\mathrm{SP}}(x)$ and

$$g(x) = \sum_{k=1}^{\infty} 2^{-L^*(k)}2^{-L(x|k)},$$

where $L(x|k)$ is given by (2.2). The predictive complexity, too, is regular, for it satisfies (3.4) with equality, as seen by putting $g(x) = 2^{-l_{\mathrm{P}}(x)}$. The nonpredictive complexity (2.7) satisfies (3.4) up to terms of order $\log n$. The inequality (3.5), then, gives a justification for the term "minimum description" in Rissanen (1983); there we described a particular coding scheme but did not prove that it produces an asymptotically shortest code length.

Further, Theorem 1 gives a lower bound for the accumulated prediction errors resulting from "honest" predictors. By an "honest" predictor we mean one where the prediction of the t'th data item is made as a function of the previous items only. Such one-sided predictions guarantee by Bayes' theorem that the resulting minimized joint density is both proper and satisfies the compatibility conditions for a random process, as was seen to be the case with the predictive complexity

above. Hence, Theorem 1 applies. The qualification of "honesty" is necessary because the cross-validation technique in Stone (1974) and Geisser and Eddy (1979) involves prediction which is "dishonest" in the sense that a data item is predicted both from the "future" and the "past" values alike. Evidently, in a given sequence, only one intermediate value can be meaningfully so predicted; the other data items are needed in this prediction and, therefore, there is no point in predicting them. A specific statement of a bound for "honest" predictions for the regression problem and the gaussian ARMA processes is given in Rissanen (1984b) and (1984a), respectively. We give here the latter statement, where the quantifications are weaker than in Theorem 1 for the reason that at that time we had not verified the condition (3.1) for gaussian ARMA processes.

THEOREM 2. *Consider the set of gaussian ARMA(p, q) processes, where the $p + q + 1$ parameters $\theta = (a_1, \ldots, a_p, b_0, \ldots, b_q)$ range over a compact subset Ω^{p+q+1} of the $(p + q + 1)$-dimensional Euclidean space with nonempty interior. In particular, $b_0 = \sigma$, where σ^2 denotes the variance of the innovation process. Let $\hat{x}_t$ be any predictor of x_t as a measurable function of the past data x^{t-1}. Then for all p and q, all positive numbers ε, and for all points $\theta \in \Omega^{p+q+1}$, except in a set $A_\varepsilon(n)$, the volume of which shrinks to zero as n grows,*

$$E_\theta \frac{1}{n} \sum_{t=1}^{n} (x_t - \hat{x}_t)^2 \geq \sigma^2 \left(1 + \frac{p + q - \varepsilon}{n} \ln n\right). \tag{3.6}$$

Many criteria for estimation appearing in the literature can be expressed as the negative logarithm of a product of conditional densities for the data items with possibly some added terms to penalize the number of parameters, to be minimized over the parameters. In view of Theorem 1 there is a growing suspicion in this author's mind that unless the minimized criterion satisfies the inequality (3.4), the estimation will run into one or another kind of trouble. Prime examples of this are the maximum likelihood function and the ordinary least-squares criterion, neither of which can be applied to estimation of complex models where the number of the parameters is also to be estimated. The cross-validation criteria were meant to rectify this problem, but they appear to be asymptotically equivalent with Akaike's AIC, which does not satisfy (3.4), and they do not allow a consistent estimation of the number of the parameters, Stone (1977a, b). Hence, it is not the idea of "cross-validation" that does the trick but a normalization (3.4). For the same reason this author is skeptical of the Bayesian attempts to introduce improper priors. In contrast, the MDL principle allows any sorts of "prior" assignments, whether they are determined from a part of the data or from no data at all so long as they produce decodable parameters. And always the optimized code length is regular.

Another form of Shannon's inequality states that the Kullback–Leibler distance between a "true" density function f and any modeled density g is nonnegative. We can sharpen the inequality if we specify only that the "true" density is one in a family, and we allow the modeled density to result from any estimation procedure; for example, we may take $g(x) = \prod_t f_{\hat{a}(t)}(x_{t+1}|x^t)$ where

$\bar{\alpha}(t) = (\bar{k}(x^t), \bar{\theta}(x^t))$ is some estimator. Then by (3.2) not only is the Kullback–Leibler distance nonnegative, but it must be strictly positive at least by an amount which reflects the uncertainty that the "true" parameter is one in the chosen class. This suggests the interpretation that this amount $(k/2)\log n$ represents the optimal model complexity. We may thus view (3.2) as providing a yardstick for the goodness of an estimator $\bar{\alpha}(x)$ by comparing the associated code length $-\log g(x)$ for long strings with the nonpredictive complexity, which represents the lower bound in Theorem 1. This is particularly appropriate because all the three complexities appear to reach the lower bound asymptotically in an "almost sure" sense; see for example the discussion in Dawid (1984), which also gives a nice justification for the term $(k/2)\log n$.

It seems to us that, indeed, the three complexities can be rigorously shown to reach the bound $-\log f_{k,\theta}(x) + (k/2)\log n$ for almost all samples, perhaps even more simply than in the "mean" sense, but we cannot see any way at all to prove that no shorter regular code length exists for almost all samples. About the reachability of the bound in Theorem 1 for the semi-predictive complexity, we can supply a proof only in selected cases. One such is the basic gaussian regression problem, Rissanen (1984b). Another is the class of Markov chains, Rissanen (1986b). A third appears to be the important ARMA class, although at this writing the job is not quite finished. Here, we prove such a result with independence conditions.

THEOREM 3. *Let the family of densities satisfy the conditions for independence for each k and $\theta \in \Omega^k$, namely, $f_{k,\theta}(x) = \prod_{t=1}^n f_{k,\theta}(x_t)$, and let $f_{k,\theta}(x_t)$ be three times continuously differentiable with respect to θ in the interior of a compact set Ω^k. Further, let the central limit theorem hold for some estimates $\hat{\theta}(x^n)$ of θ in the interior points such that the four first moments of $\sqrt{n}\,(\hat{\theta}(x^n) - \theta)$ converge. Then $l_{\mathrm{SP}}(x)$, defined by Eq. (2.4), is optimal in that for all k and all θ in Ω^k,*

$$l_{\mathrm{SP}}(x^n) \leq -E_{k,\theta}\log f_{k,\theta}(x^n) + \frac{k}{2}\log n + o(\log n), \tag{3.7}$$

where $o(\log n)/\log n$ goes to zero.

The proof is given in Appendix B.

4. Information in experiments. In this section we wish to define in a formal way such frequently used intuitive notions as "information in an experiment" and "prior information." The first of these has been defined earlier in terms of the Fisher information and in some contexts in terms of the Kullback–Leibler information, Gokhale and Kullback (1978) and Lindley (1956). Although both concepts do have the right flavor, their scope is limited to model classes with a fixed number of parameters, and a change of view is needed to generalize them for the model classes studied in this paper. As regards the "prior information", the early definition by Lindley (1956) in terms of Shannon information, is strictly restricted to the traditional way of modeling prior knowledge as a

distribution about a "true" parameter value. In other words, the only source of uncertainty stems from sampling, which, of course, is a grossly simplified view and leads to absurdities of the kind that the importance of parameters and their estimates can be measured solely in terms of how narrow the distribution of their estimates is.

Intuitively, by "useful information" in the data we mean something that we can learn. This certainly cannot be the "disorder," which we measure by complexity. Rather, such information must have the nature of regular features or constraints that we can discover, which suggests that it could be measured in terms of the reduction in the total code length below a certain neutral level, obtainable with use of no model at all. In order to calculate such a neutral level, we use a universal prior density on nonnegative real numbers, which we can define by help of the universal distribution for the integers constructed in Rissanen (1983) and given by (2.3). This universal distribution provides asymptotically the most efficient coding of integers. For example, it follows from the work in Bentley and Yao (1976) on the function log* that the length of any prefix code sequence on the natural numbers must exceed $L^*(n) - 2\hat{k}^*(n)$ infinitely many times, where $k^*(n)$ denotes the number of terms in $\log^*(n)$. Because the second term grows very slowly, we conclude that the universal sequence is not far from the least asymptotic upper bound for all probability sequences on the positive integers. That such a bound cannot be reached by any sequence is not a serious practical defect, and we feel quite free to use $L^*(n)$ as an excellent representative of the universal prior. We extend this universal distribution to a universal density for the positive real numbers as follows:

$$q^*(y) = \frac{1}{c} 2^{-\log^*(\bar{y})}, \tag{4.1}$$

where $\bar{y}$ is the smallest integer greater than or equal to y. This has the property we need: It accurately represents the complexity of any truncated real number. If a number y has r fractional decimal digits, the above density assigns to it the probability $10^{-r}q^*(y)$, and its complexity may be taken as the negative binary logarithm of this probability. For example, the number 275.233 has the complexity $10 + \log^*276 \cong 22$, and there is no way to describe this number with fewer binary digits while maintaining the additional requirement that the description can be decoded even when it is followed by other binary symbols; i.e., that the code is a prefix code. And it is precisely this property which we regard as the foremost requirement in any prior that can claim universality; for example, it follows that scale changes and other such transformations cannot reduce the complexity in a substantial way.

The universal density at the observed sequence x is taken as $q^*(x) = \prod_{t=1}^{n} q^*(x_t)$. Now we define

$$I^n(x) = -\log q^*(x) - l(x), \tag{4.2}$$

where $l(x)$ is one of the complexities defined in Section 2, to be the *information in a statistical experiment*, defined by x and the considered class of models. This represents the amount of "useful" information in the data. The word "useful" is

to be taken only in the sense of extractable regular features relative to the considered class of models. Two strings of data might well have the same amount of useful information, but we might consider one of them to be more valuable for some practical purpose. The measure also corresponds to intuition: If the data is "random" in the sense that it cannot be compressed by any model, then the useful information is zero, or near zero, as it should be. On the other hand, if we have guessed the model class right, then the best model that gives the complexity also incorporates the maximum amount of useful information; there is nothing more to learn from the data with the proposed models. However, another model class might be found which compresses the data more, and which allows us to learn more.

On intuitive grounds we would expect the information $I^n(x)$ to be positive, if the chosen model class is reasonable. To show that kind of property we calculate the mean of this information. If we consider classes of models which satisfy the assumptions in Theorems 1 and 3, then by putting $g(x) = q^*(x)$ we see that the mean information over strings of length n satisfies

$$E_{k,\theta} I^n(x) = E_{k,\theta} \log \frac{f_{k,\theta}(x)}{q^*(x)} - E_{k,\theta}[l(x) + \log f_{k,\theta}(x)] \geq -\varepsilon \log n,$$

under the qualifications regarding θ, n, and ε stated in (3.3). Here $l(x)$ denotes one of the complexities in Section 2. This is really the extreme case where we have picked a worthless class of models. In any other reasonable class, $l(x)/n$ will be below $-\log q^*(x)/n$ by the order of a constant, and the mean useful information will be strictly positive even for finite values of n.

In our general philosophy of modeling there are no data generating probabilistic systems nor "true" parameter values. Therefore, it is meaningless to measure the amount of prior knowledge in terms of Shannon information of the random variable taking values in the parameter space with some prior distribution. Indeed, a narrow concentrated distribution does not represent a great amount of prior knowledge unless the center of concentration represents a "good" parameter value; that is, one which selects a model from the family which captures well the regular features in the data. The Shannon information and hence Lindley's measure of prior knowledge are independent of the most important source of such knowledge, namely, the location of the concentration. For these reasons we define the prior information differently. Let $\hat{\alpha}(0) = (\hat{k}(0), \hat{\theta}(0))$ be a prior estimate of the number of the parameters and their values, respectively. As explained in Section 2, these estimates may define an empty parameter λ, which selects a special density $f_\lambda(x_1)$ from the family. We define

$$I^0(x) = \log \frac{f_{\hat{\alpha}(0)}(x)}{q^*(x)} - \log^* \hat{k}(0) \tag{4.3}$$

to represent the *prior information*, provided by the prior estimates. The difference

$$I_L(x) = I^n(x) - I^0(x) = -\log f_{\hat{\alpha}(0)}(x) - l(x) + L^*(\hat{k}(0)) \tag{4.4}$$

clearly represents the information provided by the likelihood function alone.

What about positivity of these last two notions of information? It seems to us that there should be no reason why just any prior parameter value ought to be able to extract useful information from the data. In fact, we might even do worse than what is achievable with the universal density $q^*(x)$, so that the prior information might be negative. However, when the sample is large we should definitely expect to learn something about it with the likelihood function so that the mean of $I^n(x) - I^0(x)$ should be positive. This, indeed, can be shown under the same qualifications as the positivity of the mean of $I^n(x)$. Just pick $g(x) = f_{\hat{\alpha}(0)}(x)$ in Theorem 1, and apply both Theorems 1 and 3.

5. Model testing. In this concluding section we wish to illustrate the wide scope of the notion of stochastic complexity by applying it to hypothesis testing. Consider a set of models $\{f_{\alpha'}(x)\}$, representing a composite null-hypothesis, and another disjoint set $\{f_\alpha(x)\}$, representing a composite alternative hypothesis. The indices $\alpha = (k, \theta)$, and $\alpha' = (k', \theta')$ are arbitrary. Let $l'(x)$ and $l(x)$ denote the semi-predictive stochastic complexities of x relative to the two classes of models, respectively. Then we take the difference $D(x) = l(x) - l'(x)$ to be a test statistic, and decide in favour of the null-hypothesis if $D(x) \geq 0$, and against it, otherwise. Notice that in case of two simple hypotheses this test statistic coincides with the likelihood ratio, which is known to be the most efficient test statistic. A related test statistic was also considered in Dawid (1984).

This testing has a number of advantages over the traditional testing procedures. First, even composite hypotheses get represented by a single model, namely, the model which gives the complexity of the data relative to the class of models. Hence, there is the possibility of learning by finding better and better model classes. Second, our test does not require knowledge of the distribution of the test statistic, and, hence, it is valid for small and large samples alike. It is clear that the validity for small samples will have to be verified by applying the test to a large number of actual cases, where the results are known. We have studied a half a dozen of them, and in each the test result appears highly reasonable. For large samples an analytic validity test can be made, and again the results appear to be good; see the examples below. Third, the size of the test is automatically adjusted to the amount of observed data. One might argue in favor of a subjectively selected size, but even if we can easily do the same we cannot see why such a thing ought to be done. After all, such a number adds nothing to the amount of information that can be extract from the data, and hence it is quite irrelevant in deciding which of the two considered hypotheses is the more likely explanation of the data. After this question is settled, we regard it as a separate matter to judge the consequences of acting on the result, which depend on issues and values that have nothing to do with the data. We illustrate our approach with two examples.

EXAMPLE 3. Consider the null-hypothesis $p = \frac{1}{2}$ against the alternative $p \neq \frac{1}{2}$ in the class of Bernoulli models. We have by Example 1 in Section 2,

$$D(x) = \log \frac{(n+1)!}{n_0! n_1!} - n \cong \tfrac{1}{2} \log n - n(1 - H(n_0/n)). \tag{5.1}$$

We accept the null-hypothesis if $D(x) \geq 0$, which means that in order for us to accept the opposite hypothesis, the ratio n_0/n must differ sufficiently from $\frac{1}{2}$ to overcome the "cost" of one parameter. One can show that for sample sizes up to about 1000, this test is close to the traditional test with the confidence level about 0.05. For longer strings, the automatically given confidence level shrinks gradually to zero, as it should. Notice that in the ordinary way of doing the testing there is no cost associated with the number of parameters, and hence the opposite hypothesis would always win by a direct comparison. This is why a direct comparison of the likelihoods cannot be made, and one must, instead, introduce an artificial threshold.

EXAMPLE 4. As a less trivial example, consider a two-way $r \times s$ contingency table with the ijth entry being n_{ij}. The observations x in reality consist with a sequence $(i_1, j_1), \ldots, (i_n, j_n)$, where $1 \leq i_k \leq r$, $1 \leq j_k \leq s$, and n_{ij} denotes the number of times (i, j) occurs in x. Let the class of models be the set of multinomials with $\theta = \{P(ij) = p_{ij}\}$ as the parameters. The null-hypothesis states that the cell probabilities are determined by $r + s$ marginal probabilities $p_{ij} = p_i p_j$ expressing independence, while the opposite hypothesis claims that no such restrictions exist. Hence, the model defined by the null-hypothesis has $r + s - 2$ free parameters while the competing "free" model requires $rs - 1$ free parameters.

We compute the predictive code lengths for the sequence x with the two hypotheses. Just as in the case of Bernoulli models in Example 1 in Section 2, with which the first symbol in a binary sequence is assigned the probability $\frac{1}{2}$, we imagine that each cell in the table has the initial content of 1 under both models. This assigns to the first cell occurrence in the string x the probability $1/rs$. After this occurrence the corresponding cell content in the table under the "free" model is incremented by 1, which gives the table for the assignment of the probability for the second occurrence, and the process is repeated. We see that the probability assigned to the entire string x under the free model is

$$P_F(x) = (rs - 1)! \frac{\Pi_{ij} n_{ij}!}{(n + rs - 1)!}. \tag{5.2}$$

Analogously the probability of the string x under the independence assumption is obtained as

$$P_I(x) = \frac{(rs - 1)!}{n!^2} \prod_i (n_{i\cdot} + s - 1)! \prod_j (n_{\cdot j} + r - 1)!, \tag{5.3}$$

where $n_{i\cdot} = \sum_j n_{ij}$, and similarly for $n_{\cdot j}$. We then have $D(x) = \log(P_I(x)/P_F(x))$, which with Stirling's formula gives

$$2D(x) = (r - 1)(s - 1)\ln n - 2\sum_{ij} n_{ij} \ln \frac{n_{ij}}{n_{i\cdot} n_{\cdot j}/n} + O(1/n). \tag{5.4}$$

We reject the independence hypothesis if $D(x) < 0$. The sum term in (5.4) is exactly the so-called Kullback G^2 measure. It has asymptotically a χ^2

distribution with $(r-1)(s-1)$ degrees of freedom. Therefore, for large values of n, our test is like the ordinary test except that the arbitrarily selected level of confidence is replaced by the first term. However, our test is valid even for small values of n, when the χ^2 distribution for the G^2 measure and, hence, the ordinary test are not justified.

APPENDIX A

PROOF OF THEOREM 1. We pick k and write f_θ and E_θ for the density and expectation, respectively, defined by a parameter vector θ with k components. We also denote the so-induced probability measure by P_θ. For each parameter θ in Ω^k let $J_n(\theta)$ denote a closed neighborhood of radius $r_n = \log n/\sqrt{n}$ with θ as its center. Define for the process determined by θ the set of its *θ-typical* strings of length n

$$Y_n(\theta) = \{x^n | \hat{\theta}(x^n) \in J_n(\theta)\}, \tag{A.1}$$

where $\hat{\theta}(x)$ denotes an estimate of θ satisfying the assumptions in the theorem. Let $P_n(\theta)$ denote the probability under P_θ of the strings that fall within $Y_n(\theta)$, which is the same as the probability that $(\hat{\theta}(x^n) - \theta)\sqrt{n}$ falls within a neighborhood of radius $\log n$ about the origin. By the assumption, this probability exceeds $1 - \delta(n)$, where $\delta(n) \to 0$.

Before continuing, we give the gist of the proof which is really quite simple. Let the parameter have only one component and let Ω be divided into N equal segments $J_n(\theta_i)$, $i = 1, \ldots, N$, of length $2r_n$. (Regard the given choice for the radius as a good guess.) Now, if a density g exists whose Kullback distance from f_{θ_1} is short, then the probability mass given by g to the set $Y_n(\theta_1)$ must exceed a certain amount, depending on how short a distance we demand. The same holds for $\theta_2, \theta_3, \ldots$. But there is only the total mass of unity available, and it so happens that g can be as close as stated only to a preciously few of the densities f_{θ_i}, which is what the theorem in essence states. A critical point in the proof is to make sure that the sets of strings for which g has to distribute its mass are indeed disjoint. The statement of the theorem is such that it is not enough to consider the sets $Y_n(\theta)$, which, of course, are disjoint by their very definition, but we have to add probabilities of sets of strings with different lengths.

We need to define another smaller set of typical sequences, namely, the set

$$X_{n,t}(\theta) = \left\{x^{n+t} | x^{n+j} \in Y_{n+j}(\theta),\ j = 0, 1, \ldots, t\right\}.$$

In words, this is the subset of sequences of $Y_{n+t}(\theta)$ such that not only the sequences themselves are typical but also all their prefixes of length ranging from n to $n + t$ are typical as well. We need to estimate the probability $P_{n,t}(\theta)$ of this subset. Let $Z_j(\theta)$, for j in the range $0 \le j \le t$, denote the set of sequences x^{n+t} in $Y_{n+t}(\theta)$ with the property that the prefix x^{n+j} of x^{n+t} is the first which is not in $Y_{n+j}(\theta)$. Hence, all the shorter prefixes $x^n, \ldots, x^{n+j-1}$ belong to $Y_{n+0}(\theta), \ldots, Y_{n+j-1}(\theta)$, respectively. The case $j = 0$ means that already the first prefix of length n does not belong to $Y_n(\theta)$. It is clear that the union of $Z_j(\theta)$ over

j is precisely the difference $Y_{n+t}(\theta) - X_{n,t}(\theta)$. Because $\delta(n)$ is summable, we deduce first that

$$P_\theta\{Y_{n+t}(\theta) - X_{n,t}(\theta)\} < \sum_{i=n}^{\infty} \delta(i) = \mu(n), \tag{A.2}$$

and then

$$P_{n,t}(\theta) > 1 - \delta(n) - \mu(n), \tag{A.3}$$

where $\delta(n) + \mu(n) \to 0$.

Let $g(x)$ be any density function defined on the data sequences, as stated in the theorem, and denote by $Q_{n,t}(\theta)$ the so induced probability over the set $X_{n,t}(\theta)$. Then by the nonnegativity of the Kullback distance between the densities $g(x)/Q_{n,t}(\theta)$ and $f_\theta(x)/P_{n,t}(\theta)$, we get

$$\int_{X_{n,t}(\theta)} f_\theta(x) \log \frac{f_\theta(x)}{g(x)}\, dx \geq P_{n,t}(\theta) \log \frac{P_{n,t}(\theta)}{Q_{n,t}(\theta)}. \tag{A.4}$$

Pick a positive integer m, and let $A_m(n, t)$ be the set of θ such that the left hand side of (A.4), denoted by $T_{n,t}(\theta)$, satisfies the inequality

$$T_{n,t}(\theta) < \left(1 - \frac{1}{m}\right) \log\left((n + t)^{k/2}\right). \tag{A.5}$$

We wish to calculate an upper bound for the volume $V_m(n)$ of $B_m(n) = A_m(n,0) + A_m(n,1) + \cdots$. To this end, let $N(n,0)$ denote the maximum number of disjoint neighborhoods $J_n(\theta)$ that can be constructed such that their centers θ lie in $A_m(n,0)$. Let $C(n,0)$ denote the set of the centers. These neighborhoods may not cover $A_m(n,0)$, because there may be points that are too close to some of the constructed neighborhoods without being covered by any. However, if we double the radius of each neighborhood in the maximal collection we get a cover $S(n,0)$ for $A_m(n,0)$. Recursively, let $N_{n,t}$ denote the maximum number of disjoint neighborhoods $J_{n+t}(\theta)$ that can be constructed such that their centers lie in the difference set $A_m(n, t) - S(n, t-1)$. Let $C(n, t)$ denote the set of the centers. This construction means, obviously, that if θ is in $C(n, i)$ and θ' in $C(n, t)$, for $i \leq t$, then the distance from θ' to $J_{n+i}(\theta)$ is not smaller than $(\log(n + i))/\sqrt{n + i}$. As above, the neighborhoods $J_{n+t}(\theta)$, for θ in $C(n, t)$, may not cover the difference set, but by doubling their radius the union of the resulting expanded neighborhoods together with $S(n, t-1)$ gives a cover $S(n, t)$ for $B_m(n, t) = A_m(n,0) + \cdots + A_m(n, t)$. Hence, the volume $V_m(n, t)$ of $B_m(n, t)$ is bounded by

$$V_m(n, t) \leq KN_{n,0}\left(\frac{\log n}{\sqrt{n}}\right)^k + \cdots + KN_{n,t}\left(\frac{\log(n + t)}{\sqrt{n + t}}\right)^k, \tag{A.6}$$

where K is a constant. We presently derive an upper bound for the right-hand side.

From (A.4) and (A.5) we conclude that

$$(A.7) \quad -\log Q_{n,t}(\theta) < \left[\frac{1-1/m}{P_{n,t}(\theta)} - \frac{\log P_{n,t}(\theta)}{\log\left((n+t)^{k/2}\right)}\right]\log\left((n+t)^{k/2}\right),$$

for θ in $A_m(n,t)$. By picking n large enough we can by (A.3) make $P_{n,t}(\theta)$ so close to unity that the expression within the brackets is less than some number β, such that $0 < \beta < 1$, uniformly in t and θ. Hence,

$$(A.8) \qquad Q_{n,t}(\theta) > (n+t)^{-k\beta/2},$$

which holds for θ in $A_m(n,t)$ and n larger than some number.

We wish to prove the inequality

$$(A.9) \qquad 1 \geq \sum_{\theta \in C(n,0)} Q_{n,0}(\theta) + \cdots + \sum_{\theta \in C(n,t)} Q_{n,t}(\theta),$$

for all t. Each sum is a probability, induced by the density g, of a set of strings with length varying from sum to sum, and we cannot directly claim the inequality. However, consider the following. The neighborhoods $J_n(\theta)$ for θ in $C(n,0)$ are disjoint by construction, which makes the corresponding $N_{n,0}$ sets $X_{n,0}(\theta)$ disjoint. Hence, surely, the first sum in (A.9) does not exceed unity. For θ in $C(n,1)$, consider the set of the prefixes with length n of the strings in $X_{n,1}(\theta)$. Denote this subset of $X_{n,0}(\theta)$ by $U_{n,1}(\theta)$. The neighborhoods $J_{n+1}(\theta)$ for θ in $C(n,1)$ are not only disjoint from each other and from all of the neighborhoods $J_n(\theta')$ for θ' in $C(n,0)$, but since the distance from θ to any neighborhood $J_n(\theta')$, θ' in $C(n,0)$, exceeds $(\log n)/\sqrt{n}$, also none of the larger neighborhoods $J_n(\theta)$ for θ in $C(n,1)$ intersects any of the neighborhoods $J_n(\theta')$, for θ' in $C(n,0)$. This means that for θ in $C(n,1)$ the set $X_{n,0}(\theta)$ does not intersect any of the sets $X_{n,0}(\theta')$, θ' in $C(n,0)$, which appear in the first sum of (A.9). Therefore, because g is a density satisfying the compatibility conditions for a random process,

$$\sum_{\theta \in C(n,1)} Q_{n,1}(\theta) \leq Q\left[\bigcup_{\theta \in C(n,1)} U_{n,1}(\theta)\right],$$

where Q denotes the probability measure defined by the density g, and (A.9) holds for $t = 1$. The same arguments apply for any t, because, as we explained above in the paragraph preceding (A.6), if θ is in $C(n,i)$ and θ' in $C(n,t)$, for $i \leq t$, then the distance from θ' to $J_{n+t}(\theta)$ is not smaller than $(\log(n+i))/\sqrt{n+i}$. Therefore, for θ in $C(n,t)$ and θ' in $C(n,i)$, $i \leq t$ and $\theta \neq \theta'$, the sets $X_{n,t}(\theta)$ and $X_{n,i}(\theta')$ consist of strings with length $n+t$ and $n+i$, respectively, such that their prefix sets of length n, $U_{n,t}(\theta)$ and $U_{n,i}(\theta')$ are disjoint. This in turn means that the two sets of strings of length n, $\bigcup_{\theta \in C(n,t)} U_{n,t}(\theta)$ and $\bigcup_{i<t,\,\theta \in C(n,i)} U_{n,i}(\theta)$, are disjoint, and (A.9) follows.

We now put the various pieces together to conclude the proof. From (A.6) we get first

$$V_m(n,t) < K \sum_{i=1}^{t} N_{n,i}(n+i)^{-k\beta/2}(n+i)^{k(\beta-1)/2}(\log(n+i))^k.$$

Because $(\log n)/\sqrt{n}$ is eventually monotone decreasing, we also have eventually

$$V_m(n,t) \le Kn^{k(\beta-1)/2}(\log n)^k \sum_{i=0}^{t} N_{n,i}(n+i)^{-k\beta/2} \le Kn^{k(\beta-1)/2}(\log n)^k, \tag{A.10}$$

where the last inequality results from (A.8) and (A.9). This holds for all t and all sufficiently large values for n. Hence the monotone increasing sequence $V_m(n,t)$, $t = 1,2,\ldots$ has a limit, which is $V_m(n)$, still bounded from above by the rightmost term in (A.10) for all sufficiently large n. Because this term converges to zero as n grows to infinity, so does $V_m(n)$.

We have shown that the measure of the set $\limsup A_m(n,t)$ is zero, or, equivalently, that for all m, the measure of the set of parameters θ for which all of the inequalities

$$T_{n,0} \ge (1-1/m)\log(n^{k/2}), \qquad T_{n,1} \ge (1-1/m)\log\big((n+1)^{k/2}\big)\ldots, \tag{A.11}$$

hold for some value of n, is the measure of Ω^k. Consider the inequality $\log y \ge a(1-(1/y))$, where $a = 1/\ln b$, b being the base of the logarithm. By putting $y = f_\theta(x)/g(x)$ we get further

$$\int_{x\in\bar{X}_{n,t}(\theta)} f_\theta(x)\log\frac{f_\theta(x)}{g(x)}\,dx > -a,$$

where $\bar{A}$ denotes the complement of the set A. Using this we get for the points θ where (A.11) hold,

$$E_\theta \log\frac{f_\theta(x^{n+t})}{g(x^{n+t})} \ge T_{n,t} - a \ge (1-1/m)\log\big((n+t)^{k/2}\big) - a.$$

Since for each such θ these hold for some n and all t, the statement in the theorem follows. □

APPENDIX B

PROOF OF THEOREM 3. Write $g(\theta,w) = -\ln f_{k,\theta}(w)$ for short, where w ranges over the reals, and denote the row vector of the first partials by $\Delta g(\theta,w) = \partial g(\theta,w)/\partial\theta$ while $J(\theta,w)$ denotes the matrix of the double derivatives of g. We now use natural logarithms. Then for each k the inequality

$$l_{\mathrm{SP}}(x) \le -\sum_{0}^{n-1} \ln f_{k,\hat{\theta}(t)}(x_{t+1}) + C_k$$

holds by the definition of the semi-predictive complexity, where $C_k = \ln 2^{\log^* k}$. Further, with $\delta_t = \hat{\theta}(t) - \theta$ we get from Taylor's expansion of $g(\theta,x)$ about θ

$$l_{\mathrm{SP}}(x) + \ln f_{k,\theta}(x) \le \sum_{t=0}^{n-1}\big[\Delta g(\theta,x_{t+1})\delta_t + \tfrac{1}{2}\delta_t' J(\theta,x_{t+1})\delta_t + R(\tilde{\theta}(t),x_{t+1})\big] + C_k,$$

where $\tilde{\theta}(t)$ is a point on the line segment connecting $\hat{\theta}(t)$ and θ, and the remainder term $R(\tilde{\theta}(t), x_{t+1})$ is proportional to a sum of terms of type

$$\frac{\partial^3 g(\theta, x_{t+1})}{\partial\theta_i\,\partial\theta_j\,\partial\theta_r}(\hat{\theta}_i(t) - \theta)(\hat{\theta}_j(t) - \theta)(\hat{\theta}_r(t) - \theta),$$

all evaluated at $\tilde{\theta}(t)$. These triple partials are uniformly bounded in the compact subset Ω^k. Since the moments of δ_t converge, we get by Schwarz's inequality that $E_{k,\theta}|R(\tilde{\theta}(t), x_{t+1})| \le Kt^{-3/2}$, for some constant K. Further,

$$E_{k,\theta}(\delta_t\delta_t') = (tJ)^{-1} + t^{-1}\rho_t,$$

where $E_{k,\theta}J(\theta, x_{t+1}) = J$ and $\rho_t \to 0$. Hence, by using the independence of x_{t+1} and $\hat{\theta}(t)$ together with the fact that the mean of the latter is θ, we get

$$n^{-1}E_{k,\theta}\left[l_{\mathrm{SP}}(x) + \ln f_{k,\theta}(x)\right] \le \frac{k}{2n}\sum_{t=1}^{n}\frac{1}{t} + r_n,$$

where

$$|r_n| \le \frac{1}{2n}\sum_{t=1}^{n}\left(\frac{|\rho_t|}{t} + Kt^{-3/2}\right) + \frac{C_k}{n}.$$

The sum of the harmonic series is $\ln n + \mu_n$, where $\mu_n/\ln n \to 0$. With a little struggling one sees that also $nr_n/\ln n \to 0$, which completes the proof. □

Acknowledgment. An anonymous referee deserves credit for carefully reading several versions of this paper and providing many helpful suggestions.

REFERENCES

AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control* **AC-19** 716–723.

BENTLEY, J. L. and YAO, A. C. (1976). An almost optimal algorithm for unbounded searching. *Inform. Process. Lett.* **5** 82–87.

CHAITIN, G. J. (1975). A theory of program size formally identical to information theory. *J. Assoc. Comput. Mach.* **22** 329–340.

DAWID, A. P. (1984). Present position and potential developments: some personal views, statistical theory, the prequential approach. *J. Roy. Statist. Soc. Ser. A* **147** 278–292.

GEISSER, S. and EDDY, W. (1979). A predictive approach to model selection. *J. Amer. Statist. Assoc.* **74** 153–160.

GOKHALE, D. V. and KULLBACK, S. (1978). *The Information in Contingency Tables*. Dekker, New York.

HJORTH, U. (1982). Model selection and forward validation. *Scand. J. Statist.* **9** 95–105.

KOLMOGOROV, A. N. (1965). Three approaches to the quantitative definition of information. *Problems Inform. Transmission* **1** 4–7.

LINDLEY, D. V. (1956). On a measure of the information provided by an experiment. *Ann. Math. Statist.* **27** 986–1005.

RISSANEN, J. (1978). Modeling 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. (1984a). Universal coding, information, prediction, and estimation. *IEEE Trans. Inform. Theory* **IT-30** 629–636.

RISSANEN, J. (1984b). A predictive least squares principle. *IMA J. of Math. Control and Information*. To appear.

RISSANEN, J. (1985). Minimum description length principle. In *Encyclopedia of Statistical Sciences* (S. Kotz and N. L. Johnson, eds.) **5** 523–527. Wiley, New York.

RISSANEN, J. (1986a). Order estimation by accumulated prediction errors. In *Essays in Time Series and Allied Processes* (J. Gani and M. B. Priestley, eds.) 55–61. Applied Probability Trust, Sheffield, England.

RISSANEN, J. (1986b). Complexity of strings in the class of Markov sources. *IEEE Trans. Inform. Theory* **IT-32** 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. *Inform. and Control* **7** 1–22; Part II. *Inform. and Control* **7** 224–254.

STONE, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B* **36** 111–147.

STONE, M. (1977a). Asymptotics for and against cross-validation. *Biometrika* **64** 29–35.

STONE, M. (1977b). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. Roy. Statist. Soc. Ser. B* **39** 44–47.

IBM ALMADEN RESEARCH CENTER
650 HARRY ROAD
SAN JOSE, CALIFORNIA 95120