Hypothesis Selection and Testing by the MDL Principle

Article *in* The Computer Journal · April 1999

DOI: 10.1093/comjnl/42.4.260 · Source: DBLP

Jorma Rissanen
Molecular Devices, LLC.
209 PUBLICATIONS **25,918** CITATIONS

Hypothesis Selection and Testing by the MDL Principle

J. RISSANEN

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

The central idea of the MDL (Minimum Description Length) principle is to represent a class of models (hypotheses) by a universal model capable of imitating the behavior of any model in the class. The principle calls for a model class whose representative assigns the largest probability or density to the observed data. Two examples of universal models for parametric classes $\mathcal{M}$ are the normalized maximum likelihood (NML) model

$$\hat{f}(x^n \mid \mathcal{M}) = f(x^n \mid \hat{\theta}(x^n)) \Big/ \int_{\Omega} f(y^n \mid \hat{\theta}(y^n))\, \mathrm{d}y^n,$$

where Ω is an appropriately selected set, and a mixture

$$f_w(x^n \mid \mathcal{M}) = \int f(x^n \mid \theta) w(\theta)\, \mathrm{d}\theta$$

as a convex linear functional of the models. In this interpretation a Bayes factor $B - f_w(x^n \mid \mathcal{M}_1)/f_v(x^n \mid \mathcal{M}_2)$ is the ratio of mixture representatives of two model classes. However, mixtures need not be the best representatives, and as will be shown the NML model provides a strictly better test for the mean being zero in the Gaussian cases where the variance is known or taken as a parameter.

1. INTRODUCTION

Although the term ‘hypothesis’ in statistics is synonymous with that of a probability ‘model’ as an explanation of data, hypothesis testing is not quite the same problem as model selection. This is because usually a particular hypothesis, called the ‘null hypothesis’, has already been selected as a favorite model and it will be abandoned in favor of another model only when it clearly fails to explain the currently available data. In model selection, by contrast, all the models considered are regarded on the same footing and the objective is simply to pick the one that best explains the data. For the Bayesians certain models may be favored in terms of a prior probability, but in the minimum description length (MDL) approach to be outlined below, prior knowledge of any kind is to be used in selecting the tentative models, which in the end, unlike in the Bayesians’ case, can and will be fitted to data and hence may be rejected. As a final comment, a model for us is a probability measure for the data, so that a likelihood function as a parametric model is actually a class of models, one for each parameter value. When a hypothesis is represented by such a function, it is often called a ‘composite’ hypothesis.

The basic idea in the MDL principle is to represent each contemplated class of models by a single probability distribution, which then may be viewed as a universal model for the class. Such a universal model should be able to imitate any particular model in the class, in which case we can meaningfully compare different model class suggestions by the probability their universal models assign to the given data. This model selection principle, then, is a global maximum likelihood principle, which allows us to compare any two classes whether or not they are of the same type or have the same number of parameters. Since by the fundamental Kraft inequality we can construct a so-called prefix code for any probability distribution, and, conversely, such a code defines a probability distribution, the comparison may equivalently be done by the code length, which explains the name ‘MDL’ principle.

There is more than one way to construct a universal representation of a model class. The earliest way was to construct a two-part code, which consists of optimally quantized maximum-likelihood estimates of the parameters, followed by a code for the data designed with the resulting model. This was done in the special case of classification models by Wallace and Boulton [1], as early as 1968 and for general parametric model classes later by Rissanen [2]. While quite crude, such a construct was shown to be asymptotically optimal in a strong sense [3]. We give below better constructs, which have certain optimum properties even non-asymptotically. These also appear to provide unsurpassed results in practice.

The same principle can also be applied to hypothesis testing, for which it brings the important simplification that composite hypotheses are reduced to simple ones. An additional bonus is that the required test statistic is always the code length, which eliminates the search for an appropriate test statistic to be used. Traditional hypothesis testing is based on the celebrated work of Pearson and Neyman, which for two simple hypotheses provides the most efficient decision rule together with an assessment of the degree of confidence in the decision. These are clearly issues related to the question of how well the null hypothesis and the alternative one can be separated. Although the decision rule extends to (very) special classes of composite hypotheses, the important problem of confidence assessment, even when just the alternative hypothesis is composite, has never been solved in a fully satisfactory manner. The difficulty of the problem is easy to see. In such a case it often happens that some of the alternative hypotheses (models in the class) are close to the null hypothesis while others are far from it so that their separability will vary. There is a Bayesian approach to the problem, which is to compute the ratio of the posterior probabilities of the two hypotheses, or the related so-called Bayes factor, to give the desired assessment. However, both will depend on the priors, which in general are not only arbitrary but also objectionable to many, even in principle. Nevertheless, one can evaluate such ratios for various extreme priors and the resulting lower and upper bounds do give useful information about the available confidence [4], which is more than can be obtained with the traditional non-Bayesian means.

The fact that the MDL principle reduces the hypotheses in the test to simple ones provides a different way to assess the confidence. In fact, it seems that such a reduction is even necessary if confidence is to be expressed in terms of probabilities. These are not, however, posterior probabilities of the hypotheses as in the Bayesian analysis but probabilities or often densities of the data that could be generated by the two hypotheses—just as if they were simple. There is a problem, however, in applying the universal models of the MDL principle to hypothesis testing. This is because such models are aimed at providing a short code length for the data, which need not coincide with the requirement in hypothesis testing due to its lopsided nature, where the null hypothesis is given a special favorite status. We study this problem below for the basic Gaussian hypotheses and we show that the MDL representative, defined by the normalized maximum likelihood (to be described), provides a different and powerful way to test hypotheses. Interestingly enough, the numerator and the denominator in a Bayes factor can also be interpreted as representatives of hypotheses. We show, however, that these representatives for all reasonable priors have a smaller power than the normalized maximum likelihood universal model. Alternatively, if in the Bayesian type of test we replace the Bayes factor by the ratio of the MDL representatives, which no longer need nor can be interpreted as posterior probabilities, we end up in a strictly sharper test.

To summarize the unique features of the MDL principle we may say that it provides a universal yardstick for assessing the quality of a model or a model class, regardless of its type or the number of its parameters. This is done in terms of the probability the model or model class assigns to the observed data. It is important to understand that the probability assigned to the data or, equivalently, the code length for them, is not an absolute quality measure, which could not even be found in a computable manner, and hence the principle provides a criterion for model comparison rather than a recipe for finding good models. When calculating the code length any type of prior knowledge may be used, but only as a part of the suggested model class, and its role is limited to shortening the code length. As a practical matter, the computation of the code lengths for complex model classes is not straightforward. For many of the usual model classes formulas exist for the shortest code length, which then serve as building blocks to be used as parts of the complex model classes.

2. MODEL SELECTION

A central problem in statistics is to discover properties of probabilistic type in an observed data set, collectively defining a probability distribution as a *model*, which leaves us with the basic question of how to measure the amount and quality of the properties. As stated in Section 1, the answer in the MDL principle is by the amount of probability mass a model or the universal representative of a model class is capable of assigning to the data. This conforms to intuition; the higher the probability the less surprise the data confront us with, the probability of unity meaning that the model can duplicate the data with perfect explanation. One must bear in mind, however, that such data-constraining properties must almost entirely be inferred from the data alone and since a finite amount of data cannot completely represent the generating distribution we cannot ask for the universal model to do its job perfectly. This is true also in the algorithmic theory if it were to be used to obtain statistical models; even the shortest program might fail to capture the distribution generating the data, which reveals the ultimate folly in attempts to capture the idea of any 'true' underlying model. What we can meaningfully ask for is to obtain a model which captures as much of the data-restricting constraints as possible, which information is obviously in the data alone although there may be prior knowledge that at least partially describes the type of the observed data.

Before proceeding to discuss the rather intricate matters in the MDL principle it may be useful to press the analogy with the algorithmic theory a bit further. We recall the idea of Kolmogorov's minimum sufficient statistic as an optimal summary of the data, see [5, pp. 176, 182]. First, a 'summarizing property' of data may be formalized as a subset A where the data belongs along with other sequences sharing this property. Hence, the property A need not specify the sequence completely. We may now think of programs consisting of two parts, where the first part describes optimally such a set A with the number of bits given by the Kolmogorov complexity $K(A)$ and the second part merely describes x^n in A with about $\log |A|$ bits, $|A|$ denoting the number of elements in A. The sequence x^n then gets described in $K(A) + \log |A|$ bits. We may now ask for a set $\hat{A}$ for which $K(\hat{A})$ is minimal subject to the constraint that for an increasing length sequence x^n, $K(\hat{A}) + \log|\hat{A}|$ agrees with the Kolmogorov complexity $K(x^n)$ to within a constant not depending on n. The set $\hat{A}$, or its defining program, may be called Kolmogorov's *minimal sufficient statistic* for the description of x^n. The bits describing $\hat{A}$ are then the 'interesting' bits in the program (code) for x^n while the rest, about $\log|\hat{A}|$ in number, are non-informative noise bits. However, notice that we cannot get the informative part $\hat{A}$ without considering the entire code length for x^n. This definition is a bit vague if we have only a fixed-length sequence x^n—just as for the Kolmogorov complexity itself because of its dependence on the particular universal computer being used. For a fixed but large n we may then require the above equality to hold to within a 'small' constant.

A more refined variant of the above ideas was discussed in a recent paper by Vitányi and Li [6], for the purpose of elucidating what the authors call an 'ideal' MDL principle and its relationship with a similar Bayesian principle, where a universal prior is taken to describe the optimal model, represented here by the set $\hat{A}$. Both developments are impractical because of the non-computability of the Kolmogorov complexities involved. Their value is in shedding light on the central issues involved, the latter development, in particular, demonstrating the limitations of both the ideal MDL principle and the Bayesian approach; see also [7].

To fix the notations let $\mathcal{M}_k = \{f(x^n;\theta)\}$ be a parametric class of densities as models, where $\theta = \theta_1, \ldots, \theta_k$ is a parameter vector ranging over a subset Θ^k of the k-dimensional Euclidean space, and $x^n = x_1, \ldots, x_n$ denotes a data sequence of length n. Further, let $\mathcal{M} = \cup_k \mathcal{M}_k$ be the union of the classes over all k, in which case the parameters range over $\Theta = \cup_k \Theta^k$.

We are interested in constructing a single density function $\hat{f}$ for data sequences y^n with the property: if a member $f(\cdot;\theta)$ in the class $\mathcal{M}_k$ generates a data sequence x^n (by sampling), then $\hat{f}(x^n)$ should be as close to $f(x^n;\theta)$ in a certain sense as possible, no matter which θ you pick. The idea, then, is to construct a universal model analogous to a universal computer or programming language in the algorithmic theory of information, which is capable of implementing any special program. Since the sequence x^n and the model class are all we have, it is clear that $\hat{f}$ must one way or another incorporate an estimate of the parameter θ in the data generating density. Let $\hat{\theta}(x^n)$ denote the maximum likelihood estimator, which we also let denote the estimate itself. This estimator clearly provides a shorter 'ideal' code length $-\log f(x^n;\hat{\theta}(x^n))$ for the data sequence x^n than any other model in the class. The term 'ideal' is used because a real code length has to be a positive integer. However, we can always construct a proper closely related code length $\lceil -\log f(x^n;\hat{\theta}(x^n))\delta^n \rceil$ for data quantized to a precision δ, which then gets eliminated in the comparisons.

Although the density function $f(z^n;\hat{\theta}(x^n))$ defines a proper model for all data sequences, including the observed one x^n, it will not satisfy the requirements of universality, because it depends on the particular sequence x^n and would poorly imitate density functions where θ differs a lot from $\hat{\theta}(x^n)$. We can fix the problem by a normalization process, which is quite similar to that in the algorithmic theory of complexity when the programs are required to satisfy the prefix property, see e.g. [8],

$$\hat{f}(x^n) = \frac{f(x^n;\hat{\theta}(x^n))}{\int_{\hat{\theta}(y^n)\in\Omega} f(y^n;\hat{\theta}(y^n))\,\mathrm{d}y^n}, \tag{1}$$

where Ω denotes a subset of the estimates that makes the integral finite. Moreover, that set ought to be small but easy to define in case it has to be communicated to the imagined decoder. Depending on the application we may make Ω depend on the observed values of $\hat{\theta}(x^n)$, in which case $\hat{f}(x^n)$ no longer would be a density function without an adjustment, or take it as all of Θ^k. We have more to say about the selection of such sets later. Quite interestingly this normalized maximum likelihood (NML) model solves the following minimax problem due to Shtarkov [9],

$$\min_q \max_{x^n} \log \frac{f(x^n;\hat{\theta}(x^n))}{q(x^n)}. \tag{2}$$

In words, it is the unique density function whose ideal code length exceeds the ideal optimal code length $-\log f(x^n;\hat{\theta}(x^n))$ by the least amount for the worst case data sequence.

Consider the identity

$$\hat{f}(x^n) \equiv f(x^n, \hat{\theta}(x^n)) = f(x^n \mid \hat{\theta}(x^n))g(\hat{\theta}(x^n)), \tag{3}$$

where the first factor is the conditional and the second the marginal density on $\hat{\theta}(x^n)$. The reader should not confuse the conditional density function $f(x^n|\hat{\theta}(x^n))$, where the range is the set of sequences y^n such that $\hat{\theta}(y^n) = \hat{\theta}(x^n)$, with the previously defined maximized likelihood $f(x^n;\hat{\theta}(x^n))$, which as a function of x^n is not a density function at all. Taking the negative logarithms we have the decomposition

$$-\log \hat{f}(x^n) = -\log f(x^n \mid \hat{\theta}(x^n)) - \log g(\hat{\theta}(x^n)), \tag{4}$$

which is very much like the decomposition in Kolmogorov's minimal sufficient statistic $g(\hat{\theta}(x^n))$, playing the role of $\hat{A}$ and the first term the role of the code length $\log|\hat{A}|$ for the non-informative noise. To see that this is not just a superficial resemblance, consider first the important exponential family of densities for which $\hat{\theta}(x^n)$ is a sufficient statistic in the sense of probability theory. Then the likelihood function factors as

$$f(y^n;\theta) = h(y^n)p(\hat{\theta}(y^n);\theta), \tag{5}$$

where $h(y^n)$ does not depend on the parameters. Comparing this with the factorization into the conditional and the marginal

$$f(y^n;\theta) = f(y^n \mid \hat{\theta}(y^n);\theta)q(\hat{\theta}(y^n);\theta), \tag{6}$$

we can identify $h(y^n)$ with the conditional density function, which leaves $p(\hat{\theta}(y^n);\theta) = q(\hat{\theta}(y^n);\theta)$ for the marginal.

Since integration of $f(y^n|\hat{\theta}(y^n)) = h(y^n)$ for each fixed $\hat{\theta}(y^n) = \hat{\theta}$ over y^n yields unity, the denominator in Equation (1) can be written as

$$C_k(n) = \int_{\hat{\theta}\in\Theta} p(\hat{\theta};\hat{\theta})\,\mathrm{d}\hat{\theta}. \tag{7}$$

Writing further $p(\hat{\theta};\hat{\theta}) = p(\hat{\theta})$ for simplicity, we then get

$$\hat{f}(x^n) = \frac{f(x^n;\hat{\theta}(x^n))}{p(\hat{\theta}(x^n))}\frac{p(\hat{\theta}(x^n))}{C_k(n)}. \tag{8}$$

This shows that the marginal density in Equation (3) is given by what we call a 'canonical' prior $g(\hat{\theta}) = p(\hat{\theta})/C_k(n)$, while the first factor is just the density $h(x^n)$ of the completely non-informative noise. We mention in passing that the canonical prior for the normal models coincides with the Jeffreys prior, but they are different in general. Notice, too, that there is a 'local' normalization in the first factor, necessary for the resulting two-part code to be complete: Once the ML estimate is encoded with the ideal code length $-\log g(\hat{\theta}(x^n))$ we already have information about the remaining data sequence to be encoded, namely, that its range is only the subset of sequences y^n such that $\hat{\theta}(y^n) = \hat{\theta}(x^n)$ and that it can be encoded with the shorter ideal code length, given by the normalized first factor, than $-\log f(x^n;\hat{\theta}(x^n))$.

For general classes of models $\mathcal{M}_k$ the ML estimator $\hat{\theta}(x^n)$ is not a sufficient statistic. However, for most of the usual model classes these estimates converge to the parameter of the data generating density in the mean square as $n \to \infty$ and even almost surely. What this means is that the ML estimator is asymptotically sufficient, and the density $g(\hat{\theta}(x^n))$ has for large n just about all the information about the data generating density while the factor $f(x^n|\hat{\theta}(x^n))$ has very little.

The factorization (8) is vaguely reminiscent of the numerator in the Bayes formula with g as the prior and $f(x^n;\theta)/p(\theta)$ as the likelihood function. However, for a fixed θ this does not integrate to unity and cannot be a density function at all. Neither is the product in (8) with $\hat{\theta}$ replaced by θ a joint density function on the data and the parameters and it is not proportional to a posterior. We may then try to view $1/C_k(n)$ as a uniform prior for the likelihood $f(x^n;\theta)$. However, for such an interpretation the posterior will be

$$f(\theta|x^n) = f(x^n;\theta)\bigg/\int_\Omega f(x^n;\theta)\,\mathrm{d}\theta,$$

whose maximum $f(\hat{\theta}(x^n) \mid x^n)$ is not $\hat{f}(x^n)$. It is clear that the MDL principle and stochastic complexity, defined by the universal density function $\hat{f}(x^n)$ in whatever form it is written, represent a quite different philosophy for model selection than the one underlying the Bayesian methods.

In [10] we evaluated the denominator $C_k(n)$ for 'smooth' model classes and obtained the following sharp formula for the ideal code length of the normalized maximum likelihood, whether or not $\hat{\theta}(x^n)$ is a sufficient statistic,

$$-\log\hat{f}(x^n) = -\log f(x^n;\hat{\theta}(x^n)) + \frac{k}{2}\log\frac{n}{2\pi} + \log\int_\Omega\sqrt{|I(\theta)|}\,\mathrm{d}\theta + o(1), \tag{9}$$

where $I(\theta)$ denotes the following Fisher information type of matrix

$$I(\theta) = \lim_{n\to\infty} -n^{-1}\left\{E\frac{\partial^2\log f(x^n;\theta)}{\partial\theta_i\partial\theta_j}\right\},$$

the expectation being with respect to the model $f(x^n;\theta)$; the remaining term $o(1)$ converges to zero as n grows. The convergence for many model classes is fast, like $O(1/n)$. The exact conditions required for Equation (9) to hold are listed in [10]. Weaker conditions, in essence requiring only that the estimates of the parameters satisfy the central limit theorem, are needed to prove that the right-hand side defines a density function.

The NML density function $\hat{f}(x^n)$ as a solution to Shtarkov's minimax problem appears to be too weak to qualify it as a universal model, which is supposed to provide the shortest code length for the data obtainable with the model class. However, there is an extension of Shannon's noiseless coding theorem [3], stating that the mean $-E_\theta\log\hat{f}(x^n)$ approaches the entropy of the data-generating process $f(x^n;\theta)$ at the fastest possible rate, no matter which process in the class you pick, except for a set of the parameters of measure zero. Moreover, this length is also asymptotically the shortest possible for all typical sequences, generated by almost all the models in the class [11]. Because of such theorems the name 'stochastic complexity' for the code length in Equation (9) seems justified. The important terms, other than the first, which represent the code length needed for the description of the informative part, the optimal model, we call the 'model complexity' or 'parametric complexity', to emphasize that the models are in this case parametric. A failure to understand the qualifications in the concept of 'shortest code length' has sometimes led to attempts to invalidate the MDL principle, see [12], by deliberately assigning a shorter codeword to a more complex model of the two models compared, the complexity measured for instance by the number of parameters. This, of course, is possible but only in a set of models of measure zero: there are really not enough short code words around to specify more than an ignorable set of models, and hence such devices pose no threat to the principle.

The significance of the decomposition (4) stems from the requirement that $\hat{\theta}(x^n)$ is close to the parameter θ in the data-generating model, which is not automatically satisfied. For instance, if k is taken to be equal or close to n so that we are trying to fit too complex a model for the given amount of data, the required condition is certainly not satisfied. We might then say, to use the terminology in [6], that we are trying to fit an inadmissible model, one that is not 'typical', even though we have here no prior distribution on the parameters to judge typicality. Instead, we do it in terms of the convergence properties of the estimates $\hat{\theta}(x^n)$ as stated. This issue becomes relevant when we try to obtain a universal model for the class $\mathcal{M}$. If we repeat the previous procedure and take the ML estimate $\hat{k} = \hat{k}(x^n)$ of the number of parameters together with the ML estimates of the parameters $\hat{\theta}^{\hat{k}}(x^n) = \hat{\theta}_1(x^n), \ldots, \hat{\theta}_{\hat{k}}(x^n)$ themselves we run into trouble. For most cases of interest the ML estimate $\hat{k}$ would be n or a number close to n and the subsequent estimates of the parameter values would not be good enough to ensure the optimality of the resulting code length for the data.

One way out is to define $\hat{k}(x^n)$ as the value which minimizes the stochastic complexity $-\log \hat{f}(x^n; k)$, where $\hat{f}(x^n; k) = \hat{f}(x^n)$ is the NML density for the class $\mathcal{M}_k$ in Equation (1). This estimate can be shown to be consistent in that it converges in many cases of interest to the number of parameters in the data-generating density both in probability and almost surely. Now we can repeat the procedure above and define

$$\hat{f}(x^n) = \frac{\hat{f}(x^n; \hat{k}(x^n))}{\sum_{k \le n} \int_{\hat{k}(y^n)=k} \hat{f}(y^n; \hat{k}(y^n))\, \mathrm{d}y^n}.$$

Writing

$$c_n(k) = \int_{\hat{k}(y^n)=k} \hat{f}(y^n; \hat{k}(y^n))\, \mathrm{d}y^n,$$

which being a probability does not exceed unity, and $\pi_n(k) = c_n(k)/\sum_{j \le n} c_n(j)$ we get analogously with Equation (8) the decomposition

$$\hat{f}(x^n) = \frac{\hat{f}(x^n; \hat{k}(x^n))}{c_n(\hat{k}(x^n))} \pi_n(\hat{k}(x^n)). \tag{10}$$

The probabilities $c_n(k)$ appear to be difficult to compute in general. However, we can obtain some information about the behavior of these numbers and hence about $\pi_n(k) = c_n(k)/\sum_{j \le n} c_n(j)$. Provided that $n \gg k$ the consistency of the estimates $\hat{k}(x^n)$ implies a high probability for the event that $\hat{k}(x^n) = k$. Since $\hat{f}(y^n; k)$ integrates to unity the numbers $c_n(k)$ will be close to unity. Hence, for such values of k and n the ideal code length $-\log \pi_n(\hat{k}(x^n))$ is nearly constant. In the other extreme, where $\hat{k}(x^n) = n$, the sequences x^n for many model classes are in one-to-one correspondence with the ML estimates $\hat{\theta}(x^n)$, which means that again $c_n(n) = 1$ or close to it. Hence, it seems that $c_n(k)$ does not deviate much from unity for any value of k, which means that $\pi_n(k)$ is close to $1/n$. Hence the MDL criterion for selecting a model class from among the collection $\{\mathcal{M}_k\}$ becomes to a good approximation simply

$$\min_k \{-\log \hat{f}(x^n; k) + \log n\}, \tag{11}$$

where the first term is given by the formula (9). The minimized ideal code length will then be close to $-\log \hat{f}(x^n)$ in Equation (10) and the optimal or near-optimal model is given by the parameters $\hat{\theta}_1(x^n), \ldots, \hat{\theta}_{\hat{k}(x^n)}(x^n)$.

For the important linear Gaussian-regression problem we can evaluate the integral in formula (9) and calculate the term $o(1)$ as accurately as desired. In the following example we describe the problem and give the result, which for small data sets provides a superior model selection criterion. A full derivation can be found in [13]. The first derivation of the exact formula for the NML density function in this special case with a different region of integration was reported in [14].

EXAMPLE. Consider the set of normal distributions $f(y; \mu, \tau)$ with variance τ and the mean written as a linear combination of a variable number of regressor variables thus $\mu = \beta_1 x_1 + \ldots + \beta_k x_k$. This is generalized to sequences $f(y^n; \mu^n, \tau)$ by independence, where $\mu^n = \mu_1, \ldots, \mu_n$, and $\mu_t = \beta_1 x_{1t} + \ldots + \beta_k x_{kt}$. Let $\theta = (\beta, \tau) = (\beta_1, \ldots, \beta_k, \tau)$ denote all the parameters. Let $\hat{\beta}_i(y^n)$ and

$$\hat{\tau}(y^n) = \frac{1}{n} \sum_t \left(y_t - \sum_i \hat{\beta}_i(y^n) x_{it} \right)^2$$

be the ML estimates. We take the set Ω in Equation (1) as $\Omega = \{\tau \ge \tau_0\} \times \{\beta' S \beta \le R\}$, where τ_0 and R are two new parameters such that Ω includes the ML estimates.

The ML estimates $\hat{\beta}(y^n)$ are normally distributed with mean β and covariance $S = (\tau/n)X'X$, where $X' = \{x_{it}\}$, while the estimates $n\hat{\tau}(y^n)/\tau$ have the χ^2-distribution with $n - k$ degrees of freedom. Moreover, $\hat{\beta}(y^n)$ and $\hat{\tau}(y^n)$ are independent and together they are minimally sufficient. By the technique in [13], outlined above and illustrated in detail in the case of hypothesis testing in the next section, we get the criterion

$$\begin{aligned} &-\log \hat{f}(y^n; X, k) \\ &\quad = \frac{n-k}{2} \log \hat{\tau} + \frac{k}{2} \log(n\hat{R}) \\ &\qquad + \frac{n-k-1}{2} \log \frac{n}{n-k} - \frac{k+3}{2} \log k + \ldots, \end{aligned}$$

where only the terms that depend on k in a relevant manner are retained. We also selected the optimal values for the parameters $R = \hat{R} = \hat{\beta}' S \hat{\beta}$ and $\tau_0 = \hat{\tau}$. We omit the code lengths required to encode these two numbers, because they are much shorter than the rest and do not depend heavily on k. This criterion can be minimized over k in order to obtain the first k most important regressor variables. We have then assumed that these have been sorted by declining importance, for instance, by a 'greedy' algorithm. Actually, a more complete calculation gives a criterion which makes such sorting unnecessary [15].

We conclude this section with a brief account of other means to construct universal representatives for model classes. One of these is the so-called Jeffreys' mixture

$$f_\pi(x^n) = \int_\Theta f(x^n \mid \theta)\, \mathrm{d}\pi(\theta), \tag{12}$$

where π is (a generalization of) the Jeffreys prior

$$\pi(\theta) = \frac{|I(\theta)|^{1/2}}{\int_{\eta \in \Theta} |I(\eta)|^{1/2}\, \mathrm{d}\eta}.$$

At least for classes of independent processes that satisfy suitable smoothness conditions [16], the negative logarithm $-\log f_\pi(x^n)$ agrees with Equation (9). Other priors than Jeffreys' can also be used, which may have computational advantage, although then the model complexity will no longer be independent of the data-generating model. Such a 'prior' actually need not be interpreted as representing prior knowledge in the Bayesian sense; rather, the entire integral may be viewed as a convex linear combination or, more accurately, functional of the members in the model class and hence simply a mathematical device to obtain a universal representative. Even this construct has a counterpart in the algorithmic theory of complexity, although there is nothing like the Jeffreys prior in that theory. The important connection between the lower bound on the code length mentioned above and the channel capacity has been revealed by Merhav and Feder [17], where, moreover, it was shown that also the mixture densities reach the lower bound except for parameters in a vanishing set.

Another important technique, which requires few conditions and is hence widely applicable, is to apply a predictive algorithm. This amounts to fitting a model, specified by the parameters $\theta = \theta_1, \ldots, \theta_k$, including k, to the 'past' data $x_1, x_2, \ldots, x_t$ for increasing $t = 0, 1, \ldots$ and summing up the accumulated ideal code lengths obtained by the conditionals $-\log f(x_{t+1} \mid x^t; \hat{\theta}^{\hat{k}(x^t)}(x^t))$ to give the code length for the entire data sequence. Clearly, the data will have to be ordered and the result will depend on the order. However, one can find a locally optimal order by a suitable 'greedy' algorithm [3], which then will give an order-independent predictive code length for the data. We note that the maximum likelihood estimates of the parameters are not always the best. Further, the updates of the past model estimates need not be made for every new datum x_t.

We conclude this section by mentioning that a two-part code length like $L_2(x^n) = -\log f(x^n; \hat{\theta}(x^n)) - \log w(\hat{\theta}(x^n))$, where $w(\theta)$ is a prior, does not define a universal density function for any model class. The reason is that $2^{-L(x^n)}$ integrates to a number less than unity, so that such a code is incomplete and hence inefficient. However, for very large data sets the loss of efficiency decreases and the code length $L_2(x^n)$ is justified as an approximation of the more efficient universal code lengths given.

3. HYPOTHESIS TESTING

We begin by outlining briefly the Neyman–Pearson approach to hypothesis testing. Consider first a test between two simple hypotheses $f_0(x)$ and $f_1(x)$, where x may well be a string of data points in a set X. The null hypothesis $f_0(x)$ will be rejected and the opposite hypothesis accepted on a level ϵ, say 5 per cent, if the data fall in a subset S of X, called the 'critical' region, such that the probability $P_0(S)$ of S under the null hypothesis is ϵ. The two density functions are (evidently) best separated by a test where the set S is so selected that under the opposite hypothesis $f_1(x)$ it has the maximum probability $P_1(S)$. Since ϵ is the probability of error when $f_0(x)$ generates the data and $1 - P_1(S)$ the error probability when $f_1(x)$ generates the data, the boundary of the critical region S separates the two hypotheses with the smallest total error (when the level ϵ is adhered to). By the Neyman–Pearson lemma the best set S consists of all points such that the ratio $f_1(x)/f_0(x)$ exceeds a number α_ϵ, selected such that level ϵ results. In this case the probability $P_1(S)$, called the 'power', serves as an excellent measure of the degree of confidence we can have in the test result.

Such a happy state of affairs regarding both the choice of the critical region and the assessment of the confidence breaks down to an extent when the hypotheses are composite. A common case is one where we have a parametric class of densities $\mathcal{M} = \{f(x;\theta)\}$ for θ ranging over a subset Θ of the k-dimensional Euclidean space. The null hypothesis $\mathcal{M}_0$ consists of a subset such that one or more of the parameters have fixed values and the alternative hypothesis $\mathcal{M}_1$ consists of the remaining density functions in $\mathcal{M}$. Now, even in the special case where the null hypothesis is taken as a single density function in the class, say $f(x;\theta_0)$, the best critical region and the power will be functions of θ. In some cases a single best critical region exists, which then is said to give the 'uniformly most powerful' test on the chosen level, but the power still will depend on θ, which makes the assessment of the degree of confidence a difficult issue.

The MDL principle reduces the test with composite hypotheses to the test with simple ones, which will be taken as universal models representing the two model classes. We mention in passing that if we take the universal models as the NML models, then the most powerful critical region so determined will automatically agree with the uniformly most powerful test whenever one exists. This is not of great interest to us, because we will determine the critical region in terms of the representatives.

Consider a composite null hypothesis $\mathcal{M}_0 = \{f(x^n;\theta_0) : \theta \in \Theta_0\}$ and a composite alternative hypothesis $\mathcal{M}_1 = \{f(x^n;\theta) : \theta \in \Theta_1\}$ consisting of the remaining models in the class $\mathcal{M}$. The null hypothesis is represented by the normalized maximum-likelihood density function, and for the moment the representative $f_1(x^n) = g(x^n)$ of the alternative hypothesis is kept free. The null hypothesis will be rejected for x^n in the most powerful critical region S_ϵ defined as follows:

$$S_\epsilon = \{x^n : g(x^n)/\hat{f}_0(x^n) \geq \alpha_\epsilon(g)\} \qquad (13)$$

$$P(S_\epsilon \mid f_0) = \int_{y^n \in S_\epsilon} f_0(y^n)\, \mathrm{d}y^n = \epsilon \qquad (14)$$

and accepted otherwise. Here, $\alpha_\epsilon(g)$ is the number such that the desired level ϵ for the test results is reached, which makes it a function of the representative g. The Neyman–Pearson type of critical region S_ϵ is most appropriate when the complement of this set contains the strings with the largest densities $f_0(y^n)$, because then these strings could be called ϵ-typical. After all, we do not want to reject the null hypothesis if a data string is ϵ-typical but still falls in the set S_ϵ.

Let

$$P_\theta = \int_{y^n \in S_\epsilon} f(y^n; \theta)\, \mathrm{d}y^n$$

denote the power as a function of θ. Clearly, in order for the function $g(x^n)$ to represent the composite alternative class well it should not only be close to the optimal model of the observed data specified by the ML estimate $\hat{\theta}(x^n)$, but it should also have power

$$P_g = \int_{y^n \in S_\epsilon} g(y^n)\, \mathrm{d}y^n,$$

which is not too far from any of the powers $P_{\hat{\theta}(x^n)}$. This leads to the condition

$$\min_g \max_{\theta \in \Theta_1} |P_\theta - P_g|, \tag{15}$$

which is satisfied for

$$P_g = \left(\epsilon + \max_{\theta \in \Theta_1} P_\theta\right)/2. \tag{16}$$

The condition (16) does not, of course, pin down the entire function. It just prevents a representative from putting too much of its probability mass in the critical region and hence poorly representing the alternative class. The question of interest which arises is how we reconcile the NML representative with the requirement of the optimal minmax power. Rather than studying this question in general cases we consider it in tests on the special but important Gaussian families, where the sufficiency of the ML estimates simplifies things.

As a final remark we add that we do not like either orthodox or Bayesian hypothesis testing, and our discussion here is somewhat tentative, guided by an attempt to satisfy requirements of common sense and logic. Of special interest to us is to show that all the information in the data is not captured by the posteriors nor Bayes factors, and that the Bayesians' methods can be improved on no matter how one interprets the results.

3.1. Testing the mean with fixed variance

Consider the normal density function for the data sequence x^n:

$$f(x^n; \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}}\, \mathrm{e}^{-1/2\sigma^2 \sum_{t=1}^n (x_t-\mu)^2} \tag{17}$$

$$= \frac{1}{\sqrt{n}(2\pi\sigma^2)^{(n-1)/2}}\, \mathrm{e}^{-1/2\sigma^2 \sum_{t=1}^n (x_t-\bar{x})^2} \times \frac{1}{\sqrt{2\pi\sigma^2/n}}\, \mathrm{e}^{-(n/2\sigma^2)(\bar{x}-\mu)^2}, \tag{18}$$

where $\bar{x} = (\sum_t x_t)/n$ is the ML estimate of the mean. We consider the simple null hypothesis, known mean μ and variance σ^2, which without loss of generality may be taken as $\mu = 0$ and $\sigma^2 = 1$. The null hypothesis then consists of the singleton class $\mathcal{M}_0 = \{f(x^n; 0, 1)\}$.

Since with a finite and often relatively small amount of data there is little point in letting the alternative hypothesis consist of all the non-zero mean values, which would just make the class far too large and hence the test less sharp, we take it as follows: $\mathcal{M}_1 = \{f(x^n; \mu, 1) : 0 < |\mu| \le c/\sqrt{n}\}$, where c is a constant to be determined later. Write the full range of the mean parameter as $\Theta = |\mu| \le c/\sqrt{n}$ and that of the alternative hypothesis as Θ_1. The reason why the range of the alternative hypothesis shrinks with growing n in the manner shown is to make the test sharp, as will become clear in a moment.

The representative of the null hypothesis class is clearly its only member. We compute the NML representative of the class $\mathcal{M}_1$ by taking advantage of the fact that $\bar{x}$ is a sufficient statistic for the family, as shown by the factorization in Equation (18) for $\sigma = 1$. The first factor gives the conditional $f(x^n \mid \bar{x}, \mu, 1)$, which is independent of μ. By putting $\mu = \bar{x}$, as required in the NML model, we get the normalizing integral by integrating first over all y^n such that their arithmetic mean is a fixed number $\bar{x}$, which gives the value $f_{\bar{x}}(\bar{x}) = \sqrt{n/(2\pi)}$ (= the second factor). Integrating then over $\bar{x}$ in the range Θ_1 gives the normalizing integral as

$$C_1 = \int_{x^n : \bar{x} \in \Theta_1} f(x^n; \bar{x}, 1)\, \mathrm{d}x^n = \frac{2c}{\sqrt{2\pi}}$$

and the NML density itself representing the alternative model class

$$\hat{f}_1(x^n) = \frac{1}{2c(2\pi)^{(n-1)/2}}\, \mathrm{e}^{-\frac{1}{2}\sum_{t=1}^n (x_t-\bar{x})^2}.$$

The ratio of the densities representing the two hypotheses becomes that of the induced densities of the sufficient statistic $z = \sqrt{n}\bar{x}$:

$$\frac{f(x^n; 0, 1)}{\hat{f}_1(x^n)} = \frac{f_0(z)}{f_1(z)} \equiv \frac{(2\pi)^{-1/2}\, \mathrm{e}^{-z^2/2}}{1/(2c)}. \tag{19}$$

Hence, the test is reduced to comparing the unit normal density with the uniform density over $[-c, c]$, and the critical region is given by the union of $(-\infty, -z_\epsilon]$ and $[z_\epsilon, \infty)$, where z_ϵ is a point such that the combined probability mass of the standard normal distribution is ϵ. Clearly, for the test to be meaningful we must have $c > z_\epsilon$. The power of this test is $1 - z_\epsilon/c$.

The maximum $P_{\max}$ of the power function P_μ in $[-c, c]$ occurs very nearly at the mid point $\mu_m = (z_\epsilon + c)/2$, and it is given by the probability mass of the normal density function $f(z; \mu_m, 1)$ falling in the intervals $[z_\epsilon, c]$ and $[-c, -z_\epsilon]$. We may ask for the value of c that makes the power of the test $1 - z_\epsilon/c$ to satisfy the minmax condition (16). This gives the equation $1 - z_\epsilon/c = (\epsilon + P_{\max})/2$ with the solution

$$c = \frac{2z_\epsilon}{2 - (\epsilon + P_{\max})}. \tag{20}$$

For small values of ϵ, which are of the greatest interest, we get the value $c \simeq 2z_\epsilon$ and the optimal power about 1/2.

The two density functions of $z = \sqrt{n}\bar{x}$ in the ratio (19) together with the size ϵ and the number c in (20) completely characterize the test. In this the test data x^n specify the generating optimal model $\mu = \bar{x}$, which is represented by the universal model $\hat{f}_1(x^n)$, whose typical sequences are the set of all sequences of length n such that $\sqrt{n}\bar{x} \in [-c, c]$. The observed ratio (19) of the two densities can be used as a measure of how much more likely the observed test data have been generated as a typical sequence by the null hypothesis than by the alternative one. Further, the number $1 - (\epsilon + P_{\max})/2$ may be interpreted as the probability of accepting the null hypothesis when in fact the data have been generated by some model in the alternative hypothesis.

The Bayesian approach to tests is based on comparing the posterior probabilities $P(\Theta_0 \mid x)$ and $P(\Theta_1 \mid x)$, given the observed data x, where Θ_0 and Θ_1 are the ranges of the parameters in the two hypotheses. These are computed by assuming prior probabilities for the two hypotheses, taken here as $1/2$ (any other choice causes no essential change), and prior density functions $g_0(\theta)$ and $g_1(\theta)$ on the ranges. The case of a simple null hypothesis will cause trouble, because $P(\{\theta_0\} \mid x)$ would be zero, but by interpreting $g_0(\theta_0)$ as a probability, the difficulty can be handled. In the current example the test leads to considering the ratio $B = f_0(z)/f_w(z)$, called the Bayes factor [4], where $f_0(z)$ and z are the same as in (19) but

$$f_w(z) = \frac{1}{\sqrt{2\pi}} \int_{-c}^{c} \mathrm{e}^{-(z-\mu)^2/2} w(\mu)\, \mathrm{d}\mu \tag{21}$$

for a prior w. This is restricted to a class of densities such that $w(|\mu|)$ is a symmetric non-increasing function with the maximum at 0. It is easily seen that for any such prior, including the uniform one, $f_w(|z|)$ is a strictly declining function with its maximum at 0.

We may view $f_w(z)$ as a representative of the alternative hypothesis rather than regard it as a posterior density and similarly view $f_0(z)/f_w(z)$ as a ratio of two densities in our interpretation. But then, because $f_w(|z|)$ is a strictly declining function while the NML density, when reduced to a function of z, is the uniform density function $1/(2c)$, the power of $f_w(|z|)$ is strictly less than the power obtained with the NML density for every ϵ and every c indicating a poorer separability of the two hypotheses and a weaker test.

A similar conclusion results even with the Bayesian interpretation, which instead of the idea of a power argues directly in terms of the ratio $f_0(z)/f_w(z)$. Indeed, for an efficient test such a testing ratio should be large for values of z where the null hypothesis is accepted and small otherwise. However, for any prior of the kind given above, $f_0(z)/f_w(z) > f_0(z)/f_1(z)$ for values of z such that the null hypothesis is rejected, while the opposite inequality holds for other values (except at the boundary, where equality holds), and again the Bayesian odds ratio gives a poorer test. Neither does the conclusion change if we assess the test in terms of the error probabilities interpreted the Bayesian way.

3.2. Testing the mean with free variance

A more general test for the mean in the Gaussian family is to let the variance be a free parameter in both hypotheses. Again the known mean in the null hypothesis may be taken as zero without loss of generality. The two hypotheses are specified by the model classes

$$\mathcal{M}_0 = \{ f(x^n; 0, \sigma^2) : \sigma_0 \le \sigma \le \sigma_1 \} \tag{22}$$

$$\mathcal{M}_1 = \Big\{ f(x^n; \mu, \sigma^2) : \sigma_0 \le \sigma \le \sigma_1,\ 0 < |\mu| \le c\sqrt{s^2/(n-1)} \Big\}, \tag{23}$$

where $s^2 = \sum_t (x_t - \bar{x})^2/n$ and the two positive limits for the variance parameter can be anything; they will cancel. Taking the limits for μ as indicated is done for the same reason as in the preceding section and will also be obvious in a moment.

We calculate first the NML density function for the null hypothesis. For this we need the ML estimate of the variance

$$\hat{\sigma}^2 = \frac{1}{n} \sum_i x_i^2 = s^2 + \bar{x}^2 = s^2 \left[1 + \left(\frac{\bar{x}}{s} \right)^2 \right]. \tag{24}$$

Then

$$f(x^n; 0, \hat{\sigma}^2) = \frac{1}{(2\pi e \hat{\sigma}^2)^{n/2}} \tag{25}$$

$$= \frac{1}{(2\pi e s^2)^{n/2}} \frac{1}{(1 + t^2/(n-1))^{n/2}} \tag{26}$$

$$= \frac{1}{(2\pi e s^2)^{n/2}} \frac{\sqrt{(n-1)\pi}\,\Gamma(\frac{n-1}{2})}{\Gamma(n/2)} \mathcal{S}_{n-1}(t), \tag{27}$$

where Γ is the gamma-function, $t = \sqrt{n-1}\frac{\bar{x}}{s}$ and $\mathcal{S}_{n-1}(t)$ is Student's *t-distribution* of $n-1$ degrees of freedom.

Write $\tau = \sigma^2$ for simplicity and consider the factorization of the normal density function in terms of the sufficient statistic $\hat{\tau} = \hat{\sigma}^2$:

$$f(x^n; 0, \tau) = h(x^n) \frac{(n/2)^{n/2}}{\tau^{n/2}\Gamma(n/2)} \hat{\tau}^{n/2-1}\, \mathrm{e}^{-\frac{1}{2} n\hat{\tau}/\tau}.$$

The first term is the conditional $f(x^n \mid \hat{\tau}; 0, \tau)$, given $\hat{\tau} = \hat{\tau}(x^n)$, and the second the marginal on the sufficient statistic. Just as in (7) we get the normalizing integral

$$J_0 = \int_{\tau_0}^{\tau_1} f(x^n; 0, \hat{\tau}(x^n))\, \mathrm{d}x^n = \frac{(n/2\,\mathrm{e})^{n/2}}{\Gamma(n/2)} \ln \frac{\tau_1}{\tau_0},$$

where τ_0 and τ_1 are the bounds σ_0^2 and σ_1^2 for the variance. The *NML* density function is then given by

$$\hat{f}_0(x^n) = \frac{1}{J_0(2\pi\, \mathrm{e}s^2)^{n/2}} \frac{\sqrt{(n-1)\pi}\,\Gamma(\frac{n-1}{2})}{\Gamma(n/2)} \mathcal{S}_{n-1}(t). \tag{28}$$

We next compute the NML density function for the alternative hypothesis. With the same technique as in the previous case by first factoring the density function $f(x^n; \mu, \tau)$ into the conditional and the marginal on the pair of sufficient statistics $(\bar{x}, s^2)$ we get

$$f(x^n; \mu, \tau) = f(x^n \mid \bar{x}, s^2; \mu, \tau) g(\bar{x}, s^2; \mu, \tau), \tag{29}$$

where the first factor does not depend on μ and τ by the sufficiency of $(\bar{x}, s^2)$. Notice that in this case $\hat{\tau} \equiv s^2$ is not the same as in the previous case. The variables $\bar{x}$ and s^2 are independent, the former having a Gaussian distribution and ns^2/τ the χ^2 distribution of $n-1$ degrees of freedom. Hence, the marginal is given by

$$g(\bar{x}, s^2; \mu, \tau) = \frac{1}{(2\pi\tau/n)^{1/2}} \mathrm{e}^{-(n/2\tau)(\bar{x}-\mu)^2} \tag{30}$$

$$\times \frac{1}{2^{(n-1)/2}\Gamma((n-1)/2)} \frac{n}{\tau} (ns^2/\tau)^{(n-3)/2} \mathrm{e}^{-ns^2/2\tau}. \tag{31}$$

Putting in (29) $\mu = \bar{x}$ and $\tau = s^2$ and integrating first over x^n such that $\bar{x}$ and $\hat{\tau} \equiv s^2$ are fixed and then over these in their range we get the normalizing integral as follows:

$$J_1 = \frac{(n/2\,\mathrm{e})^{n/2}}{\Gamma((n-1)/2)\sqrt{(n-1)\pi}} 2c \ln \frac{\tau_1}{\tau_0}.$$

This gives the NML density itself as

$$\hat{f}_1(x^n) = \frac{1}{J_1(2\pi\,\mathrm{e}s^2)^{n/2}}, \tag{32}$$

and the ratio

$$\frac{\hat{f}_0(x^n)}{\hat{f}_1(x^n)} = \frac{f_0(t)}{f_1(t)} = 2c\mathcal{S}_{n-1}(t), \tag{33}$$

where again $t = \sqrt{n-1}\bar{x}/s$. We see that we have the same type of test as in the preceding subsection. The *t-distribution* is quite similar to the Gaussian, which means that all the conclusions regarding the subsequent developments on the minmax power, the choice of c and the assessment of the confidence, remain in essence unchanged. This includes the superiority of the NML representatives over the mixture densities in the Bayes factor, when they are regarded as representatives of the two hypotheses. In order to verify the last claim we calculate the mixtures in the Bayes factor by taking the prior $w(\tau) = 1/(\tau \ln(\tau_1/\tau_0))$ for the variance and letting τ_0 and τ_1 grow to infinity. This corresponds to the Bayesians' technique of 'improper' prior $1/\tau$, [4], for which the mixture integral remains bounded. For the prior $v(\mu)$ we take one such that $v(|\mu|)$ is non-increasing in the interval $[-c/\sqrt{n}, c/\sqrt{n}]$. In the extreme case where $v(\mu)$ is uniform in the given range, the joint prior $w(\tau)v(\mu)$ will be the Jeffreys prior, which has the information-theoretic optimality property that it gives the channel capacity, [17]. This is why we adopt them; the conclusions will be similar no matter which prior one selects.

With the given priors we get the mixtures representing the null and the alternative hypotheses $f_w(x^n)$ and $f_{v,w}(x^n)$ as follows:

$$f_w(x^n) = K_n(s^2)\mathcal{S}_{n-1}(t) \tag{34}$$

$$f_{v,w}(x^n) = K_n(s^2)p(t), \tag{35}$$

where

$$p(t) = \int_{-c}^{c} \mathcal{S}_{n-1}(t - \sqrt{n-1}\mu/s)v(\mu)\,\mathrm{d}\mu,$$

and the exact expression for $K_n(s^2)$ will not be needed.

These give the Bayes factor as the ratio

$$\frac{f_w(x^n)}{f_{w,v}(x^n)} = \mathcal{S}_{n-1}(t)/p(t). \tag{36}$$

Since also the density function $p(t)$ is such that $p(|t|)$ is strictly declining the power of the mixture representative of the alternative hypothesis against the mixture representative of the null hypothesis is strictly less than the power when both hypotheses were represented by the NML densities. For the same reason the test based on comparing the Bayes factor against a number α is weaker than if the Bayes factor is replaced by the ratio $f_0(t)/f_1(t)$ of the NML densities.

We conclude this paper with two comments. First, a Bayesian might take solace in the fact that the results reached depend on the interval being finite and would not hold otherwise. However, unless one insists that tests with finite intervals are invalid we are led to challenge the belief that all information resides in the posterior, which many a Bayesian seems to hold so dear.

Secondly, the Bayes factors do not quite have a satisfactory interpretation in Bayesian analysis. If the two hypotheses are given the prior probabilities π_0 and π_1, respectively, then with the further prior densities $v(\mu)$ and $w(\tau)$ one can compute the posterior probabilities α_0 and α_1. The Bayes factor

$$\frac{f_w(x^n)}{f_{w,v}(x^n)} = \frac{\alpha_0\pi_1}{\alpha_1\pi_0}$$

is meant to express the odds in favor of the null hypothesis, given the data alone; i.e. when the effects of the priors are eliminated. However, the result still depends in a very significant way on the two priors in the mixtures, and the meaning of the Bayes factor remains obscure—except in the currently given MDL-based interpretation!

REFERENCES

[1] Wallace, C. S. and Boulton, D. M. (1968) An information measure for classification. *Comput. J.*, **11**, 185–195.

[2] Rissanen, J. (1978) Modeling by shortest data description. *Automatica*, **14**, 465–471.

[3] Rissanen, J. (1986) Stochastic complexity and modeling. *Ann. Statist.*, **14**, 1080–1100.

[4] Berger, J. O. (1985) *Statistical Decision Theory and Bayesian Analysis* (2nd edn). Springer, New York.

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

[6] Vitányi, P. and Li, M. (1997) Minimum description length induction, Bayesianism, and Kolmogorov complexity. Private communication.

[7] Vovk, V. (1997) Learning about the parameter of the Bernoulli model. *J. Comput. System Sci.*, **55**, 96–104.

[8] Li, M. and Vitányi, P. (1993) *An Introduction to Kolmogorov Complexity and Its Applications*. Springer, New York.

[9] Shtarkov, Yu. M. (1987) Universal sequential coding of single messages. Translated from *Problems of Information Transmission*, **23**, 3–17.

[10] Rissanen, J. (1996) Fisher information and stochastic complexity. *IEEE Trans. Information Theory*, **IT-42**, 40–47.

[11] Dawid, A. P. (1992) Prequential analysis, stochastic complexity and Bayesian inference (with discussion). In Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. M. (eds), *Bayesian Statistics 4*, pp. 109–125. Oxford University Press, Oxford, UK.

[12] Domingos, P. (1998) Occam's two razors: the sharp and the blunt. *The 4th Int. Conf. of Knowledge Discovery and Data Mining*, August 27–31, 1998, New York, NY.

[13] Barron, A. R., Rissanen, J. and Yu, B. (1998) The MDL principle in modeling and coding. Special issue of *IEEE Trans. Information Theory*, **IT-44**, 2743–2760. To commemorate 50 years of information theory.

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

[15] Rissanen, J. (1999) MDL denoising.
http://www.cs.tut.fi/˜rissanen/
(submitted to *IEEE Trans. Information Theory*)

[16] Clarke, B. S. and Barron, A. R. (1990) Information-theoretic asymptotics of Bayes methods. *IEEE Trans. Information Theory*, **IT-36**, 453–471.

[17] Merhav, N. and Feder, M. (1995) A strong version of the redundancy-capacity theorem of universal coding. *IEEE Trans. Information Theory*, **IT-41**, 714–722.