# Ideal MDL and Its Relation To Bayesianism[1]

Paul Vitányi
CWI and University of Amsterdam,
Kruislaan 413, 1098 SJ Amsterdam, The Netherlands
Email: paulv@cwi.nl

Ming Li
Computer Science Department, University of Waterloo,
Waterloo, Ontario N2L 3G1, Canada.
Email: mli@math.uwaterloo.ca

**Abstract:** *Statistics based inference methods like minimum message length (MML) and minimum description length (MDL), are widely applied approaches. They are the tools to use with particular machine learning praxis such as simulated annealing, genetic algorithms, genetic programming, artificial neural networks, and the like. These methods select the hypothesis which minimizes the sum of the length of the description of the hypothesis (also called 'model') and the length of the description of the data relative to the hypothesis. Ideally, MDL uses shortest effective descriptions and is expressed in terms of Kolmogorov complexity. We derive Ideal MDL from first principles and explain correspondences and differences between MDL and Bayesian reasoning, and in particular why the latter is prone to overfitting and the former isn't. In particular we show that Ideal MDL can be reduced to the Bayesian approach using the universal prior distribution, provided the minimum description length is reached for those hypotheses with respect to which the data sample is individually random in the sense of Martin-Löf.*

**Keywords:** MDL,MML, Bayes rule, Kolmogorov complexity, Randomness tests.

**Area of Interest:** Minimum Encoding Length Inference Methods, Bayesian Methodology.

## 1 Introduction

It is not always realized that most of traditional statistics is about computational learning. It is always involved with algorithms to obtain the general from the particular. Not only is this approach to machine learning mathematical, but also very much applied.

A general task of statistical learning is to select the most plausible hypothesis from among all possible hypotheses in the light of experimental evidence. The classic method to do so is Bayes' rule. The problem with applying Bayes' rule is that one requires the prior probabilities of the possible hypotheses first. Unfortunately, it is often impossible to obtain these.

One way out of this conundrum is to require inference of hypotheses to be completely data driven. The MML and MDL approaches, [14, 8], embody this idea. They are usually presented as justified in and of itself by philosophical persuasion. This novel approach in statistical learning is based on minimum description length of hypotheses and data together—one way to express the so-called MDL principle. These methods, say MDL, select the hypothesis which minimizes the sum of the length of the description of the hypothesis (also called 'model') and the length of the description of the data relative to the hypothesis.

Ideally the description lengths involved should be the shortest effective description length which is asymptotically unique and objective and known as the *Kolmogorov complexity*. We rigorously derive and justify this "Ideal MDL" via the Bayesian approach using a particular prior distribution over the hypotheses (the so-called 'universal distribution').

We derive Ideal MDL from first principles and explain correspondences and differences between MDL and Bayesian reasoning, and in particular why the latter is prone to overfitting and the former isn't. In particular we show that Ideal MDL can be reduced to the Bayesian approach using the universal prior distribution, provided the minimum description length is reached for those hypotheses with respect to which the data sample is *individually random* in the sense of Martin-Löf.

While classical probability theory cannot even express the notion of 'randomness of individual objects' the theory of Kolmogorov complexity and Martin-Löf tests for randomness was explicitly developed to rigorously express the novel notion of individual randomness. Objects which are random in in this sense will satsfy *all* effective tests for randomness properties—those which are known and those which are yet unknown alike.

The MDL discussion in this paper consists for the most part of preliminary results of ongoing work. Part of this work was reported before in preliminary form in a non-conference, non-journal setting (which cannot be referred to in view of the isis review policy).

# 2 Learning With MDL

The MDL principle is an algorithmic paradigm that is widely applied. That is, it is widely applied at least in spirit; to apply it literally may run in computation difficulties since it involves finding an optimum in a exponentially large set of candidates as noted for example in [5]. Yet in some cases one can approximate this optimum, [13, 16].

A practice oriented theory like MDL, although often lacking in justification, apparently works and is used by practitioners. The MDL principle is very easy to use in some loose sense, but it is hard to justify. A user of the MDL principle does not need to *prove* that the concept class concerned is learnable, rather he needs to choose a concept that can be described shortly and without causing too many errors (and he needs to balance these two things). In this section, we will describe the MDL principle and establish its justification from first principles.

## 2.1 Bayesian Reasoning

Consider a situation in which one has a set of observations of some phenomenon, and also a (possibly countably infinite) set of hypotheses which are candidates to explain the phenomenon. For example, we are given a coin and we flip it 100 times. We want to identify the probability that the coin has outcome 'head' in a single coin flip. That is, we want to find the bias of the coin. The set of possible hypotheses is uncountably infinite if we allow each real bias in $[0,1]$, and countably infinite if we allow each rational bias in $[0,1]$.

For each hypothesis $H$ we would like to assess the probability that $H$ is the 'true' hypothesis, given the observation of $D$. This quantity, $\Pr(H|D)$, can be described and manipulated formally in the following way.

Consider a sample space $\Omega$. Let $D$ denote a sample of outcomes, say experimental data concerning a phenomenon under investigation. Let $H_1, H_2, \ldots$ be an enumeration of countably many hypotheses concerning this phenomenon, say each $H_i$ is a probability distribution over $\Omega$. The list $\mathcal{H} = \{H_1, H_2, \ldots\}$ is called the *hypothesis space*. The hypotheses $H_i$ are exhaustive and mutually exclusive.

For example, say the hypotheses enumerate the possible rational (or computable) biases of the coin. As another possibility there may be only two possible hypotheses: hypothesis $H_1$ which says the coin has bias 0.2, and hypothesis $H_2$ which puts the bias at 0.8.

Suppose we have *a priori* a distribution of the probabilities $P(H)$ of the various possible hypotheses in $\mathcal{H}$ which means that $\sum_{H \in \mathcal{H}} P(H) = 1$. Assume furthermore that for all $H \in \mathcal{H}$ we can compute the probability $\Pr(D|H)$ that sample $D$ arises if $H$ is the case. Then we can also compute (or approximate in case the number of hypotheses with nonzero probability is infinite) the probability $\Pr(D)$ that sample $D$ arises at all

$$\Pr(D) = \sum_{H \in \mathcal{H}} \Pr(D|H)P(H).$$

From the definition of conditional probability it is easy to derive **Bayes' formula**

$$\Pr(H|D) = \frac{\Pr(D|H)P(H)}{\Pr(D)}. \tag{1}$$

The prior probability $P(H)$ is often considered as the learner's *initial degree of belief* in hypothesis $H$. In essence Bayes' rule is a mapping from *a priori* probability $P(H)$ to *a posteriori* probability $\Pr(H|D)$ determined by data $D$.

Continuing to obtain more and more data, this way the total inferred probability will concentrate more and more on the 'true' hypothesis. We can draw the same conclusion of course, using more examples, by the law of large numbers. In general, the problem is not so much that in the limit the inferred probability would not concentrate on the true hypothesis, but that the inferred probability gives as much information as possible about the possible hypotheses from only a limited number of data. Given the prior probability of the hypotheses, it is easy to obtain the inferred probability, and therefore to make informed decisions. However, in general we don't know the prior probabilities. The following MDL approach in some sense replaces an unknown prior probability by a fixed 'universal' probability.

## 2.2 Minimum Description Length

A relatively recent method in statistics was developed by C. Wallace [14, 15], who called the principle involved *Minimum Message Length (MML)* and J. Rissanen [7, 8] who named it *Minimum Description Length (MDL)*. In accordance with Occam's dictum, it tells us to go for the explanation that compresses the data the most. Here we ignore the distinctions between MML and MDL and formulate the principle involved as follows.

DEFINITION 1 *Given a sample of data, and an effective enumeration of theories, the MDL principle is to select the theory which minimizes the sum of*

- *the length, in bits, of the description of the theory; and*
- *the length, in bits, of data when encoded with the help of the theory, the* stochastic complexity *[9].*

Intuitively, with a more complex description of the hypothesis $H$, it may fit the data better and therefore decreases the misclassified data. If $H$ describes all the data, then it does not allow for measuring errors. A simpler description of $H$ may be penalized by increasing the number of misclassified data. If $H$ is a trivial hypothesis that contains nothing, then all data are described literally and there is no generalization. The rationale of the method is that a balance in between seems required. However, straightforward application of the idea may lead to obviously false results in pathological cases.

## 2.3 Derivation of MDL from Bayes' Rule

Here, we present a rigorous analysis of a pure version of MDL. While this does not rigorously prove anything about applied MDL, the relations with Bayesian inference we establish for pure MDL are corroborated by empirical evidence for applied MDL as in [6, 11, 2].

How to derive MDL from first principles, *in casu* Bayes' rule as given by Equation 1? In the latter equation we are only concerned with maximizing the term $\Pr(H|D)$ over $H$. Taking the negative logarithm at both sides of the equation, this is equivalent to *minimizing* the expression $-\log \Pr(H|D)$ over $H$:

$$\begin{aligned} -\log \Pr(H|D) &= -\log \Pr(D|H) - \log P(H) \\ &\quad + \log \Pr(D). \end{aligned}$$

Since the probability $\Pr(D)$ is constant under varying $H$, this means we want to find $H$ which minimizes:

$$-\log \Pr(D|H) - \log P(H). \tag{2}$$

Because of space limitations, we refer the reader to [5] for all definitions and analysis of auxiliary notions. This way we also do not deviate from the main argument, do not obstruct the knowledgeable reader, and do not confuse or discourage the reader who is as yet unfamiliar with Kolmogorov complexity theory.

To obtain the MDL principle it suffices to replace the probabilities involved in Equation 2 by the so-called *universal probability* $\mathbf{m}(\cdot)$. The analysis of the *conditions* under which this substitution is *justified*, or, conversely, how application of MDL is equivalent to Bayesian inference using *admissible* probabilities, is deferred to the next section.

$$\begin{aligned} \log P(H) &= \log \mathbf{m}(H) + O(1), \quad (3) \\ \log \Pr(D|H) &= \log \mathbf{m}(D|H) + O(1). \end{aligned}$$

According to [5],

$$\begin{aligned} -\log \mathbf{m}(H) &= K(H) \pm O(1), \\ -\log \mathbf{m}(D|H) &= K(D|H) \pm O(1), \end{aligned}$$

where $K(\cdot)$ is the prefix complexity. Therefore, using the substitution of Equation 3 we can replace the sum of Equation 2 by the sum of the minimum lengths of effective self-delimiting programs which compute descriptions of $H$ and $D|H$. That is, we look for the $H$ that minimizes

$$K(D|H) + K(H), \tag{4}$$

which is the code-independent, recursively invariant, absolute form of the MDL principle.

We mention that the term $-\log \Pr(D|H)$ is also known as the *self-information* in information theory and the negative log-likelihood in statistics. It can now be regarded as the number of bits it takes to redescribe or encode $D$ with an ideal code relative to $H$.

In [3] (see also [5], Theorem 3.5 on p. 195) it is proved that symmetry of information holds for individual objects in the following sense

$$\begin{aligned} K(H, D) &= K(H) + K(D|H, K(H)) + O(1) \\ &= K(D) + K(H|D, K(D)) + O(1). \end{aligned} \tag{5}$$

Since it is not more difficult to describe some object if we get more conditional information, we have $K(D|H, K(H)) \leq K(D|H) + O(1)$. Thus, by Equation 5 the quantity in Equation 4 satisfies

$$\begin{aligned} K(H) + K(D|H) &\geq K(H, D) + O(1) \\ &\geq K(D) + O(1). \end{aligned}$$

For the trivial hypothesis $H_0 = D$ or $H_0 = \emptyset$ equalities hold. At first glance this would mean that the hypothesis $H_{\rm mdl}$ which minimizes the sum of Equation 4 could be set to $D$ or $\emptyset$, which is absurd in general. However, we have only derived the validity of Equation 4 under the condition that Equation 3 holds. The crucial part of our justification of MDL is to establish precisely when Equation 3 is valid, which is the main thrust of this paper.

## 2.4 The Rôle of Universal Probability

It is well known, see for example [5], that the so-called Shannon-Fano code for an ensemble of source words distributed according to probability $Q$ is a prefix code $E_Q$ with $l(E_Q(x)) = -\log Q(x)$ satisfying

$$\sum_x Q(x) l(E_Q(x)) = \min_E \{ \sum_x Q(x) l(E(x)) : E \text{ is a prefix code} \} + O(1).$$

Therefore, without any restrictions Equation 2 is equivalent to

$$E_{\Pr(\cdot|H)}(D) + E_P(H).$$

To satisfy restriction Equation 3 we *can simply assume* that the prior probability $P()$ in Bayes' rule Equation 1 is $\mathbf{m}()$. Whether this can be justified or not is a question which we address later.

However, we *cannot assume* that the probability $\Pr(\cdot|H)$ equals $\mathbf{m}(\cdot|H)$. Namely, as explained at length in Section 2.1, probability $\Pr(\cdot|H)$ may be totally determined by the hypothesis $H$. Depending on $H$ therefore, $l(E_{\Pr(\cdot|H)}(D))$ may be *vastly* different from $K(D|H)$ indeed. This holds especially for 'simple' data $D$ which have low probability under assumption of hypothesis $H$.

EXAMPLE 1 Let us look at a simple example evidencing this discrepancy. Suppose we flip a coin of unknown bias $n$ times. Let hypothesis $H$ and data $D$ be defined by:

$$\begin{aligned} H &:= \text{Probability 'head' is } 1/2 \\ D &:= \underbrace{hh\ldots h}_{n \text{ times '}h\text{'(ead)s}} \end{aligned}$$

Then we have $\Pr(D|H) = 1/2^n$ and

$$l(E_{\Pr(\cdot|H)}(D)) = -\log \Pr(D|H) = n.$$

In contrast,

$$K(D|H) \leq \log n + 2 \log \log n + O(1).$$

$\diamond$

However, the theory dealing with randomness of individual objects states that under certain conditions $-\log \Pr(D|H)$ and $K(D|H)$ are close.

First, assume that $\Pr(\cdot|\cdot)$ is a recursive function. That is, it can be computed to any required precision for each argument $D$ and conditional $H$. Then we can appeal to the following known facts, [5]. Firstly, it is known that $\mathbf{m}(D|H) \geq 2^{-K(\Pr(\cdot|H))} \Pr(D|H)$. Therefore,

$$\begin{aligned} \log \frac{\mathbf{m}(D|H)}{\Pr(D|H)} &\geq -K(\Pr(\cdot|H)) \qquad (6) \\ &\geq -K(H) + O(1). \end{aligned}$$

The last inequality arises since from $H$ we can compute $\Pr(\cdot|H)$ by assumption on $\Pr(\cdot|\cdot)$. Secondly, for $D$ which are sufficiently *random* with respect to the distribution $\Pr(\cdot|H)$ (with respect to $H$ therefore) in the sense of Martin-Löf one can show, [5], pp. 228-229, that

$$\log \frac{\mathbf{m}(D|H)}{\Pr(D|H)} \leq 0. \qquad (7)$$

The overwhelming majority of $D$'s is random in this sense because for each $H$ we have

$$\begin{aligned} \sum_D \Pr(D|H) 2^{\log \frac{\mathbf{m}(D|H)}{\Pr(D|H)}} &= \sum_D \log \mathbf{m}(D|H) \\ &\leq 1, \end{aligned}$$

since $\mathbf{m}(\cdot|H)$ is a probability distribution. For $D$'s which are random in the appropriate sense, Equations 6, 7 mean by Equation 3 that

$$\begin{aligned} K(D|H) - K(\Pr(\cdot|H)) &\leq -\log \Pr(D|H) \\ &\leq K(D|H). \end{aligned}$$

Above, by assuming that the *a priori* probability $P(H)$ of hypothesis $H$ is in fact the universal probability we obtained $-\log P(H) = \mathbf{m}(x)$. However, we do not need to make this assumption. For recursive $P(\cdot)$, we can analyze the situation when $H$ is random in the required sense with respect to $P(\cdot)$. The first inequality below holds since $\mathbf{m}()$ majorizes $P()$ this way; the second inequality expresses the assumption of randomness of $H$,

$$\begin{aligned} \mathbf{m}(H) &\geq 2^{-K(P)} P(H) \\ \log(\mathbf{m}(H)/P(H)) &\leq 0. \end{aligned}$$

Then,

$$K(H) - K(P) \leq -\log P(H) \leq K(H). \quad (8)$$

Altogether we find

$$\begin{aligned} K(D|H) &+ K(H) - \alpha(P,H) \\ &\leq -\log \Pr(D|H) - \log P(H) \\ &\leq K(D|H) + K(H), \end{aligned}$$

with $\alpha(P,H) = K(\Pr(\cdot|H)) + K(P)$ and we note that $K(\Pr(\cdot|H)) \leq K(H) + O(1)$. We call this formula the **Fundamental Inequality (FI)** because it describes the fundamental relation between Bayes' Rule and MDL, and it is left for us to interpret it.

## 2.5 Validity Range of FI

We begin by stressing that our arguments validate the bounds in Equation 9 only in case simultaneously $H$ is $P$-random and $D$ is $\Pr(\cdot|H)$-random. What is the meaning of this?

$H$ is $P$-random means that the true hypothesis must be 'typical' for the prior distribution $P$ in the sense that it must belong to all effective majorities (sets on which the majority of $P$-probability is concentrated). In the [5] it is shown that this is the set of $H$'s such that with high $P$-probability $K(H) \approx -\log P(H)$. In case $P(H) = \mathbf{m}(H)$, that is, the prior distribution equals the universal distribution, then for *all* $H$ we have $K(H) = -\log P(H)$, that is, all hypotheses are random with respect to the universal distribution..

Let us look at an example of a distribution where some hypotheses are random, and some other hypotheses are nonrandom. Let the possible hypotheses correspond to the binary strings of length $n$, and $P$ is the uniform distribution which assigns probability $P(H) = 1/2^n$ to each hypothesis $H \in \{0,1\}^n$. Then $H = 00\ldots0$ has low complexity $K(H) \leq \log n + 2\log\log n$. However, $-\log P(H) = n$. Therefore, by Equation 8, $H$ is not $P$-random. If we obtain $H$ by $n$ flips of a fair coin however, then with overwhelming probability we will have that $K(H) = n + O(\log n)$ and therefore $-\log P(H) \approx K(H)$ and $H$ is $P$-random.

$D$ is $\Pr(\cdot|H)$-random means that the data are random with respect to the probability distribution $\Pr(\cdot|H)$ induced by the hypothesis $H$. Therefore, we require that the sample data $D$ are 'typical', that is, 'randomly distributed' with respect to $\Pr(\cdot|H)$. If, for example, $H = (\mu, \sigma)$ induces the Gaussian distribution $\Pr(\cdot|(\mu,\sigma)) = N(\mu,\sigma)$ and the data $D$ are concentrated in a tail of this distribution, like $D = 00\ldots0$, then $D$ is atypical with respect to $\Pr(\cdot|H)$ in the sense of being non-random because it violates the $\Pr(\cdot|H)$-randomness test Equation 7.

## 2.6 If Optimal Hypothesis Violates FI

The only way to violate the Fundamental Inequality is that either $D$ is not $\Pr(\cdot|H)$-random and by Equation 7, $-\log \Pr(D|H) > K(D|H)$, or that $H$ is not $P$-random and by Equation 8, $-\log P(H) > K(H)$. We give an example of the first case.

We sample a polynomial $H_2 = ax^2 + bx + c$ at $n$ arguments chosen uniformly at random from the interval $[0,1]$. The sampling process introduces Gaussian errors in the function values obtained. The set of possible hypotheses is the set of polynomials. Assume that all numbers involved are of fixed bounded accuracy.

Because of the Gaussian error in the measuring process, with overwhelming probability the only polynomials $H_{n-1}$ which fit the sample precisely are of degree $n-1$. Denote the data by $D$. Now this hypothesis $H_{n-1}$ is likely to minimize $K(D|H)$ since we have just to describe the $n$ lengths of the intervals between the sample points along the graph of $H_{n-1}$. However, for this $H_{n-1}$ data $D$ is certainly not $\Pr(\cdot|H_{n-1})$-random, since it is extremely unlikely, and hence untypical, that $D$ arises when sampling $H_{n-1}$ with Gaussian error. Therefore, Equation 7 is violated which means that $-\log \Pr(D|H_{n-1}) > K(D|H_{n-1})$ contrary to what we used in deriving the Fundamental Inequality. With prior probability $P(\cdot) := \mathbf{m}(\cdot)$, which means $-\log P(\cdot) = K(\cdot) + O(1)$, this moreover violates the Fundamental Inequality.

In contrast, $H_2$ will show the data sample $D$ random to it in which case the Fundamental Inequality holds. Now what happens if $H_{n-1}$ is the true hypothesis and the data sample $D$ by chance is as above? In that case the Fundamental Inequality is violated and Bayes' Rule and MDL may select very different hypotheses as the most likely ones, respectively.

## 2.7 If Optimal Hypothesis Satisfies FI

Suppose the $H$ that minimizes either $-\log \Pr(D|H) - \log P(H)$ (selection according to Bayes' Rule) or $K(D|H) + K(H)$ (selection according to MDL) satisfies the Fundamental Inequality. This condition restricts the set of *admissible* hypotheses with the result that not necessarily *a priori* holds $K(D|H) + K(H) \approx K(D)$ for some minimizing trivial admissible hypothesis (like $H = \emptyset$ or $H = D$ which are not admissible). When moreover $\alpha(P,H)$ as in Equation 9 is small, then Bayes' Rule and MDL select the same optimizing hypothesis. The smallness of $\alpha(P,H)$ means that both the prior distribution $P$ is simple, and that the probability distribution $\Pr(\cdot|H)$ over the data samples induced by hypothesis $H$ simple. In contrast, if $\alpha(P,H)$ is large, which means that either of the mentioned distributions is not simple, for example when $K(\Pr(\cdot|H)) = K(H)$ for complex $H$, then there may be some discrepancy. Namely, in Bayes' Rule our purpose is to maximize $\Pr(H|D)$, and the hypothesis $H$ which minimizes $K(D|H) + K(H)$ also maximizes $\Pr(H|D)$ up to a $2^{-\alpha(P,H)}$ multiplicative factor. Conversely, the $H$ that maximizes $\Pr(H|D)$ also minimizes $K(D|H) + K(H)$ up to an additive term $\alpha(P,H)$. That is, with

$$\begin{aligned} H_{\mathrm{mdl}} &:= \mathrm{minarg}_H \{K(D|H) + K(H)\} \\ H_{\mathrm{bayes}} &:= \mathrm{maxarg}_H \{\Pr(H|D)\} \end{aligned} \qquad (9)$$

we have

$$\begin{aligned} 2^{-\alpha(P,H)} &\leq \frac{\Pr(H_{\mathrm{mdl}}|D)}{\Pr(H_{\mathrm{bayes}}|D)} \leq 1 \qquad (10) \\ \alpha(P,H) &\geq K(D|H_{\mathrm{mdl}}) + K(H_{\mathrm{mdl}}) \\ &- K(D|H_{\mathrm{bayes}}) - K(H_{\mathrm{bayes}}) \geq 0. \end{aligned}$$

## 2.8 What MDL Does

We can now assess what prior distributions and assumptions about the relation between the data sample and selected hypothesis MDL assumes. That is, how we can translate MDL in terms of Bayes' Rule. Identifying application of MDL with application of Bayes' rule on some prior distribution $P$ we must assume that the Fundamental Inequality is satisfied for $H_{\mathrm{mdl}}$ as defined in Equation 9. This means that $H_{\mathrm{mdl}}$ is $P$-random for the used prior distribution $P$. One choice to guarantee this is to choose

$$P(\cdot) := \mathbf{m}(\cdot) (= 2^{-K(\cdot)}).$$

This is a valid choice even though $\mathbf{m}$ is not recursive, since the latter requirement arose from the requirement that $\mathbf{m}(\cdot)/P(\cdot)$ be enumerable, which is certainly guaranteed by choice of $P(\cdot) := \mathbf{m}(\cdot)$. This choice of prior distribution over the hypotheses is an objective and recursively invariant quantified form of Occam's razor: simple hypothesis $H$ (with $K(H) \ll l(H)$ have high probability, and complex or random hypothesis H (with $K(H) \approx l(H)$) have low probability, see [5]. This choice of prior distribution is most convenient, since the randomness test $\log \mathbf{m}(H)/P(H) = 0$ for *each* hypothesis $H$. This means that all hypotheses $H$ are random with respect to distribution $\mathbf{m}(\cdot)$, see [5].

With prior $P(\cdot)$ set to $\mathbf{m}(\cdot)$, and the Fundamental Inequality satisfied, it follows that $D$ must be $\Pr(\cdot|H_{\mathrm{mdl}})$-random. With the chosen prior, and given data sample $D$, this restriction is therefore the only constraint on the range of hypothesis from which we can choose $H_{\mathrm{mdl}}$. Hence we can interprete MDL as an application of Bayes' Rule with as prior distribution the universal distribution $\mathbf{m}(\cdot)$ and selection of a hypothesis $H_{\text{mdl}}$ which shows the given data sample random to it in the precise sense of $\Pr(\cdot|H_{\text{mdl}})$-randomness of individual objects as developed in [5].

Since the notion of individual randomness incorporates all effectively testable properties of randomness, practical application of MDL will select the simplest hypothesis which balances the $K(D|H)$ and $K(H)$ and also shows the data sample $D$ random (as far as we are able to tell) with respect to the selected hypothesis $H_{\text{mdl}}$.

This is the 'reason' why the hypothesis selected by MDL is not simply the one which perfectly fits the data. With some amount of overstatement on can say that if one obtains perfect data for a true hypothesis, then MDL interprets these data as data obtained from a simpler hypothesis subject to measuring errors. Consequently, in this case MDL is going to give you the *false simple* hypothesis and *not* the *complex true* hypothesis.

- That is, MDL only gives us the true hypothesis if the data satisfies certain conditions relative to the true hypothesis. Stated differently: there are only data and no true hypothesis for MDL. The principle simply obtains the hypothesis which is suggested by the data and it is assumed that the data are random with respect to the hypothesis. For proper application of MDL assuming a true underlying hypothesis one must also assume that the data are not completely perfect.

Proponents of the self-evident nature of MDL will not require the given treatment of the ideal algorithmic MDL to provide a justification for the nonalgorithmic real MDL procedure. It is also only to within terms of order O(1) that the MDL and the Bayesian techniques are equivalent. Indeed, in modern forms of MDL one should take $-\log[\mathbf{m}(D|\hat{H})/\sum'_D \mathbf{m}(D'|\hat{H})]$, where $\hat{H}(D)$ is the minimizing hypothesis and the summation runs through all data such that $\hat{H}(D') = \hat{H}(D)$, [9]. The probability in the denominator gets absorbed by the term O(1), however, but for smaller amounts of data it does make a difference.

# 3 Applying MDL

Unfortunately, the function $K$ is not computable, [5]. In this paper we will not address the question which encoding one uses in practice, but refer to references [8, 14, 16, 13]. For practical applications one must settle for easily computable approximations.

In statistical applications, $H$ is some statistical distribution (or model) $H = P(\theta)$ with a list of parameters $\theta = (\theta_1, \ldots, \theta_k)$, where the number $k$ may vary and influence the (descriptional) complexity of $\theta$. (For example, $H$ can be a normal distribution $N(\mu, \sigma)$ described by $\theta = (\mu, \sigma)$.) Each parameter $\theta_i$ is truncated to fixed finite precision. The data sample consists of $n$ outcomes $\mathbf{y} = (y_1, \ldots, x_n)$ of $n$ trials $\mathbf{x} = (x_1, \ldots, x_n)$ for distribution $P(\theta)$. The data $D$ in the above formulas is given as $D = (\mathbf{x}, \mathbf{y})$. By expansion of conditional probabilities we have therefore

$$\Pr(D|H) = \Pr(\mathbf{x}, \mathbf{y}|H) = \Pr(\mathbf{x}|H) \cdot \Pr(\mathbf{y}|H, \mathbf{x}).$$

In the argument above we take the negative logarithm of $\Pr(D|H)$, that is,

$$-\log \Pr(D|H) = -\log \Pr(\mathbf{x}|H) - \log \Pr(\mathbf{y}|H, \mathbf{x}).$$

Taking the negative logarithm in Bayes' rule and the analysis of the previous section now yields that MDL selects the hypothesis with highest inferred probability satisfying $\mathbf{x}$ is $\Pr(\cdot|H)$-random and $\mathbf{y}$ is $\Pr(\cdot|H, \mathbf{x})$-random.

### Learning Polynomials.

In biological modeling, we often wish to fit a polynomial $f$ of unknown degree to a set of data points $D$ such that it can predict future data $y$ given $x$. Even if the data did come from a polynomial curve of degree, say two, because of measurement errors and noise, we still cannot find a polynomial of degree two fitting all $n$ points exactly. In general, the higher the degree of fitting polynomial, the greater the precision of the fit. For $n$ data points, a polynomial of degree $n-1$ can be made to fit exactly, but probably has no predictive value. In general, we are either given the sampling position $\mathbf{x}$, or we restrict the viable models to those for which $\mathbf{x}$ fairly covers the range of arguments with typical behavior. That is, we do not want $\mathbf{x}$ to be bunched together in one corner of the representative domain. This means that we can treat the term $-\log \Pr(\mathbf{x}|H)$ as a constant. Hence, applying MDL we look for $H_{\text{mdl}} := \text{minarg}_H\{K(\mathbf{y}|H,\mathbf{x}) + K(H)\}$.

Let us apply the MDL principle where we describe all $(k-1)$-degree polynomials by a vector of $k$ entries, each entry with a precision of $d$ bits. Then, the entire polynomial is described by

$$kd + O(\log kd) \text{ bits.} \tag{11}$$

(We have to describe $k$, $d$, and account for self-delimiting encoding of the separate items.) For example, $ax^2+bx+c$ is described by $(a,b,c)$ and can be encoded by about $3d$ bits. Each datapoint $(x_i, y_i)$ which needs to be encoded separately with precision of $d$ bits per coordinate costs about $2d$ bits.

To apply the MDL principle we must trade the cost of hypothesis $H$ (Equation 11) against the cost of describing $\mathbf{y}$ with help of $H$ and $\mathbf{x}$. As a trivial example, suppose the $n-1$ out of $n$ datapoints fit a polynomial of degree 2 exactly, but only 2 points lie on any polynomial of degree 1 (a straight line). Of course, there is a polynomial of degree $n-1$ which fits the data precisely (up to precision). Then the MDL cost is $3d+2d$ for the 2nd degree polynomial, $2d+(n-2)d$ for the 1st degree polynomial, and $nd$ for the $(n-1)$th degree polynomial. Given the choice among those three options, we select the 2nd degree polynomial for all $n > 5$. For a more sophisticated approach accounting for the average encoding cost of exceptions assuming they are Gaussian distributed, see [8].

# 4 Alternative MDL-Like Principle

In the above interpretation of MDL we essentially look for a hypothesis $H$ minimizing $K(D|H)+K(H)$. This always satisfies

$$K(D|H) + K(H) \geq K(D).$$

An incorrect interpretation is sometimes confused with MDL. In the new approach the idea is that we we define $E := D - D_H$, where $D_H$ is the data set classified according to $H$. We want to minimize

$$K(H,E) \approx K(H) + K(E|H)$$

over $H$. That is, $E$ denotes the subset of the data sample $D$ which are *exceptions* to $H$ in the sense of being 'not covered' by $H$. We want to find $H$ such that the description of $H$ and the exception data $E$ *not covered* by $H$ is minimized. Note that in this case always

$$\min_H\{K(H)+K(E|H)\} \leq K(\emptyset)+K(D) = K(D),$$

in contrast to the standard interpretation of MDL above. This incarnation of MDL is not straightforwardly derived by our approach above. We may interpret it that we look for the shortest description of an accepting program for the data consisting of a classification rule $H$ and an exception list $E$. While this principle sometimes gives good results, application may lead to absurdity as the following shows.

In many problems our data sample $D$ consists of only positive examples. Suppose we want to learn (a grammar for) English language, given a corpus of data $D$. Here, data $D$ can be the Bible or an even larger sample like the so-called 'Brown Corpus'. Then, according to our new MDL rule the best hypothesis is the trivial grammar $H$ generating *all* sentences over the alphabet. Namely, this grammar gives $K(H) = O(1)$ independent of $D$ and also $E := \emptyset$. Consequently,

$$\min_H\{K(H) + K(E|H)\} = K(H) = O(1),$$

which is absurd. The principle is vindicated and reduces to the standard one in the context of interpreting $D = (\mathbf{x}, \mathbf{y})$ as in Section 3 with $\mathbf{x}$ fixed as in 'supervised learning'. We want to find $H$ minimizing

$$K(H) + K(\mathbf{y}|H, \mathbf{x}) + K(\mathbf{x}|H)$$

which is the same as minimizing

$$K(H) + K(\mathbf{y}|H, \mathbf{x})$$

if we take $K(\mathbf{x}|H)$ constant. Now, $K(\mathbf{y}|H, \mathbf{x})$ corresponds to $K(E|H)$ if we ignore the constant $\mathbf{x}$ in the conditional.

# References


[1] L. Breiman, J. Friedman, R. Olshen and C. Stone, *Classification and Regression Trees*. Wadsworth International Group, Belmont, CA, 1984.

[2] W. Buntine, Personal communication, September 1994.

[3] P. Gács, On the symmetry of algorithmic information, *Soviet Math. Dokl.*, 15 (1974) 1477-1480. Correction: ibid., 15 (1974) 1480.

[4] A.N. Kolmogorov, Three approaches to the quantitative definition of information, *Problems Inform. Transmission* 1:1 (1965) 1-7.

[5] M. Li and P.M.B. Vitányi, *An Introduction to Kolmogorov Complexity and its Applications*, Springer-Verlag, New York, 1993.

[6] D. MacKay, pp 59 in *Maximum Entropy and Bayesian Methods*, Kluwer, 1992. (Personal communication W. Buntine.)

[7] J. Rissanen, Modeling by the shortest data description, *Automatica-J.IFAC* 14 (1978) 465-471.

[8] J. Rissanen, *Stochastic Complexity and Statistical Inquiry*, World Scientific Publishers, 1989.

[9] J.J. Rissanen, Fisher information and stochastic complexity, *IEEE Trans. Inform. Theory*, IT-42:1(1996), 40-47.

[10] J. Rissanen and M. Wax, Algorithm for constructing tree structured classifiers, US Patent No. 4719571, 1988.

[11] J. Segen, *Pattern-Directed Signal Analysis*, PhD Thesis, Carnegie-Mellon University, Pittsburgh, 1980.

[12] R.J. Solomonoff, Complexity-based induction systems: comparisons and convergence theorems, *IEEE Trans. Inform. Theory* IT-24 (1978) 422-432.

[13] V. Vovk, Minimum description length estimators under the universal coding scheme, in: P. Vitányi (Ed.), *Computational Learning Theory*, Proc. 2nd European Conf. (EuroCOLT '95), Lecture Notes in Artificial Intelligence, Vol. 904, Springer-Verlag, Heidelberg, 1995, pp. 237-251.

[14] C.S. Wallace and D.M. Boulton, An information measure for classification, *Computing Journal* 11 (1968) 185-195.

[15] C.S. Wallace and P.R. Freeman, Estimation and inference by compact coding, *J. Royal Stat. Soc., Series B*, 49 (1987) 240-251. Discussion: ibid.,252-265.

[16] K. Yamanishi, Approximating the minimum description length and its applications to learning, Manuscript, NEC Research Labs, New Jersey, 1995.

[17] A.K. Zvonkin and L.A. Levin, The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms, *Russian Math. Surveys* 25:6 (1970) 83-124.