An Introduction to the MDL Principle

1 author:

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

SEE PROFILE

An Introduction to the *MDL* Principle

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

1 Introduction

The *MDL* (Minimum Description Length) principle for statistical model selection and statistical inference is based on the simple idea that the best way to capture regular features in data is to construct a model in a certain class which permits the shortest description of the data and the model itself. Here, a model is a probability measure, and the class is a parametric collection of such models; an example is the likelihood function. Despite its simplicity the idea represents a drastically different view of modeling. First, the model class has to be such that its members can be described or encoded in terms of a finite number of symbols, say the binary. We give a brief description of the elementary coding theory needed in the appendix. This requirement also means that the traditional nonparametric models as some sort of idealized and imagined data generating distributions cannot be used unless they can be fitted to data. In the *MDL* approach we just fit models to data, and no assumption that the data are a sample from a 'true' random variable is needed. This in one stroke eliminates the difficulty in the other approaches to modeling that the more complex a model we fit the better estimate of the 'truth' we get, a problem that has had only ad hoc solutions.

As discussed in the appendix a code in which objects can be encoded defines a probability distribution for the objects, and, conversely, a probability distribution defines a code in such a way that a code length for an object in essence may be identified with the negative binary logarithm of the probability of the object. This immediately implies that the *MDL* principle may also be seen to be a global maximum likelihood principle, global because it applies even to the class of models which have different numbers of parameters. An example is curve fitting, in which the class consists of gaussian density functions for the data having the mean as a linear combination of any number of some basis functions and the variance also being a parameter.

A fundamental construct in *MDL* approach to modeling is a universal model in the universe defined by the considered class of models. Much as a universal computer or computer language, which is capable of representing any special program as a 'model', a universal model can imitate any model in the class in the sense that it assigns to any long data string about as large a probability as any model in the class. Hence, it also permits us to encode long data strings with about as few binary

[*] The author's work was supported by the Academy of Finland under contract 53884.

digits as what could be done with any model in the class. The code length of the data obtainable by a special universal model is called the *stochastic complexity* of the data, relative to the model class. We can show that for any long string it is not possible to obtain a code length that is shorter for all intents and purposes than the string's stochastic complexity - without changing the class of models, of course. Since the universal model must incorporate all the regular features in the data that can be expressed in terms of the models in the class we can compare different selections of model classes in terms of the stochastic complexity they assign to the observed data sequence: the shorter the stochastic complexity the better the model class. But more is true: We now have means to formally measure the amount of useful and learnable 'information' in the data as well as the amount that cannot be explained as incompressible 'noise'. This may be compared with the usual definition of noise as the high frequency part in the data. The problem, of course, is how high must the frequency be in order to qualify.

We conclude this introduction with a discussion of the relationship between the *MDL* principle and the Bayesian techniques, which sometimes erroneously are claimed to be the same. The broad *MDL* principle is sometimes also confused with an implementation of it as a particular and not so powerful model selection criterion BIC, see [2] and [5]. If there is any Bayes principle it probably is associated with the paradigm prior-posterior probabilities in the Bayes' identity, or actually 'degrees of belief' since in the real case where none of the hypotheses is true the usual interpretation of probability such as 'the probability of the event that the hypothesis is true', is vacuous. Putting such religious sounding interpretations aside, the Bayes identity states

$$Q(\theta|x) = \frac{P(x|\theta)w(\theta)}{\sum_\eta P(x|\eta)w(\eta)},$$

where $w(\theta)$ is called the prior probability of the hypothesis or model labelled by the parameter θ and $Q(\theta|x)$ the posterior, conditional on the data x. A fundamental objective in Baysian statistics is to pick the hypothesis, labelled as $\bar{\theta}(x)$, which maximizes the posterior or the numerator, while the objective of the *MDL* principle is to maximize the code length for the data. It is true that we can encode the data with the 2-part code length

$$L(x, \bar{\theta}(x)) = -\log P(x|\bar{\theta}(x)) - \log w(\bar{\theta}(x),$$

given a class of models $\{P(x|\theta)\}$ and the prior $w(\theta)$, but this is not the shortest code length for the data. In fact, the following code length is shorter

$$L(x) = -\log \frac{P(x|\bar{\theta}(x))}{P(\bar{\theta}(x))} - \log w(\bar{\theta}(x)),$$

where

$$P(\bar{\theta}(x)) = \sum_{y:\bar{\theta}(y)=\bar{\theta}(x)} P(y|\bar{\theta}(x)).$$

Moreover, in case the parameters range over a continuum they have to be quantized, which process plays a fundamental role in the *MDL* theory related to the complexity of the models, issues that are completely outside of Bayesian statistics.

It also turns out that the denominator in the Bayes' identity can be shown to have excellent asymptotic properties as a universal model, [3], properties alien to Bayesian statistics. We discuss in the next section other universal model constructions for model classes, with or without distributions on the parameters, which have certain optimality properties even for short strings. Such multitude of attempts to get short code lengths appear to support another criticism of the *MDL* principle, namely that it cannot be used to find the best model class and the universal model. However, this is as it should be since by a theorem of Kolmogorov in the algorithmic theory of complexity the problem to find the best computable model is noncomputable, which should put an end to the futile attempts to estimate the 'true' data generating distribution.

As a final comment, unlike in the Bayesian approach, where the posterior cannot be maximized over a family of priors in the very real case where there are more than one person's idea of prior knowledge, we can find the best prior and the corresponding universal model in the *MDL* theory.

The *MDL* principle in two-part code form, generalizing an earlier work in [17], which was restricted to a classification problem, was published in [8]. The modern theory is patterned after the algorithmic theory of complexity with more powerful implementations and information theoretic performance theorems developed over the years; for the most recent survey we refer to [14]. For excellent introductory treatments of the principle we refer to [6] and [4].

2 Universal Models

We consider model classes of parametric density or probability functions as models

$$\mathcal{M}_k = \{f(x^n;\theta) : \theta \in \Omega_k \subseteq R^k\}, \;\; \mathcal{M} = \bigcup_k \mathcal{M}_k,$$

where $\theta = \theta_1, \ldots, \theta_k$, the components being real numbers. In case of density models they must be converted to probabilities for code length interpretation. This is done by letting δ be the precision with which the data points are written and writing $P(x^n;\theta) = f(x^n;\theta)\delta^n$ so that the code length $-\log P(x^n;\theta)$ differs from $-\log f(x^n;\theta)\delta^n$ by $n \log 1/\delta$, which can be ignored. With an innocent abuse of notations we call also a negative logarithm of a density function a code length. The treatment of a density function $w(\theta)$, if present, is similar but the precision issue will be dealt with differently; it can be optimized.

2.1 Normalized maximum likelihood models

We begin with a construction of a universal model for he model class $\mathcal{M}_k$ without any density function $w(\theta)$. It is known as the *NML* (normalized maximum likelihood) model, [1], [12]:

$$\hat{f}(x^n;\mathcal{M}_k) = \frac{f(x^n;\hat{\theta}(x^n))}{C_{n,k}} \tag{1}$$

$$\begin{aligned} C_{n,k} &= \int_{\hat{\theta}(y^n)\in\Omega^\circ} f(y^n;\hat{\theta}(y^n))dy^n \qquad (2) \\ &= \int_{\hat{\theta}\in\Omega^\circ} (\int_{\hat{\theta}(y^n)=\hat{\theta}} f(y^n;\hat{\theta})dy^n)d\hat{\theta}, \end{aligned}$$

where $\hat{\theta} = \hat{\theta}(x^n)$ is the maximum likelihood estimate and the inner integral is seen to be the statistic $g(\hat{\theta}; \theta)$ evaluated at $\theta = \hat{\theta}$. In order for the integrals to be finite the space of parameters Ω is assumed to be closed and bounded with Ω° its interior. The notations dy^n and $d\hat{\theta}$ refer to differential volumes.

We take the expression

$$-\log \hat{f}(x^n; \mathcal{M}_k) = -\log f(x^n; \hat{\theta}(x^n)) + \log C_{n,k} \tag{3}$$

as the 'shortest code length' for the data x^n that can be obtained with the model class $\mathcal{M}_k$ and call it the **Stochastic Complexity** of x^n, given $\mathcal{M}_k$, [12]. It cannot represent literally the shortest code length for every data sequence. Rather, it will be that in a probabilistic sense, which however is strong enough to mean the shortest for all intents and purposes unless n is small.

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

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

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

The second fact strengthens the maxmin property such that if we restrict the data generating distributions to the class $\mathcal{M}_k$ then the mean of the stochastic complexity cannot be beaten by any code whatsoever, except 'accidentally' in the sense that the code defining model q must 'guess' the data generating model parameter to a precision ϵ, which goes to zero as the data length increases. For a precise statement we refer to [9].

We see that the lower bound for the code length is determined by the normalizing coefficient $C_{n,k}$. For some important model classes it can be calculated exactly up to reasonable data sizes as well as for the normal density models in linear quadratic regression. If the Central Limit Theorem holds for $\hat{\theta}(y^n)$ we have the excellent estimate

$$\log C_{n,k} = \frac{k}{2} \log \frac{n}{2\pi} + \log \int_\Omega |J(\theta)|^{1/2} d\theta + o(1), \tag{4}$$

where

$$J(\theta) = \lim_{n\to\infty} -n^{-1}\{E_\theta \frac{\partial^2 \log f(X^n; \theta)}{\partial\theta_i \partial\theta_j}\} \tag{5}$$

is a generalization of Fisher's information matrix. The mean is taken with respect to a model determined by θ. We assume the elements to be continuous functions of θ. Since the last term $o(1)$ goes to zero as $n \to \infty$ and the second term is a constant we see that asymptotically $\log C_{n,k}$ behaves like the first term. In fact, the criterion

$$-\log f(x^n; \hat{\theta}(x^n)) + \frac{k}{2} \log n, \tag{6}$$

called *BIC* for Bayes Information Criterion, was derived in [15] by Bayesian arguments. It was also derived in [8] by optimization of the precision for the parameters as an approximation of a 2-part code. However, for smaller amounts of data the second term through the Fisher information matrix modifies the behavior of the stochastic complexity criterion in an important manner, because it includes the effect of the sensitivity of the likelihood function as a function of the parameters. In the important linear quadratic regression problem the Fisher information matrix depends on the covariance of the data, which lends the model selection criterion based on the minimization of the stochastic complexity excellent properties. We should add that the models $f(x^n;\theta)$ in general not only depend on the number of parameters but also on the structure where the parameters lie, which we have ignored for the sake of simplicity of notations.

The calculation of the Fisher information in general is not a simple matter, and various approximations may have to be resorted to. Whenever it can be calculated it pays off to do so for we get excellent results in any statistical problem requiring modeling such as finding which regressor variables are the most important in predicting the dependent variables. A number of applications of the exact *NML* criterion to model discrete data in microbiology have been done by I. Tabus and his students; they can be found in the web. For an application to denoising we refer to [7].

2.2 The predictive *MDL* model

We next describe a predictive universal model, which is particularly attractive for time series where the data have a natural time order thus $x^n = x_1, \ldots, x_n$. Consider the conditional probability or density function

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

Suppose we model the first two data points by some model obtained by prior knowledge; write it as $f(x_1|x^0, \hat{\theta}(x^0))f(x_2|x^1, \hat{\theta}(x^1))$. Suppose then that we can fit one parameter $\hat{\theta}(x^2)$ to these two data points, so that the next can be modeled by $f(x_3|x^2, \hat{\theta}(x^2))$. Continue in this manner until k parameters can be fitted. The code length

$$L(x^n) = -\sum_{t=0}^{n-1} \log f(x_{t+1}|x^t, \hat{\theta}(x^t)),$$

can be taken as a universal model for the class $\mathcal{M}_k$. Such a model was introduced independently by P. Dawid as a prequential principle and me as a predictive *MDL* principle at about the same time the early 1980's. A justification of it is the argument that if the future data behave like the past, then the conditionals $f(x_{t+1}|x^t, \hat{\theta}(x^t))$ ought to be increasingly good, while if this is not the case no principle can perform well. In fact, one can show with some difficulty that $E_\theta L(x^n)/n$ converges to the entropy $H(\theta)$ of any model in the class at the rate $(k/2n)\log n$, which we know is optimal. Notice that by summation by parts we have the identity

$$L(x^n) = -\log f(x^n), \hat{\theta}(x^n)) + \sum_t \log \frac{f(x^{t+1}, \hat{\theta}(x^{t+1}))}{f(x^{t+1}, \hat{\theta}(x^t))},$$

where the sum is clearly positive, representing the length $(k/2)\log n$ needed to encode the optimally quantized parameters. This may be compared with the 2-part code in which the encoder looks at all the data and finds the parameters, which then must be communicated to the decoder. In the predictive version the parameters are determined only from the past so that the decoder can replicate what the encoder did.

As a special case of gaussian models with fixed variance the minimization of the predictive *MDL* code length with respect to the number of parameters amounts to a predictive least squares technique.

3 A Comparizon of three criteria on AR models

For linear least squares regression problems the universal *NML* criterion was derived in [13], which for AR (autoregressive models) of type

$$y_t = \theta_1 y_{t-1} + \ldots + \theta_k y_{t-k} + e_t,$$

$t = k+1, \ldots, n$, with gaussian density function for the errors e_t of variance τ, becomes

$$NML(k) = \{\frac{n-2k}{2}\ln\hat{\tau} + \frac{k}{2}\ln(\hat{\theta}'\frac{X'X}{n-k}\hat{\theta}) - \Gamma\left(\frac{n-2k}{2}\right) - \Gamma\left(\frac{k}{2}\right)\}. \tag{7}$$

Here $X = \{x_{ij}\}$, $x_{ij} = y_{i-j+k}$, $1 \leq i \leq n-k$, $1 \leq j \leq k$, and

$$\hat{\theta} = (X'X)^{-1}X'y_{k+1}^n \tag{8}$$

$$\hat{\tau} = \frac{1}{n-k}\sum_{t=k+1}^{n}(y_t - \sum_{j=1}^{k}\hat{\theta}_j y_{t-j})^2. \tag{9}$$

The criterion (6) is given by

$$BIC(k) = \{\frac{n-k}{2}\ln\hat{\tau} + \frac{k+1}{2}\ln(n-k)\}. \tag{10}$$

In the following tables, worked out by Ciprian Giurcaneanu, we give results of applying the *NML*, *BIC*, and the predictive *MDL* criteria, marked PLS, on data of various length generated by the indicated AR models by simulations.

The results show clearly the superiority of the *NML* criterion on short data sequences.

Appendix

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

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

by the operation of *concatenation*: $C(xx_t) = C(x)C(x_t)$, where xy denotes the string obtained when symbol y is appended to the end of the string x. We want the extended code, also written as C,

	$n = 50$			$n = 100$			$n = 300$		
	NML	BIC	PLS	NML	BIC	PLS	NML	BIC	PLS
1*	**191**	**175**	**188**	**197**	**182**	**194**	**196**	**188**	**195**
2	6	14	10	3	11	5	4	10	4
3	3	9	2	0	5	1	0	2	1
4	0	0	0	0	2	0	0	0	0
5	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0
8	0	1	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0
10	0	1	0	0	0	0	0	0	0

Table 1: Results obtained for 200 runs when the autoregressive model is order 1 with parameters $a_1 = -0.5$ and $\tau = 1.0$. The criteria NML, BIC and PLS are evaluated for $k \in \{1, \dots, 10\}$, and we report how many times every value of k is selected by each criterion.

	$n = 50$			$n = 100$			$n = 300$		
	NML	BIC	PLS	NML	BIC	PLS	NML	BIC	PLS
1	0	0	1	0	0	0	0	0	0
2*	**200**	**174**	**194**	**199**	**179**	**196**	**200**	**196**	**200**
3	0	14	4	1	15	4	0	4	0
4	0	10	0	0	2	0	0	0	0
5	0	0	1	0	1	0	0	0	0
6	0	0	0	0	2	0	0	0	0
7	0	0	0	0	0	0	0	0	0
8	0	1	0	0	0	0	0	0	0
9	0	1	0	0	1	0	0	0	0
10	0	0	0	0	0	0	0	0	0

Table 2: Results obtained for the order 2 autoregressive model with parameters $a_1 = -1.8$, $a_2 = 0.97$, $\tau = 1.0$.

	$n = 50$			$n = 100$			$n = 300$		
	NML	BIC	PLS	NML	BIC	PLS	NML	BIC	PLS
1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0
3	0	0	7	0	0	2	0	0	0
4*	**198**	**143**	**188**	**200**	**172**	**190**	**200**	**186**	**194**
5	2	12	3	0	19	6	0	12	6
6	0	21	2	0	6	2	0	1	0
7	0	10	0	0	2	0	0	1	0
8	0	9	0	0	0	0	0	0	0
9	0	4	0	0	1	0	0	0	0
10	0	1	0	0	0	0	0	0	0

Table 3: Results obtained for the order 4 autoregressive model with parameters $a_1 = -2.7607$, $a_2 = 3.8106$, $a_3 = -2.6535$, $a_4 = 0.9238$, $\tau = 1.0$.

to be not only invertible but also such that the codewords $C(a_i)$ can be separated and recognized in the code string $C(x)$ without a comma. This implies an important restriction on the codes, the so-called prefix property, which states that *no codeword is a prefix of another.* This requirement, making the code a *prefix code*, implies the important Kraft inequality

$$\sum_{a\in A} 2^{-n_a} \leq 1, \tag{11}$$

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

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

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

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

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

References

[1] Barron, A., Rissanen, J., Yu, B., (1998), 'The Minimum Description Length Principle in Coding and Modeling', (invited paper), **Information Theory 50 Years of Discovery**, (Sergio Verdu, Editor, Steven McLaughlin, Coeditor), IEEE Press.

[2] Burnham and Anderson (2002), *Model Selection and Multi-model Inference: A Practical Information-theoretic Approach*, Springer Verlag.

[3] Clarke, B. S. and Barron, A. R. (1990), 'Information-theoretic asymptotics of Bayes methods', *IEEE Trans. Information Theory*, Vol. **IT-36**, No. 3, pp 453-471

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

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

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

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

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

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

[10] Rissanen, J. (1984), 'Universal Coding, Information, Prediction, and Estimation', *IEEE Trans. Information Theory*, Vol. **IT-30**, Nr. 4, 629-636

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

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

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

[14] Rissanen, J. (2005), 'Complexity and Information in Modeling', Chapter IV in the book *Computability, Complexity and Constructivity in Economic Analysis*, (Editor: K Vela Velupillai

ISBN 1405130784), Blackwell Publishings, Oxford, Publications Dates: US: May, 2005; Rest of the World - March, 2005; Pages: 324

[15] Schwarz, G. (1978), 'Estimating the dimension of a model', *Annals of Statistics*, **6**, (2)

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

[17] Wallace, C. S. and Boulton, D. M. (1968), 'An information measure for classification', *Comput. J.* **11**, (2), 185-194