Model Selection Using the Minimum Description Length Principle

Peter G. BRYANT and Olga I. CORDERO-BRAÑA

The minimum description length (MDL) principle articulated in the last decade by Rissanen and his co-workers yields new criteria for statistical model selection. MDL criteria permit data-based choices from among alternative statistical descriptions of data without necessarily assuming that the data were sampled randomly. This article explains the MDL principle informally, indicates the criteria it yields in the common cases of multinomial distributions and Gaussian regression, and illustrates MDL's use with numerical examples. We hope thereby to stimulate experimentation and debate about the pedagogical and practical implications of the MDL approach.

KEY WORDS: Data analysis; Information theory; Minimum description length; Model selection; Stochastic complexity.

1. INTRODUCTION

Formal statistical models provide structure for many data analyses. In his classic paper, Fisher (1922) defined the object of statistics as "reduction of data," and stated

> this object is accomplished by constructing a hypothetical infinite population, of which the actual data are regarded as constituting a [random] sample. The law of distribution of this hypothetical population is specified by relatively few parameters. . ..

In this scenario, a *statistical model* is a hypothesized form for the population distribution, reflecting our external knowledge or assumptions about the source of the data. The parameters capture its essential features, and we summarize our data by estimating them.

Typical guidelines for choosing among *alternative* statistical models for the same data use Fisher's assumption that the data constitute a random sample. In this article, we present an expository introduction to the *minimum description length* (MDL) principle, a different approach to model selection. Many of the ideas and quantities involved are related to such familiar ideas of statistics and information theory as maximum likelihood estimation, but they appear here in a different context, and with slightly different interpretations. Their use for statistical model selection has developed over the last decade or so, largely as a result of the work of Rissanen. [For a review of the theoretical developments, see Rissanen (1986, 1987, 1988, 1989) and the references therein.] Much of this work has been inspired by work in the overlapping areas of Kolmogorov complexity, information theory and computer science, as well as statistics. See Li and Vitányi's (1997) admirably comprehensive text for full discussion, references, applications, and a thorough history of the ideas. A recent technical report by Hansen and Yu (1997) provides an overview of recent theoretical work and more advanced applications. The papers by Qian, Gabor, and Gupta (1994, 1996a, 1996b) provide other applications of similar ideas. Relatively little of this material has made its way into elementary textbooks, though. In hopes of stimulating debate and experimentation with the practical criteria MDL suggests, we present here MDL criteria for simple cases derived from the multinomial distribution and Gaussian regression models.

In the remainder of this section, we discuss model selection and define our notation. Section 2 introduces the idea of *representing* data and derives the MDL principle informally. Section 3 gives MDL criteria for model classes derived from the multinomial distribution and applies them to problems of comparing proportions and testing for the independence of the rows and columns in a contingency table. Section 4 derives the corresponding criteria for Gaussian regression models and applies them to problems of comparing means and selecting variables in linear regression. Section 5 discusses a few other aspects and implications of MDL. The appendix contains various tables useful for applying MDL criteria.

The technical derivations for the results given here were given by Bryant (1996). We omit them here. While occasionally complicated, they usually involve familiar quantities (though often with unfamiliar interpretations). The Rissanen references cited earlier may also be consulted for details.

1.1 Models and Model Selection

Let the data to be summarized be $Y = (y_1, y_2, \ldots, y_n)$, $y_i \in \mathcal{Y}$. Suppose we wish to summarize Y by a *statistical model* from a class $\mathcal{M} = \{m(Y|\theta) : \theta \in \Theta\}$ consisting of probability densities $m(Y|\theta)$ indexed by a parameter θ taking values in some parameter space Θ. We restrict attention to models of the form $m(Y|\theta) = \prod_{i=1}^{n} f(y_i|\theta)$, but the principles developed below apply more generally. We consider two issues:

- How can we best summarize the data using a statistical model from the class $\mathcal{M}$; that is, find the value $\hat{\theta} \in \Theta$ that provides the best summary in the class?

Peter G. Bryant is Professor of Management Science and Information Systems, College of Business, University of Colorado at Denver, Campus Box 165, Denver, CO 80217-3364 (E-mail: Peter.Bryant@cudenver.edu). Olga I. Cordero-Braña is Assistant Professor, Mathematics Department, University of Hawaii, 200 West Kawili Street, Hilo, HI 96720-4091 (E-mail: olgacb@hawaii.edu). Research supported by College of Business sabbatical funds (PB) and by National Science Foundation grant DMS-9632745 (OC-B). The authors thank the editors and referees for several helpful suggestions.

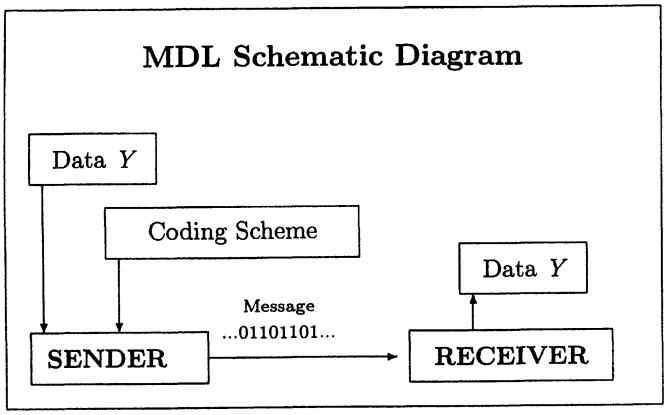

Figure 1. Basic Coding and Decoding Processes.

• How should we choose from among alternative model classes $\mathcal{M}_1$ and $\mathcal{M}_2$? Is the best model from $\mathcal{M}_1$ better or worse than the best model from $\mathcal{M}_2$?

Many common statistical procedures, such as comparing a sample mean to an externally specified value μ_0, involve *model selection:* we ask whether a model class $\mathcal{M}_1$, consisting only of models with mean μ_0, is better or worse than model class $\mathcal{M}_2$, containing models with arbitrary means. Similarly, comparing the means (or proportions) from two samples to see if they are "significantly" different, amounts to asking whether a model class $\mathcal{M}_2$ that allows for possibly different means or proportions in the two samples is better than a model class $\mathcal{M}_1$ that forces the two samples to have a common mean or proportion. Selecting variables to include (or exclude) from a regression model corresponds to choosing among model classes which include (or exclude) those variables.

Some model classes are more complicated than others. For example, the model class that permits the means of two samples to differ is more complicated than the one which requires a common mean. It fits the data at least as well as the common-mean model (and usually better), but it has more parameters. Does the better fit justify the additional complexity?

Occam's razor suggests that we assume the simpler model as a "null hypothesis," to be adopted unless the more complicated model fits the data substantially better: the burden of proof is on the more complicated model. In MDL (among other) approaches to model selection, though, no particular model class is privileged: each proposed model for the data fits the data to some extent and has a certain complexity. A model selection *criterion* must somehow combine both fit *and* model complexity to produce an aggregate figure of merit, which we can use to select the better model.

If we consider a statistical model at all, we in effect assume the data items are in some sense comparable, of course, but that need not imply that we necessarily think the sample actually was drawn randomly. In such cases, the justifications for the usual model selection criteria are not particularly compelling: we wish to use a statistical model as a useful summary, a descriptive tool. In such situations, the MDL principle offers an organized method of selecting from among alternative statistical models for a given dataset.

2. THE MDL PRINCIPLE

2.1 Representing Data

2.1.1 Codes and Code Lengths

Figure 1 depicts the task of a (perhaps imaginary) *sender*, who must represent $Y = (y_1, \ldots, y_n)$ as a sequence of the symbols 0 and 1.

The message must be sent to a (perhaps equally imaginary) *receiver*, who must decode the message, recreating the original data Y. The sender must choose a *coding scheme* to represent Y. The coding scheme assigns to each possible y_i a particular *codeword*, a sequence of 0's and 1's in a unique and unambiguous manner. In particular, the code must be a *prefix* code: no sequence of 0's and 1's used to represent one value of y may appear as a prefix in a sequence representing some other value of y. For example, if "0" is used to represent, say, y_1, then "00" may not be used to represent any other value, y_2, for then the receiver would not be able to tell whether the received message "00" represented y_2 or y_1 followed by a second copy of y_1. This involves no loss of generality. All the results hold for any uniquely decodable code. See, for example, Li and Vitányi (1997, pp. 71–74). The message consists of the codeword for y_1, followed by the codeword for y_2, and so on, arranged as a string of symbols.

2.1.2 The Kraft Inequality

Figure 2 gives the possible code words as a tree.

A coding scheme specifies some of these possible code words to be *selected*: the selected code words will denote values of Y. In Figure 2, code words "1", "00", "011" and "010" are selected. For prefix codes, if a code word such as "00" is selected for use, then no codeword *below* it in the tree may be selected, for "00" is a prefix of all codewords below it in the tree.

For each fixed level of the tree, if $l(X)$ denotes the length (in symbols) of code word X, then $\sum 2^{-l(X)} = 1$. Combining this with the prefix restriction, we obtain the *Kraft Inequality* (Fano 1961, pp. 67ff) of coding theory:

$$\sum_{X \text{ selected}} 2^{-l(X)} \leq 1, \tag{1}$$

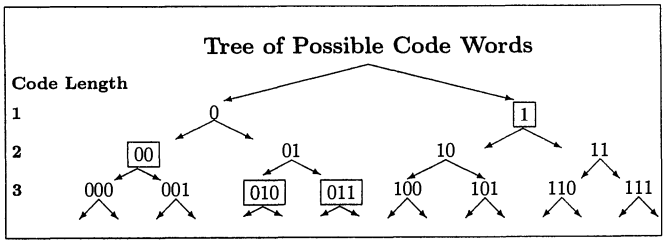

Figure 2. Selecting Code Words for Prefix Codes. Selected code words are in boxes.

which applies to any prefix code. If the code is efficient, then "$\leq$" in (1) may be replaced by "$=$". If our coding scheme chooses a code word for each value of y, and does so efficiently, the lengths $l(y)$ of the code words selected to represent y must satisfy $\sum_{y\in\mathcal{Y}} 2^{-l(y)} = 1$; that is, $2^{-l(y)}$ behaves like a probability function: it is non-negative and sums to unity. *There is thus a one-to-one relationship between coding schemes and frequency functions,* and we shall speak of the "coding scheme corresponding to" a frequency function.

2.1.3 *Efficient Codes and Statistical Models*

The following lemma provides the motivation for the MDL approach to model selection.

Lemma 1. The code length $L = \sum_{i=1}^{n} l(y_i)$ required to represent Y satisfies

$$L \geq -n \sum_{y\in\mathcal{Y}} f(y) \log_2 f(y),$$

where $f(y)$ denotes the relative frequency with which y occurs in the data Y. That is, the most efficient code for Y is the one corresponding to the actual frequencies in the data, with codewords of length $-\log_2 f(y)$.

Proof. If we represent Y by some other scheme, say, the one corresponding to $\psi(y)$, the resulting code length will be $L = -n \sum_{y\in\mathcal{Y}} f(y) \log_2 \psi(y)$. The excess code length (the penalty we pay for using the code corrresponding to ψ instead of f), is

$$-n \sum_{y\in\mathcal{Y}} f(y) \log_2 f(y) - \left[-n \sum_{y\in\mathcal{Y}} f(y) \log_2 \psi(y) \right] = \left[-n \sum_{y\in\mathcal{Y}} f(y) \log_2 \frac{\psi(y)}{f(y)} \right],$$

or n times the Kullback–Leibler "distance" from f to ψ, which is non-negative, and this proves the lemma. Such a code can actually be constructed (to a close approximation), using, say, the Huffman algorithm (Fano 1961, pp. 75–81).

By minimizing the code length, then, we are choosing the model that comes closest in this information-theoretic sense to the observed frequencies (though we have thus far taken no account of the model's complexity). Note that while many sampling-based approaches take a hypothesized *model* as given and measure how far away the data lie (from the point of view of that model), the MDL approach takes the *data* as given and asks how far away the model lies (from the point of view of the data).

2.1.4 *Model Selection Based on Code Length*

In determining which model of $\mathcal{M}$ best represents Y, we choose the one yielding the smallest code length; that is, we choose $\theta \in \Theta$ to minimize $-\sum_{i=1}^{n} \log_2 f(y_i|\theta)$. This corresponds to the model in $\mathcal{M}$ that is closest to the observed distribution of the data. We denote the minimizing value of θ by $\hat{\theta} = \hat{\theta}(Y)$, which is formally equal to the maximum likelihood estimator for θ.

When $\mathcal{M}$ contains only one model specified by θ_0, then $\hat{\theta} = \theta_0$: the code produced by the coding scheme corresponding to $f(Y \mid \theta_0)$ is the shortest (and only) code for the data Y, and its length is $-\sum_i \log_2 f(y_i|\theta_0)$.

In general, though, while the code produced by the coding scheme corresponding to $f(Y \mid \hat{\theta})$ is the shortest code for the data Y producible by a coding scheme corresponding to a member of $\mathcal{M}$, it is not yet a legitimate prefix code. The receiver in Figure 1 doesn't know which θ (i.e., which encoding scheme) was used to encode Y, and without this information, the receiver will be unable to decode the message. The sender must thus encode the value of $\hat{\theta}$ somehow, and place the code for $\hat{\theta}$ in front of the code for Y. If the sender encodes $\hat{\theta}$ by the coding scheme corresponding to some probability function $g(\hat{\theta})$, then $-\log_2 g(\hat{\theta})$ is the extra code length required to specify $\hat{\theta}$. Complicated models, such as those with many parameters or large parameter spaces Θ, will tend to have large values of $-\log_2 g(\hat{\theta})$, and thus longer messages. Coding theory thus measures both the model's ability to describe the data and its complexity on a common scale: code length.

Once the receiver knows that the best value $\hat{\theta}$ is some particular value t, though, the range of values of Y to be considered is reduced. Only those values of Y which lead to $\hat{\theta}(Y) = t$ need be considered for encoding, and these may be encoded by the coding scheme corresponding to m restricted to the range $A_t = \left\{ Y : \hat{\theta}(Y) = t \right\}$. The corresponding frequency function is the conditional frequency function of Y given $\hat{\theta}(Y) = t$, or

$$h(Y \mid t) = \frac{m(Y \mid t)}{m^*(t)}, \tag{2}$$

where

$$m^*(t) = \sum_{Y\in A_t} m(Y \mid t)$$

(or the equivalent integral form $\int_{A_t} m(Y \mid t)\, dY$ in the continuous case).

The message, then, consists of three parts:

1. Some coded representation of g, of the model class $\mathcal{M}$ used to encode Y and of any auxiliary information necessary to describe it, if such information is not assumed known to the receiver;

2. a coded representation of t, the optimal value of θ for this Y, coded by the coding scheme corresponding to $g(t)$, with code length $-\log_2 g(t)$; and

3. a coded representation of the data Y, coded by the coding scheme corresponding to $h(Y \mid t)$, of length $-\log_2 h(Y \mid t)$.

The code length for part 1 is usually about the same for all model classes under consideration. As it is common to all models, it plays no eventual role in model selection. We shall typically omit it in what follows. The binary alphabet and the corresponding use of logarithms to the base 2 are

arbitrary, and it is more convenient to use natural logarithms in what follows.

These considerations yield the *MDL criterion* for measuring the desirability of model class $\mathcal{M}$ for describing data Y, encoding the parameter $\hat{\theta}$ via the coding scheme corresponding to $g(\hat{\theta})$:

$$\mathrm{MDL}\left(Y|\mathcal{M},g\right)=-\ln h\left(Y|t\right)-\ln g\left(t\right),$$

whenever $\hat{\theta}\left(Y\right)=t$. The *MDL principle* asserts that we prefer that model class $\mathcal{M}$ for our data Y that yields the smallest value of $\mathrm{MDL}\left(Y|\mathcal{M},g\right)$.

To summarize, for MDL model selection, we must specify:

- The model class $\mathcal{M}$.

When the model class contains more than one model, we must also specify

- A set T of values of $\hat{\theta}$ to be considered. In general $T\neq\Theta$. For example, when $\mathcal{Y}$ is discrete, the obtainable values of $\hat{\theta}$ form a discrete set, too (and $\hat{\theta}$ must be encoded, not θ). In many cases, we shall take T to be finite or bounded for purposes of defining an MDL criterion, even if Θ was continuous or infinite.

- An encoding scheme corresponding to some probability function $g(t)$, $t\in T$.

We shall use the "natural" encoding scheme which defines g as proportional to $m^{*}\left(t\right)$:

$$g(t)=g^{*}\left(t\right)=\frac{m^{*}\left(t\right)}{\sum_{t\in T}m^{*}\left(t\right)};t\in T,$$

or $g^{*}\left(t\right)=m^{*}\left(t\right)/\int_{T}m^{*}\left(t\right)dt$ in the continuous case. This yields an MDL criterion

$$\mathrm{MDL}^{*}\left(Y|\mathcal{M}\right)=-\ln m\left(Y\mid t\right)+\ln\sum_{t\in T}m^{*}\left(t\right),\qquad(3)$$

(or

$$\mathrm{MDL}^{*}\left(Y|\mathcal{M}\right)=-\ln m\left(Y\mid t\right)+\ln\int_{T}m^{*}\left(t\right)dt$$

in the continuous case).

It is also possible to encode $t\in T$ by a method corresponding to some other probability function $g\left(t\right)$, though we make no further use of such an approach here. Such a g corresponds in some ways to a prior distribution in Bayesian analysis, and many of the issues concerning appropriate choice of a prior appear in modified form in this approach to MDL criteria.

3. MDL CRITERIA FOR MULTINOMIAL MODELS

The multinomial model with k categories is

$$m_{M}\left(Y|p_{1},\ldots,p_{k}\right)=\binom{n}{n_{1}\ldots n_{k}}\prod_{j=1}^{k}p_{j}^{n_{j}},\qquad(4)$$

where $\theta=\left(p_{1},\ldots,p_{k}\right),0<p_{j}<1,\sum_{j=1}^{k}p_{j}=1,\mathcal{Y}=\left(1,2,\ldots k\right)$, and $n_{j}=n_{j}\left(Y\right)=$ the number of y_{j} that are equal to j. The optimal values of p_{j} are known to be $\hat{p}_{j}=n_{j}/n$, so it suffices to take the set of parameter values T to be

$$T=\left\{\left(t_{1},\ldots,t_{k}\right)|t_{j}\in\left(0,\frac{1}{n},\ldots\frac{n-1}{n},1\right),\ \sum_{j=1}^{k}t_{j}=1\right\}.$$

From this, it follows directly from (3) that the MDL criterion is

$$\mathrm{MDL}_{M}^{*}\left(Y\right)=m_{M}\left(Y|\hat{p}_{1},\ldots,\hat{p}_{k}\right)+\ln B_{k}\left(n\right),\qquad(5)$$

where

$$B_{k}\left(n\right)=n^{-n}\sum\binom{n}{n_{1}\ldots n_{k}}\prod_{j=1}^{k}\left(n_{j}\right)^{n_{j}},\qquad(6)$$

and the summation is over all $\left(n_{1},\ldots,n_{m}\right)$ with $n_{j}\geq0,j=1,\ldots k$ and $\sum_{j=1}^{k}n_{j}=n$. Values of $B_{k}\left(n\right)$ may be derived recursively by the binomial mixture

$$B_{m}\left(n\right)=\sum_{i=0}^{n}\binom{n}{i}\left(\frac{i}{n}\right)^{i}\left(1-\frac{i}{n}\right)^{n-i}B_{m-1}\left(n-i\right)$$

with $B_{1}\left(n\right)\equiv1$. Table 4 gives values of $\ln B_{m}\left(n\right)$ for some common values of m and n.

3.1 Comparing a Proportion to a Standard

3.1.1 Formulas

The MDL criterion for describing a Bernoulli model with a fixed (standard) proportion p_{0} is

$$\mathrm{MDL}_{0}\left(p_{0}\right)=-n\left[\hat{p}\log p_{0}+\left(1-\hat{p}\right)\log\left(1-p_{0}\right)\right]$$

(no penalty term is required if we assume that "standard" means the value p_{0} is known to the receiver). The corresponding MDL criterion using the best-fitting proportion $\hat{p}$ is

$$\mathrm{MDL}_{M}\left(\hat{p}\right)=-n\left[\hat{p}\log\hat{p}+\left(1-\hat{p}\right)\log\left(1-\hat{p}\right)\right]+B_{2}\left(n\right)\qquad(7)$$

We prefer the model with a best-fitting proportion if $\mathrm{MDL}_{M}\left(\hat{p}\right)<\mathrm{MDL}_{0}\left(p_{0}\right)$.

3.1.2 Numerical Example

Consider testing whether an observed proportion $\hat{p}=1/3$ based on $n=249$ observations is significantly different from a hypothesized value $p_{0}=.4$. A conventional approximate significance test based on

$$z=\frac{\hat{p}-p_{0}}{\sqrt{\frac{p_{0}\left(1-p_{0}\right)}{n}}}=\frac{\left(\frac{1}{3}\right)-.4}{\sqrt{\frac{.4\left(1-.4\right)}{249}}}=-2.1473$$

yields a p value of .032 for testing the hypothesis that $p_{0}=.4$ .We would thus reject the hypothesis at significance level $\alpha=.05$, though not at $\alpha=.01$.

Using MDL, we must compare the MDL criteria

$$\begin{aligned}\mathrm{MDL}_{0}\left(p_{0}\right)&=-249\left[\frac{1}{3}\ln.4+\left(1-\frac{1}{3}\right)\ln\left(1-.4\right)\right]\\&=160.85\end{aligned}$$

for the model using p_0 and

$$\begin{aligned}&\mathrm{MDL}_M(\hat{p})\\ &= -249\left[\frac{1}{3}\ln\frac{1}{3}+\left(1-\frac{1}{3}\right)\ln\left(1-\frac{1}{3}\right)\right]+\ln B_2(249)\\ &= 158.49+\ln B_2(249)\end{aligned}$$

for the model with an arbitrary p. From Table 4 we see that $2.58 < \ln B_2(249) < 3.36$, so that $\mathrm{MDL}_{\mathrm{Bernoulli}}(\hat{p}) > 158.49+2.58 = 161.07 > MDL_{\mathrm{Bernoulli}}(p_0)$. We conclude that we prefer the hypothesized value $p_0 = .4$. The MDL conclusion differs from the $\alpha = .05$ significance test.

While it is difficult to generalize, in many cases models selected by MDL tend to be simpler than or the same as those selected by conventional methods and significance levels. Alternatively, MDL requires more evidence to select the more complicated model than do conventional methods. In this example, the $\alpha = .05$ significance test will reject the hypothesis that $p = .4$ for $n \geq 208$, while MDL will only do so when $\ln B_2(n) > .00947n$, which is false for $n = 250$ (though true for $n = 500$, for example). Many examples show a similar interval of sample sizes for which MDL selects the simpler model while conventional methods reject it.

3.2 Comparing Two Proportions

If the data are split into two independent datasets of size n_1 and n_2 ($n_1 + n_2 = n$), we may wish to decide if the proportions of successes within the two groups are "significantly" different from each other.

3.2.1 Formulas

The MDL criterion using a single best-fitting proportion $\hat{p}$ for the entire dataset is (7), and this model would be preferred to separate models with best-fitting proportions $\hat{p}_1$ and $\hat{p}_2$ if $\mathrm{MDL}_M(\hat{p})$ is smaller than the sum of the two MDL criteria obtained by applying the same approach to separate subsets of the data:

$$\begin{aligned}&\mathrm{MDL}_{M2}(n_1,\hat{p}_1,n_2,\hat{p}_2)\\ &= n_1\left[\hat{p}_1\log\hat{p}_1+(1-\hat{p}_1)\log(1-\hat{p}_1)\right]+B_2(n_1)\\ &\quad +n_2\left[\hat{p}_2\log\hat{p}_2+(1-\hat{p}_2)\log(1-\hat{p}_2)\right]+B_2(n_2)\end{aligned}$$

The algebra for this comparison can be expressed in different ways, but the simplest approach is simply to compute both criteria and compare them. If $\mathrm{MDL}_{M2}(n_1,\hat{p}_1,n_2,\hat{p}_2) < \mathrm{MDL}_M(\hat{p})$, we conclude the proportions are different enough to warrant the separate summaries.

3.2.2 Numerical Example

From customer satisfaction data consisting of 25 women (of whom 5 were satisfied) and 20 men (of whom 10 were satisfied), should we conclude that the satisfaction rates of men and women are substantially different? We have $n_1 = 20, \hat{p}_1 = .50, n_2 = 25, \hat{p}_2 = .20$, and $n = 45$, $\hat{p} = 1/3$. The calculations yield $\mathrm{MDL}_M(\hat{p}) = 30.85$, and $\mathrm{MDL}_{M2}(n_1,\hat{p}_1,n_2,\hat{p}_2) = 30.152$. We conclude that the proportions are sufficiently different for men and women to warrant separate summaries, though the decision is (perhaps surprisingly) a close one. A classical hypothesis testing approach based on the central limit theorem approximation would consider

$$z = \frac{.50-.20}{\sqrt{\frac{1}{3}\left(1-\frac{1}{3}\right)\left(\frac{1}{20}+\frac{1}{25}\right)}} = 2.12,$$

and would reject the hypothesis of equal proportions using significance level .05. In this case, the two methods agree.

3.3 Testing for Independence in a Contingency Table

3.3.1 Formulas

The independence model for a contingency table with r rows and c columns is

$$\begin{aligned}&m_{\mathrm{ind}}(Y|p_1,\ldots,p_r;q_1,\ldots,q_c)\\ &\quad=\binom{n}{n_{11}(Y)\cdots n_{rc}(Y)}\left[\prod_{j=1}^{r}p_j^{n_{j\cdot}}\right]\left[\prod_{k=1}^{c}q_k^{n_{\cdot k}}\right],\end{aligned}$$

where, for $j = 1,\ldots,r$ and $k = 1,\ldots,c$, p_j and q_k denote the marginal probabilities associated with row j and column k, respectively, n_{jk} denotes the number of observations in the cell determined by row j and column k, and $n_{j\cdot}$ and $n_{\cdot k}$ denote the total number of observations in rows j and k, respectively. The corresponding MDL criterion is similar to (5):

$$\begin{aligned}&\mathrm{MDL}^*_{\mathrm{ind}}(Y)\\ &\quad=-\ln\left[\binom{n}{n_{1\cdot}(Y)\cdots n_{r\cdot}(Y)}\prod_{j=1}^{r}\hat{p}_j^{n_{j\cdot}}\right]\\ &\qquad-\ln\left[\binom{n}{n_{\cdot 1}(Y)\cdots n_{\cdot c}(Y)}\prod_{k=1}^{c}\hat{q}_k^{n_{\cdot k}}\right]\\ &\qquad+\ln B_r(n)+\ln B_c(n),\end{aligned}$$

where $\hat{p}_j = \frac{n_{j\cdot}}{n}$ and $q_{\cdot k} = \frac{n_{\cdot k}}{n}$.

The more general model permitting dependence is the multinomial model (4) with number of categories equal to rc, the total number of interior cells in the table, and $\hat{p}_{jk} = \frac{n_{jk}}{n}$, $j = 1,\ldots,r; k = 1,\ldots,c$. We reject the independence model in favor of the more general model when its MDL criterion is larger than the criterion for the general model. After some algebra, we see that this is equivalent to choosing a model with dependent rows and columns whenever

$$\sum_{i=1}^{r}\sum_{j=1}^{c}n_{ij}(Y)\ln\left(\frac{n\cdot n_{ij}(Y)}{n_{i\cdot}(Y)\cdot n_{\cdot j}(Y)}\right) > \ln\frac{B_{rc}(n)}{B_r(n)\,B_c(n)}. \tag{8}$$

If we let $e_{ij}(Y) = n_{i\cdot}(Y)\,n_{\cdot j}(Y)/n$ denote the "expected" number in each cell under the independence model, then (8) becomes

$$\sum_{i=1}^{r}\sum_{j=1}^{c}n_{ij}(Y)\ln\left(\frac{n_{ij}(Y)}{e_{ij}(Y)}\right) > \ln\frac{B_{rc}(n)}{B_r(n)\,B_c(n)}. \tag{9}$$

For $n_{ij}(Y)$ near the values $e_{ij}(Y)$ predicted by the independence model, (9) becomes approximately

$$\frac{\chi^2}{2} > \ln \frac{B_{rc}(n)}{B_r(n) B_c(n)},$$

where $\chi^2 = \sum_i \sum_j (n_{ij}(Y) - e_{ij}(Y))^2 / n_{ij}(Y)$ is Neyman's χ^2 for testing the independence of rows and columns. It differs from the more usual χ^2 statistic in that it has $n_{ij}(Y)$ in the denominator rather than $e_{ij}(Y)$. If any $n_{ij}(Y) = 0$, the corresponding cell should be omitted from the summation. Table 5 gives values of the cutoff value $\ln(B_{rc}(n) / [B_r(n) B_c(n)])$ for tables up to size 5×5.

3.3.2 Numerical Example

Consider the customer satisfaction data above, represented as a contingency table

	Women	Men	Total
Satisfied	5	10	15
Unsatisfied	20	10	30
Total	25	20	45

The corresponding table of expected values under independence is

	Women	Men	Total
Satisfied	8.33333	6.66667	15
Unsatisfied	16.66667	13.33333	30
Total	25	20	45

The MDL criterion is

$$\begin{aligned} & 5 \ln\left(\frac{5}{8.333333}\right) + 10 \ln\left(\frac{10}{6.66667}\right) \\ & \quad + 20 \ln\left(\frac{20}{16.66667}\right) + 10 \ln\left(\frac{10}{13.33333}\right) \\ & = 2.27. \end{aligned}$$

From Table 5, the cutoff value for a 2×2 table and $n = 45$ is 1.14, so we conclude that a model consisting of dependent rows and columns should be preferred. While the decision here agrees in this case with the result for comparing the two proportions directly, note that the contingency table test used here is effectively *not* conditional on fixed marginal frequencies—the penalty term is summed over all possible combinations, not just those with the given marginals.

4. MDL CRITERIA FOR REGRESSION MODELS

In Gaussian regression, we have $\mathcal{Y} = \mathcal{R}^1$ and we model $Y = (y_1, y_2, \ldots, y_n)^t$ as normally distributed with mean vector $X\beta$, for some $n \times p$ matrix X (assumed known and, for convenience here, of full rank $p < n$), an unknown $p \times 1$ vector β, and covariance matrix $\sigma^2 I_n$, for some unknown $\sigma > 0$. We have $\theta = (\sigma, \beta)$ and

$$\begin{aligned} \mathcal{M} &= \{m(Y|\theta) : \theta \in \Theta\} \\ &= \left\{ (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2} \|Y - X\beta\|^2\right] \right. \\ &\qquad \left. : \sigma > 0, \beta \in \mathcal{R}^p \right\}, \end{aligned}$$

where $\|z\|^2 = z^t z$ denotes the Euclidean norm of a vector z. The optimal values $\hat{\theta}$ are well known to be

$$\hat{\beta} = \left(X^t X\right)^{-1} X^t Y,$$

and

$$\hat{\sigma} = \left[n^{-1} \left(Y - X\hat{\beta}\right)^t \left(Y - X\hat{\beta}\right)\right]^{\frac{1}{2}},$$

respectively.

The Kraft inequality applies to most countable situations, and in the case of Gaussian regression, we have a continuous parameter space, so the detailed derivations become more complicated. Also, unless some care is taken, the results may depend on the particular parameterization used, and the (finite) precision to which the parameters are represented. The particular formulation presented here was derived by Bryant (1996), but others are possible, too. We consider a range of possible values of $\hat{\sigma}$ given by

$$\sigma_{\min} < \hat{\sigma} < \sigma_{\max}$$

and encode the transformed (standardized) parameter

$$\hat{\gamma} = (\hat{\gamma}_1, \ldots \hat{\gamma}_p)^t = \frac{(X^t X)^{\frac{1}{2}} \hat{\beta}}{\sqrt{n}\hat{\sigma}},$$

rather than $\hat{\beta}$ itself, where $(X^t X)^{\frac{1}{2}}$ denotes a nonsingular $p \times p$ matrix A such that $A^t A = (X^t X)$, and $(X^t X)^{-\frac{1}{2}}$ denotes its inverse. This ensures that the resulting criteria do not depend on the scale of measurement of Y or of the design matrix X. We consider a range of γ values of total width R_0 centered at some arbitrary value γ_0. After some calculus, we find that the corresponding MDL criterion is

$$\begin{aligned} & \mathrm{MDL}^*_{\mathrm{regr}}(Y) \\ & = \frac{n}{2} \ln\left(\hat{\sigma}^2\right) - \ln \Gamma\left(\frac{n-p}{2}\right) + p \ln\left(R_0/\sqrt{\pi}\right) \\ & \quad + \ln \ln\left(\sigma_{\max}/\sigma_{\min}\right)^2 + G(n), \end{aligned} \tag{10}$$

where

$$G(n) = \frac{n}{2}\left[\frac{n-1}{n} \ln n + \ln \pi\right].$$

MDL criteria corresponding to various cases of interest follow from the appropriate choices for the design matrix X in the normal regression model.

The terms $\ln \ln\left(\sigma_{\max}/\sigma_{\min}\right)^2 + G(n)$ are typically common to all model classes under consideration, and in those cases they play no role in the eventual choice of model. The quantity R_0 denotes the range of standardized parameters $\hat{\gamma}$ we think are worthy of possible consideration. For example, for the model specifying a normal distribution with an arbitrary mean μ, $\hat{\gamma} = \bar{x}/s$, and we consider a range of $\hat{\gamma}$'s of width R_0. Our experience suggests that values of R_0 between perhaps 10 and 100 give reasonable results.

Table 1. Variable Selections for Miles per Gallon Data

Variables included	p	MDL* $R_0 = 1$	MDL* $R_0 = 10$	MDL* $R_0 = 00$	C_p	R^2	R^2_{adj}	AIC	BIC
Length,Weight	3	8.3†	15.2	22.1	3.0†	.78†	.74	74.0†	76.6†
Weight	2	10.2	14.8†	19.4†	6.2	.74	.77†	77.3	79.0
Length	2	18.6	23.2	27.8	29.7	.57	.55	94.0	93.8

NOTE: Selected models are indicated by †.

4.1 Variable Selection in Regression

To compare alternative regression models for the same data Y, we run the competing regressions and compute the corresponding MDL criteria. The regression yielding the lowest MDL criterion is preferred. Commonly we must choose whether to remove a particular variable from a model containing p independent variables (including the constant, if applicable). After some algebra, we see that such a variable should be retained in the model whenever

$$|t| > t_{p;n}(R_0),$$

where t is that variable's t statistic from the regression printout, and

$$t_{p;n}(R_0) = \sqrt{(n-p)\left\{\left[\frac{\Gamma\left(\frac{n-p+1}{2}\right)R_0}{\Gamma\left(\frac{n-p}{2}\right)\sqrt{\pi}}\right]^{2/n} - 1\right\}}. \quad (11)$$

Table 6 gives typical values of $t_{p;n}(R_0)$. For values of $n \leq 1000$ and $R \leq 500$ outside the range of the table, the approximation

$$t_{p;n}(R_0) \approx \sqrt{\log(n) - 1.838 + 2\log(R_0)}$$

is helpful. Note that the cutoff values *increase* with the sample size n. Thus, while the model selection decision may be phrased in terms of a familiar statistic, the decision process is of a different nature than in classical significance testing. Generally, MDL requires larger values of t than we are used to before it will declare a difference "significant." Two special cases of interest are discussed in the following sections.

4.1.1 Numerical Examples of Variable Selection

To illustrate the kinds of variable selections made by MDL criterion, we present summaries of regression variable selections for a number of published datasets, together with the corresponding results using other criteria: C_p, R^2, adjusted R^2, AIC, and BIC. The MDL* criteria listed omit the terms $G(n) + \ln\ln(\sigma_{\max}/\sigma_{\min})^2$, which are common to all the models being compared. The models chosen by the various criteria are marked by the symbol †.

Miles per gallon data (from Montgomery and Peck 1992). The dependent variable in this case is rated miles per gallon for the automobile. The independent variables are the length and weight of the auto. There are $n = 32$ observations. The results are in Table 1.

Crime rate data (from U. S. Bureau of the Census, cited in Neter, Kutner, Nachtsheim, and Wasserman 1996) These data consist of $n = 141$ observations for predicting crime rates in terms of the independent variables (I)ncome, % (H)igh School Graduates, and (P)opulation Density. The results are in Table 2.

In this case the choice of variables from the C_p criterion is, perhaps, ambiguous. The model that minimizes apparent bias is the model including P and I, with a value of $C_p = 2.77$, near $p = 3$. On the other hand, as suggested by Montgomery and Peck (1992, p. 274), the model containing only I, with $C_p = .89$ may have smaller *total* error, and this is the choice we have indicated in Table 2.

Gas consumption data (from Bryant and Smith 1999). The data consist of $n = 60$ observations on monthly resource usage in a chemical factory. The goal is to predict natural gas consumption as a function of the weather, measured by heating degree days, and production of two types of chemicals, generically labeled X and Y. The results are in Table 3.

At least with R in the range of 10 to 100, the MDL choice is generally either consistent with the choices of the other methods or more conservative.

4.2 Testing a Single Mean Against a Standard

If we assume, as before, that an externally specified mean μ_0 is known to the receiver, a model corresponding to a normal distribution corresponds to the regression model with

Table 2. Variable Selections for Crime Rate Data

Variables included	p	MDL* $R_0 = 1$	MDL* $R_0 = 10$	MDL* $R_0 = 100$	C_p	R^2	R^2_{adj}	AIC	BIC
I	2	1160†	1165†	1170†	.89†	.955	.955†	2775†	2777†
I,H	3	1161	1168	1175	2.28	.955	.954	2776	2779
P,I	3	1162	1169	1176	2.77	.955	.954	2777	2779
P,I,H	4	1163	1172	1181	4.0	.955†	.954	2778	2781
P,H	3	1364	1371	1378	2288.0	.208	.196	3181	3176
P	2	1365	1369	1374	2369.0	.181	.175	3184	3181
H	2	1378	1383	1387	2890.0	.010	.003	3211	3207

NOTE: Selected models are indicated by †.

Table 3. Variable Selections for Gas Consumption Data

Variables included	p	MDL* $R_0 = 1$	MDL* $R_0 = 10$	MDL* $R_0 = 100$	C_p	R^2	R^2_{adj}	AIC	BIC
X,Y,Degday	4	460†	469†	478	4.0†	.73†	.72†	1061†	1064†
X,Degday	3	463	470	477†	10.8	.69	.68	1068	1070
Y,Degday	3	465	472	479	16.1	.67	.65	1073	1074
Degday	2	480	484	489	61.3	.44	.43	1102	1101
X	2	488	493	497	98.3	.26	.25	1118	1117
X,Y	3	489	496	503	98.9	.27	.24	1120	1117
Y	2	492	497	502	122.5	.15	.13	1127	1125

NOTE: Selected models are indicated by †.

$p = 0$. The alternative model allowing an arbitrary mean μ corresponds to the case $p = 1$ with $X = (1, \ldots, 1)^t$. We prefer the general mean to μ_0 whenever $|t| > t_{1;n}(R_0)$.

For example, the specified rate for a certain machine producing small metal parts is 42,000 parts per hour. A sample of $n = 28$ such machines on a particular day yields an average production rate of 38,346 parts per hour, with a corresponding standard deviation of 3629 parts per hour. [The data were given by Bryant and Smith (1999, pp. I-31–33).] The t statistic for testing $\mu = 42{,}000$ is

$$t = \frac{38346 - 42000}{3629/\sqrt{28}} = -5.328$$

so that we would reject the hypothesis that on average the machines are producing at their specified rate at ordinary significance levels. From Table 6, any value of t bigger than, say, 2.5 to 3.2 should lead us to conclude the same thing, so in this case MDL and conventional methods agree.

4.3 Testing the Equality of Two Means

For comparing two means from independent samples of size n_1 and n_2 ($n_1 + n_2 = n$) (assuming a common variance) the model selection decision can be expressed in terms of the usual statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

where s_p is the pooled-within-groups estimate of the standard deviation. We prefer a description with different means whenever $|t| > t_{2;n}(R)$.

For example, a sample of $n_1 = 44$ women and $n_2 = 48$ women yielded average systolic blood pressures of $\hat{\mu}_1 = 114.77$ and $\hat{\mu}_2 = 128.54$. The corresponding standard deviations were $s_1 = 12.28$ and $s_2 = 16.24$. [The data were given by Bryant and Smith (1999, pp. III-93–94).] If we assume a common variance, its estimate is $s = \sqrt{\left(43\,(12.28)^2 + 47\,(16.24)^2\right)/90} = 14.484$, and the usual t statistic is

$$t = \frac{128.54 - 114.77}{14.484\sqrt{\frac{1}{44} + \frac{1}{48}}} = 4.56.$$

From Table 6, the cutoff value for t for $n = 92$ would be between perhaps 2.7 and 3.3, depending on the particular value of R_0, so we conclude that a model allowing for separate means for the men and women is justified.

If we do not assume equal variances, the Satterthwaite approximate t statistic is

$$t = \frac{128.54 - 114.77}{\sqrt{\frac{(12.28)^2}{44} + \frac{(16.24)^2}{48}}} = 4.61$$

with 86 degrees of freedom, which leads to the conclusion that the means were different at conventional significance levels. To make the decision using the MDL criterion, we may compare the sum of the MDL criteria for two separate populations to the corresponding criterion for a single population, and make a decision on this basis. From (10), we obtain

$$\begin{aligned}
\mathrm{MDL}_1^* &= 172.5372 + \ln\left(R_0/\sqrt{\pi}\right) + \ln\ln\left(\sigma_{\max}/\sigma_{\min}\right)^2; \\
\mathrm{MDL}_2^* &= 201.7069 + \ln\left(R_0/\sqrt{\pi}\right) + \ln\ln\left(\sigma_{\max}/\sigma_{\min}\right)^2; \\
\text{and} \\
\mathrm{MDL}^* &= 385.6415 + \ln\left(R_0/\sqrt{\pi}\right) + \ln\ln\left(\sigma_{\max}/\sigma_{\min}\right)^2,
\end{aligned}$$

for the first, second and combined samples, respectively. We prefer the model class permitting separate means and variances if $\mathrm{MDL}_1^* + \mathrm{MDL}_2^* < \mathrm{MDL}^*$, which happens whenever $\ln\left(R_0/\sqrt{\pi}\right) + \ln\ln\left(\sigma_{\max}/\sigma_{\min}\right)^2 < 385.6415 - 201.7069 - 172.5372 = 11.397$. If, for example, $R_0 = 100$ and $(\sigma_{\max}/\sigma_{\min}) \leq 1000$, this is easily satisfied, so we prefer the model with separate means and variances, as before.

5. REMARKS

5.1 Derivations

For the familiar multinomial and Gaussian regression models considered here, we can derive explicit MDL criteria. Other simple cases, such as choosing the number of classes in a histogram, also yield relatively tractable criteria. In the cases of selecting the order, say, of an ARMA model for time series, further developments are available, too. See Rissanen (1989) for some examples. For some other models, though, this is difficult. Much remains to do in discovering useful general characterizations and approximations.

5.2 Relation to other methods

MDL has appeared in a number of different variations, many without the conditioning in (2). It is also related to various other ideas and approaches.

5.2.1 *Minimum Message Length (MML)*

The minimum message length criterion described by Wallace and his co-workers (e.g., Wallace and Boulton 1968, Wallace and Freeman 1987) produces results similar to MDL, though some of the rationale differs. In particular in some of the detailed derivations, these authors apparently assumed that θ is a random variable for purposes of deriving a penalty to minimize an "expected" loss. Compared with MDL approaches, MML has more of a Bayesian flavor and less of a data-analytic flavor.

5.2.2 *Bayesian Approaches*

The MDL criterion used here is similar to using the Bayesian posterior mode as an estimate and basing model selection on the posterior distribution evaluated at that mode. MDL differs from such an approach by conditioning on $\hat{\theta} = t$ (i.e., the denominator in (2)). This conditioning has no obvious interpretation in a Bayesian context, though it is natural in the data-analytic context considered here.

The function $g(\theta)$ used here to encode the parameters is the analogue of the prior distribution in Bayesian analysis, and many of the issues and controversies surrounding the selection of a Bayesian prior appear in somewhat disguised form in MDL. Discussion of the MDL equivalents of such ideas as non-informative and reference priors appear, for example, in Rissanen (1987).

5.2.3 *Maximum Likelihood*

The MDL estimate $\hat{\theta}$ is formally identical to the maximum likelihood estimator, and the MDL criterion is formally a form of penalized maximum likelihood model selection, though its motivation is different. Various alternative penalized maximum likelihood with penalties such as Akaike's (1973) Information Criterion or ICOMP (Bozdogan 1988, 1990, 1993) have been proposed, but all are justified from the idea that the data are a random sample. When augmented by prior information as in the Bayesian approach, criteria such as the Bayesian Information Criterion (Schwarz 1978) appear, also.

If our data really do consist of a random sample from some hypothetical population, then variants of the usual asymptotic results for maximum likelihood estimators will apply. For example, one can show under appropriate conditions that if $Y \sim N(Z\gamma, \sigma^2 I)$, but we model it as $Y \sim N(X\beta, \sigma^2 I)$, the MDL criterion derived from the "true" model $Y \sim N(Z\gamma, \sigma^2 I)$ will be less than the one derived from any other model $Y \sim N(X\beta, \sigma^2 I)$ from some point on with probability one unless $X = ZP$ for some nonsingular matrix P (in which case the criteria are equal). Given the data-analytic motivation of the MDL principle as articulated here, though, it is not clear that such limiting results based on the assumption of random sampling are really a fair way of evaluating MDL summaries, which are not specifically designed to seek a population "truth."

In addition, a variety of theorems on the optimality of MDL results in the context of complexity theory have been derived. See Li and Vitányi (1997), for example.

5.3 Challenges to Practice

To prompt debate and experimentation, we mention some issues and challenges posed by the MDL approach to model selection.

- How should the various hyperparameters (such as R_0 in the case of normal regression) be interpreted? How should they be chosen? How much does the choice matter in practice?

- Commonly MDL chooses simpler models than other approaches for a certain range of sample sizes. When the choices differ, what does this tell us in practical terms?

- MDL suggests that organized approaches to data analysis and summarization are possible without specifically assuming random sampling. Can Fisher's "object" be "accomplished" by MDL or its equivalent, too? When the assumption of random sampling is implausible, perhaps in such situations as data mining, could alternative approaches such as MDL yield more interpretable results?

5.4 Implications for Teaching

Finally, we hope that some implications for teaching and curricula may emerge as more people experiment with MDL criteria. The possibilities for courses in mathematical statistics seem fairly clear. The implications for the elementary service course in statistical methods are less clear. We speculate that the availability of model selection criteria not specifically justified by the assumption of random sampling might lead to increased emphasis on formal methods of data analysis. For example, instead of limiting coverage to the frequentist ideas of confidence and significance, such courses might compare and contrast four alternatives:

1. Frequentist approaches such as confidence intervals and hypothesis tests for situations involving repeated random sampling (e.g., process control), for which long-run average performance has a natural interpretation;

2. likelihood methods and interpretations for individual (rather than repeated) random samples;

3. elementary Bayesian methods for situations in which the prior distribution seems easy to obtain and interpret; and

4. MDL-based model selection criteria where assumptions of random sampling, long-run average performance and/or meaningful prior distributions are not satisfied, and the primary purpose is to select appropriate models for descriptive purposes.

We invite discussion on these possibilities.

APPENDIX: TABLES FOR MDL MODEL SELECTION

For reference, the various tables used in the text for critical values and penalties for MDL model selection are gathered together here. For descriptions of the tables and their uses, consult the text.

Table 4. Table of the Multinomial Penalty Function $\ln B_m(n)$ for Common Values of $m \leq 21$

	Number of categories, m															
n	*m* = 2	*m* = 3	*m* = 4	*m* = 5	*m* = 6	*m* = 7	*m* = 8	*m* = 9	*m* = 10	*m* = 12	*m* = 14	*m* = 15	*m* = 16	*m* = 18	*m* = 20	*m* = 21
1	.69	1.10	1.39	1.61	1.79	1.95	2.08	2.20	2.30	2.48	2.64	2.71	2.77	2.89	3.00	3.04
2	0.92	1.50	1.95	2.30	2.60	2.86	3.09	3.30	3.48	3.81	4.09	4.21	4.33	4.55	4.74	4.84
3	1.06	1.77	2.32	2.78	3.17	3.51	3.81	4.09	4.34	4.78	5.17	5.34	5.51	5.82	6.09	6.22
4	1.17	1.98	2.61	3.15	3.61	4.02	4.38	4.72	5.02	5.57	6.05	6.27	6.47	6.86	7.21	7.37
5	1.26	2.14	2.85	3.45	3.97	4.44	4.86	5.24	5.60	6.23	6.79	7.05	7.29	7.75	8.16	8.36
6	1.33	2.28	3.05	3.70	4.28	4.80	5.26	5.69	6.09	6.81	7.44	7.73	8.01	8.53	9.00	9.23
7	1.39	2.40	3.22	3.93	4.55	5.11	5.62	6.09	6.53	7.32	8.02	8.34	8.65	9.23	9.76	10.01
8	1.45	2.51	3.38	4.13	4.79	5.39	5.94	6.45	6.92	7.77	8.54	8.89	9.23	9.86	10.44	10.72
9	1.49	2.60	3.51	4.30	5.01	5.64	6.23	6.77	7.27	8.19	9.01	9.39	9.76	10.44	11.07	11.37
10	1.54	2.69	3.64	4.46	5.20	5.87	6.49	7.06	7.60	8.57	9.45	9.85	10.24	10.97	11.65	11.97
15	1.71	3.02	4.13	5.11	5.99	6.79	7.54	8.24	8.90	10.11	11.21	11.73	12.22	13.16	14.03	14.45
20	1.84	3.27	4.49	5.58	6.57	7.48	8.33	9.13	9.88	11.28	12.55	13.16	13.73	14.83	15.86	16.35
25	1.94	3.46	4.78	5.95	7.03	8.02	8.96	9.83	10.67	12.22	13.64	14.31	14.96	16.19	17.35	17.91
30	2.02	3.63	5.02	6.27	7.41	8.48	9.48	10.43	11.32	13.00	14.55	15.28	15.99	17.34	18.62	19.23
35	2.09	3.76	5.22	6.53	7.74	8.87	9.93	10.94	11.89	13.68	15.34	16.13	16.89	18.34	19.71	20.37
40	2.15	3.88	5.40	6.77	8.03	9.21	10.33	11.38	12.39	14.28	16.03	16.87	17.67	19.22	20.68	21.39
45	2.21	3.99	5.56	6.98	8.29	9.52	10.68	11.78	12.84	14.82	16.66	17.53	18.38	20.01	21.55	22.30
50	2.26	4.09	5.70	7.16	8.52	9.79	11.00	12.14	13.24	15.30	17.22	18.13	19.02	20.73	22.34	23.13
100	2.58	4.73	6.65	8.42	10.08	11.65	13.15	14.58	15.96	18.59	21.06	22.24	23.40	25.64	27.78	28.82
500	3.36	6.27	8.95	11.47	13.87	16.18	18.41	20.57	22.67	26.73	30.62	32.51	34.36	37.99	41.50	43.23

NOTES: Tabled entries are computed as $\ln B_m(n)$, where $B_m(n)$ is defined by Equation (6) in the text.
Column headings: n = sample size; m = number of categories.

Table 5. Table of MDL Contingency Table Cutoffs

	Table size									
n	*2 by 2*	*3 by 2*	*3 by 3*	*4 by 2*	*4 by 3*	*4 by 4*	*5 by 2*	*5 by 3*	*5 by 4*	*5 by 5*
2	.11	.18	.30	.22	.36	.43	.26	.41	.49	.56
3	.20	.34	.55	.43	.69	.87	.50	.79	.99	1.13
4	.27	.46	.76	.60	.98	1.25	.70	1.14	1.45	1.66
5	.33	.57	.96	.75	1.24	1.59	.89	1.46	1.86	2.17
6	.39	.67	1.13	.88	1.48	1.91	1.06	1.75	2.25	2.64
7	.44	.76	1.29	1.01	1.70	2.21	1.21	2.01	2.61	3.07
8	.48	.83	1.43	1.11	1.88	2.47	1.34	2.25	2.93	3.47
9	.53	.92	1.57	1.23	2.08	2.74	1.48	2.49	3.26	3.87
10	.56	.97	1.68	1.31	2.24	2.96	1.6 0	2.70	3.55	4.23
15	.71	1.26	2.20	1.70	2.96	3.96	2.08	3.60	4.79	5.77
20	.81	1.46	2.59	2.00	3.52	4.75	2.46	4.31	5.79	7.03
25	.90	1.63	2.91	2.24	3.98	5.40	2.78	4.90	6.62	8.09
30	.98	1.76	3.17	2.44	4.35	5.95	3.03	5.38	7.33	8.98
35	1.04	1.89	3.42	2.62	4.70	6.45	3.27	5.84	7.96	9.80
40	1.10	2.00	3.62	2.78	5.00	6.87	3.47	6.22	8.51	10.51
45	1.14	2.09	3.80	2.91	5.27	7.26	3.65	6.56	9.01	11.15
50	1.18	2.17	3.96	3.04	5.51	7.62	3.82	6.88	9.48	11.76
60	1.27	2.34	4.27	3.26	5.95	8.25	4.11	7.45	10.30	12.81
70	1.34	2.46	4.51	3.46	6.31	8.79	4.36	7.93	11.00	13.73
80	1.38	2.57	4.75	3.63	6.65	9.28	4.58	8.37	11.64	14.55
90	1.44	2.68	4.94	3.79	6.95	9.72	4.78	8.74	12.20	15.27
100	1.49	2.77	5.12	3.92	7.21	10.10	4.96	9.09	12.71	15.94
500	2.23	4.24	8.03	6.10	11.51	16.46	7.84	14.77	21.08	26.96
1000	2.56	4.90	9.34	7.09	13.46	19.38	9.15	17.35	24.93	32.05
2000	2.90	5.57	10.67	8.09	15.46	22.35	10.49	20.00	28.88	37.29

NOTES: Table entries are cutoff values for the MDL criterion for testing the independence of the rows and columns of a contingency table, from Equation (9) in the text.
Column headings are:

Table size : Number of rows and columns

n : Total number of observations in the table

Table 6. Table of MDL Regression cutoffs for $t_{p;n}(R_0)$.

n	p	$R_0=10$	$R_0=20$	$R_0=50$	n	p	$R_0=10$	$R_0=20$	$R_0=50$	n	p	$R_0=10$	$R_0=20$	$R_0=50$
2	1	1.48	2.32	3.86	10	1	2.39	2.81	3.36	30	1	2.56	2.87	3.25
3	1	1.96	2.70	3.89	10	2	2.22	2.61	3.13	30	2	2.51	2.81	3.19
3	2	1.08	1.56	2.31	10	3	2.03	2.41	2.89	30	3	2.45	2.75	3.13
4	1	2.14	2.78	3.73	10	4	1.84	2.19	2.64	30	4	2.40	2.70	3.06
4	2	1.57	2.08	2.83	10	5	1.63	1.95	2.36	30	5	2.35	2.63	2.99
4	3	.89	1.23	1.73	10	6	1.41	1.70	2.06	30	6	2.29	2.57	2.92
5	1	2.23	2.80	3.61	10	7	1.16	1.41	1.73	30	7	2.23	2.51	2.85
5	2	1.81	2.30	3.00	10	8	.87	1.08	1.34	30	8	2.17	2.45	2.78
5	3	1.34	1.74	2.29	10	9	.51	.67	.86	30	9	2.11	2.38	2.71
5	4	.77	1.05	1.42						30	10	2.05	2.32	2.64
6	1	2.28	2.80	3.53	15	1	2.46	2.82	3.29	50	1	2.64	2.92	3.27
6	2	1.96	2.42	3.06	15	2	2.35	2.70	3.15	50	2	2.61	2.89	3.23
6	3	1.60	2.00	2.55	15	3	2.23	2.57	3.01	50	3	2.58	2.85	3.19
6	4	1.19	1.52	1.96	15	4	2.12	2.44	2.86	50	4	2.54	2.82	3.16
6	5	.69	.92	1.23	15	5	1.99	2.31	2.71	50	5	2.51	2.78	3.12
7	1	2.32	2.80	3.47	15	6	1.87	2.17	2.54	50	6	2.48	2.75	3.08
7	2	2.05	2.50	3.10	15	7	1.73	2.02	2.37	50	7	2.45	2.71	3.04
7	3	1.76	2.16	2.70	15	8	1.59	1.86	2.20	50	8	2.41	2.68	3.00
7	4	1.45	1.79	2.25	15	9	1.44	1.69	2.01	50	9	2.38	2.64	2.96
7	5	1.08	1.36	1.74	15	10	1.28	1.51	1.80	50	10	2.35	2.60	2.92
7	6	.63	.83	1.10	20	1	2.50	2.84	3.26	100	1	2.75	3.01	3.32
8	1	2.35	2.81	3.42	20	2	2.42	2.75	3.16	100	2	2.73	2.99	3.30
8	2	2.12	2.55	3.11	20	3	2.34	2.66	3.06	100	3	2.72	2.97	3.29
8	3	1.88	2.27	2.79	20	4	2.25	2.57	2.96	100	4	2.70	2.96	3.27
8	4	1.62	1.97	2.43	20	5	2.17	2.47	2.85	100	5	2.69	2.94	3.25
8	5	1.33	1.63	2.03	20	6	2.08	2.37	2.74	100	6	2.67	2.92	3.23
8	6	1.00	1.25	1.57	20	7	1.98	2.27	2.63	100	7	2.65	2.90	3.21
8	7	.58	.77	1.00	20	8	1.89	2.17	2.51	100	8	2.64	2.89	3.19
9	1	2.37	2.81	3.39	20	9	1.79	2.06	2.39	100	9	2.62	2.87	3.17
9	2	2.17	2.58	3.13	20	10	1.69	1.94	2.26	100	10	2.60	2.85	3.15
9	3	1.97	2.35	2.85	25	1	2.53	2.86	3.26	200	1	2.86	3.10	3.40
9	4	1.74	2.09	2.55	25	2	2.47	2.79	3.18	200	2	2.85	3.09	3.39
9	5	1.50	1.82	2.22	25	3	2.41	2.71	3.10	200	3	2.84	3.08	3.38
9	6	1.24	1.51	1.86	25	4	2.34	2.64	3.02	200	4	2.84	3.08	3.37
9	7	.93	1.16	1.45	25	5	2.27	2.57	2.94	200	5	2.83	3.07	3.36
9	8	.54	.71	.92	25	6	2.20	2.49	2.85	200	6	2.82	3.06	3.35
					25	7	2.13	2.41	2.76	200	7	2.81	3.05	3.34
					25	8	2.06	2.34	2.68	200	8	2.80	3.04	3.33
					25	9	1.99	2.25	2.59	200	9	2.79	3.03	3.32
					25	10	1.91	2.17	2.49	200	10	2.79	3.02	3.31

NOTES: Table entries are cutoff values $t_{p;n}\left(R_0\right)$ computed from Equation (11) in the text. If the usual t-statistic from a regression printout is less than $t_{p;n}$ in absolute value, the corresponding variable should be removed from the regression equation, according to the MDL criterion.

Column headings are as follows:

n : Number of observations in regression

p : Number of predictors in regression, including constant term, if any (before possibly removing the variable in question).

R_0 : MDL regression hyperparameter as described in the text.

[Received February 1999. Revised April 2000.]

REFERENCES

Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle," in *Proceedings of the Second International Symposium on Information Theory*, eds. B. N. Petrov and F. Csaki, Budapest: Akademia Kiedo, pp. 267–281.

Bozdogan, H. (1988), "ICOMP: A New Model-Selection Criterion," in *Classification and Related Methods of Data Analysis*, ed. H. H. Bock, Amsterdam: International Federation of Classification Societies, Elsevier Science Publishers (North-Holland), pp. 599–608.

——— (1990), "On the Information-Based Measure of Covariance Complexity and its Application to the Evaluation of Multivariate Linear Models," *Communications in Statistics, Theory, and Methods*, 19, 221–278.

——— (1993), "Choosing the Number of Component Clusters in the Mixture-Model Using a New Informational Complexity Criterion of the Inverse Fisher Information Matrix," in *Information and Classification: Concepts, Methods and Applications*, eds. O. Opitz, B. Lausen, and R. Klar, number 16 in *Studies in Classification, Data Analysis, and Knowledge Organization, Proceedings of the Annual Conference of the Gesellschaft fur Klassifikation*, Berlin: Gesellschaft für Klassifikation, Springer-Verlag, pp. 40–54.

Bryant, P. G. (1996), "The Minimum Description Length Principle for Gaussian Regression," Technical Report UCD-CBA Working Paper 1996-08, University of Colorado at Denver, College of Business, Campus Box 165, Denver, Colorado 80217-3364.

Bryant, P. G., and Smith, M. A. (1999), *Practical Data Analysis: Case Studies in Business Statistics* (vols. I, II, and III), Burr Ridge, IL: Richard D. Irwin, Inc.

Fano, R. M. (1961), *Transmission of Information: A Statistical Theory of Communications*, Cambridge, MA: The MIT Press.

Fisher, R. A. (1922), "On the Mathematical Foundations of Theoretical Statistics," *Philosophical Transactions of the Royal Society of London*, Series A, 222, 309–368.

Hansen, M. H., and Yu, B. (1997), "Model Selection and the Principle of Minimum Description Length," Technical Report, Bell Laboratories, Murray Hill, NJ.

Li, M., and Vitányi, P. (1997), *An Introduction to Kolmogorov Complexity and Its Applications* (2nd ed.), New York: Springer-Verlag.

Montgomery, D. C., and Peck, E. A. (1992), *Introduction to Linear Regression Analysis* (2nd ed.), New York: Wiley.

Neter, J., Kutner, M. H., Nachtsheim, C., and Wasserman, W. (1996), *Applied Linear Regression Models* (4th ed.), Burr Ridge, IL: Richard D. Irwin, Inc.

Qian, G., Gabor, G., and Gupta, R. P. (1994), "Principal Components Selection by the Criterion of the Minimum Mean Difference of Complexity," *Journal of Multivariate Analysis*, 49, 55–75.

——— (1996a), "Generalised Linear Model Selection by the Predictive Least Quasi-Deviance Criterion," *Biometrika*, 83, 41–54.

Qian, G., Gupta, R. P., and Gabor, G. (1996b), "Test for Homogeneity of Several Populations by Stochastic Complexity," *Journal of Statistical Planning and Inference*, 53, 133–151.

Rissanen, J. (1986), "Stochastic Complexity and Modeling," *The Annals of Statistics*, 14, 1080–1100.

——— (1987), "Stochastic Complexity," *Journal of the Royal Statistical Society*, Series B, 49, 223–265.

——— (1988), "Minimum Description Length Principle," in *Encyclopedia of Statistical Sciences* (vol. 5), eds. S. Kotz and N. Johnsos, New York: Wiley, pp. 523–527.

——— (1989), *Stochastic Complexity in Statistical Inquiry*, Singapore: World Scientific Publishing Company.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

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

Wallace, C. S., and Freeman, P. R. (1987), "Estimation and Inference by Compact Coding," *Journal of the Royal Statistical Society*, Series B, 49, 240–252.