5-1999

Local Asymptotics and the Minimum Description Length

Dean P. Foster
University of Pennsylvania

Robert A. Stine
University of Pennsylvania

Recommended Citation

Foster, D. P., & Stine, R. A. (1999). Local Asymptotics and the Minimum Description Length. *IEEE Transactions on Information Theory, 45* (4), 1289-1293. http://dx.doi.org/10.1109/18.761287

Local Asymptotics and the Minimum Description Length

Abstract

Common approximations for the minimum description length (MDL) criterion imply that the cost of adding a parameter to a model fit to n observations is about (1/2) log *n* bits. While effective for parameters which are large on a standardized scale, this approximation overstates the parameter cost near zero. A uniform approximation and local asymptotic argument show that the addition of a small parameter which is about two standard errors away from zero produces a model whose description length is shorter than that of the comparable model which sets this parameter to zero. This result implies that the decision rule for adding a model parameter is comparable to a traditional statistical hypothesis test. Encoding the parameter produces a shorter description length when the corresponding estimator is about two standard errors away from zero, unlike a model selection criterion like BIC whose threshold increases logarithmically in *n*.

Keywords

BIC, hypothesis test, model selection, two-part code, universal code

Disciplines

Statistics and Probability

Local Asymptotics and the Minimum Description Length

Dean P. Foster and Robert A. Stine

Department of Statistics

The Wharton School of the University of Pennsylvania

Philadelphia, PA 19104-6302

March 27, 1998

Abstract

Common approximations for the minimum description length (*MDL*) criterion imply that the cost of adding a parameter to a model fit to n observations is about $(1/2)\log n$ bits. While effective for parameters which are large on a standardized scale, this approximation overstates the parameter cost near zero. A uniform approximation and local asymptotic argument show that the addition of a small parameter which is about two standard errors away from zero produces a model whose description length is shorter than that of the comparable model which sets this parameter to zero. This result implies that the decision rule for adding a model parameter is comparable to a traditional statistical hypothesis test. Encoding the parameter produces a shorter description length when the corresponding estimator is about two standard errors away from zero, unlike a model selection criterion like *BIC* whose threshold increases logarithmically in n.

Key Phrases: BIC, hypothesis test, model selection, two-part code, universal code.

1 Introduction

The description length of a sequence of n random variables $Y_1, Y_2, \ldots, Y_n$ based on a model with k parameters $\theta_1, \ldots, \theta_k$ is defined as (Rissanen 1983)

$$L_{n,k}(Y,\theta) = \log^* \left(C(k) \parallel \theta \parallel^k \right) + \log \frac{1}{P(Y|\theta)} , \tag{1}$$

where $P(Y|\theta)$ is the likelihood function, $C(k)$ is the volume of the k dimensional unit ball, the parameter norm is

$$\|\theta\|^2 = \theta' H \theta, \quad H = \frac{-\partial^2 \log P(Y|\theta)}{\partial \theta^2} ,$$

and $\log^*$ denotes the iterated logarithm

$$\log^* x = \log x + \log\log x + \cdots , \tag{2}$$

where the sum extends only over positive terms. All logs are base 2. The description length approximates the number of bits required to encode both the parameters of the model and the associated compressed data. In order to select the best parametric model from among several of possibly varying dimension k, Rissanen proposes that one choose the model which obtains the minimum description length (*MDL*).

As a model selection criterion, *MDL* presents an explicit trade-off of model complexity and goodness of fit to the data. The two summands which define the description length (1) can be associated with the lengths of the two components of a two-part code for Y: a preamble which indicates the value of the parameter θ used to encode the data Y in the second part of the code. This balance guards against overfitting. A complex model with many parameters might obtain high data compression (the second part of the code being rather short), but its complexity would necessitate a long preamble. The need to encode the model parameters as part of the two-part code thus avoids the tendency to overfit.

Our results focus on a set of parameter values which are within $\log n/\sqrt{n}$ of the origin. Although this set is not studied in the usual asymptotic analysis of *MDL*, we show that the decision of whether or not to code a parameter is made on this set. A particularly important approximation (Rissanen 1989) shows that the minimum description length of a parametric model is

$$L_{n,k}(Y,\theta) \approx \frac{k}{2} \log n + \log \frac{1}{P(Y|\theta)} , \tag{3}$$

for all θ in a compact subset of $\mathbf{R}^k$, *with the exception* of a small set of vanishing measure. In the one-dimensional case, we show that the cost of coding a nonzero parameter from the exceptional set near zero is considerably less than $(1/2)\log n$. Thus, adding such a parameter is "easier" than the approximation (3) would suggest. The disagreement follows from a lack of uniform convergence in the asymptotics which produce (3).

The example in the next section gives the explicit correspondence between the description length and two-part codes for Gaussian data with an unknown mean. This generic context provides the setting for a detailed look at the code length and so offers the opportunity to see the origin of the components that make up the *MDL* criterion (1). We have included in §2 several coding methods whose lengths approximate $\log^*$. Using codebooks, we demonstrate in §3 that the minimization of the description length produces a fixed threshold at $\pm c/\sqrt{n}$. We conclude in §4 with a brief summary discussion.

2 Example: Coding a normal mean

Suppose that the data are normally distributed $Y_i \overset{\text{iid}}{\sim} N(\mu_0, \sigma^2)$ with unknown mean μ_0 and $\sigma^2 = 1$. These data are to be compressed into a two-part code, whose first part indicates the value for μ used to encode the data which make up the second part of the code. Ignoring the issue of quantizing the data to some finite precision, the number of bits required to encode the Y_i in the second part of the code using parameter μ is

$$\begin{aligned} \log \frac{1}{P(Y|\mu)} &= \frac{n}{2}\log(2\pi) + \frac{\log e}{2}\sum_i (Y_i - \mu)^2 \\ &= \log \frac{1}{P(Y|\overline{Y})} + R_n(\mu - \overline{Y}) \,, \end{aligned} \tag{4}$$

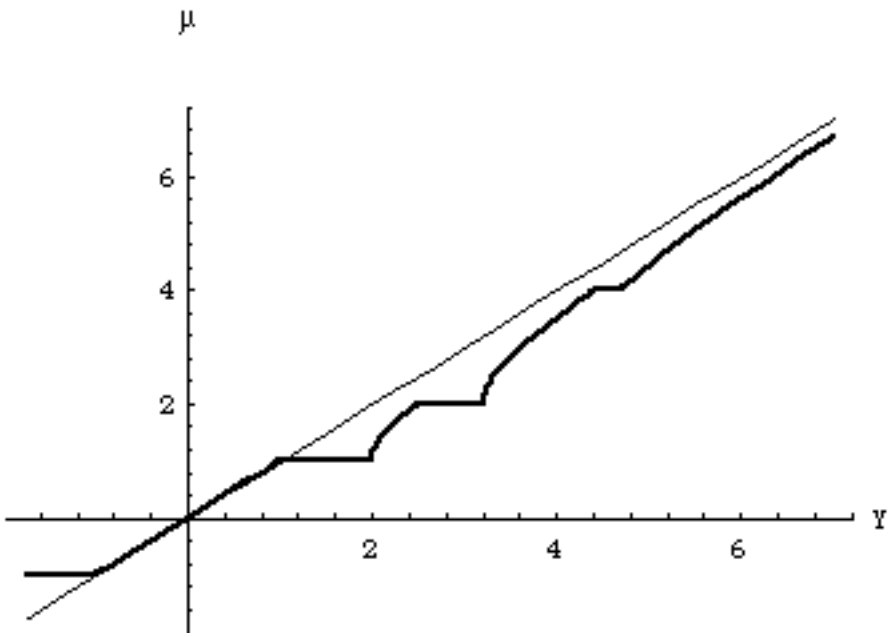
where the 'regret' for not using the maximum likelihood estimator $\overline{Y} = \sum Y_i/n$ is

$$R_n(\delta) = \frac{n\,\delta^2 \log e}{2} \,. \tag{5}$$

Here and elsewhere we ignore fractional bits in the calculations.

For those familiar with the often asserted connection between *MDL* and the *BIC* criterion, the estimator for μ which minimizes the description length is unexpected.

Figure 1: The estimator which minimizes the description length (1) is shown as a function of the mean of the input data, on a standardized scale. This estimator offers no shrinkage at the origin.



For this univariate problem, the leading term in the *MDL* criterion simplifies since $C(1) = 1$ and by using (4) we have

$$L_{n,1}(Y, \mu) = \log^* (\sqrt{n}\, \mu) + R_n(\mu - \overline{Y}) + \log \frac{1}{P(Y|\overline{Y})} \, . \tag{6}$$

To see the impact of this criterion on the choice of the best estimator, Figure 1 shows a plot of

$$\arg\min_{\mu} L_{n,1}(Y, \mu)$$

versus the input mean on a standardized scale defined by the standard error $\mathrm{SE}(\overline{Y}) = 1/\sqrt{n}$. For *MDL* to produce a parsimonious model, in this case a model with mean zero, the criterion needs to penalize non-zero values. As Figure 1 makes clear, this formulation of the description length produces no shrinkage near the origin since the penalty term $\log^*(x) = 0$ for $|x| < 1$. Flat spots in the graph originate from jumps in the $\log^*$ function when an additional summand appears.

In the rest of this paper, we show that a carefully formulated version of the description length does indeed imply shrinkage at the origin, but of limited magnitude.

The *MDL* estimator $\hat{\mu}$ defined in equation (15) of §3 below is shrunken to zero for $\sqrt{n}\,|\overline{Y}| < 2.4$. In contrast, the *BIC* criterion produces an estimate of zero for data with mean satisfying $\sqrt{n}\,|\overline{Y}| < \sqrt{\ln n}$. Our arguments require a very close accounting of the message length obtained in a two-part code for the data, and we now turn to these issues.

Although (4) implies that coding the data using $\mu = \overline{Y}$ produces the most data compression, we need to round the encoding parameter to finite precision in order to form the preamble of the two-part code. It turns out that we get a shorter overall message length by considerable rounding.

Without specific prior information that would imply a coding scheme, one is left with a somewhat arbitrary choice of how to encode the parameter value. Rissanen (1983) argues for the the use of an optimal universal representation of the integers, building on the work of Elias (1975). Suppose that we encode the data using a rounded parameter of the form $\tilde{\mu} = j/n^m$. The total code length obtained in this way is then

$$L_{n,1}(Y, \tilde{\mu}) = \log \frac{1}{P(Y|\tilde{\mu})} + L(j) \;, \tag{7}$$

where $L(j)$ denotes the length of the code for the integer j. Rissanen (1983) shows that if this length is that of an optimal universal code as defined by Elias (1975), then the overall code length is minimized by rounding the encoded parameter to terms of order $O(1/\sqrt{n})$ — that is, with $m = 1/2$ and $\overline{Y}$ rounded to a grid with spacing on the order of its standard error. The integer j defining $\tilde{\mu}$ is, in effect, the z score used when testing the null hypothesis $H_0 : \mu = 0$. The code length (7) thus reduces to

$$L_{n,1}(Y, \mu) = \log \frac{1}{P(Y|\overline{Y})} + \ell_n(\overline{Y}, \mu) \tag{8}$$

where the length in excess of the minimum determined by the log likelihood evaluated at the MLE is

$$\ell_n(\overline{Y}, \mu) = R_n(\mu - \overline{Y}) + L^*(\langle\, \sqrt{n}\,\mu \,\rangle) \;, \tag{9}$$

with $\langle\, x \,\rangle$ equal to the integer closest to x and L^* denoting the length of any such optimal code. Before moving on, we remark that rounding to this precision alters the length of the encoded data by much less than a single bit,

$$R_n(\mu - \langle\, \sqrt{n}\mu \,\rangle/\sqrt{n}) < \frac{\log e}{8} \approx 0.18 \;. \tag{10}$$

Thus, one can obtain a slightly shorter message by rounding to a more coarse grid. Such details have been discussed elsewhere (e.g., Wallace and Freeman 1987), and for our purposes any such grid with spacing to order $O(1/\sqrt{n})$ suffices.

Specific features of the universal representation determine which grid element provides the shortest code length. The need to encode the parameter does not imply simply rounding $\overline{Y}$ to the nearest grid element. Rather, one must round the MLE to minimize the excess length ℓ_n. Depending on the universal code being used, such rounding occasionally shifts the estimator because of changes in the length of codes for adjacent integers. Three universal codes for representing the parameter are illustrated in Table 1; all three are optimal in the sense of Elias (1975) who proposed and named the first two. Following the convention of Rissanen (1983), all of the codes assign the one bit symbol "0" to represent zero and treat the remaining integers symmetrically so that $L(j) = L(-j)$. Our examples show the codes for $|j|$; a trailing bit would be added for $j \neq 0$ to give its sign, adding one more bit to the descriptions in the table.

It is useful to consider the structure of several universal codes. The doubly compound code combines a prefix code of about $2 \log \log j$ bits for the length of the binary representation for $|j|$, followed by the binary representation itself. This notion of a code which combines the binary representation together with a prefix representation for $\log j$ is typical of optimal representations, and the doubly compound is perhaps the simplest of these and serves as an accessible example. Odd bits in the first portion of the representation shown in Table 1 indicate the number of bits in the binary representation of $|j|$. (Spaces in the table are useful for the human reader, but are not needed by the decoder.) The length of this code is then (with one added for the sign bit)

$$L_c^*(j) = 4 + \lfloor \log |j| \rfloor + 2 \lfloor \log(1 + \lfloor \log |j| \rfloor) \rfloor \quad j \neq 0 \,. \tag{11}$$

Elias (1975) offers some enhancements of this code which save several bits. The penultimate code offers a shorter representation for large integers. Each block in the penultimate code beginning with a 1 gives the binary representation for one plus the length of the following block, with the last block giving the bits for $j + 1$ (except in the case of the code for 0). A single zero bit denotes the end of the code, indicating that the previous block gives the sought value. The length of this code for $j \neq 0$ is (again,

adding one for the sign bit)

$$L_p^*(j-1) = 2+(1+\lfloor \log|j| \rfloor)+(1+\lfloor \log^{(2)}|j| \rfloor)+\cdots+(1+\lfloor \log^{(k)}|j| \rfloor) , \quad \log^{(k)}|j| \geq 1, \tag{12}$$

where terms are included in the sum so long as the k-fold iterated log (e.g., $\log^{(2)} x = \log\log x$) is at least one. Thus, L_p^* resembles a discretized version of $\log^*$. However, $L_p^*(j)$ is not a uniform approximation because it jumps by several bits at integers of the form $j = 2^{2^{\cdots}} - 1$, with the jump equal to the number of logarithmic summands in (12). Table 1 shows these jumps in comparing $L_p^*(2)$ to $L_p^*(3)$ and $L_p^*(14)$ to $L_p^*(15)$.

The final universal representation in Table 1 is similar to the penultimate code, but uses arithmetic coding to avoid sudden changes in code length. Arithmetic coding systematically determines an optimal code for a given probability distribution. (e.g. Cover and Thomas 1991, Chapter 5). Since (Rissanen 1983)

$$\sum_{j=1}^{\infty} 2^{-\log^* j} = c \approx 2.865 ,$$

we can associate a discrete probability measure with $\log^*$, Rissanen's (1983) universal prior for the integers. Defined for zero and treating negative and positive values symmetrically, the universal prior assigns probabilities as

$$Q^*(j) = \begin{cases} 1/2, & j = 0, \\ \frac{2^{-\log^*|j|}}{4c} & j \neq 0 . \end{cases} \tag{13}$$

Table 1 shows the codes produced by an arithmetic coder for this distribution. This implementation simply aligns the underlying encoded intervals on dyadic rationals so that the associated output code lengths are monotone. The properties of arithmetic coding guarantee that the length of the code L_a^* is within a bit of that implied by the underlying probabilities,

$$\sup_{j \in \mathbf{Z}} |\log 1/Q^*(j) - L_a^*(j)| < 1 , \tag{14}$$

avoiding the jumps in the lengths L_p^* of the penultimate code.

Table 1: *Examples of three optimal universal codes for nonnegative integers.* Spaces are for the reader and are not needed in the actual codes. A sign bit would be appended for $j \neq 0$. The doubly compound and penultimate codes are from Elias (1975); the third is an arithmetic coder for the probabilities $Q^*(|j|) = Q^*(j) + Q^*(-j),\ j \neq 0$.

j	$\log 1/Q^*(\|j\|)$	Doubly Compound	Penultimate Code	Arithmetic Code for Q^*
0	1.0	0	0	0
1	2.5	10 1	10 0	100
2	3.5	1100 10	11 0	1010
3	4.8	1100 11	10 100 0	10110
4	5.5	1110 100	10 101 0	101110
5	6.3	1110 101	10 110 0	1011110
6	6.9	1110 110	10 111 0	1011111
7	7.4	1110 111	11 1000 0	11000000
8	7.8	110100 1000	11 1001 0	11000001
14	9.2	110100 1110	11 1111 0	1100010001
15	9.4	110100 1111	10 100 10000 0	1100010010
...				
256	15.8	17 bits	16 bits	16 bits
1024	18.4	19 bits	18 bits	19 bits
65534	25.5	26 bits	23 bits	26 bits
65535	25.5	26 bits	28 bits	26 bits

3 Model selection via *MDL*

The best two-part code for the data and the *MDL* estimator for μ are together determined by minimizing the excess bit length ℓ_n. The *MDL* estimator is thus

$$\hat{\mu} = \arg \min_{\mu \in \mathbf{Z}/\sqrt{n}} \ell_n(\overline{Y}, \mu) \ , \tag{15}$$

where the minimization is over discrete parameter values in the set $\mathbf{Z}/\sqrt{n} = \{j/\sqrt{n} : j \in \mathbf{Z}\}$. We find that this minimization and the associated rounding is most easily understood graphically. Each quadratic shown in Figures 2 and 3 indicates the excess bit length $\ell_n(\overline{Y}, j/\sqrt{n})$ obtained when data with mean $\overline{Y} = z/\sqrt{n}$ are coded with parameter $\mu = j/\sqrt{n}$. In Figure 2, the height of the base of each quadratic is displaced by $L_p^*(j)$, reflecting the code lengths of the penultimate code. Were it the case that $\overline{Y} = j/\sqrt{n}$ and the data coded with $\mu = j/\sqrt{n}$, then $L^*(j)$ bits would be needed to encode this parameter value. Figure 3 presumes that μ is coded using the arithmetic code with length L_a^*. For example, suppose $z = \overline{Y} = 0$. Then the minimum excess length in both cases is one bit and $\hat{\mu} = 0$. If the mean is $\overline{Y} = 1/\sqrt{n}$, one standard error above zero (z=1) , coding with $\mu = 0$ inflates the message length by about 1.6 bits, one bit for coding zero and 0.6 bits due to the relative entropy. However, the excess length obtained by coding $\mu = 1/\sqrt{n}$ is longer, $\ell_n(1/\sqrt{n}, 1/\sqrt{n}) = L_p^*(1) = 4$, even though the latter codes using the "true" parameter value so that the relative entropy component of ℓ_n is zero. For $z = 2$, the minimum code length is again obtained by coding $\mu = 0$. For both $z = 1$ and $z = 2$, the number of bits required to encode a non-zero parameter in the preamble is greater than the corresponding gain in data compression because $L^*(0) = 1$ is so much less than the lengths $L^*(1) = 4$. Thus, the *MDL* estimate remains $\hat{\mu} = 0$ for $|z| < 2$, the region in the figures for which the quadratic centered at zero attains the minimum.

Now consider the implications for model selection. One approach is to consider how the universal code lengths penalize non-zero estimates and shrink toward the origin, as discussed in §2. Figure 4 shows the plot of $\hat{\mu}$ when the integer z score for the parameter is encoded using the arithmetic version of the universal code. It is useful to contrast this figure with Figure 1 defined from (1). The *MDL* estimator $\hat{\mu} = 0$ over the region shown in Figure 3 where the quadratic center at the origin provides the minimum code

length. That is,

$$\hat{\mu} = 0 \quad \Longleftrightarrow \quad L_a^*(0) + R_n(z) < L_a^*(2) + R_n(z-2) \; ,$$

or where $|z| < 1 + 2/\log e \approx 2.4$. The remaining stairstep features of the estimator are due to the discrete rounding needed for coding the estimator. For data with a mean just exceeding this value, we would code with $\hat{\mu} = 2$, in effect rejecting $H_0 : \mu_0 = 0$. This procedure resembles the decision rule of the usual statistical test which rejects H_0 when $|z| > 1.96$. Thus a strict code-length interpretation of the *MDL* principle implies that the parameter penalty is fixed on the z-score scale and does not grow with the sample size n.

An alternative approach (Rissanen 1983) is to compare $L_{n,1}$ to the description length of the null model which forces $\mu = 0$. Since the null model does not require encoding of a parameter, its length function is simply

$$L_{n,0}(Y) = \log \frac{1}{P(Y|\overline{Y})} + R_n(\overline{Y}) \; .$$

Thus, a one-parameter model obtains a shorter code length than the null model, $L_{n,1}(Y,\mu) < L_{n,0}(Y)$, when

$$\ell_n(\overline{Y}, \hat{\mu}) < R_n(\overline{Y}) \; ,$$

or, in terms of the z score and arithmetic code, whenever

$$|z| = |\sqrt{n}\,\overline{Y}| > 1 + \frac{2.5}{\log e} \approx 2.7 \; . \tag{16}$$

This cutoff differs slightly from that implied by shrinkage since the comparison of $L_{n,0}$ to $L_{n,1}$ avoids the single bit needed for coding zero. In general, one obtains a slightly different cutoff value depending upon the specifics of the implementation of the universal code. In each case, however, the cutoff is fixed on the standard error scale.

Remark. Since the quadratics associated with coding $\hat{\mu} = \pm 1/\sqrt{n}, \pm 3/\sqrt{n}$ in the penultimate code never attain the lower boundary in Figure 2, the corresponding codes ($j = \pm 1, \pm 3$) would never be used. Thus, the coding procedure just described will be somewhat inefficient in that it would not use some of the available symbols. The code can be improved by using a more coarse coding grid, as noted previously. Such a

change might alter the critical value in the decision rule (16), most likely increasing the threshold slightly.

Universal length functions like L_p^* or L_a^* are rather unwieldy to manipulate, and this complexity suggests other approximations. As shown in Rissanen (1983, Theorem 2), the length function for any optimal universal code for the integers is bounded as

$$\log j < L^*(j) < \log j + r(j) \tag{17}$$

where $r(j)/\log j \to 0$ as $|j| \to \infty$. In this sense, the lengths of all of the optimal universal codes are logarithmic. This property, together with the ease of manipulating log rather than $\log^*$, has led to the most common approximation to the code length $L_{n,k}$. It is this approximation, rather than an intrinsic property of the *MDL* principle itself, that leads to a logarithmic parameter penalty.

In various papers (e.g., 1983 §4, 1986, 1989), Rissanen approximates $L^*(\langle\, z \,\rangle)$ as $\log \langle\, z \,\rangle$, where again we denote $z = \sqrt{n}\,\overline{Y}$. In the context of coding a mean, the excess length is then about

$$\ell_n(Y, \mu) \approx \log \mu + (\log n)/2 + R_n(\mu - \overline{Y}) \; . \tag{18}$$

If we fix μ and take the limit of the approximation as $n \to \infty$, we obtain the *BIC* penalty

$$\min_{\mu \in \mathbf{Z}/\sqrt{n}} \ell_n(\overline{Y}, \mu) \approx \log |\sqrt{n}\,\overline{Y}| = \frac{1}{2} \log n + O_p(1) \; . \tag{19}$$

Under these conditions, z grows with n and one can exploit the relationship that $L^*(\langle\, z \,\rangle) - \log z = o(\log z)$ implied by (17). Interpreted as a coding procedure, this approximation is the code length that we would obtain were the parameter space compact, say $|\mu| < M/2$, and each parameter value on a grid over this space coded with $(1/2)\log n + \log M$ bits. That is, coding with a discrete uniform prior over a grid with spacing $1/\sqrt{n}$ on the parameter space. This approximation thus leads to a fixed code length for representing a parameter rather than the varying length implied by $L^*(\langle\, z \,\rangle)$. It exacts a large penalty for any parameter, regardless of how close that parameter lies to zero.

As a coding procedure, the use of a fixed-length logarithmic code has advantages. In particular, one can show that the code length obtained by this representation (over a compact parameter space) is about as short as possible. The logarithmic penalty

provides a lower asymptotic bound for the excess length ℓ_n. For example in the mean coding problem we have been discussing, let Ω denote a compact subset of $\mathbf{R}$ and let A_n denote a set whose measure tends zero as $n \to \infty$. Then for all $\mu \in \Omega - A_n$ and any $\epsilon > 0$, it follows from Rissanen (1989, Theorem 3.1) that there exists an n such that

$$E_n \inf_{\mu} \ell_n(\overline{Y}, \hat{\mu}) \geq \frac{1-\epsilon}{2} \log n \,, \tag{20}$$

where the expectation E_n is with respect to the density of $Y_1, \ldots, Y_n$. These are powerful results; however, issues of model selection are only relevant for the set of small z scores near the origin. Although the absolute size of this set diminishes with increasing sample size (and so can be ignored in this theorem as the set A_n), its size remains fixed on a standard error scale. The perspective of using asymptotics on a fixed standard error scale (so that $\sqrt{n}\mu$ is constant as $n \to \infty$) is not novel and forms the essential ingredient of so-called local asymptotics as advocated by LeCam (e.g., LeCam and Yang 1990) and Ibragimov and Hasminskii (1981).

Returning to model selection, the excess length of the null model, $R_n(\overline{Y})$, is shorter than this approximation to ℓ_n unless (with $M = 1$)

$$z^2 > \frac{\log n}{\log e} = \ln n \,. \tag{21}$$

In comparison to the decision rule generated by direct application of *MDL*, this approximation tests $H_0 : \mu = 0$ by comparing the classical test statistic $\sqrt{n}\,\overline{Y}$ to an increasing critical value. This type of logarithmic penalty is also found in the *BIC* criterion introduced by Schwarz (1986). As suggested by the illustration of the previous section, this approximation is not accurate for small z. In particular, the approximation (19) is not uniform in μ and, almost surely,

$$\lim_{n\to\infty} \frac{\ell_n(\overline{Y}, \langle\, \sqrt{n}\,\overline{Y}\,\rangle/\sqrt{n})}{\log n} = \begin{cases} 1/2, & \mu \neq 0 \,, \\ 0, & \mu = 0 \,. \end{cases} \tag{22}$$

Thus, one cannot locate the *MDL* estimator $\hat{\mu}$ by minimizing this approximation to the description length when μ is near zero.

4 Discussion and summary

The decision of whether to code the mean parameter μ is resolved within a vanishing set of parameter values $|\mu| \leq c/\sqrt{n}$ near the origin. Once the absolute z score $\sqrt{n}\,\overline{Y}$

is about 2.4, the description length for the model is shorter when this parameter is included than when it is forced to zero. The use of *MDL* for testing a single parameter thus leads to a decision rule that resembles a traditional hypothesis test: there is a fixed threshold lying about 2 standard errors from the origin rather than a threshold which grows with the logarithm of the sample size.

This discrepancy from a logarithmic penalty arises because standard approximations for *MDL* overstate the model cost for parameters near the origin. For example, consider a model whose parameter lies on the boundary suggested by the approximate description length $L_{n,1}(Y) \approx (1/2)\log n + \log 1/P(Y|\overline{Y})$, namely $z = \sqrt{\log n}$. The description length of such a model is considerably less than that implied by the approximation,

$$\begin{aligned} L_{n,1}(Y) &= \log^* z + \log \frac{1}{P(Y|\overline{Y})} + \epsilon \\ &= (1/2)\log\log n + \log \frac{1}{P(Y|\overline{Y})} + o(\log\log n) \ , \end{aligned}$$

for $|\epsilon| < 1$. Since the contribution of the parameter to the description length at this point is less than $(1/2)\log n$, it raises the issue of where one should begin coding, which is resolved in §3.

References

Foster, D. P. and R. A. Stine (1997). An information theoretic comparison of model selection criteria. Discussion paper 1180, Center for Mathematical Studies in Economics and Management Science, Northwestern University.

Le Cam, L. and G. Yang (1990). *Asymptotics in Statistics*. Springer, New York.

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

Elias, P. (1975) Universal codeword sets and representations of the integers. *IEEE Transactions on Information Theory*, **21**, 194-203.

Ibragimov, I. A. and R. Z. Has'minskii (1981). *Statistical Estimation*. Springer, New York.

Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *Annals of Statistics*, **11**, 416-431.

— (1986). Minimum description length. In *Encyclopedia of Statistics*, Kotz and Johnson, Ed, pp . Wiley, New York.

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

Schwarz, G. (1986). Estimating the dimension of a model. *Annals of Statistics*, **6**, 416-446.

Wallace, C. S. and P. R. Freeman (1987). Estimation and inference by compact coding. *J. Royal Statistical Society Ser. B*, **49**, 240-252.

Figure 2: *The penultimate codebook.* Quadratics indicate the excess message length above $\log 1/P(Y|\overline{Y})$ for estimates $\hat{\mu} = j/\sqrt{n}$ when the parameter is encoded using the penultimate code.

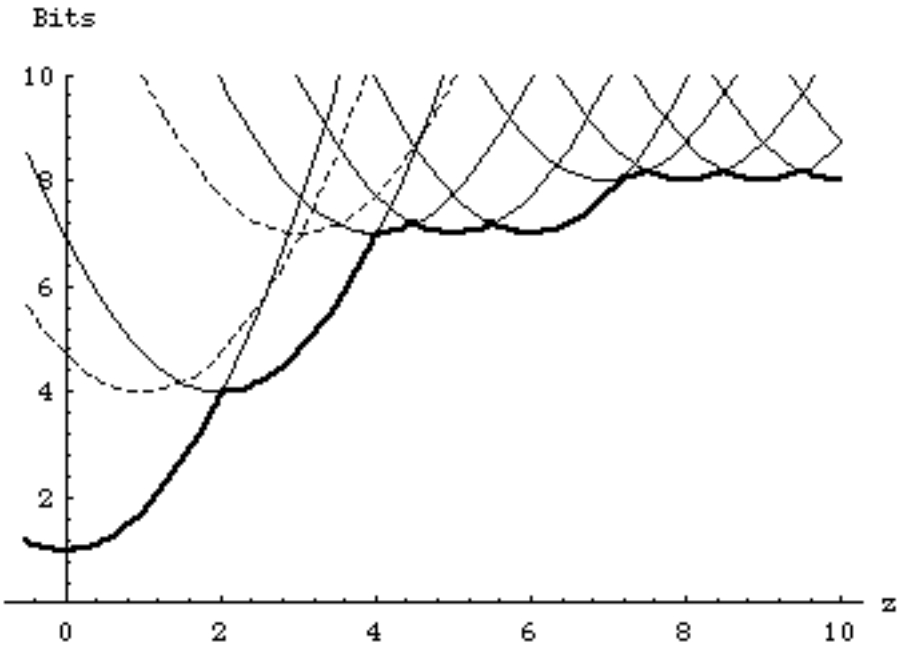


Figure 3: *The arithmetic codebook.* Quadratics indicate the excess message length above $\log 1/P(Y|\overline{Y})$ for estimates $\hat{\mu} = j/\sqrt{n}$ when the parameter is encoded using the arithmetic code for Q^*.

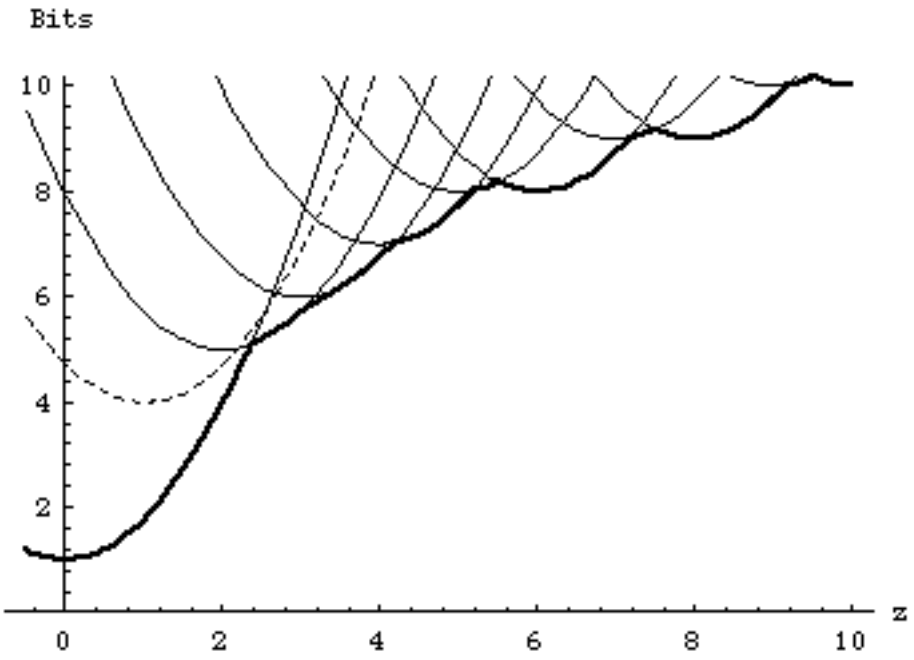


Figure 4: The estimator which minimizes the code length with penalty L_a^* is shown as a function of the mean of the input data, on a standardized scale. The estimator shrinks values to zero when $|z| < 1 + 2/\log e$.