Information Theory and Model Selection

Dean Foster & Robert Stine
Dept of Statistics, Univ of Pennsylvania

May 11, 1998

- Data compression and coding
- Duality of code lengths and probabilities
- Model selection via coding: testing $H_0 : \mu = 0$
  - Local asymptotic coding
- Coding interpretation of selection criteria in regression:
  - Mallows' C_p, Akaike information criterion (AIC)
  - Bayesian information criterion (BIC, SIC)
  - Risk inflation, thresholding (RIC)
  - Empirical Bayes criterion, multiple testing ($eBIC$)
- Discussion, extensions

Overview

Ultimate problem for today

Which variables ought to be used in a regression, particularly when the number of potential predictors p is large (data mining).

Model selection = data compression

Model selection via popular criteria

$$AIC,\ BIC,\ RIC,\ eBIC$$

is equivalent to choosing the model which offers the greatest compression of the data.

Two-part codes

The compressed data are represented by a two-part code

Model Parameters || Compressed Data

Selection criteria differ in how they encode the parameters.

Information/coding theory

Coding view of selection as data compression offers

- Consistent, alternative perspective for the various criteria.
- Tangible comparison of criteria.
- Suggests new criteria, customized to specific problems.

Representative problem

Test the null hypothesis $\mu_0 = 0$.

So many choices — is any one right?

Context: orthogonal regression with n observations and p predictors.

Threshold: choose X_j if $|z_j| > \tau$, criterion's threshold.

$\tau = 0$	OLS, max R^2	Gauss
$\tau = 1$	$\max \overline{R}^2$, $\min s^2$	Theil 1961
$\sqrt{2}$	*Unbiased est of out-of-sample error*	
	C_p	Mallows 1964,1973
	AIC	Akaike 1973
	Cross-valid	Stone 1974
$\sqrt{\log n}$	*Model averaging*	
	BIC, *SIC*	Bayes (Schwarz 1978)
	"*MDL*"	Inf. thry. (Rissanen 1978)
$\sqrt{2 \log p}$	*Minimax risk (Bonferroni)*	
	RIC	Foster & George 1994
	Wavethresh	Donoho & Johnstone 1994
$\sqrt{2 \log p/q}$	*Adaptive selection*	
	eBIC	Foster & George 1996
	Mult tests	Benjamini & Hochberg 1996

Data Compression

File compression

Disk compression utilities: WinZip, Stacker, Stuffit, compress.

How do they work?

How to compress a file of characters into a sequence of bits (0's and 1's) without losing information (lossless compression)?

Sample problem

File (message) composed of 4 characters: *a, b, c, d.*

What would you need to know in order to compress a file of these characters?

Question rephrased

View file as a sequence Y_1, Y_2, $\ldots, Y_n$ of *iid* discrete r.v.'s,

$$Y_1,\ Y_2, \ldots, Y_n \overset{\text{iid}}{\sim} p(y)\ .$$

Let $\ell(y)$ denote code length for y. What is the smallest compressed file length (on average),

$$\min_{\ell} E \sum_{i=1}^{n} \ell(Y_i) = n\ \min_{\ell} E\, \ell(Y_1)\ ,$$

and what code achieves this limit?

Alternative Coding Methods

Two codes

- Code I: a fixed-length code (like ASCII, but with 2 bits each)
- Code II: a variable-length code, matching length to exponent

Symbol y	$p(y)$	Code I	Code II
a	$1/2 = 1/2^1$	00	0
b	$1/4 = 1/2^2$	01	10
c	$1/8 = 1/2^3$	10	110
d	$1/8 = 1/2^3$	11	111

Examples

String	P(String)	Code I	Code II
baa	$\frac{1}{4}\frac{1}{2}^2 = \frac{1}{2^4}$	010000	1000
dad	$\frac{1}{8}\frac{1}{2}\frac{1}{8} = \frac{1}{2^7}$	110011	1110111

Prefix codes and delimiters

- Unlike Morse codes, neither code requires a delimiter.
- Code II is a "prefix code"; the code for no symbol is a prefix to any other. Despite varying length, such codes are 'instantaneous'.

Optimal Code?

Symbol y	$p(y)$	Code I	Code II
a	$1/2 = 1/2^1$	00	0
b	$1/4 = 1/2^2$	01	10
c	$1/8 = 1/2^3$	10	110
d	$1/8 = 1/2^3$	11	111

Expected lengths

For $Y \in \{a, b, c, d\}$, the length for Code I is fixed, $E\,\ell_1(Y) = 2$, whereas for Code II,

$$E\ \ell_2(Y) = 1(\frac{1}{2}) + 2(\frac{1}{4}) + 3(\frac{1}{8}) + 3(\frac{1}{8}) = 1.75 < E\ \ell_1(Y)$$

Big question

Can you do any better?

Specifically, retaining the assumptions of *i.i.d.* data,

- independence and
- identical distribution (strong stationarity),

is there a code with shorter average length than Code II?

Kraft Inequality

Code length implies sub-probability

For any instantaneous binary code over discrete symbols y, assigning length $\ell(y)$ to the symbol y,

$$\sum_y 2^{-\ell(y)} \leq 1 \ .$$

Tree-based interpretation for Code II

- Associate probability $2^{-\text{depth}}$ with each leaf node.
- Code for a symbol determined by the sequence of left branches (0) and right branches (1) followed to the node.
- Inequality since you need not use all branches.

Optimal Codes

Entropy determines minimum bit length

The minimum expected number of bits needed to encode a discrete r.v. $Y \sim p(y)$ is

$$H(Y) \le E\,\ell(Y) < H(Y) + 1\,,$$

where the entropy $H(Y)$ is defined (all logs are base 2)

$$H(Y) = E \underbrace{\log 1/p(Y)}_{\text{opt len}} = \sum_y \left(\log \frac{1}{p(y)}\right) p(y)$$

Relative entropy (*aka*, Kullback-Leibler divergence)

A 'distance' between two probability distributions $p(y)$ and $q(y)$,

$$D(\underbrace{p}_{\text{truth}} \,\|\, \underbrace{q}_{\text{fit}}) = \sum_y \left(\log \frac{p(y)}{q(y)}\right) p(y) \ge 0$$

with the inequality following from Jensen's inequality.

Interpretation of relative entropy

Suppose the true distribution is $p(y)$ but we use a code based on the wrong model $q(y)$. Then the expected cost in excess bits is the relative entropy,

$$\sum_y \left(\log \frac{1}{q(y)} - \log \frac{1}{p(y)}\right) p(y) = \sum_y \left(\log \frac{p(y)}{q(y)}\right) p(y) = D(p\|q)$$

Derivations

Why does entropy give the limit?

The entropy bound

$$H(Y) \leq E\,\ell(Y) < H(Y) + 1\ ,$$

is a consequence of:

- Kraft inequality: $\sum_y 2^{-\ell(y)} \leq 1$
- Relative entropy: $D(p\|q) \geq 0$

Proof outline

For any code with lengths $\ell(y)$ associate the sub-probability $q(y) = 2^{-\ell(y)}$ and define $c \geq 1$ such that $\sum_y c\,q(y) = 1$.

Then for the lower bound,

$$\begin{aligned}
E\ \ell(Y) - H(Y) &= \sum_y (\ell(y) + \log p(y))\,p(y) \\
&= \sum_y (\log 1/q(y) + \log p(y))\,p(y) \\
&= \sum_y (\log 1/(c\,q(y)) + \log p(y))\,p(y) + \log c \\
&= D(p\|c\,q) + \log c \geq 0
\end{aligned}$$

The upper bound follows by using a code with length $\ell(y) = \lceil \log 1/p(y) \rceil < 1 + \log 1/p(y)$. Such a code may be obtained by Huffman coding or arithmetic coding.

Arithmetic Coding

Goal

Generate a prefix code for a discrete random variable,

$$Y \sim p(y) \;, \quad y = 0, 1, \ldots \qquad P(y) = \sum_{j \le y} p(j)$$

Assume probabilities are monotone, $p(y) \ge p(y+1)$.

Approach Rissanen & Langdon

Partition unit interval $[0, 1]$ according to $P(y)$. How many bits does it take to uniquely identify the interval associated with y?

Key step

Recursively refine a binary partition, until "fractional" binary value uniquely indicates the interval asociated with y.

Issues

Unless $p(y) = 2^j$

- Not typically Kraft tight.
- Not always monotone (ie, $p(y) > p(x)$ but $\ell(y) > \ell(x)$).

Example

On next page...

Example of Arithmetic Coding

y	$p(y)$	$P(y)$	$\log p(y)$
0	0.55	0.55	0.9
1	0.25	0.80	2
2	0.15	0.95	2.7
3	0.05	1.00	4.3

Summary of Relevant Coding Theory

Entropy

Entropy determines min expected message length (discrete),

$$\min_{\ell} E \sum_{i=1}^{n} \ell(Y_i) = nH(Y), \quad H(Y) = \sum_y \left(\log \frac{1}{p(y)} \right) p(y)$$

Optimal obtained (within one bit) using a code with lengths

$$\ell(y) = \log \frac{1}{p(y)}$$

Implications

- High compression requires short codes for likely symbols.
- Kraft-tight codes are synonymous with pdfs,

$$p(y) = 2^{-\ell(y)}$$

Relative entropy

Cost for coding using wrong model is $nD(p\|q)$ bits, where

$$D(p\|q) = E_p(\log p / \log q) = \sum_y \underbrace{\left(\log \frac{p(y)}{q(y)} \right)}_{\text{log L.R.}} p(y) \geq 0$$

Achievable?

Yes, within one bit on average, via arithmetic coding.

Coding Bernoulli Random Variables

Bernoulli observations

Suppose data consists of n Bernoulli r.v.'s,

$$Y_1, \ldots, Y_n \sim B(p), \quad k = \sum_i Y_i, \quad \hat{p} = k/n$$

How can you compress a Boolean?

Since each Y_i is just a bit, how can you compress anything?

Code $Y = (Y_1, \ldots, Y_n)$ as a *block*, using joint density

$$p_n(Y) = \prod_i p(Y_i) = p^k (1-p)^{n-k} \ .$$

Coding efficiency

Optimal code compresses n bits down to $n\,H(\hat{p})$

- $n\,H(1/2) = n$
- $n\,H(1/8) \approx n/2$
- $n\,H(1/n) \approx \log n \qquad \Leftarrow$ give its index

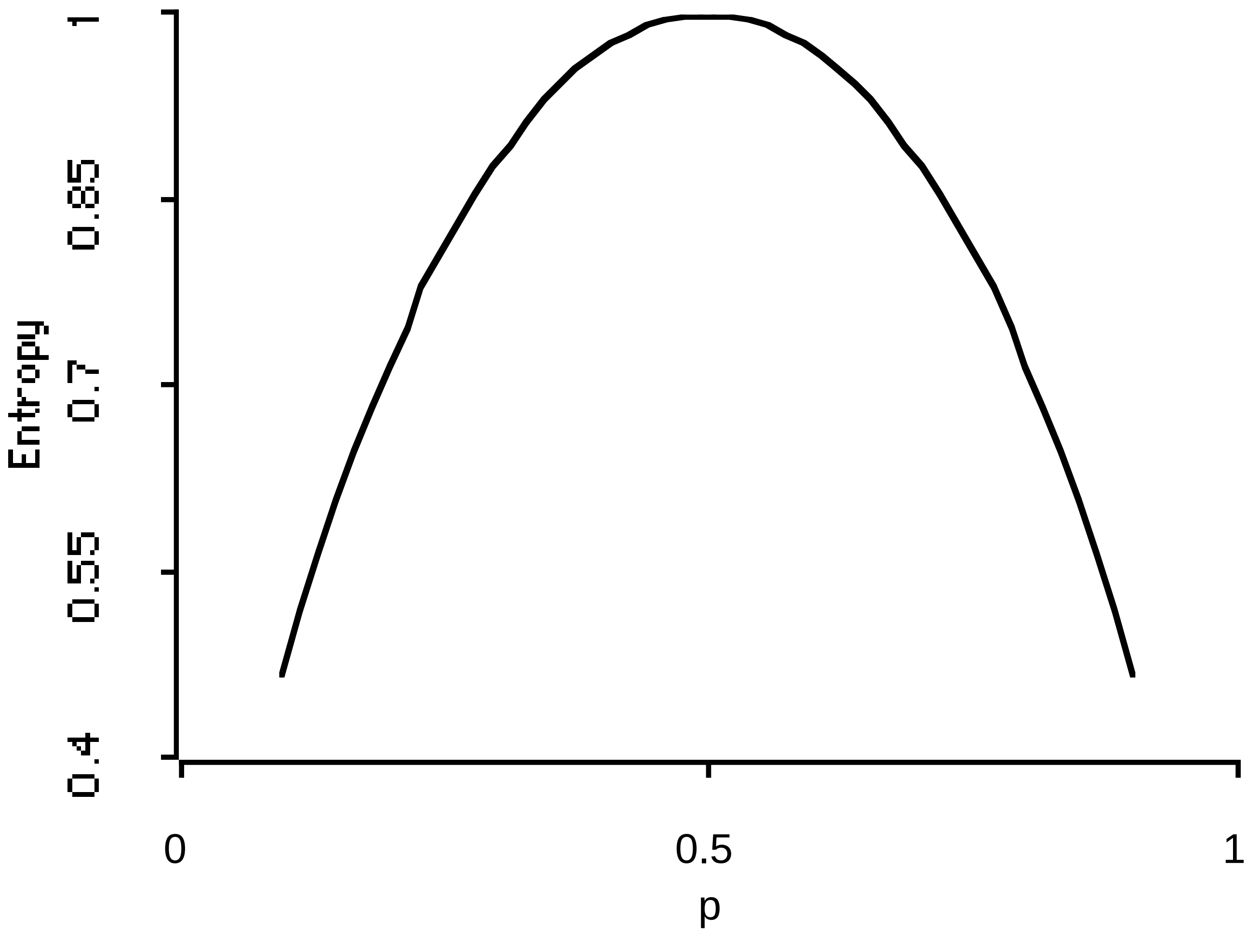
Log-likelihood

Log-likelihood determines the compressed length

$$\begin{aligned} n\ H(\hat{p}) &= n \left(\hat{p} \log \frac{1}{\hat{p}} + (1-\hat{p}) \log \frac{1}{1-\hat{p}} \right) \\ &= k \log \frac{1}{\hat{p}} + (n-k)\ \log \frac{1}{1-\hat{p}} = \log \frac{1}{P(Y|\hat{p})} \end{aligned}$$

Bernoulli Entropy Function

$$H(p) = p \log p + (1-p) \log(1-p)$$
$$\approx 1 - 3(p - \tfrac{1}{2})^2$$



Coding Continuous Random Variables

Continuous data?

Solution is to 'quantize', rounding to a discrete grid.

Relative entropy for quantizing

Continuous r.v. Y rounded to precision 2^{-Q} requires

$$H(Y) + Q \qquad \text{bits, on average.}$$

Net effect: add a constant number of bits for each obs.

Normal data compression

$Y_1, \ldots, Y_n \sim N(\mu, 1)$ with mean $\overline{Y} = \sum_i Y_i/n$.

$$\text{Minimum bits} = \underbrace{\log 1/P(Y|\overline{Y})}_{\text{log-like at MLE}} + \underbrace{n\,Q}_{\text{quantized}}$$

Relative entropy and testing

Additional bits if we code with m as the mean rather than the MLE, (known as the 'regret')

$$\begin{aligned} R_n(m - \overline{Y}) &= \log \frac{P(Y|\overline{Y})}{P(Y|m)} \\ &= \frac{n(m - \overline{Y})^2}{2 \ln 2} = \frac{z_m^2}{2 \ln 2} \end{aligned}$$

where $z_m = \sqrt{n}(\overline{Y} - m)$ is the test statistic for $H_0 : \mu = m$.

Normal Location Problem

Task

Transmit $Y_1, \ldots, Y_n \sim N(\mu, 1)$ to a receiver using as few bits as possible. Receiver knows $Y_i \sim N(\cdot, 1)$ and n, but nothing else.

Complication

If we encode the data using the optimal code defined by $P(Y|\overline{Y})$, the receiver will need $\overline{Y}$ in order to decode the message.

Solution via a two-part code

- Add $\overline{Y}$ as a prefix to the message, then
- Compress data into $\log 1/P(Y|\overline{Y})$ bits (ignore quantization).

$$\text{Total message length} = \underbrace{\text{Parameter Prefix}}_{?} + \underbrace{\text{Compressed Data}}_{\log\ 1/P(Y|\overline{Y})}$$

How to represent $\overline{\mathbf{Y}}$ in the prefix?

Quantizing suggests rounding $\overline{Y}$ to some precision. Rissanen shows that rounding $\overline{Y}$ to SE scale is optimal,

$$\hat{\mu} = \frac{\langle \sqrt{n}\,\overline{Y} \rangle}{\sqrt{n}} = \frac{\langle z_0 \rangle}{\sqrt{n}}\ ,$$

adding less than one bit to data since $R_n(\hat{\mu} - \overline{Y}) < 1$.

- How to represent the *integer* z-score, $\langle z_0 \rangle = \langle \sqrt{n}\,\overline{Y} \rangle$?
- Can you be clever if $\overline{Y}$ is near zero?

Bayesian Perspective

How to represent the rounded z-score?

How to encode rounded z_0 from $\hat{\mu} = \langle z_0 \rangle / \sqrt{n}$.

Bayesian view

Code choice for z_0 implies a prior probability,

$$\begin{aligned} \text{Total length} &= \text{Parameter Prefix} + \text{Compressed Data} \\ &= \log 1/P(\mu) + \log\ 1/P(Y|\mu) \\ \Rightarrow \quad P(Y,\mu) &= P(\mu) \times P(Y|\mu) \\ &= \underbrace{\text{Prior for } \mu}_{?} \times \text{Likelihood} \end{aligned}$$

Universal prior Elias 1975, Rissanen 1983

- Code "as well as" true distribution, assuming monotonicity
- Robust, proper prior roughly comparable to a *log-Cauchy*

How to represent $\overline{\mathbf{Y}}$ in the prefix?

- Find the integer z score that produces the shortest message, maximizing the joint probability.
- Total message length is

$$\underbrace{\ell[U_s(z)] + R_n\left(\frac{z}{\sqrt{n}} - \overline{Y}\right)}_{\text{arg min } z} + \log \frac{1}{P(Y|\overline{Y})}$$

Universal Priors

Simple example

Interleave continuation bits with binary form,

$$5 = 101_2 \quad \Rightarrow \quad 1\mathbf{1}\ 0\mathbf{1}\ 1\mathbf{0}$$

Length is roughly $2 \log z$, implying $p(z) \approx 1/z^2$, or Cauchy-like tails.

Recursive log

Send a sequence of blocks,each giving length of next. Define

$$\log^* x = \log x + \log\log x + \log\log\log x + \cdots$$

where sum includes only positive terms. Series is summable,

$$\sum_{j=1}^{\infty} 2^{-\log^* j} \approx 2.8 = 2^{1.5} < \infty$$

Probabilities

Define $p^*(0) = 1/2$ and for $j = 1, 2, 3, \ldots,$

$$p^*(j) = 2^{-(\log^* j + 2.5)} = c \times \left(\frac{1}{j}\right) \times \frac{1}{\log j} \times \frac{1}{\log\log j} \times \cdots$$

Very, very thick tails

$$\log^*(x) \approx \log x + 2\log\log x \Rightarrow \log \text{ Cauchy}$$

Universal Codes

j	Cauchy	$U(j)$	$\ell[U(j)]$
0	0	0	1 bit
1	10	100	3
2	1100	1010	4
3	1110	10110	5
4	110100	101110	6
5	110110	1011110	7
6	111100	1011111	7
...			
100	14 bits	14 bits	
1000	20	19	
10000	28	23	

- Length of Cauchy code is $2 \log j$
- Length $\ell[U(x)] = c + \log\langle x\rangle + \log\log\langle x\rangle + \log\log\log\langle x\rangle + \cdots$, with rounding embedded, $U(x) = U(\langle x\rangle)$.
- Signed universal appends sign bit, $U_s(j) = U(j) \parallel (+/-)$

Optimal Parameter Code

Optimal estimate

$$\hat{\mu} = z/\sqrt{n}, \quad \arg\min_z \ell[U_s(z)] + R_n(z/\sqrt{n} - \overline{Y})$$

Table on SE grid

$\overline{Y}$	$z = 0$	1	2	3	4
0	**1.0**	4.7	7.9	12.5	18.5
$1/\sqrt{n}$	**1.7**	4.0	5.7	8.9	13.5
$2/\sqrt{n}$	**3.9**	4.7	5.0	6.7	9.9
$3/\sqrt{n}$	7.5	6.9	**5.7**	6.0	7.7
$4/\sqrt{n}$	12.5	10.5	7.9	**6.7**	7.0

Note

- Code a non-zero parameter once $|z| > 2.4$.
- Decision rule resembles familiar normal test.
- Shrinkage stops once $|z| = \sqrt{n}\,\overline{Y} > 5$.

Reference

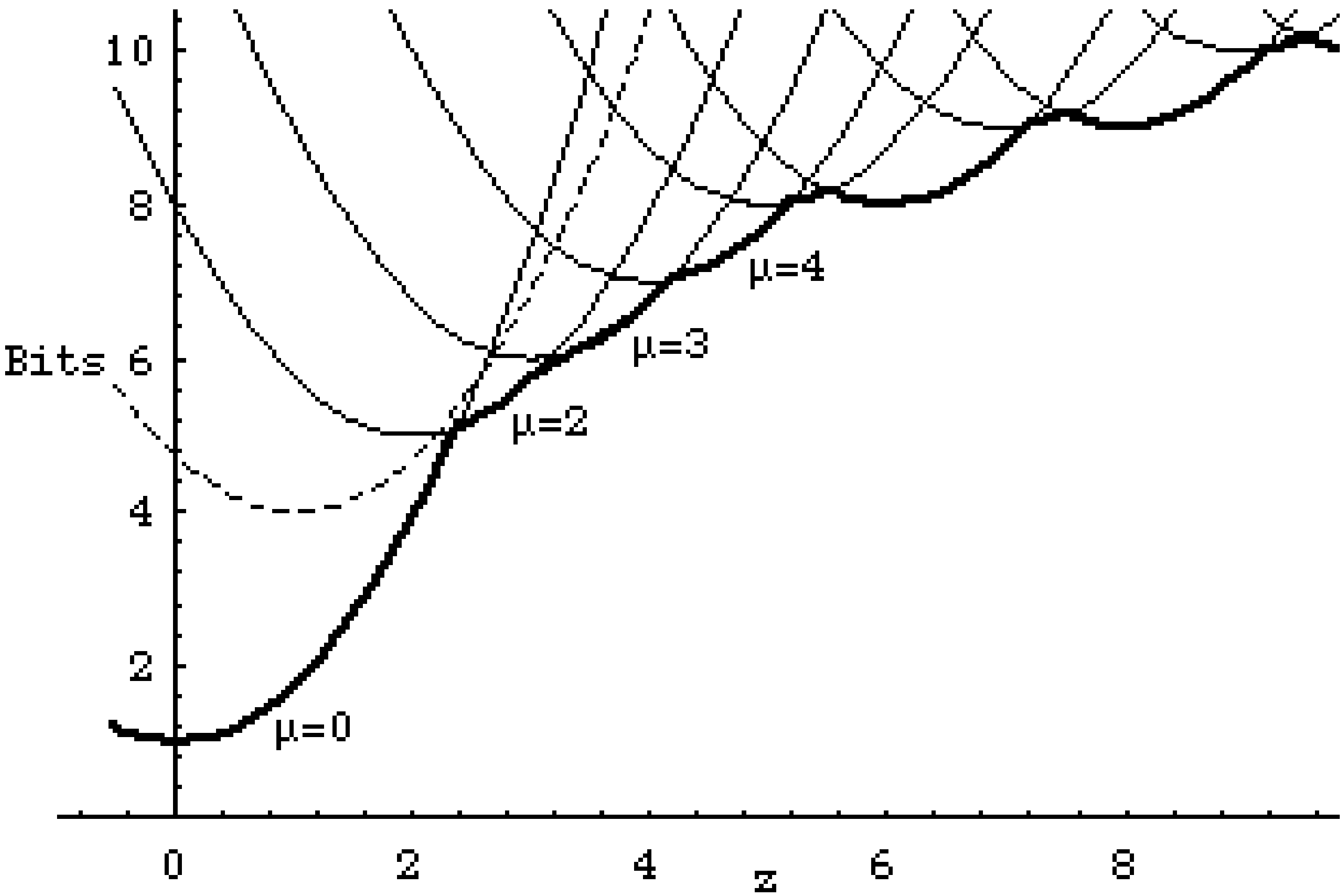
"Local asymptotics and the minimum description length"

http:www-stat.wharton.upenn.edu/~bob

Graph of Codebook

Vertical Axis: Bits are the excess $\ell[U_s(z_\mu)] + R_n(\overline{Y} - \hat{\mu})$ over minimum determined by the log likelihood at $\overline{Y}$.

Horizontal Axis: $z = \sqrt{n}\,\overline{Y}$, the usual z-score.



Alternative Asymptotic Analysis

Asymptotic code length Rissanen's *MDL* (1983)

Asymptotic analysis of optimal code length, with $n \to \infty$ and $\mu = EY$ fixed so that $z = \sqrt{n}\,\overline{Y}$ is large:

$$\begin{aligned} \text{Code length} &= \ell[U_s(\sqrt{n}\,\overline{Y})] + \log \frac{1}{P(Y|\overline{Y})} + c \\ &\approx \log \sqrt{n}\,\overline{Y} + \log \frac{1}{P(Y|\overline{Y})} \\ &= \tfrac{1}{2}\log n + \log \frac{1}{P(Y|\overline{Y})} + O_p(1) \end{aligned}$$

Implication for prefix length

To code *any* mean value requires $\frac{1}{2}\log n$ bits.

Model selection

Use a special one-bit code for zero. Code any non-zero parameter using $1 + \frac{1}{2}\log n$ bits:

Parameter	Prefix
0	0
$z \neq 0$	$1 \;\|\; \frac{1}{2}\log n$ bits for z

Penalized likelihood *BIC*

Reject $H_0 : \mu = 0$ and code a non-zero mean only if

$\log P(Y|\overline{Y}) - \log P(Y|\mu = 0) > \frac{1}{2}\log n$ or $|z| > \sqrt{\log n}$.

Spike and Slab Prior

Code = Probability

Recall that the choice of coding method implies a probability model. This applies to the parameter codes as well.

$\Rightarrow$ Very Bayesian point of view.

Implicit assumption

If we knew that $|\mu| < \frac{1}{2}$, then to grid this interval to precision $1/\sqrt{n}$ requires $\log\sqrt{n} = \frac{1}{2}\log n$ bits. The larger the range allowed for μ, the larger the number of bits.

Associated prior on μ

- If we do not code a mean, then we represent $\mu = 0$ with just 1 bit, implying a probability of 1/2.
- If we do code a mean, then we represent μ using $1 + 1/2 \log n$ bits, corresponding to a uniform distribution on $|\mu| < \frac{1}{2}$.

Natural prior?

Parameter is either *exactly zero*, or anywhere in allowed range. Asymptotics essentially force large z score for any $\mu \neq 0$.

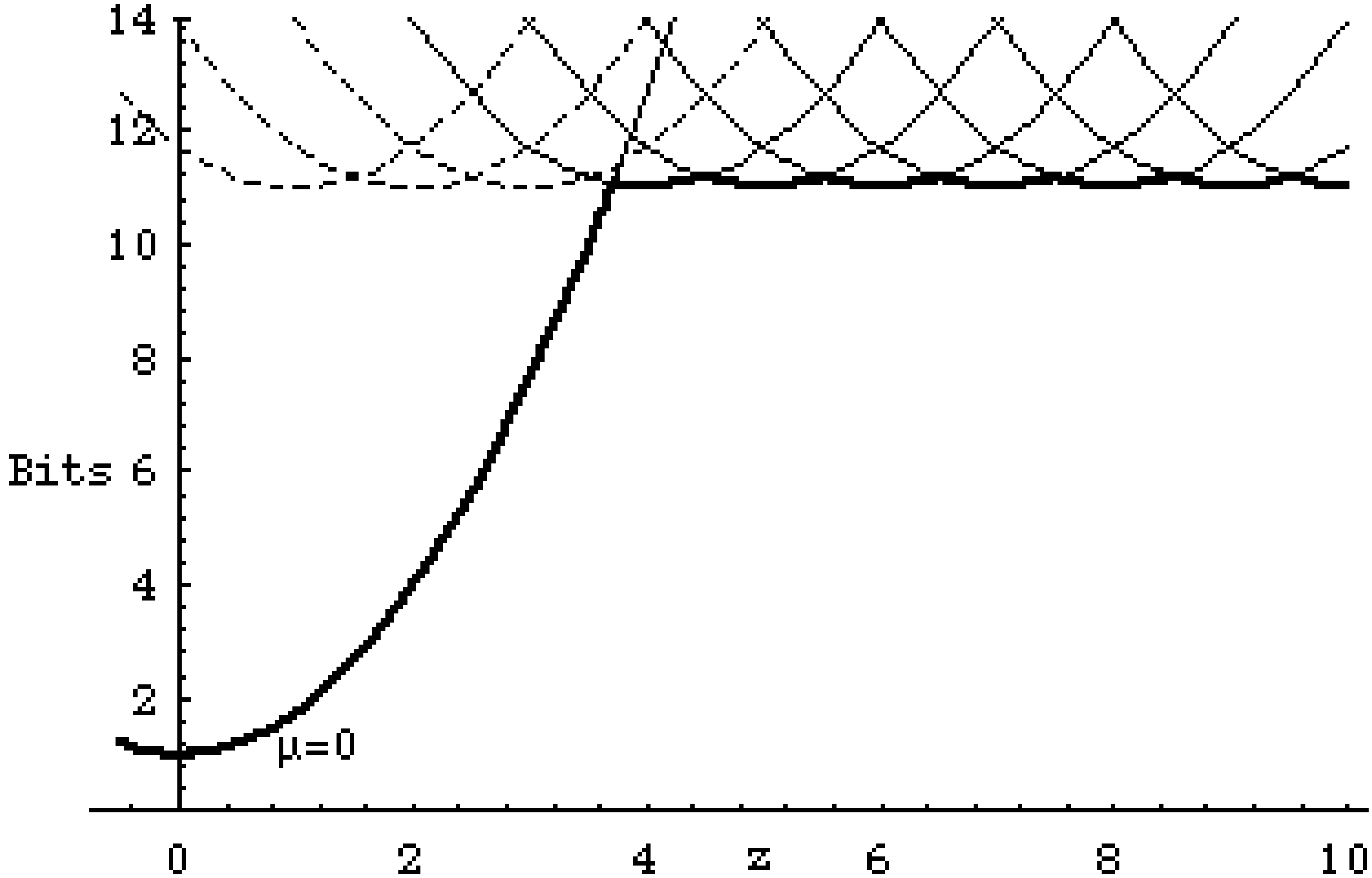
Impact of prior

Priors are much more important in model selection than elsewhere.

Graph of BIC Codebook

Vertical Axis: Bits are the excess $\frac{1}{2}\log n + R_n(\overline{Y} - \hat{\mu})$ over minimum determined by the log likelihood at $\overline{Y}$, with $n = 1024$ and $-16 < \mu \leq 16$

Horizontal Axis: $z = \sqrt{n}\,\overline{Y}$, the usual z-score.



Comparison of Coding Decisions

Attributes

	Local Asym Code	Traditional
as $n \to \infty$	$\mu \to 0,\ z$ fixed	$z \to \infty,\ \mu$ fixed
code $z \neq 0$ if	$\|z\| > 2.4$	$\|z\| > \sqrt{\log n}$
consistency	irrelevant	consistent
prior on z	log-Cauchy	spike-and-slab

Contradiction?

Traditional asymptotic analysis is not uniformly convergent, and must exclude a set of parameters of vanishing size — precisely those near the origin.

$$\Rightarrow \quad \lim_n \arg\min_z \text{CodeLength}(z) \neq \arg\min_z \lim_n \text{CodeLength}(z)$$

Model selection lives in the small set near 0.

Philosophical

Sample sizes are chosen to detect certain features.

Gather large samples to find features undetected in small samples.

$\Rightarrow$ Still have small z scores, even though n is large.

Review and Next Steps

So far

Information theory provides another view of modeling: good models produce short codes.

Parameter coding

Method of coding *rounded* parameter corresponds to a prior on the parameter space, with coding making the prior very explicit. Different codes/priors lead to different modeling criteria:

- Local asymptotics suggest fixed threshold near 2.4.
- Large z arguments lead to BIC with a threshold $\sqrt{\log n}$.

Regression

Same coding ideas, but now with multiple parameters.

Again, choose the model producing the shortest message (parameters + data).

Additional feature in regression

Codes for regression must also identify the chosen predictors as well as give the values of any parameter estimates.