Overview

Focus on regression problem

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

Reproduce criteria

Model selection via *AIC*, *BIC*, *RIC*, *eBIC* are equivalent to choosing the model which offers the greatest compression *according to a specific* two-part code. Selection criteria differ in how they encode the parameters.

Composition of prefix

Prefix must indicate two features:

1. Coefficient estimates
2. Which variables are being used

Similar to location problem

Again will encode parameters as integer z-scores, adding enough information to associate estimates with variables.

Regression Model

True Model

Rather than assume $EY = X\beta$, mean is unspecified:

$$Y = \eta + \epsilon, \quad E\,\epsilon = 0, \quad \mathrm{Var}\,\epsilon = \sigma^2 I_n,$$

Working Model

View $X\beta$ as the projection of η into the space defined by the available set of predictors,

$$Y = X\beta + \epsilon \quad \text{where} \quad X\beta = (X(X'X)^{-1}X)\,\eta\ ,$$

and treat $\epsilon \sim N(0, \sigma^2)$.

Subset/selection coefficients

Let $\gamma = (\gamma_1, \ldots, \gamma_p)$ denote a sequence of p 0's and 1s. Denote a subset of β by (miss APL notation!)

$$\beta_\gamma \quad \text{defined by} \quad \beta_j \neq 0 \iff \gamma_j = 1$$

Simplifying assumptions

- $p \leq n$ possible **orthogonal** regressors X_j, with $X_j' X_j = n$.
- σ^2 is known.
- Receiver knows n and X, so needn't send either.

Coding Regression

Model prefix

Prefix encodes γ and associated estimates:

1. Code for $\gamma = (1, 0, 0, \ldots, \gamma_j, \ldots, 1)$, the selection indicator
2. Code for fitted $\hat{\beta}_\gamma$ estimates
3. Compressed data

Goal and protection

The goal remains to construct the shortest message. Note the automatic penalty for over-fitting: the more variables used, the longer the prefix since more estimates must be added.

Estimates

$$b_j = \beta_j + \frac{X_j' \epsilon}{X_j' X_j} = \beta_j + \frac{\sigma}{\sqrt{n}} Z, \quad Z \sim N(0,1)$$

so that $SE(b_j) = \sigma/\sqrt{n}$

Rounding coefficients

Round coefficient estimates b_j to a standard error scale, as in the location problem,

$$\hat{\beta}_j = \frac{\sigma \langle z_j \rangle}{\sqrt{n}}\ , \qquad z_j = \frac{\sqrt{n}\, b_j}{\sigma}\ .$$

Variable Selection via Coding

Trade-off

Add additional variable X_j if

(Gain in data compression) $>$ (Increase in prefix length).

Gain in data compression

Log-likelihood based on q predictors is (ignoring constants)

$$\log \frac{1}{P(Y|b_{\gamma_q})} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i(q))^2}{\sigma^2 2 \ln 2} = \frac{RSS(q)}{\sigma^2 2 \ln 2} \, .$$

If add another predictor, say X_j, then

$$\triangle RSS = RSS(q) - RSS(q+1) = n\, b_j^2 \, ,$$

so the gain in data compression is

$$\frac{\triangle RSS}{\sigma^2 2 \ln 2} = \frac{n b_j^2}{\sigma^2 \, 2 \ln 2} = \frac{z_j^2}{2 \ln 2} \text{ fewer bits.}$$

Parameter cost differs

Least squares, *AIC*, *BIC*, *RIC*, and *eBIC* code the parameters

$$\gamma, \quad \hat{\beta}$$

differently, and so reach different compromises of data compression and model complexity.

Least Squares Coding

Fixed format code

Fixed format with reserved, fixed-length fields for each predictor:

1. p bits for the indicator γ,
2. $\frac{k}{2}\log n$ for each parameter $(2|\beta_j| < k\sigma/\sqrt{n})$, and
3. tag on the fully compressed data.

Parameters of code

$$\underbrace{\gamma_1,\ldots,\gamma_p}_{p \text{ bits}} \,\|\, \underbrace{\underbrace{\langle z_1\rangle}_{(k/2)\log n \text{ bits}} \quad \cdots \quad \underbrace{\langle z_p\rangle}_{(k/2)\log n \text{ bits}}}_{\text{all } p}$$

Data

Encode data using the associated rounded parameter estimates. This requires about

$$\log \frac{1}{P(Y_1,\ldots,Y_n|b_1,\ldots,b_p)} + \underbrace{nQ}_{\text{quantized}} \quad \text{bits}$$

Resulting selection

With fixed-length fields, regardless of selected variables, one obtains the shortest message by encoding *all* of the parameters.

BIC/SIC Code

Partly-fixed format code

Simple modification of the OLS code, with a fixed format for each *chosen* predictor rather than all predictors:

1. p bits for the indicator γ,
2. $\frac{k}{2}\log n$ for each chosen parameter, and
3. tag on the fully compressed data.

Parameters of code Assuming $|\gamma| = \sum_j \gamma_j = q$,

$$\underbrace{\gamma_1, \ldots, \gamma_p}_{p \text{ bits}} \| \underbrace{\underbrace{\langle z_{j_1} \rangle}_{k/2 \log n \text{ bits}} \quad \cdots \quad \underbrace{\langle z_{j_q} \rangle}_{k/2 \log n \text{ bits}}}_{q \text{ chosen}}$$

Resulting selection

Add X_j if

$$\underbrace{\frac{z_j^2}{2\ln 2}}_{\text{Increased compression}} > \underbrace{\frac{k}{2}\log n}_{\text{Increased parm bits}} ,$$

implying that one selects X_j if (with $k = 1$)

$$|z_j| > \sqrt{\log n}\,,$$

as when using *BIC*.

Interpreting the BIC/SIC Code

Code

Assuming $|\gamma| = \sum_j \gamma_j = q$,

$$\underbrace{\gamma_1, \ldots, \gamma_p}_{p \text{ bits}} | \quad \underbrace{\underbrace{\langle z_{j_1} \rangle}_{k/2 \log n \text{ bits}} \quad \cdots \quad \underbrace{\langle z_{j_q} \rangle}_{k/2 \log n \text{ bits}}}_{q \text{ chosen}}$$

Spike and slab for each slope

When X_j is

Excluded: 1 bit to denote zero (in the code for γ_j).

Included: $1 + \frac{1}{2} \log n$ bits for γ_i and z_j.

Prior on "complexity"

Since γ is always coded in p bits, as though *iid* coin tosses, this code assigns equal probability to all 2^p possible models.

$$\text{Prior prob}(q = 0) = \frac{1}{2^p}, \quad \text{Prior prob}(q = 1) = \frac{p}{2^p}$$

In general

$$\text{Prior prob}(q) = \frac{\binom{p}{q}}{2^p} \ ,$$

so that the most favored model (highest prior) is $q = p/2$.

$\Rightarrow$ We expect *half* of the variables to enter the model, albeit with a high threshold, $|z_j| > \sqrt{\log n}$.

AIC Regression Coding

Variable length code

Fixed p-bit prefix for γ with varying fields for each predictor:

- Prefix γ embedded in parameter codes,
- Concatenate universal codes $U_s(z_j),\ j = 1, \ldots, p$ for z_j.

Parameters of code

$$\underbrace{U_s(z_1),\ U_s(z_2),\ \ldots, U_s(z_p)}_{p \text{ univ codes}}$$

EG: $p = 6$ and simpler to read Cauchy codes,

$$\mathbf{0} \quad \mathbf{1}\ 0 + \quad \mathbf{0} \quad \mathbf{1}\ 1\ 1\ 0 \text{ -} \quad \mathbf{0} \quad \mathbf{0} \quad \Rightarrow 0\ 1\ 0\ \text{-}3\ 0\ 0$$

Leading bits of the universal codes *are* indicators γ_j.

Resulting selection

Add X_j if improved goodness of fit compensates for adding $U_s(z_j)$ bits for the additional parameter,

$$\underbrace{z_j^2/(2\ln 2)}_{\text{Increased compression}} > \underbrace{U_s(z_j) - 1 \approx 2\log\langle z_j\rangle}_{\text{Increased param bits}}$$

which implies that one codes once $|z_j| > 2.4$ (approximately).

Resembles AIC

Threshold fixed on z scale, as with AIC or C_p.

Interpreting the AIC Code

Associated prior on β_j

The associated prior on coordinates is a “rectangular” log-Cauchy distribution, and is *not* spherically symmetric.

Coefficients are not artificially constrained to some interval.

Natural prior?

Suggests many small coefficients, regardless of the sample size.

Prior on complexity

Though embedded into universal codes, γ still uses p bits for all models, as in the *BIC* code; expect half of the variables to enter.

Local asymptotics

Motivated by local asymptotics in which one fixes z as $n \to \infty$ rather than letting the z score grow to infinity.

Robustness of Universal Codes

What about the message lengths?

If in fact the z scores are large, won't the AIC model codes be a lot longer than the BIC model codes?

Oracle

Suppose that it is known that $-M\sigma \le \beta_j \le M\sigma$.

Data

Suppose further that the fitted coefficients for q variables attain this upper limit, $b_1 = \cdots = b_q = M\sigma$ so that $z_j = \sqrt{n}M$.

Large value exceeds AIC and BIC thresholds.

Prefix length for BIC code Since the grid of z-scores has

$$\frac{2M\sigma}{\sigma/\sqrt{n}} = 2\sqrt{n}\,M \quad \text{positions,}$$

uniform coding requires

$$q\ \log 2\sqrt{n}M = q(\log\sqrt{n}M + 1) \text{ bits.}$$

Prefix length for AIC code

$$q\ \log^* \sqrt{n}M \approx q\ (\log\sqrt{n}M + 2\log\log\sqrt{n}M)$$

which shares the dominant term with the BIC code.

MDL in Regression

Definition of minimum description length Rissanen (1983)

In its original asymptotic form, the *MDL* for a model with q orthogonal predictors is

$$\begin{aligned} MDL(q) &= \log^*(V(k)\|\hat\theta\|^q) + \log\frac{1}{P(Y|\hat\theta)} \\ &\approx \frac{q}{2}\log n + \frac{q}{2}\log\sum\hat\theta_i^2 + \log V(q) + \log\frac{1}{P(Y|\hat\theta)} \\ &\approx \frac{q}{2}\log n + \log\frac{1}{P(Y|\hat\theta)} \\ &= BIC(q) \end{aligned}$$

where $V(k) = \text{Vol}(k\text{-dim ball, radius one})$ and $\|\theta\|^2 = n\sum\theta_j^2$.

Implicit selection indicator

Since γ does not appear in this definition, its as though it is coded with a fixed number of bits for all models.

Local coding interpretation

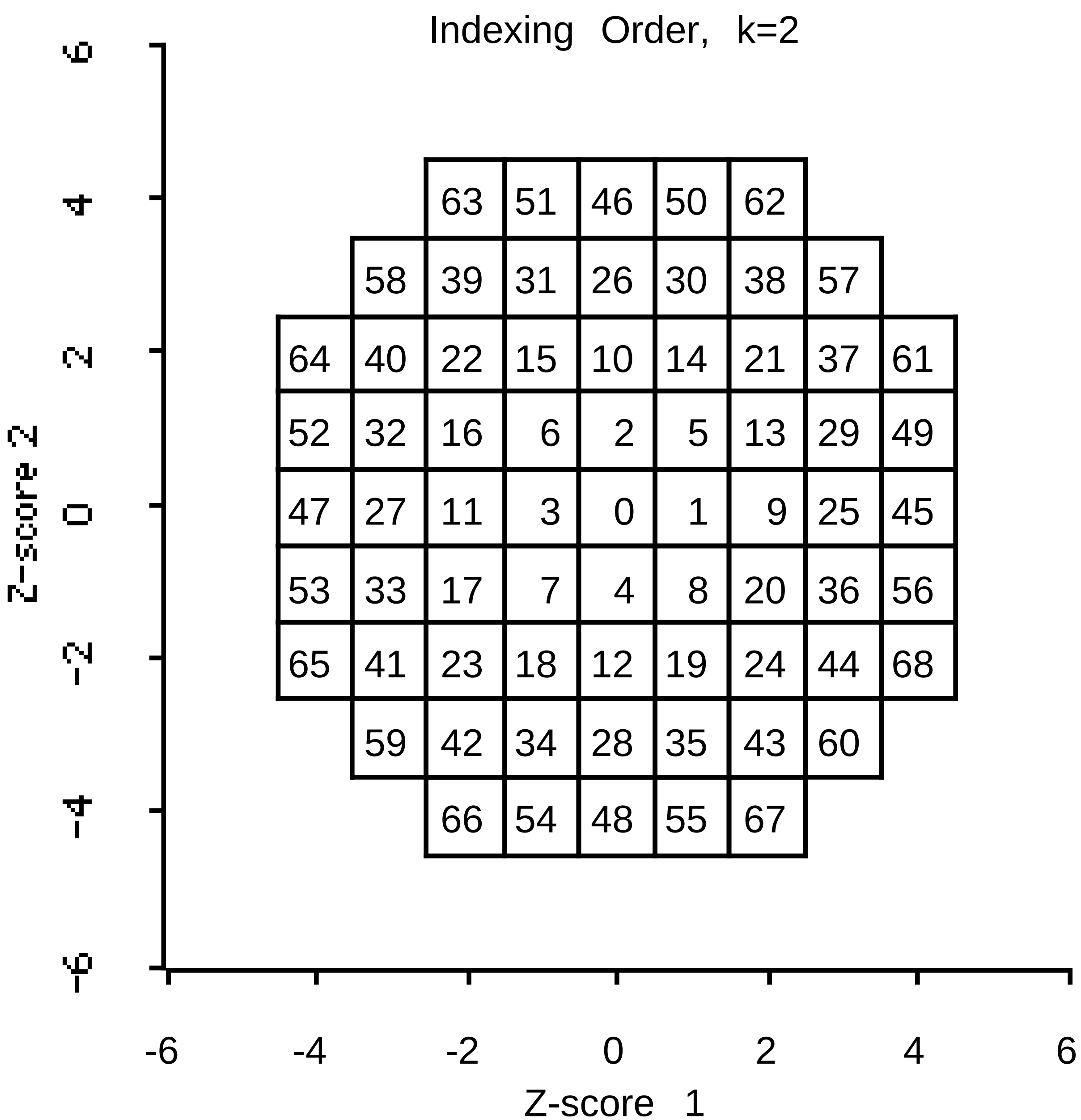
The prefix encodes an index for *vector* $\mathbf{z} = (z_1, \ldots, z_q)$ using a spiraling code:

$$\begin{aligned} MDL(q) &\approx \log^*(V(q)\|\hat\theta\|^q) + \log 1/P(Y|\hat\theta) \\ &= U(i(\mathbf{z})) + \log 1/P(Y|\hat\theta) \end{aligned}$$

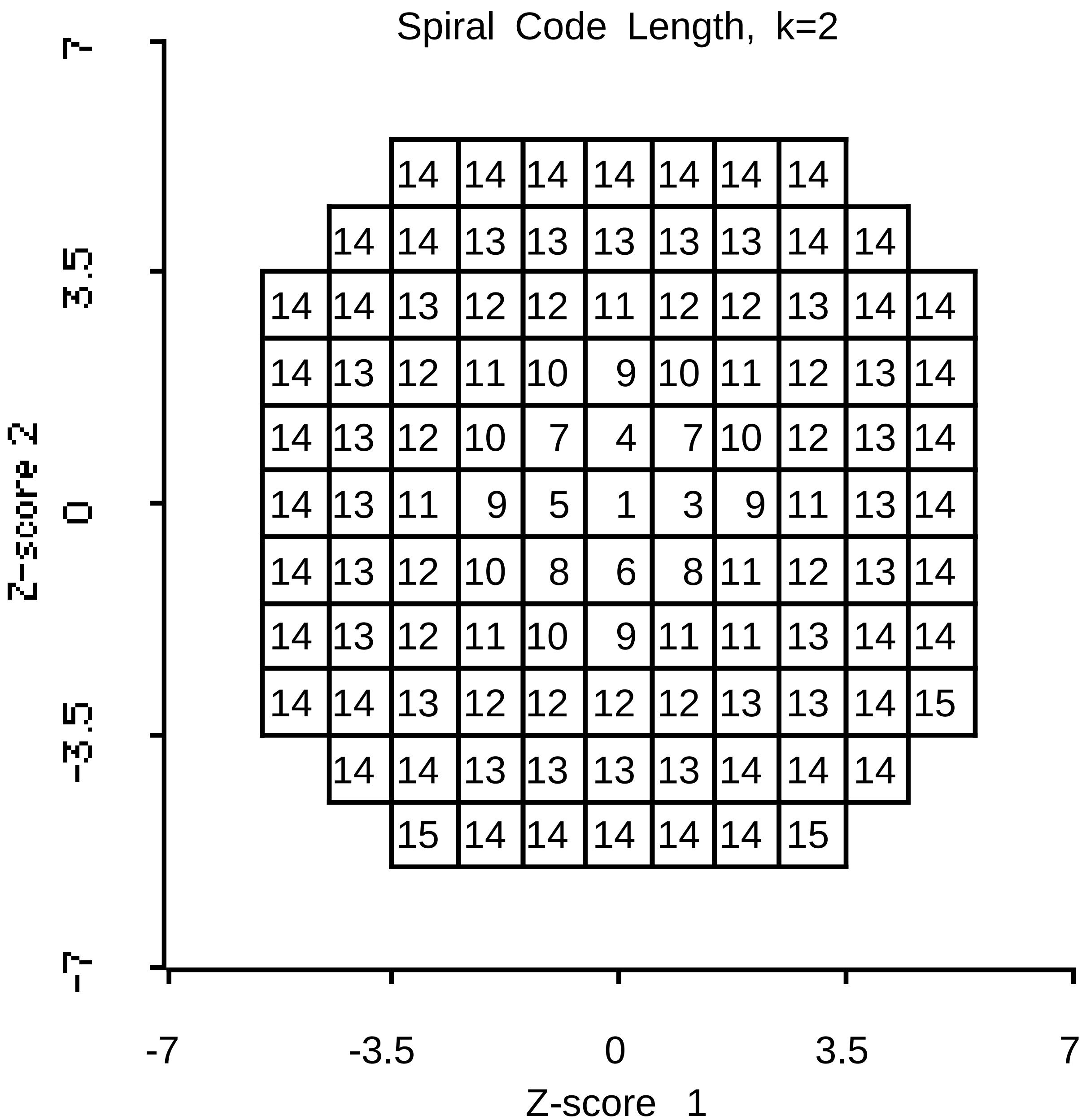
where $i(\mathbf{z})$ denotes index identifying the vector z score.

Spiral MDL Index Path

- "Spiral" indexing using universal code on SE scale.
- Plots of index (below), bits for index (next).



Code Lengths for Spiral MDL Index



Implications of Spherical Prior/Code

Projection to subspace

Following plot: code lengths with subspaces show that threshold for adding another variable *increases* with $\|\mathbf{z}\|$.

Example: Use one or two variables?

$q = 2$ gives shorter code than $q = 1$ when

$$\underbrace{\frac{z_2^2}{2 \ln 2}}_{\text{Gain in compression}} > \underbrace{U(i(z_1, z_2)) - U(i(z_1))}_{\text{Increase parm bits}}$$

For large $z_1 >> z_2 > 0$,

$$\begin{aligned} U(i(z_1, z_2)) - U(i(z_1)) &\approx \log^* \|z_1, z_2\|^2 - \log^* \|z_1\| \\ &\approx 2 \log z_1 - \log z_1 \\ &= \log z_1 \end{aligned}$$

Add X_2 to model with just X_1 when

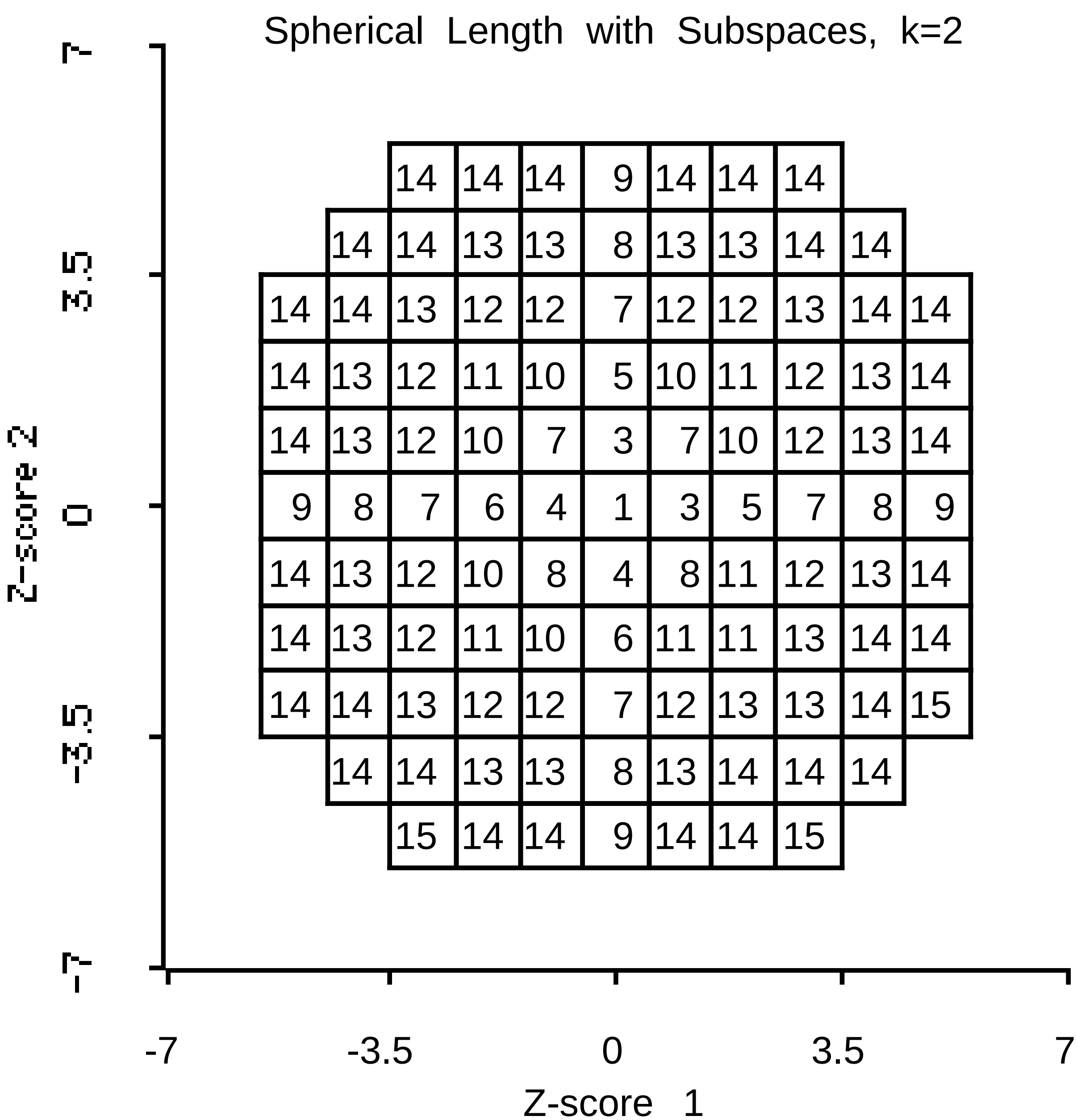
$$\frac{z_2^2}{2 \ln 2} > \log z_1$$

Maximum coefficient determines threshold

q large coefficients, $\mathbf{z} = (z, z, \ldots, z)$, and one smaller coefficient, $z >> \tilde{z} > 0$. Add $\tilde{z}$ if

$$\frac{\tilde{z}^2}{2 \ln 2} > \log z + \frac{1}{2} \log q$$

Code Lengths with Projections



Examples

Add $\tilde{z}$?

Add $\tilde{z}$ to model with q coefficients when

$$\underbrace{\frac{\tilde{z}^2}{2\ln 2}}_{\text{gain}} > \underbrace{\log z + \tfrac{1}{2}\log q}_{\text{penalty}}$$

Local coding = BIC in special case

If $\max \mathbf{z} = \sqrt{n}$, then penalty $= \frac{1}{2}(\log n + \log q)$ and threshold is about $\sqrt{\log n}$.

Explicit examples "Elephants and mice"

Model A Two small coefficients

$$\mathbf{z} = (3,4) \quad \Rightarrow \text{ pick } \Rightarrow \quad (3,4)$$

Model B Two small, plus one large

$$\mathbf{z} = (3,4,10) \quad \Rightarrow \text{ pick } \Rightarrow \quad (3,4,10)$$

$$\mathbf{z} = (3,4,100) \quad \Rightarrow \text{ pick } \Rightarrow \quad (4,100)$$

$$\mathbf{z} = (3,4,1000) \quad \Rightarrow \text{ pick } \Rightarrow \quad (1000)$$

Model C Two small, plus many large

$$\mathbf{z} = (3,4,100,\ldots,100) \quad \Rightarrow \text{ pick } \Rightarrow \quad (4,100,\ldots,100)$$

Indexed Parameter Coding: *RIC*

Complexity prior

Suppose expect very few coefficients to enter model, $|\gamma| \approx 1$.

Bernoulli compression

Compress $Y_1, \ldots, Y_n \sim B(1/n)$ by giving indices for i s.t. $Y_i = 1$,

$$n\, H(Y_1) \approx \log n\ .$$

Encode γ as a sequence of indices rather than 0/1 indicators.

Parameters of code

$$q \mid \underbrace{(j_1,\, U_s(z_{j_1}))}_{\log p + \ell(U_s(z_{j_1}))} \mid \cdots \mid (j_q,\, U_s(z_{j_q}))$$

Resulting selection

Add X_j if (approximately)

$$\underbrace{z_j^2/(2\ln 2)}_{\text{Increased compression}} \quad > \quad \underbrace{\log p + 2\log \langle z_j \rangle}_{\text{Increased parm bits}}$$

or roughly once z_j exceeds the *Bonferroni* bound,

$$|z_j| > \sqrt{2\,\log p} \approx \Phi^{-1}(1 - 1/p)$$

Adaptive Coding: eBIC

Coding methods

Method	Code for γ	Expected Predictors
BIC	p bit prefix	$p/2$
AIC	Embedded p bit prefix	$p/2$
RIC	Indexing	1

Why *a priori* assume the complexity — code adaptively.

Compress γ

Compress $\gamma_1, \ldots, \gamma_p$, treating as a Bernoulli sequence. In effect, modify the *AIC* code by compressing the leading bits of the sequence of universal codes.

Such a code will produce slightly longer messages than

- *RIC* code if indeed $q = 1$ is best
- *AIC* code if $q = p/2$ is best

but the added length in these cases is very small.

Two-part code

$$\underbrace{\gamma_1, \ldots, \gamma_p}_{p\ H(q/p)} \mid \underbrace{U_s(z_{j_1}), \ldots, U_s(z_{j_q})}_{\sum_{k=1}^{q} \ell(U(z_{j_k} - q))}$$

Adaptive Coding: eBIC

Adaptive selection criterion

Add X_j if (approximately)

$$\frac{z_j^2}{2\ln 2} > p\left(H\left(\frac{q+1}{p}\right) - H\left(\frac{q}{p}\right)\right) + 2\log z_j$$

or once $|z_j|$ exceeds the adaptive thresholding bound,

$$\frac{z_j^2}{2\ln 2} > \log\frac{p-q}{q+1} + 2\ \log z_j \quad \Rightarrow \quad |z_j| > \sqrt{2\log p/q}$$

Comparable to

- Simes, Step-up/Step-down Testing: Compare $\max z_j$ to Bonferroni, second largest to next normal order stat, etc.
- Empirical Bayes prior for the number of parameters.

Further variations

Can compresses other bits in the z scores as well.

Big question

Is coding useful when used in this more extensive manner?

- Fine for offering another way to think about model selection criteria.
- But, is it appropriate to use bit lengths to judge which is a better selection criterion?

Discussion

Model selection message formats

Criterion	Threshold	Parameter	Complexity
BIC, *SIC*	$\sqrt{\log n}$	Spike-slab	half
AIC, C_p	$\sqrt{2}$	Log-Cauchy	half
RIC, hard	$\sqrt{2\log p}$	Log-Cauchy	1
eBIC	$\sqrt{2\log p/q}$	Log-Cauchy	adaptive

Discussion

- *Interplay of information theory and Bayesian ideas*
  Another way to think about priors, particularly in harder problems (priors for continuous functions).
- *Robustness of the priors*
  Universal priors are uncommon, but quite powerful.
- *Spherical priors*
  Common, but appropriate in variable selection?
- *Is coding a realistic criterion?*
  Seems fine as a way to characterize methods, but can it suggest which are really better?

Research Continues

Other types of models

Application to smoothing splines, piecewise models, or regression trees?

Varying parameter effects on compression, and how to code?

Dependent processes

Context tree models are capable of capturing dependence in a nonparametric way and have been used to develop, for example, model-free bootstrap resampling methods.

Reducing assumptions: Collinearity, variance, normality

Collinearity Drop the assumption of orthogonal parameters.

Variance Drop the assumption of known variance.

Robustness CLT for coefficients, but what about compression of the data via likelihood? Using wrong distribution makes a big difference in *data compression*, implying $z^2/(2 \ln 2)$ may not be the right trade-off.

File compression vs model selection

Coding seeks big gains in compression, on the order of 10%.

Testing/parameter selection deals with several bits.

Can a method used to obtain 10% gains be applied to evaluate changes of several bits?