Codes and Models

Bin Yu
Bell Labs, Lucent and UC Berkeley

Thanks to:
Mark Hansen, Youngjun Yoo, Antonio Ortega,
Andrew Barron, Jorma Rissanen

IMS-ENAR Meeting, Atlanta
March, 1999

Outline:

- Introduction

  Prelude

  A “Bit” of Information Theory

- From Coding to Modeling:
  Minimum Description Length (MDL) Principle

  Foundation of MDL: Coding Theorems

  $gMDL$ in Regression: Bridging AIC and BIC

  $lMDL$ in Wavelet Image Denoising:

  Simultaneous Denoising and Compression

- From Modeling to Coding: Wavelet Image Coder

  Statistical Models in an Image Coder

- Concluding Remarks

Introduction

Prelude

Computer technologies make easy the collection of data, driving the need for effective ways to

- transmit and store data, and
- analyze data.

The former is the subject of

- Information theory

  $\rightarrow$ data compression/coding.

  Claude Shannon:
  A Mathematical Theory of Communication (1948)

The latter is the subject of

- Statistics

  $\rightarrow$ estimation/inference;

  ... Fisher, Neyman, Tukey, ...

The two fields shared a long history of interactions.

A personal and biased list:

- Kullback (’51) Mutual information and sufficiency
- Jaynes (’57) Maximum entropy method
- Kullback and Leibler (’59) KL divergence
- Kolmogorov (’65) K-sufficiency in algorithmic complexity theory
- Wallace and Boulton (’68) Minimum Message Length (MML)
- Rissanen (’78) Minimum Description Length (MDL)
- Csiszär and Tusnädy (’84) Alternating minimization – EM algorithm
- Berrou, Glavieux and Thitimajshima’s (’93) Turbo decoding
- Barron, Birgé and Massart (’98) Nonparametric estimation by complexity regularization (’98)

A “Bit” of Information Theory

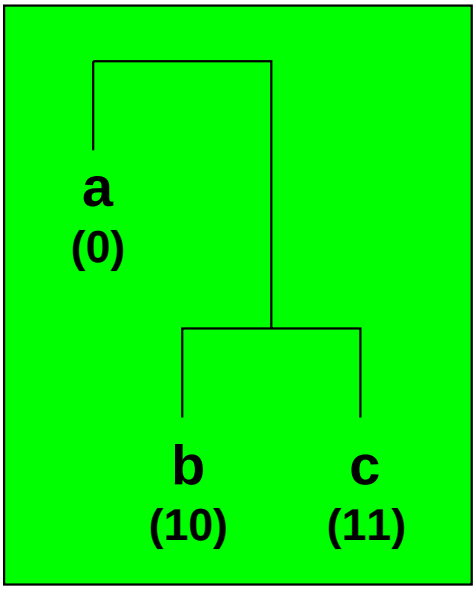
Claude Shannon
(A Mathematical Theory of Communication, 1948):

> *The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point.*

What is a code?

Given a finite alphabet (or set) $\mathcal{A}$, a binary code $\mathcal{C}$ is a map from $\mathcal{A}$ to strings of 0’s and 1’s.

Example: $\mathcal{A} = \{a, b, c\}$



$$\begin{aligned} \mathcal{C} : \mathcal{A} &\rightarrow \{0,1\}^* \\ a &\rightarrow 0 \\ b &\rightarrow 10 \\ c &\rightarrow 11 \end{aligned}$$

L is the code length function of $\mathcal{C}$ in
bits – for *b*inary dig*its* (Tukey):

$$L(0) = 1,\ L(10) = 2,\ L(11) = 2$$

$\mathcal{C}$ is "prefix" requiring no separating symbols:

0001110 must have come from *aaacb*.

Moreover, $L(\cdot) = \Leftrightarrow \log P(\cdot)$,

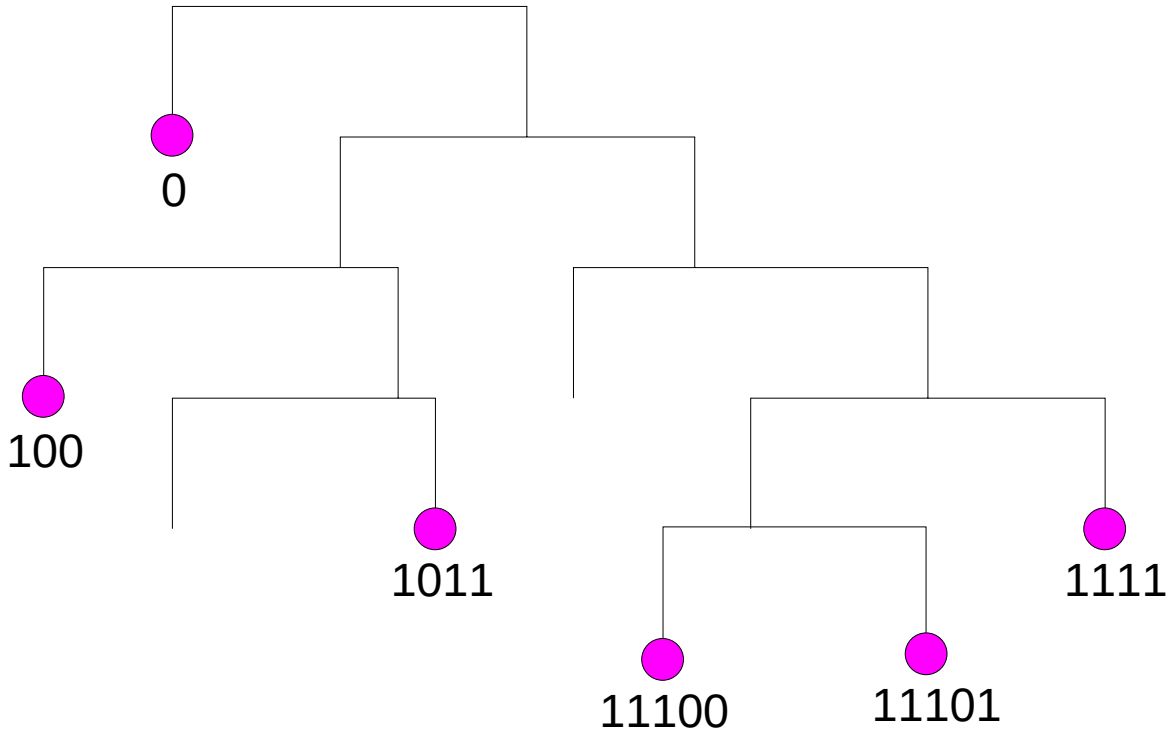
for $P(a) = 1/2,\ P(b) = P(c) = 1/4.$

Kraft's inequality (1949):

$L(\cdot)$ is the code length function of a (binary) prefix code on $\mathcal{A}$ iff

$$\sum_x 2^{-L(x)} \leq 1.$$

Key: Map a code to a binary tree, then "prefix" iff the codewords are all end-nodes.



$$2^{-1} + 2^{-3} + 2^{-4} + 2^{-5} + 2^{-5} + 2^{-4} < 1$$

It follows that for any probability distribution $Q(\cdot)$,

$-\log Q(x)$ may be regarded as the code length of x of a prefix code.

Claude Shannon
(A Mathematical Theory of Communication, 1948):

> *The significant aspect is that the actual message is one* **selected** *from* **a set** *of possible messages.*

How good is a code?

Assume the data string is generated from a source distribution P, then we can define

Entropy of P:

$$H(P) = \sum P(x)[\Leftrightarrow \log P(x)].$$

Redundancy: Given a true (source) distribution P, the expected redundancy of a code L is defined as

$$R(L, P) = E_P L(X) \Leftrightarrow H(P).$$

Re-write as

$$\begin{aligned} R(L, P) &= E_P \log[P(X)/Q(X)] \\ &\equiv KL(P, Q) \geq 0, \end{aligned}$$

where $Q(X) = 2^{-L(X)}$, viewed as a probability distribution by Kraft's inequality.

Connection to Fisher's Information:

$$\lim_{t \to 0} \frac{KL(P_\theta, P_{\theta+t})}{t^2} = \frac{I(\theta)}{\ln 4}.$$

For iid data strings from P,

$$H(P^n) = nH(P).$$

Shannon's Coding Theorem (for iid data strings) :
For any code $L_n(x^n)$, the per symbol redundancy

$$\frac{R(L_n, P^n)}{n} = \frac{E_P L_n(X^n)}{n} \Leftrightarrow H(P) \geq 0;$$

and for code $L_n^*(x^n) = [\Leftrightarrow \log P^n(x^n)]_+$,

$$\frac{R(L_n, P^n)}{n} = \frac{E_P L_n(X^n)}{n} \Leftrightarrow H(P) \leq 1/n.$$

$H(P)$ **is the compression limit.**

Universal Coding Theorem (for iid data strings):

Without knowing P, there is a code L_n that achieves the entropy rate asymptotically,

$$\frac{L_n(X^n)}{n} \to H(P) \quad \text{in probability.}$$

References on Information Theory:

Shannon (1948),
A Mathematical Theory of Communication

Book: Cover and Thomas (1990),
Elements of Information Theory

Review: Verdú (1998),
Fifty Years of Shannon Theory

From Coding to Modeling: MDL

Rissanen's ('78) Minimum Description Length (MDL) Principle:

Choose the model that gives the shortest description of data.

References on MDL:

Book: Rissanen, 1989,
Review: Barron, Rissanen and Yu, 1998, IEEE-IT
Review: Hansen and Yu, 1998

Precursors to MDL:

Algorithmic complexity
(Kolmogorov, Solomonoff, Chaitin, 60's);

Shortest description length for classification
(Wallace and Boulton, '68)

Related earlier statistical works:

C_p model selection criterion (Mallows, '73)

AIC model selection criterion (Akaike, '74)

BIC model selection criterion (Schwarz, '78)

Recall for any probability distribution $P(\cdot)$,

$-\log P(x^n)$ may be regarded as the code length of x^n of a prefix code via Kraft's inequality.

Hence for any density function $f(\cdot)$ and precision δ,

$$\Leftrightarrow \log[f(x^n)\delta^n] = \Leftrightarrow \log f(x^n) \Leftrightarrow n \log \delta$$

is an (approximate) code length. Therefore,

$\Leftrightarrow \log f(x^n)$ may be regarded as the (idealized) code length of x^n.

What description form to use for MDL?

For one parametric family

$$\mathcal{M}_k = \{f_\beta : \beta \in \Theta_k \subset R^k\}$$

Foundation for MDL: Shannon's Coding Theorem

It implies that $\Leftrightarrow \log P(x^n)$ is the code length to use in MDL IF P is known.

$$L(x^n) = \Leftrightarrow \log P_\beta(x^n) + L(\beta).$$

Hence MDL is the same as Maximum Likelihood, IF $L(\beta)$ is independent of β,

$$\min_\beta\{\Leftrightarrow \log P_\beta(x^n)\} \quad \longleftrightarrow \quad \max_\beta\{P_\beta(x^n)\}$$

Over a collection of parametric families, this is the **model selection problem.**

$$\mathcal{M}_k = \{f_\beta : \beta \in \Theta_k \subset R^k\}, \quad \mathcal{M}_k \in \mathcal{C}$$

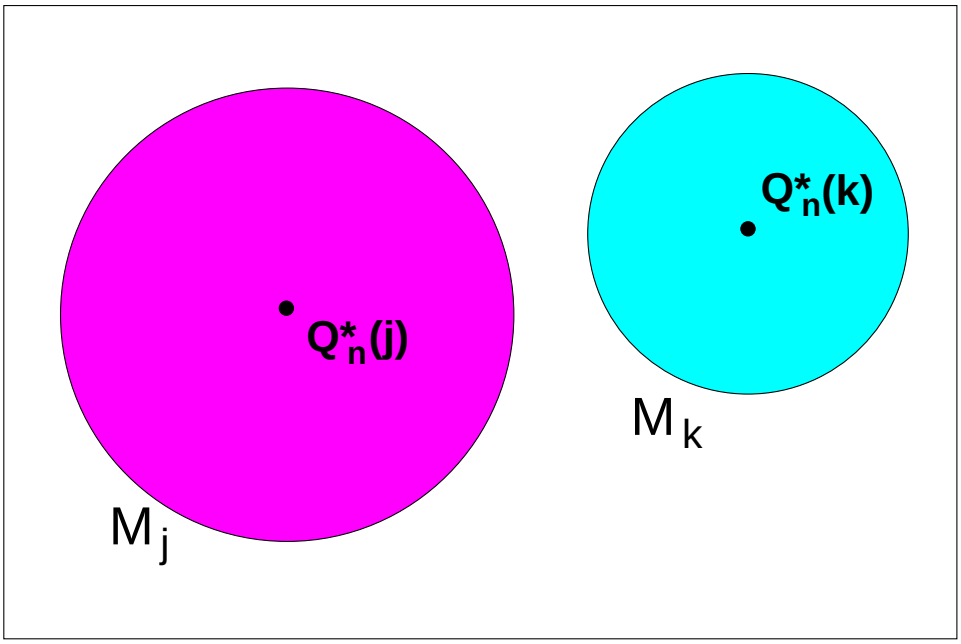
Select one $\mathcal{M}_k$ from $\mathcal{C}$ based on n samples.

Foundation for MDL: A universal coding theorem justifies the description form based on a model class.

Universal Coding Theorem over $\mathcal{M}_k$ (Rissanen, 1986):

- $R(L_n, P^n)/n \geq \frac{k}{2}(\log n/n + o(1))$; and
- a universal code L_n^* which achieves the lower bound.

A universal code L_n^* based on $\mathcal{M}_k$ should be used for model selection in MDL.



Here, $Q_n^* = 2^{-L_n^*}$

Examples of Universal Codes or DL's

(i)Two-stage Description Length

$$L(\text{data}|\widehat{\beta}) + L(\widehat{\beta}) + L(\mathcal{M}_k)$$

$$\Leftrightarrow \text{maximum likelihood} + \text{penalty},$$

IF $L(\mathcal{M}_k)$ is chosen independent of k,

where penalty = code length for the parameter estimate ($\approx k/2 \times \log n$),

(ii) Mixture Form of Description Length

$$\Leftrightarrow \log m(x^n) = \Leftrightarrow \log \int_\beta f_\beta(x^n) w(\beta) d\beta + L(\mathcal{M}_k).$$

This form connects to Bayesian model selection.

(iii) Predictive

$$L(x^n) + L(\mathcal{M}_k)$$
$$= \sum_{t}^{n} \Leftrightarrow \log f(x_t|\widehat{\beta}_{t-1}) + L(\mathcal{M}_k),$$

where $\widehat{\beta}_{t-1}$ is a good estimator, say MLE, based on the first $(t \Leftrightarrow 1)$ observations $x_1, ..., x_{t-1}$.

It connects to prequential statistics of P. Dawid and to learning theory/machine learning.

(iv) Normalized Maximum Likelihood (NML)

Example: What is the NML description length for a 0-1 sequence of length n based on the iid Bernoulli model?

k= number of 1's.

Use $\log n$ to code k.

a. Two-stage: $-k \log \frac{k}{n} - (n-k) \log(1-\frac{k}{n}) + \log n$

b. NML: $\log \binom{n}{k} + \log n$

It first appeared in the coding literature and was brought into MDL by Rissanen (1996).

MDL in Normal Linear Regression: Bridging AIC and BIC

Ref: Hansen and Yu (1999a)

Model:

$$y = \sum_{\gamma_m=1} \beta_m x_m + \epsilon,$$

where

- $\epsilon \equiv N(0, \sigma^2)$,
- $\gamma = (\gamma_1, \ldots, \gamma_M) \in \{0,1\}^M$ index for the 2^M possible models.

Recall

$$BIC(\gamma) = \frac{n}{2} \times \log RSS(\gamma) + \frac{k_\gamma}{2} \times \log n.$$

$$AIC(\gamma) = \frac{n}{2} \times \log RSS(\gamma) + \frac{k_\gamma}{2} \times 2.$$

- If the model is finite dimensional (parametric), BIC is consistent and prediction optimal;

- If the model is infinite dimensional (nonparametric), AIC is prediction-optimal.

One Mixture Form of MDL: $gMDL$

- For any given γ, take an inverted gamma prior on $\tau = \sigma^2$

$$p(\tau) = \sqrt{\frac{a}{2\pi}}\tau^{-3/2}\exp\left(\frac{\Leftrightarrow a}{2\tau}\right),$$

- given τ, β has a multivariate Gaussian prior

$$p(\beta|\tau) \sim N(0, c\tau\Sigma).$$

- Use Zellner's g-prior (1986) $\Sigma = (X^t X)^{-1}$
- Minimize over a_γ and c_γ according to MDL for each γ

Then we get

$$L(y^n|\gamma) = \begin{cases} \frac{n}{2}\log\frac{RSS(\gamma)}{(n-k_\gamma)} + \frac{k_\gamma}{2}\log F_\gamma, & R^2_\gamma \geq k/n, \\ \frac{n}{2}\log(y^t y/n) & \text{otherwise,} \end{cases}$$

where $F_\gamma = \frac{(y^t y - RSS(\gamma))}{k_\gamma S_\gamma}$.

Hence

$$gMDL(\gamma) = L(y^n|\gamma) + L(\widehat{a}_\gamma) + L(\widehat{c}_\gamma) + L(\gamma).$$

- $L(\widehat{a}_\gamma) + L(\widehat{c}_\gamma)$ could be made indep. of γ if we use a fixed precision say $1/\sqrt{n}$ for all models; and

- $L(\gamma)$ could be that of a two-stage code on $\{0,1\}^M$:

$$L(\gamma) = \log \left(\begin{array}{c} M \\ k_\gamma \end{array} \right) + \log n.$$

Example 1: Number of Bristles on the Fruit Fly

Original data collected by Long et al (1995) to identify genetic loci that influence the number of bristles on the fruit fly.

As a linear regression variable selection problem: (e.g. Broman, 1997).

y: number of bristles

x: gender indicator, on-and-off indicators at 19 genetic markers, and their interaction terms with gender (39 variables total).

Broman (1997) used

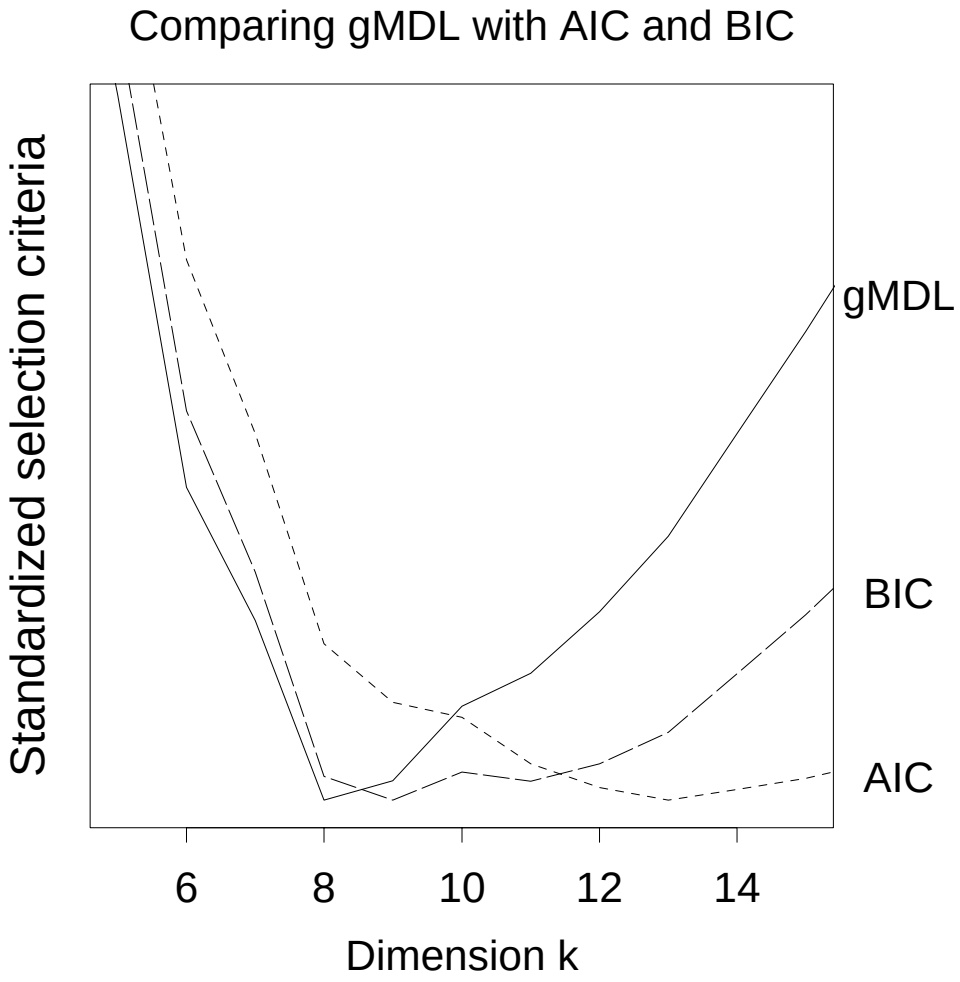
$$BIC_\eta = \frac{n}{2} \times \log RSS + \eta \times \frac{k}{2} \times \log n,$$

with $\eta = 2, 2.5, 3$ to adapt to this problem.

All three give rise to an 8-term model:

- constant term
- main effect for gender
- five marker main effects (markers 2, 5, 9, 13 and 17), and
- one gender $\times$ marker interaction (at marker 5).

	Estimate	StdErr
intercept	12.9	0.2
sex	-1.4	0.1
M13.5	1.2	0.2
M35	1.1	0.3
M46	1.8	0.3
M69.5	1.1	0.2
M90	1.7	0.2
sex $\times$ M35	0.9	0.1


Comparing gMDL with AIC and BIC
Standardized selection criteria
gMDL
BIC
AIC
6
8
10
12
14
Dimension k

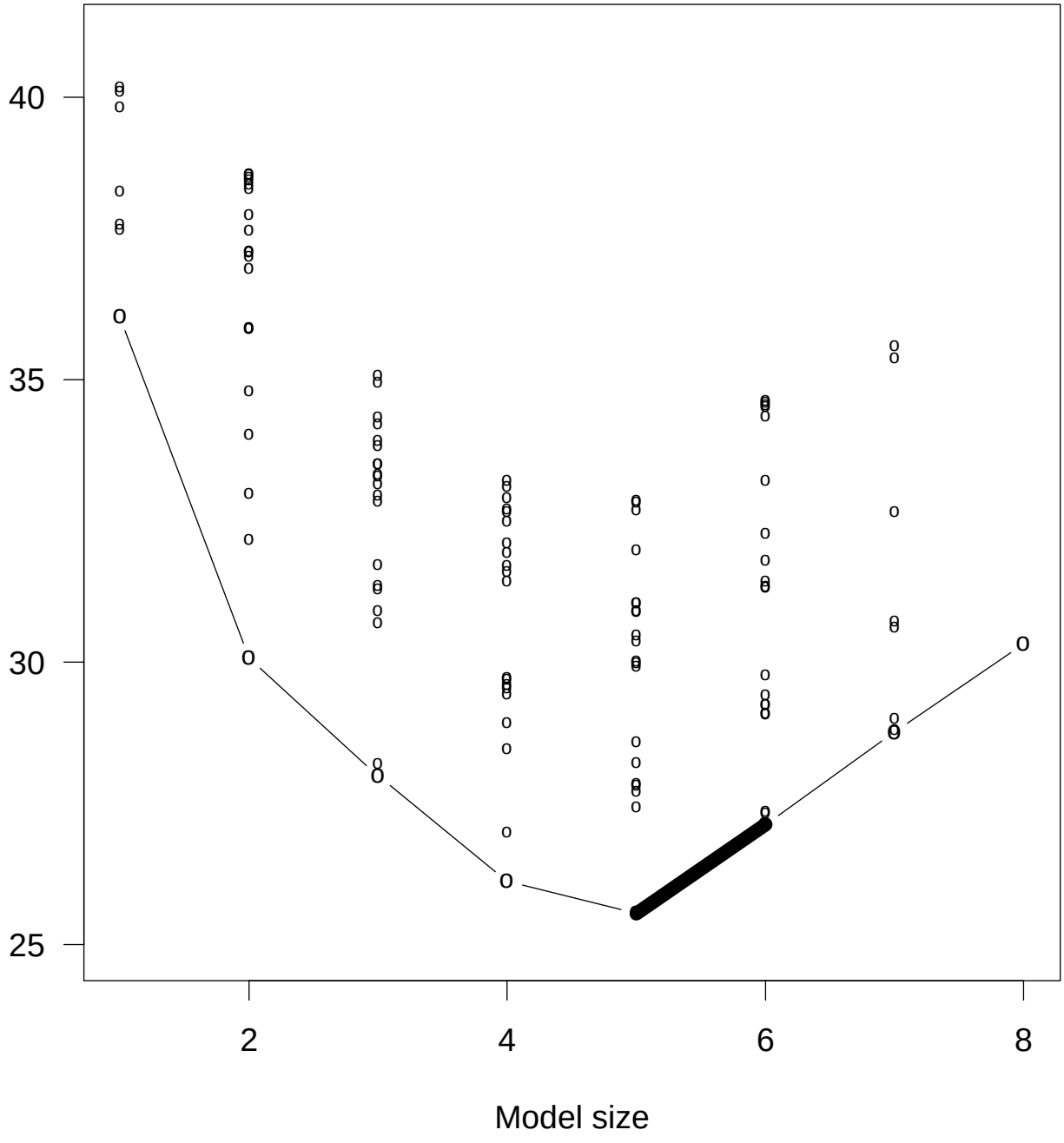
Example 2: A Simulation Study

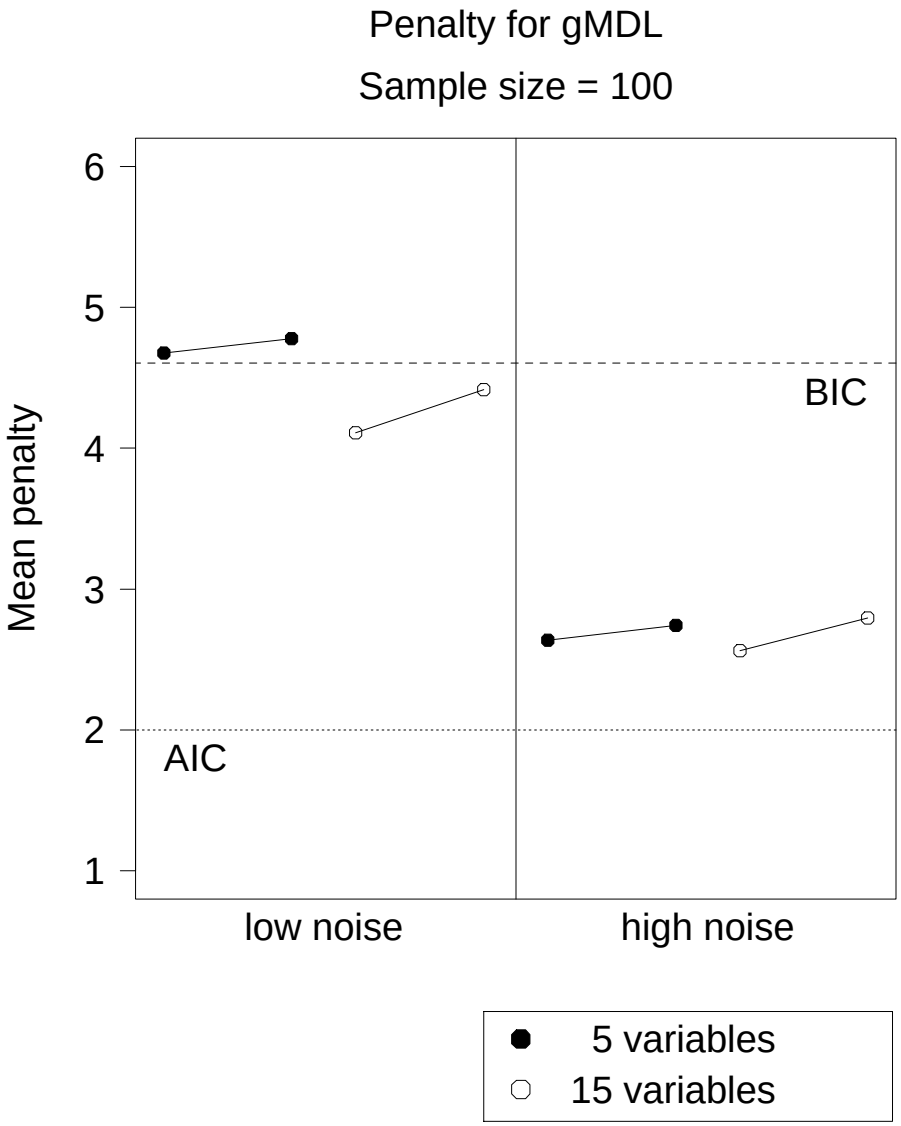
Set-up: $y = X\beta + \epsilon$

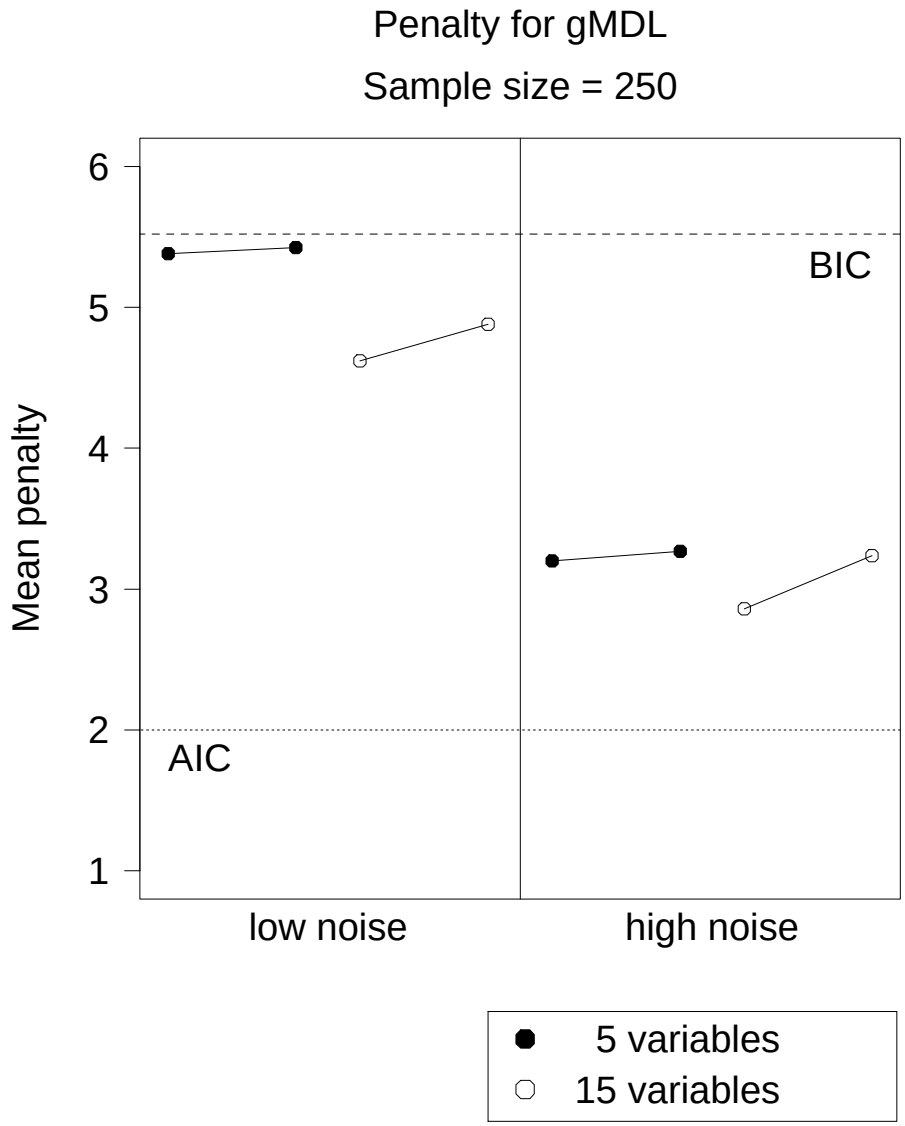
Experiment design:

- $X \sim N(0, \Sigma_{20\times 20})$, $\Sigma_{i,j} = \rho^{|i-j|}$,
- β built from k IID exp(1) and rest zero,
- $\epsilon \sim N(0, \sigma^2 I_{n\times n})$.

$$\rho(\{0.25, 0.77\})\times$$
$$k(\{5, 15\})\times$$
$$\sigma^2(\{1/3, 3\})\times$$
$$n(\{100, 250\}).$$


40
35
30
25
2
4
6
8
Model size





Equivalent Penalty = $2\{G(\mathcal{M}^*) - \frac{n}{2} \log RSS(\mathcal{M}^*)\}$

where criterion G=AIC, or BIC, or $gMDL$, and $\mathcal{M}^*$ is the optimal model according to G.

What is going on?

It can be shown (Hansen and Yu, 1999a) under the model in Breiman and Freedman (1986) that the $gMDL$ penalty is approximately

$$\log[nC_k]$$

where

$$C_k = \text{average SNR for Model} \mathcal{M}_k.$$

Adjusting the penalty with this factor, $gMDL$ adapts to act like AIC or BIC depending on the underlying bias and variance trade-off and hence exhibits "bridging" behavior.

Wavelet Image Denoising

Wavelet Transform: Compacts energy better – sparse wavelet coefficients.

Book References on Wavelets:

Daubechies (1992)

Vetterli and Kovačević (1995)

Strang and Nguyen (1996)

Mallat (1998)

2-D wavelet transform: product of 2 1-D transforms with one horizontal and one vertical.

A 3-level wavelet transform is used.

H - High pass filter or Detail, e.g. differencing in Haar;

L - Low pass filter or Smooth, e.g. averaging in Haar.

Level 1: Apply the 2-D transform to the original image, we get four subbands HH_1, HL_1, LH_1 and LL_1;

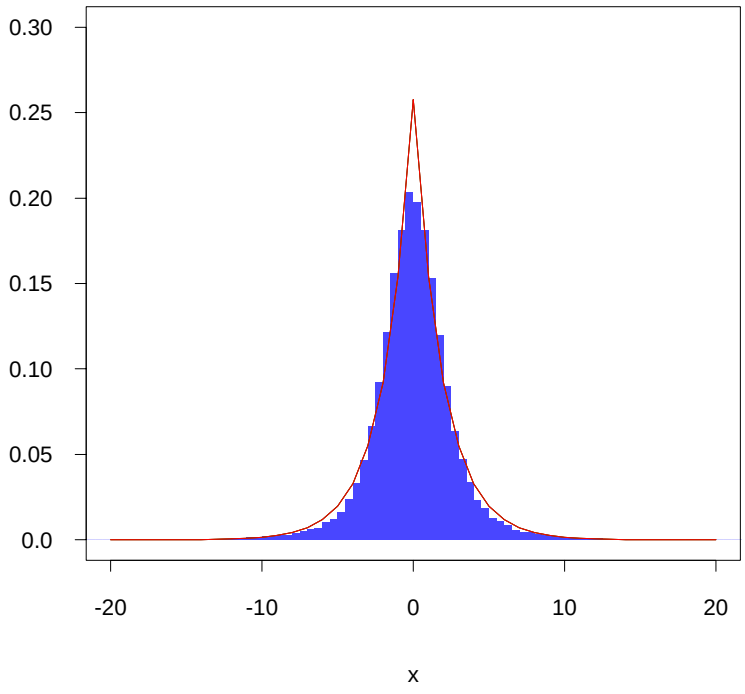
Level 2: Take LL_1 ...

Level 3: Take LL_2 ...

Ordinary 1-level, 2-d Wavelet Transform

Ordinary 3-level, 2-d Wavelet Transform

Key empirical fact: histograms of coefficients by subband suggest a Laplacian distribution (Simoncelli and Adelson, 1996):



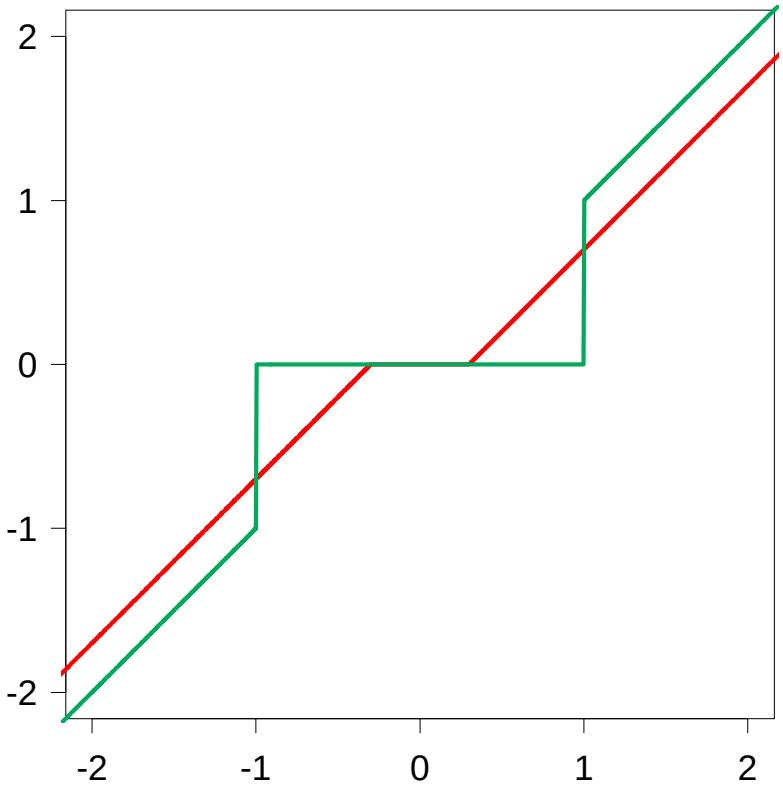
MDL in Wavelet Denoising and Compression

Donoho and Johnstone's denoising model in the wavelet domain:

$$y = \beta + \epsilon$$

where ϵ iid $N(0, \sigma^2)$.

- Thresholding denoises and also "compresses" because it sets coefficients to zero.
- Model selection also sets coefficients to zero and estimates for the non-zero ones.



Adding iid Gaussian Noise

- Donoho and Johnstone's (1994) VisuShrink uses hard threshold $\sigma\sqrt{2\log n}$, and

- Saito's (1994) two-stage MDL uses hard threshold $\sigma\sqrt{3\log n}$.

For images they both set too many coefficients to zero, even though they possess minimax optimality properties over Besov spaces.

$lMDL$ for simultaneous wavelet image denoising and compression

Ref: Hansen and Yu, (1999b)

Assume noise $\sigma = 1$.

IID model for each subband:

β's are either 0 or from a Lap(λ).

Marginally,

$$\beta \sim p\delta_0 + (1 \Leftrightarrow p)\text{Lap}(\lambda).$$

Let γ be the 0-1 model or state vector.

If $\gamma = 0$, $y \sim \phi(y)$ and use $\widehat{\beta} = 0$.

If $\gamma = 1$, $y \sim m_\lambda(y) = Lap * \phi$
and use $\widehat{\beta} =$ posterior mean.

Fortunately, both $m(y)$ and $\widehat{\beta}$ are closed-form.

$lMDL$ Multi-stage coding (subband-dependent)

Total code length:

$$\text{lMDL}(y) = L_{\text{total}}(y) = L(y|\gamma) + L(\gamma|\widehat{p}) + L(\widehat{p})$$

$lMDL$: choose γ to minimize $L_{\text{total}}(y)$.

Hansen and Yu (1999b) show that $lMDL$ thresholds at

$$T_{lMDL} = h^{-1}((1 \Leftrightarrow \widehat{p})/\widehat{p}),$$

where $h(y) = m(y)/\phi(y)$ (increasing), and

$$\widehat{p} = \text{ proportion of } \{y, h(y) > 1\}.$$

$lMDL$ Multi-stage coding (subband-dependent)

- Estimate p by $\widehat{p} = \#\{y : \frac{m(y)}{\phi(y)} > 1\}/n$ and code $\widehat{p}$ with

$$L(\widehat{p}) = \log n/2.$$

- Given $\widehat{p}$, code the state variable or model γ using a Bernoulli coder

$$L(\gamma|\widehat{p}) = -\sum_{ij}(1-\gamma_{ij})\log(\widehat{p}) - \sum_{ij}\gamma_{ij}\log(1-\widehat{p})$$

- Given γ, code y

$$L(y_{ij}|\gamma) = -\log\phi(y_{ij}) \quad \text{if } \gamma_{ij} = 0;$$

$$L(y_{ij}|\gamma) = -\log m(y_{ij}) \quad \text{if } \gamma_{ij} = 1$$

$$L(y|\gamma) = \sum L(y_{ij}|\gamma)$$

Results (test image Lena; SNR=4.4)

Methods in comparison:

1. MAP Soft-thresholding (Moulin and Liu, 1998):
$T_{MAP} = \sqrt{2}\sigma^2/\sigma_\beta$.

2. Optimal MSE Soft-thresholding (Chang, Yu and Vetterli , 1997):
$T_{MSE} = \sigma^2/\sigma_\beta$.

It outperforms most of the time SureShrink (by up to 6%), and is simpler to compute.

3. lMDL

$T_{lMDL} = h^{-1}((1 \Leftrightarrow \widehat{p})/\widehat{p})$.

All are iid model and Lap-based and of computation of order O(n).

Number of Coefficients Kept
($512 \times 512 = 262,144$ total pixels)

subband (level)	Moulin and Liu	Chang et al	lMDL
LL (3)	4096	4096	4096
HH (3)	1296	1909	591
HL (3)	3010	3298	1491
LH (3)	1697	2299	758
HH (2)	29	242	88
HL (2)	2716	5047	1563
LH (2)	283	1102	324
HH (1)	0	0	0
HL (1)	94	384	215
LH (1)	0	0	0
Total	13,221	18,377	9,126
Percentage (%)	5.0	7.0	3.5

measure	Moulin and Liu	Chang et al	lMDL
MSE	71.12	63.84	70.56
MSE/σ^2	0.178	0.160	0.176
PSNR	29.61	30.08	29.65

where PSNR $= 10 \log_{10}[255^2/MSE]$.

Performance Comparison for Lena with SNR =4.4

	Moulin and Liu	Chang et al	lMDL
% of coef. kept	5.0	7.0	3.5
MSE	0.178	0.160	0.176

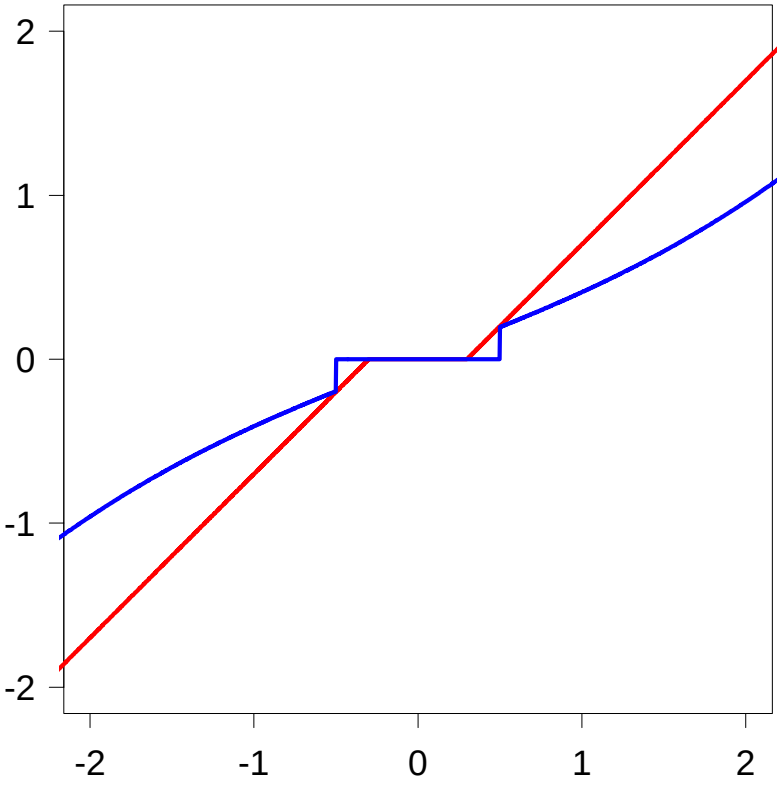
Moulin and Liu (1998): MAP Soft Thresholding
$T_{MAP} = \sqrt{2}\sigma^2/\sigma_\beta$.

Chang, Yu and Vetterli (1997): Optimal MSE Soft Thres.
$T_{MSE} = \sigma^2/\sigma_\beta$.

Chang et al. (1997)

lMDL (1999)

Moulin and Liu (1998)



Recap: lMDL achieves a good trade-off between denoising and compression.

Note: under the Lap model,

entropy of kept coefficients $\propto$ # of kept coefficients.

Thus lMDL uses half of the bit rate of Chang et al while losing only 0.4 dB PSNR (or 10% MSE), and uses 70 % of the bit rate as Moulin and Liu while having the same PSNR (or MSE).

(In comparison, in compression literature, cutting the bit rate into half results in about 3 dB of distortion loss.)

From Modeling to Coding:
Wavelet Image Coder

Yoo, Ortega and Yu (1999): one of the best image compression schemes.

Key idea: use Lap. model to quantize wavelet coefficients:

- Optimal quantization becomes a uniform quantizer under entropy constraint (Sullivan, 1996);
- Estimation based on quantized data is ML estimation;
- Bit allocation is done by table-look-up; and
- Spatial adaptivity is achieved by predicting the variance of the current pixel using the neighboring quantized pixels.

Performance Comparison in MSE For Lena

Rate (bpp)	EZW	SPIHT	EQ	CBCAQ
0.25	31.34	25.12	23.34	23.29
0.50	15.31	12.28	11.35	11.40
1.00	7.21	5.85	5.34	5.47

(Average Power or Energy Level β^2: 19,760.66)

EZW: Shapiro (1993).

SPIHT: Said and Pearlman (1996).

EQ: LoPresto et al (1997).

CBCAQ: Yoo et al (1999).

Concluding Remarks

We have seen at various levels the interaction between “codes” and “models”, or between information theory and statistics.

Future will see more of this interaction...

The goals of information theory and statistics will become more entangled. For example, statisticians have to take into account formally the compression aspect of their data; and the information/coding theorists have to compress while keeping inference in mind.

At the concrete coding level, models will be useful for

- image
- video
- speech
- hyperspectral (or curve) data
- ...

At the meta-coding level, MDL principle gives rise to $gMDL$ and $lMDL$ with impressive performances in their respective problem.

It will be useful for

- linear models
- time series
- classification problems based on hyperspectral data (e.g. forestry)
- nonparametric estimation
- ...