Model Selection using Predictive Risk

Bob Stine

May 11, 1998

- Outline
  - Predictive risk (out-of-sample accuracy) as criterion
  - Unbiased estimates:
    Mallows' C_p, Akaike's AIC, C-V $\Rightarrow$ $\quad |z| > \sqrt{2}$
  - Adjusting for selection:
    Risk inflation, hard thresholding $\Rightarrow$ $\quad |z| > \sqrt{2 \log p}$
- Goals
  - Convey origins of the methods
  - Characterize strengths, weaknesses

Regression Model

True Model

Rather than assume $E\,Y = X\beta$, leave mean unspecified:

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

Out-of-sample prediction error

Given p covariates $X = [X_1, X_2, \ldots, X_p]$, prediction MSE

$$PMSE(X) = E\|Y^* - X\hat{\beta}\|^2/n, \quad Y^* \text{ ind } Y$$

where norm is the sum of squares, $\|Y\|^2 = Y'Y = \sum_i y_i^2$.

Projection error Denote "hat matrix" $H_x = X(X'X)^{-1}X$, then

$$\begin{aligned} n\,PMSE(X) &= E\|Y^* - \eta\|^2 + E\|\eta - X\hat{\beta}\|^2 \\ &= n\sigma^2 + E\|\eta - H_x\eta + H_X\eta - X\hat{\beta}\|^2 \\ &= n\sigma^2 + \|\eta - H_x\eta\|^2 + E\|H_x\eta - H_xY\|^2 \\ &= \underbrace{n\sigma^2}_{\text{common}} + \underbrace{\|(I - H_x)\eta\|^2}_{\text{wrong X's}} + \underbrace{(E\|H_x\epsilon\|^2 = p\sigma^2)}_{\text{est error}} \end{aligned}$$

Working Model

Avoid common projection error $(I - H_x)\eta$, and let β denote projection of η into column span of X:

$$Y = X\beta + \epsilon \quad \text{where} \quad X\beta = H_x\eta\ .$$

More on the Regression Model

Covariates

Collection of p *potential* predictors, $X = [X_1, \ldots, X_p]$.

Working Model Add normality,

$$Y = X\beta + \epsilon\ , \qquad \epsilon_i \sim N(0, \sigma^2)$$

Robustness?

Central limit theory handles estimates, but one might question squared error as the right measure of loss.

Subset/selection coefficients

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

$$X_\gamma,\ \beta_\gamma \quad \text{defined by} \quad \beta_j \in \beta_\gamma \iff \gamma_j = 1$$

The number of fitted coefficients is $q = \sum_j \gamma_j = |\gamma|$.

True subset

Some of the members of β are possibly zero. Want to avoid this subset (perhaps) and isolate the meaningful predictors. Denote the subset of $\beta_j \neq 0$ by γ^*.

Orthogonal Regression

Selecting basis elements

n orthogonal predictors X_j, $\qquad X'X = n\,I_n$

Estimates

$$\begin{aligned}\hat{\beta}_j &= \frac{X_j'Y}{X_j'X_j} = \frac{X_j'(X\beta+\epsilon)}{n} \\ &= \beta_j + \frac{X_j'\epsilon}{n} \qquad \text{CLT} \\ &= \beta_j + \frac{\sigma Z}{\sqrt{n}} \qquad Z \sim N(0,1)\end{aligned}$$

Test statistic

Note the "mean-like" standard error $\text{SE}(\hat{\beta}_j) = \sigma/\sqrt{n}$. If we know σ^2, then test $H_0 : \beta_j = 0$ with

$$z_j = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} = \frac{\sqrt{n}\hat{\beta}_j}{\sigma}$$

Contribution to fit

Regression SS is

$$\hat{\beta}'(X'X)\hat{\beta} = n\sum \hat{\beta}_j^2$$

so X_j improves fit by adding

$$n\hat{\beta}_j^2 = \sigma^2 \underbrace{(\frac{\sqrt{n}\beta_j}{\sigma} + Z)^2}_{\text{non-central } \chi^2}$$

Mallow's C_p

Problem (Mallows 1964, *Technometrics* 1973)

Given a model with p covariates, $Y = X\beta + \epsilon$, find an unbiased estimate of the prediction MSE.

Prediction MSE

$$\begin{aligned} nPMSE(\hat{\beta}) &= E\|Y^* - X\hat{\beta}\|^2 \\ &= n\sigma^2 + E\|X\beta - X\hat{\beta}\|^2 \\ &= n\sigma^2 + E\|H_x\epsilon\|^2 \\ &= (n+p)\sigma^2 \end{aligned}$$

Residual SS suggests an estimator:

$$\begin{aligned} E(RSS_p) &= E\|Y - X\hat{\beta}\|^2 \\ &= E\|(I - H_x)\epsilon)\|^2 \\ &= (n-p)\sigma^2 \end{aligned}$$

leading to the unbiased estimator

$$pmse(X) = \frac{RSS_p + 2p\hat{\sigma}^2}{n}, \quad \hat{\sigma}^2 = RSS_p/(n-p)$$

Mallows' C_p

$$C_p = \frac{RSS_p}{\hat{\sigma}^2} + 2p - n$$

so that assuming we have the right model $C_p \approx p$.

C_p in Orthogonal Regression

Orthogonal setup

$$X_j \text{ adds } n\hat{\beta}_j^2 = \sigma^2 \left(\frac{\sqrt{n}\beta_j}{\sigma} + Z \right)^2 \text{ to Regr SS}$$

Coefficient threshold

- Add X_{p+1} to a model with p coefficients?
- Minimum C_p criterion implies

$$\text{Add } X_{p+1} \iff C_{p+1} < C_p$$

$$\begin{aligned} 0 < C_p - C_{p+1} &= \frac{RSS_p - RSS_{p+1}}{\sigma^2} + 2p - 2(p+1) \\ &= \frac{n\hat{\beta}_{p+1}^2}{\sigma^2} - 2 \\ &= z_{p+1}^2 - 2 \end{aligned}$$

- Add X_{p+1} when $|z_{p+1}| > \sqrt{2}$, (In the null case one chooses about 16% of variables, $\mathrm{P}\{|N(0,1)| > \sqrt{2}\} = 0.157$.)

Adjusted $\mathbf{R^2}$ criterion (Theil 1961)

Add variables which increase adjusted $\overline{R}^2$ (or decrease $\hat{\sigma}^2$):

$$\text{Add } X_{p+1} \iff \hat{\sigma}_p^2 > \hat{\sigma}_{p+1}^2 \iff 1 < \frac{n\hat{\beta}_{p+1}^2}{\hat{\sigma}_p^2} = z_{p+1}^2$$

Discussion of Mallows' C_p

Objective

Find unbiased estimate of PMSE for a given regression model.

Selection criterion

Minimize C_p (or unbiased estimate of $PMSE$).

Mallows' caveats

"[These results] should give pause to workers who are tempted to assign significance to quantities of the magnitude of a few units or even fractions of a unit on the C_p scale...
Thus using the 'minimum C_p' rule to select a subset of terms for least squares fitting cannot be recommended universally."

Issues

- Consistency
  Since testing at $\alpha = 0.16$, will asymptotically overfit.
- Where'd you get $\hat{\sigma}^2$?
  Fit the "full" regression model, assuming $p << n$.
- Effects of selection bias:
  Estimate of $PMSE$ for model with smallest observed $pmse$ is no longer unbiased.
- How to apply in problems other than regression?

Akaike's Information Criterion

Generalization (Akaike 1973)

Extends model selection beyond regression, motivated by notion of model approximation rather than prediction. Origins in FPE criterion for picking order of autoregression.

Kullback-Leibler divergence (aka, relative entropy)

How close are two models?

Let f_{θ^*} denote density of *true* model. How close is another model/density f_θ using parameters θ?

Akaike uses *expected* divergence, averaged over samplng dist of $\hat{\theta}$:

$$E_Y D(f_{\theta^*} \| f_{\hat{\theta}}) = E_Y \int \underbrace{\log \frac{f(Y^*, \theta^*)}{f(Y^*, \hat{\theta})}}_{\text{log L.R.}} f(Y^*, \theta^*) dY^* \geq 0$$

where $Y^* \sim f_{\theta^*}$ is independent of $\hat{\theta} = \hat{\theta}(Y)$.

Notes

Likelihood ratio Divergence is integrated log of the likelihood ratio of the true model to the fitted model.

Out of sample LR evaluated at data Y^* using estimate $\hat{\theta}(Y)$ from *independent* sample Y.

AIC in Gaussian Problems

Divergence

Abbreviate densities and focus on parameters,

$$E_Y K(\theta^*, \hat{\theta}) = E_Y \int \log \frac{f(Y^*, \theta^*)}{f(Y^*, \hat{\theta})} f(Y^*, \theta^*) dY^*$$

Gaussian regression

Assume σ^2 given, then with $E\,Y = \eta$ and β_p denoting projection of η into p dimensional subspace,

$$\begin{aligned} E_Y K(\beta, \hat{\beta}_p) &= E_{Y^*,Y} \log \frac{e^{-1/2\sigma^2 \|Y^* - \eta\|^2}}{e^{-1/2\sigma^2 \|Y^* - X\hat{\beta}_p\|^2}} \\ &= \frac{1}{2\sigma^2} E_{Y^*,Y} \|Y^* - X\hat{\beta}_p\|^2 - \|Y^* - \eta\|^2 \\ &= \frac{1}{2\sigma^2} E_Y \|\eta - X\hat{\beta}_p\|^2 \\ &= \frac{1}{2\sigma^2} (p\sigma^2 + \|\eta - X\beta_p\|^2) \end{aligned}$$

Unbiased estimate

Add $2p\sigma^2$ to residual SS ($-2\sigma^2 \times$ log likelihood) as in C_p,

$$\begin{aligned} E\ RSS_p &= E\|Y - X\beta + X\beta - X\hat{\beta}_p\|^2 \\ &= (n-p)\sigma^2 + \|X\beta - X\beta_p\|^2 \end{aligned}$$

$$\Rightarrow \hat{K} = 2\sigma^2 (\underbrace{p - \log f(Y, \hat{\beta}_p)}_{\text{minimize}})$$

General Form of AIC

Focus on varying part of criterion

$$\begin{aligned} E_Y K(\theta^*, \hat{\theta}_p) &= E_Y \int \log \frac{f(Y^*, \theta^*)}{f(Y^*, \hat{\theta}_p)} f(Y^*, \theta^*) dY^* \\ &= E_Y \int \log f(Y^*, \theta^*) f(Y^*, \theta^*) dY^* \\ &\quad -E_Y \int \log f(Y^*, \hat{\theta}_p) f(Y^*, \theta^*) dY^* \end{aligned}$$

How to estimate $E_Y \log f(Y^*, \hat{\theta}_p)$?

Use sample log likelihood (as in using RSS to estimate PMSE)

$$\sum \log f(Y_i, \hat{\theta}_p), \quad Y_i \sim f_{\theta^*}$$

Penalty Use the quadratic approximation,

$$\log f(Y, \theta) = \ell(Y, \theta) \approx \ell(Y, \hat{\theta}) + \tfrac{1}{2} \|\theta - \hat{\theta}\|_I^2 , \quad \hat{\theta} = \text{MLE} ,$$

and I is the information matrix at $\hat{\theta}$ with $||x||_I^2 = x' I x$.

On avg $E\,\ell(Y^*, \hat{\theta}_p^*) = E\,\ell(Y, \hat{\theta}_p)$,

$$\begin{aligned} E\ell(Y^*, \hat{\theta}_p) - \ell(Y, \hat{\theta}_p) &\approx (E\,\|\hat{\theta}_p - \hat{\theta}_p^*\|_I^2)/2 \\ &= (E\,\|\hat{\theta}_p - \theta_p + \theta_p - \hat{\theta}_p^*\|_I^2)/2 \\ &= E\,\|\hat{\theta}_p - \theta_p\|_I^2 = E\chi_p^2 = p , \end{aligned}$$

where θ_p is projection of θ into p dimensions (e.g., Brown, *Geometry of Exponential Families*).

Discussion of AIC

Objective

Minimize unbiased estimate of divergence via

$$\textit{Penalized log-likelihood:} \quad p - \sum \log f(Y_i, \hat{\theta}_p)$$

Comments

- **Equivalence to C_p**:
  Out-of-sample log-likelihood $\propto$ prediction MSE for normal.
  Threshold for orthogonal regression remains at $|z_j| > \sqrt{2}$.
- **Parametric**:
  Nested parametric models with known form of likelihood.
- **Consistency**:
  Since amounts to a test with low threshold, will make some type I errors regardless of n. Hence, not consistent.
  Do I care?
- **Origins and true model**:
  Fitting (nested) autoregressions, $AR(1)$, $AR(2), \ldots$. One seldom believes any such model is "the true model."
- **Selection bias**:
  Estimate of relative entropy for model with smallest observed penalized likelihood is no longer unbiased.

Cross-Validation

Motivation Stone 1974

Estimate properties of prediction rule by direct calculation.

Leave-one-out

Estimate out-of-sample prediction squared error from

$$CVSS = \sum_i (y_i - x_i' \hat{\beta}_{(-i)})^2$$

where $\hat{\beta}_{(-i)}$ denotes slope estimate *without* using the ith observation.

Simplified calculation

Use expressions for $\hat{\beta}_{(-i)}$ in terms of $\hat{\beta}$ and residuals,

$$\begin{aligned} CVSS &= \sum_i (y_i - x_i\hat{\beta} + x_i'(\hat{\beta} - \hat{\beta}_{(-i)}))^2 \\ &= \sum_i \left(\underbrace{y_i - x_i'\hat{\beta}}_{e_i} + \underbrace{x_i'(X'X)^{-1}x_i}_{h_i} \frac{e_i}{1-h_i} \right)^2 \\ &= \sum_i e_i^2 \left(1 + \frac{h_i}{1-h_i}\right)^2 \\ &= \sum_i \frac{e_i^2}{(1-h_i)^2} \ , \end{aligned}$$

where h_i are the leverages associated with the fitted model.

Cross-Validation $\approx C_p$

Generalized cross-validation

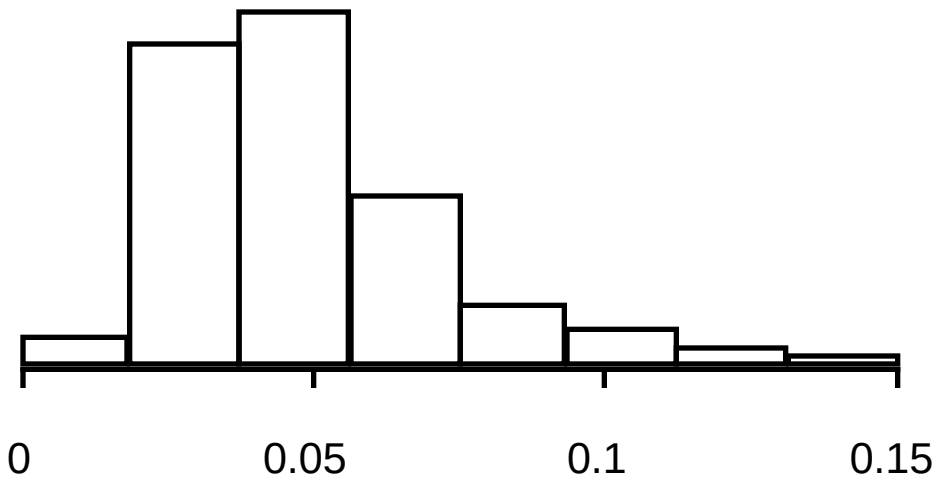
Replace h_i by its average p/n,

$$\begin{aligned}\frac{CVSS}{n} &= \frac{1}{n}\sum_i \frac{e_i^2}{(1-h_i)^2} \\ &\approx \frac{1}{n}\sum_i \frac{e_i^2}{(1-p/n)^2} \\ &= \frac{RSS}{n-p}\left(\frac{n}{n-p}\right) = s_p^2\left(1+\frac{p}{n}\right) ,\end{aligned}$$

inflating s_p^2 by the C_p adjustment $(1+p/n)$.

How good are the approximations?

Histogram of h_i for simulated analysis with $n = 100$, fitting a constant and 4 "near orthogonal" predictors ($p = 5$):



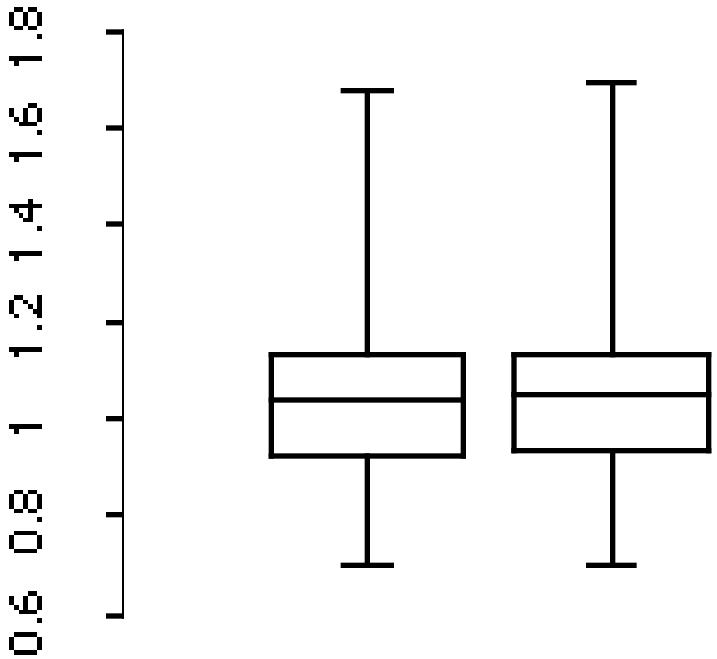
Cross-Validation Simulation

Estimates of PMSE

In a simulation of 250 trials, C_p and $CVSS$ quite similar ($p = 5$, $\sigma^2 = 1$, and standard errors for means $\approx .01$).

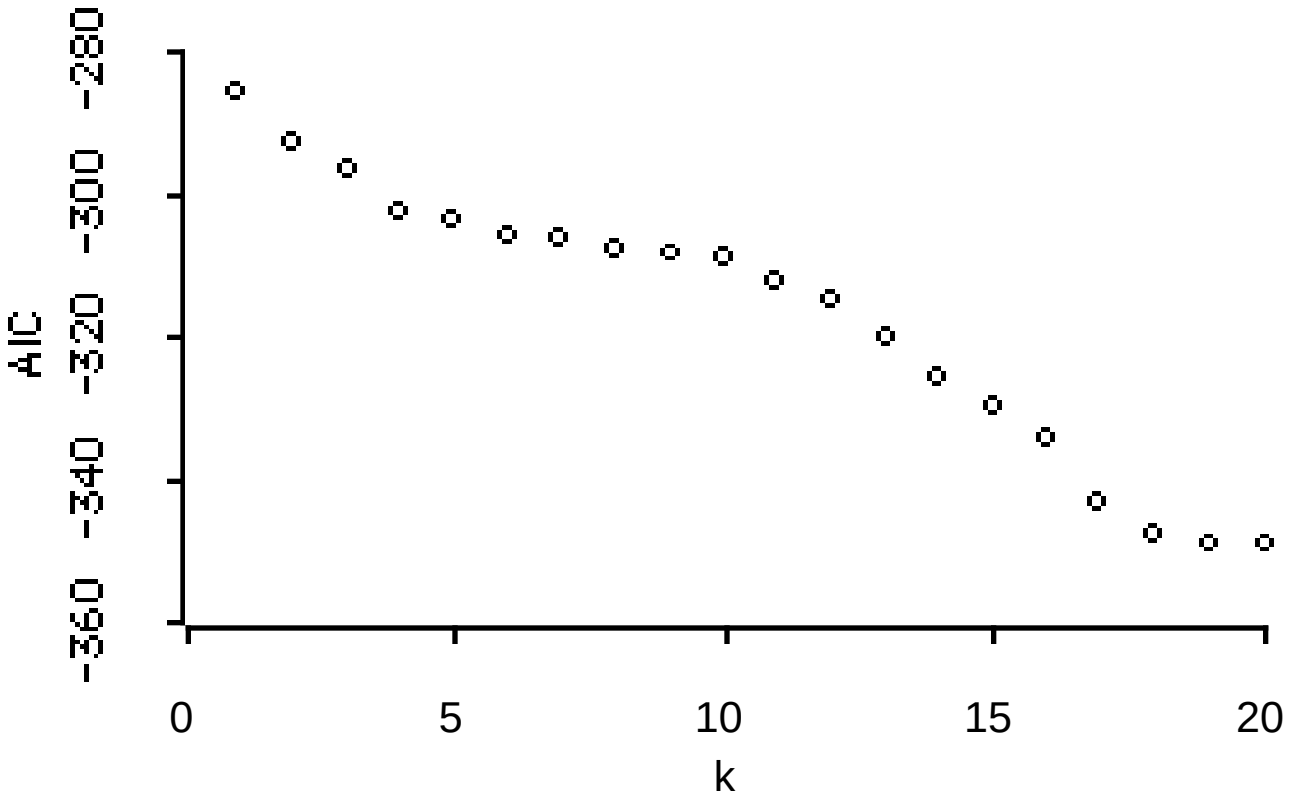
	Mean	SD
s^2	0.996	0.143
C_p	1.046	0.150
CV	1.050	0.151

and $\text{corr}(CV, C_p) = 0.998$.



Criteria in McDonald's Example

AIC



Impact of Selection

Variable selection

Suppose that q variables chosen from collection of p available predictors using a stepwise method, then assessed using C_p.

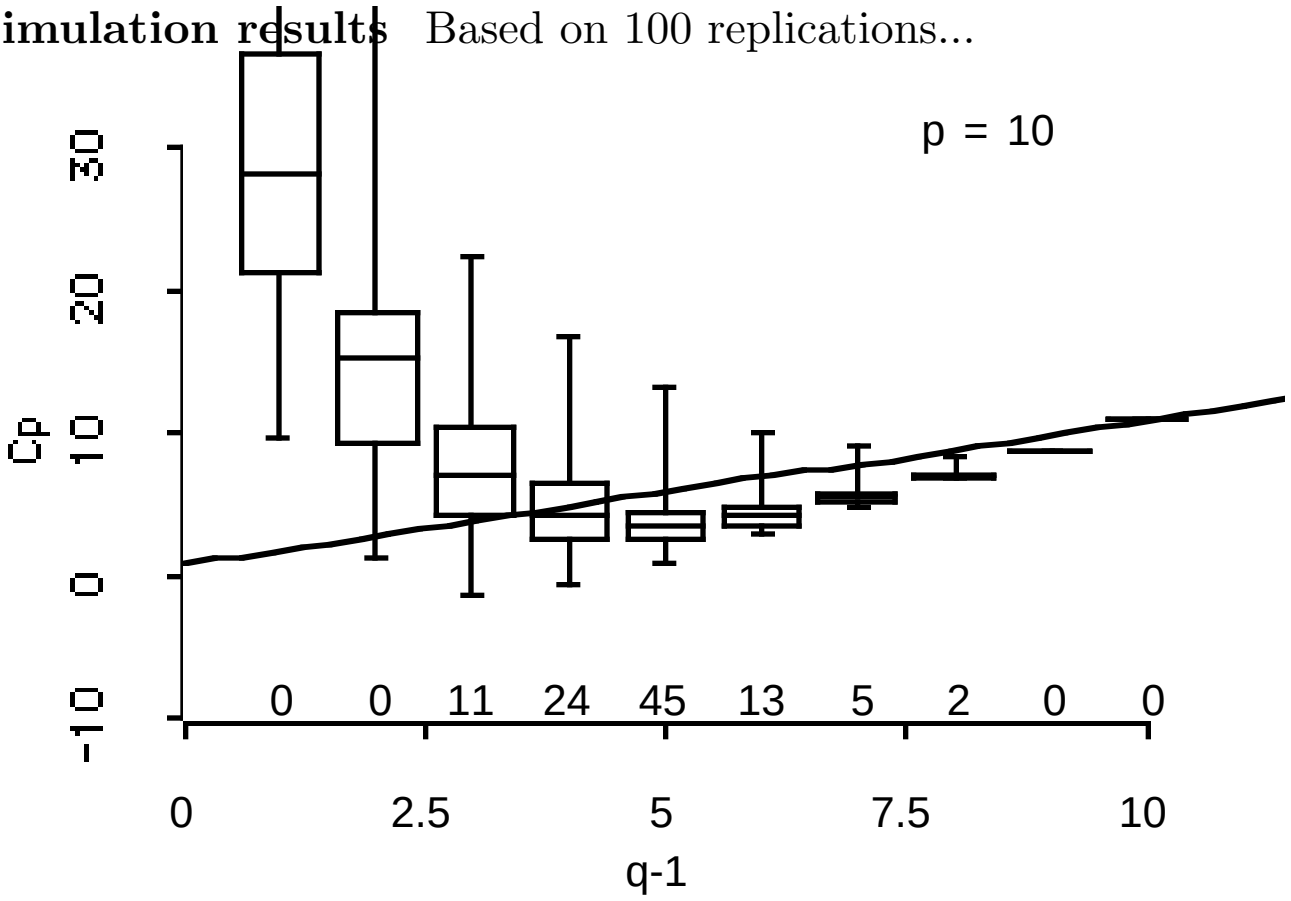
Model In addition to fitting a constant,

$p = 10$ possible "near orthogonal" predictors, $n = 100$

True coefficient vector, on z score scale is

$$z_\beta = (5,\ 4,\ 3,\ 2,\ 1,\ 0, \ldots 0)$$

Simulation results Based on 100 replications...



Impact of More Selection

Variable selection

Suppose now that q variables chosen from *larger* collection of 25 predictors using a stepwise method, again assessed using C_p.

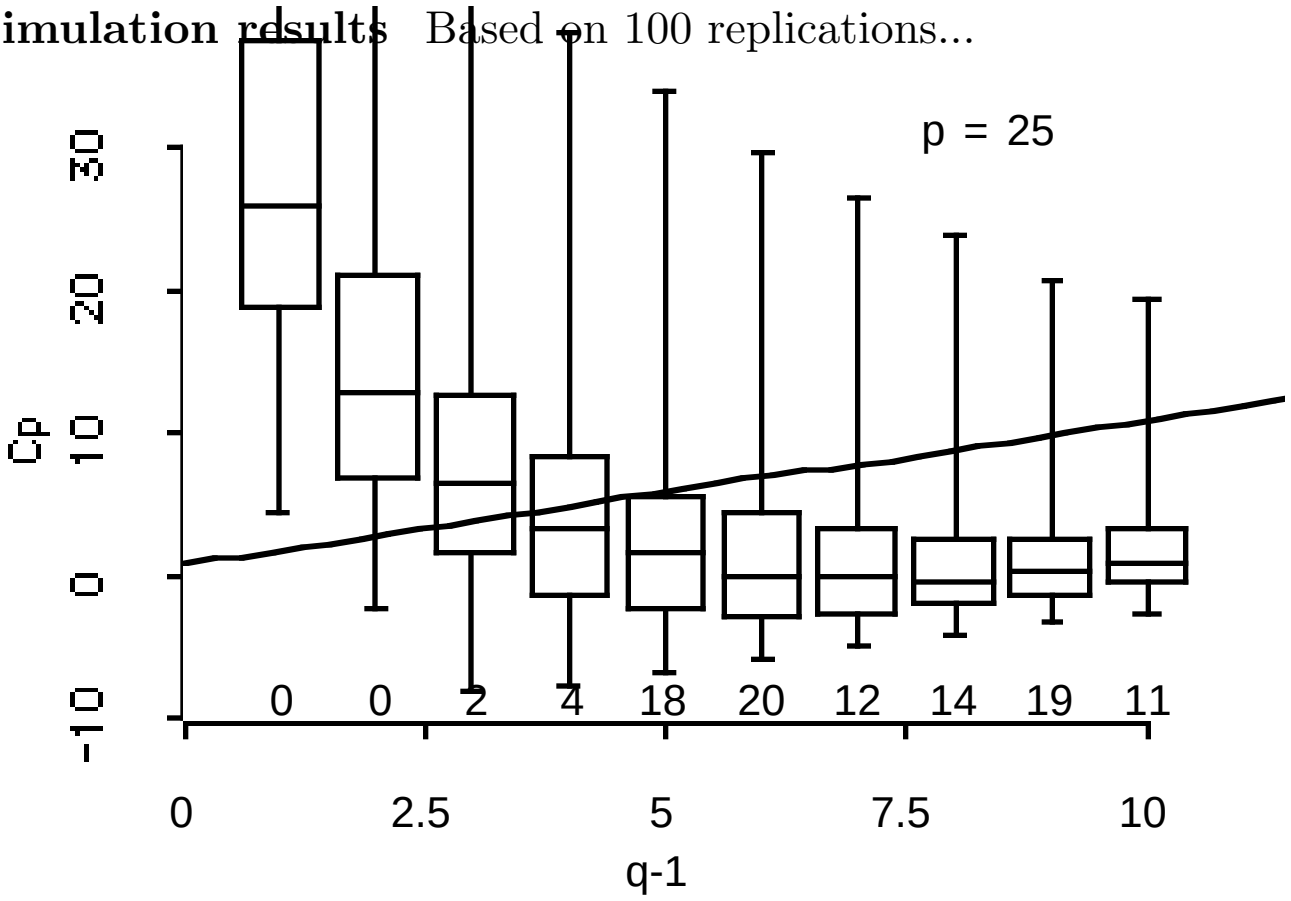
Model In addition to fitting a constant,

$p = 25$ possible "near orthogonal" predictors, $n = 100$

True coefficient vector, on z score scale is

$$z_\beta = (5,\ 4,\ 3,\ 2,\ 1,\ 0, \ldots 0)$$

Simulation results Based on 100 replications...



Summary of Predictive Risk

Penalize in-sample estimates

In-sample estimates of prediction error, e.g. residual SS, are optimistic, suggesting the model predicts better than it with.

Unbiased estimates

C_p and AIC provide simple adjustments that lead to unbiased estimates of predictive risk and relative entropy, *for a given model.*

Cross-validation

Direct computation by leave-one-out cross-validation duplicates C_p and AIC in the normal case.

Inconsistent model selection

Since choosing variables whose $|z| > \sqrt{2}$, 16% of predictors enter in null case. Asymptotically overfits *if one is willing to fix the model as n grows.*

Selection effects

Though unbiased for a given model, C_p and AIC are biased when applied to the model which minimizes the criterion.

Incorporating Selection in the Criterion

Problem in using C_p, AIC

Unbiased estimate of the predictive risk of *one* model, but

- Estimate of risk *is biased* in presence of selection, and
- If all $\beta_j = 0$, about 16% are accepted ($P(|Z| > \sqrt{2}) \approx 0.16$).

Alternative threshold

If cannot construct an unbiased estimate in presence of selection, can you guarantee a level of performance?

What threshold obtains this performance?

Minimax

Can we at least bound the worst-case predictive risk?

Minimax and model selection

Don't always work too well together:

$$\text{Use} \quad c\,\overline{Y} \quad \text{to estimate} \quad \mu$$

Squared error risk is

$$\begin{aligned} R(\mu,\, \overline{Y}) &= E(c\overline{Y} - \mu)^2 = E(c\overline{Y} - c\mu)^2 + (c\mu - \mu)^2 \\ &= c^2 \frac{\sigma^2}{n} + \mu^2 (c-1)^2 \end{aligned}$$

Unless $c = 1$, minimax risk $R(\mu,\, \overline{Y}) \to \infty$ as $\mu \to \infty$.

Testimators

Idea

Construct an estimator for μ from a test of $H_0 : \mu = 0$:

$$\hat{\mu} = \begin{cases} 0, & \text{accept } H_0, \qquad |\sqrt{n}\,\overline{Y}/\sigma| < \tau \\ \overline{Y}, & \text{reject } H_0\ . \end{cases}$$

Also known as "hard thresholding" with threshold τ.

Graph of testimator

Connection to model selection

In variable selection, each slope estimate is a testimator.

Key questions

What is the effect of the choice of the threshold τ on the predictive risk of the regression estimator?

What can be guaranteed of the testimator with threshold τ? Under what conditions?

Risk of Testimator

Model and estimator

Orthogonal regression with n observations,

$$\hat{\beta}_j = \beta_j + \frac{\sigma Z_j}{\sqrt{n}}, \qquad \frac{n\hat{\beta}_j^2}{\sigma^2} = \left(\underbrace{\frac{\sqrt{n}\beta_j}{\sigma}}_{\zeta_j} + \underbrace{Z_j}_{N(0,1)} \right)^2$$

Predictive risk

If I exclude β_j, then have a bias term:

$$E\|X_j\beta_j - X_j\hat{\beta}_j\|^2 = E\|X_j\beta_j\|^2 = n\beta_j^2$$

If I include β_j, then have a variance term:

$$E\|X_j\beta_j - X_j\hat{\beta}_j\|^2 = \sigma^2$$

The risk combines these, weighted by probabilities of occurrence,

$$\begin{aligned} R(\beta, \hat{\beta}_\tau) &= E\,\|X\beta - X\hat{\beta}_\tau\|^2 \\ &= \sigma^2 + n\sum_j \beta_j^2 \underbrace{P\left(\frac{n\hat{\beta}_j^2}{\sigma^2} \le \tau^2\right)}_{\text{exclude}} \\ &\quad + n\sum_j E((\beta_j - \hat{\beta}_j)^2 \underbrace{I\{\frac{n\hat{\beta}_j^2}{\sigma^2} > \tau^2\}}_{\text{include}}) \end{aligned}$$

Risk Function

Essential risk function

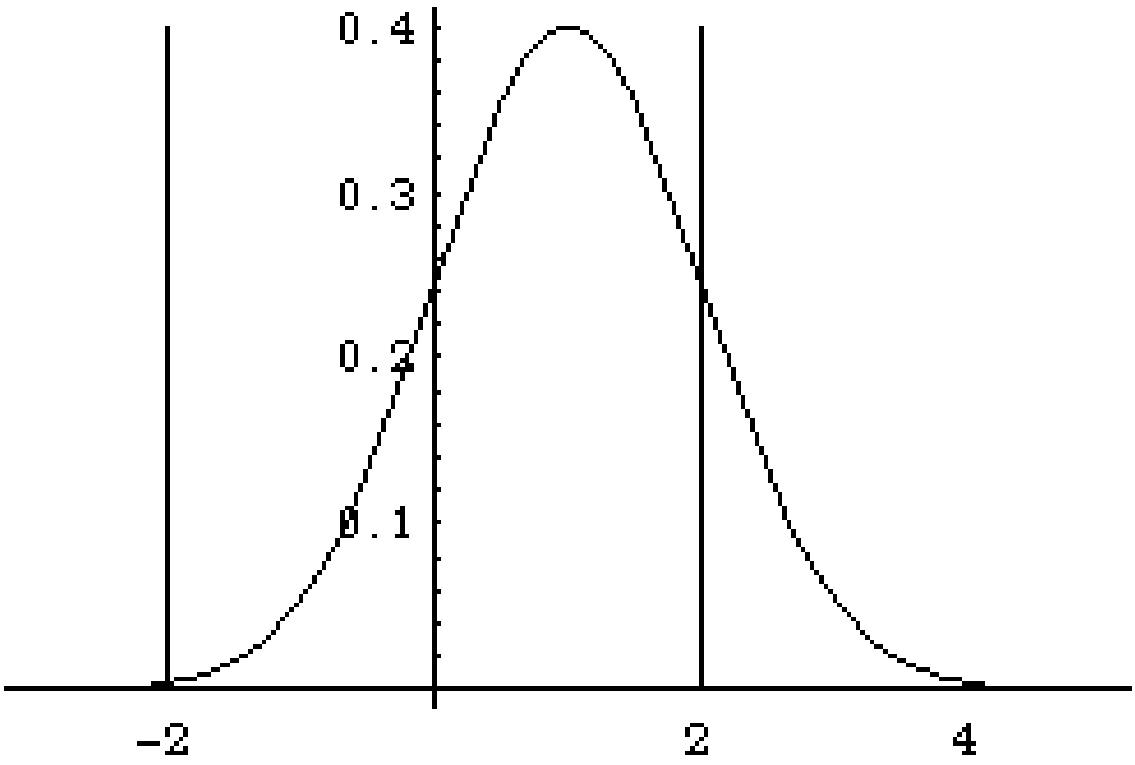
$$R(\beta, \hat{\beta}_\tau) = E \left\| X\beta - X\hat{\beta}_\tau \right\|^2 = \sigma^2 \left(1 + \sum_j R^*(\zeta_j, \tau) \right)$$

where for $Z \sim N(0,1)$,

$$R^*(\zeta, \tau) = \underbrace{\zeta^2 P((\zeta + Z)^2 \le \tau^2)}_{\text{exclude} \to \text{bias}} + \underbrace{E\left(Z^2 I\{(\zeta + Z)^2 > \tau^2\}\right)}_{\text{include} \to \text{var}}$$

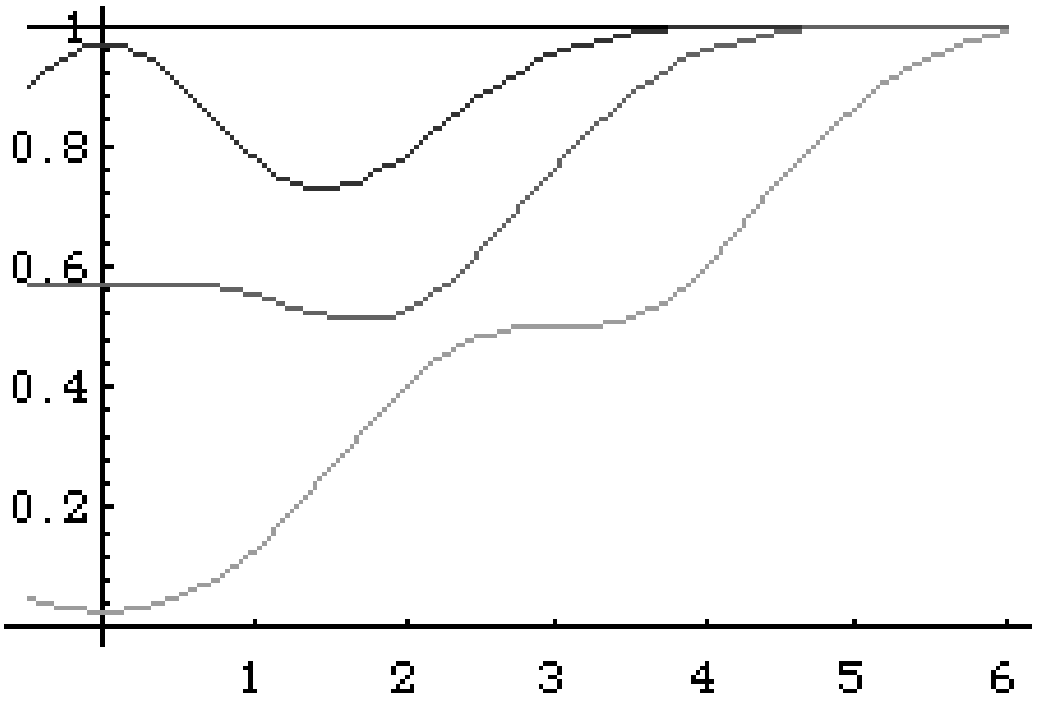
Plot

Distribution of observed z-score $\sqrt{n}\hat{\beta}/\sigma$ centered at $\zeta = \sqrt{n}\beta/\sigma = 1$ and $\tau = 2$:

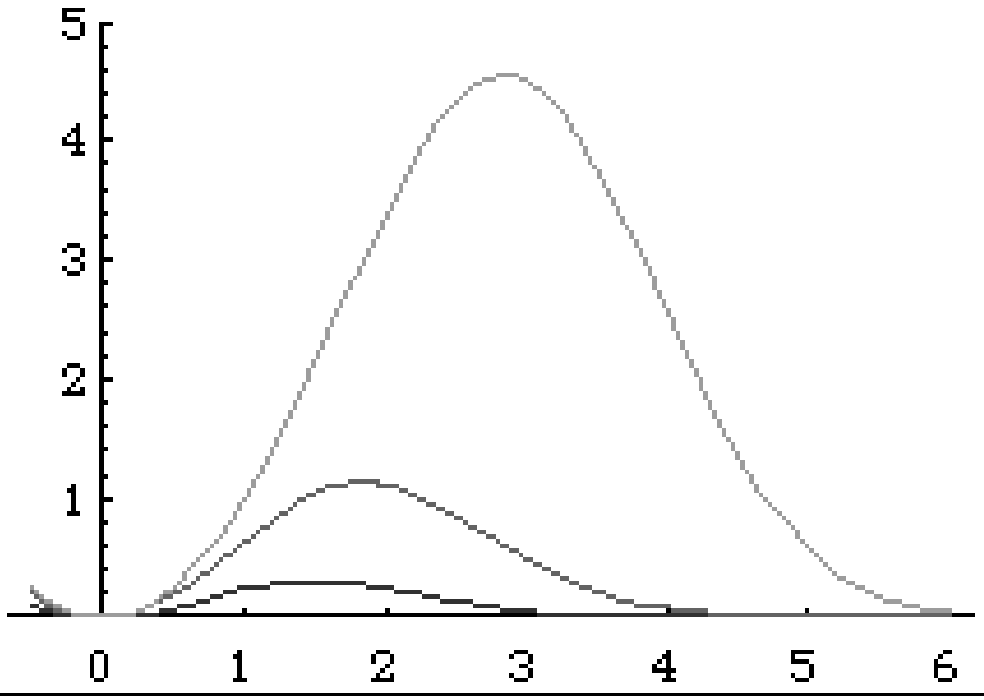


Risk Components

Variance $E\left(Z^2 I\{(\zeta+Z)^2 > \tau^2\}\right)$ $\tau = 0, \frac{1}{2}, \sqrt{2}, 3$

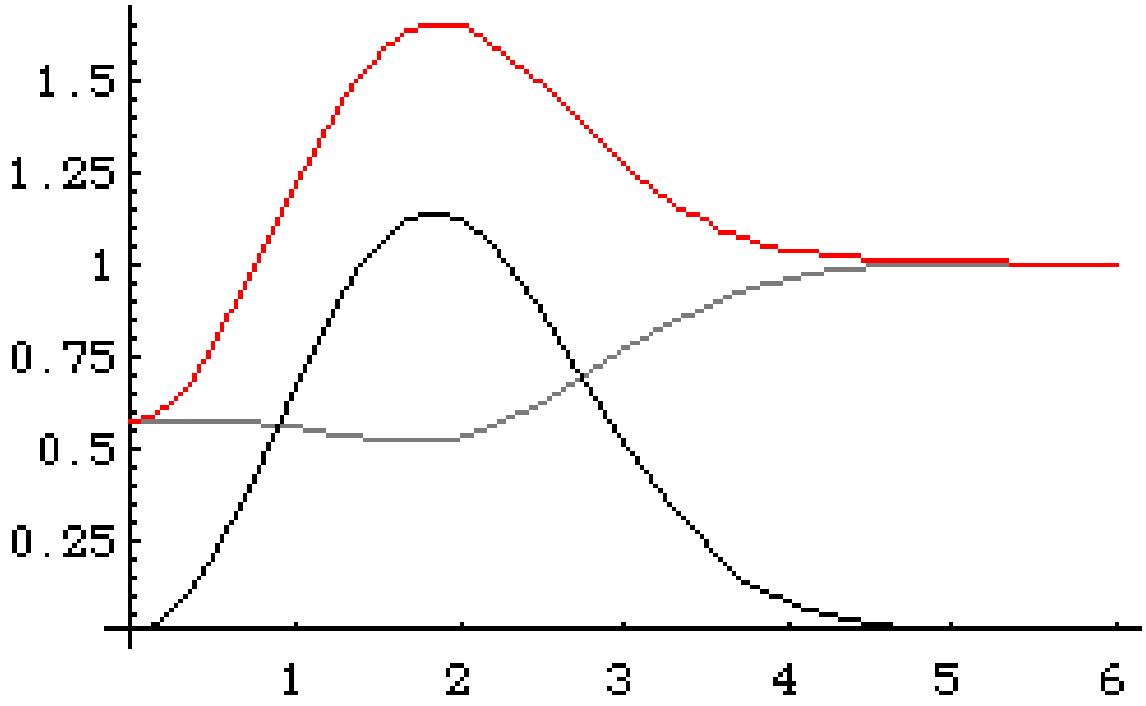


Bias $\zeta^2 P((\zeta+Z)^2 \leq \tau^2)$

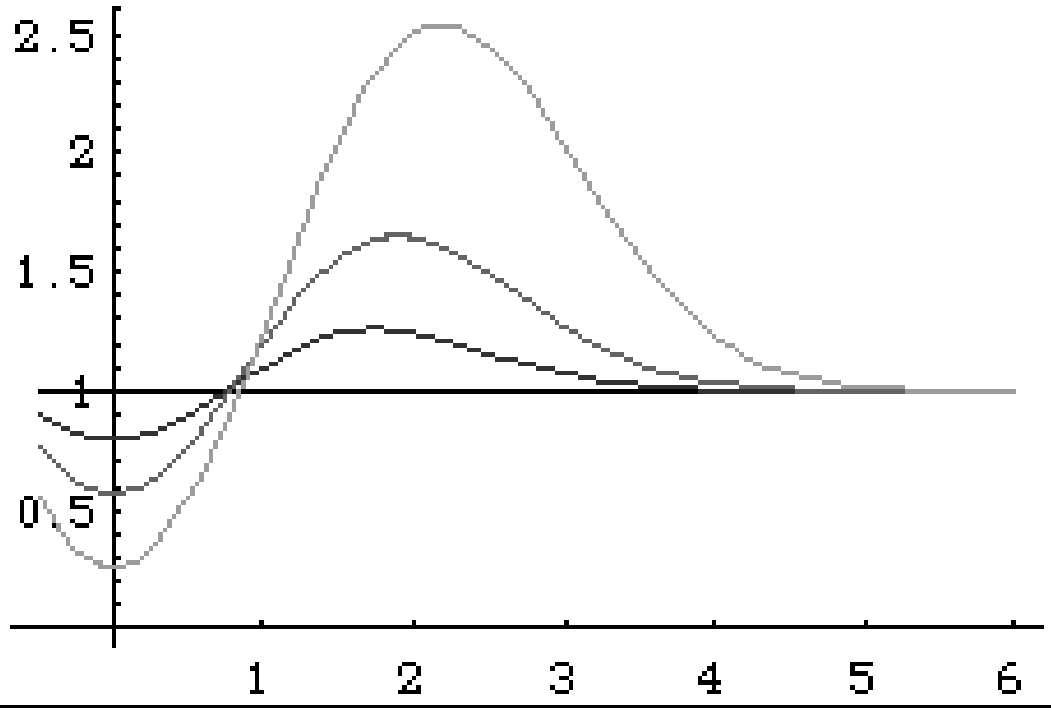


Risk Function

Components for $\tau = \sqrt{2}$, threshold for C_p



Risk $\tau = 0,\ 1,\ \sqrt{2},\ 3$ (Mallows 1973)



Where should we put the threshold?

For $Z \sim N(0,1)$,

$$R^*(\zeta,\tau) = \zeta^2 P((\zeta+Z)^2 \le \tau^2) + E\left(Z^2 I\{(\zeta+Z)^2 > \tau^2\}\right),$$

Minimax

Set threshold $\tau = 0$, using all variables: no bias, all variance.

$$R^*(\zeta,0) = 1 \quad \Rightarrow \quad R(\beta,\hat{\beta}_{\tau=0}) = p\sigma^2 \ .$$

Large Thresholds

Bias dominates, with relatively little variance since

$$E\left(Z^2 I\{(\zeta+Z)^2 > \tau^2\}\right) \le E\, Z^2 = 1$$

If $\zeta = \tau$, miss half: $R^* = \tau^2/2$.

If $\zeta = \tau - 2$, miss most:$R^* \approx (\tau-2)^2 \approx \tau^2$

Heuristic

For a large threshold, the maximum risk when fitting p coefficients is near

$$\sup_{\beta} R(\beta,\hat{\beta}_\tau) \approx p\ \sigma^2\ \tau^2$$

Lower Bound for Minimax Risk

Theorem (Foster & George, 1994)

For *any* estimator $\hat{\beta}$, with $|\gamma| = q$ nonzero true values,

$$\sup_{\beta_\gamma} R(\beta, \hat{\beta}) \geq \sigma^2 (2q \log p - o(\log p)) \ ,$$

asymptotically as $p \to \infty$ for fixed q.

Simpler problem Help from an oracle...

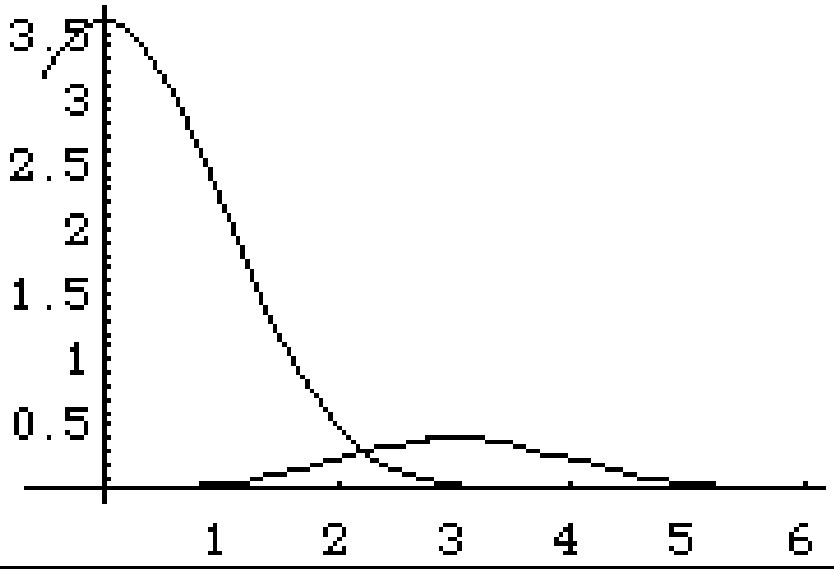
Suppose you know $q = 1$ *and* that the non-zero $\beta_j = C > 0$.

Do *not* know which coefficient $\neq 0$, and further treat γ_j as independent trials, with prob $1/p$.

What's the minimax risk in this case?

Utopian estimator via Bayes (Donoho & Johnstone, 1994)

Bayes gives the best estimator via posterior mean, and will use a rough approximation to this estimator.



Lower Bound for Minimax Risk, cntd

Utopian estimator via Bayes

Assuming $\gamma_1, \ldots, \gamma_p \sim B(1/p)$, Bayes gives the best estimator via posterior mean. Let $z_j = \sqrt{n}\hat{\beta}_j/\sigma$.

$$\begin{aligned} E(\beta_j|\hat{\beta}_j) &= 0 \times P(\beta_j = 0|\hat{\beta}_j) + C\ P(\beta_j = C|\hat{\beta}_j) \\ &= C\frac{P(\hat{\beta}_j|\beta_j = C)P(\beta_j = C)}{P(\hat{\beta}_j|\beta_j = C)P(\beta_j = C) + P(\hat{\beta}_j|\beta_j = 0)P(\beta_j = 0)} \\ &= C\frac{1}{1 + \frac{(p-1)\ N_{0,1}(z_j)}{N_{C,1}(z_j)}} = \frac{C}{1 + (p-1)e^{-C(z_j - C/2)}}\ , \end{aligned}$$

Posterior mode step-function approx to the posterior mean,

$$\hat{M}_j = \begin{cases} 0, & z_j < \frac{\log p}{C} + \frac{C}{2} \\ C, & \text{otherwise.} \end{cases}$$

Risk $(\sigma^2 = 1)$

$$\begin{aligned} R(\beta, \hat{M}) &= p\, C^2[P(z_1 > \frac{\log p}{C} + \frac{C}{2}|\beta_1 = 0)P(\beta_1 = 0) \\ &\quad + P(z_1 \leq \frac{\log p}{C} + \frac{C}{2}|\beta_1 = C)P(\beta_1 = C)] \\ &= C^2\left[(p-1)P(Z > \frac{\log p}{C} + \frac{C}{2}) + P(Z \leq \frac{\log p}{C} - \frac{C}{2})\right] \\ &= C^2\left[(p-1)\left(1 - \Phi\left(\frac{\log p}{C} + \frac{C}{2}\right)\right) + \Phi\left(\frac{\log p}{C} - \frac{C}{2}\right)\right] \end{aligned}$$

How large can "nature" make this risk by choice of C?

Minimax Risk Threshold

Maximum risk $(\sigma^2 = 1)$

If locate the non-zero value $C = \sqrt{2 \log p}$, then

$$\begin{aligned} R(\beta, \hat{M}) &\approx C^2 \left[p \left(1 - \Phi \left(\frac{\log p}{C} + \frac{C}{2} \right) \right) + \Phi \left(\frac{\log p}{C} - \frac{C}{2} \right) \right] \\ &\approx 2 \log p \left(\frac{1}{\sqrt{2 \log p}} + \Phi(0) \right) \\ &= \log p + \sqrt{2 \log p} \end{aligned}$$

At a slightly smaller value, say $\sqrt{2 \log p} - 2$, increases to

$$\sup_C R(\beta, \hat{M}) \approx 2 \log p$$

Results

- For small $|\gamma|$ and any $\hat{\beta}$, $R(\beta_\gamma, \hat{\beta}) \geq \sigma^2 |\gamma| \, (2 \log p)$.
- For large thresholds, $\sup_\beta R(\beta, \hat{\beta}_\tau) \approx p \sigma^2 \tau^2$.

Hard threshold, *RIC* criteria

Assume $|\gamma|$ is small (as with wavelets) and pick τ to obtain minimax risk:

$$\tau = \sqrt{2 \log p}$$

Close to the Bonferroni bound:

$$\Phi(x) \approx \frac{\phi(x)}{x} \;\Rightarrow\; \Phi^{-1}(\frac{1}{p}) \approx \sqrt{2 \log p} - \tfrac{1}{2} \frac{\log(2 \log p)}{\sqrt{2 \log p}}$$

"Ancient Model Selection"

Finding a cycle hidden in noise

- Power: sum of squares associated with pairs of coefficients in a "full" orthogonal harmonic regression (n even)

$$Y_t = A_0 + \sum_{j=1}^{n/2-1} A_j \cos\frac{2\pi jt}{n} + B_j \sin\frac{2\pi jt}{n} + A_{n/2}(-1)^t$$

$$A_j = (2/n)\sum_t Y_t \cos\frac{2\pi jt}{n}$$

- Regression SS for jth frequency:

$$SS_j = n\left(\frac{A_j^2 + B_j^2}{2}\right)$$

- Question: Does the $\max_j SS_j$ indicate significant variation?

R. A. Fisher's 1929 method (Bloomfield 1976, *Time Series*)

- Under null model and normality, $SS_j/\sigma^2 \overset{\text{iid}}{\sim} \mathcal{E},\ (\frac{1}{2}\chi^2_2)$.
- $X = \max_j SS_j/\sigma^2$, max of $m = n/2$ standard exponentials
- $P(X < x) = (1 - e^{-x})^m \Rightarrow P(X < x + \log m) \approx \exp(-e^{-x})$

$$\Rightarrow X \approx \log m$$

- Find "signal" if $X > \log m$.
- Corresponds to RIC threshold $2\log p$ for regression SS, with 2 dropped since looking at the average of two coefficients.

Less Conservative Procedure

Bonferroni For large p

$$RIC \text{ threshold} \approx \Phi^{-1}\left(\tfrac{p}{p+1}\right)$$

Why use this hard threshold for all of the coefficients?

Half-normal method C. Daniel 1959

Order the absolute z scores,

$$|z_{(1)}| > |z_{(2)}| > \cdots > |z_{(p)}|$$

Compare

$$|z_{(1)}| > \Phi^{-1}\left(\frac{p}{p+1}\right) \approx \sqrt{2\log p}$$

$$|z_{(2)}| > \Phi^{-1}\left(\frac{p-1}{p+1}\right) \approx \sqrt{2\log p/2}$$

$$|z_{(q)}| > \Phi^{-1}\left(\frac{p-q+1}{p+1}\right) \approx \sqrt{2\log p/q}$$

Adaptive criterion

Leads to a selection criterion similar to those I'll more carefully formulate in empirical Bayes and information theory.

Multiple testing

Simes (1986) result from testing multiple hypotheses adapted to variable selection by Abromovich and Benjamini (*aka*, step-up, step-down tests).

Conclusions

Orthogonal thresholds

Assuming n independent observations from identical model, p potential predictors, thresholds for coefficient z-scores are

Method	Threshold τ
C_p, AIC, cross-validation	$\sqrt{2}$
RIC, hard thresholding	$\sqrt{2 \log p}$

Selection criteria

Built-in prejudices for certain kinds of models:

RIC: Ideal basis should have only a few large coefficients, and obtains minimum risk against worst case model. (Oracle idea: Does as well as knowing which to use in the worst case problem.)

Hidden biases

Other selection method have hidden biases toward certain types of models, as suggested by *RIC*'s preference for few coefficients.

Bayesian ideas and information theory reveal more of these as well as ways to adapt to problem at hand.

Remaining issue

Once you have chosen a model, how well will it predict?