Bayesian Model Selection

Bob Stine

May 11, 1998

- Methods
  - Review of Bayes ideas
  - Shrinkage methods (ridge regression)
  - Bayes factors: threshold $|z| > \sqrt{\log n}$
  - Calibration of selection methods
  - Empirical Bayes (EBC) $|z| >\approx \sqrt{\log p/q}$
- Goals
  - Characteristics, strengths, weaknesses
  - Think about priors in preparation for next step

Bayesian Estimation

Parameters as random variables

- Parameter θ drawn randomly from prior $p(\theta)$.
- Gather data Y conditional on some fixed θ.
- Combine data with prior to form posterior,

$$p(\theta|Y) = \frac{p(Y|\theta)p(\theta)}{p(Y)}$$

- Drop marginal $p(Y)$ since constant given Y

$$\underbrace{p(\theta|Y)}_{\text{posterior}} \propto \underbrace{p(Y|\theta)}_{\text{likelihood}} \times \underbrace{p(\theta)}_{\text{prior}}$$

Key Bayesian example Inference on normal mean μ

- Observe $Y_1, \ldots, Y_n|\mu \sim N(\mu, \sigma^2)$, with $\mu \sim N(M, \nu^2 = c^2\sigma^2)$.
- Posterior is normal with (think about regression)

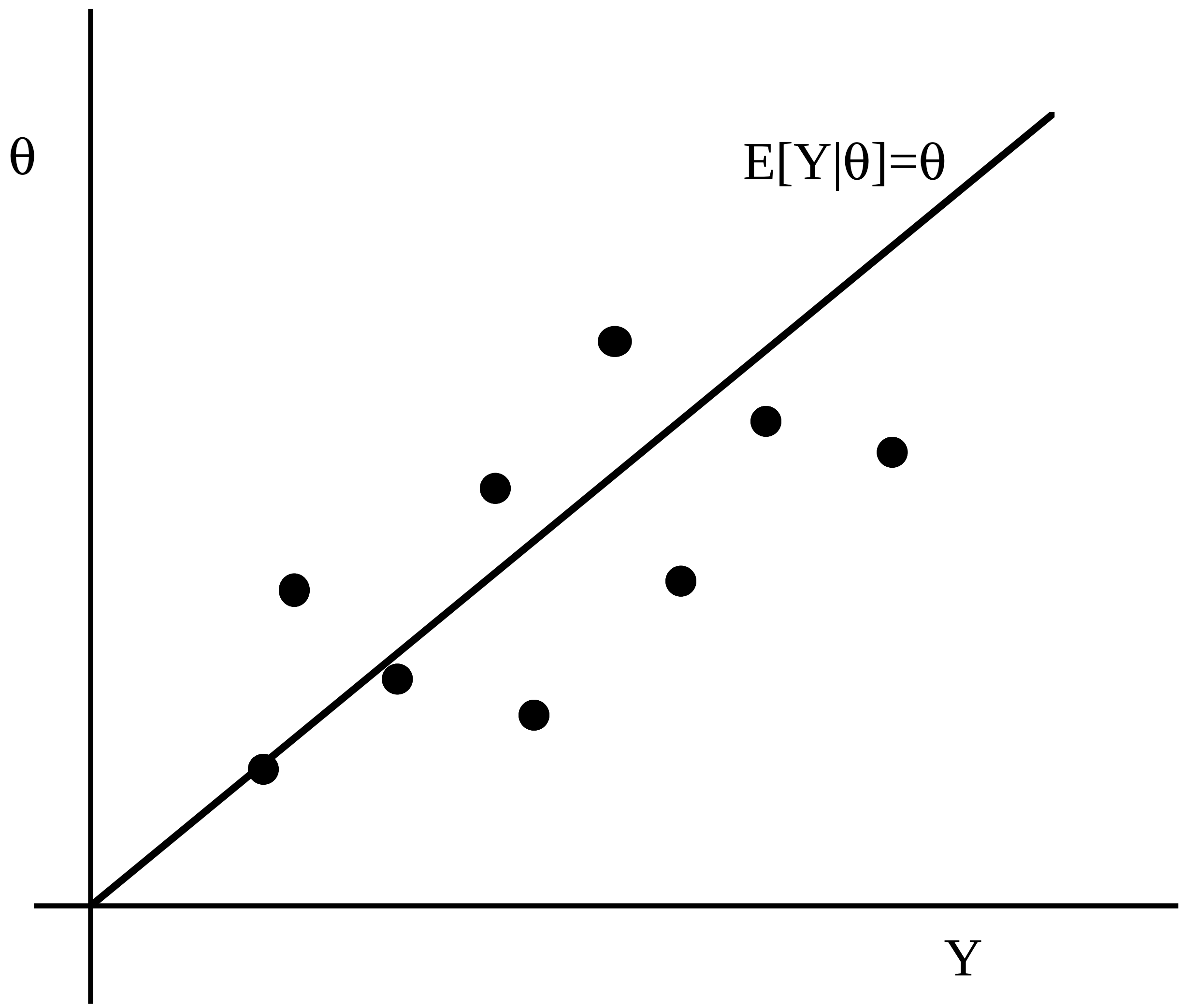
$$\begin{aligned} E(\mu|Y) &= M + \frac{\nu^2}{\nu^2 + \sigma^2/n}(\overline{Y} - M) \\ &= \left(\frac{n/\sigma^2}{n/\sigma^2 + 1/\nu^2}\right)\overline{Y} + \left(\frac{1/\nu^2}{n/\sigma^2 + 1/\nu^2}\right) M \end{aligned}$$

- In special case of $\sigma^2 = \nu^2$, "prior is worth one observation":

$$E(\mu|Y) = \left(\frac{n}{n+1}\right)\overline{Y} + \left(\frac{1}{n+1}\right)M$$

Why Bayes Shrinkage Works

Steigler (1983) discussion of article by C. Morris in *JASA*.



Regression slope is

$$\frac{\mathrm{Cov}(Y,\theta)}{\mathrm{Var}\ Y} = \frac{c^2\ \sigma^2}{(1+c^2)\sigma^2} = \frac{c^2}{1+c^2}$$

Ridge Regression

Collinearity

Originates in optimization problems where Hessian matrix in a Newton procedure becomes ill-conditioned. Marquardt's idea is to perturb the diagonal, avoiding singular matrix.

Hoerl and Kennard (1970) apply to regression analysis, with graphical aides like the ridge trace:

$$\text{plot} \quad \hat\beta = (X'X + \lambda I_p)^{-1} X'Y \quad \text{on} \quad \lambda$$

Bayes hierarchical model Lindley and Smith 1972

Data follows standard linear model, with a normal prior on slopes:

$$Y \sim N(X\beta, \sigma^2 I_n) \ , \quad \beta \sim N(0, c^2\sigma^2 I_p) \ .$$

The posterior mean shrinks toward 0,

$$\begin{aligned} E(\beta|\hat\beta) &= (X'X + \frac{1}{c^2} I_p)^{-1} X'Y \\ &= (I_p + \frac{1}{c^2}(X'X)^{-1})^{-1}\hat\beta \end{aligned}$$

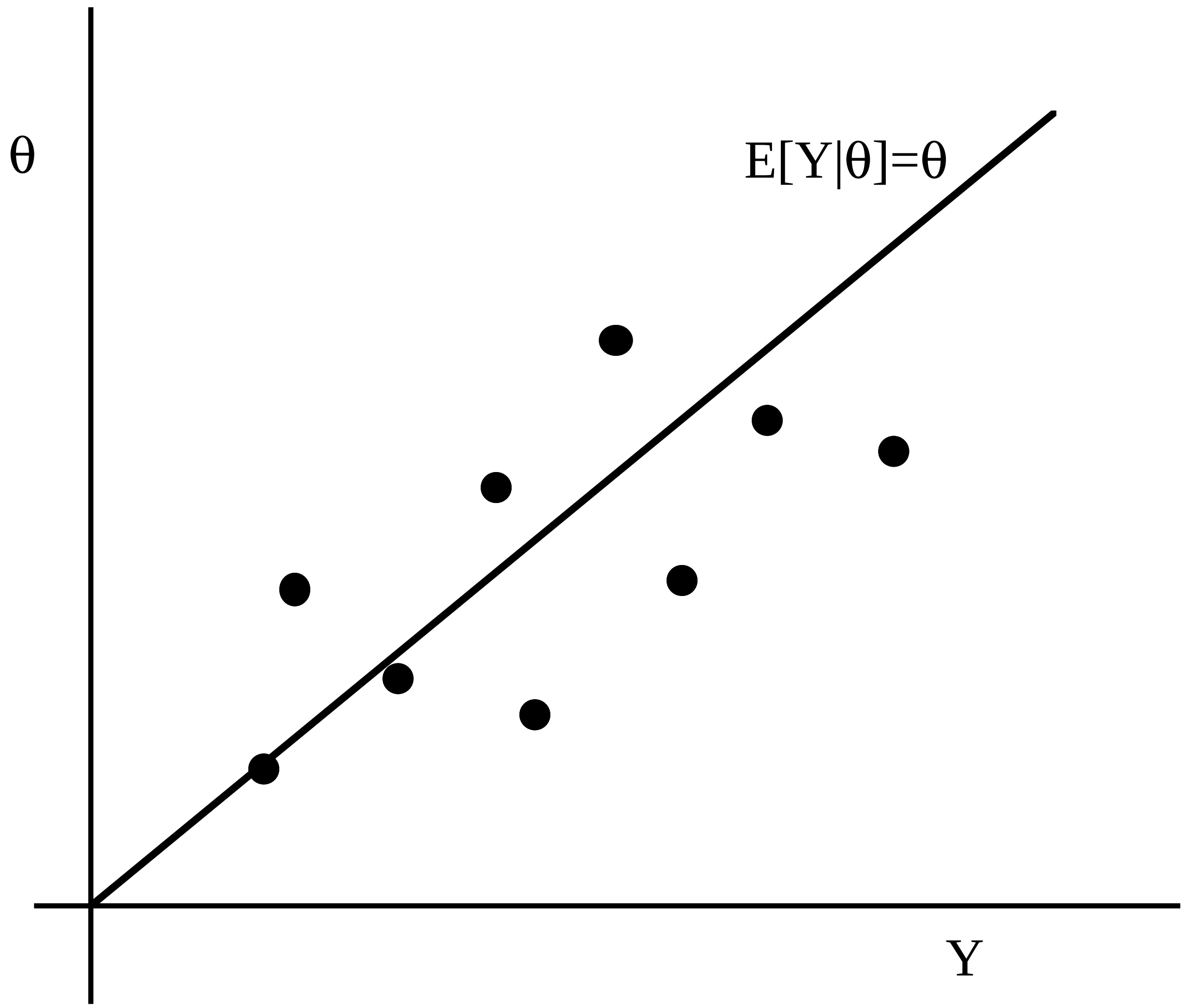
Shrinkage larger as $c^2 \to 0$.

Picture Collinear and orthogonal cases

Where is the variable selection?

How to Obtain Selection from Bayes

Main point To do more than just shrink linearly to prior mean, we have to get away from bivariate normal.



Three alternatives

- Spike-and-slab methods and *BIC*.
- Normal mixtures.
- Cauchy methods in information theory.

Bayesian Model Selection

Conditioning on models

- Jeffreys (1935), recently Raftery, Kass, and others
- Which model is most likely given the data? Think in terms of models M rather than parameters θ:

$$p(M|Y) = \underbrace{p(Y|M)}_{?} p(M)/p(Y)$$

- Express $p(Y|M)$ as average of the usual likelihood $p(Y|M,\theta)$ over the parameter space,

$$p(M|Y) = \int_\theta \underbrace{p(Y|\theta, M)}_{\text{usual like}} p(\theta|M) d\theta \; p(M)/p(Y)$$

Bayes factors

- Compare two models, M_0 and M_1.
- *Posterior odds* in favor of model M_0 over alternative M_1 are

$$\underbrace{\frac{p(M_0|Y)}{p(M_1|Y)}}_{\text{Posterior odds}} \; = \; \underbrace{\left(\frac{p(Y|M_0)}{p(Y|M_1)}\right)}_{\text{Bayes factor}} \times \underbrace{\frac{p(M_0)}{p(M_1)}}_{\text{Prior odds}}$$

- Since often $p(M_0) = p(M_1) = 1/2$, prior odds $= 1$ so that

Posterior odds $=$ Bayes factor $= K$

Using Bayes Factors

Bayes factor for comparing null model M_0 to alternative M_1

$$\underbrace{\frac{p(M_0|Y)}{p(M_1|Y)}}_{\text{Posterior odds }=} = \underbrace{\left(\frac{p(Y|M_0)}{p(Y|M_1)}\right)}_{\text{Bayes factor } K \,\times} \underbrace{1}_{\text{Prior odds}}$$

What's a big Bayes factor? Jeffreys' Appendix (1961):

$$\begin{array}{rcl} K > 1 & \Rightarrow & \text{"Null hypothesis supported."} \\ K < 1 & \Rightarrow & \text{"Not worth more than bare mention."} \\ K < 1/\sqrt{10} & \Rightarrow & \text{"Evidence against } H_0 \text{ substantial."} \\ K < 1/10 & \Rightarrow & \text{"strong"} \\ K < 1/10^{3/2} & \Rightarrow & \text{"very strong"} \\ K < 1/100 & \Rightarrow & \text{"Evidence against } H_0 \text{ decisive."} \end{array}$$

Computing

$$p(Y|M) = \int_\theta p(Y|\theta, M)\, p(\theta|M) d\theta$$

can be very hard to evaluate, especially in high dimensions in problems lacking a neat, closed-form solution.

Details in Kass and Raftery (1995).

Approximations?

Approximating Bayes Factors: BIC

Goal Schwarz (1976), *Annals of Statistics*

Approximate $p(Y|M) = \int_\theta p(Y|\theta, M)\, p(\theta|M) d\theta$.

Approach Make the integral look like a normal integral.

Define $g(\theta) = \log p(Y|\theta, M) p(\theta|M)$ so that

$$p(Y|M) = \int e^{g(\theta)} d\theta$$

Quadratic expansion around max at $\tilde{\theta}$ (post. mode):

$$\begin{aligned} g(\theta) &\approx g(\tilde{\theta}) + (\theta - \tilde{\theta})' H(\tilde{\theta})(\theta - \tilde{\theta})/2 \\ &\approx g(\tilde{\theta}) - (\theta - \tilde{\theta})'(I_\theta)^{-1}(\theta - \tilde{\theta})/2 \end{aligned}$$

where $H = [\partial^2 g/\partial\theta_i \partial\theta_j]$ and I_θ is information matrix.

Laplace's method For $\theta \in \mathbf{R}^p$, posterior becomes

$$\begin{aligned} p(Y|M) &\approx \exp(g(\tilde{\theta})) \int_\theta \exp[(-1/2)(\theta - \tilde{\theta})'(I_\theta)^{-1}(\theta - \tilde{\theta})] d\theta \\ &= \exp(g(\tilde{\theta}))\, (2\pi)^{p/2} |I_\theta|^{1/2} \end{aligned}$$

Log posterior approximately penalized log likelihood at MLE $\hat{\theta}$,

$$\begin{aligned} \log P(Y|M) &= \log p(Y|\tilde{\theta}, M) + \log p(\tilde{\theta}|M) + (1/2) \log |I_\theta| + O(1) \\ &= \log p(Y|\hat{\theta}, M) + \log p(\hat{\theta}|M) - (p/2) \log n + O(1) \\ &= \underbrace{\log p(Y|\hat{\theta}, M)}_{\text{log-likelihood at MLE}} - \underbrace{(p/2) \log n}_{\text{penalty}} + O(1) \end{aligned}$$

BIC Threshold 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?
- *BIC* criterion implies

$$\text{Add } X_{p+1} \iff \text{penalized like increases}$$

$$\log p(Y|\hat{\theta}_{p+1}) - \frac{p+1}{2}\log n > \log p(Y|\hat{\theta}_p) - \frac{p}{2}\log n$$

which implies that the change in the residual SS must satisfy

$$\frac{RSS_p - RSS_{p+1}}{2\sigma^2} = \frac{n\hat{\beta}_{p+1}^2}{2\sigma^2} = \frac{z_{p+1}^2}{2} > \frac{\log n}{2}$$

- Add X_{p+1} when

$$|z_{p+1}| > \sqrt{\log n}\ ,$$

so that effective "α level" $\to 0$ as $n \to \infty$.

Consistency of criterion

If there is a "true model" of dimension p, as $n \to \infty$ *BIC* will identify this model w.p. 1. In contrast, *AIC* asymptotically overfits since it tests with a fixed α level.

Bayes Hypothesis Test

Problem

Test $H_0 : \mu = 0$ vs $H_1 : \mu \neq 0$ given $Y_1, \ldots, Y_n \sim N(\mu, \sigma^2)$.

Pick the hypothesis with higher posterior odds.

Test of point null (Berger, 1985)

Under H_0, $\mu = 0$ and integration reduces to

$$p(H_0|\overline{Y}) = \underbrace{p(\overline{Y}|\mu_0 = 0)}_{N(0,\sigma^2/n)} p(H_0)/p(\overline{Y})$$

Under alternative $H_1 : \mu \sim N(0, \sigma^2)$,

$$p(H_1|\overline{Y}) = \left(\int_\mu p(\overline{Y}|\mu)p(\mu|H_1)d\mu \right) p(H_1)/p(Y)$$

Bayes factor Assume $p(H_0) = p(H_1) = \frac{1}{2}$ and $\sigma^2 = 1$,

$$\begin{aligned} \frac{p(H_0|\overline{y})}{p(H_1|\overline{y})} &= \frac{\overline{Y} \sim N(0, 1/n)}{\overline{Y} \sim N(0, 1 + 1/n)} \\ &= \sqrt{\frac{1 + 1/n}{1/n}} \frac{e^{-n\overline{y}^2/2}}{e^{-(n/(n+1))\overline{y}^2/2)}} \\ &\approx \sqrt{n} \, \frac{e^{-n\overline{y}^2/2}}{e^{-\overline{y}^2/2}} \approx \sqrt{n} \, e^{-n\overline{y}^2/2} \end{aligned}$$

which is equal to one when

$$n\overline{y}^2 = \log n \quad \text{or in } z \text{ scores} \quad |z| = \sqrt{\log n}$$

Bayes Hypothesis Test: The Picture

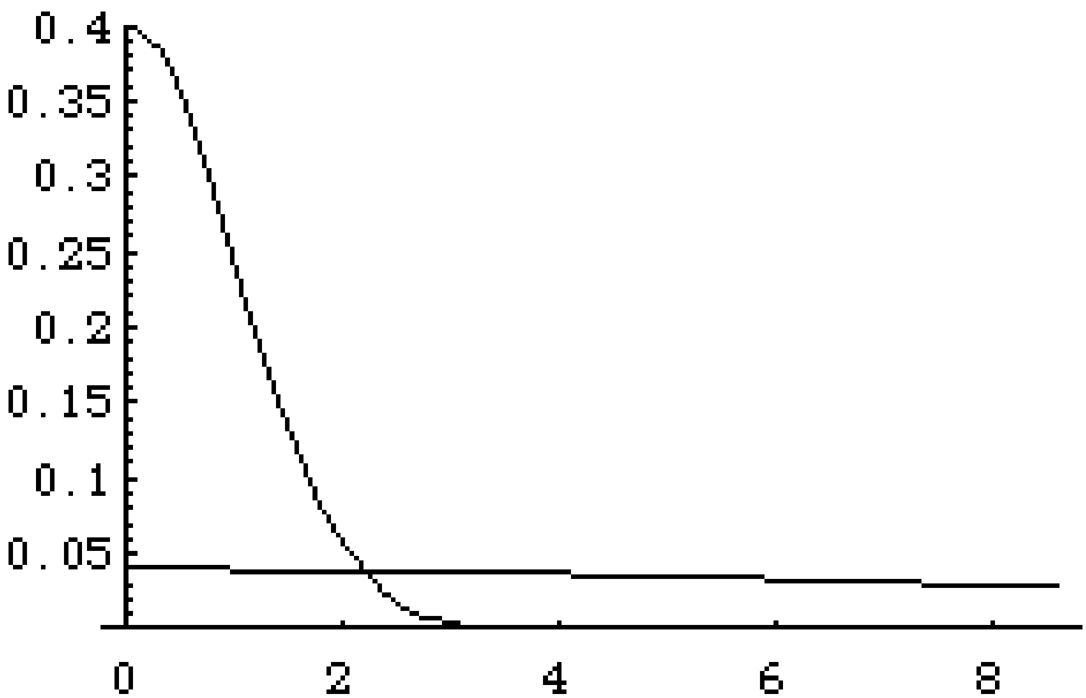
Scale

Think on scale of z-score, so in effect the Bayes factor

$$\frac{p(H_0|Y)}{p(H_1|Y)} = \frac{N(0, \sigma^2/n)}{N(0, \sigma^2(1 + 1/n))}$$

is $\approx$ ratio of $N(0, 1)$ to $N(0, n)$.

Spike and slab With $\sigma^2 = 1$ and $n = 100$



Thus, density of $\overline{Y}$ under H_1 is incredibly diffuse relative to H_0. This leads to the notion of a "spike and slab" prior when ideas are applied in estimation.

Discussion of BIC

Bayes factor

- Posterior odds to compare one hypothesis to another.
- Requires a likelihood and various approximations.

Comparison to other criteria

	Penalty	Test Level $\alpha(n)$	z-statistic
BIC	$\frac{1}{2}\log n$	Decreasing in n	$\lvert z\rvert > \sqrt{\log n}$
AIC	1	Fixed	$\lvert z\rvert > \sqrt{2}$
RIC	$\log p$	Decreasing in p	$\lvert z\rvert > \sqrt{2\log p}$

$\Rightarrow$ BIC tends to pick very parsimonious models.

Consistency

Strong penalty leads to consistency for fixed θ as $n \to \infty$.

Important?

e.g., Is the time series $AR(p)$ for any *fixed* order p?

Spike and slab prior for testimator

For p models, $M_1, \ldots, M_p$, with equal priors $p(M_j) = 1/p$

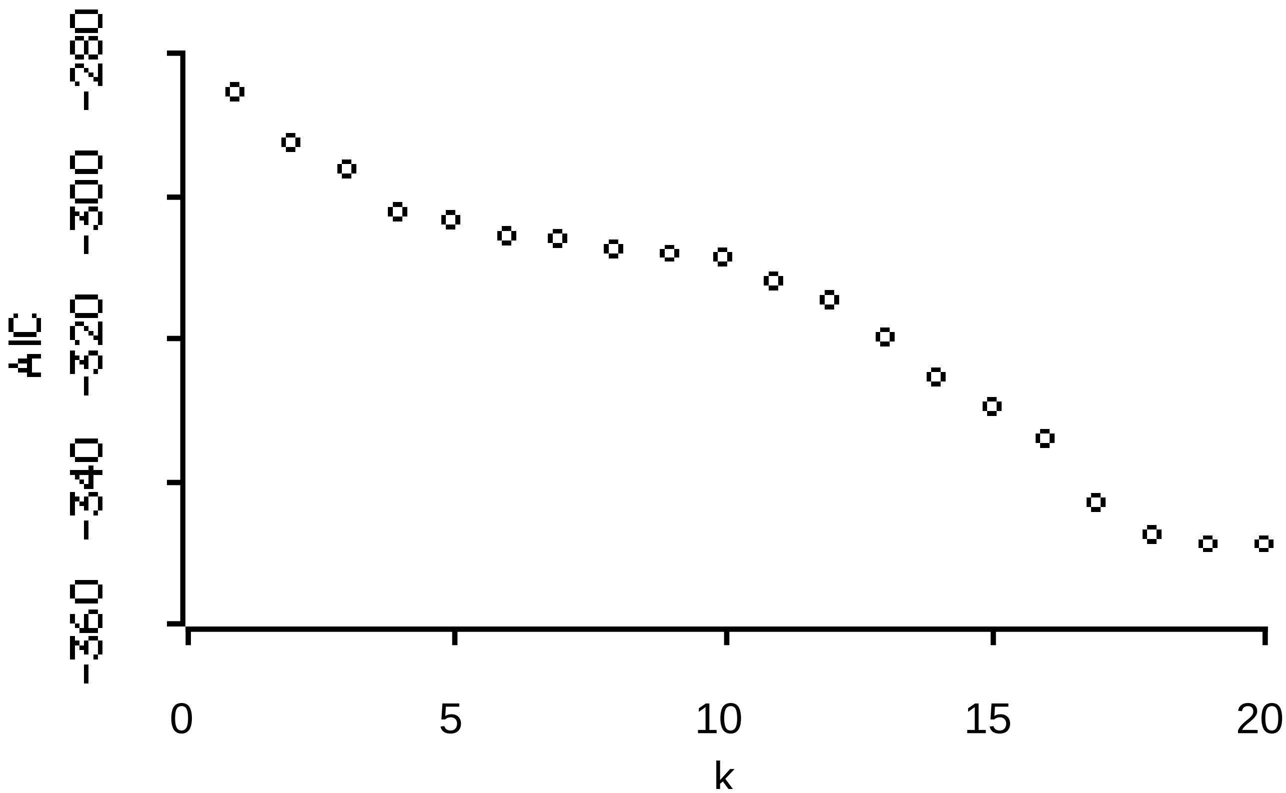
indexed by $\theta = (\theta_1, \ldots, \theta_p)$,

Bayes factors implies equal probability for 2^p models with prior

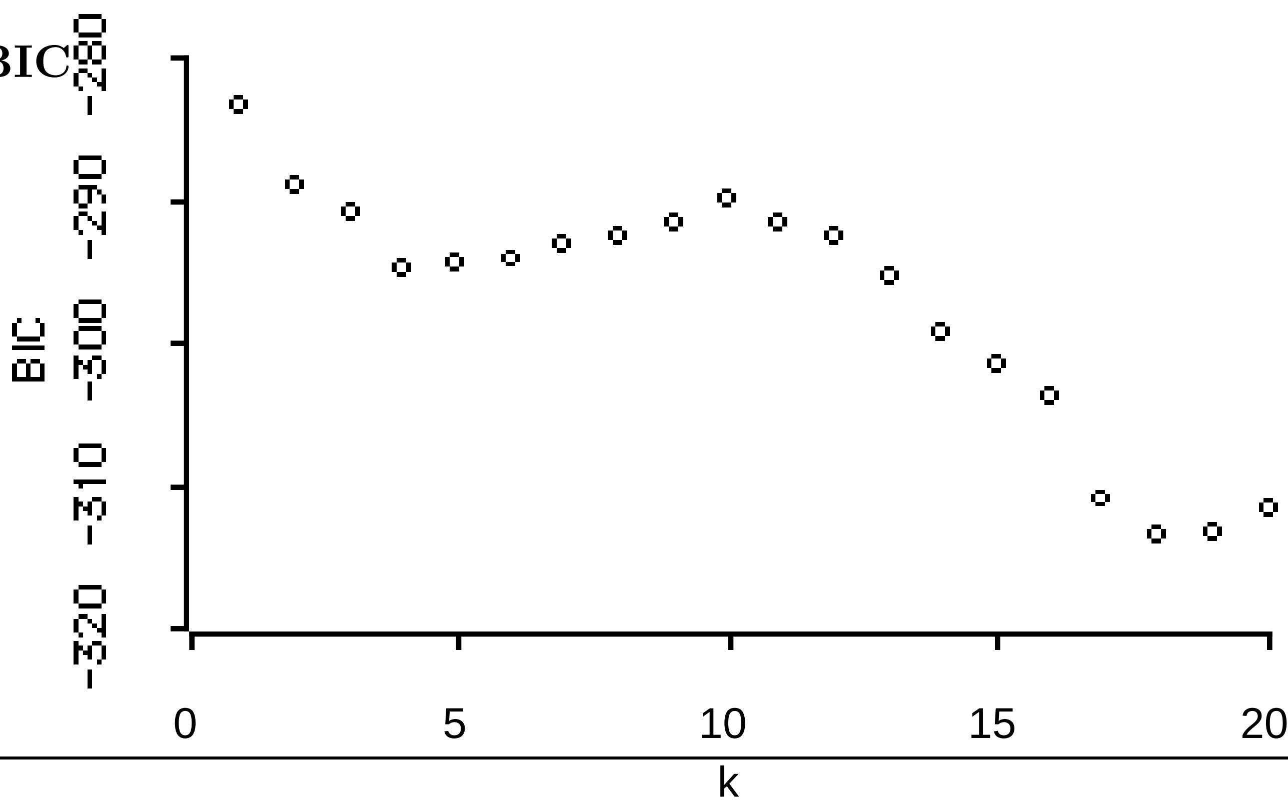
$$p(\theta_j) = \begin{cases} 1/2, & \theta_j = 0 \\ c, & \theta_j \neq 0 \end{cases}$$

Criteria in McDonald's Example

AIC



BIC



Detour to Random Effects Model

Model Simplified version of variable selection problem.

$$Y_j = \theta_j + \epsilon_j, \quad \epsilon_j \sim N(0, \sigma^2)\ , \quad j = 1, \ldots, p$$

or

$$Y_j | \theta_j \sim N(\theta_j,\ \sigma^2) \quad , \quad \theta_j \sim N(0,\ \nu^2 = c^2\ \sigma^2)$$

Decomposition of the marginal variance

$$\text{Var}\ Y = \sigma^2 + \nu^2 = \sigma^2(1 + c^2)$$

so that c^2 measures "signal strength".

Examples

- Repeated measures ($Y_{ij},\ i = 1, \ldots, n_j,\ Y_j \Rightarrow \overline{Y}_j$)
- Center effects in a clinical trial
- Growth curves (longitudinal models)

Relationship to regression

- Parameters: $\theta_j \iff X_j\, \beta_j$
- Residual SS: Sum of squares not fit, $\sum_{\gamma_j = 0} Y_j^2$

MLE Fits everything, so

$$\hat\theta_j = Y_j \quad \Rightarrow \quad \textstyle\sum_j E(\hat\theta_j - \theta_j)^2 = p\sigma^2$$

Bayes Estimator for Random Effects

Model

$$\begin{aligned} Y_j|\theta_j &\sim N(\theta_j,\ \sigma^2) \qquad j = 1, \ldots, p \\ \theta_j &\sim N(0,\ \nu^2 = c^2\,\sigma^2) \end{aligned}$$

Posterior mean (given normality, c^2)

$$E(\theta_j|Y) = Y_j \times \left(\frac{1/\sigma^2}{1/\sigma^2 + 1/\nu^2} = \frac{c^2}{1+c^2}\right)$$

Risk of Bayes estimator

$$\begin{aligned} \sum_j E(\theta_j - E[\theta_j|Y])^2 &= \frac{p}{1/\sigma^2 + 1/\nu^2} \\ &= \left(\frac{c^2}{1+c^2}\right) p\sigma^2 \\ &< p\sigma^2 = \text{Risk of MLE} \end{aligned}$$

Relative risk

$$\frac{E(\theta - \hat{\theta})^2}{E(\theta_j - E[\theta_j|Y])^2} = 1 + \frac{1}{c^2}$$

If $c^2 \approx 0$, Bayes estimator does much better... but what's c^2?

Approaches

Could put a prior on c^2, or try to estimate from data...

Shrink to Zero?

Model and estimator

$$Y_j|\theta_j \sim N(\theta_j, \sigma^2) \qquad \theta_j \sim N(0,\ c^2\,\sigma^2)$$

$$E(\theta_j|Y) = \left(1 - \frac{1}{1+c^2}\right)\ Y_j$$

Example

$$c^2 = 1 \quad \Rightarrow \quad E(\theta_j|Y) = \frac{Y_j}{2}$$

How to shrink all the way to zero?

Revise model and use a mixture prior,

$$\begin{aligned} Y_j|\theta_j &\sim N(\theta_j,\ \sigma^2) \\ \theta_j &\sim \pi\, N(0,\ c^2\,\sigma^2) + (1-\pi)\ \mathbf{1}_0\ , \end{aligned}$$

Where π denotes the probability of non-zero θ_j.

Bayes estimator?

Posterior mean $E(\theta_j|Y_j) =?$, but surely not zero.

Prior estimates?

Need π and c^2.

Alternative Strategy

Approach

Avoid direct evaluation of posterior mean $E[\theta_j|Y_j]$.

Indicator variables

Introduce indicators as used in regression

$$\gamma = (\gamma_1, \ldots, \gamma_p) \ , \quad \gamma_j \in \{0, 1\} \ ,$$

and write the revised model as

$$\begin{aligned} Y_j|\theta_j &\sim N(\theta_j,\ \sigma^2) \\ \gamma_j &\sim \text{Bernoulli}(\pi) \end{aligned}$$

where

$$\theta_j \sim \begin{cases} N(0,\ c^2\,\sigma^2) & \gamma_j = 1 \\ = 0 & \gamma_j = 0 \end{cases}$$

Three-step process George & Foster, 1996

1. Estimate prior parameters c^2 and π.
2. Maximize posterior over γ rather than find posterior mean, necessitating the next step.
3. Shrink estimates identified by γ.

Calibration of Selection Criteria

Revised model with mixture prior

$$Y_j|\theta_j \sim N(\theta_j,\ \sigma^2)\ , \quad \theta_j \sim \underbrace{\pi\ N(0,\ c^2\,\sigma^2)}_{\gamma_j=1} + \underbrace{(1-\pi)\ \mathbf{1_0}}_{\gamma_j=0}$$

Calibration idea

- How do π and c^2 affect choice of $\gamma_j = 1$?
- Result in a penalized likelihood, with previous penalties?

Posterior for γ Given c^2, σ^2, and π,

$$\begin{aligned}
p(\gamma|Y) &\propto p(\gamma)p(Y|\gamma) \\
&\propto \pi^q(1-\pi)^{p-q}\,\frac{1}{\sigma^p}\left(\frac{1}{1+c^2}\right)^{q/2} \\
&\quad \exp -\tfrac{1}{2}\left(\frac{\sum_{j=1}^q Y_j^2}{(1+c^2)\sigma^2} + \frac{\sum_{j=q+1}^p Y_j^2}{\sigma^2}\right) \\
&\propto \left(\frac{\pi}{1-\pi}\right)^q \left(\frac{1}{1+c^2}\right)^{q/2} \exp\left(\frac{-RSS}{2\sigma^2}\frac{c^2}{1+c^2}\right) \\
&\propto \exp\left[\frac{c^2}{1+c^2}\left(\frac{RegrSS}{2\sigma^2} - qR(\pi,c^2)\right)\right]
\end{aligned}$$

where the penalty for each parameter to the log-likelihood is

$$R(\pi,c^2) = \frac{1+c^2}{c^2}\left(\log\frac{1-\pi}{\pi} + \tfrac{1}{2}\log(1+c^2)\right)$$

Calibration of Selection Criteria

Revised model with mixture prior

$$Y_j|\theta_j \sim N(\theta_j,\ \sigma^2)\ , \quad \theta_j \sim \pi\ N(0,\ c^2\,\sigma^2) + (1-\pi)\ \mathbf{1_0}$$

Penalty

$$R(\pi, c^2) = \frac{1+c^2}{c^2}\left(\log\frac{1-\pi}{\pi} + \tfrac{1}{2}\log(1+c^2)\right)$$

Matching terms

Maximizing the posterior gives:

π	$\approx c^2$	Criterion	Comments
1/2	3.92	AIC	Lots of small coefs
1/2	n	BIC	
1/2	p^2	RIC	Few large coefs

But is the value $\pi = \frac{1}{2}$ reasonable?

Insight

$$\text{Each criterion} \quad \Longleftrightarrow \quad \begin{cases} \text{prob on number } \theta_j \neq 0 \\ \text{prior for nonzero coefficients} \end{cases}$$

Empirical Bayes for Random Effects

Model

$$Y_j|\theta_j \sim N(\theta_j, \sigma^2) \qquad \theta_j \sim N(0,\ c^2\ \sigma^2)$$

Bayes estimator

$$E(\theta_j|Y) = \frac{c^2}{1+c^2}\ Y_j = \left(1 - \frac{1}{1+c^2}\right)\ Y_j$$

Idea

Assume you know σ^2 or have a good estimate.

Then estimate prior from marginal distribution

$$Y_j \sim N(0, (1+c^2)\sigma^2)\ .$$

So define

$$\hat{c}^2 = \frac{s^2}{\sigma^2} - 1 \quad \text{where} \quad s^2 = \frac{\sum Y_j^2}{p}$$

Almost James-Stein

Plug in estimator is

$$\widehat{E}(\theta_j|Y) = \left(1 - \frac{1}{1+\hat{c}^2}\right)\ Y_j = \left(1 - \frac{\sigma^2}{s^2}\right)\ Y_j$$

James-Stein adjusts for fact that $E1/\hat{c}^2 \neq 1/E\hat{c}^2$.

Empirical Bayes Criterion

Model with mixture prior

$$Y_j|\theta_j \sim N(\theta_j,\, \sigma^2)\ , \quad \theta_j \sim \pi\, N(0,\, c^2\,\sigma^2) + (1-\pi)\, \mathbf{1}_0$$

Step 1

Estimate for prior parameters via MLE for approximate likelihood (no avg)

$$L^*(c^2, \pi) = \pi^q (1-\pi)^{p-q} (1+c^2)^{q/2} \exp(c^2 RegrSS/(2\sigma^2(1+c^2))$$

$$\widehat{c^2}_\gamma = \left(\frac{\sum_{j=1}^q Y_j^2}{q\sigma^2} - 1 \right)_+ \qquad \hat{\pi}_\gamma = q/p$$

Clearly need large number of parameters. (large p)

Step 2

Iteratively identify nonzero θ_j as those maximizing posterior for most recent set of estimates $\widehat{c^2}$ and $\hat{\pi}$

$$\max_\gamma p(\gamma \mid Y) \propto p(Y \mid \gamma)\, p(\gamma)$$

Step 3

Shrink nonzero θ_j to compensate for maximization.

Simulations show that this step is essential.

Properties of EBC

Maximized likelihood George & Foster 1996

Maximimum value of the approximate likelihood L^* is

$$\frac{RegrSS}{\sigma^2} - q\left(1 + \log\frac{RegrSS}{q\sigma^2}\right) - 2pH(q/p)$$

where the Boolean entropy function is

$$H(\pi) = -\pi\log\pi - (1-\pi)\log(1-\pi) \geq 0 \ .$$

Adaptive penalty

Penalty depends on $\hat{\pi}$ and estimated strength of signal in $\widehat{c^2}$.

Features

- Large $RegrSS$ relative to $q \Rightarrow BIC$ type behavior.
- Small $\hat{\pi} = 1/p \Rightarrow RIC$ behavior since

$$pH(1/p) \approx \log p$$

Simulation of Predictive Risk

Conditions Normal with random effects setup $X = I_n$, $n = p = 1000$. Full least squares scores 1000 in each case.

Weak signal $c^2 = 2$

q	0	10	50	100	500	1000
C_p	572	577	599	627	845	1118
BIC	74.8	87.8	146	217	781	1488
RIC	2.9	21.0	93.3	183	899	1799
EBC	503	611	787	902	1000	999
EBC_δ	20.3	61.7	474	872	1000	999
$\hat{\beta}$	1.1	19.9	91.8	168	501	667

Strong signal $c^2 = 100$

q	0	10	50	100	500	1000
C_p	572	578	597	623	824	1072
BIC	75.4	88.7	143	213	771	1460
RIC	3.3	26.0	116	229	1142	2274
EBC	496	26.5	106	194	737	999
EBC_δ	15.9	26.5	106	194	737	999
$\hat{\beta}$	1.1	26.4	106	193	731	990

Discussion

Calibration

Model selection criteria correspond to priors on θ_i, with typically many coefficients (p large):

AIC Prior variance **four** times $\sigma^2 \Rightarrow$ Lots of little θ_i

RIC Prior variance $\mathbf{p^2}$ times $\sigma^2 \Rightarrow$ Fewer, very large θ_i

Adaptive selection

- Tune selection criterion to problem at hand.
- With shrinkage, does better than OLS with any fixed criterion.

Weakness of normal prior

- Cannot handle mixture of large and small nonzero θ_i
- Some effects are very significant (eg: age and osteoporosis)
- Others are more speculative (eg: hormonal factors)
- Mixing these "confuses" normal prior $\theta_i \sim N(0, \nu^2)$

Cauchy prior

- Tolerates mixture of large and small θ_j (Jeffreys, 1961)
- Not conjugate
- Useful in information theoretic context

Empirical Bayes Estimators, Method 2

Full Bayes model

$$
\begin{aligned}
Y_j|\theta_j &\sim N(\theta_j,\ \sigma^2) \\
\theta_j|\gamma &\sim \begin{cases} N(0, \tau^2 = c^2\,\sigma^2) & \gamma_j = 1 \\ 0 & \text{otherwise} \end{cases} \\
\gamma &\sim \pi^q (1-\pi)^{p-q}, \quad q = |\gamma| \\
\pi &\sim U[0,1]
\end{aligned}
$$

Conjugate priors lead to all the needed conditional distributions...

$$
\begin{aligned}
(\pi \mid \gamma,\ ,\theta,\ \nu^2,\ \sigma^2,\ Y) &\sim (\pi \mid \gamma) \sim \text{Beta}(q+1, p-q+1) \\
\nu^2 \mid \text{others} &\sim \text{Inverted gamma}_1 \\
\sigma^2 \mid \text{others} &\sim \text{Inverted gamma}_2 \\
(\theta_j, \gamma_j) \mid \text{others} &\sim (\text{Bernoulli}, \text{Normal})
\end{aligned}
$$

Bayes probability

Bernoulli probability via likelihood ratio

$$
P\{\gamma_j = 1\} = \frac{\pi\ N(0, \sigma^2 + \nu^2)[Y_j]}{\pi\ N(0, \sigma^2 + \nu^2)[Y_j] + (1-\pi)\ N(0, \sigma^2)[Y_j]}
$$

Gibbs

Gibbs sampler avoids maximizing likelihood, but much slooooooowwwwweeeeerrrrr.

Calibration in Regression

Goal

Think about various methods in a common setting, comparing more deeply than just via a threshold.

Comparison

Method	Threshold	Prior for p	Prior for β
Bayes factors (BIC)	$\sqrt{\log n}$	$\mathrm{Bi}(\frac{1}{2}, p)$	Spike-and-slab
Predictive risk (AIC)	$\sqrt{2}$	?	?
Minimax risk (RIC)	$\sqrt{2\log p}$	?	?

Bayesian variable selection George & Foster 1996

Put a prior on everything, particularly ($\gamma_j = 1$ implies fit β_j)

$$p(\gamma), \quad p(\beta_\gamma|\gamma) \quad \Rightarrow \quad p(\gamma|Y)$$

In particular, consider beta prior for γ given some constant w

$$p(\gamma|w) \propto w^q(1-w)^{p-q} \ , \quad q = |\gamma| \ ,$$

and normal for β given γ, obs. variance σ^2, and a constant c^2:

$$p(\beta|\gamma, c, \sigma^2) = N(0, \ c^2\sigma^2(X'_\gamma X_\gamma)^{-1}) \ .$$

Bayesian Calibration Results

Posterior for γ

$p(\gamma|Y, \sigma, c, w)$ is increasing as a function of

$$RegrSS_\gamma/\sigma^2 - F(c^2, w)\, |\gamma|$$

where

$$F(c^2, w) = \frac{1+c^2}{c^2}\left(\log(1+c^2) + 2\log\frac{1-w}{w}\right)$$

Match constants to known penalties

Since *AIC*, *BIC*, and *RIC* are selecting model based on penalized likelihood (penalty factors 2, $\frac{1}{2}\log n$, and $\log p$) calibrate by solving

$$F(c^2, w) = 2\ \times\ \text{penalty factor}$$

Which solution?

Pick $w = \frac{1}{2}$ and solve for c^2?

If p is very large (or perhaps very small, do you expect half of the coefficients to enter the model?

Next steps

- Put another prior on c^2 and w?
- Choose c^2 and w to maximize the likelihood $L(c^2, w|Y)$.

Empirical Bayes Criterion

Criterion

Pick the subset of p possible predictors γ which maximizes the penalized likelihood (approximately)

$$\frac{SS_\gamma}{\hat\sigma^2} - q \underbrace{\left(1 + \log \frac{SS_\gamma}{q\hat\sigma^2}\right)}_{\text{BIC}} + \underbrace{2\left((p-q)\log(p-q) + q\log q\right)}_{\text{RIC}}$$

where $q = |\gamma|$ chosen predictors and SS_γ denotes the *regression* sum-of-squares for this collection of predictors.

Empirical Bayes estimates

$$\hat c_\gamma = \left(\frac{SS_\gamma}{\hat\sigma^2 q} - 1\right)_+, \quad \hat w_\gamma = q/p$$

BIC component

Typically $tr(X'X) = O(n)$ so that for fixed γ,

$$1 + \log \frac{SS_\gamma}{q\hat\sigma^2} = O(\log n)$$

RIC component

With $q = 1$, reduces to $\approx 2\log p$.

Key property

Threshold varies with the complexity of model: Once q gets large, decreasing threshold allows other variables into model.