Problem statement

- The data: Orthonormal regression with lots of X's (possible lots of β's are zero:

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \sigma Z_i, \qquad Z_i \sim N(0,1) \ ,$$

- Equivalent form: Normal mean problem (known σ)

$$Y_i = \mu_i + Z_i, \qquad Z_i \sim N(0,1) \ ,$$

- Unimodal prior for μ: $\pi \in \mathcal{M}$ iff
  - $\pi(\mu)$ is a symmetric
  - $|\mu| \leq |\mu'|$ implies $\pi(\mu) \geq \pi(\mu')$.

- Risk function: Kullback-Liebler divergence.

$$\mathcal{R}_n(\vec{\mu}, \pi) = \int \log \frac{P_{\vec{\mu}}(Y|X)}{P_\pi(Y|X)} P_{\vec{\mu}}(Y|X) dY \ .$$

- Problem: Find a universal π.

Risk lower bounds

Theorem 1 *For all n, for all $\vec{\mu}$, and $\pi \in \mathcal{M}$,*

$$\mathcal{R}_n(\vec{\mu}, \pi) \geq c \sum_i \min\left(\mu_i^2 + \epsilon(\pi), \frac{1}{\mu_i^2} + \log \frac{\mu_i}{\epsilon(\pi)}\right)$$

But, how is $\epsilon(\pi)$ defined?

- Marginal distribution of Y_i:

$$\phi_\pi(y) = \int \phi(y - \mu)\pi(\mu) d\mu \; .$$

- $\tau(\pi)$ says when ϕ_π's tail gets fat relative to a nromal tail:

$$\tau(\pi) = \inf_\tau \{\tau : \frac{\int_\tau^\infty \phi_\pi(y)dy}{\int_\tau^\infty \phi(y)dy} > 7.38... = e^2\} \; .$$

- $\epsilon(\pi)$ measures how big this fat tail is:

$$\epsilon(\pi) = \int_{\tau(\pi)}^\infty \phi_\pi(y)dy \; .$$

Knowing $\epsilon(\pi)$ is as good as knowing π

- Goal: find a single prior that can do almost as well as any unimodal prior with a fixed value of $\epsilon(\pi)$

- Spike and slab (Cauchy slab)

$$\widehat{\pi}_\epsilon(\mu) = (1-\epsilon) \text{ Spike} + \epsilon \text{ Cauchy}$$

Theorem 2 *For all n, for all $\vec{\mu}$, and $\epsilon \leq .5$,*

$$\mathcal{R}_n(\vec{\mu}, \widehat{\pi}_\epsilon) \leq 2 \sum_i \min\left(\mu_i^2 + \epsilon, \frac{1}{\mu_i^2} + \log\frac{\mu_i}{\epsilon}\right)$$

Note: Same shape as lower bound. So it is only off by a constant factor.

Suppose $p = 1$.
Our risk compared to the lower bound.

Figure 1: Risk of the Cauchy mixture $\hat{\pi}_{0.001}$ and the lower bound for the divergence attainable by any Bayes prior with $\epsilon(\pi) = 0.001$.

Figures 2 and 3: The ratio is bounded by 6 in these examples for $\epsilon = 0.01$ (left) and $\epsilon = .00001$ (right).

Empirical Bayes: Doing without ϵ

- Put prior on ϵ: $\epsilon \sim \text{Beta}(0, p)$
  - strongly biased towards "null" model
  - Puts most of the weight near $\epsilon = 0$
  - $P(\epsilon < 1/p) > .5$
  - Induces an exchangable prior over μ. call it $\tilde{\pi}$.

-

Theorem 3

$$\mathcal{R}_n(\vec{\mu}, \tilde{\pi}) \leq \omega_0 + \omega_1 \inf_{\pi \in \mathcal{M}} \mathcal{R}_n(\vec{\mu}, \pi)$$

Key point: $\tilde{\pi}$ has "almost" as good a risk as the best unimodal prior.

Do there exist other procedures that have ω_0 and ω_1 both constant?

Normal is bad: A spike and normal slab has unbounded ω_1 (even if calibration is used like in George and Foster).

Tradition rules are bad: AIC / BIC / C_p have unbounded ω_1.

Risk inflation is better: The best a testimator can achieve is $\omega_1 = O(\log p)$. (Donoho and Johnstone / Foster and George).

Jefferies is competitive: If $\epsilon \sim$ Beta(.5,.5) then ω_1 is constant, *but* $\omega_0 = O(\log p)$. So still not linear.

Adaptive rules work: Some adaptive procedures might work (nothing has been proven though):

- Simes-like methods (Benjamini and Hochberg)
- estimated degrees of freedom (Ye)
- Empirical Bayes? (Zhang)

Everyone likes a good forecast.

- If you don't like the risk perspective, how about a forecasting perspective?
- Dawid's prequential approach
- Predict successive observations
- Use so-called "log-loss"
  - decision-maker gives a forecast of $P(\cdot)$
  - Y is observed
  - Loss = $\log \frac{1}{P(Y)}$

$$\text{our total loss} = \underbrace{\text{intrinsic loss}}_{\mu \text{ known}} + O(\text{best Bayes excess})$$

$$\sum_{i=1}^{n} \log \frac{1}{P_{\widehat{\pi}}^{i-1}(Y_i)} = \underbrace{\sum_{i=1}^{n} \log \frac{1}{P_{\mu}^{i-1}(Y_i)}}_{O(n)} + O_p(\underbrace{\inf_{\pi \in \mathcal{M}} \mathcal{R}_n(\vec{\mu}, \pi)}_{O(\log n)})$$

Take home messages

- Don't worry about eliciting the shape of a IID prior for variable selection. It can be done well enough by automatic methods so the effort isn't justified.

- Bias your priors toward *not* including variables.
  - "Pretend" you have seen p insignificant variables before you start.
  - Make sure about 1/2 of your probability is on the "no signal" model.

- Cauchy priors are cool!

Adaptive Variable Selection with Bayesian Oracles

Dean Foster & Bob Stine
Department of Statistics, The Wharton School
University of Pennsylvania, Philadelphia PA
diskworld.wharton.upenn.edu

November 3, 2002

Adaptive Variable Selection with Bayesian Oracles

Dean Foster & Bob Stine
Department of Statistics, The Wharton School
University of Pennsylvania, Philadelphia PA
diskworld.wharton.upenn.edu

November 3, 2002

Abstract

We analyze the performance of adaptive variable selection with the aid of a Bayesian oracle. A Bayesian oracle supplies the statistician with a distribution for the unknown model parameters, here the coefficients in an orthonormal regression. We derive lower bounds for the predictive risk of regression models constructed with the aid of a class of Bayesian oracles, those that are unimodal and symmetric about zero. These bounds are not asymptotic and obtain for all sample sizes and model parameters. We then construct a model whose predictive risk is bounded by a linear function of the risk obtained by the best Bayesian oracle. The procedure that achieves this performance is related to an empirical Bayes estimator and those derived from step-up/step-down testing.

Abstract

We analyze the performance of adaptive variable selection with the aid of a Bayesian oracle. A Bayesian oracle supplies the statistician with a distribution for the unknown model parameters, here the coefficients in an orthonormal regression. We derive lower bounds for the predictive risk of regression models constructed with the aid of a class of Bayesian oracles, those that are unimodal and symmetric about zero. These bounds are not asymptotic and obtain for all sample sizes and model parameters. We then construct a model whose predictive risk is bounded by a linear function of the risk obtained by the best Bayesian oracle. The procedure that achieves this performance is related to an empirical Bayes estimator and those derived from step-up/step-down testing.