Bayesian Model Selection and Model Averaging

Yingbo Li

Clemson University

MATH 9810

Normal Linear Regression

$$\mathbf{y} = \beta_0 + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \epsilon_i \overset{\text{iid}}{\sim} \mathsf{N}\left(0, \sigma^2\right), \quad i = 1, \ldots, n$$

- Observation $\mathbf{y}$: n-dim
- Predictors $\mathbf{X}_1, \ldots, \mathbf{X}_p$: centered and scaled
- OLS: smallest variance among all unbiased estimators.
- Bias-variance trade-off: biased estimators with smaller variance $\longrightarrow$ improve overall prediction and estimation accuracy

$$E(\tilde{\beta} - \beta)^2 = Var(\tilde{\beta}) + [E(\tilde{\beta}) - \beta]^2$$

- Variable selection:
  - Avoid the use of redundant variables (problems with interpretations)
  - Occam's Razor
  - Inclusion of un-necessary terms yields less precise estimates, particularly if explanatory variables are highly correlated with each other

Classical Variable Selection Procedures

- Stepwise Regression: Forward, Stepwise, Backward add/delete variables until all t-statistics are significant (easy, but not recommended)
- Use a Model Selection Criterion to pick the best model
  - Adjusted R^2 (since R^2 picks largest model)
  - Mallow's C_p
  - Akaike Information Criterion (AIC): the smaller the better

$$AIC = -2 \log f(\hat{\theta}) + 2p$$

  - Bayesian Information Criterion (BIC, or Schwarz criterion): the smaller the better

$$BIC = -2 \log f(\hat{\theta}) + p \log(n)$$

- Trade off between model complexity p with goodness of fit $f(\hat{\theta})$.
- Between AIC and BIC: AIC tends to prefer larger models, while BIC tends to prefer smaller models.

Penalization Methods

Estimates are obtained by minimizing SSE plus a penalty function

$$(\hat{\beta}_0, \hat{\boldsymbol{\beta}})_\lambda = \min_{\beta_0, \boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \mathbf{x}_i' \boldsymbol{\beta})^2 + f_\lambda(\boldsymbol{\beta}) \right\}$$

- Ridge regression: $f_\lambda(\beta) = \lambda \sum_{j=1}^{p} \beta_j^2$.
- All columns of the full design matrix $\mathbf{X}_j$ should be centered as rescaled to one (so that β_j's are on the same scale).
- The tuning parameter $\lambda > 0$ is usually chosen via cross-validation (maximize the average out of sample prediction accuracy)
- Bayesian counterparts: **posterior mode** under the prior with density

$$p(\boldsymbol{\beta} \mid \lambda) = \exp\{-\frac{f_\lambda(\boldsymbol{\beta})}{2\sigma^2}\}$$

Ridge regression: L_2 penalty

$$f_\lambda(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} \beta_j^2 \Longleftrightarrow \beta_j \mid \lambda, \sigma^2 \overset{\text{iid}}{\sim} \mathsf{N}\left(0, \frac{\sigma^2}{\lambda}\right), \quad j = 1, 2, \ldots, p$$

- Ridge regression solution is available in closed form (suppose Y is centered so $\hat{\beta}_0 = 0$),

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

- For orthogonal design $\mathbf{X}^T\mathbf{X} = \mathbf{I}$, then $\hat{\beta}_j^{\text{ridge}} = \frac{1}{1+\lambda}\hat{\beta}_j^{\text{mle}}$.
- Stabilize the estimate when columns of $\mathbf{X}$ are highly-correlated.
- Can shrink coefficients towards zero, but not exactly to zero.

- Bayesian counterpart of ridge regression: independent normal prior
- Fully Bayes estimate of λ, we can let it have a prior distribution, e.g., $\lambda \sim G(1, 1)$, and estimate it via its posterior distribution (using MCMC method to draw posterior samples).

Lasso: L_1 penalty

$$f_\lambda(\beta) = \lambda \sum_{j=1}^{p} |\beta_j| \Longleftrightarrow p(\beta_j \mid \sigma^2, \lambda) = \frac{\lambda}{2\sigma} e^{-\frac{\lambda}{\sigma}|\beta_j|}$$

- Lasso solution is not available in closed form, but has fast algorithm.
- Can shrink coefficients exactly to zero.

Bayesian Lasso [Park and Casella 2008]

- Independent double exponential distribution (i.e., Laplace distr.).
- Hierarchical representation:

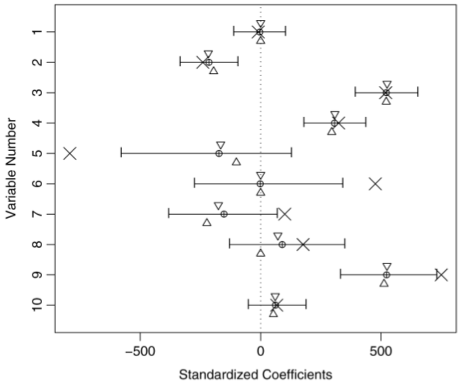
$$\beta_j \mid \sigma^2, \tau_j^2 \overset{\text{ind}}{\sim} \mathsf{N}(0, \sigma^2\tau_j^2), \quad \tau_j^2 \mid \lambda \overset{\text{iid}}{\sim} \mathsf{Exp}\left(\lambda^2/2\right)$$

- Non-informative prior $p(\sigma^2) = 1/\sigma^2$, or inverse Gamma prior on σ^2.
- Two options to choose the parameter λ:
  - Empirical Bayes (maximize marginal likelihood): using Monte Carlo EM
  - Gamma prior on λ^2 (enables Gibbs updating).
    The improper prior $p(\lambda^2) = 1/\lambda^2$ leads to improper posterior.

Bayesian Lasso

Closed form full conditionals

- $p(\beta_j \mid \tau_j, \sigma^2, \lambda^2)$: normal
- $p(\tau_j \mid \beta_j, \sigma^2, , \lambda^2)$: inverse-Gaussian
- $p(\sigma^2 \mid \beta_j, \tau_j, \lambda^2)$: inverse Gamma
- $p(\lambda^2 \mid \beta_j, \tau_j, \sigma^2)$: Gamma



- ⊕ posterior median of Bayesian lasso, and its 95% equal-tail CI
- × mle
- △ lasso estimate based on n-fold CV
- ▽ lasso estimate to match L_1 norm of Bayesian lasso

Examples of Independent Shrinkage Priors

Suppose β_j has independent normal priors

$$\beta_i \mid \omega_j, \sigma^2 \sim \mathsf{N}(0, \omega_j \sigma^2),$$

Different hyper priors on ω_i's lead to different (marginal) prior distributions on β_j's.

- Student's t (df$= v$, scale$= \eta$): $\omega_j \sim IG(v/2, v\eta^2/2)$
- Cauchy: $\omega_j \sim IG(1/2, \eta^2/2)$
- Bayesian lasso: $\omega_j \sim Exp(\eta^2/2)$
- Normal-Gamma prior: $\omega_j \sim Gamma$
- Horseshoe: $\sqrt{\omega_j} \sim$ half Cauchy

Good shrinkage priors should has

- heavy tails (than normal), to avoid over-shrinking large coefficients
- large probability mass around zero: shrink small coefficients to be very close to zero

Spike and Slab Prior

Pure shrinkage priors cannot select variables, since the estimates of all β_j's (posterior mean, median or mode) are usually non-zero.

Spike-and-slab prior: the prior distribution of β_j is a mixture of a point mass at zero (spike) and a continuous distribution centered at zero (slab):

$$p(\beta_j \mid \rho) \overset{\text{iid}}{=} (1-\rho)\delta_0(\beta_j) + \rho f(\beta_j)$$

- $\beta_j = 0$ with probability $1 - \rho$.
- $\beta_j \sim f(\cdot)$ with probability ρ.
  - In the original spike-and-slab prior paper [Mitchell and Beauchamp, 1988], $f(\cdot)$ is Uniform$(-c, c)$.
  - Most popular choice of $f(\cdot)$: normal with mean 0.
  - Or $f(\cdot)$ can be any shrinkage prior we've mentioned before.

Parameter ρ: prior marginal inclusion probability $p(\beta_j \neq 0)$

- Usually, results are sensitive to the choice of ρ.
- Hyperprior $\rho \sim$ Beta(a, b).

Posterior Distribution under the Spike and Slab Prior

Let indicator $\gamma_j = \delta(\beta_j \neq 0)$, then the marginal posterior distribution

$$p(\beta_j \mid \mathbf{y}) = P(\gamma_j = 0 \mid \mathbf{y})\delta_0(\beta_j) + P(\gamma_j = 1 \mid \mathbf{y})f(\beta_j \mid \mathbf{y})$$

- Posterior probabilities $P(\gamma_j = 0 \mid \mathbf{y})$ is updated
- So is $P(\gamma_j = 1 \mid \mathbf{y})$: posterior marginal inclusion probabilities
- $f(\beta_j \mid \mathbf{y})$ is the posterior density of the continuous component.
- The posterior distributions of $\beta_1, \ldots, \beta_p$ are not independent.

Point estimates of β_j

- Posterior mean $P(\gamma_j = 1 \mid \mathbf{y})E_f(\beta_j \mid \mathbf{y})$: non-zero
- Posterior median: can be zero if $P(\gamma_j = 0 \mid \mathbf{y}) \geq 0.5$

Bayesian Model Selection

- Models for the variable selection problem are based on a subset of the $\mathbf{X}_1, \ldots, \mathbf{X}_p$ variables.
- Encode models with a vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$ where $\gamma_j \in \{0, 1\}$ is an indicator for whether variable $\mathbf{X}_j$ should be included in the model $\mathcal{M}_{\boldsymbol{\gamma}}$.

$$\gamma_j = 0 \Longleftrightarrow \beta_j = 0$$

- Each value of $\boldsymbol{\gamma}$ represents one of the 2^p models.
- Under model $\mathcal{M}_{\boldsymbol{\gamma}}$:

$$\mathbf{y} \mid \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2, \boldsymbol{\gamma} \sim \mathsf{N}(\mathbf{X}_{\boldsymbol{\gamma}} \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2 \mathbf{I}_n)$$

where $\mathbf{X}_{\boldsymbol{\gamma}}$ is design matrix using the columns in $\mathbf{X}$ where $\gamma_j = 1$ and $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ is the subset of $\boldsymbol{\beta}$ that are non-zero.

Model Selection Criteria

If selecting a single best model is important (dependent on the type of application), then we can view this as a hypothesis testing problem with 2^p hypotheses; each hypothesis H_{γ} is a subset model γ.

Posterior probabilities under $H_{\boldsymbol{\gamma}}$ is

$$P(\boldsymbol{\gamma} \mid \mathbf{y}) = \frac{m(\mathbf{y} \mid \boldsymbol{\gamma})p(\boldsymbol{\gamma})}{\sum_{\boldsymbol{\gamma}'} m(\mathbf{y} \mid \boldsymbol{\gamma}')p(\boldsymbol{\gamma}')}$$

- Maximum a posteriori (MAP): $\hat{\boldsymbol{\gamma}}^{\mathsf{MAP}} = \arg\max_{\boldsymbol{\gamma}} P(\boldsymbol{\gamma} \mid \mathbf{y})$
  Sometimes there are several top models have very close posterior probabilities; hard to say which one is better.
- Median probability model: based on marginal inclusion probabilities $\hat{\boldsymbol{\gamma}}^{\mathsf{median}} = \{\gamma_j = 1 : P(\gamma_j \mid \mathbf{y}) \geq 0.5\}$

Bayesian Model Averaging (BMA)

Rather than use a single model, BMA uses all (or potentially a lot) models, but weights model predictions by their posterior probabilities (measure of how much each model is supported by the data)

$$E(\triangle \mid \mathbf{y}) = \sum_{\boldsymbol{\gamma}} E(\triangle \mid \mathbf{y}, \boldsymbol{\gamma}) P(\boldsymbol{\gamma} \mid \mathbf{y})$$

Examples of BMA estimates:

- Posterior marginal inclusion probabilities: $P(\beta_j \neq 0 \mid \mathbf{y}) = \sum_{\gamma_j = 1} \; P(\boldsymbol{\gamma} \mid \mathbf{y})$
- Posterior mean of coefficients: $\tilde{\beta}_j = \sum_{\boldsymbol{\gamma}} E(\beta_j \mid \mathbf{y}, \boldsymbol{\gamma}) \; P(\boldsymbol{\gamma} \mid \mathbf{y})$
- Posterior prediction: $\tilde{y}^* \mid \mathbf{y} = \sum_{\boldsymbol{\gamma}} \tilde{y}^*_{\boldsymbol{\gamma}} \; P(\boldsymbol{\gamma} \mid \mathbf{y})$

Stochastic Search Variable Selection (SSVS)

The potential model space of $\boldsymbol{\gamma}$ is very large: 2^p different subset models. George and McCulloch [1993, 1997] propose the SSVS method: suppose $p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}), p(\sigma^2)$ are conjugate, so that the marginal likelihood

$$m(\mathbf{y} \mid \boldsymbol{\gamma}) = \int f(\mathbf{y} \mid \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2, \boldsymbol{\gamma}) p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2) d(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2)$$

has closed form, then we can draw posterior samples of $\boldsymbol{\gamma}$ using MCMC. In each iteration, (1) first random select $j \sim \text{Unif}\{1, 2, \ldots, p\}$, (2) then

- Gibbs sampler: draw a new γ_j from Bernoulli distribution with success probability $P(\gamma_j = 1 \mid \boldsymbol{\gamma}_{(-j)}, \mathbf{y}) =$

$$\frac{\rho \cdot m(\mathbf{y} \mid (\gamma_j = 1, \boldsymbol{\gamma}_{(-j)}))}{\rho \cdot m(\mathbf{y} \mid (\gamma_j = 1, \boldsymbol{\gamma}_{(-j)})) + (1 - \rho) \cdot m(\mathbf{y} \mid (\gamma_j = 0, \boldsymbol{\gamma}_{(-j)}))}$$

- Alternatively, Metropolis-Hastings algorithm, propose $\boldsymbol{\gamma}^*$ s.t.,

$$\gamma_j^* = 1 - \gamma_j^{(s)}, \quad \boldsymbol{\gamma}_{(-j)}^* = \boldsymbol{\gamma}_{(-j)}^{(s)}$$

Prior Distributions on $\boldsymbol{\beta}_\gamma, \beta_0, \sigma^2$

When view the model selection problem as a hypothesis testing problem with 2^p hypothesis,

- Priors on σ^2, β_0 (common parameters in all hypotheses) can be improper, i.e., $p(\sigma^2) \propto 1/\sigma^2$ and $p(\beta_0) \propto 1$.
- Prior distributions $p(\boldsymbol{\beta}_\gamma)$ cannot be improper.
  When compare H_γ to the null model (only has intercept) H_{γ_0},

$$B_{\gamma,\gamma_0} = \frac{\int p(\mathbf{y} \mid \boldsymbol{\beta}_\gamma, \beta_0, \sigma^2)\ c_1\ p(\boldsymbol{\beta}_\gamma)\ c_2\ p(\beta_0, \sigma^2) d(\boldsymbol{\beta}_\gamma, \beta_0, \sigma^2)}{\int p(\mathbf{y} \mid \boldsymbol{\beta}_\gamma, \beta_0, \sigma^2)\ c_2\ p(\beta_0, \sigma^2) d(\beta_0, \sigma^2)}$$

- Vague but proper priors may also lead to paradoxes!
- Conjugate Normal-Gammas lead to closed form expressions for marginal likelihoods. Zellners g-prior is the most popular.

Zellner's g-prior

For a regression problem $\mathbf{y} \sim \mathsf{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$.
Zellner [1986] develops the g-prior via imaginary samples:

- Treat $\mathbf{X}$ as fixed (not random). Suppose before the real dataset $\mathbf{y}$ is collected, using expects knowledge we may have some idea about the values of the responses, denoted by $\mathbf{y}_0$, a n-dim vector.
- We are more uncertain about these imaginary samples, so let the residuals variance be $g\sigma^2$ instead of σ^2 (usually $g > 0$ exceeds 1), i.e,

$$\mathbf{y}_0 \sim \mathsf{N}(\mathbf{X}\boldsymbol{\beta}, g\sigma^2\mathbf{I}_n)$$

- g-prior is obtained as the posterior distribution $\boldsymbol{\beta} \mid \sigma^2$ under the Jeffreys prior $p(\boldsymbol{\beta}) \propto 1$ updated by the imaginary samples $(\mathbf{y}_0, \mathbf{X})$:

$$p(\boldsymbol{\beta} \mid \sigma^2) \propto p(\mathbf{y}_0 \mid \boldsymbol{\beta}, \sigma^2)p(\boldsymbol{\beta}) = N(\boldsymbol{\beta}_0, g\sigma^2(\mathbf{X}'\mathbf{X})^{-1})$$

  where $\boldsymbol{\beta}_0 = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}_0$.

Zellner's g-prior for Model Selection

Centered model, i.e., columns of $\mathbf{X}$ are centered:

$$\mathbf{y} = \mathbf{1}_n\beta_0 + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- $\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \sigma^2, \boldsymbol{\gamma} \sim \mathsf{N}(\mathbf{0}, g\sigma^2(\mathbf{X}'\mathbf{X})^{-1})$
- $p(\beta_0) \propto 1$
- $p(\sigma^2) \propto 1/\sigma^2$

Advantage of g-prior:

- Invariance to transformation of $\mathbf{X}_j$
- Marginal likelihood $m(\mathbf{y} \mid \boldsymbol{\gamma})$ has a closed form.

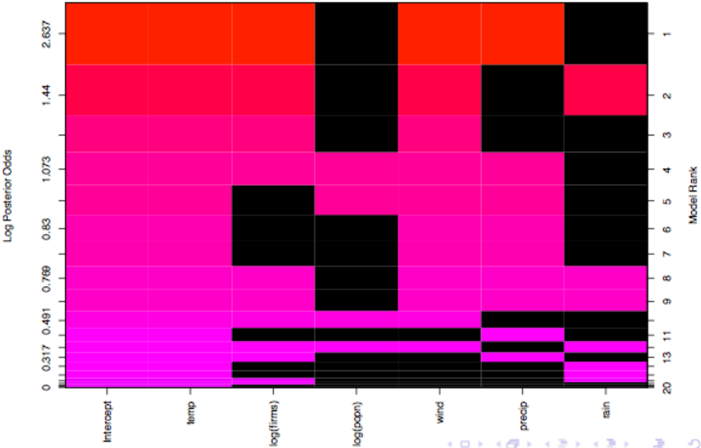
USair Data

```
library(BAS)
poll.bma = bas.lm(log(SO2) ~ temp + log(firms) +
                              log(popn) + wind +
                              precip+ rain,
                  data=pollution,
                  prior="g-prior",
                  alpha=41,
                  n.models=2^7,
                  update=50,
                  initprobs="Uniform")

par(mfrow=c(2,2))
plot(poll.bma, ask=F)
```

Model Space

image(poll.bma)



Coefficients

```
beta = coef(poll.bma)
par(mfrow=c(2,3));  plot(beta, subset=2:7,ask=F)
```