STA 601: Bayesian Model Choice in Linear Regression

STA 601 Fall 2021

Merlise Clyde

October 19, 2021

Bayesian Model Choice

General setting:

1. Define a list of models; let Γ be a "finite" set of different possible models.

2. Each model γ is in Γ, including the "true" model. Also, let θ_γ represent the parameters in model γ.

3. Put a prior over the set Γ. Let $\Pi_\gamma = p[\gamma] = \Pr[\gamma \text{ is true}]$, for all $\gamma \in \Gamma$.

4. Put a prior on the parameters in each model, that is, each $\pi(\theta_\gamma)$.

5. Compute marginal posterior probabilities $\Pr[\gamma|Y]$ for each model, and select a model based on the posterior probabilities or use the full posterior over all models!

Bayesian Model Probabilities

- For each model $\gamma \in \Gamma$, we need to compute $\Pr[\gamma | Y]$.

- Let $p_\gamma(Y)$ denote the marginal likelihood of the data under model γ, that is, $p[Y|\gamma]$. As before,

$$\hat{\Pi}_\gamma = \Pr[\gamma|Y] = \frac{p[Y|\gamma] \cdot p[\gamma]}{\sum_{\gamma^\star \in \Gamma} p[Y|\gamma^\star] \cdot p[\gamma^\star]} = \frac{p_\gamma(Y)\Pi_\gamma}{\sum_{\gamma^\star \in \Gamma} p_{\gamma^\star}(Y)\Pi_{\gamma^\star}}$$

$$= \frac{\Pi_\gamma \cdot \left[\int_{\Theta_\gamma} p_\gamma(Y|\theta_\gamma) \cdot \pi(\theta_\gamma) \mathrm{d}\theta_\gamma\right]}{\sum_{\gamma^\star \in \Gamma} \Pi_{\gamma^\star} \cdot \left[\int_{\Theta_{\gamma^\star}} p_{\gamma^\star}(Y|\theta_{\gamma^\star}) \cdot \pi(\theta_{\gamma^\star}) \mathrm{d}\theta_{\gamma^\star}\right]}.$$

- If we assume a uniform prior on Γ, that is, $\Pi_\gamma = \frac{1}{\#\Gamma}$, for all $\gamma \in \Gamma$, then

$$\hat{\Pi}_\gamma = \frac{p_\gamma(Y)}{\sum_{\gamma^\star \in \Gamma} p_{\gamma^\star}(Y)} = \frac{\left[\int_{\Theta_\gamma} p_\gamma(Y|\theta_\gamma) \cdot \pi(\theta_\gamma) \mathrm{d}\theta_\gamma\right]}{\sum_{\gamma^\star \in \Gamma} \left[\int_{\Theta_{\gamma^\star}} p_{\gamma^\star}(Y|\theta_{\gamma^\star}) \cdot \pi(\theta_{\gamma^\star}) \mathrm{d}\theta_{\gamma^\star}\right]}$$

Bayesian Model Selection

- How should we choose the Bayes optimal model?
- We can specify a loss function. The most common is

$$L(\hat{\gamma}, \gamma) = \mathbf{1}(\hat{\boldsymbol{\gamma}} \neq \boldsymbol{\gamma}),$$

  that is,

  1. Loss equals zero if the correct model is chosen; and
  2. Loss equals one if incorrect model is chosen.

- Next, select $\hat{\gamma}$ to minimize Bayes risk. Here, Bayes risk (expected loss over posterior) is

$$R(\hat{\gamma}) = \sum_{\gamma \in \Gamma} \mathbf{1}(\hat{\boldsymbol{\gamma}} \neq \boldsymbol{\gamma}) \cdot \hat{\Pi}_{\gamma} = 0 \cdot \hat{\Pi}_{\gamma_{\text{true}}} + \sum_{\gamma \neq \gamma_{\text{true}}} \hat{\Pi}_{\gamma} = \sum_{\gamma \neq \hat{\gamma}} \hat{\Pi}_{\gamma} = 1 - \hat{\Pi}_{\hat{\gamma}}$$

- To minimize $R(\hat{\gamma})$, choose $\hat{\gamma}$ such that $\hat{\Pi}_{\hat{\gamma}}$ is the largest! That is, select the model with the largest posterior probability.

Inference vs prediction

- What if the goal is prediction? Then maybe we should care more about predictive accuracy, rather than selecting specific variables.

- For predictions, we care about the posterior predictive distribution, that is

$$
\begin{aligned}
p(y_{n+1}|Y=(y_1,\ldots,y_n)) &= \int_\Gamma \int_{\Theta_\gamma} p(y_{n+1}|\gamma,\theta_\gamma)\cdot\pi(\gamma,\theta_\gamma|Y)\,\mathrm{d}\theta_\gamma \mathrm{d}\gamma \\
&= \int_\Gamma \int_{\Theta_\gamma} p(y_{n+1}|\gamma,\theta_\gamma)\cdot\pi(\theta_\gamma|Y,\gamma)\cdot\Pr[\gamma|Y]\,\mathrm{d}\theta_\gamma \mathrm{d}\gamma \\
&= \sum_{\gamma\in\Gamma} \int_{\Theta_\gamma} p(y_{n+1}|\gamma,\theta_\gamma)\cdot\pi(\theta_\gamma|Y,\gamma)\cdot\hat{\Pi}_\gamma\,\mathrm{d}\theta_\gamma \\
&= \sum_{\gamma\in\Gamma} \hat{\Pi}_\gamma \cdot \int_{\Theta_\gamma} p(y_{n+1}|\gamma,\theta_\gamma)\cdot\pi(\theta_\gamma|Y,\gamma)\,\mathrm{d}\theta_\gamma \\
&= \sum_{\gamma\in\Gamma} \hat{\Pi}_\gamma \cdot p(y_{n+1}|Y,\gamma),
\end{aligned}
$$

which is just averaging out the predictions from each model, over all possible models in Γ, with the posterior probability of each model, and this is known as Bayesian model averaging (BMA).

Bayesian Linear Regression

Practical Issues: the posterior probability that the model is true

$$\hat{\Pi}_\gamma = \frac{\Pi_\gamma p_\gamma(Y)}{\sum_{\gamma^\star \in \Gamma} \Pi_{\gamma^\star} p_{\gamma^\star}(Y)}.$$

- We need to calculate marginal likelihoods for ALL models in Γ
- In general for, we cannot calculate the marginal likelihoods unless we have a proper or conjugate priors (Normal-Gamma priors within each model)
- We need to specify proprer prior distributions on all common parameters θ_γ in each models! Conventional priors such as Zellner's g-prior or Ridge Regression to reduce elicitation of prior covariances
- Can put priors on hyperparameters cases and integrate or use numerical approximations!
- May not be able to enumerate! Gibbs or MCMC for more flexibility!

Bayesian Variable Selection (BVS)

- Rewrite each model $\gamma \in \Gamma$ as

$$\boldsymbol{Y} \mid \alpha, \boldsymbol{\beta}_\gamma, \gamma, \phi \sim \mathcal{N}_n(\mathbf{1}_n \alpha + \boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma, \phi^{-1} \boldsymbol{I}_{n \times n})$$

- γ represents the set of predictors we want to include in our model.

- $\gamma = (\gamma_1, \ldots, \gamma_p) \in \{0, 1\}^p$, so that the cardinality of Γ is 2^p, the number of models in Γ.

$$\gamma_j = \begin{cases} 1 & \text{if the j'th predictor is included in the model} \\ 0 & \text{if it is not} \end{cases}$$

- $p_\gamma \equiv \sum_{j=1}^{p} \gamma_j$, so that p_γ is the number of predictors included in model γ

- $\boldsymbol{X}_\gamma$ ($n \times p_\gamma$) is the matrix of predictors with $\gamma_j = 1$ (wolg design matrix with centered columns)

- $\boldsymbol{\beta}_\gamma$ ($p_\gamma \times 1$)is the corresponding vector of predictors with $\gamma_j = 1$

BVS

- Recall that we can also write each model as

$$Y_i = 1\alpha + \boldsymbol{\beta}_\gamma^T \boldsymbol{x}_{i\gamma} + \epsilon_i; \quad \epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \phi^{-1}).$$

- As an example, suppose we had data with 5 potential predictors including the intercept, so that each $\boldsymbol{x}_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5})$, and $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$.

- Then for model with $\gamma = (1, 0, 0, 0, 0)$, $Y_i = \boldsymbol{\beta}_\gamma^T \boldsymbol{x}_{i\gamma} + \epsilon_i$

$$\implies Y_i = \alpha + \beta_1 x_{i1} + \epsilon_i; \quad \epsilon_i \overset{iid}{\sim} \mathcal{N}(0, 1/\phi),$$

  with $p_\gamma = 1$.

- Whereas for model with $\gamma = (0, 0, 1, 1, 0)$, $Y_i = \alpha + \boldsymbol{\beta}_\gamma^T \boldsymbol{x}_{i\gamma} + \epsilon_i$

$$\implies Y_i = \alpha + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i; \quad \epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2 1/\phi),$$

  with $p_\gamma = 2$.

Steps

The outline for variable selection would be as follows:

1) Write down likelihood under model γ. That is,

$$p(\boldsymbol{y}|\boldsymbol{X}, \gamma, \alpha, \boldsymbol{\beta}_\gamma, \phi) \propto (\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{\phi}{2}(\boldsymbol{y} - \mathbf{1}\alpha - \boldsymbol{X}_\gamma\boldsymbol{\beta}_\gamma)^T(\boldsymbol{y} - \mathbf{1}\alpha - \boldsymbol{X}_\gamma\boldsymbol{\beta}_\gamma)\right\}$$

2) Define a prior for γ, $\Pi_\gamma = \Pr[\gamma]$.

- $p(\gamma_j = 1) = .5 \Rightarrow p(\gamma) = .5^p$ Uniform on space of models and $p_\gamma \sim \text{Bin}(p, .5)$
- $\gamma_j \mid \pi \overset{iid}{\sim} \text{Ber}(\pi)$ and $\pi \sim \text{Beta}(a, b)$ then $p_\gamma \sim \text{Beta-Binomial}(a, b)$

$$p(p_\gamma \mid p, a, b) = \frac{\Gamma(p+1)\Gamma(p_\gamma + a)\Gamma(p - p_\gamma + b)\Gamma(a+b)}{\Gamma(p_\gamma + 1)\Gamma(p - p_\gamma + 1)\Gamma(p + a + b)\Gamma(a)\Gamma(b)}$$

$$p_\gamma \sim \text{Beta-Binomial}(1, 1) \sim \text{Unif}(0, p)$$

Prior on model specific parameters

3) Using independent Jeffrey's priors on common parameters and the g-prior we have

$$
\begin{aligned}
\pi(\alpha, \phi) &= \phi^{-1} \\
\pi(\boldsymbol{\beta}_\gamma | \phi) &= \mathsf{N}_p \left(\boldsymbol{\beta}_{0\gamma} = \mathbf{0}, \Sigma_{0\gamma} = \frac{g}{\phi} \left[\boldsymbol{X}_\gamma{}^T \boldsymbol{X}_\gamma \right]^{-1} \right)
\end{aligned}
$$

Posteriors

- With those pieces, the conditional posteriors are straightforward

$$
\begin{aligned}
\alpha \mid \phi, y &\sim \mathsf{N}\left(\bar{y}, \frac{1}{n\phi}\right) \\
\boldsymbol{\beta}_\gamma \mid \gamma, \phi, g, y &\sim \mathsf{N}\left(\frac{g}{1+g}\hat{\boldsymbol{\beta}}_\gamma, \frac{g}{1+g}\frac{1}{\phi}\left[\boldsymbol{X}_\gamma{}^T\boldsymbol{X}_\gamma\right]^{-1}\right) \\
\phi \mid \gamma, y &\sim \mathsf{Gamma}(\cdot, \cdot) \\
p(\gamma \mid y) &\propto p(y \mid \gamma)p(\gamma)
\end{aligned}
$$

- due to conjugacy, the marginal likelihood of γ is proportional to

$$
p(Y \mid \gamma) = C(1+g)^{\frac{n-p_\gamma-1}{2}}(1+g(1-R^2_\gamma))^{-\frac{(n-1)}{2}}
$$

- R^2_γ is the usual coefficient of determination for model γ,

$$
R^2_\gamma = 1 - \frac{(y-\hat{y}_\gamma)^T(y-\hat{y}_\gamma)}{(y-\mathbf{1}\bar{y})^T(y-\mathbf{1}\bar{y})}
$$

- we can run a collapsed Gibbs or MH sampler over just Γ!

Summaries

- We can then compute marginal posterior probabilities $\Pr[\gamma|Y]$ for each model and select model with the highest posterior probability.
- We can also compute posterior $\Pr[\gamma_j = 1 \mid Y]$, the posterior probability of including the j'th predictor, often called marginal inclusion probability (MIP), allowing for uncertainty in the other predictors.
- Also straightforward to do model averaging once we all have posterior samples.
- The Hoff book works through one example and you can find the Gibbs sampler for doing inference there. I strongly recommend you go through it carefully!
- Also paper by Liang et al (2008) JASA
- we will focus on using R packages for implementing

Examples with BAS

```
library(BAS)
data(usair, package="HH")
poll.bma = bas.lm(log(SO2) ~ temp + log(mfgfirms) +
                              log(popn) + wind +
                              precip + raindays,
                  data=usair,
                  prior="g-prior",
                  alpha=nrow(usair), # g = n
                  n.models=2^6,
                  modelprior = uniform(),
                  method="deterministic")
```

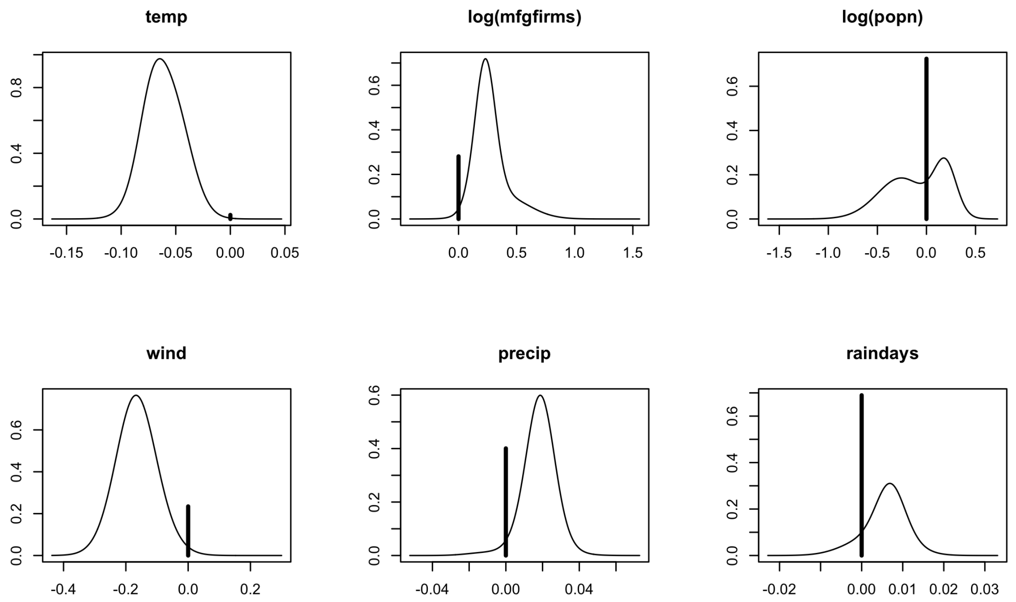
Summaries

```
poll.bma
```

```
##
## Call:
## bas.lm(formula = log(SO2) ~ temp + log(mfgfirms) + log(popn) +
##     wind + precip + raindays, data = usair, n.models = 2^6, prior = "g-p
##     alpha = nrow(usair), modelprior = uniform(), method = "deterministic
##
##
##  Marginal Posterior Inclusion Probabilities:
##     Intercept          temp  log(mfgfirms)      log(popn)          win
##        1.0000        0.9755         0.7190         0.2757        0.765
##        precip      raindays
##        0.5994        0.3104
```

Plots of Coefficients

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



Summary of Coefficients

```
beta
```

```
##
##  Marginal Posterior Summaries of Coefficients:
##
##  Using  BMA
##
##  Based on the top  64 models
##               post mean  post SD   post p(B != 0)
## Intercept      3.153004   0.082872   1.000000
## temp          -0.059724   0.020675   0.975504
## log(mfgfirms)  0.195716   0.177190   0.719031
## log(popn)     -0.026093   0.164277   0.275681
## wind          -0.126379   0.090777   0.765449
## precip         0.010821   0.011497   0.599380
## raindays       0.001803   0.004023   0.310357
```

Iterated Expectations!

Model Space Visualization

```
image(poll.bma, rotate=FALSE)
```