STA 601: Bayesian Model Averaging

STA 601 Fall 2021

Merlise Clyde

October 19, 2021

Posteriors

Likelihood under model γ

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

Posteriors

Likelihood under model γ

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

Independent Jeffrey's priors on common parameters and the g-prior

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

Posteriors

With those pieces, the conditional posteriors are straightforward

$$\alpha \mid \phi, y \sim \mathsf{N}\left(\bar{y}, \frac{1}{n\phi}\right)$$

$$\boldsymbol{\beta}_\gamma \mid \gamma, \phi, g, y \sim \mathsf{N}\left(\frac{g}{1+g}\hat{\boldsymbol{\beta}}_\gamma, \frac{g}{1+g}\frac{1}{\phi}\left[\boldsymbol{X}_\gamma{}^T\boldsymbol{X}_\gamma\right]^{-1}\right)$$

$$\phi \mid \gamma, y \sim \mathsf{Gamma}\left(\frac{n-1}{2}, \frac{\mathsf{TotalSS} - \frac{g}{1+g}\mathsf{RegSS}}{2}\right)$$

$$p(\gamma \mid y) \propto p(y \mid \gamma)p(\gamma)$$

$$\mathsf{TotalSS} \equiv \sum_i (y_i - \bar{y})^2 \qquad \mathsf{RegSS} \equiv \hat{\boldsymbol{\beta}}_\gamma^T \boldsymbol{X}_\gamma^T \boldsymbol{X}_\gamma \hat{\beta}\gamma$$

$$R_\gamma^2 = \frac{\mathsf{RegSS}}{\mathsf{TotalSS}} = 1 - \frac{\mathsf{ErrorSS}}{\mathsf{TotalSS}}$$

Posteriors

With those pieces, the conditional posteriors are straightforward

$$\alpha \mid \phi, y \sim \mathsf{N}\left(\bar{y}, \frac{1}{n\phi}\right)$$

$$\boldsymbol{\beta}_\gamma \mid \gamma, \phi, g, y \sim \mathsf{N}\left(\frac{g}{1+g}\hat{\boldsymbol{\beta}}_\gamma, \frac{g}{1+g}\frac{1}{\phi}\left[\boldsymbol{X}_\gamma{}^T \boldsymbol{X}_\gamma\right]^{-1}\right)$$

$$\phi \mid \gamma, y \sim \mathsf{Gamma}\left(\frac{n-1}{2}, \frac{\mathsf{TotalSS} - \frac{g}{1+g}\mathsf{RegSS}}{2}\right)$$

$$p(\gamma \mid y) \propto p(y \mid \gamma) p(\gamma)$$

$$\mathsf{TotalSS} \equiv \sum_i (y_i - \bar{y})^2 \qquad \mathsf{RegSS} \equiv \hat{\boldsymbol{\beta}}_\gamma^T \boldsymbol{X}_\gamma^T \boldsymbol{X}_\gamma \hat{\beta}\gamma$$

$$R^2_\gamma = \frac{\mathsf{RegSS}}{\mathsf{TotalSS}} = 1 - \frac{\mathsf{ErrorSS}}{\mathsf{TotalSS}}$$

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

Find Posteriors

Continued

Summaries

- We can run a collapsed Gibbs or MH sampler over just Γ!
- We can then compute marginal posterior probabilities $\Pr[\gamma|Y]$ for each model and select model with the highest posterior probability.

Summaries

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

Summaries

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

Summaries

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

Summaries

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

Summaries

- We can run a collapsed Gibbs or MH sampler over just Γ!
- 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")
```

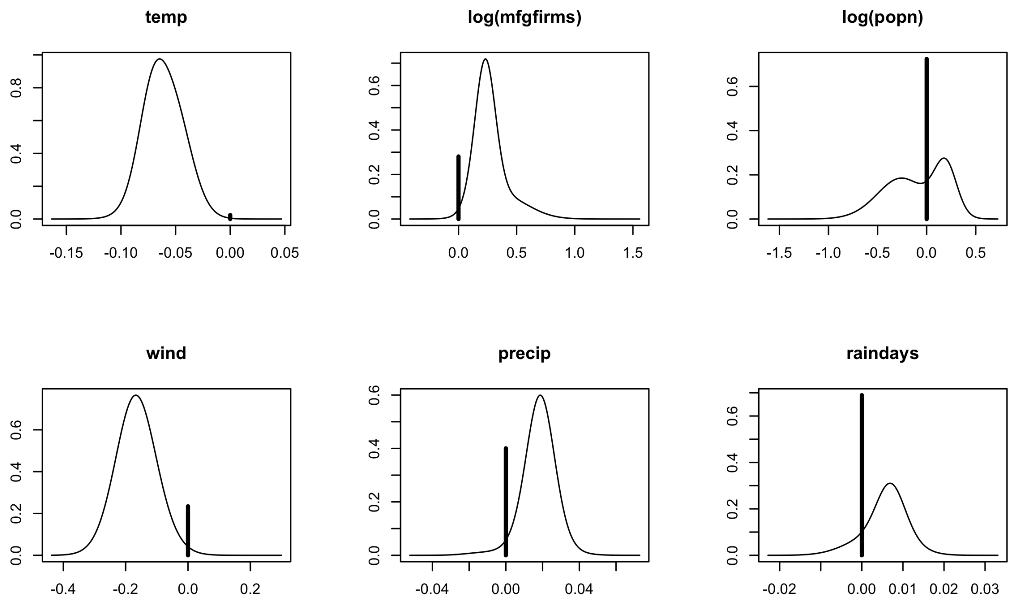
Summaries

```
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

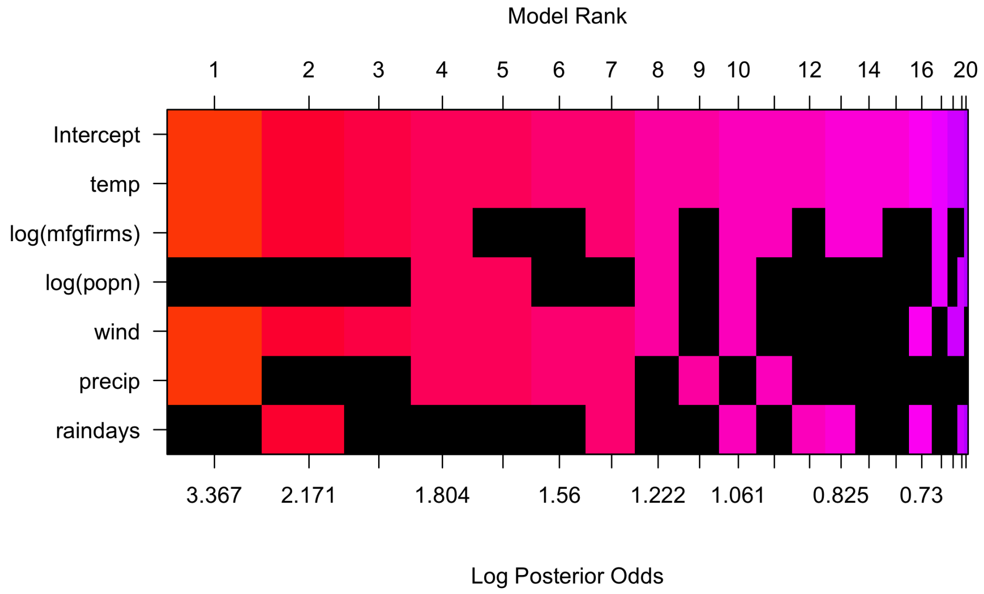
```
beta
```

```
##
##  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)
```



Bartlett's Paradox

$$\text{BF}(\gamma : \gamma_0) = (1+g)^{(n-1-p_\gamma)/2}(1+g(1-R^2_\gamma))^{-(n-1)/2}$$

- What happens to Bayes Factors or posterior probabilites of γ as $g \to \infty$? (for fixed data)

Bartlett's Paradox

$$\mathsf{BF}(\gamma : \gamma_0) = (1+g)^{(n-1-p_\gamma)/2}(1+g(1-R^2_\gamma))^{-(n-1)/2}$$

- What happens to Bayes Factors or posterior probabilites of γ as $g \to \infty$? (for fixed data)
- What happens to Bayes Factor as $g \to 0$

Information Paradox

$$\text{BF}(\gamma : \gamma_0) = (1+g)^{(n-1-p_\gamma)/2}(1+g(1-R^2_\gamma))^{-(n-1)/2}$$

- Let g be a fixed constant and take n fixed imagine a sequence of data such that $R^2_\gamma \to 1$ (increasing explained variation)

Information Paradox

$$\mathsf{BF}(\gamma : \gamma_0) = (1+g)^{(n-1-p_\gamma)/2}(1+g(1-R^2_\gamma))^{-(n-1)/2}$$

- Let g be a fixed constant and take n fixed imagine a sequence of data such that $R^2_\gamma \to 1$ (increasing explained variation)

- Let $F = \frac{R^2_\gamma/p_\gamma}{(1-R^2_\gamma)/(n-1-p_\gamma)}$

Information Paradox

$$\text{BF}(\gamma : \gamma_0) = (1+g)^{(n-1-p_\gamma)/2}(1+g(1-R^2_\gamma))^{-(n-1)/2}$$

- Let g be a fixed constant and take n fixed imagine a sequence of data such that $R^2_\gamma \to 1$ (increasing explained variation)
- Let $F = \frac{R^2_\gamma / p_\gamma}{(1-R^2_\gamma)/(n-1-p_\gamma)}$
- As $R^2_\gamma \to 1$, $F \to \infty$ LR test would reject γ_0 where F is the usual F statistic for comparing model γ to γ_0

Information Paradox

$$\text{BF}(\gamma : \gamma_0) = (1+g)^{(n-1-p_\gamma)/2}(1+g(1-R^2_\gamma))^{-(n-1)/2}$$

- Let g be a fixed constant and take n fixed imagine a sequence of data such that $R^2_\gamma \to 1$ (increasing explained variation)
- Let $F = \frac{R^2_\gamma/p_\gamma}{(1-R^2_\gamma)/(n-1-p_\gamma)}$
- As $R^2_\gamma \to 1$, $F \to \infty$ LR test would reject γ_0 where F is the usual F statistic for comparing model γ to γ_0
- BF converges to a fixed constant $(1+g)^{-p_\gamma/2}$ (does not go to infinity

Information Paradox

$$\text{BF}(\gamma : \gamma_0) = (1+g)^{(n-1-p_\gamma)/2}(1+g(1-R^2_\gamma))^{-(n-1)/2}$$

- Let g be a fixed constant and take n fixed imagine a sequence of data such that $R^2_\gamma \to 1$ (increasing explained variation)
- Let $F = \frac{R^2_\gamma/p_\gamma}{(1-R^2_\gamma)/(n-1-p_\gamma)}$
- As $R^2_\gamma \to 1$, $F \to \infty$ LR test would reject γ_0 where F is the usual F statistic for comparing model γ to γ_0
- BF converges to a fixed constant $(1+g)^{-p_\gamma/2}$ (does not go to infinity
- one predictor example

Information Paradox

$$\text{BF}(\gamma : \gamma_0) = (1+g)^{(n-1-p_\gamma)/2}(1+g(1-R^2_\gamma))^{-(n-1)/2}$$

- Let g be a fixed constant and take n fixed imagine a sequence of data such that $R^2_\gamma \to 1$ (increasing explained variation)
- Let $F = \frac{R^2_\gamma / p_\gamma}{(1-R^2_\gamma)/(n-1-p_\gamma)}$
- As $R^2_\gamma \to 1$, $F \to \infty$ LR test would reject γ_0 where F is the usual F statistic for comparing model γ to γ_0
- BF converges to a fixed constant $(1+g)^{-p_\gamma/2}$ (does not go to infinity
- one predictor example

Information Inconsistency see Liang et al JASA 2008

Mixtures of g-priors & Information Consistency

Need $BF \to \infty$ if $R^2_\gamma \to 1 \Leftrightarrow \mathsf{E}_g[(1+g)^{-p_\gamma/2}]$ diverges for $p_\gamma < n-1$ (proof in Liang et al)

Mixtures of g-priors & Information Consistency

Need $BF \rightarrow \infty$ if $R^2_\gamma \rightarrow 1 \Leftrightarrow \mathsf{E}_g[(1+g)^{-p_\gamma/2}]$ diverges for $p_\gamma < n - 1$ (proof in Liang et al)

- Zellner-Siow Cauchy prior, $1/g \sim \mathsf{Gamma}(1/2, 1/2))$

Mixtures of g-priors & Information Consistency

Need $BF \to \infty$ if $R^2_\gamma \to 1 \Leftrightarrow \mathsf{E}_g[(1+g)^{-p_\gamma/2}]$ diverges for $p_\gamma < n-1$ (proof in Liang et al)

- Zellner-Siow Cauchy prior, $1/g \sim \mathsf{Gamma}(1/2, 1/2))$
- hyper-g prior or hyper-g/n (Liang et al JASA 2008)

Mixtures of g-priors & Information Consistency

Need $BF \to \infty$ if $R^2_\gamma \to 1 \Leftrightarrow \mathsf{E}_g[(1+g)^{-p_\gamma/2}]$ diverges for $p_\gamma < n-1$ (proof in Liang et al)

- Zellner-Siow Cauchy prior, $1/g \sim \mathsf{Gamma}(1/2, 1/2))$
- hyper-g prior or hyper-g/n (Liang et al JASA 2008)
- robust prior (Bayarrri et al Annals of Statistics 2012

Mixtures of g-priors & Information Consistency

Need $BF \rightarrow \infty$ if $R^2_\gamma \rightarrow 1 \Leftrightarrow \mathsf{E}_g[(1+g)^{-p_\gamma/2}]$ diverges for $p_\gamma < n-1$ (proof in Liang et al)

- Zellner-Siow Cauchy prior, $1/g \sim \textsf{Gamma}(1/2, 1/2))$
- hyper-g prior or hyper-g/n (Liang et al JASA 2008)
- robust prior (Bayarrri et al Annals of Statistics 2012

  All have tails that behave like a Cauchy distribution (robustness)