

# Cauchy Priors: Mixtures of Normals & MCMC

STA721 Linear Models Duke University

Merlise Clyde

October 8, 2019

# Bayesian Estimation with 2 Block $g$ -prior (Normal-Jeffreys)

Model in centered parameterization

$$\begin{aligned}\mathbf{Y} &= \mathbf{1}\beta_0 + (\mathbf{I}_n - \mathbf{P}_1)\mathbf{X}_1\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ p(\beta_0, \phi) &\propto 1/\phi \\ \boldsymbol{\beta} \mid \beta_0, \phi &\sim \mathbf{N}(\mathbf{0}, \frac{g}{\phi}(\mathbf{X}^T(\mathbf{I}_n - \mathbf{P}_1)\mathbf{X})^{-1})\end{aligned}$$

Log Likelihood

$$\mathcal{L}(\beta_0, \beta_1, \phi) \propto \frac{n}{2} \log(\phi) - \frac{\phi}{2} (n(\beta_0 - \bar{y})^2 + (\mathbf{Y}_c - \mathbf{X}_c\boldsymbol{\beta})^T(\mathbf{Y}_c - \mathbf{X}_c\boldsymbol{\beta}))$$

Since

$$\mathbf{Y} = (\mathbf{I} - \mathbf{P}_1)\mathbf{Y} + \mathbf{P}_1\mathbf{Y}$$

Integrated Likelihood after integrating  $\beta_0$

$$\mathcal{L}(\beta_1, \phi) \propto \frac{n-1}{2} \log(\phi) - \frac{\phi}{2} (\mathbf{Y}_c - \mathbf{X}_c\boldsymbol{\beta})^T(\mathbf{Y}_c - \mathbf{X}_c\boldsymbol{\beta})$$

## Prior Data

Note

$$(\mathbf{X}^T(\mathbf{I}_n - \mathbf{P}_1)\mathbf{X}) = (\mathbf{X}^T(\mathbf{I}_n - \mathbf{P}_1)^T(\mathbf{I}_n - \mathbf{P}_1)\mathbf{X}) = (\mathbf{X} - \mathbf{1}_n\bar{\mathbf{X}}^T)^T(\mathbf{X} - \mathbf{1}_n\bar{\mathbf{X}})$$

$$\text{Let } (\mathbf{X} - \mathbf{1}_n\bar{\mathbf{X}}^T)^T(\mathbf{X} - \mathbf{1}_n\bar{\mathbf{X}}) = SS_{\mathbf{X}} = \mathbf{U}^T\mathbf{U}$$

Quadratic contribution to the log likelihood from prior after integrating out  $\beta_0$

$$(\mathbf{Y}_c - \mathbf{X}_c\beta)^T(\mathbf{Y}_c - \mathbf{X}_c\beta) + (\beta^T \frac{\mathbf{U}^T\mathbf{U}}{g}\beta)$$

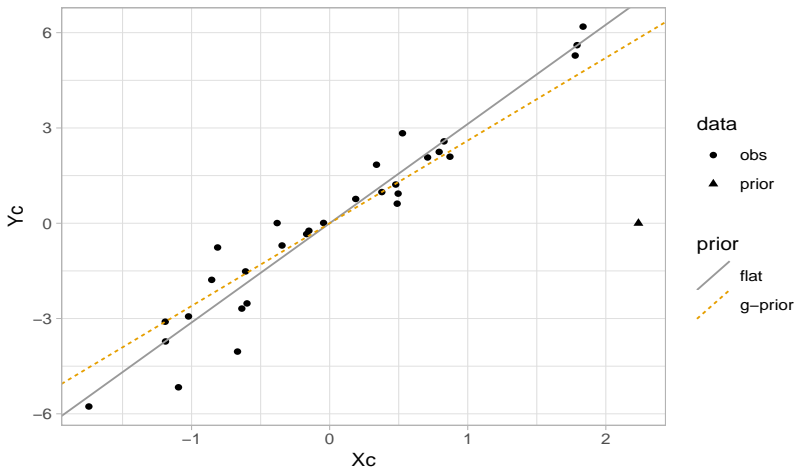
$$(\mathbf{Y}_c - \mathbf{X}_c\beta)^T(\mathbf{Y}_c - \mathbf{X}_c\beta) + (\mathbf{0}_p - \frac{\mathbf{U}}{\sqrt{g}}\beta)^T(\mathbf{0}_p - \frac{\mathbf{U}}{\sqrt{g}}\beta)$$

Prior observations with  $Y_c = 0$ .

## Example: $g=5$ , $n=30$

In SLR it is like an extra  $Y_0 = 0$  at  $\mathbf{X}_0 = \sqrt{\frac{SS_x}{g}}$ :

$$(\mathbf{Y}_c - \mathbf{X}_c\beta)^T(\mathbf{Y}_c - \mathbf{X}_c\beta) + (0 - \sqrt{\frac{SS_x}{g}}\beta)^T(0 - \sqrt{\frac{SS_x}{g}}\beta)$$



# Disadvantages of Conjugate Priors

## Disadvantages:

- ▶ Results may have be sensitive to prior “outliers” due to linear updating
- ▶ Problem potentially with all Normal priors, not just the  $g$ -prior.
- ▶ Cannot capture all possible prior beliefs
- ▶ Mixtures of Conjugate Priors

# Mixtures of Conjugate Priors

## Theorem (Diaconis & Ylvisaker 1985)

*Given a sampling model  $p(y \mid \theta)$  from an exponential family, any prior distribution can be expressed as a mixture of conjugate prior distributions*

- ▶ Prior  $p(\theta) = \int p(\theta \mid \omega)p(\omega) d\omega$
- ▶ Posterior

$$\begin{aligned} p(\theta \mid \mathbf{Y}) &\propto \int p(\mathbf{Y} \mid \theta)p(\theta \mid \omega)p(\omega) d\omega \\ &\propto \int \frac{p(\mathbf{Y} \mid \theta)p(\theta \mid \omega)}{p(\mathbf{Y} \mid \omega)} p(\mathbf{Y} \mid \omega)p(\omega) d\omega \\ &\propto \int p(\theta \mid \mathbf{Y}, \omega)p(\mathbf{Y} \mid \omega)p(\omega) d\omega \\ p(\theta \mid \mathbf{Y}) &= \frac{\int p(\theta \mid \mathbf{Y}, \omega)p(\mathbf{Y} \mid \omega)p(\omega) d\omega}{\int p(\mathbf{Y} \mid \omega)p(\omega) d\omega} \end{aligned}$$

## Zellner-Siow prior (assume $\mathbf{X}$ is centered)

Zellner's g-prior  $\beta \mid \phi \sim N(\mathbf{0}_p, g(\mathbf{X}_c^T \mathbf{X}_c)^{-1} / \phi)$

- ▶ Choice of  $g$ ?
- ▶  $\frac{g}{1+g}$  weight given to the data
- ▶ Let  $\tau = 1/g$  assign  $\tau \sim G(1/2, n/2)$
- ▶ Marginal prior on  $\beta \sim C(0, \phi^{-1}(\mathbf{X}_c^T \mathbf{X}_c / n)^{-1})$
- ▶ Can express posterior as a mixture of g-priors

$$p(\tau \mid \mathbf{Y}) = \frac{p(\mathbf{Y} \mid \tau)p(\tau)}{\int p(\mathbf{Y} \mid \tau)p(\tau) d\tau}$$

- ▶ Problem: no analytic expression for integral
- ▶ Need 2 one dimensional integrals to obtain posterior.
- ▶ What about credible intervals?

# Markov Chain Monte Carlo

- ▶ We know that  $\beta_0, \beta, \phi \mid \mathbf{Y}, g = 1/\tau$  has a Normal-Gamma distribution
- ▶ We can show that  $\tau \mid \beta_0, \beta, \phi, \mathbf{Y}$  has a Gamma distribution

$$p(\tau \mid \beta, \phi, \mathbf{Y}) \propto \mathcal{L}(\beta_0, \beta, \phi) \tau^{p/2} e^{(-\tau \frac{\phi}{2} \beta^T (\mathbf{X}^T \mathbf{X}) \beta)} \tau^{1/2-1} e^{-\tau n/2}$$

- ▶ alternate sampling from full conditional distributions given current values of other parameters. (STA 601)
- ▶ JAGS or STAN



## JAGS Code: library(R2jags)

```
model = function(){  
  for (i in 1:n) {  
    Y[i] ~ dnorm(beta0+ (X[i] -Xbar)*beta, phi)  
  }  
  beta0 ~ dnorm(0, .000001*phi) #precision is 2nd arg  
  beta ~ dnorm(0, phi*tau*SSX) #precision is 2nd arg  
  phi ~ dgamma(.001, .001)  
  tau ~ dgamma(.5, .5*n)  
  g <- 1/tau  
  sigma <- pow(phi, -.5)  
}  
data = list(Y=Y, X=X, n=length(Y), SSX=sum(Xc^2),  
            Xbar=mean(X))  
ZSout = jags(data, inits=NULL,  
              parameters.to.save=c("beta0", "beta", "g",  
                                   "sigma"),  
              model=model, n.iter=10000)
```

## HPD intervals

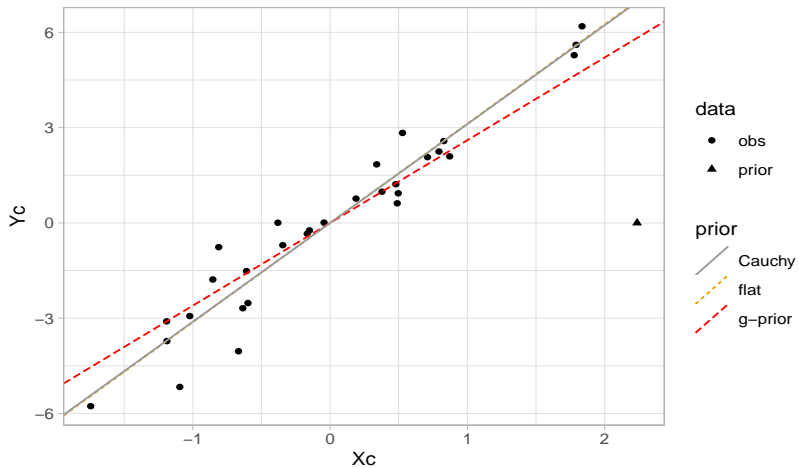
```
confint(lm(Y ~ Xc))
```

```
##                2.5 %    97.5 %  
## (Intercept) -0.3985359 0.2048303  
## Xc           2.7945824 3.4555162
```

```
HPDinterval(as.mcmc(ZSout$BUGSoutput$sims.matrix))
```

```
##                lower      upper  
## beta           2.7823047    3.4453690  
## beta0          -0.3764027    0.2095465  
## deviance       70.2043917    78.4813041  
## g              19.4503373 3782.7134974  
## sigma          0.6171029    1.0504892  
## attr(,"Probability")  
## [1] 0.95
```

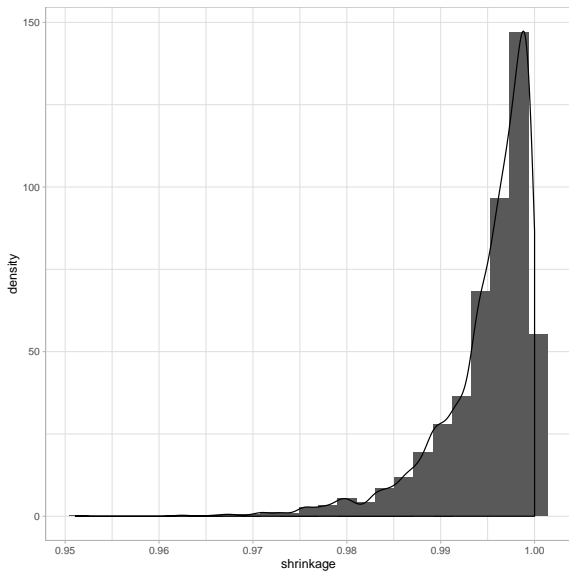
# Compare



ZSout

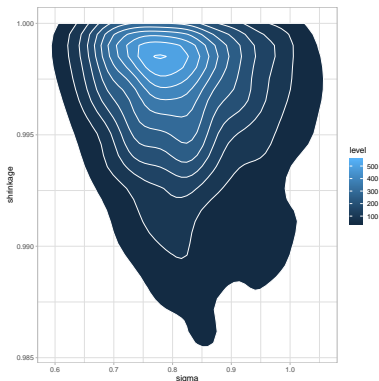
```
## Inference for Bugs model at "/var/folders/n4/nj1122xj6bn5_xgbptv7bml40000gp/T//RtmpmL3pQi/modelc086a1a
## 3 chains, each with 10000 iterations (first 5000 discarded), n.thin = 5
## n.sims = 3000 iterations saved
##          mu.vect    sd.vect    2.5%    25%    50%    75%    97.5%  Rhat
## beta          3.112      0.170  2.782   2.997   3.115   3.225   3.445  1.001
## beta0         -0.099      0.152 -0.384  -0.204  -0.099   0.001   0.204  1.002
## g            2263.147 38967.029 48.273 146.129 282.298 697.063 9018.709 1.001
## sigma          0.827      0.114  0.636   0.747   0.816   0.896   1.079  1.001
## deviance       73.347      2.563 70.390  71.458  72.680  74.500  79.882  1.002
##          n.eff
## beta          3000
## beta0         1200
## g             3000
## sigma         3000
## deviance      1600
##
## For each parameter, n.eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor (at convergence, Rhat=1).
##
## DIC info (using the rule, pD = var(deviance)/2)
## pD = 3.3 and DIC = 76.6
## DIC is an estimate of expected predictive error (lower deviance is better).
```

# Posterior Distribution of shrinkage



# Joint Distribution of $\sigma$ and $g/(1+g)$

```
ggplot(postdf, aes(x=sigma, y=shrinkage) ) +  
  stat_density_2d(aes(fill = ..level..),  
                  geom = "polygon", colour="white") +  
  theme_light()
```



# Cauchy Summary

- ▶ Cauchy rejects prior mean if it is an "outlier"
- ▶ robustness related to "bounded" influence (more later)
- ▶ requires numerical integration or Monte Carlo sampling (MCMC)

## How Good are these Estimators?

Quadratic loss for estimating  $\beta$  using estimator  $\mathbf{a}$

$$L(\beta, \mathbf{a}) = (\beta - \mathbf{a})^T (\beta - \mathbf{a})$$

- ▶ Consider our expected loss (before we see the data) of taking an “action”  $\mathbf{a}$
- ▶ Under OLS or the Reference prior the Expected Mean Square Error

$$\begin{aligned} E_{\mathbf{Y}}[(\beta - \hat{\beta})^T (\beta - \hat{\beta})] &= \sigma^2 \text{tr}[(\mathbf{X}^T \mathbf{X})^{-1}] \\ &= \sigma^2 \sum_{j=1}^p \lambda_j^{-1} \end{aligned}$$

where  $\lambda_j$  are eigenvalues of  $\mathbf{X}^T \mathbf{X}$ .

- ▶ If smallest  $\lambda_j \rightarrow 0$  then  $\text{MSE} \rightarrow \infty$
- ▶ Note: estimate is unbiased!



## Is the $g$ -prior better?

Explore Frequentist properties of using a Bayesian estimator

$$E_{\mathbf{Y}}[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_g)^T(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_g)]$$

but now  $\hat{\boldsymbol{\beta}}_g = g/(1 + g)\hat{\boldsymbol{\beta}}$

when is the  $g$  prior better than the Reference prior or OLS? Is it always better?

# Estimator Properties

- ▶ Bias
- ▶ Variability
- ▶  $MSE = Bias^2 + Variance$  (multivariate analogs)
- ▶ Problems with OLS, g-priors & mixtures of g-priors with collinearity
- ▶ Solutions:
  - ▶ removal of terms
  - ▶ other shrinkage estimators