More Prior Distributions

STA721 Linear Models Duke University

Merlise Clyde

September 22, 2014

Bayesian Estimation

Model

$$\mathbf{Y} \sim \mathsf{N}(\mathbf{X}\beta, \mathbf{I}_n/\phi)$$

with precision $\phi = 1/\sigma^2$.

More Prior Choices:

- Jeffreys' Priors
- More on g-priors
- Zellner-Siow Cauchy Prior

Jeffreys Prior

Jeffreys proposed a default procedure so that resulting prior would be invariant to model parameterization

$$p(\boldsymbol{\theta}) \propto |\mathcal{I}(\boldsymbol{\theta})|^{1/2}$$

where $\mathcal{I}(\boldsymbol{\theta})$ is the Expected Fisher Information matrix

$$\mathcal{I}(\theta) = -\mathsf{E}[\left[\frac{\partial^2 \log(\mathcal{L}(\boldsymbol{\theta}))}{\partial\theta_i \partial\theta_j}\right]]$$

Log Likelihood

$$\log(\mathcal{L}(\boldsymbol{\beta}, \phi)) = \frac{n}{2}\log(\phi) - \frac{\phi}{2}\|(\mathbf{I} - \mathsf{P}_{\mathbf{x}})\mathbf{Y}\|^2 - \frac{\phi}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T(\mathbf{X}^T\mathbf{X})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$$

$$\begin{aligned}
\frac{\partial^2 \log \mathcal{L}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} &= \begin{bmatrix} -\phi(\mathbf{X}^T\mathbf{X}) & -(\mathbf{X}^T\mathbf{X})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \\ -(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T(\mathbf{X}^T\mathbf{X}) & -\frac{n}{2}\frac{1}{\phi^2} \end{bmatrix} \\
\mathsf{E}\left[\frac{\partial^2 \log \mathcal{L}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right] &= \begin{bmatrix} -\phi(\mathbf{X}^T\mathbf{X}) & \mathbf{0}_p \\ \mathbf{0}_p^T & -\frac{n}{2}\frac{1}{\phi^2} \end{bmatrix} \\
\mathcal{I}((\boldsymbol{\beta}, \phi)^T) &= \begin{bmatrix} \phi(\mathbf{X}^T\mathbf{X}) & \mathbf{0}_p \\ \mathbf{0}_p^T & \frac{n}{2}\frac{1}{\phi^2} \end{bmatrix}
\end{aligned}$$

Jeffreys Prior

Jeffreys Prior

$$
\begin{aligned}
p_J(\boldsymbol{\beta}, \phi) &\propto |\mathcal{I}((\boldsymbol{\beta}, \phi)^T)|^{1/2} \\
&= |\phi(\mathbf{X}^T\mathbf{X}|^{1/2} \left(\frac{n}{2}\frac{1}{\phi^2}\right)^{1/2} \\
&\propto \phi^{p/2-1}|\mathbf{X}^T\mathbf{X}|^{1/2} \\
&\propto \phi^{p/2-1}
\end{aligned}
$$

Improper prior $\int\int p_J(\boldsymbol{\beta}, \phi)\, d\boldsymbol{\beta}\, d\phi$ not finite

Formal Bayes Posterior

$$
\begin{aligned}
p(\boldsymbol{\beta}, \phi \mid \mathbf{Y}) &\propto p(\mathbf{Y} \mid \boldsymbol{\beta}, \phi)\phi^{p/2-1} \\
&\propto \phi^{n/2}\phi^{p/2-1} \exp -\frac{\phi}{2}\text{SSE} \exp(-\frac{\phi}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T\mathbf{X}^T\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}))
\end{aligned}
$$

Formal Bayes Posterior

If $p(\mathbf{Y} \mid \boldsymbol{\beta}, \phi)\phi^{p/2-1}$ can be renormalized to obtain formal posterior distribution

$$\begin{aligned} \boldsymbol{\beta} \mid \phi, \mathbf{Y} &\sim \mathsf{N}(\hat{\boldsymbol{\beta}}, (\mathbf{X}^T\mathbf{X})^{-1}\phi^{-1}) \\ \phi \mid \mathbf{Y} &\sim \mathbf{G}(n/2, \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2/2)\boldsymbol{\beta} \mid \mathbf{Y} \sim t_n(\hat{\boldsymbol{\beta}}, \frac{\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2}{n}\mathbf{X}^T\mathbf{X}) \end{aligned}$$

Limiting case of Conjugate prior with $\mathbf{b}_0 = 0$, $\Phi = \mathbf{0}$, $\nu_0 = 0$ and $\mathsf{SS}_0 = 0$
Posterior does not depend on dimension p;

Jeffreys did not recommend using this

Independent Jeffreys Prior

- Treat $\boldsymbol{\beta}$ and ϕ separately ("orthogonal parameterization")
- $p_{IJ}(\boldsymbol{\beta}) \propto |\mathcal{I}(\boldsymbol{\beta})|^{1/2}$
- $p_{IJ}(\phi) \propto |\mathcal{I}(\phi)||^{1/2}$

$$\mathcal{I}((\boldsymbol{\beta}, \phi)^T) = \begin{bmatrix} \phi(\mathbf{X}^T\mathbf{X}) & \mathbf{0}_p \\ \mathbf{0}_p^T & \frac{n}{2}\frac{1}{\phi^2} \end{bmatrix}$$

$$p_{IJ}(\boldsymbol{\beta}) \propto |\phi\mathbf{X}^T\mathbf{X}|^{1/2} \propto 1$$

$$p_{IJ}(\phi) \propto \phi^{-1}$$

Independent Jeffreys Prior is

$$p_{IJ}(\beta, \phi) \propto p_{IJ}(\boldsymbol{\beta})p_{IJ}(\phi) = \phi^{-1}$$

Formal Posterior Distribution

With Independent Jeffreys Prior

$$p_{IJ}(\beta, \phi) \propto p_{IJ}(\boldsymbol{\beta}) p_{IJ}(\phi) = \phi^{-1}$$

Formal Posterior Distribution

$$\begin{aligned}
\boldsymbol{\beta} \mid \phi, \mathbf{Y} &\sim \mathrm{N}(\hat{\boldsymbol{\beta}}, (\mathbf{X}^T\mathbf{X})^{-1}\phi^{-1}) \\
\phi \mid \mathbf{Y} &\sim \mathbf{G}((n-p)/2, \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2/2) \\
\boldsymbol{\beta} \mid \mathbf{Y} &\sim t_{n-p}(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1})
\end{aligned}$$

Bayesian Credible Sets $p(\boldsymbol{\beta} \in C_\alpha) = 1 - \alpha$ correspond to frequentist Confidence Regions

$$\frac{\boldsymbol{\lambda}^T\boldsymbol{\beta} - \boldsymbol{\lambda}\hat{\boldsymbol{\beta}}}{\sqrt{\hat{\sigma}^2\boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{\lambda}}} \sim t_{n-p}$$

Partitioned Zellner's g-prior

Zellner recognized that some parameters might have less information

$$\mathbf{Y} = \mathbf{X}_0\boldsymbol{\beta}_0 + \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}$$

- $\mathbf{X}_0^T\mathbf{X}_1 = \mathbf{0}$ (orthogonal columns)
- Fisher information block diagonal
- $\boldsymbol{\beta}_0 \sim N(\mathbf{b}_0, g_0(\mathbf{X}_0^T\mathbf{X}_0)^{-1}/\phi)$
- $\boldsymbol{\beta}_1 \sim N(\mathbf{b}_1, g_1(\mathbf{X}_1^T\mathbf{X}_1)^{-1}/\phi)$
- limiting case $g_0 \to \infty$, $\mathbf{b}_0 = 0$
- $p(\phi) \propto 1/\phi$

HW $\mathbf{X}_0 = \mathbf{1}_n$

Decompose

Disadvantages of Conjugate Priors

Disadvantages:

- Results may have be sensitive to prior "outliers" due to linear updating
- Cannot capture all possible prior beliefs
- Mixtures of Conjugate Priors

Mixtures of Conjugate Priors

Theorem (Diaconis & Ylivisaker 1985)

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

- Prior $p(\boldsymbol{\theta}) = \int p(\boldsymbol{\theta} \mid \omega) p(\omega)\, d\omega$
- Posterior

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

Zellner-Siow prior

Zellner's g-prior $\beta \mid \phi \sim \mathsf{N}(\mathbf{b}_0, g(\mathbf{X}^T\mathbf{X})^{-1}/\phi)$

- Choice of g?
- $\frac{g}{1+g}$ weight given to the data
- Let $\tau = 1/g$ assign $\tau \sim G(1/2, n/2)$
- Find prior distribution
- Can expres posterior as a mixture of g-priors

How Good are these Estimators?

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

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

- 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

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

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

- If smallest $\lambda_j \to 0$ then MSE $\to \infty$
- Note: estimate is unbiased!

Is the g-prior better?

Explore Frequentist properties of using a Bayesian estimator

$$\mathsf{E}_{\mathbf{Y}}[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_g)^T(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_g)$$

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

Estimator Properties

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