Lecture 12: Choice of Priors in Regression

STA702

Merlise Clyde
Duke University

https://sta702-F23.github.io/website/

Conjugate Priors in Linear Regression

- Regression Model (Sampling model)

$$\mathbf{Y} \mid \boldsymbol{\beta}, \phi \sim \mathsf{N}(\mathbf{X}\boldsymbol{\beta}, \phi^{-1}\mathbf{I}_n)$$

- Conjugate Normal-Gamma Model: factor joint prior $p(\boldsymbol{\beta}, \phi) = p(\boldsymbol{\beta} \mid \phi)p(\phi)$

$$\boldsymbol{\beta} \mid \phi \sim \mathsf{N}(\mathbf{b}_0, \phi^{-1}\boldsymbol{\Phi}_0^{-1}) \qquad p(\boldsymbol{\beta} \mid \phi) = \frac{|\phi\boldsymbol{\Phi}_0|^{1/2}}{(2\pi)^{p/2}} e^{\left\{-\frac{\phi}{2}(\boldsymbol{\beta}-\mathbf{b}_0)^T\boldsymbol{\Phi}_0(\boldsymbol{\beta}-\mathbf{b}_0)\right\}}$$

$$\phi \sim \mathsf{Gamma}(v_0/2, \mathsf{SS}_0/2) \qquad p(\phi) = \frac{1}{\Gamma(\nu_0/2)}\left(\frac{\mathsf{SS}_0}{2}\right)^{\nu_0/2}\phi^{\nu_0/2-1}e^{-}$$

$$\Rightarrow (\boldsymbol{\beta}, \phi) \sim \mathsf{NG}(\mathbf{b}_0, \boldsymbol{\Phi}_0, \nu_o, \mathsf{SS}_0)$$

- Need to specify the 4 hyperparameters of the Normal-Gamma distribution!
- hard in higher dimensions!

Choice of Conjugate Prior

Seek default choices

- Jeffreys' prior
- unit-information prior
- Zellner's g-prior
- ridge regression priors
- mixtures of conjugate priors
  - Zellner-Siow Cauchy Prior
  - (Bayesian) Lasso
  - Horseshoe

Which? Why?

Jeffreys' Prior

- Jeffreys prior is invariant to model parameterization of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \phi)$

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

- $\mathcal{I}(\boldsymbol{\theta})$ 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 expressed as function of sufficient statistics

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

- projection $\mathbf{P}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ onto the column space of $\mathbf{X}$

Information matrix

$$\frac{\partial^2 \log \mathcal{L}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} = \begin{bmatrix} -\phi(\mathbf{X}^T\mathbf{X}) & -(\mathbf{X}^T\mathbf{X})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \\ -(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T(\mathbf{X}^T\mathbf{X}) & -\frac{n}{2}\frac{1}{\phi^2} \end{bmatrix}$$

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

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

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

Jeffreys' did not recommend - marginal for ϕ dies not account for dimension p

Recommended Independent Jeffreys Prior

- Treat $\boldsymbol{\beta}$ and ϕ separately (*orthogonal parameterization* which implies asymptoptic independence of $\boldsymbol{\beta}$ and ϕ)
- $p_{IJ}(\boldsymbol{\beta}) \propto |\mathcal{I}(\boldsymbol{\beta})|^{1/2}$ and $p_{IJ}(\phi) \propto |\mathcal{I}(\phi)|^{1/2}$

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

$$\begin{aligned} p_{IJ}(\boldsymbol{\beta}) &\propto |\phi\mathbf{X}^T\mathbf{X}|^{1/2} \propto 1 \\ p_{IJ}(\phi) &\propto \phi^{-1} \\ p_{IJ}(\beta, \phi) &\propto p_{IJ}(\boldsymbol{\beta})p_{IJ}(\phi) = \phi^{-1} \end{aligned}$$

Two group *reference prior*

Formal Posterior Distribution

- Use Independent Jeffreys Prior $p_{IJ}(\beta, \phi) \propto p_{IJ}(\boldsymbol{\beta})p_{IJ}(\phi) = \phi^{-1}$
- Formal Posterior Distribution

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

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

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

- conditional on $\mathbf{Y}$ for Bayes and conditional on $\boldsymbol{\beta}$ for frequentist

Unit Information Prior

Unit information prior $\boldsymbol{\beta} \mid \phi \sim \mathsf{N}(\hat{\boldsymbol{\beta}}, n(\mathbf{X}^T\mathbf{X})^{-1}/\phi)$

- Based on a fraction of the likelihood $p(\boldsymbol{\beta}, \phi) \propto \mathcal{L}(\boldsymbol{\beta}, \phi)^{1/n}$

$$\log(p(\boldsymbol{\beta}, \phi) \propto \frac{1}{n}\frac{n}{2}\log(\phi) - \frac{\phi}{2}\frac{\|(\mathbf{I}_n - \mathbf{P}_{\mathbf{x}})\mathbf{Y}\|^2}{n} - \frac{\phi}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T\frac{(\mathbf{X}^T\mathbf{X})}{n}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$$

- ``average information'' in one observation is $\phi\mathbf{X}^T\mathbf{X}/n$ or "unit information"
- Posterior mean $\frac{n}{1+n}\hat{\boldsymbol{\beta}} + \frac{1}{1+n}\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$
- Posterior Distribution

$$\boldsymbol{\beta} \mid \mathbf{Y}, \phi \sim \mathsf{N}\left(\hat{\boldsymbol{\beta}}, \frac{n}{1+n}(\mathbf{X}^T\mathbf{X})^{-1}\phi^{-1}\right)$$

Unit Information Prior

- Advantages:
  - Proper
  - Invariant to model parameteriztion of $\mathbf{X}$ (next)
  - Equivalent to MLE (no bias) and tighter intervals
- Disadvantages
  - cannot represent *prior* beliefs;
  - double use of data!
  - no shrinkage of $\boldsymbol{\beta}$ with noisy data (larger variance than biased estimators)

Exercise for the Energetic Student

- What would be a "Unit information prior" for ϕ?
- What is the marginal posterior for $\boldsymbol{\beta}$ using both unit-information priors?

Invariance and Choice of Mean/Precision

- the model in vector form $Y \mid \beta, \phi \sim \mathsf{N}_n(X\beta, \phi^{-1} I_n)$
- What if we transform the mean $X\beta = XHH^{-1}\beta$ with new X matrix $\tilde{X} = XH$ where H is $p \times p$ and invertible and coefficients $\tilde{\beta} = H^{-1}\beta$.
- obtain the posterior for $\tilde{\beta}$ using Y and $\tilde{X}$

$$Y \mid \tilde{\beta}, \phi \sim \mathsf{N}_n(\tilde{X}\tilde{\beta}, \phi^{-1} I_n)$$

- since $\tilde{X}\tilde{\beta} = XH\tilde{\beta} = X\beta$ invariance suggests that the posterior for β and $H\tilde{\beta}$ should be the same
- plus the posterior of $H^{-1}\beta$ and $\tilde{\beta}$ should be the same

Exercise for the Energetic Student

With some linear algebra, show that this is true for a normal prior if $b_0 = 0$ and Φ_0 is kX^TX for some k

Zellner's g-prior

- Popular choice is to take $k = \phi/g$ which is a special case of Zellner's g-prior

$$\beta \mid \phi, g \sim \mathsf{N}\left(\mathbf{b}_0, \frac{g}{\phi}(X^TX)^{-1}\right)$$

- Full conditional

$$\beta \mid \phi, g \sim \mathsf{N}\left(\frac{g}{1+g}\hat{\beta} + \frac{1}{1+g}\mathbf{b}_0, \frac{1}{\phi}\frac{g}{1+g}(X^TX)^{-1}\right)$$

- one parameter g controls shrinkage
- invariance under linear transformations of $\mathbf{X}$ with $\mathbf{b}_0 = 0$ or transform mean $\tilde{\mathbf{b}}_0 = H^{-1}\mathbf{b}_0$
- often paired with the Jeffereys' reference prior for ϕ
- allows an informative mean, but keeps the same correlation structure as the MLE

Zellner's Blocked g-prior

- Zellner also realized that different blocks might have different degrees of prior information
- Two blocks $\mathbf{X}_1$ and $\mathbf{X}_2$ with $\mathbf{X}_1^T\mathbf{X}_2 = 0$ so Fisher Information is block diagonal
- Model $\mathbf{Y} = \mathbf{X}_1\boldsymbol{\alpha} + \mathbf{X}_2\boldsymbol{\beta} + \boldsymbol{\epsilon}$
- Priors

$$\boldsymbol{\alpha} \mid \phi \sim \mathsf{N}(\boldsymbol{\alpha}_1, \frac{g_{\boldsymbol{\alpha}}}{\phi}(\mathbf{X}_1^T\mathbf{X}_1)^{-1})$$

$$\boldsymbol{\beta} \mid \phi \sim \mathsf{N}(\mathbf{b}_0, \frac{g_{\boldsymbol{\beta}}}{\phi}(\mathbf{X}_2^T\mathbf{X}_2)^{-1})$$

- Important case $\mathbf{X}_1 = \mathbf{1}_n$ corresponding to intercept with limiting case $g_{\boldsymbol{\alpha}} \to \infty$

$$p(\boldsymbol{\alpha}) \propto 1$$

Potential Problems

- The posterior in Jeffereys' prior(s), the unit information prior, and Zellner's g-priors depend on $(\mathbf{X}^T\mathbf{X})^{-1}$ and the MLE $\hat{\boldsymbol{\beta}}$
- If $\mathbf{X}^T\mathbf{X} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$ is nearly singular ($\lambda_j \approx 0$ for one or more eigenvalues), certain elements of β or (linear combinations of β) may have huge posterior variances and the MLEs (and posterior means) are highly unstable!
- there is no unique posterior distribution if any $\lambda_j = 0$! ($p > n$ or non-full rank)
- Posterior Precision and Mean in conjugate prior

$$\begin{aligned}\boldsymbol{\Phi}_n &= \mathbf{X}^T\mathbf{X} + \boldsymbol{\Phi}_0 \\ \mathbf{b}_n &= \boldsymbol{\Phi}^{-1}(\mathbf{X}^T\mathbf{Y} + \boldsymbol{\Phi}_0\mathbf{b}_0)\end{aligned}$$

- Need a proper prior with $\boldsymbol{\Phi}_0 > 0$ (OK for $\mathbf{b}_0 = 0$)
- Simplest case: take $\boldsymbol{\Phi}_0 = \kappa\mathbf{I}_p$ so that $\boldsymbol{\Phi}_n = \mathbf{X}^T\mathbf{X} + \kappa\mathbf{I}_p = \mathbf{U}(\boldsymbol{\Lambda} + \kappa\mathbf{I}_p)\mathbf{U}^T > 0$

Ridge Regression

Model: $\mathbf{Y} = \mathbf{1}_n\alpha + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$

- WLOG assume that $\mathbf{X}$ has been centered and scaled so that $\mathbf{X}^T\mathbf{X} = \text{corr}(\mathbf{X})$
- typically expect the intercept α to be a different order of magnitude from the other predictors.
  - Adopt a two block prior with $p(\alpha) \propto 1$
  - If $\mathbf{X}$ is centered, $\mathbf{1}_n^T\mathbf{X} = \mathbf{0}_p$
- Prior $\boldsymbol{\beta} \mid \phi \sim \text{N}(\mathbf{0}_b, \frac{1}{\phi\kappa}\mathbf{I}_p)$ implies the $\mathbf{b}$ are exchangable *a priori* (i.e. distribution is invariant under permuting the labels and with a common scale and mean)
  - if different predictors have different variances, rescale $\mathbf{X}$ to have variance 1
- Posterior for $\boldsymbol{\beta}$

$$\boldsymbol{\beta} \mid \phi, \kappa, \mathbf{Y} \sim \text{N}\left((\kappa I_p + X^TX)^{-1}X^TY, \frac{1}{\phi}(\kappa I_p + X^TX)^{-1}\right)$$

Bayes Ridge Regression

- Posterior mean (or mode) given κ is biased, but can show that there **always** is a value of κ where the frequentist's expected squared error loss is smaller for the Ridge estimator than MLE!
- Unfortunately the optimal choice depends on "true" $\boldsymbol{\beta}$!
- related to penalized maximum likelihood estimation

$$-\frac{\phi}{2}\left(\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \kappa\|\boldsymbol{\beta}\|^2\right)$$

- Choice of κ ?
  - Cross-validation (frequentist)
  - Empirical Bayes? (frequentist/Bayes)
  - fixed *a priori* Bayes (and how to choose)
- Should there be a common κ? Or a κ_j per variable? (or shared in a group?)

Mixture of Conjugate Priors

- can place a prior on κ or κ_j for fully Bayes
- similar issue for g in the g priors
- often improved robustness over fixed choices of hyperparameter
- may not have cloosed form posterior but sampling is still often easy!
- Examples: Bayesian Lasso, Double Laplace, Horseshoe prior, mixtures of g-priors