STAT 525 Lecture 17

October 26, 2017

1 Last time

$$\begin{aligned}
\boldsymbol{y} &\sim N\left(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \mathbb{I}\right) \\
\boldsymbol{\beta} &\sim N\left(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta\right)
\end{aligned}$$

$$\boldsymbol{\mu}_\beta = 0 \Rightarrow \begin{cases} \text{g-prior} & \boldsymbol{\Sigma}_\beta = g\sigma^2 \left[\boldsymbol{X}^T \boldsymbol{X}\right]^{-1} \\ \text{ridge} & \boldsymbol{\Sigma}_\beta = \sigma_\beta^2 \mathbb{I} \end{cases}$$

Let $\tau = \sigma^{-2}$ then

$$\begin{aligned}
\boldsymbol{y} &\sim N\left(\boldsymbol{X}\boldsymbol{\beta}, \tau^{-1} \mathbb{I}_n\right) \\
\boldsymbol{\beta} &\sim N\left(\boldsymbol{\mu}_\beta, \tau^{-1}\boldsymbol{\Sigma}_\beta\right) \\
\tau &\sim \text{Gamma}\left(\alpha_\tau, \beta_\tau\right)
\end{aligned}$$

Then the posterior would be

$$\begin{aligned}
p\left(\tau \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta}\right) &\propto p\left(\tau, \boldsymbol{\beta}\right) p\left(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta}, \tau\right) \\
&= p\left(\tau\right) p\left(\boldsymbol{\beta} \mid \tau\right) p\left(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta}, \tau\right) \\
&\propto \tau^{\alpha_\tau - 1} \exp\left(-\beta_\tau \tau\right) \left|2\pi\tau^{-1}\mathbb{I}_n\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\tau \left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right\|^2\right) \times \\
&\left|2\pi\tau^{-1}\boldsymbol{\Sigma}_\beta\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\left(\boldsymbol{\beta} - \boldsymbol{\mu}_\beta\right)^T \left[\tau^{-1}\boldsymbol{\Sigma}_\beta\right]^{-1} \left(\boldsymbol{\beta} - \boldsymbol{\mu}_\beta\right)\right) \\
&\propto \tau^{\alpha_\tau + \frac{n}{2} + \frac{p}{2} - 1} \exp\left(-\tau\left[\beta_\tau + \frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right\|^2 + \frac{1}{2}\left(\boldsymbol{\beta} - \boldsymbol{\mu}_\beta\right)^T \left[\boldsymbol{\Sigma}_\beta\right]^{-1} \left(\boldsymbol{\beta} - \boldsymbol{\mu}_\beta\right)\right]\right)
\end{aligned}$$

Therefore,

$$\begin{aligned}
\tilde{\alpha}_\tau &= \alpha_\tau + \frac{n}{2} + \frac{p}{2} \\
\tilde{\beta}_\tau &= \beta_\tau + \frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right\|^2 + \frac{1}{2}\left(\boldsymbol{\beta} - \boldsymbol{\mu}_\beta\right)^T \left[\boldsymbol{\Sigma}_\beta\right]^{-1} \left(\boldsymbol{\beta} - \boldsymbol{\mu}_\beta\right)
\end{aligned}$$

Instead if we don't have τ in the prior on beta then

$$\boldsymbol{\beta} \sim N\left(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta\right)$$

$$\begin{aligned}\tilde{\alpha}_\tau &= \alpha_\tau + \frac{n}{2} \\ \tilde{\beta}_\tau &= \beta_\tau + \frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{X\beta}\right\|^2\end{aligned}$$

If we have the following g prior

$$\begin{aligned}\boldsymbol{\mu}_\beta &= 0 \\ \boldsymbol{\Sigma}_\beta &= g\sigma^2\left[\boldsymbol{X}^T\boldsymbol{X}\right]^{-1}\end{aligned}$$

After marginizing $\boldsymbol{\beta}$ out, we have

$$[\tau \mid \boldsymbol{y}, \boldsymbol{X}] \sim \text{Gamma}\left(\alpha_\tau + \frac{n}{2}, \beta_\tau + SSR_g\right)$$

where

$$\begin{aligned}SSR_g &= \boldsymbol{y}^T\left[\mathbb{I}_n - \frac{g}{g+1}\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}\right]\boldsymbol{y} \\ &= \|y\|^2 - \frac{g}{g+1}\boldsymbol{y}^T\left[\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{ols}}\right]\end{aligned}$$

2 Prediction

Given new $\tilde{\boldsymbol{X}}$ to predict $\tilde{\boldsymbol{y}}$ or $p\left(\tilde{\boldsymbol{y}} \mid \boldsymbol{y}\right)$

Sources of uncertainty

1. Model variability σ^2 (not accounted for by $\boldsymbol{X\beta}$)
2. Posterior uncertainty in $p\left(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y}\right)$ (due to finite sample size)

$$\left[\tilde{\boldsymbol{y}} \mid \tilde{\boldsymbol{X}}, \beta, \sigma^2\right] \sim N\left(\tilde{\boldsymbol{X}}\boldsymbol{\beta}, \sigma^2\mathbb{I}\right)$$

where we simulate $\boldsymbol{\beta}$, σ and $\tilde{\boldsymbol{y}}$.

3 Bayesian Robustness

Instead of $\epsilon_i \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right)$ try $\epsilon_i \overset{\text{iid}}{\sim} t_\upsilon\left(0, \sigma^2\right)$ where ν is the degree of freedom. Can be augmented as

$$\begin{aligned}\epsilon_i \mid \xi_i &\overset{\text{iid}}{\sim} N\left(0, \sigma^2/\xi_i\right) \\ \xi_i &\overset{\text{iid}}{\sim} \text{Gamma}\left(\frac{\nu}{2}, \frac{\nu}{2}\right)\end{aligned}$$

(Scaled Mixture of Gaussian)

1. (Asymmertic) Laplace
2. Skew normal
3. Discrete mixtures of Gaussian

4 Shrinkage and penalized regression

- Many predictors ($p \gg n$) but may be unrelated to $\boldsymbol{y}$.
- Including unnecessary predictors. Can cause poor performance.
- Nice to represent a small sets of predictors.

Key trade off for several methods

- Discrete (or two groups model)
$$p\left(\beta_j = 0\right) > 0$$
  - Can select $\{j : \beta_j \neq 0\}$
  - Problem: computationally feasible for moderate p
- Continuous (One group)
  - no true zero but small $|\beta_j| \approx 0$
  - Scalable

Penalized Regression

Setting:
$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \mathcal{L}\left(\boldsymbol{y}, \boldsymbol{X}\boldsymbol{\beta}\right) + \lambda P\left(\boldsymbol{\beta}\right)$$

where $\mathcal{L}\left(\boldsymbol{y}, \boldsymbol{X}\boldsymbol{\beta}\right)$ is the loss function (corresponding to (-log) likelihood) and $P\left(\boldsymbol{\beta}\right)$ is penalty (corresponding to (-log) prior). λcontrol the trade off (corresponding to prior precision)

$$\lambda \to 0 \mathcal{L} \text{ dominates}$$
$$\lambda \to \infty P \text{ dominates}$$

4.1 Lasso Regression

$$\hat{\boldsymbol{\beta}}_L = \arg\min_{\boldsymbol{\beta}} \mathcal{L}\left(\boldsymbol{y}, \boldsymbol{X}\boldsymbol{\beta}\right) + \lambda \sum_{j=1}^{p} |\beta_j|$$

1. unpenalized intercept ($\bar{\boldsymbol{y}}$ centered)
2. variable X should be on the same scale
3. penalized MLE (post mode) produces sparse solution.

4.2 Bayesian Lasso (Park & Casella 2008)

$$\left[\beta_j \mid \sigma^2, \lambda\right] \overset{\text{iid}}{\sim} DE\left(\lambda/\sigma\right)$$

where
$$p\left(\beta_j\right) = \frac{\lambda}{2\sigma} \exp\left(-\lambda \left|\beta_j\right| / \sigma\right)$$

notice that

$$
\begin{aligned}
-\log P(\boldsymbol{\beta}) &= -\sum_{j=1}^{p} \log\left(p\left(\beta_j\right)\right) \\
&= A^{-1} \sum_{j=1}^{p} |\beta_j|
\end{aligned}
$$

which is equivalent to

$$
\begin{gathered}
[\beta_j \mid \sigma, \eta_j] \overset{\text{indep}}{\sim} N\left(0, \sigma^2/\eta_j\right) \\
\left[\eta_j^{-1} \mid \lambda\right] \overset{\text{indep}}{\sim} Exp\left(\lambda^2/2\right) \\
\boldsymbol{\beta} \sim N\left(0, \boldsymbol{\Sigma}_\beta\right)
\end{gathered}
$$

where

$$
\boldsymbol{\Sigma}_\beta = \begin{bmatrix} \sigma^2/\eta_1 & & \\ & \sigma^2/\eta_2 & \\ & & \\ & & \sigma^2/\eta_p \end{bmatrix}
$$

- η_j is inverse-Gaussian
- do not get true zeros in $\boldsymbol{\beta}$
- but do get SEs (even when $|\beta_j| \approx 0$)

4.3 Horseshoe prior

$$
\begin{aligned}
\left[\beta_j \mid \sigma^2, \lambda_j^2\right] &\overset{\text{indep}}{\sim} N\left(-, \sigma^2\lambda_j^2\right) \\
\lambda_j &\overset{\text{indep}}{\sim} C^+(0, A)
\end{aligned}
$$

Priors for A:

- $A \sim C^+(0,1)$
- $A \sim \text{Uniform}(0,1)$

$$
\begin{aligned}
\boldsymbol{\beta} &\sim N\left(0, \boldsymbol{\Sigma}_\beta\right) \\
\boldsymbol{\Sigma}_\beta &= \text{diag}\left(\sigma^2\lambda_1, \sigma^2\lambda_2, \cdots, \sigma^2\lambda_p\right)
\end{aligned}
$$

Comments

- Many priors in this family: global local priors

  $$\beta_j \sim N\left(0, \sigma^2 A^2 \lambda_j^2\right)$$

  where $p(A, \lambda_1, \cdots, \lambda_p)$ determines behavior of $\boldsymbol{\beta}$.
- Do we need $\beta_j = 0$ or is $|\beta_j| \approx 0$ good enough?
- Threshold procedures:

- Recall
$$\kappa_j = \frac{1}{1+\lambda_j^2} \sim \text{Beta}\left(\frac{1}{2}, \frac{1}{2}\right)$$
is the amount of shrinkage to zero
- Reasonable:
$$\hat{b}_j = \mathbb{I}\left(\kappa_j < \frac{1}{2}\right)\hat{\beta}_j$$
where $\hat{\beta}_j$ is the posterior mean.