STA 365: Applied Bayesian Statistics

Boris Babic
Assistant Professor, University of Toronto

Week 9A: Bayesian Regression

Bayesian Regression

The multiple linear regression model is

$$Y_i \sim \mathrm{N}\left(\beta_0 + \sum_{j=1}^{p} X_{i,j}\beta_j, \sigma^2\right),$$

for $i = 1, \ldots, n$. Y_i are independently across the n observations.

Bayesian Regression

The least squares estimate of $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is

$$\hat{\beta}_{\mathrm{OLS}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (Y_i - \mu_i)^2,$$

where $\mu_i = \beta_0 + X_{i,1}\beta_1 +, \ldots, +X_{i,p}\beta_p$.

$\hat{\beta}_{\mathrm{OLS}}$ is unbiased even if the errors are non-Gaussian.

If the errors are Gaussian then the likelihood is proportional to

$$\prod_{i=1}^{n} \exp\left\{-\frac{(Y_i - \mu_i)^2}{2\sigma^2}\right\} = \exp\left\{-\frac{\sum_{i=1}^{n}(Y_i - \mu_i)^2}{2\sigma^2}\right\}.$$

Therefore, if the errors are Gaussian $\hat{\beta}_{\mathrm{OLS}}$ is also the MLE.

Bayesian Regression

Linear regression is often simpler to describe using linear algebra notation.

Let $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ be the response vector and $\mathbf{X}$ be the $n \times (p+1)$ matrix of covariates.

Then the mean of $\mathbf{Y}$ is $\mathbf{X}\boldsymbol{\beta}$ and the least squares solution is

$$\boldsymbol{\beta}_{\mathrm{OLS}} = \arg\min_{\boldsymbol{\beta}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}.$$

If the errors are Gaussian then the sampling distribution is

$$\hat{\boldsymbol{\beta}}_{\mathrm{OLS}} \sim \mathrm{N}\left[\boldsymbol{\beta}, \sigma^2 (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\right].$$

If the variance σ^2 is estimated using the mean squared residual error then the sampling distribution is multivariate t.

Bayesian Regression

- The Jeffrey's prior is flat $\pi(\boldsymbol{\beta}) \propto 1$.
- This is improper, but the posterior is proper under the same conditions required by least squares.
- If σ is known then

$$\boldsymbol{\beta} \mid \mathbf{Y} \sim \mathrm{N}\left[\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}, \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\right].$$

Bayesian Regression

We rarely know σ^2 in practice.

The Jeffreys prior for $(\boldsymbol{\beta}, \sigma^2)$ is

$$\pi(\boldsymbol{\beta}, \sigma^2) \propto 1/\sigma^2,$$

which is the limit case of an inverse gamma distribution with shape and rate parameters approaching zero.

Then the posterior of $\boldsymbol{\beta}$ follows a multivariate t centered on $\hat{\boldsymbol{\beta}}_{\text{OLS}}$.

Bayesian Regression

If there are many covariates or the covariates are collinear, then $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ is unstable.

Independent priors can counteract collinearity

$$\beta_j \sim \text{N}(0, \sigma^2/g)$$

independent over j.

The posterior mode is

$$\arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (Y_i - \mu_i)^2 + g \sum_{j=1}^{p} \beta_j^2.$$

In classical statistics, this is known as the ridge regression solution and is used to stabilize the least squares solution.

Bayesian Regression

An increasingly-popular prior is the double exponential or Bayesian LASSO prior

The prior is $\beta_j \sim \mathrm{DE}(\tau^2)$ which has the probability density function

$$\pi(\beta_j) \propto \exp\left(-\frac{|\beta|}{\tau^2}\right).$$

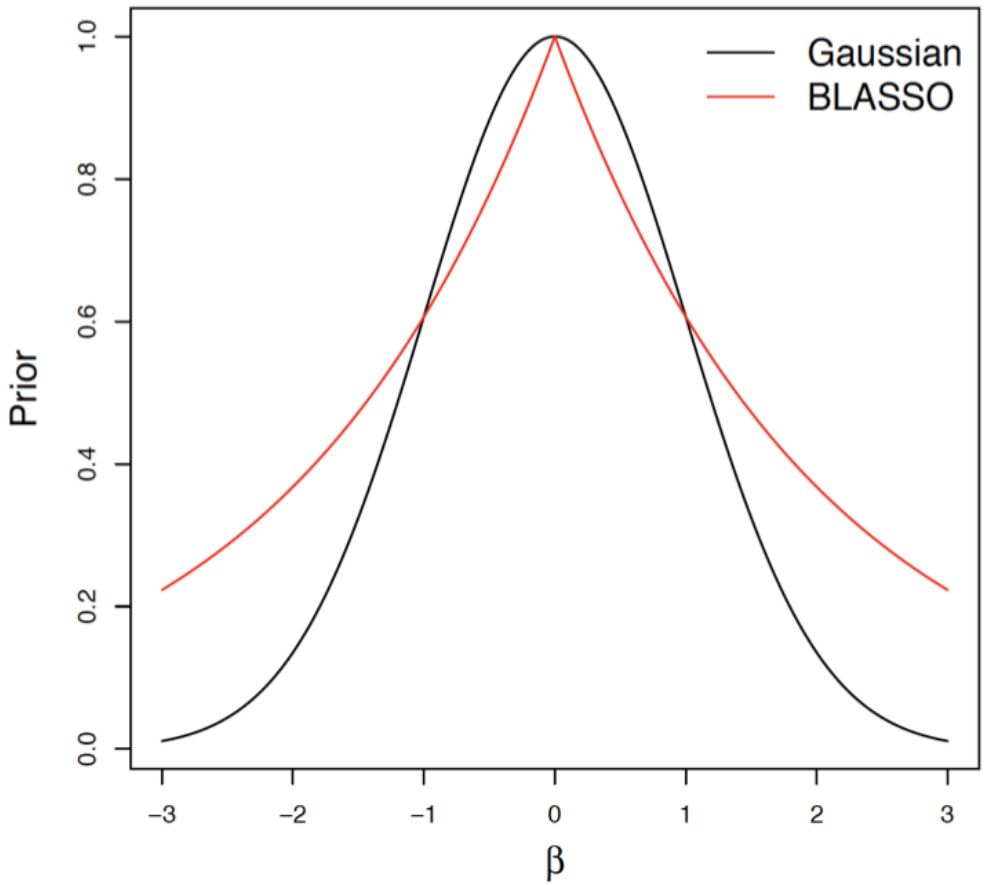
The square in the Gaussian prior is replaced with an absolute value

The shape of the PDF is thus more peaked at zero

The BLASSO prior favors settings where there are many β_j near zero and a few large β_j.

That is, p is large but most of the covariates are noise.

Bayesian Regression



Bayesian Regression

The posterior model is

$$\arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (Y_i - \mu_i)^2 + \tau^{-2} \sum_{j=1}^{p} |\beta_j|.$$

In classical statistics, this is known as the LASSO solution

It is popular because it adds stability by shrinking estimates towards zero, and also sets some coefficient to zero

Covariates with coefficients set to zero and can be excluded from the model.

LASSO performs variable selection and estimation simultaneously.

Bayesian Regression

Mixture Prior

$$\beta_j \mid \gamma_j \sim (1-\gamma_j)\mathrm{N}(0, c_0^2) + \gamma_j \mathrm{N}(0, c_1^2).$$

$$\gamma_j \sim \mathrm{Bernoulli}(\pi).$$

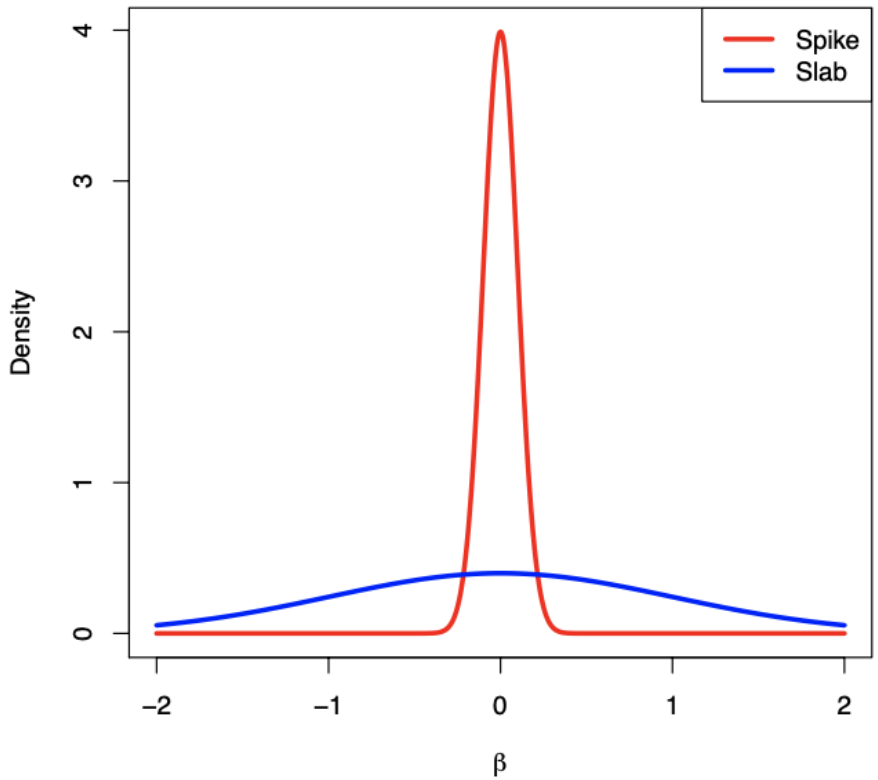
The constant c_0^2 is small, so that if $\gamma_j = 0$, "β_j could be safely estimated by 0".

The constant c_1^2 is large, so that if $\gamma_j = 1$, "a non-zero estimate of β_j should probably be included in the final model".

It works well for computing the marginal inclusion probability of each covariate and for model averaging

This model is computationally convenient and extremely flexible

Bayesian Regression



Bayesian Regression

- With flat or Gaussian (with fixed prior variance) priors the posterior is available in closed-form and Monte Carlo sampling is not needed
- With Gaussian priors all full conditionals are Gaussian or inverse gamma, and so Gibbs sampling is simple and fast
- With the BLASSO prior the full conditionals are more complicated
  - There is a trick to make all full conditional conjugate so that Gibbs sampling can be used
  - Metropolis sampling works fine too
- With the Spike Slab prior the full conditionals are available
- JAGS can handle all of them

Bayesian Regression

Say we have a new covariate vector $\mathbf{X}_{\text{new}}$ and we would like to predict the corresponding response Y_{new}.

A plug-in approach would fix $\boldsymbol{\beta}$ and σ at their posterior means $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}$ to make predictions

$$Y_{\text{new}} \mid \boldsymbol{\beta}, \hat{\sigma}^2 \sim \text{N}(\mathbf{X}_{\text{new}}\hat{\boldsymbol{\beta}}, \hat{\sigma}^2).$$

However, this plug-in approach suppresses uncertainty about $\boldsymbol{\beta}$ and σ^2.

Therefore these prediction intervals will be slightly too narrow leading to under coverage.

Bayesian Regression

- We should really account for all uncertainty when making predictions, including our uncertainty about $\boldsymbol{\beta}$ and σ^2.
- We really want to PPD

$$\begin{aligned} \pi(Y_{\text{new}} \mid \mathbf{Y}) &= \int \pi(Y_{\text{new}}, \boldsymbol{\beta}, \sigma^2 \mid \mathbf{Y}) d\boldsymbol{\beta} d\sigma^2 \\ &= \int \pi(Y_{\text{new}} \mid \boldsymbol{\beta}, \sigma^2) \pi(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{Y}) d\boldsymbol{\beta} d\sigma^2 \end{aligned}$$

- Marginalizing over the model parameters accounts for their uncertainty

Bayesian Regression

MCMC naturally gives draws from Y_{new}'s PPD

- For MCMC iteration t we have $\boldsymbol{\beta}^{(t)}$ and $\sigma^{2(t)}$.
- For MCMC iteration t we sample

$$Y_{\text{new}}^{(t)} \sim \text{N}\left(\mathbf{X}\boldsymbol{\beta}^{(t)}, \sigma^{2(t)}\right).$$

- $Y_{\text{new}}^{(1)}, \ldots, Y_{\text{new}}^{(S)}$ are samples from the PPD.

Thus, "Bayesian methods" naturally quantify uncertainty.

JAGS can handle it.