

# (7) Bayesian linear regression

ST440/540: Applied Bayesian Statistics

Spring, 2018

# Bayesian linear regression

- ▶ Linear regression is by far the most common statistical model
- ▶ It includes as special cases the t-test and ANOVA
- ▶ The multiple linear regression model is

$$Y_i \sim \text{Normal}(\beta_0 + X_{i1}\beta_1 + \dots + X_{ip}\beta_p, \sigma^2)$$

independently across the  $i = 1, \dots, n$  observations

- ▶ As we'll see, Bayesian and classical linear regression are similar if  $n \gg p$  and the priors are uninformative.
- ▶ However, the results can be different for challenging problems, and the interpretation is different in all cases

## Review of least squares

- ▶ The least squares estimate of  $\beta = (\beta_0, \beta_1, \dots, \beta_p)^T$  is

$$\hat{\beta}_{OLS} = \operatorname{argmin}_{\beta} \sum_{i=1}^n (Y_i - \mu_i)^2$$

where  $\mu_i = \beta_0 + X_{i1}\beta_1 + \dots + X_{ip}\beta_p$

- ▶  $\hat{\beta}_{OLS}$  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  $\hat{\beta}_{OLS}$  is also the MLE

## Review of least squares

- ▶ Linear regression is often simpler to describe using linear algebra notation
- ▶ Let  $\mathbf{Y} = (Y_1, \dots, Y_n)^T$  be the response vector and  $\mathbf{X}$  be the  $n \times (\rho + 1)$  matrix of covariates
- ▶ Then the mean of  $\mathbf{Y}$  is  $\mathbf{X}\beta$  and the least squares solution is

$$\hat{\beta}_{OLS} = \underset{\beta}{\operatorname{argmin}} (\mathbf{Y} - \mathbf{X}\beta)^T (\mathbf{Y} - \mathbf{X}\beta) = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

- ▶ If the errors are Gaussian then the sampling distribution is

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

- ▶ If the variance  $\sigma^2$  is estimated using the mean squared residual error then the sampling distribution is multivariate t

# Bayesian regression

- ▶ The likelihood remains

$$Y_i \sim \text{Normal}(\beta_0 + X_{i1}\beta_1 + \dots + X_{ip}\beta_p, \sigma^2)$$

independent for  $i = 1, \dots, n$  observations

- ▶ As with a least squares analysis, it is crucial to verify this is appropriate using qq-plots, added variable plots, etc.
- ▶ A Bayesian analysis also requires priors for  $\beta$  and  $\sigma$
- ▶ We will focus on prior specification since this piece is uniquely Bayesian.

# Priors

- ▶ For the purpose of setting priors, it is helpful to standardize both the response and each covariate to have mean zero and variance one.
- ▶ Many priors for  $\beta$  have been considered:
  1. Improper priors
  2. Gaussian priors
  3. Double exponential priors
  4. Many, many more...

## Improper priors

- ▶ The Jeffreys' prior is flat  $p(\beta) = 1$
- ▶ This is improper, but the posterior is proper under the same conditions required by least squares
- ▶ If  $\sigma$  is known then

$$\beta | \mathbf{Y} \sim \text{Normal} \left[ \hat{\beta}_{OLS}, \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1} \right]$$

- ▶ See “Post beta” in <http://www4.stat.ncsu.edu/~reich/ABA/Derivations7.pdf>
- ▶ Therefore, the results should be similar to least squares
- ▶ How are they different?

## Improper priors

- ▶ Of course we rarely know  $\sigma$
- ▶ Typically the error variance follows an  $\text{InvGamma}(a, b)$  prior with  $a$  and  $b$  set to be small, say  $a = b = 0.01$ .
- ▶ In this case the posterior of  $\beta$  follows a multivariate  $t$  centered on  $\hat{\beta}_{OLS}$
- ▶ Again, the results are similar to OLS
- ▶ The objective Bayes Jeffreys prior for  $\theta = (\beta, \sigma)$  is

$$p(\beta, \sigma^2) = \frac{1}{\sigma^2}$$

which is the limit as  $a, b \rightarrow 0$



## Multivariate normal prior

- ▶ Another common prior for is Zellner's g-prior

$$\beta \sim \text{Normal} \left[ 0, \frac{\sigma^2}{g} (\mathbf{X}^T \mathbf{X})^{-1} \right]$$

- ▶ This prior is proper assuming  $\mathbf{X}$  is full rank
- ▶ The posterior mean is

$$\frac{1}{1+g} \hat{\beta}_{OLS}$$

- ▶ This shrinks the least estimate towards zero
- ▶  $g$  controls the amount of shrinkage
- ▶  $g = 1/n$  is common, and called the unit information prior

# Univariate Gaussian priors

- ▶ If there are many covariates or the covariates are collinear, then  $\hat{\beta}_{OLS}$  is unstable
- ▶ Independent priors can counteract collinearity

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

independent over  $j$

- ▶ The posterior mode is

$$\underset{\beta}{\operatorname{argmin}} \sum_{i=1}^n (Y_i - \mu_i)^2 + g \sum_{j=1}^p \beta_j^2$$

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

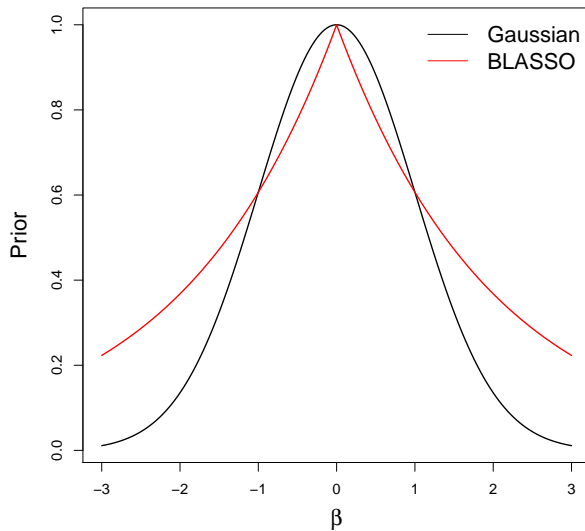
# BLASSO

- ▶ An increasingly-popular prior is the double exponential or Bayesian LASSO prior
- ▶ The prior is  $\beta_j \sim \text{DE}(\tau)$  which has PDF

$$f(\beta) \propto \exp\left(-\frac{|\beta|}{\tau}\right)$$

- ▶ The square in the Gaussian prior is replaced with an absolute value
- ▶ The shape of the PDF is thus more peaked at zero (next slide)
- ▶ The BLASSO prior favors settings where there are many  $\beta_j$  near zero and a few large  $\beta_j$
- ▶ That is,  $p$  is large but most of the covariates are noise

# BLASSO



# BLASSO

- ▶ The posterior mode is

$$\operatorname{argmin}_{\beta} \sum_{i=1}^n (Y_i - \mu_i)^2 + g \sum_{j=1}^p |\beta_j|$$

- ▶ 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 coefficients to zero
- ▶ Covariates with coefficients set to zero can be removed
- ▶ Therefore, LASSO performs variables selection and estimation simultaneously

# Computing

- ▶ With flat or Gaussian (with fixed prior variance) priors the posterior is available in closed-form and Monte Carlo sampling is not needed
- ▶ With normal priors all full conditionals are Gaussian or inverse gamma, and so Gibbs sampling is simple and fast
- ▶ JAGS works well, but there are R (and SAS and others) packages dedicated just to Bayesian linear regression that are preferred for big/hard problems
- ▶ BLR is probably the most common
- ▶ `http://www4.stat.ncsu.edu/~reich/ABA/code/regJAGS`

# Computing for the BLASSO

- ▶ For 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
- ▶ BLR works well for BLASSO and is super fast
- ▶ JAGS can handle this as well,
- ▶ `http://www4.stat.ncsu.edu/~reich/ABA/code/BLASSO`

## Summarizing the results

- ▶ The standard summary is a table with marginal means and 95% intervals for each  $\beta_j$
- ▶ This becomes unwieldy for large  $p$
- ▶ Picking a subset of covariates is a crucial step in a linear regression analysis.
- ▶ We will discuss this later in the course.
- ▶ Common methods include cross-validation, information criteria, and stochastic search.



# Logistic regression

- ▶ Other forms of regression follow naturally from linear regression
- ▶ For example, for binary responses  $Y_i \in \{0, 1\}$  we might use logistic regression

$$\text{logit}[\text{Prob}(Y_i = 1)] = \eta_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip}$$

- ▶ The logit link is the log-odds  $\text{logit}(x) = \log[x/(1 - x)]$
- ▶ Then  $\beta_j$  represents the increase in the log odds of an event corresponding to a one-unit increase in covariate  $j$
- ▶ The expit transformation  $\text{expit}(x) = \exp(x)/[1 + \exp(x)]$  is the inverse, and

$$\text{Prob}(Y_i = 1) = \text{expit}(\eta_i) \in [0, 1]$$

# Logistic regression

- ▶ Bayesian logistic regression requires a prior for  $\beta$
- ▶ All of the prior we have discussed for linear regression (Zellner, BLASSO, etc) apply
- ▶ Computationally the full conditional distributions are no longer conjugate and so we must use Metropolis sampling
- ▶ The R function `MCMClogit` does this efficiently
- ▶ It is fast in JAGS too, for example <http://www4.stat.ncsu.edu/~reich/ABA/code/GLM>

# Predictions

- ▶ Say we have a new covariate vector  $\mathbf{X}_{new}$  and we would like to predict the corresponding response  $Y_{new}$
- ▶ A plug-in approach would fix  $\beta$  and  $\sigma$  at their posterior means  $\hat{\beta}$  and  $\hat{\sigma}$  to make predictions

$$Y_{new} | \hat{\beta}, \hat{\sigma} \sim \text{Normal}(\mathbf{X}_{new} \hat{\beta}, \hat{\sigma}^2)$$

- ▶ However this plug-in approach suppresses uncertainty about  $\beta$  and  $\sigma$
- ▶ Therefore these prediction intervals will be slightly too narrow leading to undercoverage

## Posterior predictive distribution (PPD)

- ▶ We should really account for all uncertainty when making predictions, including our uncertainty about  $\beta$  and  $\sigma$
- ▶ We really want the PPD

$$\begin{aligned} p(Y_{new}|\mathbf{Y}) &= \int f(Y_{new}, \beta, \sigma | \mathbf{Y}) d\beta d\sigma \\ &= \int f(Y_{new} | \beta, \sigma) f(\beta, \sigma | \mathbf{Y}) d\beta d\sigma \end{aligned}$$

- ▶ Marginalizing over the model parameters accounts for their uncertainty
- ▶ The concept of the PPD applies generally (e.g., logistic regression) and means the distribution of the predicted value marginally over model parameters

# Posterior predictive distribution (PPD)

- ▶ MCMC naturally gives draws from  $Y_{new}$ 's PPD
  - ▶ For MCMC iteration  $t$  we have  $\beta^{(t)}$  and  $\sigma^{(t)}$
  - ▶ For MCMC iteration  $t$  we sample

$$Y_{new}^{(t)} \sim \text{Normal}(\mathbf{X}\beta^{(t)}, \sigma^{(t)2})$$

- ▶  $Y_{new}^{(1)}, \dots, Y_{new}^{(S)}$  are samples from the PPD
- ▶ This is an example of the claim that “Bayesian methods naturally quantify uncertainty”
- ▶ `http://www4.stat.ncsu.edu/~reich/ABA/code/Predict`