# Assignment 1: AMS 268 (Due Date 2/10)

January 23, 2016

Consider the high dimensional linear regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon, \;\; \epsilon \sim N(0, \sigma^2) \tag{1}$$

Let $\boldsymbol{x} = (x_1, ..., x_p)' \sim N(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a $p \times p$ positive definite matrix. Assume that we have observed a sample of size $n$, $(y_i, \boldsymbol{x}_i)_{i=1}^n$ and assume $\sigma^2 = 1$. Consider simulating data by taking various combinations of $(n, p, \boldsymbol{\Sigma}, \boldsymbol{\beta})$ as follows

(a) $n = 50, 500$

(b) $p = 100, 1000$

(c) $\boldsymbol{\Sigma} = \boldsymbol{I}, \boldsymbol{S}_{0.1}, \boldsymbol{S}_{0.6}$,

(d) (i) $\beta_1 = \cdots = \beta_5 = 3$, $\beta_j = 0$ for any other $j$; (ii) $\beta_1 = \cdots = \beta_5 = 5$, $\beta_6 = \cdots = \beta_{10} = -2$, $\beta_{11} = \cdots = \beta_{15} = 0.5$, $\beta_j = 0$ for any other $j$; (iii)$\beta_j = 1$ for all $j$,

where $\boldsymbol{S}_{\rho,ii} = 1, \boldsymbol{S}_{\rho,ij} = \rho^{|i-j|}$ for $i \neq j$. Altogether they give rise to 36 different combinations.

- Simulate data for all 36 combinations described as above.
- Run Lasso and Ridge regression for all 36 combinations and compare the results.
- Run Bayesian models with spike and slab, Bayesian Lasso and Generalized double pareto prior distributions on $\boldsymbol{\beta}$ respectively for all 36 combinations. (Write your own code)

- Numerically obtain $E(\beta_j|\boldsymbol{y})$ for all the competing Bayesian models for all $j$. Discuss accuracy of the Bayesian models w.r.t a metric.

- Compare lasso and spike and slab prior as methods for selecting variables.

- Let $L_j$ be the length of 95% credible interval for the $j$th predictor. Let $M_{zero} = mean(L_j : \beta_j^0 \neq 0)$ and $M_{zero} = mean(L_j : \beta_j^0 = 0)$ where $\beta_j^0$ is the true value of $\beta_j$. Calculate $M_{zero}$ and $M_{nonzero}$ for all the competing Bayesian shrinkage priors.

- Take a particular combination $n = 50$, $p = 100$, $\boldsymbol{S}_{0.6}$ and (i), out of the 36 combinations. Simulate 50 additional responses and predictors $(y_{pred,i}, \boldsymbol{x}_{pred,i})_{i=1}^{50}$ for this combination with (1). Draw 1000 samples from the posterior predictive distribution $\pi(y|y_1, ..., y_n, \boldsymbol{x}_{pred,i})$, $\forall \ i = 1, ..., 50$ for all the shrinkage priors. Calculate posterior predictive mean $y_{est,i}$ at every $\boldsymbol{x}_{pred,i}$ for Bayesian lasso and generalized double pareto. Calculate MSPE $= \frac{1}{50}\sum_{i=1}^{50}(y_{pred,i} - y_{est,i})^2$ for both shrinkage priors.