Assignment 2: AMS 268 (Due Date 2/21)

February 9, 2018

1. 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) $n = 50, 200$

(b) $p = 20$

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

(d) $\beta_1 = \cdots = \beta_5 = 3$, $\beta_j = 0$ for any other j where $\boldsymbol{S}_{\rho,ii} = 1, \boldsymbol{S}_{\rho,ij} = \rho^{|i-j|}$ for $i \neq j$.

- Simulate data for both combinations described as above. You already have the code from Hw 1.
- Run Bayesian high dimensional model with g prior with a fixed g of your choice from the classnotes. (Write your own code)
- Test the hypotheses (i) $H_0 : \beta_1 = 0$ (ii) $H_0 : \beta_{10} = 0$.

2. Simulate $x_{1i}, x_{2i}, \ldots, x_{pi} \sim N(0, 1)$ for $i = 1, ..., n$. Simulate the response from

$$y_i = 10 sin(\pi x_{1i} x_{2i}) + 20(x_{2i} - 0.5)_+^2 + 10 x_{4i} + \epsilon_i, \;\; \epsilon_i \sim N(0, \sigma^2).$$

Consider, $\sigma^2 = 0.5$ and fit Random Forest and BART when (i) $p = 200, n = 200$ and (ii) $n = 500, p = 100$. Use the same number trees for fitting BART and Random Forest. Simulate an additional 100 responses.

- Provide MSPE, length and coverage of 95% predictive intervals for both BART and Random Forest.
- Show the analysis for the number of trees equal to 10 and 500.
- Add noise from $N(0, 0.1)$ and add the noise to predictors x_1, x_2 and x_9. Repeat the analysis.