STA 5107: Final Project

Spring 2022/Due Date: April 28

Reversible Jump Markov Chain Monte Carlo Algorithm for Model Selection in Linear Regression

1. **Goal**: We are interested in solving for the regression coefficients in a standard linear model but the selection of predictors is not known. In other words, we are given a large number, say m, of predictors and we have to select an appropriate subset to obtain the optimal model. This is called the problem of **model selection**. We will restrict to a smaller problem where the optimal subset is simply the first n predictors, we just don't know what n is. (Note that this reduces the possible number of models from 2^m to m.)

2. **Problem Specification**: To be specific, we seek coefficients for the model

$$y = \sum_{i=1}^{n} x_i b_i + \epsilon \ ,$$

where $n < m$, x_is are the predictors, y is the response variable, and ϵ is the measurement noise. We are given k independent measurements, denoted in bold by $\mathbf{y}$, $\mathbf{X}$ and ϵ. We will seek a Bayesian solution to the joint estimation of $\{n, b_1, \ldots, b_n\}$.

To setup a Bayesian formulation we need to define a joint posterior density of the type:

$$f(n, \mathbf{b}_n|\mathbf{y}) \propto f(\mathbf{y}|n, \mathbf{b}_n) f(\mathbf{b}_n|n) f(n) \ .$$

We will use the notation $\mathbf{X}_n = \mathbf{X}(:, 1:n)$ and $\mathbf{b}_n = \{b_1, \ldots, b_n\}$. We will use the following terms:

- The likelihood function is given by: $f(\mathbf{y}|n, \mathbf{b}_n) = (\frac{1}{\sqrt{2\pi\sigma_0^2}})^k e^{\frac{-1}{2\sigma_0^2}\|\mathbf{y}-\mathbf{X}_n\mathbf{b}_n\|^2}$.
- The prior on $\mathbf{b_n}$ given n is: $f(\mathbf{b}_n|n) = (\frac{1}{\sqrt{2\pi\sigma_p^2}})^n e^{\frac{-1}{2\sigma_p^2}\|\mathbf{b}_n-\mu_b\|^2}$.
- The prior on n is simply uniform: $f(n) = \frac{1}{m}$.

3. **Sampling from the Posterior**: We will use an RJMCMC technique for sampling from the posterior. Here is the algorithm for implementing this algorithm: Let $(n, \mathbf{b}_n)$ be the current samples from the posterior.

   (a) Select a candidate number n^* from the probability $f(n)$.

   (b) If $n^* \geq n$, generate a random vector $\mathbf{u} \sim N(0, \sigma_r I_{n^*})$. The candidate coefficient vector is given by:

$$\mathbf{b}_{n^*} = \begin{bmatrix} \mathbf{b}_n \\ 0 \end{bmatrix} + \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix} \ , \quad \mathbf{u} = \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix} \ .$$

   Compute the likelihoods:

$$h_1(\mathbf{u}) = (\frac{1}{\sqrt{2\pi\sigma_r^2}})^{n^*} e^{\frac{-1}{2\sigma_r^2}\|\mathbf{u}\|^2}, \quad h_2(\mathbf{u}_1) = (\frac{1}{\sqrt{2\pi\sigma_r^2}})^{n} e^{\frac{-1}{2\sigma_r^2}\|\mathbf{u}_1\|^2} \ .$$

(c) If $n^* < n$, generate a random vector $\mathbf{u}_1 \sim N(0, \sigma_r I_{n^*})$. The candidate coefficient vector is given by:

$$\mathbf{b}_{n^*} = \mathbf{b}_n^1 + \mathbf{u}_1 \ , \quad \mathbf{b}_n = \begin{bmatrix} \mathbf{b}_n^1 \\ \mathbf{b}_n^2 \end{bmatrix} \ ,$$

and form $\mathbf{u} = \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{b}_n^2 \end{bmatrix}$.

Compute the likelihoods:

$$h_2(\mathbf{u}) = (\frac{1}{\sqrt{2\pi\sigma_r^2}})^n e^{\frac{-1}{2\sigma_r^2}\|\mathbf{u}\|^2}, \quad h_1(\mathbf{u}_1) = (\frac{1}{\sqrt{2\pi\sigma_r^2}})^{n^*} e^{\frac{-1}{2\sigma_r^2}\|\mathbf{u}_1\|^2} \ .$$

(d) Compute the acceptance-rejection function:

$$\begin{aligned} \rho &= \min\{1, \frac{f(n^*, \mathbf{b}_{n^*}|\mathbf{y})h_2}{f(n, \mathbf{b}_n|\mathbf{y})h_1}\} = \min\{1, \frac{f(\mathbf{y}|n^*, \mathbf{b}_{n^*})f(\mathbf{b}_{n^*}|n^*)h_2}{f(\mathbf{y}|n^*, \mathbf{b}_n)f(\mathbf{b}_n|n)h_1}\} \\ &= \min\{1, \frac{e^{-(E_1-E_2)}(2\pi\sigma_p^2)^{(n-n^*)/2} e^{\frac{-1}{2\sigma_p^2}(\|\mathbf{b}_{n^*}-\mu_b\|^2 - \|\mathbf{b}_n-\mu_b\|^2)} h_2}{h_1}\} \end{aligned}$$

where $E_1 = \frac{1}{2\sigma_0^2}\|\mathbf{y} - \mathbf{X}_{n^*}\mathbf{b}_{n^*}\|^2$ and $E_2 = \frac{1}{2\sigma_0^2}\|\mathbf{y} - \mathbf{X}_n\mathbf{b}_n\|^2$.

(e) If $U \sim U[0,1]$ is less than ρ then set $(n, \mathbf{b}_n) = (n^*, \mathbf{b}_{n^*})$. Else, return to Step (a).

4. **Experiment**: Simulate a dataset with the following code:

```
m = 10;
n0 = ceil(rand*m);
k = 10;
σ0 = 0.2;
σp = 0.3;
μb = 2*ones(n0,1);
b = μb + σp*randn(n0,1);
X = 5*randn(k,m);
y = X(:,1:n0)*b + σ0*randn(k,1);
```

For this data (y,X) implement the RJMCMC algorithm (with $\sigma_r = 0.2$) to sample from the posterior and using $N = 100,000$ samples from the posterior: Show a histogram of n values visited by the Markov chain. This estimates the posterior probability $f(n|\mathbf{y})$.

Repeat this experiment for 10 different realizations of (y, X) and analyze your results.

5. **Report**: Prepare a report for this project describing completely all parts of the project. Use the same format as in the report for the mid-term project.