STAT534:Statistical Computing
Homework 1

Po-An, Chen(Andy Chen)

March 27, 2022

Problem 1

Write a function in R that computes the logarithm of the determinant of a matrix R. The function call should be:

> logdet(R)

You can exploit the fact that the determinant of a matrix is given by the product of the eigenvalues of the matrix. You can use the function "eigen" to find these eigenvalues.

Answer:

The code would be like below.

```
# function define
logdet <- function(R){
  log(prod(eigen(R)$values))
}

# test matrix
R <- matrix(c(13, -4, 2, -4, 11, -2, 2, -2, 8), 3, 3, byrow=
   TRUE)
R
logdet(R) # 6.858565
```

Problem 2

The data file "erdata.txt" has $n = 158$ samples (rows) and $p = 51$ variables (columns). The first variable (column 1) is your response variable (levels of a probe corresponding with the estrogen receptor transcription factor). The remaining 50 columns represent the expression levels of 50 genes that are highly correlated with the response. All these variables are continuous.

You need to write a function in R that computes the logarithm of the marginal likelihood associated with the regression of the response on a subset of explanatory variables $A \subset \{2, ..., p\}$. The function call should be:

> logmarglik(data, A)

This assumes that "data" represents the $n \times p$ data matrix. The marginal likelihood is defined by formula (3) below. You essentially need to calculate $\log(p(D_1, D_A|[1|A]))$.

You could use the function "lgamma" to calculate the log of the Γ function, the function "solve" to calculate the inverse of a matrix and the function "logdet" from Problem 1 to compute the logarithm of the determinant of a matrix.

For example, the logarithm of the marginal likelihood of the regression of the response $Y = X_1$ on the explanatory variables X_2, X_5 and X_{10}, namely

$$y = \beta_2 x_2 + \beta_5 x_5 + \beta_{10} x_{10} + \epsilon,$$

is obtained using the call

> logmarglik(data, c(2, 5, 10))

You need to explicitly calculate this number in order to complete your assignment.

Answer:

The code is like below, I use the "inv" function in the "matlib" package in R to calculate inverse matrix. The answer I get is -59.97883.

```
1  library(matlib)  # in order to calculate the inverse matrix
2  data <- read.table('erdata.txt', head = FALSE)
3  A <- c(2, 5, 10)
4
5  logmarglik <- function(data, A){
6    # A is a vector
7    M_A <- diag(length(A)) + t(data[A])%*%as.matrix(data[A])
8    ans <- lgamma((nrow(data) + length(A) + 2)/2) -
9           lgamma((length(A) + 2)/2) -
10          0.5*logdet(M_A) -
11          ((nrow(data) + length(A) + 2)/2)*log(1 + t(data[1])%
              *%as.matrix(data[1]) - t(data[1])%*%as.matrix(data
              [A])%*%inv(M_A)%*%t(data[A])%*%as.matrix(data[1]))
12   return(ans)
13 }
14
15 logmarglik(data, A) # minus 59.97883
```

1 Bayesian Inference in Normal Linear Regression

Let $Y = X_1$ be a continuous response variable and $X_{(2:p)} = (X_2, ..., X_p)$ be the vector of explanatory variables. Denote by D_1 the first column of the $n \times p$ data matrix D, by $D_{(2:p)}$ the columns $(2 : p) = \{2, ..., p\}$ of D. The linear regression containing all the explanatory variables X_1 is denoted by $[1|(2 : p)]$ and has coefficients $\beta_{(2:p)} = (\beta_2, ..., \beta_p)$. We center and scale the observed variables such that their sample means are zero and their sample standard deviations are one. As such, there is no need for an intercept parameter.
The full regression $[1|(2 : p)]$ is given by

$$p(Y|X_{(2:p)} = x_{(2:p)}, \beta_{(2:p)}) = N\left(\sum_{j=2}^{p} \beta_j x_j, \sigma^2\right). \tag{1}$$

Now assume that only some explanatory variables $X_A,\ A \subset (2 : p)$, are present in the linear regression (1). That is, we set to zero the regression coefficients associated with the rest of the explanatory variables:

$$\beta_j = 0,\ j \in (2 : p) \subset A.$$

We denote with $[1|A]$ the regression that involves only the explanatory variables X_A. Moreover, we denote by $|A|$ the number of elements of A and by D_A the columns of D indexed by A. The equation associated with regression $[1|A]$ is

$$p(Y|X_A = x_A, \beta_A) = N\left(\sum_{j\in A} \beta_j x_j, \sigma^2\right). \tag{2}$$

We consider the following Bayesian specification of the regression model (2). The prior for σ^2 is $p(\sigma^2) =$ inverse-Gamma$((|A| + 2)/2, 1/2$ and, conditional on σ^2, the regression coefficients β_A have independent priors $p(\beta_j) = N(0, \sigma^2),\ j \in A$). The corresponding posterior distributions are

$$p(\sigma^2|D_1, D_A) = \text{inverse-Gamma}((n + |A| + 2)/2, (1 + D_1^T D_1 - D_1^T D_A M_A^{-1} D_A^T D_1)),$$
$$p(\beta_A|\sigma^2, D_1, D_A) = N_{|A|}(M_A^{-1} D_A^T D_1, \sigma^2 M_A^{-1}),$$

where $M_A = I_{|A|} + D_A^T D_A$. The marginal likelihood of $[1|A]$ therefore given by

$$p(D_1, D_A|[1|A]) = \frac{\Gamma((n + |A| + 2)/2)}{\Gamma((|A| + 2)/2)} (\det M_A)^{-1/2} (1 + D_1^T D_1 - D_1^T D_A M_A^{-1} D_A^T D_1)^{-(n+|A|+2)/2} \tag{3}$$