assigment

abc

11/29/2019

Question 1

a)

Recall that the *Fisher information* is defined as:

$$I(\theta) = E_\theta[(\frac{\partial ln[L(\theta;x)]}{\partial\theta})^2] = -E_\theta(\frac{\partial^2 ln[L(\theta;x)]}{\partial\theta^2})$$

In our regression model, $\epsilon \sim N(0,\sigma^2)$ which means we want to find the *Jeffreys prior* for the normal variance with known mean.

For known mean equal to 0 and unknown variance σ^2, we have log-likelihood

$$f(\epsilon|\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}exp(-\frac{\epsilon^2}{2\sigma^2})$$

$$L(\sigma;\epsilon) = \prod_{i=1}^{n} f(x_i|\sigma) = (2\pi\sigma^2)^{-n/2}exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2)$$

$$l(\sigma;\epsilon) = ln(L(\sigma;\epsilon) = -\frac{n}{2}ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\epsilon_i^2$$

Then we can calculate the fisher information for σ is

$$\begin{aligned}
I(\sigma) &= -E_\sigma(\frac{\partial^2 ln[L(\sigma;\epsilon)]}{\partial\sigma^2}) \\
&= -E_\sigma(\frac{\partial^2 - \frac{n}{2}ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\epsilon_i^2}{\partial\sigma^2}) \\
&= -E_\sigma(\frac{\partial}{\partial\sigma}(\frac{n}{\sigma} + \sigma^{-3}\sum_{i=1}^{n}\epsilon_i^2)) \\
&= -E_\sigma(-n\sigma^{-2} - 3\sigma^{-4}\sum_{i=1}^{n}\epsilon_i^2)
\end{aligned}$$

Because $var(\epsilon) = E(\epsilon^2) - E^2(\epsilon) = \sigma^2$ and $E(\epsilon) = \mu = 0$
so $E[\sum_{i=1}^{n}\epsilon_i^2)] = \sum_{i=1}^{n}(E[\epsilon_i^2]) = n(var(\epsilon) + 0^2) = n\sigma^2$ Then

$$\begin{aligned}
I(\sigma) &= -E_\sigma(-n\sigma^{-2} - 3\sigma^{-4}\sum_{i=1}^{n}\epsilon_i^2) \\
&= n\sigma^{-2} + 3\sigma^{-4}n\sigma^2 \\
&= \frac{4n}{\sigma^2}
\end{aligned}$$

So the *Jeffreys prior* is

$$\pi(\sigma) = \sqrt{I(\sigma)} = \frac{2\sqrt{n}}{\sigma} \propto \frac{1}{\sigma}$$

b)

Suppose that X is full rank: $rank(X) = k$

And the $\beta|\sigma^2, X \sim N_{k+1}(\beta_0, g\sigma^2(X^TX)^{-1})$

The liklihood of ordinary normal linear model $y|\beta, \sigma^2, X \sim N_n(X\beta, \sigma^2 I_n)$ is

$$l(\beta, \sigma^2|y, X) = (2\pi\sigma^2)^{-n/2} exp[-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)]$$

Also the MLE of β is the solution of the least square problem and defined as

$$\hat{\beta} = (X^TX)^{-1}X^Ty$$

Unbiased estimator of σ^2 is

$$s^2 = (y - X\hat{\beta})^T(y - X\hat{\beta})$$

$$\hat{\sigma}^2 = \frac{1}{n-k-1}(y - X\hat{\beta})^T(y - X\hat{\beta})) = \frac{s^2}{n-k-1}$$

Since the design matrix X is known and fixed, the g is constant, follow the process in question (a) we can directly conclude that the *g-prior* is

$$\pi(\sigma^2|X) \propto \frac{1}{\sigma^2}$$

Because the X^TX is used in both prior and likelihood, then joint posterior distribution can be simplified into

$$\begin{aligned}
&p(\beta, \sigma^2|y, X) \\
&= \pi(\beta)\pi(\sigma^2)\prod_{i=1}^{n} f(y_i, x_i, \sigma^2, \beta) \\
&\propto \frac{1}{\sqrt{2\pi\sigma^2(X^TX)^{-1}g}} exp(-\frac{1}{2\sigma^2 gX^TX}(\beta - \beta_0)^T X^TX(\beta - \beta_0)) * \frac{1}{\sigma^2} * \\
&(\frac{1}{\sqrt{2\pi\sigma^2}})^n exp(-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)) \\
&\propto (\sigma^2)^{-(\frac{n}{2}+1+\frac{k+1}{2})} exp(-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta) - \frac{1}{2\sigma^2}(\beta - \beta_0)^T X^TX(\beta - \beta_0)) \\
&\propto (\sigma^2)^{-(\frac{n}{2}+1+\frac{k+1}{2})} exp(-\frac{1}{2\sigma^2}(y - X(X^TX)^{-1}X^Ty)^T(y - X(X^TX)^{-1}X^Ty) \\
&- \frac{1}{2\sigma^2}(X\beta - y)^T(X\beta - y) - \frac{1}{2g\sigma^2}(\beta - \beta_0)^T X^TX(\beta - \beta_0))
\end{aligned}$$

Since $y - X(X^TX)^{-1}X^Ty = 0$

$$\begin{aligned}
&p(\beta, \sigma^2|y, X) \\
&\propto (\sigma^2)^{-(\frac{n}{2}+1+\frac{k}{2})} exp(-\frac{1}{2\sigma^2}(y - X\hat{\beta})^T(y - X\hat{\beta}) - \frac{1}{2\sigma^2}(X\beta - X(X^TX^{-1})X^Ty)^T(X\beta - X(X^TX)X^Ty) \\
&- \frac{1}{2g\sigma^2}(\beta - \beta_0)X^TX(\beta - \beta_0))
\end{aligned}$$

We know that

$$X\beta - X(X^TX^{-1})X^Ty)^T(X\beta - X(X^TX)X^Ty) = (\beta - \hat{\beta})^T X^TX(\beta - \hat{\beta})$$

$$
\begin{aligned}
&p(\beta, \sigma^2|y, X) \\
&\propto (\sigma^2)^{-(\frac{n}{2}+1+\frac{k}{2})} exp[-\frac{1}{2\sigma^2}(y - X\hat{\beta})^T(y - X\hat{\beta}) - \\
&\frac{1}{2\sigma^2}(\beta - \hat{\beta})^T X^T X(\beta - \hat{\beta})] exp[-\frac{1}{2g\sigma^2}(\beta - \beta_0) X^T X(\beta - \beta_0)]
\end{aligned}
$$

c)

From the joint posterior $p(\beta, \sigma^2|y, X)$, we obtain a Gaussian posterior on β.

$$
\begin{aligned}
&p(\beta|\sigma^2, y, X) \\
&= exp(\frac{1}{2\sigma^2}(\beta - \hat{\beta})^T X^T X(\beta - \hat{\beta}) - \frac{1}{2g\sigma^2}(\beta - \beta_0) X^T X(\beta - \beta_0)) \\
&= exp(\frac{g}{2g\sigma^2}(\beta - \hat{\beta})^T(\beta - \hat{\beta}) - \frac{1}{2g\sigma^2}(\beta - \beta_0)(\beta - \beta_0)) X^T X \\
&= exp(\frac{[(g+1)\beta^2 - (2g\hat{\beta} - 2\beta_0)\beta] X^T X}{2g\sigma^2}) \\
&= exp(-\frac{g+1}{2g\sigma^2}(\beta - \frac{g}{g+1}\hat{\beta} - \frac{1}{g+1}\beta_0)^2 X^T X)
\end{aligned}
$$

$$
\beta|\sigma^2, y, X \sim N_k(\frac{g}{g+1}(\frac{\beta_0}{g} + \hat{\beta}), \frac{\sigma^2 g}{g+1}(X^T X)^{-1})
$$

Since prior is inverse and likelihood is normal distribution, then posterior is inverse Gamma distribution. So we have an Inverse gamma posterior on σ^2

$$
\begin{aligned}
&p(\sigma^2|y, X) \\
&= (\sigma^2)^{\frac{n}{2}} exp(-\frac{1}{2\sigma^2}(y - X\hat{\beta})^T(y - X\hat{\beta}) - \frac{1}{2\sigma^2}(\beta - \hat{\beta})^T X^T X(\beta - \hat{\beta}))
\end{aligned}
$$

Since $s^2 = (y - X\hat{\beta})^T(y - X\hat{\beta})$, we also replace β by it's posterior mean.

Which means

$$
\begin{aligned}
&p(\sigma^2|y, X) \\
&= (\sigma^2)^{\frac{n}{2}} exp(-\frac{s^2}{2\sigma^2} - \frac{1}{2(g+1)\sigma^2}(\beta_0 - \hat{\beta})^T X^T X(\beta_0 - \hat{\beta}))
\end{aligned}
$$

Hence $\sigma^2|y, X \sim IG(a, b)$, $a = \frac{n}{2}$ and $b = \frac{s^2}{2} + \frac{1}{2(g+1)}(\beta_0 - \hat{\beta})^T X^T X(\beta_0 - \hat{\beta})$

$$
\sigma^2|y, X \sim IG(\frac{n}{2}, \frac{s^2}{2} + \frac{1}{2(g+1)}(\beta_0 - \hat{\beta})^T X^T X(\beta_0 - \hat{\beta}))
$$

d)

Since there is no precise prior information about β_0 and g. Try $g = 10$ and $\beta_0 = 0_k$

For the problem of setting g, we can find that if $g \to \infty$ the influence of prior will be vanish and we recover the frequentist estimate of $\beta : E(\beta|y, X) = \hat{\beta}$. Let $g \to 0$ takes the posterior to the prior distribution. Some other options for choosing g include using BIC, empirical Bayes, and full Bayes.

Initialise σ, i.e. find starting values $\beta_i^{(1)}$ for $i = 1, ..., k$. For $j = 1, ..., M$

1. Draw $\beta_1^{(j+1)}$ from $p(\beta|\sigma_1^{(j)}, y, X)$
2. Draw $\sigma_1^{(j+1)}$ from $\pi(\sigma^2|x, y, \beta_1^{(j+1)})$
3. Draw $\beta_2^{(j+1)}$ from $p(\beta|\sigma_1^{(j+1)}, y, X)$
4. Draw $\sigma_2^{(j+1)}$ from $\pi(\sigma^2|x, y, \beta_2^{(j+1)})$
5. ...
6. Draw $\beta_k^{(j+1)}$ from $p(\beta|\sigma_k^{(j)}, y, X)$
7. Draw $\sigma_k^{(j+1)}$ from $\pi(\sigma^2|x, y, \beta_k^{(j+1)})$
8. Put $\beta^{(j+1)} = (\beta_1^{(j+1)}, ..., \beta_k^{(j+1)})$ and $\sigma^{(j+1)}$, set $j + 1 = j$

Here we used $\pi(\sigma^2|y, X, \beta)$ as the posterior for the $sigma^2$. And $p(\sigma^2|y, X, \beta)$ as the posterior for the β

$$
\begin{aligned}
&\pi(\sigma^2|x, y, \beta) \\
&\propto \pi(\sigma^2|y, x)\pi(\beta|\sigma^2, y, x) \\
&\propto (\sigma^2)^{-n/2-1} exp(-\frac{s^2}{2\sigma^2} - \frac{(\hat{\beta}-\beta_0)^2/(g+1)}{2\sigma^2(X^TX)^{-1}})(\sigma^2)^{-1/2} exp(-\frac{(g+1)(\beta - \frac{g\hat{\beta}+\beta_0}{g+1})^2}{2g\sigma^2(X^TX)^{-1}})) \\
&\propto (\sigma^2)^{-n/2-3/2} exp(\frac{-s^2/2 + (\beta-\hat{\beta})^T(X^TX)(\beta-\hat{\beta}) + \frac{(\beta-\beta_0)^T(X^TX)(\beta-\beta_0)}{2g}}{\sigma^2}) \\
&\propto InverseGamma(\frac{n+1}{2}, \frac{s^2}{2} + \frac{(\beta-\hat{\beta})^T(X^TX)(\beta-\hat{\beta})}{2} + \frac{(\beta-\beta_0)^T(X^TX)(\beta-\beta_0)}{2g})
\end{aligned}
$$

```
library("invgamma")
library("mvtnorm")

## input
## response variable y
## predictors data X

gibbs = function(y, X){
    beta = matrix(0, nrow=11, ncol = 1100)
    beta_0 = matrix(1, nrow=11, ncol=1)

    sigma2 = rep(0, 1100)
    T = 100  # burn-in

    n = dim(X)[1]
    k = dim(X)[2]
    g = 100

    beta_hat = solve(t(X) %*% X) %*% t(X) %*% y
    s2 = t(y - X %*% beta_hat) %*% (y - X %*% beta_hat)

    ## initialisation
    sigma2[1] = t(y-X%*%beta_hat)%*%(y-X%*%beta_hat)/(n-k-1)

    for(i in 2:1000){
        beta[,i] = rmvnorm(n=1, mean = g/(g+1) *(beta_0/g + beta_hat),
```

```
                           sigma = g*sigma2[i-1]/(g+1)*solve(t(X)%*%X))
        sigma2[i] = rinvgamma(n=1, shape = (n+1)/2, rate = s2/2 +
          (t(beta[,i]-beta_hat)%*%t(X)%*%X%*%(beta[,i]-beta_hat))/2)
    }

    par(mfrow=c(3,4))
    # remove burn-in
    beta = beta[,-(1:T)]
    sigma2 = sigma2[-(1:T)]

    for(j in 1:11){
        hist(beta[j,], xlab = paste0("simulated beta ", j),
             mai1 = paste("Histogram of beta" , j), nclass = 50)
      }
    hist(sigma2, xlab = "simulated sigma", main = "Histogram of sigma",
             nclass = 50)
    for(k in 1:11){
      print(paste0("beta ", k-1, " has mean ", mean(beta[k,]), " has variance ",
                   var(beta[k,])/(1100 - 1)) )
    }
    print(paste0("sigma^2 has mean ", mean(sigma2), " has variance ", var(sigma2)))
}
```

e)

For this dataset, we have $p = 10$ predictors with sample size $n = 442$. The regression model can be written like:

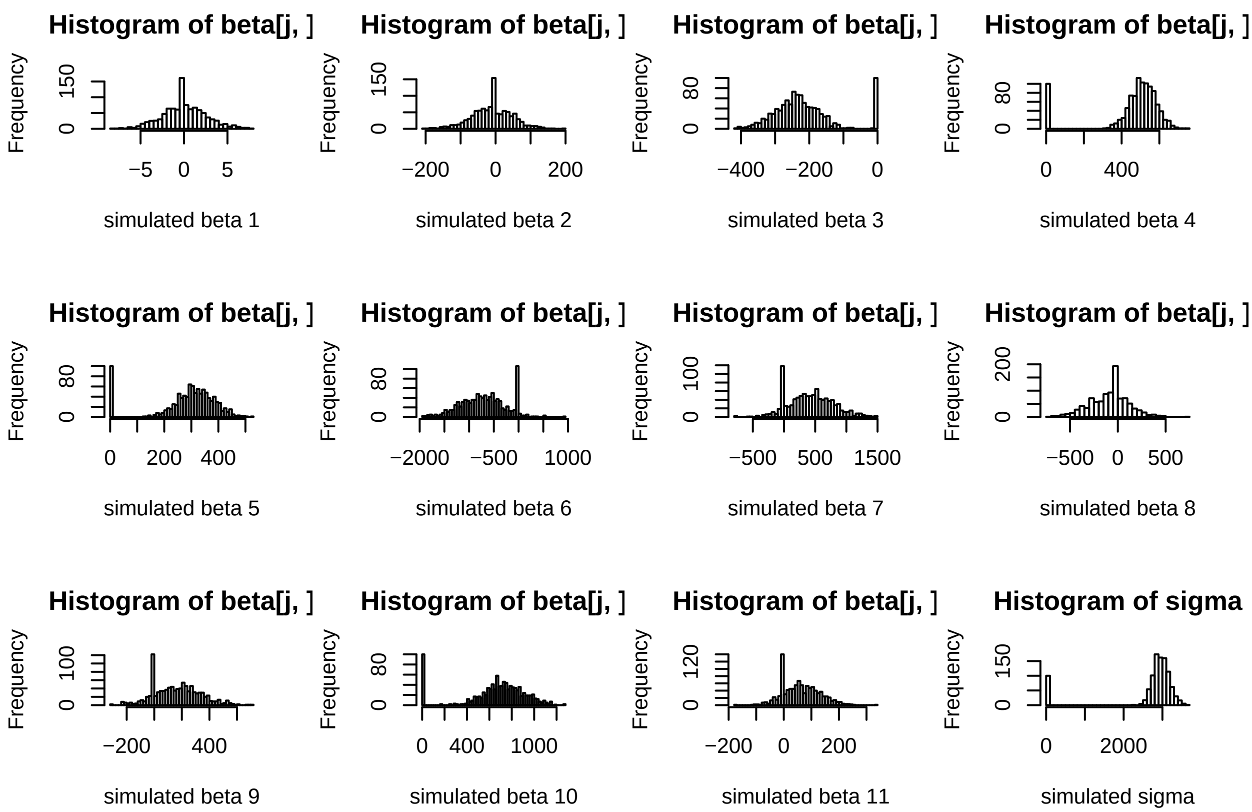
$$y = \beta_1 + \beta_2 X + \beta_3 X + ... + \beta_{10} X + \epsilon$$

To find the posterior distribution of (β, σ), we used the function defined in (d):

```
## Read data
diabete = read.table("https://web.stanford.edu/~hastie/Papers/LARS/diabetes.sdata.txt",
                     skip = 12)

X = as.matrix(diabete[, 1:10])
ONE =  rep(1,nrow(X))
X = cbind(ONE, X)

y = as.matrix(diabete[, 11])
## implement function
gibbs(y, X)
```




```
## [1] "beta 0 has mean 0.0582183100706382 has variance 0.00564086722800956"
## [1] "beta 1 has mean -9.69955296694 has variance 2.99399118395689"
## [1] "beta 2 has mean -215.868823460727 has variance 7.61839865702898"
## [1] "beta 3 has mean 463.502813089834 has variance 25.4067769007592"
## [1] "beta 4 has mean 287.048421245189 has variance 11.8153798351263"
## [1] "beta 5 has mean -694.705996003791 has variance 191.622036185762"
## [1] "beta 6 has mean 413.257431877747 has variance 116.004283869554"
## [1] "beta 7 has mean -85.4633015830116 has variance 37.7895005174702"
## [1] "beta 8 has mean 160.63451255459 has variance 25.357257659749"
## [1] "beta 9 has mean 664.562341593239 has variance 69.360799918068"
## [1] "beta 10 has mean 61.4890116523835 has variance 3.9868180517822"
## [1] "sigma^2 has mean 2674.20214490108 has variance 831700.885688921"
```

Question 2

(i)

For the case $K = 2$ in logistic regression, we have the model form:

$$log\frac{P(G=1|x)}{P(G=2|X=x)} = \beta_{10} + \beta_1^T x$$

(ii)

```r
library("statmod")
```

```
## Warning: package 'statmod' was built under R version 3.6.1
```

```r
library("stats")
## input:
## target variable y
## data matrix X

bayeslasso = function(y, X, lambda){
  y_centered = scale(y)
  ybar = mean(y)
  n = dim(X)[1]
  p = dim(X)[2]

  tau2 = rep(0, p)
  D = matrix(0, p, p)
  lambda = rep(0, 1000)
  sigma2 = rep(0, 1000)
  beta = matrix(0, nrow = p, ncol = 1000)

  ## initial
  r = 1
  sigma2[1] = 1.78
  beta[,1] = rep(1, p)

  for(j in 2:1000){

    for(i in 1:p){
      tau2[i] = rinvgauss(1, sqrt(lambda^2 * sigma2[j-1] / beta[i, j-1]),
                          lambda^2)^(-1)
    }

    tau2[is.na(tau2)] = 10^(-10)
    diag(D) = tau2
    A = t(X)%*%X + solve(D)
    beta[,j] = rmvnorm(n=1, mean = solve(A)%*%t(X)%*%y_centered,
                       sigma = sigma2[j-1] * solve(A))
    sigma2[j] = rinvgamma(n=1, shape=(n-1)/2+p/2,
                          rate=t(y_centered-X%*%beta[,j])%*%(y_centered-X%*%beta[,j])/2+
                            t(beta[,j])%*%solve(D)%*%beta[,j]/2)
    # Using Hyperpriors for the Lasso Parameter lambda
    #lambda[j] = sqrt(rgamma(n = 1, shape = p+r, rate = sum(tau2 / 2) + sqrt(sigma2[j])))
  }
  return(list(beta, sigma2, lambda))
}
```