18.S096 Problem Set 7 Spring 2018
Due Date: 4/30/2018
Where: On Stellar, prior to 11:59pm

Collaboration on homework is encouraged, but you will benefit from independent effort to solve the problems before discussing them with other people. **You must write your solution in your own words. List all your collaborators.**

Problem 1. O-Ring Failures As stated in Robert and Casella (2010, p. 196):

> In 1986, the space shuttle Challenger exploded during takeoff, killing the seven astronauts aboard. The explosion was the result of an O-ring failure, a splitting of a ring of rubber that seals the parts of the ship together. The accident was believed to have been caused by the unusually cold weather (31 degrees Fahrenheit or 0 degrees Celsius) at the time of launch, as there is reason to believe that the O-ring failure probabilities increase as temperature decreases. Data on previous space shuttle launches and O-ring failures is given in the dataset *challenger* provided with the *mcsm* package. The first column corresponds to the failure indicators y_i and the second column to the corresponding temperature x_i, $(1 \leq i \leq 24)$.

1(a) Consider the logistic regression model for this data:

$$P(Y_i = 1 \mid x_i) = p(x_i) = e^{(\alpha+\beta x_i)}/[1 + e^{(\alpha+\beta x_i)}]$$

Fit this model by maximum likelihood using the R function $glm()$:

```
> library("mcsm")
> data(challenger)
> fit.logistic<-glm(oring ~ temp, data=challenger,
+                   family=binomial(link="logit"))
> # Print out the results using the default print and summary()
> #fit.logistic
> #summary(fit.logistic)
```

Report the MLEs and construct approximate 95% confidence intervals for α and β.

1(b) The likelihood function for the logistic model is given by

$$L(\alpha, \beta) = \prod_{i=1}^{n} \left(\frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}\right)^{y_i} \left(\frac{1}{1+e^{\alpha+\beta x_i}}\right)^{1-y_i}.$$

Compute the posterior distribution for (α, β) corresponding to the prior distribution for (α, β) which assumes:

$$\alpha \sim N(0, 25) \text{ and } \beta \sim Normal(0, 25/s_x^2), \text{ (independent)}$$

where s_x^2 is the sample variance of the temperatures.

The following r script implements the Metropolis Hastings algorithm using $a[t-1] + Laplace(0, ascal)$ and $b[t-1] + Laplace(0, bscal)$ distributions for the *candidate* densities (where the parameters *ascal* and *bscal* are set based on the MLEs in part a).

```
> set.seed(1)
> Nsim=10^4
> x=challenger$temp
> y=challenger$oring
> sigmaa=5 ; sigmab=5/sd(x)
> lpost=function(a,b){sum(y*(a+b*x)-log(1+exp(a+b*x)))+ dnorm(a,sd=sigmaa,log=TRUE)+dnc
+ }
> # Initialize a and b to equal the MLEs
> beta=as.vector(fit.logistic$coefficients)
> a=b=rep(0,Nsim)
> a[1]=beta[1]
> b[1]=beta[2]
> #As scale for the proposal densities consider the square root of the
> #  cov.unscaled from the ml fit to the logistic model
> fit.logistic.summary<-summary(fit.logistic)
> scala=sqrt(fit.logistic.summary$cov.unscaled[1,1])
> scalb=sqrt(fit.logistic.summary$cov.unscaled[2,2])
> for (t in 2:Nsim){
+   propa=a[t-1]+sample(c(-1,1),1)*rexp(1)*scala
+   if (log(runif(1))<lpost(propa,b[t-1])- lpost(a[t-1],b[t-1]))
+     a[t]=propa else
+       a[t]=a[t-1]
+   propb=b[t-1]+sample(c(-1,1),1)*rexp(1)*scalb
+   if (log(runif(1))<lpost(a[t],propb)- lpost(a[t],b[t-1]))
+     b[t]=propb
+ else b[t]=b[t-1]
+ }
```

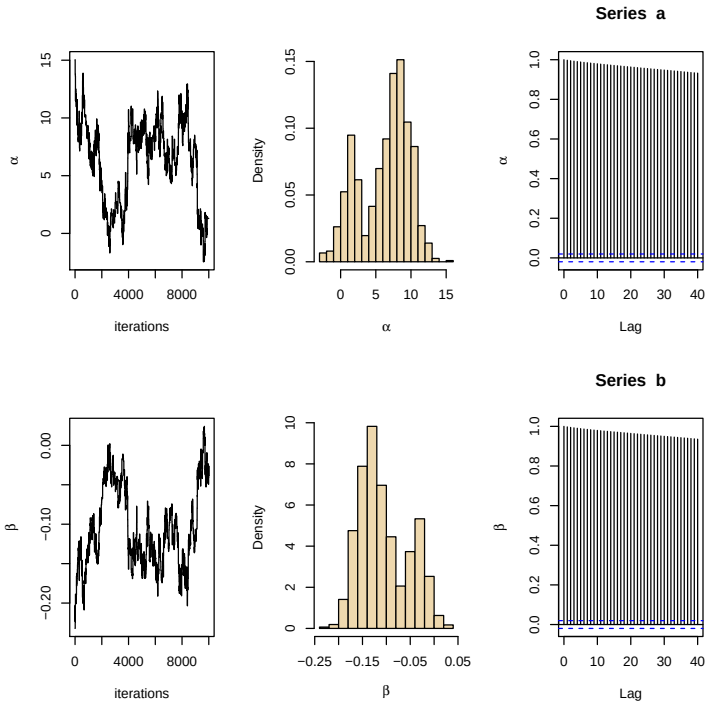
For this Monte Carlo Markov Chain approximation to the posterior distribution we can conduct the following evalutions:

- Compute the acceptance rates

  ```
  > length(unique(a))/Nsim
  [1] 0.1001
  > length(unique(b))/Nsim
  [1] 0.0987
  ```

- Explore the results graphically:

```
> par(mfrow=c(2,3))
> plot(a,type="l",xlab="iterations",ylab=expression(alpha))
> hist(a,prob=TRUE,col="wheat2",xlab=expression(alpha),main="")
> acf(a,ylab=expression(alpha))
> plot(b,type="l",xlab="iterations",ylab=expression(beta))
> hist(b,prob=TRUE,col="wheat2",xlab=expression(beta),main="")
> acf(b,ylab=expression(beta))
```

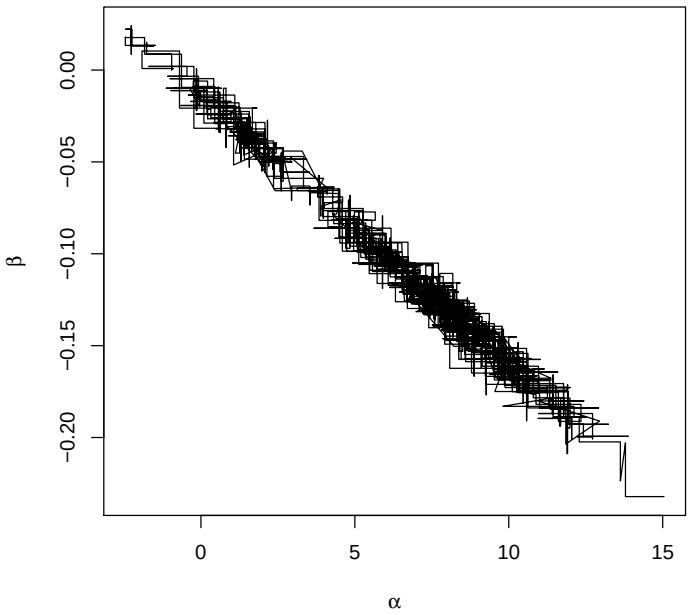


- Summarize the means and standard deviations of the posterior distribution for α and β

```
> print (c(mean(a), sd(a))); print (c(mean(b), sd(b)))
[1] 6.294067 3.497769
[1] -0.10533114  0.05123078
>
```

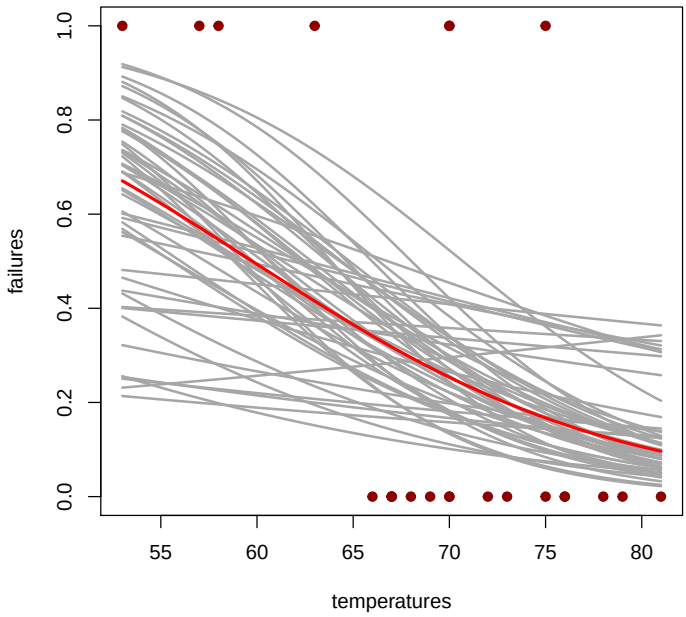
- Plot the path of the Monte Carlo Markov Chain:

```
> par(mfcol=c(1,1))
> plot(a,b,type="l",xlab=expression(alpha),ylab=expression(beta))
```



- On one plot, display: (i) the sample data (failure versus temperature); (ii) super-pose the logistic curve for probability of failure at 50 realizations of the posterior distribution for (α, β); (iii) add the logistic curve corresponding to the logistic curve

```
> plot(challenger$temp,challenger$oring,pch=19,col="red4",
+ xlab="temperatures",ylab="failures")
> for (t in seq(1000,Nsim,le=50)) curve(1/(1+exp(-a[t]-b[t]*x)),
+ add=TRUE,col="grey65",lwd=2)
> curve(1/(1+exp(-mean(a)-mean(b)*x)),add=TRUE,col="red",lwd=2.5)
> postal=rep(0,1000);i=1
```



- Estimate the probability of failure along with a standard error at $x = 60, 50, 40,$ and 30 (degrees Fahrenheit).

```
>   for (x0 in c(60,50,40,30)){
+     print(c(x0,
+             mean(1/(1+exp(-a-b*x0))),
+             sd(1/(1+exp(-a-b*x0)))))
+   }
[1] 60.0000000  0.4944781  0.1431809
[1] 50.0000000  0.7012076  0.1959169
[1] 40.0000000  0.8134438  0.2077082
[1] 30.0000000  0.8673593  0.1983737
```

1(c) Repeat part (b) with less informative Normal priors for (α, β) which have higher variances by a factor of 100. Comment specifically on:

- Comparing the acceptance rates of the algorithm for α and β
- Comparing the means/standard deviations of the marginal posterior distributions for α and for β.
- Comparing the probability of failure and its standard error at 30 Degrees Fahrenheit.

1(d) Repeat part (c), changing the scale of the *candidate* Laplace densities to be one-tenth their specificaitons in (a).

Problem 2. Failure-Time Data: Nuclear Plant Pumps

Gaver and O'Muircheartaigh (1987) reported data on the number of failures and times of observations for 10 pumps in a nuclear plant:

Pump	1	2	3	4	5	6	7	8	9	10
Failures	5	1	5	14	3	19	1	1	4	22
Time	94.32	15.72	62.88	125.76	5.24	31.44	1.05	1.05	2.10	10.48

Formalize a model for the data as follows:

- $i = 1, \ldots, 10$ Pumps
- For each pump, observe

  t_i = observation time of ith pump

  X_i = number of failures
- Data Model

  $X_i \sim Poisson(\lambda_i t_i)$ $i = 1, \ldots, 10$

:

- Hierarchical Prior for $\{\lambda_1, \ldots, \lambda_{10}\}$

  $\lambda_i \sim Gamma(\alpha, \beta)$ (i.i.d.) $i = 1, \ldots, 10$

  $\beta \sim \Gamma(\gamma, \delta)$.

  where the prior model parameters α, γ, δ are constants set by the statistical modeler.

The joint posterior distribution for $\theta = (\lambda_1, \ldots, \lambda_{10}, \beta)$:

$$\pi(\lambda_1, \ldots, \lambda_{10}, \beta \mid t_1, \ldots, t_{10}, x_1, \ldots, x_{10}) \propto \Big[\prod_{i=1}^{10} (\lambda_i t_i)_i^x e^{-\lambda_i t_i} \lambda_i^{\alpha-1} e^{-\beta\lambda_i} \Big] \beta^{10\alpha} \beta^{\gamma-1} e^{-\delta\beta}$$

2(a) Simplify the proportional expression for the joint posterior density, combining similar factors when possible.

2(b) Derive Full Conditional distributions:

$\pi(\lambda_i \mid \beta, t_i, x_i)$ $i = 1, \ldots, 10$

$\pi(\beta \mid \lambda_1, \ldots, \lambda_{10})$

2(c) Consider the same prior specification of Robert and Casella (2010) see demo("Chapter.7",library="mcsm")

$\alpha = 1.8$, $\gamma = 0.01$, and $\delta = 1$

- Generate a Monte-Carlo sample ($n = 10000$) from the prior distribution for λ_1
- Plot a histogram of the simulated prior sample

- Compute summary statistics for the prior distribution sample: sample mean, sample median, standard deviation.
- Compute maximum-likelihood estimates of $\lambda_1, \ldots, \lambda_{10}$, ignoring any prior distribution for the parameters.
- Replot the histogram and add vertical bars at the MLE values.
- Compute the quantile/percentile of the MLE values with respect to the simulated distribution.
- Comment on the degree of consistency of the MLEs with the prior specification.

2(d) Apply the Gibbs Sampler to approximate marginal posterior distributions of $\lambda_1, \ldots, \lambda_1$ and β. (You can just copy the code from the demo script).

2(e) Compute 95% posterior credible intervals for the parameters.

Based on these computations, can you identify any pumps that are more or less reliable than the others?

2(f) Repeat the analysis of 2(c), 2(d) and 2(e) with one or more alternate specifications of the prior distribution. (If possible, motivate your choice of specification, e.g., based on trying to specify a non-informative prior or based on informative prior that has knowledge of the range of the MLEs.)