# Assignment 10

1. One of the characteristics of leukemia is an excess of white blood cells. The white blood cell count at diagnosis can be used to aid in predicting a patient's survival time after diagnosis, with high white blood cell counts indicating a low expected survival time. Feigl and Zelen (Biometrics, 1965) show survival times in weeks and white blood cell counts (WBC) at diagnosis for 33 patients who died of acute leukemia. The patients were classified as AG positive or AG negative depending on the presence or absence of certain characteristics in the white blood cells. In this problem consider the data for the 17 patients classified as AG positive. The data set is also available online or the full data is available as `leuk` in the `MASS` package.

   Feigl and Zelen model the conditional distribution of the lifetimes given the white blood cell count as exponentially distributed with mean

   $$\theta_1 \exp(-\theta_2 x_i),$$

   where

   $$x_i = \log(\text{WBC}_i/10000)$$

   The maximum likelihood estimates for this model are $\hat{\theta}_1 = 56.85$ and $\hat{\theta}_2 = 0.482$.

   Consider a proper but vague prior distribution with $\theta_1$ and $\theta_2$ independent a priori and

   $$\begin{aligned} 1/\theta_1 &\sim \text{Exponential with mean } 1000 \\ \theta_2 &\sim \text{Uniform on } [0, 1000]. \end{aligned}$$

   (a) Verify the maximum likelihood estimates graphically by plotting the log likelihood function and numerically using a nonlinear optimizer such as `optim` in R.

   (b) Compute the unnormalized marginal posterior density of $\theta_2$ by analytically integrating out $\theta_1$ and plot the resulting unnormalized density.

   (c) Use either the rejection sampling or the ratio-of-uniforms approach to construct a method to sample from the marginal posterior density of $\theta_2$. Use graphical, numerical, or analytical methods to choose a reasonable bound for rejection sampling or the shift $\mu$ and to choose bounding rectangle for ration-of-uniforms sampling. Generate a sample of size 10000 and graphically compare a density estimate for your sample to the analytic form of the unnormalized density.

   (d) Augment your sample of $\theta_2$ values to a sample of $\theta_1, \theta_2$ pairs by drawing $\theta_1$ values from the appropriate conditional distribution.

(e) Use your sample to construct estimates of the posterior means and variances of $\theta_1$ and $\theta_2$, the posterior covariance of $\theta_1$ and $\theta_2$, and the marginal posterior density of $\theta_1$. Obtain standard errors for your moment estimates and provide a plot of the approximate marginal posterior density of $\theta_1$. Use any variance reduction methods that seem appropriate.

You should submit your assignment electronically using Icon. Submit your work as a single compressed tar file. If your work is in a directory `mywork` then you can create a compressed tar file with the command

```
tar czf mywork.tar.gz mywork
```

## Solutions and Comments

1. The likelihood for the model is

$$L(\theta_1, \theta_2) = \frac{\exp\{\theta_2 \sum x_i\}}{\theta_1^n} \exp\{-\sum t_i \exp\{\theta_2 x_i\}/\theta_1\}$$

   (a) A contour plot of the log likelihood can be used to graphically confirm the MLE values by successively reducing the axis ranges.

   (b) Let $\psi_1 = 1/\theta_1$. Then the posterior density of $f(\psi_1, \theta_2)$ is proportional to

$$f(\psi_1, \theta_2 | t_i) \propto \psi_1^n \exp\{\theta_2 \sum x_i\} \exp\{-\sum t_i \psi_1 \exp\{\theta_2 x_i\} - \psi_1/1000\}$$

   As a function of $\psi_1$ this has the form of a Gamma density, and the unnormalized marginal posterior density of $\theta_2$ is therefore

$$f(\theta_2 | t_i) \propto \frac{\exp\{\theta_2 \sum x_i\}}{[\sum t_i \exp\{\theta_2 x_i\} + 1/1000]^{n+1}}$$

   for $0 < \theta_2 < 1000$. The data can be read in and a plot of the unnormalized density constructed using

```
fz <- read.table("feigzel.dat", head = T)
x <- log(fz[, 1] / 10000)
y <- fz[, 2]
lf2 <- function(t2)
    ifelse(0 < t2 & t2 < 1000,
           t2 * sum(x) - (length(x) + 1) * log(sum(y * exp(t2 * x)) + 1 / 1000),
           -Inf)
plot(function(x) exp(sapply(x, lf2)), 0, 1)
```

   (c) Inspection of the plot shows that the mode is roughly at 0.45. A plot of the ratio of uniforms region is obtained by

```
mu <- .45
lf2m <- lf2(mu)
tt <- seq(-4, 4, len = 1000)
u <- exp((sapply(tt + mu, lf2) - lf2m)/ 2)
v <- tt * u
plot(v, u, type = "l")
polygon(v, u, col = "grey")
```

   Computing the density on the log scale and subtracting the log density value at the approximate mode $\mu$ should help with numerical stability.

   Inflating the minimum and maximum of the discretized boundary values by 2% produces a rectangle that encloses the region:

```
vmax <- max(v) * 1.02
vmin <- min(v) * 1.02
umax <- max(u) * 1.02
rect(vmin,0,vmax,umax)
```

The ratio of uniforms sample is constructed by rejection sampling from this rectangle.

```
ru <- function(n = 10000) {
    x <- double(n)
    for (i in 1:n) {
        repeat {
            u <- runif(1, 0, umax)
            v <- runif(1, vmin, vmax)
            if (u <= exp((lf2(v / u + mu) - lf2m) / 2)) break
        }
        x[i] <- v / u
    }
    x + mu
}
t2 <- ru(10000)
```

For the graphical comparison we need to scale the plots appropriately; scaling both to have maximum at or near one is simplest:

```
d <- density(t2)
plot(d$x, d$y/ max(d$y), type = "l")
lines(tt + mu, exp((sapply(tt + mu, lf2) - lf2m)), col = "red")
```

(d) The conditional distribution $\psi_1|\theta_2$ is Gamma with exponent $n+1$ and rate $\sum t_i \exp\{\theta_2 x_i + 1/1000\}$. A sample of $\theta_1 = 1/\psi_1$ values is generated by

```
t1 <- 1/rgamma(length(t2), length(x) + 1,
               rate = sapply(t2, function(t) sum(y * exp(t * x)) + 1 / 1000))
```

(e) The estimated posterior means, variances, and covariance are computed by

```
mean(t1)
mean(t2)
var(cbind(t1,t2))
```

The mean of $\theta_1$ could be estimated more accurately using Rao-Blackwellization based on its inverse Gamma full conditional distribution.

Standard errors for the means are given by

```
sd(t1) / sqrt(length(t1))
sd(t2) / sqrt(length(t2))
```

and asymptotic standard errors for the variances and covariance by

```
sd((t1 - mean(t1))^2) / sqrt(length(t1))
sd((t2 - mean(t2))^2) / sqrt(length(t2))
sd((t1 - mean(t1)) * (t2 - mean(t2))) / sqrt(length(t1))
```

Marginal density estimates for $\theta_1$ can be computed more accurately using Rao-Blackwellization.