Bayesian Statistics
Statistics 4224/5224 — Fall 2023

Homework 6

The following problems are taken from *A First Course in Bayesian Statistical Methods*, by Peter D. Hoff.

1. A population of 532 women living near Phoenix, Arizona were tested for diabetes. Other information was gathered from these women at the time of testing, including number of pregnancies, glucose level, blood pressure, skin fold thickness, body mass index, diabetes pedigree, and age. The information appears in a data file that can be read into R by

```
file <- "http://www2.stat.duke.edu/~pdh10/FCBS/Exercises/azdiabetes.dat"
Data <- read.table(file=file, header=T);  rm(file);
dim(Data);  names(Data);  table(Data$diabetes);
y.D <- Data[Data$diabetes=="Yes", 1:7]
y.N <- Data[Data$diabetes== "No", 1:7]
rm(Data)
```

   Model the joint distribution of these variables for the diabetics and non-diabetics separately, using a multivariate normal distribution: $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n | \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim$ iid Normal$_7(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Assume the prior distributions $\boldsymbol{\mu} \sim$ Normal$_7(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0)$ and $\boldsymbol{\Sigma} \sim$ Inverse-Wishart$_7(\nu_0, \mathbf{S}_0^{-1})$ with $p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = p(\boldsymbol{\mu}) \times p(\boldsymbol{\Sigma})$.

   For both groups, separately, use the following type of *unit information prior*: $\boldsymbol{\mu}_0 = \bar{\boldsymbol{y}}$, the sample mean vector; $\boldsymbol{\Lambda}_0 = \mathbf{S}_0 = \hat{\boldsymbol{\Sigma}}$, where $\hat{\boldsymbol{\Sigma}}$ is the sample covariance matrix, and $\nu_0 = p + 2 = 9$.

   Choose reasonable starting values and run $S = 10{,}000$ iterations of the Gibbs sampler for $\{\boldsymbol{\mu}_d, \boldsymbol{\Sigma}_d\}$ and $\{\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n\}$, the model parameters for diabetics and non-diabetics, respectively.

   (a) For each of the seven variables, $j \in \{1, \ldots, 7\}$, compare the marginal posterior distributions of $\mu_{d,j}$ and $\mu_{n,j}$. Which variables seem to differ between the two groups? Also obtain $\Pr(\mu_{d,j} > \mu_{n,j} | \mathbf{Y})$ for each $j \in \{1, \ldots, 7\}$.

   R *Hint: If you have posterior simulations saved as* `muD.chain` *and* `SigmaD.chain`*, and* `muN.chain` *and* `SigmaN.chain`*, respectively, then*

```
op <- par(mfrow=c(2,4))
for(j in 1:7){
 den.D <- density(muD.chain[,j], adj=2)
 den.N <- density(muN.chain[,j], adj=2)
 plot(NA, xlim=range(c(den.D$x, den.N$x)),
   ylim=c(0, max(c(den.D$y, den.N$y))),
   main=paste("Posteriors for expected value of ",
     colnames(y.D)[j], sep=""),
   xlab=paste("mu_", j, sep=""), ylab="Density")
 lines(den.D, col="red", lwd=2);
 lines(den.N, col="blue", lwd=2, lty=2);  }
legend("topright", inset=.05, lwd=2, lty=1:2,
   col=c("red", "blue"),
   legend=c("Diabetics", "Non-diabetics") )
par(op)
apply(muD.chain > muN.chain, 2, mean)
```

will produce the graphical display and numerical summaries called for by this question.

(b) Obtain the posterior means of $\mathbf{\Sigma}_d$ and $\mathbf{\Sigma}_n$ and compare their diagonal entries. What do the results suggest about the differences between the variances for diabetics versus non-diabetics?

R *Hint: If* `Sigma.chain` *contains the posterior simulations* $\mathbf{\Sigma}^s$ *for* $s = 1, \ldots, S$*, then*

```
vars <- colnames(y)
Sigma.hat <- matrix(apply(Sigma.chain, 2, mean), p, p)
rownames(Sigma.hat) <- vars;  colnames(Sigma.hat) <- vars;
round(Sigma.hat, 2)
```

will return the MCMC approximation to $\mathrm{E}(\mathbf{\Sigma}|\mathbf{Y})$.

(c) Obtain $\Pr(\tilde{Y}_{d,j} > \tilde{Y}_{n,j}|\mathbf{Y})$ for each $j \in \{1, \ldots, 7\}$. Clearly explain what these values represent.

2. The file `agehw.dat` contains data on the ages of $n = 100$ opposite-sex married couples sampled from the U.S. population.

```
file <- "http://www2.stat.duke.edu/~pdh10/FCBS/Exercises/agehw.dat"
Data <- read.table(file=file, header=T);  rm(file);
```

Assume the sampling model $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n | \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim$ iid $\text{Normal}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and use a semiconjugate "diffuse prior" with $\boldsymbol{\mu}_0 = \mathbf{0}$ and $\boldsymbol{\Lambda}_0 = 10^5 \times \mathbf{I}$ and $\nu_0 = 3$ and $\mathbf{S}_0 = 1000 \times \mathbf{I}$.

Obtain an MCMC approximation to $p(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$.

(a) Plot the joint posterior distribution of μ_h and μ_w. Obtain the posterior median and 50% and 95% posterior confidence intervals for μ_h and μ_w.

R *Hint: Assuming* `mu.chain` *contains the posterior simulations, the scatterplot*

```
plot(mu.chain, xlab="Husband age", ylab="Wife age", cex=.5)
```

will give a suitable visual summary of the joint distribution, and the table

```
probs <- c(.025, .25, .5, .75, .975)
Quants <- apply(mu.chain, 2, quantile, probs=probs)
Quants <- t(Quants); rownames(Quants) <- c("Husband", "Wife")
round(Quants, 2)
```

will neatly summarize the posterior quantiles called for in this problem.

(b) Plot the marginal posterior density of ρ, the correlation coefficient between Y_h and Y_w (the ages of a husband and wife). Obtain the posterior median and 50% and 95% posterior confidence intervals for ρ.

Hint: If `Sigma.chain` *contains the posterior simulations* $\boldsymbol{\Sigma}^s$ *for* $s = 1, \ldots, S$*, then*

```
rho.chain <- Sigma.chain[,2]/sqrt(Sigma.chain[,1]*Sigma.chain[,4])
```

generates samples from the posterior $p(\rho | \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$.

(c) Plot the marginal posterior density of $\mu_h - \mu_w$, and obtain $\Pr(\mu_h > \mu_w | \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$.

(d) Plot the posterior predictive distribution of $(\tilde{Y}_h, \tilde{Y}_w)$, the husband's and wife's ages for a randomly selected opposite-sex married couple, and compare this with your answer to part (a). Obtain $\Pr(\tilde{Y}_h > \tilde{Y}_w | \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$, and compare with your answer to part (c).

Hint: To compare the posterior distribution $p(\mu_h, \mu_w | \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ *with the posterior predictive distribution* $p(\tilde{y}_h, \tilde{y}_w | \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$*, produce the graphical display*

```
plot(y.tilde, xlab="Age of husband", ylab="Age of wife", cex=.5)
points(mu.chain, pch=19, cex=.25, col="red")
```

and explain why it appears the way it does.

3. The file `http://www2.stat.duke.edu/~pdh10/FCBS/Exercises/swim.dat` contains data on the amount of time, in seconds, it takes each of four high school swimmers to swim 50 yards. Each swimmer has six times, taken on a biweekly basis.

   Perform the following data analysis for each swimmer separately:

   - Fit a linear regression model with swimming time as the response and week as the explanatory variable. To formulate your prior, use the information that competitive times for this age group generally range from 22 to 24 seconds.
   - For each swimmer k, obtain a posterior distribution for their expected time, and a posterior predictive distribution for their realized time, if they were to swim two weeks from the last recorded time.

   Letting $Y_{k,x}$ denote the kth swimmer's time at bi-weekly period x, posit the model that

$$Y_{k,x} = \beta_{1,k} + \beta_{2,k}x + \epsilon_{k,x}$$

   for $k \in \{1,2,3,4\}$ and $x \in \{1,2,3,4,5,6\}$, where the $\epsilon_{k,x} \sim iid$ Normal$(0,\sigma_k^2)$.

   Analyze the four swimmers' performances separately and independently. Based on the 'prior information' provided, a reasonable prior distribution would be

$$\boldsymbol{\beta}_k \sim \text{Normal}(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0) \quad \textit{and} \quad \sigma_k^2 \sim \textit{Inv-}\chi^2(\nu_0, \sigma_0^2)$$

   setting

$$\boldsymbol{\beta}_0 = \begin{pmatrix} 23 \\ 0 \end{pmatrix} \quad \textit{and} \quad \boldsymbol{\Sigma}_0 = \begin{pmatrix} 1/4 & 0 \\ 0 & 1/4 \end{pmatrix} \quad \textit{and} \quad \nu_0 = 1 \quad \textit{and} \quad \sigma_0^2 = 1/4\,.$$

   Now you can approximate the joint posterior $p(\boldsymbol{\beta}_k, \sigma_k^2 | y_{k,1}, \ldots, y_{k,6})$ using the Gibbs sampler as illustrated in the 'Ex14a1' example, and obtain $\{(\boldsymbol{\beta}_k^{(s)}, \sigma_k^{2(s)}) : s = 1, \ldots, S\}$ for $k = 1,2,3,4$.

   For the posterior predictive simulation, let $\tilde{Y}_k = Y_{k,7} = \beta_{1,k} + 7\beta_{2,k} + \epsilon_{k,7}$ where $\epsilon_{k,7} \sim$ Normal$(0,\sigma_k^2)$. Generate samples from $p(y_{k,7}|y_{k,1}, \ldots, y_{k,6})$ by

$$\tilde{y}_k^{(s)} \sim \text{Normal}(\beta_{1,k}^{(s)} + 7\beta_{2,k}^{(s)}, \sigma_k^{2(s)})$$

   for $s = 1, \ldots, S$, for $k = 1,2,3,4$.

   (a) Prepare a graphical summary to facilitate comparison between the posterior distributions for the four swimmers' expected times.

(b) Prepare a graphical summary to compare the four posterior predictive distributions.

Part (a) calls for a comparison of the distributions $p(\beta_{k,1} + 7\beta_{k,2}|\boldsymbol{y}_k)$*, and part (b) for a comparison of the distributions* $p(\tilde{y}_k|\boldsymbol{y}_k)$*, across* $k = 1, 2, 3, 4$*.*

Graphical summaries of the comparisons can be made by either posting four plots in a single display, taking care to use consistent sales on both x*- and* y*-axes; or by plotting four density curves on a single pair of axes.*

(c) The coach of the team has to decide which of the four swimmers will compete in a swimming meet in two weeks. Use your analysis to make a recommendation to the coach

i. if the team is only allowed to enter one swimmer;

ii. if the team is allowed to enter two swimmers; and

iii. if the team is allowed to enter three of their four swimmers.

Justify your answer.

For part (c), compute $\Pr(\tilde{Y}_{\pi_1} < \tilde{Y}_{\pi_2} < \tilde{Y}_{\pi_3} < \tilde{Y}_{\pi_4}|\mathbf{Y})$ *for each permutation* $\boldsymbol{\pi}$ *of* $\{1, 2, 3, 4\}$*, to determine which swimmer or pair of swimmers is most likely to post the best (worst) time(s) at this meet in two weeks.*