Stat 641 Bayesian Statistics

Exam 2 (take home), due Friday, November 6, 2020.

This exam is a concrete example of a situation where we have bivariate normal data, consisting of the ages of 100 married heterosexual couples sampled from the U.S. population. The problem is essentially an analysis of the relationship between the ages of husband-wife pairs, but not on the specific ages themselves. For example, are husbands typically older or younger or the same age as their wives? And are the ages of husbands and wives typically positively correlated, negatively correlated, or uncorrelated? (The questions below are an adapted version of a length question in the text book by Peter Hoff for an intro Bayes class.)

The file `agehw.txt` contains data on the ages of 100 married heterosexual couples sampled from the U.S. population. It is posted on Blackboard.

1. Before you look at the data, use your own knowledge to formulate a semiconjugate prior distribution for $\boldsymbol{\mu} = (\mu_h, \mu_w)^T$ and Σ, where μ_h, μ_w are mean husband and wife ages, and Σ is the covariance matrix.

2. Generate a *prior predictive data set* of size $n = 100$, by sampling $(\boldsymbol{\mu}, \Sigma)$ from your prior distribution and then simulating $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n \sim$ iid MVN$(\boldsymbol{\mu}, \Sigma)$. Generate several such data sets, make bivariate scatterplots for each data set, and make sure they roughly represent your prior beliefs about what such a data set would actually look like. If necessary, repeat (a). Report the prior you eventually decide upon, and provide scatterplots for at least three prior predictive data sets.

   * By "prior predictive data sets", I mean "sample from the prior distribution for ages of husband-wife pairs". In class I showed you how to do this using the R function `mvrnorm`, but it's actually easier to use JAGS to obtain a sample, by simply omitting from your model statement the likelihood. The model plus R code below are what I'd suggest you use; they are a slightly tidier version of code from pages 165-170 of the lecture notes.

```
model{
  # (Note: I removed the likelihood statement here, so I'd be
  # sampling from the prior predictive distribution for ypred below)
  mu[1:2] ~ dmnorm( mu0[],Lambda0.inv[,] )
  # (And I allow for a correlated prior distribution by using
  # dmnorm instead of two dnorm statements.)

  Sigma.inv[1:2,1:2] ~ dwish( S0[,], nu0 )
  Sigma[1:2,1:2] <- inverse( Sigma.inv[1:2,1:2] )
  ypred[1:2] ~ dmnorm( mu[],Sigma.inv[,] )
}
```

```
########## Here's R code, which you will need to modify:
n <- 50  # sample of this many couples?
library(MASS) # for mvrnorm;  MASS is in the default R installation

n <- 100 # sample of this many couples?

# prior for (mu.h,mu.w):  N( mu0, Lambda0 )
mu0 <- c( 40,39 )  # mean(h,w)
Lambda0 <- matrix( c(25,20,20,25), nrow=2,ncol=2 )
Lambda0.inv <- solve(Lambda0)

# prior for Sigma:  IW( nu.0, S0.inv )
v.h.0 <- 10 # 186 # 10
v.w.0 <- 10 # 164 # 10
rho.0 <- 0.1
cov.0 <- rho.0*sqrt( v.h.0 * v.w.0 )
S0 <- matrix( c( v.h.0, cov.0,cov.0, v.w.0 ),
              nrow=2,ncol=2 )
S0.inv <- solve(S0)
nu0 <- 4 # wide prior for covariance matrix?

my.data <- list( S0=S0, nu0=nu0,
                 mu0=mu0, Lambda0.inv=Lambda0.inv )
my.inits <- list( mu=c(40,40) )
my.fname <- "agehw-model.txt"
library(rjags)
my.jags.model <- jags.model(
  file=my.fname, data=my.data, inits=my.inits,
  n.chains=1, n.adapt=1000, quiet=FALSE)
my.variables <- c("mu","Sigma","ypred")
my.coda.samples <- coda.samples(my.jags.model,
                                my.variables, 1200)

# Extract just a few samples, and this will be one simulated data set:
which <- seq( 100,1200, length=n )
agesh <- my.coda.samples[[1]][,"ypred[1]"][which] ##############
agesw <- my.coda.samples[[1]][,"ypred[2]"][which] ##############
plot(agesw,agesh) # construct scatterplot of the ages for pairs.
```

3. In the R code above, what am I using for the joint prior distribution for (μ_h, μ_w), the average ages of husbands/wives? What is the (prior) correlation between their ages?

4. Read the data set in the file `agehw.txt` into R and create a scatterplot of the data. Comment briefly. Sample R code:

```
dat <- read.table("agehw.txt",header=TRUE)
names(dat)
nrow(dat)
hhh <- dat$ageh
www <- dat$agew
apply(dat,2,mean)
cov(dat)
cor(dat)
summary( hhh )
summary( www )
plot( www,hhh )
```

5. Use the prior distribution you selected in (1), together with the data in `agehw.txt` to fit your Bayesian model. You should use JAGS for this part, by modifying the model statement to include the likelihood. (See model statement on page 166 of the lecture notes.)

   Summarize your results: Traceplots, density plots, summary statistics, and a brief discussion. Be sure to discuss (briefly) the posterior correlation between μ_h and μ_w, and include a scatterplot.

6. Obtain 200 samples from the posterior predictive distribution for couples; provide a scatterplot, and discuss in a sentence or two. (These are again 'ypred', but now sampling from the posterior distribution.)

7. Find a 95% credible interval for the age of husbands whose wives are 43.0 years old. (You will need to modify the code on pages 169-170; here, wives are y2, we're assuming y2 is known and we wish to obtain samples of y1.)

8. Find a 95% credible interval for the age of wives whose husbands are 43.0 years old. (This is similar to problem 7, except now we're assuming y1 is known and we'd like to obtain samples of y2.)

9. Be sure to include your model statement(s) and your R code.