Paul Cubre

Math 9810 Homework 3

7.3 Australian crab data: The files bluecrab.dat and orangecrab.dat contain measurements of body depth (Y_1) and read width (Y_2), in millimeters, made on 50 male crabs from each of two species, blue and orange. We will model these data using a bivariate normal distribution.

a) For each of the two species, obtain posterior distributions of the population mean θ and covariance matrix Σ as follows: Using the semi-conjugate prior distributions for θ and Σ, set μ_0 equal to the sample mean of the data, Λ_0 and S_0 equal to the sample covariance matrix and $\nu_0 = 4$. Obtain $10,000$ posterior samples of θ and Σ. Note that this "prior" distribution loosely centers the parameters around empirical estimates based on the observed data (and is very similar to the unit information prior described in the previous exercise). It cannot be considered as our true prior distribution, as it was derived from the observed data. However it can me roughly considered as the prior distribution of someone with weak but unbiased information.

b) Plot values of $\theta = (\theta_1, \theta_2)'$ for each group and compare. Describe any size differences between the two groups.

We see in Figure 1 and 2 that the blue crabs are highly correlated and orange crabs are not

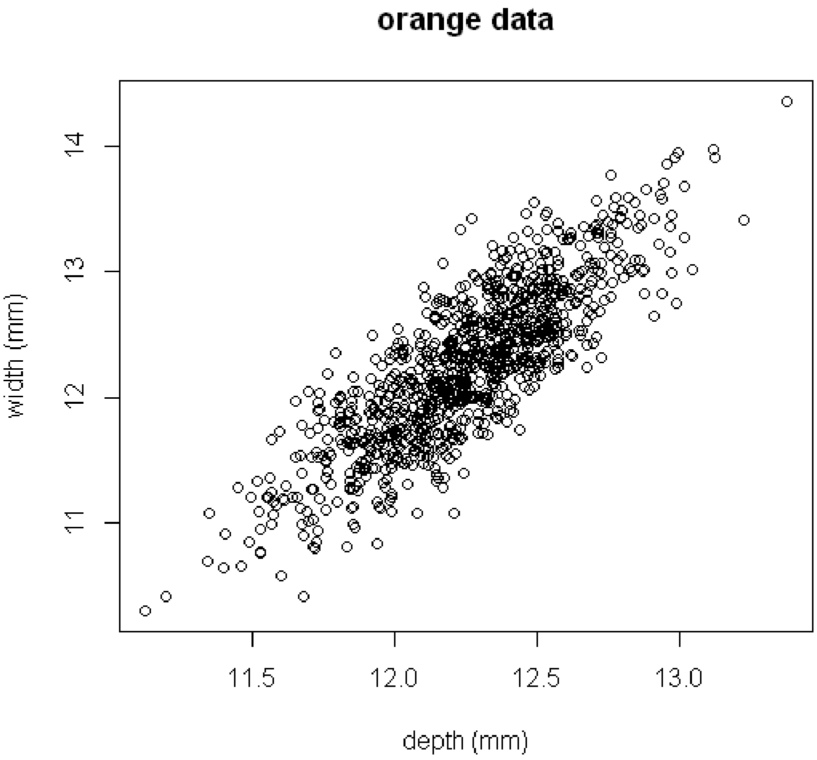

Figure 1: Plots for θ_1 and θ_2

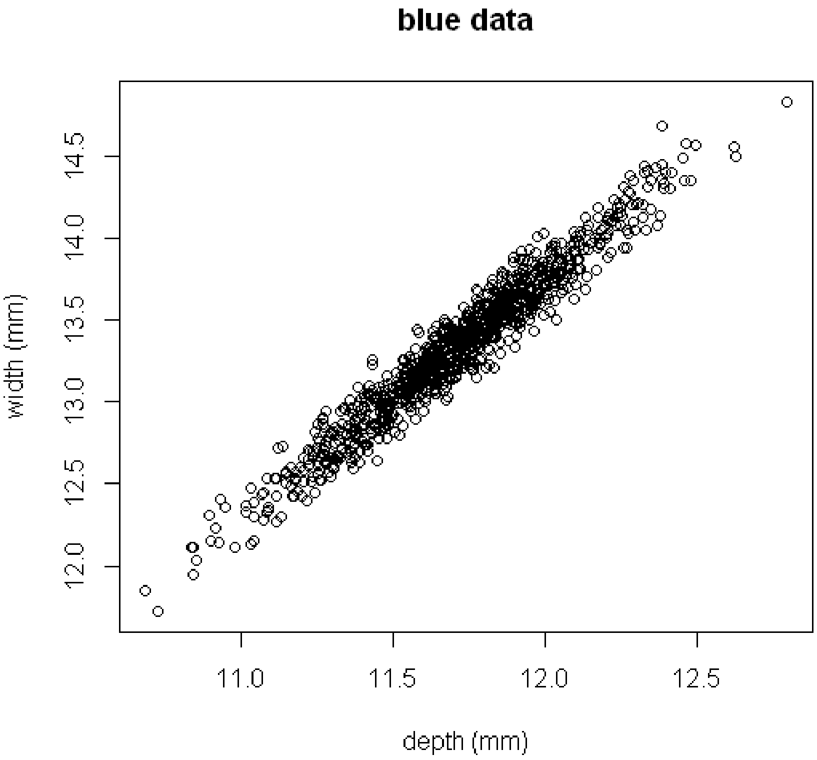

Figure 2: Plots for θ_1 and θ_2

c) From each covariance matrix obtained from the Gibbs sampler, obtain the corresponding correlation coefficient. From these values, plot posterior densities of the correlations ρ_{blue} and ρ_{orange} for the two groups. Evaluate differences between the two species by comparing these posterior distributions. In particular, obtain an approximation to $Pr(\rho_{blue} < \rho_{orange} | y_{blue}, y_{orange})$. What do the results suggest about differences between the two populations?
We have that $Pr(\rho_{blue} < \rho_{orange} | y_{blue}, y_{orange}) = 0$. Therefore the blue crab population is highly correlated and the orange crab population is less correlated.

9.1 Extrapolation: The file 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.

a) Perform the following data analysis for each swimmer separately:

i. Fit a linear regression model of 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.

We have that assuming a Jeffrey's prior:

$$\begin{aligned}
\pi(\beta_0, \beta_1, \sigma^2|Y) &\propto p(Y|\beta_0, \beta_1, \sigma^2)\pi(\beta_0, \beta_1, \sigma^2) \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{\|Y-X\beta\|^2}{2\sigma^2}} \frac{1}{\sigma^2} \\
&\propto \left(\frac{1}{\sigma^2}\right)^{n/2+1} e^{-\frac{\|Y-X\beta\|^2}{2\sigma^2}}
\end{aligned}$$

Therefore

$$\begin{aligned}
\pi(\beta_0, \beta_1|\sigma^2, Y) &= \frac{\pi(\beta_0, \beta_1, \sigma^2|Y)}{\pi(\sigma^2|Y)} \\
&\propto \pi(\beta_0, \beta_1, \sigma^2|Y) \\
&\propto \left(\frac{1}{\sigma^2}\right)^{n/2+1} e^{-\frac{\|Y-X\beta\|^2}{2\sigma^2}} \\
&\propto e^{-\frac{\|Y-X\beta\|^2}{2\sigma^2}} \\
&= e^{-\frac{(Y-X\beta)'(Y-X\beta)}{2\sigma^2}} \\
&= e^{-\frac{Y'Y-2\beta'X'Y+\beta'X'X\beta}{2\sigma^2}} \\
&= e^{-\frac{(\beta-\hat{\beta}_{OLS})'(X'X)(\beta-\hat{\beta}_{OLS})+Y'(1-P_X)Y}{2\sigma^2}} \\
&\propto e^{-\frac{(\beta-\hat{\beta}_{OLS})'(X'X)(\beta-\hat{\beta}_{OLS})}{2\sigma^2}}
\end{aligned}$$

Therefore we have that $\beta_0, \beta_1|\sigma^2, Y \sim N_2(\hat{\beta}_{OLS}, \sigma^2(X'X)^{-1} = \Sigma)$. Furthermore we have:

$$\begin{aligned}
\pi(\sigma^2|Y) &= \frac{\pi(\beta, \sigma^2|Y)}{\pi(\beta|\sigma^2, Y)} \\
&\propto \frac{\left(\frac{1}{\sigma^2}\right)^{n/2+1} e^{-\frac{\|Y-X\beta\|^2}{2\sigma^2}}}{\left(\frac{1}{\sigma^2}\right) e^{-(\beta-\hat{\beta}_{OLS})'\Sigma^{-1}(\beta-\hat{\beta}_{OLS})}} \\
&\propto \left(\frac{1}{\sigma^2}\right)^{\frac{n-2}{2}+1} e^{-\frac{\left\|Y-X\hat{\beta}_{OLS}\right\|^2}{2\sigma^2}}
\end{aligned}$$

Therefore we have that $\sigma^2|Y \sim$Inv-Gamma($\frac{n-2}{2}, S^2/2 = \left\|Y - X\beta\right\|^2/2$). Now we compute

$\beta|Y$

$$\begin{aligned}
\pi(\beta_0,\beta_1|Y) &= \int_0^\infty \pi(\beta_0,\beta_1|\sigma^2,Y)\pi(\sigma^2|Y)d\sigma^2 \\
&= \int_0^\infty \left(\frac{1}{\sqrt{(2\pi)^2\,|\Sigma|}}\right) e^{-\frac{1}{2}(\beta-\hat\beta_{OLS})'\Sigma^{-1}(\beta-\hat\beta_{OLS})} \left(\frac{(S^2/2)^{(n-2)/2}}{\Gamma((n-2)/2)}\right)(\sigma^2)^{-\frac{n-2}{2}-1}exp\{-S^2/2\sigma^2 \\
&= \int_0^\infty \left(\frac{1}{\sqrt{(2\pi)^2}\sigma^2\,|X'X|^{1/2}}\right) e^{-\frac{(\beta-\hat\beta_{OLS})'(X'X)^{-1}(\beta-\hat\beta_{OLS})}{2\sigma^2}} \left(\frac{(S^2/2)^{(n-2)/2}}{\Gamma((n-2)/2)}\right)(\sigma^2)^{-\frac{n-2}{2}-1}exp \\
&= \left(\frac{(S^2/2)^{(n-2)/2}}{\Gamma((n-2)/2)\,|X'X|^{1/2}\sqrt{(2\pi)^2}}\right)\int_0^\infty e^{-\frac{2(\beta-\hat\beta_{OLS})'(X'X)^{-1}(\beta-\hat\beta_{OLS})+Y'(1-P_X)Y}{2\sigma^2}}(\sigma^2)^{-\frac{n-2}{2}-2}d \\
&= \left(\frac{(S^2/2)^{(n-2)/2}}{\Gamma((n-2)/2)\,|X'X|^{1/2}\sqrt{(2\pi)^2}}\right)\frac{(\frac{n-2}{2})!2^{\frac{n}{2}}}{[(\beta-\hat\beta_{OLS})'(X'X)^{-1}(\beta-\hat\beta_{OLS})+S^2]^{\frac{n}{2}}} \\
&= \left(\frac{1}{\Gamma((n-2)/2)\,|X'X|^{1/2}\,\pi}\right)\frac{\Gamma(n/2)}{[\frac{1}{S^2}(\beta-\hat\beta_{OLS})'(X'X)^{-1}(\beta-\hat\beta_{OLS})+1]^{\frac{n}{2}}}
\end{aligned}$$

Thus we have that $\pi(\beta_0,\beta_1|Y)\sim t(2,\hat\beta_{OLS},(n-2)(X'X)^{-1}/S^2)$. Therefore a good estimate for β is $\hat\beta_{ols}$. Therefore to fit the model we may use ordinary least squares estimate with $X=[x_1,x_2]$ with $x_1=(1,1,1,1,1,1)'$ and $x_2=(1,2,3,4,5,6)'$.

ii. For each swimmer j, obtain a posterior predictive distribution for Y_j^*, their time if they were to swim two weeks from the last recorded time.
We have that

$$\begin{aligned}
P(Y_j^*|Y_j) &= \int\int P(Y_j^*|Y_j,\sigma^2,\beta)P(\beta,\sigma^2|Y_j)d\beta d\sigma^2 \\
&= \int \frac{1}{\sqrt{2\pi}}e^{-\frac{(y_j-\beta_1-8\beta_2)^2}{2}}\frac{1}{\sqrt{2\pi}}e^{\frac{(\beta-\hat\beta_{ols})'(X'X)(\beta-\hat\beta_{ols})}{2}}d\beta \\
&\propto
\end{aligned}$$

b) The coach of the team has to decide which of the four swimmers will compete in a swimming meet in two weeks. Using your predictive distributions, compute $P(Y_j^*=\max\{Y_1^*,\ldots,Y_4^*\}|Y))$ for each swimmer j, and based on this make a recommendation to the coach

```
#PRoblem 1
getwd();
setwd('Google␣Drive/m9810Fall2014');
yblue=scan('bluecrab.dat');
yorange=scan('orangecrab.dat');
yblue.mat=matrix(yblue,nrow=50,ncol=2,byrow=TRUE)
yorange.mat=matrix(yorange,nrow=50,ncol=2,byrow=TRUE)
y=matrix(c(yblue.mat[,1],yblue.mat[,2],yorange.mat[,1],yorange.mat[,2]),nrow=50,ncol=
library(coda)
install.packages('mvtnorm')
library(mvtnorm)
install.packages('MCMCpack');
library(MCMCpack);
#initialize parameters
mu0=c(mean(yblue.mat[,1]),mean(yblue.mat[,2]),mean(yorange.mat[,1]),mean(yorange.mat[
Ybar=mu0;
cb=cov(yblue.mat);
co=cov(yorange.mat);
lambda0=matrix(c(cb[1,1],cb[1,2],0,0,cb[2,1],cb[2,2],0,0,0,0,co[1,1],co[1,2],0,0,co[2
S0=lambda0;
nu0=4;
n=dim(y)[1];
# save mcmc
iter=1e4;
thin = 10;
Mu = matrix(NA, nrow = iter / thin, ncol = 4);
Phi = matrix(NA, nrow = iter / thin, ncol = 16);
Sigma = matrix(NA, nrow = iter / thin, ncol = 16);

# intial values
mu = rmvnorm(1, mu0, lambda0);
Mu[1, ] = mu;
phi = rwish(nu0, solve(S0));
Phi[1, ] = c(phi);
Sigma[1, ] = c(solve(phi));

inv.lambda0 = solve(lambda0);

## Gibbs sampling
for(t in 2:iter){

  # t = 2;
  ## update mu
  mu.cov = solve( inv.lambda0 + n * phi );
  mu.mean = mu.cov %*% (n * phi %*% Ybar + inv.lambda0 %*% mu0);
  mu = rmvnorm(1, mu.mean, mu.cov);

  ## update phi
  sse = t(sweep(y, 2, mu)) %*% sweep(y, 2, mu);
  phi = rwish(n + nu0, solve(S0 + sse));

  ## save record
  if(t %% thin == 0){
    Mu[t / thin, ] = mu;
    Phi[t / thin, ] = c(phi);
    Sigma[t / thin, ] = c(solve(phi));
  }
}

## traceplot
Mu.mcmc = as.mcmc(Mu);
Phi.mcmc = as.mcmc(Phi);

plot(Mu.mcmc);
autocorr.plot(Mu.mcmc);
```

```
plot(Phi.mcmc);
autocorr.plot(Phi.mcmc);

plot(Mu[,1],Mu[,2],main="blue␣data", xlab="depth␣(mm)", ylab="width␣(mm)")
plot(Mu[,3],Mu[,4],main="orange␣data", xlab="depth␣(mm)", ylab="width␣(mm)")
##
#P(rho blue < rho orange| data)
mean(Sigma[501:1000, 2] / sqrt(Sigma[501:1000, 1] * Sigma[501:1000, 6]) < Sigma[501:1



##problem 2
getwd();
setwd('Google␣Drive/m9810Fall2014');
yswim=scan('swim.dat');
yswim.mat=matrix(yswim,nrow=4,ncol=6,byrow=TRUE)
yswim.one=yswim.mat[1,]
yswim.two=yswim.mat[2,]
yswim.three=yswim.mat[3,]
yswim.four=yswim.mat[4,]
X.mat=matrix(c(rep(1,6),seq(1,6,1)),nrow=6,ncol=2)
beta.ols.one=solve((t(X.mat) %*% X.mat))%*%t(X.mat)%*% yswim.one
beta.ols.two=solve((t(X.mat) %*% X.mat))%*%t(X.mat)%*% yswim.two
beta.ols.three=solve((t(X.mat) %*% X.mat))%*%t(X.mat)%*% yswim.three
beta.ols.four=solve((t(X.mat) %*% X.mat))%*%t(X.mat)%*% yswim.four
```