Introduction to Bayesian Data Analysis
Tutorial 6

(1) Problem 7.3 (Hoff) Australian crab data: The files
`bluecrab.dat` and `orangecrab.dat` contain measurements of body depth (Y_1) and rear width (Y_2), in millimetres 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 $\boldsymbol{\theta}$ and covariance matrix Σ as follows: Using the semi-conjugate prior distributions for $\boldsymbol{\theta}$ and Σ, set $\boldsymbol{\mu}_0$ equal to the sample mean of the data and Λ_0 and $\mathbf{S}_0$ equal to the sample covariance matrix and $\nu_0 = 4$. Obtain 10,000 posterior samples of $\boldsymbol{\theta}$ and Σ. Note that this "prior" distribution loosely centres the parameters around the empirical estimates based on the observed data (and is very similar to the unit information prior for the multivariate normal distribution). It cannot be considered as our true prior distribution, as it was derived from the observed data. However, it can be roughly considered as the prior distribution of someone with weak but unbiased information.

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

(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 $Pr(\rho_{\text{blue}} < \rho_{\text{orange}} | \mathbf{y}_{\text{blue}}, \mathbf{y}_{\text{orange}})$. What do the results suggest about differences between the two populations?

(2) Problem 7.4 (Hoff) Marriage data: The file `agehw.dat` contains data on the ages of 100 married couples from the US population.

(a) Before you look at the data, use your own knowledge to formulate a semiconjugate prior distribution for $\boldsymbol{\theta} = (\theta_h, \theta_w)^T$ and Σ, where θ_h, θ_w are mean husband and wife ages, and Σ is the covariance matrix.

(b) Generate a *prior predictive dataset* of size $n = 100$, by sampling $(\boldsymbol{\theta}, \Sigma)$ from your prior distribution and then simulating $\mathbf{Y}_1, ...\mathbf{Y}_n \overset{\text{iid}}{\sim}$ multivariate normal$(\boldsymbol{\theta}, \Sigma)$. Generate several such datasets, make bivariate scatter plots for each dataset, and make sure they roughly represent your prior beliefs about what such a dataset would actually look like. If your prior predictive datasets do not conform to your beliefs, go back to part (a) and formulate a new prior. Report the prior that you eventually decide upon, and provide scatter plots for at least three prior predictive datasets.

(c) Using your prior distributions and the 100 values in the dataset, obtain an MCMC approximation to $p(\boldsymbol{\theta}, \Sigma|\mathbf{y}_1, ...\mathbf{y}_{100})$. Plot the joint posterior distribution of θ_h and θ_w, and also the marginal posterior density of the correlation between Y_h and Y_w, the ages of a husband and wife. Obtain 95% posterior confidence intervals for θ_h and θ_w and the correlation coefficient.