Bayesian Statistics
Statistics 4224/5224 — Fall 2023

Homework 7

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

1. The file `http://www.stat.duke.edu/~pdh10/FCBS/Exercises/interexp.dat` contains data from an experiment that was interrupted before all the data could be gathered. Of interest was the difference in reaction times of experimental subjects when they were given stimulus A versus stimulus B. Each subject is tested under one of the two stimuli on their first day of participation in the study, and is tested under the other stimulus at some later date. Unfortunately the experiment was interrupted before it was finished, leaving the researchers with 26 subjects with both A and B responses, 15 subjects with only A responses, and 17 subjects with only B responses.

   (a) *Using only complete records:* For the first part of this problem, retain only the complete data cases, and discard the records corresponding to missing observations. Assuming a bivariate normal sampling model, and using the 'unit information prior' described in Homework 6 Problem 1, run 5000 iterations of the Gibbs sampler to obtain an MCMC approximation to $p(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{26})$. Compute the posterior expectation and a 95% posterior confidence interval for $\mu_B - \mu_A$.

   (b) *Treating imputed values as 'real' data:* Calculate empirical estimates of μ_A, μ_B, σ_A, σ_B, and ρ using the R functions `mean()`, `sd()`, and `cor()`. Use all the A responses to get $\hat{\mu}_A$ and $\hat{\sigma}_A$, and use all the B responses to get $\hat{\mu}_B$ and $\hat{\sigma}_B$; use only the complete data cases to get $\hat{\rho}$. For each subject with only an A response, $i \in \{27, \ldots, 41\}$, impute a B response as

$$\hat{y}_{i,B} = \hat{\mu}_B + \hat{\rho}\frac{\hat{\sigma}_B}{\hat{\sigma}_A}(y_{i,A} - \hat{\mu}_A) \; .$$

   For each subject with only a B response, $i \in \{42, \ldots, 58\}$, impute an A response as

$$\hat{y}_{i,A} = \hat{\mu}_A + \hat{\rho}\frac{\hat{\sigma}_A}{\hat{\sigma}_B}(y_{i,B} - \hat{\mu}_B) \; .$$

   You now have two 'observations' for each individual. Assuming a bivariate normal sampling model and a 'unit information prior' distribution, run 5000 iterations of the

Gibbs sampler to obtain an MCMC approximation to $p(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{26}, \hat{\boldsymbol{y}}_{27}, \ldots, \hat{\boldsymbol{y}}_{58})$. Compute the posterior expectation and a 95% posterior confidence interval for $\mu_B - \mu_A$.

(c) *Imputation done correctly:* Assuming a bivariate normal sampling model and a 'unit information prior' distribution, implement a Gibbs sampler (at least 10,000 iterations) that approximates the joint posterior distribution of the parameters and the missing data. Compute the posterior expectation as well as a 95% posterior confidence interval for $\mu_B - \mu_A$.

(d) Compare the results of parts (a), (b), and (c), and discuss.

The Gibbs sampler for missing data imputation, as needed for part (c) of this problem, is implemented as follows.

Given starting values $\boldsymbol{\mu}^1$ *and* $\boldsymbol{\Sigma}^1$*; for* $s = 1, \ldots, S-1$*:*

- *Impute the missing* y_B*-values: For* $i = 27, \ldots, 41$*,*

$$y_{i,B}^{s+1} \sim \text{Normal}[\mu_B^s + \rho^s \frac{\sigma_B^s}{\sigma_A^s}(y_{i,A} - \mu_A^s),\ \sigma_B^{2(s)}(1 - \rho^{2(s)})] \ .$$

- *Impute the missing* y_A*-values: For* $i = 42, \ldots, 58$*,*

$$y_{i,A}^{s+1} \sim \text{Normal}[\mu_A^s + \rho^s \frac{\sigma_A^s}{\sigma_B^s}(y_{i,B} - \mu_B^s),\ \sigma_A^{2(s)}(1 - \rho^{2(s)})] \ .$$

- *Update* $\boldsymbol{\mu}$ *and* $\boldsymbol{\Sigma}$ *in the usual way,*

$$\boldsymbol{\mu}^{s+1} \sim p(\boldsymbol{\mu}|\boldsymbol{\Sigma}^s, \boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}^{s+1})$$

and

$$\boldsymbol{\Sigma}^{s+1} \sim p(\boldsymbol{\Sigma}|\mu^{s+1}, \boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}^{s+1})$$

using the appropriate bivariate normal and inverse Wishart distributions.

2. Researchers interested in identifying the optimal planting density for a type of perennial grass performed the following randomized experiment: Ten different plots of land were divided into eight subplots, and planting densities of 2, 4, 6 and 8 plants per square meter were randomly assigned to the subplots, so that there are two subplots at each density in each plot. At the end of the growing season the amount of plant matter yield was recorded in metric tons per hectare. The data are available at `http://www2.stat.duke.edu/~pdh10/FCBS/Exercises/pdensity.dat`.

Letting $Y_{i,j}$ denote the yield and $x_{i,j}$ the planting density in the ith subplot of the jth plot, for $i = 1, \ldots, 8$ and $j = 1, \ldots, m = 10$, we will analyze these data using the hierarchical linear model

$$Y_{i,j}|\boldsymbol{\beta}_j, \sigma^2 \sim \text{indep Normal}(\beta_{1,j} + \beta_{2,j}x_{i,j} + \beta_{3,j}x_{i,j}^2, \sigma^2) \ .$$

The across-plot heterogeneity among the regression coefficients will be modeled by

$$\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m|\boldsymbol{\mu}, \boldsymbol{\Sigma} \sim \text{iid Normal}_3(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \ .$$

(a) First fit the model $y = \beta_1 + \beta_2 x + \beta_3 x^2 + \epsilon$ using OLS within each group.

   i. Make a plot showing the heterogeneity of the estimated regression curves.

   ii. From the least squares coefficients, find *ad hoc* estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Also obtain an estimate of σ^2 by combining the information from the residuals across the groups. Denote these estimates by $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\Sigma}}$, $\hat{\sigma}^2$.

Now we will conduct a Bayesian analysis of the data using independent priors

$$\begin{aligned}
\boldsymbol{\mu} &\sim \text{Normal}_3(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0) \\
\boldsymbol{\Sigma} &\sim \text{Inverse-Wishart}(\eta_0, \mathbf{S}_0^{-1}) \\
\sigma^2 &\sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2) \ ,
\end{aligned}$$

with $\boldsymbol{\mu}_0 = \hat{\mu}$, $\eta_0 = p + 2 = 5$ and $\boldsymbol{\Lambda}_0 = \mathbf{S}_0 = \hat{\boldsymbol{\Sigma}}$; $\nu_0 = 1$ and $\sigma_0^2 = \hat{\sigma}^2$.

Such a prior distribution can be roughly interpreted as the belief of an individual who has weak but unbiased prior information.

(b) Use a Gibbs sampler to approximate posterior expectations of $\boldsymbol{\beta}_j$ for each group, and plot the resulting regression curves. Compare to the curves in part (a) and describe why you see any differences between the two sets of regression curves.

(c) From your posterior samples, plot marginal posterior and prior densities of the elements of $\boldsymbol{\mu}$. Plot the posterior and prior densities of σ^2.

(d) Suppose we want to identify the planting density that maximizes the expected yield, $\mu_1 + \mu_2 x + \mu_3 x^2$. Let $x_{\max}$ denote the value of x that maximizes expected yield.

   i. Plot the posterior density of $x_{\max}$ and give a 95% confidence interval for $x_{\max}$.

   ii. Obtain $\hat{x}_{\max}$, the posterior expectation of $x_{\max}$.

   iii. Provide a 95% prediction interval for the yield on a randomly selected plot planting density $\hat{x}_{\max}$.

In 50 words or less, summarize what was learned from this analysis about the optimal plotting density and potential yield per hectare.

3. Younger male sparrows may or may not nest during a mating season, perhaps depending on their physical characteristics. Researchers have recorded the nesting success of 43 young male sparrows of the same age, as well as their wingspan, and the data appear in the file `http://www2.stat.duke.edu/~pdh10/FCBS/Exercises/msparrownest.dat`.

   Let Y_i be the binary indicator that the ith sparrow successfully nests, and let x_i denote their wingspan. Our model for Y_i is

   $$\text{logit}[\Pr(Y_i = 1|\alpha, \beta, x_i)] = \alpha + \beta x_i \ ,$$

   where the logit function is given by $\text{logit}(\theta) = \log[\theta/(1-\theta)]$.

   Use independent prior distributions $\alpha \sim \text{Normal}(0, 10^2)$ and $\beta \sim \text{Normal}(0, 2^2)$.

   (a) Implement a Markov chain Monte Carlo algorithm to approximate $p(\alpha, \beta|\boldsymbol{y}, \boldsymbol{x})$. Run the algorithm long enough so that the effective sample size is at least 1000 for each parameter. Display trace plots and sample autocorrelation functions, and report effective sample sizes.

   (b) Compare the posterior densities of α and β to their prior densities.

   (c) Using output from the MCMC algorithm, come up with a way to make a 90% confidence band for the following function $f_{\alpha,\beta}(x)$ of wingspan:

   $$f_{\alpha,\beta}(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} \ ,$$

   where α and β are the parameters in your sampling model. Make a plot of this band, and discuss your results.

   It would be good practice to code your own Metropolis algorithm for this problem.

   Letting (α^s, β^s) denotes the current state of the chain, the updated state $(\alpha^{s+1}, \beta^{s+1})$ is generated as follows:

   - *Simulate the jump proposal by*

   $$\begin{pmatrix} \alpha^* \\ \beta^* \end{pmatrix} \sim \text{Normal}_p \left[\begin{pmatrix} \alpha^s \\ \beta^s \end{pmatrix}, \ \delta^2(\mathbf{X}^T\mathbf{X})^{-1} \right] \ .$$