The Chinese University of Hong Kong, Shenzhen
SDS · School of Data Science

· Fall Semester 2022

DDA 4010 – Bayesian Statistics

Exercise Sheet 4

This exercise is due on **Nov. 25th., 5:00 pm.**

Assignment A4.1 (7.2 in Textbook):
Unit information prior: Letting $\Psi=\Sigma^{-1}$, show that a unit information prior for $(\boldsymbol{\theta}, \Psi)$ is given by $\boldsymbol{\theta} \mid \Psi \sim$ multivariate normal $\left(\overline{\boldsymbol{y}}, \Psi^{-1}\right)$ and $\Psi \sim$ Wishart $\left(p+1, \mathbf{S}^{-1}\right)$, where $\mathbf{S}=\sum\left(\boldsymbol{y}_i-\overline{\boldsymbol{y}}\right)\left(\boldsymbol{y}_i-\overline{\boldsymbol{y}}\right)^T / n$. This can be done by mimicking the procedure outlined in Exercise 5.6 as follows:

- Reparameterize the multivariate normal model in terms of the precision matrix $\Psi=\Sigma^{-1}$. Write out the resulting log likelihood, and find a probability density $p_U(\boldsymbol{\theta}, \Psi)=p_U(\boldsymbol{\theta} \mid \Psi) p_U(\Psi)$ such that $\log p(\boldsymbol{\theta}, \Psi)=l(\boldsymbol{\theta}, \Psi \mid \mathbf{Y}) / n+c$, where c does not depend on $\boldsymbol{\theta}$ or Ψ. Hint: Write $\left(\boldsymbol{y}_i-\boldsymbol{\theta}\right)$ as $\left(\boldsymbol{y}_i-\overline{\boldsymbol{y}}+\overline{\boldsymbol{y}}-\boldsymbol{\theta}\right)$, and note that $\sum \boldsymbol{a}_i^T \mathbf{B} \boldsymbol{a}_i$ can be written as $\operatorname{tr}(\mathbf{A B})$, where $\mathbf{A}=\sum \boldsymbol{a}_i \boldsymbol{a}_i^T$.
- Let $p_U(\Sigma)$ be the inverse-Wishart density induced by $p_U(\Psi)$. Obtain a density $p_U\left(\boldsymbol{\theta}, \Sigma \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\right) \propto p_U(\boldsymbol{\theta} \mid \Sigma) p_U(\boldsymbol{\Sigma}) p\left(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \mid \boldsymbol{\theta}, \Sigma\right)$. Can this be interpreted as a posterior distribution for θ and Σ ?

Assignment A4.2 (7.3 in Textbook):
JAustralian crab data: The files bluecrab. dat and orangecrab. dat contain measurements of body depth (Y_1) and rear 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.

- For each of the two species, obtain posterior distributions of the population mean $\boldsymbol{\theta}$ and covariance matrix Σ as follows: Using the semiconjugate prior distributions for $\boldsymbol{\theta}$ and Σ, set $\boldsymbol{\mu}_0$ equal to the sample mean of the data, 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 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 be roughly considered as the prior distribution of someone with weak but unbiased information.
- Plot values of $\boldsymbol{\theta}=(\theta_1, \theta_2)'$ for each group and compare. Describe any size differences between the two groups.
- 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 $\operatorname{Pr}\left(\rho_{\text{blue}}<\rho_{\text{orange}} \mid \boldsymbol{y}_{\text{blue}}, \boldsymbol{y}_{\text{orange}}\right)$. What do the results suggest about differences between the two populations?

Assignment A4.3 (7.5 **in Textbook):**
Imputation: The file 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.

- Calculate empirical estimates of $\theta_A, \theta_B, \rho, \sigma_A^2, \sigma_B^2$ from the data using the commands mean, cor and var. Use all the A responses to get $\hat{\theta}_A$ and $\hat{\sigma}_A^2$, and use all the B responses to get $\hat{\theta}_B$ and $\hat{\sigma}_B^2$. Use only the complete data cases to get $\hat{\rho}$.

- For each person i with only an A response, impute a B response as

  $$\hat{y}_{i,B} = \hat{\theta}_B + \left(y_{i,A} - \hat{\theta}_A\right) \hat{\rho} \sqrt{\hat{\sigma}_B^2 / \hat{\sigma}_A^2}$$

  For each person i with only a B response, impute an A response as

  $$\hat{y}_{i,A} = \hat{\theta}_A + \left(y_{i,B} - \hat{\theta}_B\right) \hat{\rho} \sqrt{\hat{\sigma}_A^2 / \hat{\sigma}_B^2}$$

  You now have two "observations" for each individual. Do a paired sample t-test and obtain a 95% confidence interval for $\theta_A - \theta_B$.

- Using either Jeffreys' prior or a unit information prior distribution for the parameters, implement a Gibbs sampler that approximates the joint distribution of the parameters and the missing data. Compute a posterior mean for $\theta_A - \theta_B$ as well as a 95% posterior confidence interval for $\theta_A - \theta_B$. Compare these results with the results from b) and discuss.