AUSTRALIAN NATIONAL UNIVERSITY

RESEARCH SCHOOL OF FINANCE ACTUARIAL STUDIES, AND APPLIED STATISTICS

INTRODUCTION TO BAYESIAN DATA ANALYSIS (STAT3016/4116/7016)

SEMESTER 2 2016

ASSIGNMENT 3

DUE DATE: Thursday 22 September 2016, by 4pm

(12.5% of total course grade)

INSTRUCTIONS:

1. All students must hand in an assignment of their own writing.
2. The assignment should be handed in to the assignment box for STAT3016/7016 available on level 4 of the ANUCBE Building 26C. There will be no online submission facility.
3. Ensure you also complete and attach a cover sheet to your assignment (available on the course website)
4. Begin each question on a new page.
5. Where required, provide sufficient computer output to support your answers. Provide enough intermediate numerical calculations to justify working for your final answer.
6. Computer output must be interpreted in written format. A solution solely highlighting the computer output is not acceptable.
7. No late assignments will be accepted.

COLLABORATION POLICY (as stated in the course outline)

University policies on plagiarism will be **strictly** enforced. You are encouraged to (orally) discuss your assignments with your classmates, but each student must write up solutions separately. Be sure that you have worked through each problem yourself and that all answers you submit are the results of your own efforts. This includes all computer code and output.

Problem 1 [STAT4116/7016 ONLY]

Parameter expansion and the t model

Suppose we have n independent data points from the $t_\nu(\mu, \sigma^2)$ distribution and we assume the degrees of freedom ν is known. The t likelihood for each data point is equivalent to the model:

$$y_i \sim N(\mu, V_i)$$
$$V_i \sim \text{InvGamma}(\nu/2, \nu\sigma^2/2)$$

A Gibbs sampler can be used to obtain posterior draws of μ, σ^2, and each V_i. However, convergence will be slow if a simulation draw of σ^2 is close to zero. We can add an extra parameter as follows:

$$y_i \sim N(\mu, \alpha^2 U_i)$$
$$U_i \sim \text{InvGamma}(\nu/2, \nu\tau^2/2)$$

where $\alpha > 0$ is an additional scale parameter. The parameter α has no meaning and its only role is to allow the Gibbs sampler to move in more directions to avoid getting stuck.

Assuming a uniform prior distribution on μ, $\log\tau$ and $\log\alpha$, derive the four steps of the Gibbs sampler for the expanded model. That is, derive the conditional distributions for U_i $(i = 1, ..., n)$, μ, τ^2 and α^2.

Problem 2

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.

(a) Calculate empirical estimates of θ_A, θ_B, ρ, σ_A^2, σ_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 $\hat{\rho}$.

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

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

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

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

You now have two "observations" for each individual. Do a paired sample t-test and obtain a 95% confidence interval for $\hat{\theta}_A - \hat{\theta}_B$.

(c) Use the Jeffreys' prior for multivariate normal data ($p_J(\boldsymbol{\theta}, \Sigma) \propto |\Sigma|^{-(p+2)/2}$ where p is the dimension of each observable vector $\mathbf{y_i}$) for the parameters, and 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 part (b) and discuss.

Note, you may use the following results for the posterior distributions:

$$\Sigma|\mathbf{y_1}, ..., \mathbf{y_n} \sim \text{Inverse-Wishart}(n-1, S^{-1})$$

.

$$\boldsymbol{\theta}|\Sigma, \mathbf{y_1}, ..., \mathbf{y_n} \sim \text{Multivariate Normal}(\bar{\mathbf{y}}, \Sigma/n)$$

where $S = \sum_{i=1}^{n}(\mathbf{y_i} - \bar{\mathbf{y}})(\mathbf{y_i} - \bar{\mathbf{y}})^T$

(d) Contrast the two expressions for the imputed values $\hat{y}_{i,B} = \hat{\theta}_B$ and $\hat{y}_{i,B} = \hat{\theta}_B + (y_{i,A} - \hat{\theta}_A)\hat{\rho}\sqrt{\hat{\sigma}_B^2/\hat{\sigma}_A^2}$. Explain (in words) the differences in any underlying assumptions between the two expressions.

Problem 3

A population of 532 women living near Phoenix, Arizona were tested for diabetes. Other information was gathered from these women at the time of testing, including number of pregnancies, glucose level, blood pressure, skin fold thickness, body mass index, diabetes pedigree and age. This information appears in the file `azdiabetes.dat`. Model the joint distribution of these variables for the diabetics and non-diabetics separately, using a multivariate normal distribution:

(a) For both groups separately, use the following type of unit information prior, where $\hat{\Sigma}$ is the sample covariance matrix.

(i) $\boldsymbol{\mu}_0 = \bar{\boldsymbol{y}}$, $\Lambda_0 = \hat{\Sigma}$

(ii) $S_0 = \hat{\Sigma}$, $\nu_0 = p + 2 = 9$

Generate at least 10,000 (Markov Chain) Monte Carlo samples for $\{\boldsymbol{\theta}_d, \Sigma_d\}$ and $\{\boldsymbol{\theta}_n, \Sigma_n\}$, the model parameters for diabetics and non-diabetics respectively. For each of the seven variables $j \in \{1, ..., 7\}$, compare the marginal posterior distributions of $\theta_{d,j}$ and $\theta_{n,j}$. Which variables seem to differ between the two groups? Also obtain $Pr(\theta_{d,j} > \theta_{n,j}|\mathbf{Y})$ for each $j \in \{1, ..., 7\}$.

(b) Obtain the posterior means of Σ_d and Σ_n, and plot the entries versus each other. What are the main differences, if any?

Problem 4

Probit regression: A panel study followed 25 married couples over a period of five years. One item of interest is the relationship between divorce rates and the various characteristics of the couples. For example, the researchers would like to model the probability of divorce as a function of age differential, recorded as the man's age minus the women's age. The data can be found in the file `divorce.dat`. We will model these data with a probit regression, in which a binary variable Y_i is described in terms of an explanatory variable x_i via the following latent variable model:

$$Z_i = \beta x_i + \epsilon_i$$
$$Y_i = \delta_{(c,\infty)}(Z_i) \ ,$$

where β and c are unknown coefficients, $\epsilon_1, ..., \epsilon_n \overset{\text{iid}}{\sim} \text{normal}(0,1)$ and $\delta_{(c,\infty)} = 1$ if $z > c$ and equals zero otherwise.

(a) Assuming $\beta \sim \text{normal}(0, \tau_\beta^2)$ obtain the full conditional distribution $p(\beta|\mathbf{y}, \mathbf{x}, \mathbf{z}, c)$.

(b) Assuming $c \sim \text{normal}(0, \tau_c^2)$, show that $p(c|\mathbf{y}, \mathbf{x}, \mathbf{z}, \beta)$ is a constrained normal density, i.e proportional to a normal density but constrained to lie in an interval. Similarly, show that $p(z_i|\mathbf{y}, \mathbf{x}, \mathbf{z}_{-i}, \beta, c)$ is proportional to a normal density but constrained to be either above c or below c, depending on y_i.

(c) Letting $\tau_\beta^2 = \tau_c^2 = 16$, implement a Gibbs sampling scheme that approximates the joint distribution of $\mathbf{Z}$, β, and c. Compute the effective sample sizes of all unknown parameters (including all the Z_i's). Also compute the autocorrelation function of the parameters and discuss the mixing of the Markov chain. (See Hoff Section 12.1.1 for a method on sampling from a constrained normal distribution).

(d) Obtain a 95% posterior confidence interval for β, as well as $Pr(\beta > 0|\mathbf{y}, \mathbf{x})$.