University of Edinburgh
School of Mathematics

Bayesian Data Analysis

Assessed Problem 2 Due 5 April, 11pm **Semester 2, 2016–2017**

Variance models for protein expression with missing data.

Description

Concentration of 460 proteins was measured for each of 9 rats using images of 2D gels. Sensitivity of the equipment resulted in concentration values less than 5 to be treated as missing.

The aim is to create an appropriate model for modelling variability of the concentration of proteins and of the missing data.

Notation

Let X_{gr} be the observed concentration of protein g in rat r, $g = 1, \ldots, 460$ and $r = 1, \ldots, 9$. It is known that the (unobserved) true concentration is different for different proteins however not much is known about their variability. A reasonable assumption is that their variability is similar or the same.

Typically, logarithm of the concentration X_{gr} is assumed to have a normal distribution: $Y_{gr} = \log(X_{gr})$,

$$Y_{gr} \mid \mu_g, \sigma_g^2 \sim N(\mu_g, \sigma_g^2) \text{ independently for all } g, r.$$

Statistical analysis

Modelling the variance

1. Propose a non-informative prior for μ_g, independently for all g.
2. Propose the following types prior distributions for the variance in your model:
   (a) equal variance for all proteins: $\sigma_g^2 = \sigma^2$ for all g, with a non-informative prior for σ^2
   (b) hierarchical models for protein-specific variance $\sigma_g^2 \mid \theta \sim p(\theta)$, considering at least two different families of distributions $p(\theta)$ (e.g. log normal and gamma), and propose non-informative priors for hyperparameters θ.
3. Fit the corresponding Bayesian model for each variance model in Question 2 (demonstrating convergence and justifying your choices of burnin and thinning) to the proteins without missing data (413 proteins, data file DataProtein413.txt for WinBUGS and DataProtein413.csv for R).
4. Check sensitivity to the prior of the parameters θ of the hierarchical model for the variances.
5. Determine the best model for the variances using DIC and posterior predictive checks, using mixed predictive checks where possible (include histograms of the posterior predictive p-values).

Missing values

6. Discuss whether the information available a priori indicates that the missing values are missing at random.

7. Fit the best model identified in Question 5 to the data for the remaining 47 proteins with missing observations in two ways:

   (a) ignore missing data and use only present observations as the data (data is in file DataProteinsMissing.txt for WinBUGS),

   (b) impute missing values from the model (data is in file DataProteinsMissingNA.txt for WinBUGS and DataProteinsMissingNA.csv for R).

   Use the information from the posterior distribution of the hyperparameters θ under the best model that you fitted to the 413 proteins without missing observations in Question 5 as the prior information.

8. Compare the posterior distributions of the variances (you can use summaries of the posterior distributions) under the two different models for handling the missing data. Discuss whether your conclusion is expected under the indicated missing mechanism discussed in Question 6.