University of Edinburgh
School of Mathematics

Bayesian Data Analysis

Solution to Exercise 1

Report on gene expression data analysis.

1 Description of the problem

Expression (concentration) of each of 50 genes was observed in the pancreas tissue of two groups of people: in a case group of 12 people who have pancreatic cancer, and in a control group of 10 people who do not have pancreatic cancer (the groups are approximately matched by age and gender). The study was performed in the same hospital.

The question is to determine whether there is a difference between the gene expression in the pancreas tissue between the groups of people with and without the cancer.

2 Likelihood

Introduce the following random variables. Denote the logarithm of the gene expression of the kth gene for individual i in the case group by X_{ik}, and for the jth individual in the control group by Y_{jk}, $k = 1, 2, \ldots, N = 50$, $i = 1, \ldots, n = 12$, $j = 1, 2, \ldots, m = 10$.

Possible likelihood:

$$X_{ik} \mid \mu_1, \sigma_1 \sim N(\mu_1, \sigma_1^2),\ i = 1, 2, \ldots, n \quad \textit{independently (given } \mu_1, \sigma_1^2),$$
$$Y_{jk} \mid \mu_2, \sigma_2 \sim N(\mu_2, \sigma_2^2),\ j = 1, 2, \ldots, m \quad \textit{independently (given } \mu_2, \sigma_2^2).$$

Also, X_{ik} and Y_{jk} are independent for all i, j, k.

This data set can be summarised using the following sufficient statistics for parameters μ_1, σ_1, μ_2 and σ_2. For cases,

$$\bar{X} \sim N(\mu_1, \sigma_1^2/(nN)), \quad S_x^2/\sigma_1^2 \sim \chi^2_{Nn-1}$$

and for controls,

$$\bar{Y} \sim N(\mu_2, \sigma_2^2/(mN)), \quad S_y^2/\sigma_2^2 \sim \chi^2_{Nm-1},$$

where

$$\bar{X} = \frac{1}{nN}\sum_{i=1}^{n}\sum_{k=1}^{N} X_{ik}, \quad \bar{Y} = \frac{1}{mN}\sum_{j=1}^{m}\sum_{k=1}^{N} Y_{jk},$$
$$s_X^2 = \frac{1}{nN}\sum_{i=1}^{n}\sum_{k=1}^{N}(X_{ik} - \bar{X})^2, \quad s_Y^2 = \frac{1}{mN}\sum_{j=1}^{m}\sum_{k=1}^{N}(Y_{jk} - \bar{Y})^2.$$

The assumptions made are normality of the log concentrations, independence of all observations, and that in each group, with and without cancer, all genes and all subjects

have the same group-specific mean and variance. The assumption of normality of log concentrations is commonly used in gene expression studies, as well as of the independence of observations between different subjects. Other assumptions will be discussed in the discussion (Section 6).

Means μ_1 and μ_2 can be interpreted as the population log concentration of the considered 50 genes among subjects with and without cancer, respectively, under the assumption that all 50 genes have the same mean and variance in each group of subjects. Hence, $\exp(\mu_1)$ and $\exp(\mu_2)$ can be interpreted as the population concentration of the considered 50 genes among subjects with and without cancer, respectively, under the same assumption.

To address the question whether there is a difference between the gene expression in the pancreas tissue between the groups of people with and without the cancer, we can test the hypothesis $\exp(\mu_1) = \exp(\mu_2)$, or, equivalently, since exp is a one-to-one function, hypothesis $\mu_1 = \mu_2$.

Below, we assume that $\sigma_1 = 0.3$ and $\sigma_2 = 0.1$ are known, hence there are two unknown parameters: μ_1 and μ_2.

3 Prior distribution

3.1 Available prior information

For the mean of log concentration in the cases group, there is no a priori information, hence a non-informative prior for μ_1 will be used.

For the mean of log concentration in the controls group μ_2, two types of statistical analysis will be run: one is without a priori information, and the other one is to use available prior information from a study in several other hospitals that a similar Bayesian analysis of the log gene expression of the same 50 genes in control groups only produced the posterior distribution of μ_2 to be $N(2.38, 0.04^2)$. In the second case, the posterior distribution from a previous study will be used as a prior distribution for μ_2. The implications of such choice of the prior are that the results of the corresponding statistical analysis will based on joint analysis of the data from the current study and of the control data from the previous studies.

3.2 Prior distribution(s)

For μ_1, we consider two non-informative priors:

$$\begin{aligned} &1) \quad p(\mu_1) = 1, \quad \mu_1 \in (-\infty, +\infty), \\ &2) \quad \mu_1 \sim N(0, A^2), \quad \mu_1 \in (-\infty, +\infty) \quad \text{for large } A. \end{aligned}$$

The first one is the Jeffreys prior for μ_1 which is improper, and the second one is a proper approximation to the Jeffreys prior from the conjugate family of priors which is a set of normal distributions, as A is large. We will take $A = 100$.

For μ_2, we consider the same two non-informative priors as for μ_1, and the informative prior from the previous studies:

$$\begin{aligned} &1) \quad p(\mu_2) = 1, \quad \mu_2 \in (-\infty, +\infty), \\ &2 \quad) \ \mu_2 \sim N(0, A^2), \quad A = 100, \\ &3) \quad \mu_2 \sim N(2.38, 0.04^2). \end{aligned}$$

	Mean	Median	s.d.	5% quantile	95% quantile
Uniform	4.030	4.030	0.0122	4.010	4.050
$N(0, A^2)$	4.030	4.030	0.0122	4.010	4.050

Table 1: Summaries of the posterior distribution of μ_1 for the considered priors.

	Mean	Median	s.d.	5% quantile	95% quantile
Uniform	2.590	2.590	0.004	2.583	2.597
$N(0, A^2)$	2.590	2.590	0.004	2.583	2.597
Informative	2.587	2.587	0.004	2.580	2.595

Table 2: Summaries of the posterior distribution of μ_2 for the considered priors.

4 Posterior inference

4.1 Posterior distribution

For each prior distribution for μ_1 and μ_2, the corresponding posterior distributions are stated below, using the following formula.

If $Z \mid \theta \sim N(\theta, \sigma^2)$ and $\theta \sim N(\theta_0, \sigma_0^2)$ then $\theta \mid Z \sim N(\mu_P, \sigma_P^2)$ where $\mu_P = Z\sigma^{-2}/(\sigma^{-2} + \sigma_0^{-2})$ and $\sigma_P^2 = 1/(\sigma^{-2} + \sigma_0^{-2})$.

For μ_1:

$$1) \quad \mu_1 \mid (x_{ik}) \sim N\left(\bar{x}, \frac{\sigma_1^2}{(nN)}\right) = N(4.030, 0.0122^2),$$

$$2) \quad \mu_1 \mid (x_{ik}) \sim N\left(\frac{\bar{x}}{1 + \sigma_1^2 A^{-2}/(nN)}, \frac{\sigma_1^2}{nN(1 + \sigma_1^2 A^{-2}/(nN))}\right) = N(4.030, 0.0122^2).$$

Note that for the considered precision, both types of non-informative priors give the same posterior distribution.

For μ_2:

$$1) \quad \mu_2 \mid (y_{jk}) \sim N\left(\bar{y}, \frac{\sigma_2^2}{(mN)}\right) = N(2.590, 0.0045^2),$$

$$2) \quad \mu_2 \mid (y_{jk}) \sim N\left(\frac{\bar{y}}{1 + \sigma_2^2 A^{-2}/(mN)}, \frac{\sigma_2^2}{mN(1 + \sigma_2^2 A^{-2}/(mN))}\right) = N(2.590, 0.0045^2),$$

$$3) \quad \mu_2 \mid (y_{jk}) \sim N\left(\frac{mN\sigma_2^{-2}\bar{y} + 0.04^{-2}2.38}{mN\sigma_2^{-2} + 0.04^{-2}}, \frac{\sigma_2^2}{mN(1 + \sigma_2^2 0.04^{-2}/(mN))}\right) = N(2.587, 0.0044^2).$$

Note that for the considered precision, both types of non-informative priors give the same posterior distribution, which is slightly affected by the prior in case 3).

4.2 Posterior summaries and plots

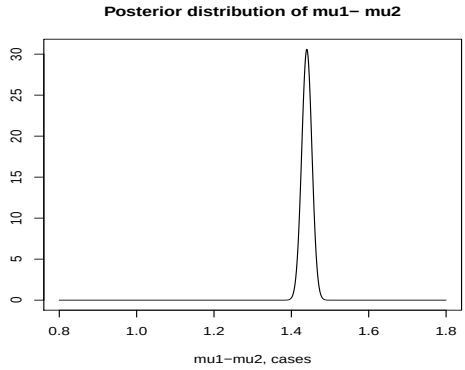
Summaries of the posterior distributions of μ_1 and μ_2 for each considered prior are given below.

4.3 Decisions

Here we consider only the improper uniform (Jeffreys) priors for both μ_1 and μ_2, as their posterior distributions are almost the same for the considered priors (see Section 4.2 and

Section 5 with further investigation of sensitivity to the prior).

The posterior distribution of $\mu_1 - \mu_2 \mid (x_{ik}), (y_{jk}) \sim N(1.44, 0.01362)$ which is plotted in the figure below. Two-sided equitailed 90% credible interval for $\mu_1 - \mu_2$ is $[1.4185541.461446]$ which excludes 0, hence the available data provides a strong evidence that $\mu_1 \neq \mu_2$. The posterior probability $P(\mu_1 - \mu_2 > 0 \mid (x_{ik}), (y_{jk})) = 1$ confirming this conclusion.

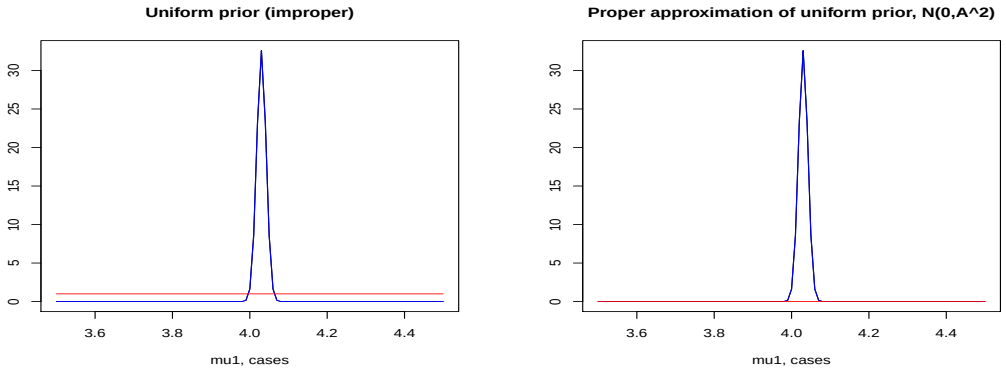


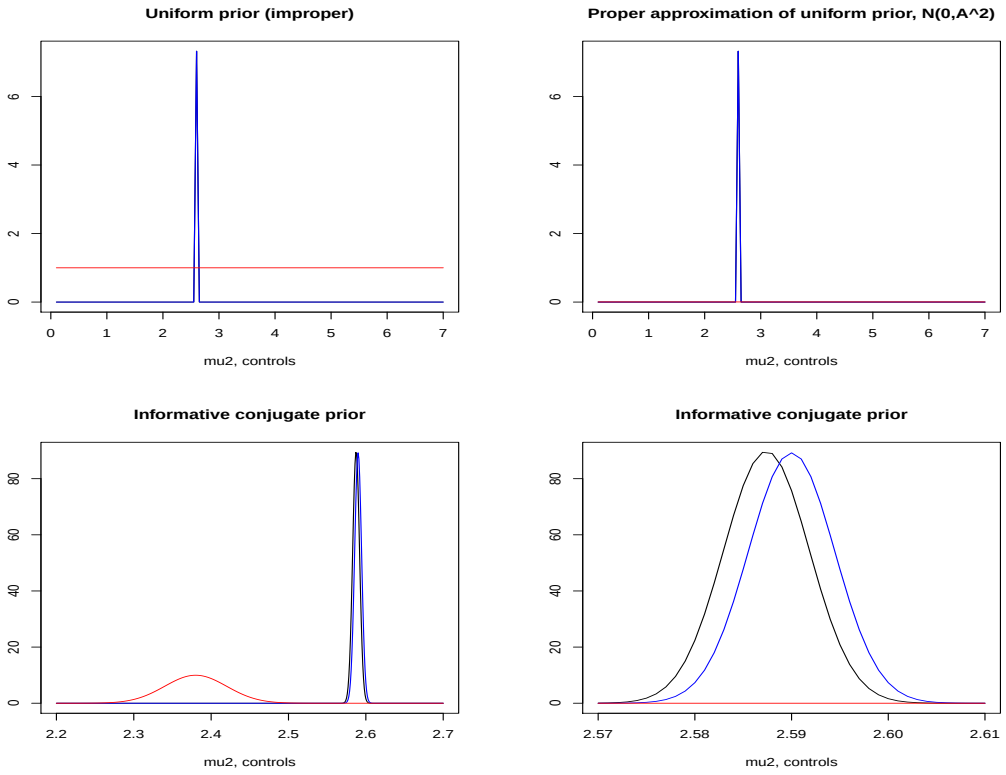
Also, we can consider the individual 90% credible intervals for μ_1 and μ_2, $[4.010, 4.050]$ and $[2.583, 2.597]$, which do not intersect, hence, with 80% confidence we reject hypothesis that $\mu_1 = \mu_2$. To improve the level of confidence, consider 1% and 99% quantiles of each of the posterior distribution which gives us 98% credible intervals $[4.001508, 4.058492]$ and $[2.579596, 2.600404]$ for μ_1 and μ_2 respectively. They do not overlap, hence, with 96% confidence, we reject the hypothesis that $\mu_1 = \mu_2$.

Therefore, the observed data provided strong evidence that there is a difference in gene expression between subjects with and without pancreatic cancer.

5 Sensitivity of posterior inference with respect to the choice of prior

Prior/likelihood/ posterior plots.





None of the plots shows a conflict between the non-informative priors and the likelihood for either parameters μ_1 and μ_2.

The major part of the support of the informative prior for μ_2 lies outside of the major part of the support of the likelihood, although the effect of this on the posterior distribution is quite weak due to a high amount of information in the likelihood, $mN = 500$ giving the observed precision $mN/\sigma_2^2 = 50000$ which is much higher than the precision of the prior $1/0.04^2 = 625$. Nevertheless, a further investigation may improve understanding of why this happened, and may allow to improve the statistical model and hence the precision of the statistical inference.

6 Discussion

The assumption of normality of the observations needs to be investigated for this data. The assumption of independence between subjects seems reasonable unless there are additional factors (covariates) which may result in closer similarity between subjects within group that between groups.

The assumption that the gene expression is the same for all genes may be questioned, and a model taking this into the account can improve the quality of statistical analysis, particularly of predictive inference. The same applies to independence between the genes as it is known that some genes work together and can be correlated.

Finally, using the posterior distribution of μ_2 from another study as a prior for current study may not be appropriate, as the previous study was carried out in several different hospitals, which may influence the quality of the data (e.g. due to different ways of collecting or processing data).

7 Summary

The observed data provided strong evidence that there is a difference in gene expression between subjects with and without pancreatic cancer.

Sensitivity of the posterior analysis to several priors was investigated, without changing the conclusion or the strength of the supporting evidence.

The assumption of independence between the genes and their distributions being identical may be too strong. Available a priori information about the gene expression in controls may need to be investigated further, e.g. by including a hospital effect, if this additional data to be incorporated in the current data analysis.

Since plug-in values of σ_1 and σ_2 were used, this may lead to overconfidence. Further analysis including their estimated values should give a more realistic estimation of uncertainty about μ_1 and μ_2.