ST3457 Statistical Inference I (Bayesian Statistics)

Lab 2 - 23/11/17

Bernardo Nipoti

Lab 2 - Normal model: joint inference for mean and variance

Read the sections *DATA*, *MODEL* and *POSTERIOR DISTRIBUTION*, and address the points listed at the end of this sheet, by working on the R script `LAB_Normal.R` that you can download from the course webpage `http://www.bernardonipoti.com/teaching/st3457_17_18`

DATA. As we did last week, we analyse the Iris dataset, available in the `Datasets` package in `R`, which collects different measurements of three species of iris flowers. Today we focus on the sepal length of these flowers. We have three species: *versicolor*, *setosa*, *virginica*. We let $Y_{1,1}, \ldots, Y_{n_1,1}$ be the observations for the first group (versicolor), that is $Y_{i,1}$ is the sepal length in *cm* of the ith observed flower of species versicolor. Similarly, $Y_{1,2}, \ldots, Y_{n_2,2}$ and $Y_{1,3}, \ldots, Y_{n_3,3}$ are the observations for second and third group (setosa and virginica, respectively).

MODEL. We assume a Normal sampling model for each group. Unlike last week, we consider a *two-dimensional* parameter (θ, σ^2), that is both the mean and the variance of each group are assumed unknown. We choose a conjugate normal/inverse-gamma prior for mean and variance. The priors for the three groups are assumed independent and identical. That is

$$
\begin{aligned}
Y_{1,1}, \ldots, Y_{n_1,1} \mid \theta_1, \sigma_1^2 &\overset{\text{iid}}{\sim} \text{Normal}(\theta_1, \sigma_1^2) \\
Y_{1,2}, \ldots, Y_{n_2,2} \mid \theta_2, \sigma_2^2 &\overset{\text{iid}}{\sim} \text{Normal}(\theta_2, \sigma_2^2) \\
Y_{1,3}, \ldots, Y_{n_3,3} \mid \theta_3, \sigma_3^2 &\overset{\text{iid}}{\sim} \text{Normal}(\theta_3, \sigma_3^2) \\
\sigma_i^2 &\overset{\text{iid}}{\sim} \text{inverse-gamma}(\nu_0/2, \sigma_0^2\nu_0/2) \\
\theta_i \mid \sigma_i^2 &\overset{\text{iid}}{\sim} \text{Normal}(\mu_0, \sigma_i^2/k_0).
\end{aligned}
$$

Recall that the parameters $\mu_0, k_0, \sigma_0^2, \nu_0$ have a neat interpretation:

- μ_0: prior guess for the mean,
- k_0: prior sample size (upon which, *ideally*, we base our prior guess μ_0),
- σ_0^2: prior guess for the variance,
- ν_0: prior sample size (upon which, *ideally*, we base our prior guess σ_0^2).

POSTERIOR DISTRIBUTION. The posterior distribution of each pair (θ_i, σ_i^2), for $i = 1, 2, 3$, is given by

$$
p(\theta_i, \sigma_i^2 \mid y_{i,1}, \ldots, y_{n_i,i}) = p(\sigma_i^2 \mid y_{i,1}, \ldots, y_{n_i,i}) p(\theta_i \mid \sigma_i^2, y_{1,i}, \ldots, y_{n_i,i})
$$

where

$$
\begin{aligned}
\sigma_i^2 \mid y_{1,i}, \ldots, y_{n_i,i} &\sim \text{inverse-gamma}(\nu_{n_i,i}/2, \sigma_{n_i,i}^2 \nu_{n_i,i}/2), \\
\theta_i \mid \sigma_i^2, y_{1,i}, \ldots, y_{n_i,i} &\sim \text{Normal}(\mu_{n_i,i}, \sigma_{n_i,i}^2/(k_0 + n_i)),
\end{aligned}
$$

where

$$\nu_{n_i,i} = \nu_0 + n_i, \qquad \text{and} \qquad \mu_{n_i,i} = \frac{k_0}{k_0 + n_i}\mu_0 + \frac{n_i}{k_0 + n_i}\bar{y}_i,$$

$$\sigma^2_{n_i,i} = \frac{1}{\nu_{n_i,i}}\left[\nu_0\sigma_0^2 + (n_i - 1)s_i^2 + \frac{k_0 n_i}{k_0 + n_i}(\bar{y}_i - \mu_0)^2\right]$$

and

$$\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{j,i} \qquad \text{and} \qquad s_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(y_{j,i} - \bar{y}_i)^2$$

are sample mean and sample variance for each group.

POINTS TO ADDRESS.

1. A botanist tells you that the mean sepal length is $6cm$ and that the variance across sepal lengths is about $0.2cm^2$. Set the values of the parameters $\mu_0, k_0, \sigma_0^2, \nu_0$ so to translate the botanist's guess and so to make the prior non-informative (i.e. small values for k_0 and ν_0).

2. Compute $\mu_{n_i,i}$, $\nu_{n_i,i}$ and $\sigma^2_{n_i,i}$ for the three groups $i = 1, 2, 3$.

3. Generate three Monte Carlo samples of size 2,000 from the joint posterior distribution of (θ_i, σ_i^2) for the three groups.

4. *Point estimates.* Starting from the Monte Carlo samples, estimate the posterior mean for θ_i and for σ_i^2, for each group, together with quantile-based 95% posterior credible intervals.

   $\hat{\theta}_1 = \ldots\ldots$ with 95% c.i. $(\ldots\ldots, \ldots\ldots)$; $\hat{\sigma}_1^2 = \ldots\ldots$ with 95% c.i. $(\ldots\ldots, \ldots\ldots)$

   $\hat{\theta}_2 = \ldots\ldots$ with 95% c.i. $(\ldots\ldots, \ldots\ldots)$; $\hat{\sigma}_2^2 = \ldots\ldots$ with 95% c.i. $(\ldots\ldots, \ldots\ldots)$

   $\hat{\theta}_3 = \ldots\ldots$ with 95% c.i. $(\ldots\ldots, \ldots\ldots)$; $\hat{\sigma}_3^2 = \ldots\ldots$ with 95% c.i. $(\ldots\ldots, \ldots\ldots)$

5. Compare the estimated values $\hat{\sigma}_i^2$ with the theoretical values that you can obtain by observing that the mean of a random variable with inverse-gamma(a, b) distribution is given by $b/(a - 1)$, provided that $a > 1$.

6. Generate a Monte Carlo sample of size 2,000 from the prior joint distribution of (θ, σ^2). Make a single plot showing four scatterplots, one from the joint prior and one from the joint posterior distribution of (θ_i, σ_i^2) of each group (use different colours).

7. Modify the parameters k_0 and ν_0 so to give more weight to the botanist's prior guess (e.g. $k_0 = \nu_0 = 50$) and see how the plot made at point 6 changes.

8. Say you are interested in comparing the variance of sepal length of setosa and virginica iris flowers ($i = 2$ and $i = 3$). By using the generated Monte Carlo samples, estimate $P(\sigma_2^2 > \sigma_3^2)$.

9.* (*optional*) Check how the estimate of $P(\sigma_2^2 > \sigma_3^2)$ changes as the prior sample sizes k_0 and ν_0 change. Consider a grid of values for $\{1, 2, \ldots, 50\}$ for both k_0 and ν_0, and produce a contour plot of $P(\sigma_2^2 > \sigma_3^2)$ as a function of k_0 and ν_0. Which parameter has a stronger impact on $P(\sigma_2^2 > \sigma_3^2)$?