ST3457 Statistical Inference I (Bayesian Statistics)

Lab 3 - 30/11/17

Bernardo Nipoti

Lab 3 - Multivariate normal model

Read the sections *DATA*, *MODEL*, *POSTERIOR DISTRIBUTION* and *PARAMETERS SPECIFICATION*, and address the points listed at the end of this sheet, by working on the R script `LAB_mult_normal.R` that you can download from the course webpage

`http://www.bernardonipoti.com/teaching/st3457_17_18`

DATA. We use again the Iris dataset, available in the `Datasets` package in `R`, which collects different measurements of three species of iris flowers. Today we model jointly the petal and the sepal length of these flowers. We have three species: *versicolor*, *setosa*, *virginica*. We let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n_1}$ be the observations for the first group (versicolor), that is $\boldsymbol{X}_i = (X_{i,1}, X_{i,2})$ is the vector composed by petal length ($X_{i,1}$) and sepal length ($X_{i,2}$), measured in *cm*, of the ith observed flower of species versicolor. Similarly, $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{n_2}$ and $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{n_3}$ are the observations for second and third group (setosa and virginica, respectively).

MODEL. We assume a multivariate Normal sampling model for each group. We choose a semi-conjugate normal/inverse-Wishart prior for mean vector and covariance matrix. The priors for the three groups are assumed independent and identical. That is

$$\begin{aligned}
\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n_1} \mid \boldsymbol{\theta}_1, \Sigma_1 &\overset{\text{iid}}{\sim} \mathrm{N}_2(\boldsymbol{\theta}_1, \Sigma_1) \\
\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{n_2} \mid \boldsymbol{\theta}_2, \Sigma_2 &\overset{\text{iid}}{\sim} \mathrm{N}_2(\boldsymbol{\theta}_2, \Sigma_2) \\
\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{n_3} \mid \boldsymbol{\theta}_3, \Sigma_3 &\overset{\text{iid}}{\sim} \mathrm{N}_2(\boldsymbol{\theta}_3, \Sigma_3) \\
\boldsymbol{\theta}_i &\overset{\text{iid}}{\sim} \mathrm{N}_2(\boldsymbol{\mu}_0, \Lambda_0) \\
\Sigma_i &\overset{\text{iid}}{\sim} \text{inverse-Wishart}(\nu_0, S_0^{-1}).
\end{aligned}$$

POSTERIOR DISTRIBUTION. The posterior distribution of the vectors $(\boldsymbol{\theta}_i, \Sigma_i)$ is not directly tractable, so we resort to a Gibbs sampler algorithm, based on the full conditional distributions for $\boldsymbol{\theta}_i$ and Σ_i. For the first group, we have

$$\begin{aligned}
\boldsymbol{\theta} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \Sigma &\sim \mathrm{N}_p(\boldsymbol{\mu}_n, \Lambda_n) \\
\Sigma \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{\theta} &\sim \text{inverse-Wishart}(\nu_n, S_n^{-1})
\end{aligned}$$

where the posterior parameters are specific to each group. For the first group they are given by

$$\begin{aligned}
\Lambda_n &= (\Lambda_0^{-1} + n_1 \Sigma^{-1})^{-1} \\
\boldsymbol{\mu}_n &= \Lambda_n^{-1}(\Lambda_0^{-1}\boldsymbol{\mu}_0 + n_1 \Sigma^{-1} \bar{\boldsymbol{x}}) \\
\nu_n &= \nu_0 + n_1 \\
S_n &= S_0 + \sum_{j=1}^{n_1} (\boldsymbol{x}_j - \boldsymbol{\theta}_1)(\boldsymbol{x}_j - \boldsymbol{\theta}_1)^{\intercal}.
\end{aligned}$$

PARAMETERS SPECIFICATION. Previous studies on iris flowers indicate that the average petal length is $4cm$ and the average sepal length is $6cm$ with a variance which in both cases is equal to $1cm^2$. Moreover, it is know that petal and sepal length are positively correlated. All this considered we set the parameters of the model as follows.

$$\begin{aligned}\boldsymbol{\mu}_0 &= (4, 6)\\ \nu_0 &= 3\\ S_0 = \Lambda_0 &= \begin{pmatrix}1 & 0.5\\ 0.5 & 1\end{pmatrix}.\end{aligned}$$

By setting $\nu_0 = 3$ (recall that $\nu_0 > 2$ in a bidimensional model), we assign little weight to our prior guess S_0 for Σ.

POINTS TO ADDRESS.

1. Read and understand the `R` script where, for the first group, the following MCMC samples are generated by means of a Gibbs sampler:
   - a sample from the posterior distribution of $(\boldsymbol{\theta}_1,\Sigma_1)$
   - a sample from the posterior predictive distribution of a new versicolor flower $\boldsymbol{X}_{n_1+1}$
2. Generate the same type of MCMC samples for second and third group.
3. Use the MCMC samples to estimate $\boldsymbol{\theta}_i$ (compute the posterior mean) and to compute 95% posterior credible intervals.
4. Based on the MCMC samples for Σ_i, estimate the correlation $\rho_{petal,sepal} = \sigma_{1,2}/\sqrt{\sigma_1^2\sigma_2^2}$ and compute compute 95% posterior credible intervals for such correlations.
5. Using the MCMC samples, estimate the following quantities:
   - $P(|\boldsymbol{\theta}_1| > |\boldsymbol{\theta}_3|)$, where $|(a_1, a_2)| = \sqrt{a_1^2 + a_2^2}$
   - $P(|\boldsymbol{X}_{n_1+1}| > |\boldsymbol{Z}_{n_3+1}|)$
6. Does the initial value for S_0 have a big impact on your estimates? Try different values for S_0.