Applied Bayesian Statistics
Practical 1
Bayes factors. Monte Carlo methods.

Robin Ryder

Aim: Monte Carlo methods. Application to model choice via Bayes factors and model averaging.

Reference: *Bayesian Essentials with R* (Marin & Robert), chapter 2.

In this practical, we shall analyse data on activity of French members of Parliament (*députés*). The data come from `www.nosdeputes.fr` and represent activity between January 2018 and January 2019. Download the data from `http://bit.ly/MASH-BCS`, then read them in R:

```
> deputes=read.csv2('/path/to/deputes2019.csv')
> attach(deputes)
```

Initially, we shall focus on the column `questions_orales`, which represents the number of oral questions Y_i asked by each MP. We would like to test whether it depends on a binary variable Z_i. For Z_i, you can use the gender of the MP (column `sexe`).

1. Explore briefly the data (number of individuals, size of groups 1 and 2, histograms...). Choose a parametric family $(\mathbb{P}_\lambda)$ which seems suitable for these data.

   We shall study the two following models, where π is a prior distribution:

$$
\begin{array}{ccc|ccc}
 & \mathcal{M}_1 & & & \mathcal{M}_2 & \\
 & & & Y_i|Z_i = j & \sim_{i.i.d} & \mathbb{P}_{\lambda_j} \\
Y_i & \sim_{i.i.d} & \mathbb{P}_\lambda & \lambda_1 & \sim & \pi \\
\lambda & \sim & \pi & \lambda_2 & \sim & \pi
\end{array}
$$

2. Find a conjugate prior for the chosen family of distributions. Is this family of priors flexible enough? If not, which prior would you choose?

3. Find Jeffrey's prior for this model. What is the associated posterior?

4. Decide what your prior distribution will be.

5. Plot the prior of the parameters, and the posterior for the parameters of each model. Repeat with different values of the prior hyperparameters.

6. Give a 95% credibility interval for the parameters in each model.

7. In model 2, let $r_\lambda = \frac{\lambda_1}{\lambda_2}$. Give a Monte Carlo estimate of the prior and posterior expectation and variance of r_λ.

   We would now like to compute the Bayes factor

$$B_{21} = \frac{m_2(\boldsymbol{y})}{m_1(\boldsymbol{y})} \quad \text{where} \quad m_k(\boldsymbol{y}) = \int_{\Theta_k} L_k(\theta_k|\boldsymbol{y})\pi_k(\theta_k)d\theta$$

   We propose several Monte Carlo methods to calculate the Bayes factor; we would like to compare the methods. For each method, write a script to visualize the convergence of the method.

8. **Method 1: vanilla Monte Carlo**

   Give an approximation of B_{21} based on an M-sample of parameters simulated from the prior distribution.

9. **Method 2: Importance sampling**

   Compute the posterior mean and variance of the parameters for each model. Deduce a reasonable instrumental distribution g to perform importance sampling. Give an approximation of B_{21} in this case.

   Try again, using a different instrumental distribution. What do you observe?

10. **Method 3: explicit computation**

    Give the explicit expression of B_{21}, and write an R script to evaluate it.

11. Compare the 4 methods and select the best one.

12. Now that you have chosen your method, compute the Bayes factor and conclude on which model is the best.

13. Suppose that our prior probability for each model is 0.5. What is the posterior probability of each model?

14. We wish to predict the number of questions that will be asked by a new female individual. Draw a posterior sample from the corresponding λ parameter in each of the following cases:

    a) We choose model $\mathcal{M}_1$.

    b) We choose model $\mathcal{M}_2$.

    c) We take a mixture of the two models, with weights equal to the posterior probabilities.

    Give the posterior mean and variance. Comment.

15. Perform model choice for other columns: for Y_i, you might look at any of the quantitative variables. For Z_i, you could also use `groupe_sigle`, which gives political affiliation, or `nb_mandats`, which gives the number of other elective offices held. In those cases, Z_i can take more than 2 values.