Université Paris Dauphine
Département MIDO
Master 2 MASH – Méthodes MCMC

2022–2023

Projet R

stoehr@ceremade.dauphine.fr

Instructions.

Due to 20th January, 2023, before 8 p.m.

- Must be sent by email: a report including comments, code and output (RMarkdown, knitr, Anaconda).
- **Each day of delay is penalised by 1 point.**

The report:

- should contains answers to the questions and comments on the study lead. Clear and concise drafting will be appreciated. The report should not exceed 20 pages.
- Graphs should be carefully annotated.
- When explaining algorithms, you can use pseudo-code/Rmarkdown/knitr to refer to your code. But, **do not simply copy-paste the R code in the report.**

The R code:

- should be well commented and fairly optimised to use the specificity of the language.
- Your code should run without errors and allow to reproduce the results presented in the report (give the random seed used). In R, the function `set.seed` allows to specify the random seed to use.

Exercise 1.

Model The haptoglobine has 3 different possible configurations AA, aa, aA. If the population is randomly mating, genes frequencies are following Hardy Weinberg Equilibrium. In this exercise, we assume that the environment favors certain genes associations (*e.g.*, geographically-structured population). We describe this using an inbreeding model parametrised by $\theta \in [0,1]$ and $p \in [0,1]$

$$\mathbb{P}[X = \mathrm{AA}] = p(1-\theta) + (1-p)(1-\theta)^2, \qquad \mathbb{P}[X = \mathrm{aA}] = 2(1-p)\theta(1-\theta),$$
$$\mathbb{P}[X = \mathrm{aa}] = p\theta + (1-p)\theta^2.$$

p is referred to as the inbreeding coefficient and represents the probability that the allele types from parents are identical by descent.

Dataset: x^{obs}

Genotype	AA	aa	aA
Count	302	125	73

EM algorithm.

1. Write a latent variable representation of the model using a latent Bernoulli random variable of parameter p, denoted Z.

2. Give the objective function of the EM algorithm associated to this latent representation and compute the updates for estimating p and θ.

3. Implement the corresponding EM algorithm and run it on the data x^{obs} explaining your setting (*e.g.*, initialisation, stopping criterion). Check graphically the convergence of your algorithm and the fit between the estimated model and the empirical distribution of the data.

For the remainder of the Exercise, we consider the Bayesian paradigm and use independent prior distributions on p and θ. As both parameters are constrained to lie between 0 and 1, we can use either a uniform distribution on $[0,1]$ or a Beta distribution. For the various algorithms below, you will test the sensitivity to the prior distributions used (and to their parameters when using Beta distributions).

Metropolis Hastings .

4. Implement a Metropolis Hastings scheme using a Gaussian distribution as proposal density to sample from the posterior distribution of (p,θ).

5. Test the sensitivity of the scheme to the variance of the Gaussian kernel. Discuss briefly your result.

Gibbs sampler.

6. Using the latent representation from Question 1, compute the following conditionnal distributions: $\pi(z \mid p,\theta,x^{\text{obs}})$ and $\pi(p,\theta \mid z,x^{\text{obs}})$.

7. Implement a Gibbs sampler to sample from the posterior distribution of (z,p,θ).

8. Compare the empirical posterior distribution of (p,θ), the marginal distributions of p and θ as well as their posterior means obtained with Metropolis Hastings scheme and Gibbs sampler. Which algorithm do you recommend to use?

ABC algorithm.

9. Implement a vanilla ABC algorithm (using a L^2-standardised distance) to sample from the posterior distribution of (p,θ). Justify the choice of the summary statistics and study the influence of the size of the neighborhood around x^{obs} on the posterior approximation.

10. Compare the empirical posterior distribution of (p,θ), the marginal distributions of p and θ as well as their posterior means to the ones obtained with the previous MCMC schemes.

11. Hardy Weinberg Equilibrium (model $m = 1$) is embeded in the inbreeding model we used (model $m = 2$). We would like to study if it was reasonnable to use a more complex model for our data. Using an ABC routine, compute estimates of the posterior probabilities of models 1 and 2, *i.e.* $\pi(m \mid x^{\text{obs}})$, under the assumption of a uniform prior on model index m. What is your conclusion?