Département MIDO

MASTER 2 MASH
2016–2017
Bayesian Case Studies
Robin Ryder

Final examination

Documents: you may use a single A4 sheet with notes and no other document. Any attempt to use the Internet, including any form of e-mail or messenging, will result in immediate disqualification. No phones allowed.
Duration: 3 hours.
Students may answer in English or French. All code must be written in the R language.
At the end of the examination, you must hand in your answers written on paper AND send your R code to `ryder@ceremade.dauphine.fr`.
Please contact the examiner if you wish to hand in your answers early. Please make sure that your R code has been correctly received before leaving the room.
Make sure to save your code on a regular basis. Loss of data following computer failure shall not entitle you to extra time.
The sections are independent.

Following standard notation, we say that the real-valued continuous random variables $X \sim \mathcal{E}(\lambda)$ and $Y \sim \Gamma(a, b)$ if X and Y have respective probability density functions

$$f_X(x; \lambda) = \lambda e^{-\lambda x} \mathbb{I} x > 0 \qquad f_Y(y; a, b) = \frac{b^a}{\Gamma(a)} y^{a-1} e^{-by} \mathbb{I}_{y>0}$$

where $\Gamma(a)$ denotes the Gamma function implemented by R's `gamma()`.
We say that the real-valued continuous random variable $Z \sim Pareto(m, \alpha)$ if Z has probability density function

$$f_Z(z; m, \alpha) = \alpha \frac{m^\alpha}{z^{\alpha+1}} \mathbb{I}_{\{z \geq m\}}.$$

In the French public secondary school system, teachers who apply for a new position are ranked using a system of points based on their status, career and personal situation. The data we wish to analyse lists the number of points necessary to obtain a position in a number of secondary schools in the Académie de Versailles in 2012 and for a number of subjects (Mathematics, English, Physics...),

as well as some covariates about those schools. Each line of the data corresponds to a school and a subject, so there may be more than one line for the same school.

The data can be downloaded from `http://bit.ly/MASH-BCS` and read using the command `read.csv('examdata2017.csv')`
and the column of interest in the resulting data frame is labelled `Barre`. In addition to the subject, covariates listed in this data set include the type of school[1], and for the schools where it is relevant, various metrics about the success at the Baccalauréat examination and the size of the school.

1 Introduction

1. Show that if $X \sim Exp(\alpha)$ and $Z = me^X$ then $Z \sim Pareto(m, \alpha)$.
2. Write a function `rpareto(n, m, alpha)` which samples n realizations of the $Pareto(m, \alpha)$ distribution.

2 Case with m known and α unknown

In this section, suppose that $m = 21$ is known but α is unknown.

3. We wish to estimate α in a Bayesian fashion. Show that the $(\Gamma(a, b))$ family of distributions is a conjugate family of priors for α.
4. What is Jeffrey's prior for α? Is it proper? If the prior is improper, is the associated posterior proper?
5. For the prior of your choice, compute the posterior mean and variance of α. Check the impact of the prior.
6. Give (on paper) the analytical value of the marginal likelihood of this model.
7. For this section, we consider only the data in the two categories "MATHS" and "ANGLAIS" (as given by the column `Matiere` in the data). We consider two possible models : (1) the data in the two categories come from a common distribution $Pareto(m, \alpha)$; (2) the data in the two categories come from different distributions $Pareto(m, \alpha_M)$ and $Pareto(m, \alpha_A)$. Compute and interpret the Bayes' factor between these two models.

3 Case with m unknown and α known

We now suppose that m is unknown but $\alpha = \frac{1}{2}$ is known.

8. What is Jeffrey's prior for m? Is it proper? If the prior is improper, is the associated posterior proper?
   From now on, we use this prior for m. If you did not find the prior, you may use the improper uniform prior on $\mathbb{R}_+$, $\pi(m) \propto 1$.
9. Compute the posterior mean and variance of m.

[1] in particular, CLG=collège, LYC=lycée, LP=lycée professionnel

4 Mixture model

In this section, we ignore the categories. We assume that there are unknown parameters $\alpha_1, \alpha_2, m_1, m_2$ and for each observation i there is a latent variable $W_i \in \{1, 2\}$ such that

$$Z_i | W_i = j \sim Pareto(m_j, \alpha_j)$$

We take independent $\Gamma(a, b)$ priors on the (α_j) and the prior from question **??** on the (m_j). We note $p = P[W_i = 1]$ and take a $U([0, 1])$ prior on p.

10. Visualize the data. What seem like reasonable values for m_1 and m_2? Use these values are initial values in you Gibbs algorithm.

11. Give the conditional posterior distribution of α_1, α_2 given the (m_j), the (Z_i) and the (W_i).

12. Give the conditional distribution of W_i given $Z_i, \alpha_1, \alpha_2, m_1, m_2$.

13. Give the conditional distribution of m_1, m_2 given the (α_j), the (Z_i) and the (W_i). *[If you struggle with this distribution, you may take m_1 and m_2 to be fixed as constants set to the values chosen above and proceed with the other questions for partial credit.]*

14. Write a Gibbs' sampler to produce a posterior sample of $\alpha_1, \alpha_2, m_1, m_2$.

15. Explain on paper how you checked that your Gibbs' sampler reached stationarity.

16. Compute the Effective Sample Size of your output.

17. What happens when you choose different initial values for your parameters? Comment.

18. Write an Importance Sampling scheme to estimate the marginal likelihood of this model.

19. We wish to compare this model to the simpler model where $\forall i, Z_i \sim Pareto(m, \alpha)$. Compute and interpret the relevant Bayes' factor.

20. Bonus question: perform any model validation operations that seem relevant. If you deem it necessary, propose and implement a more complex model.

5 Generalized Pareto distribution

In this section, we consider the Generalized Pareto distribution. We say that $T \sim GPD(m, \alpha, \tau)$ if T has density

$$f_T(t; m, \alpha, \tau) = \frac{\alpha}{m\tau} \left(1 + \frac{t - m}{\tau m}\right)^{-\alpha - 1} \mathbb{I}_{t \geq m}.$$

It is easy to verify that if $\tau = 1$ this corresponds to the standard Pareto distribution.
We take the priors $\alpha \sim \Gamma(2, 2)$, $\tau \sim Exp(1)$ and assume as before that $m = 21$.

21. Write a Metropolis-Hastings algorithm or a Metropolis-within-Gibbs algorithm to sample from the posterior distribution of (α, τ).

22. Explain how you checked that your Metropolis-Hastings algorithm has converged.

23. Compute the Effective Sample Size of your output.

24. Give a 95% credible interval for τ.

25. Validate your inference procedure: simulate synthetic data from the $Pareto(21, \alpha)$ distribution for some value of α; estimate the parameters using your Metropolis-Hastings scheme and verify that the credible interval contains the value $\tau = 1$. If you deem it necessary, perform other validation procedures.

26. Write an Importance Sampling scheme to estimate the marginal likelihood of this model.

27. We wish to compare this model to the simpler model where $\tau = 1$. Compute and interpret the relevant Bayes' factor.

28. Bonus question: perform any model validation operations that seem relevant. If you deem it necessary, propose and implement a more complex model.

6 Pareto regression

The questions in this section are much less detailed and should probably be attempted last.
In this section, we consider that $\alpha = \frac{1}{2}$ is known. The data are assumed to follow a distribution

$$Y_i \sim Pareto(m_i, \alpha) \qquad \text{with} \qquad m_i = X_i\beta$$

where X_i is a vector of covariates and β is a vector of regression parameters to be estimated. We use a $\mathcal{N}(0, 1)$ prior of each component of β.

29. Write a Metropolis-Hastings algorithm to sample from the posterior distribution of β.

30. Write a procedure to perform model choice between all the possible subsets of covariates.

31. Using this procedure, give a 95% credible interval for a prediction of Y at a new point of your choosing.

32. Perform any model validation operations that seem relevant. If you deem it necessary, propose and implement a more complex model.