Bayesian case studies
Final examination

Robin Ryder

March 6, 2018

Documents: you may use notes, books and other documents, as well as access the Internet. Any attempt to use any form of e-mail or messenging or to post on a forum 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 mostly independent.

We say that the real-valued continuous random variable $X \sim \mathcal{W}(\theta, k)$ (Weibull distribution) if X has probability density function

$$f_X(x; \theta, , k) = \frac{k}{\theta} x^{k-1} \exp\left(-\frac{x^k}{\theta}\right) \mathbb{I}_{x>0}$$

where $k > 0$ and $\theta > 0$.

The moments of this distribution are given by

$$E[X^n] = \theta^{n/k} \Gamma\left(1 + \frac{n}{k}\right)$$

where Γ is the Gamma function implemented in R by the function `gamma()`.

The data at your disposal were collected by Thu Anh Le based on data from the Free Internet Chess Server. You have information about chess games between a human and a computer. We are interested in modelling the number of moves in a game[1]. For each game, you have:

- player names, date and time
- White.Elo and Black.Elo: a measure of the strength of the players

[1] In chess, two players, called white and black, take turns moving pieces. The game ends when one player wins by checkmate, or a player resigns, or a draw. We only include games that ended with a checkmate.

- WhiteIsComp and BlackIsComp: Bernoulli variables indicating whether white or black is played by the computer
- TimeControl, White.Clock, Black.Clock: indicate how much time the players have to think about their moves
- ECO: a nomenclature of the opening (the first few moves)
- PlyCount: number of moves[2]
- Result

We are mostly interested in the columns PlyCount and WhiteIsComp/BlackIsComp.

The data can be downloaded from `http://bit.ly/MASH-BCS` and read using the command
`read.csv('examdata2018.csv')`
and the column modelled by the random variable X in the resulting data frame is labelled `PlyCount`.

1 Introduction

1. Verify that if $Y \sim \mathcal{E}(1)$ (standard exponential distribution) and $X = (\theta Y)^{1/k}$ then $X \sim \mathcal{W}(\theta, k)$.
2. Use this property to write a function `my.rweibull(n, θ, k)` which generates n realizations of the Weibull distribution.
   If you do not succeed, you may use the function `rweibull()` which exists in R with a different parametrization: shape $a = k$ and scale $b = \theta^{1/k}$.
3. Using this function and the values of your choice for θ and k, perform a Monte Carlo experiment to estimate $E[X]$. Verify that you recover the theoretical value.

2 Case with k known

In this section, we suppose that we know the value of k.

4. Give Jeffreys' prior for θ. Is this prior proper? If not, is the associated posterior proper?
5. Verify that the Inverse Gamma is a conjugate prior for θ and give the associated posterior. Recall that the $\mathcal{IG}(\alpha, \beta)$ distribution has mean $\frac{\beta}{\alpha-1}$ and probability density function
$$\frac{\beta}{\Gamma(\alpha)} t^{-\alpha-1} \exp\left(-\frac{\beta}{t}\right) \mathbb{I}_{\{t>0\}}.$$
6. For the prior of your choice and $k = 1.85$, compute the prior and posterior mean. Check the impact of the prior.
7. We now wish to choose between two models: (1) all the data come from the same $\mathcal{W}(\theta, 1.85)$ distribution and (2) the data come from $\mathcal{W}(\theta_1, 1.85)$ or $\mathcal{W}(\theta_2, 1.85)$ depending on the value of WhiteIsComp. Propose a way to get a Bayes factor to choose between these two models, and interpret it.

[2] The column PlyCount actually corresponds to what chess players call *half-moves*: if white plays 40 times and black plays 40 times, we record the length of the game as 80.

3 Case with θ known

In this section, suppose that we know the value $\theta = 3317$. Take the improper prior $\pi(k) \propto \frac{1}{k}$.

8. Write a function that gives the posterior distribution of k.
9. Write a Metropolis-Hastings algorithm to simulate from the posterior distribution of k.
10. Explain how you checked that your Metropolis-Hastings algorithm has converged.
11. Compute the Effective Sample Size of your output.
12. Give a 95% credible interval for k.
13. Validate your inference procedure: simulate data from the $\mathcal{W}(\theta, k)$ for your choice of the values of the parameters, and verify that the 95% credible interval contains the true value of k.
14. We now wish to choose between two models: (1) all the data come from the same $\mathcal{W}(3317, k)$ distribution and (2) the data come from $\mathcal{W}(3317, k_1)$ or $\mathcal{W}(3317, k_2)$ depending on the value of WhiteIsComp. Propose a way to get a Bayes factor to choose between these two models, and interpret it.

4 Joint estimation

In this section, we wish to estimate θ and k jointly. We take the priors $\theta \sim \mathcal{IG}(1, 1)$ and $k \sim \mathcal{E}(1)$.

15. Write a Metropolis-Hastings or a Metropolis-within-Gibbs algorithm to draw from the posterior distribution.
    If you use Metropolis-Hastings, use the previous sections to decide what a reasonable proposal kernel standard deviation is for each parameter. If you use Metropolis-within-Gibbs, describe on paper what the conditional distribution is.
16. Explain how you checked that your Metropolis-Hastings algorithm has converged.
17. Compute the Effective Sample Size of your output.
18. Write code to compute the marginal likelihood of the model.
19. We now wish to choose between two models: (1) all the data come from the same $\mathcal{W}(\theta, k)$ distribution and (2) the data come from $\mathcal{W}(\theta_1, k_1)$ or $\mathcal{W}(\theta_2, k_2)$ depending on the value of WhiteIsComp. Propose a way to get a Bayes factor to choose between these two models, and interpret it.
20. Think about model misspecifications. How can you handle them?

5 Weibull regression

Choose a covariate Y, for example the value of White.Elo. We now write

$$X_i \sim \mathcal{W}(\theta + \beta Y_i, k)$$

21. Write a function to compute the likelihood of (θ, β, k).
22. Propose a prior for β. Write a function to compute the corresponding posterior.
23. Write an algorithm to simulate samples from the posterior of (θ, β, k). If necessary and for partial credit, you may fix θ and/or k to a constant value.
24. Propose and implement a methodology to test whether $\beta \neq 0$.
25. Generalize this methodology to the case with more than one covariate.

6 Censored data

We now consider the dataset `data2018bis.csv`, which also includes games which did not end by a checkmate. We shall model the games that ended by a resignation. Our model is:

- $X_i \sim \mathcal{W}(\theta, k)$ is, as before, the number of moves the game would have had if it had gone all the way to checkmate.
- $Z_i \sim \mathcal{TE}(\lambda, X_i)$ is the number of moves before the end of the game that resignation occured, where $\mathcal{TE}(\lambda, X_i)$ is the the truncated exponential distribution, with parameter λ and truncated to values less than X_i.
- We observe $Y_i = X_i - Z_i$.

26. Propose a prior for λ.
27. Write a function that simulates realizations of the truncated exponential distribution.
28. Give the conditional distributions of (θ, k) knowing (Y_i, Z_i) (for all i), of Z_i knowing the Y_i, λ, θ and k, and of λ knowing the Z_i. State any modelling assumptions you make.
29. Write a Gibbs' sampler to simulate from the joint posterior.
30. Give a 95% confidence interval for λ.
31. Think about model misspecifications. How can you handle them?