Bayesian case studies
Final examination

Robin Ryder

February 28, 2019

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.
Sections 1 to 3 are independent. Section 4 is an extension of section 3, and sections 5, 6 and 7 assume that you have completed at least one of the previous sections.

In September 2017, hurricane Maria hit Puerto Rico, devastating the island. The White House stated that 64 people died because of the hurricane, but this number has been greatly contested. The aim of this exam is to estimate the mortality due to hurricane Maria[1].

Official records give information on all deaths on each day, but cause of death is rarely registered. We shall therefore compare the number of deaths before and after the hurricane, to estimate the number of deaths attributable to the hurricane.

1 Bayesian inference for the binomial distribution

Let Y_t be the number of deaths on day t. We assume that the (Y_t) are iid of distribution $Bin(K, p)$ with p unknown and $K = 3\,350\,000$ known. We wish to infer on p, the probability of death on a given day. We assume that the population size K is constant through time.

1. What is Jeffreys' prior for p?

2. Give a family of distributions that is conjugate for this model. Within this family, propose distribution parameters which represent your prior beliefs about p.
   In the following, I advise you to use one of the following priors: (a) Jeffreys' prior, (b) a conjugate prior, or (c) a $\mathcal{U}([0, 1])$ prior.

[1] This exam is inspired by Rivera & Rolke (2018). Data from the Puerto Rico Department of Health.

3. What is the analytical posterior distribution of p?

4. Compute the posterior mean and variance of p.

5. Give the marginal likelihood of the data under this model.

We focus on D_1 the data for the summer of 2015 and D_2 the data from the autumn of 2015. We wish to test whether the probability of death is the same during the two periods. Let p_1, resp. p_2 be the daily probability of death during the summer, resp. the autumn, of 2015.

6. Give the posterior mean and variance of $p_1|D_1$ and $p_2|D_2$.

7. Compute a Bayes factor to choose between the models (1) $p_1 = p_2$ and (2) $p_1 \neq p_2$. Interpret your result. What do you think of the assertion "In normal times, the death rate is the same during the summer and the autumn"?

8. Repeat for D_3 the data for the summer of 2017 and D_4 the data for the autumn of 2017 (just before and after hurricane Maria). Interpret your result. What do you think of the assertion "the death rate increased due to hurricane Maria"?

2 Number of deaths due to the hurricane

We now propose an alternative modelization:

- before the hurricane, $Y_t = Y_t^1$ with $Y_t^1 \sim Bin(K, p)$;
- after the hurricane, $Y_t = Y_t^1 + Y_t^2$ with $Y_t^1 \sim Bin(K, p)$ and with $Y_t^2 \sim Bin(K, p')$.

In this modelization, Y_t^2 represents the number of people who died because of the hurricane, and Y_t^1 the number of people who died because of other causes. These two variables are latent.

For p and p', you can use the same priors as in section 1. Use the data D_3 and D_4 defined in question (8) as samples from before and after the hurricane.

9. What is the joint posterior distribution of the parameters?

10. What is the conditional distribution of p and p' given the (Y_t), the (Y_t^1) and the (Y_t^2)?

11. What is the conditional distribution of (Y_t^1, Y_t^2) given Y_t, p and p'?

12. Write a Gibbs' sampler to sample from the joint posterior.

13. Explain how you have verified that your algorithm has converged.

14. Compute the Effective Sample Size of your output.

15. Give the posterior mean and a 95% credible interval for p and p'.

16. Validate your inference procedure: simulate data from the binomial distributions of your choice, and verify that the 95% credible region covers the true parameter value.

17. Let $Z = \sum_t Y_t^2$ be the total number of deaths due to the hurricane. Give the posterior mean and variance of Z. Estimate $P[Z \leq 64]$. What do you think of the White House's claim that $Z = 64$?

18. Why would the model $Y_t \sim Bin(K, p + p')$ after the hurricane be a good approximation of the model we have described?

3 Hierarchical model

We no longer wish to assume that p is the same throughout the year. Instead, we assume there are 12 parameters $(p_1, \ldots, p_{12})$ for the 12 months of the year. We propose a hierarchical model, wherein $Y_t \sim Bin(K, p_m)$ if t is in month m, and the p_m are iid from a $Beta(a, b)$ distribution. Take the prior of your choice for a and b.

19. What is the joint posterior distribution of the parameters?

20. Write an MCMC scheme to sample from the posterior – it can be a Gibbs' sampler, a Metropolis-Hastings algorithm, Metropolis-within-Gibbs...

21. Explain how you have verified that your algorithm has converged.

22. Compute the Effective Sample Size of your output.

23. Give the posterior mean of $\frac{a}{a+b}$ and a 95% credible interval.

4 Hierarchical model with hurricane effect

We extend the model from the previous section: before the hurricane, we model $Y_t \sim Bin(K, p_m)$. After the hurricane, we model $Y_t \sim Bin(K, p_m + p_h)$ where p_h is a constant representing the deaths due to the hurricane.

24. What is the joint posterior distribution of the parameters?

25. Write an MCMC scheme to sample from the posterior – it can be a Gibbs' sampler, a Metropolis-Hastings algorithm, Metropolis-within-Gibbs...

26. Explain how you have verified that your algorithm has converged.

27. Compute the Effective Sample Size of your output.

28. Give the posterior mean of p_h and a 95% credible interval.

5 Model choice

29. For each of the models considered previously, write a function to estimate the marginal likelihood.
    Credit will be given even if only some of the previous models are considered.

30. Which model is the best fit to the data?

6 Dampening effect

31. The effect of the hurricane on the mortality can be expected to disappear with time. We now assume that the death probability due to the hurricane on day t can be written as $p' e^{-\beta t}$. For the model of your choice from one of the previous sections, adapt your code to sample from the posterior of this model.

7 Posterior predictive distribution

Recall that the Beta function is defined as

$$B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$$

where Γ is the Gamma function. The Beta function and its logarithm are available in R through the functions `beta()` and `lbeta()`.

Let $n \in \mathbb{N}$, $\alpha > 0$ and $\beta > 0$. We say that $X \sim BB(n, \alpha, \beta)$ (Beta-Binomial distribution) if

$$P[X = j] = \binom{n}{j} \frac{B(j+\alpha, n-j+\beta)}{B(\alpha, \beta)} \quad \text{for } j \in \mathbb{N}.$$

32. Prove or accept that the posterior predictive distribution for the model of section 1 is a Beta-Binomial distribution. Give the values of the parameters of this distribution.

33. Write a function to sample realizations of the Beta-Binomial distribution.

34. Using point estimates given by your answers to one of the previous sections, simulate posterior predictive values for the number of daily deaths. Compare the distribution of your sample to the distribution of the observations.