Département MIDO

MASTER 2 MASH
2015–2016
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.

We wish to study the number of fatalities in traffic accidents in the French Orne département.
Our model for the number of fatalities in year t is Poisson:

$$Y_t \sim \mathcal{P}(\alpha + \beta t) \tag{1}$$

with α and β unknown. We observe realizations (y_t) for $1979 \leq t \leq 2012$.
(To be precise: Y_t represents the number of people who have died of their injuries within the 30 days that follow a traffic accident.)
The data can be downloaded from `http://bit.ly/MASH-BCS` and contain 3 columns, corresponding to t, Y_t and Z_t respectively. The last column is only useful for Section 4.
(Source of the data: Inserm CépiDC and Écosanté.)

1. Calculate the likelihood function of (α, β).

1 Prior specification

2. In this question, we consider the special case $\beta = 0$ so that $\forall t, Y \sim \mathcal{P}(\alpha)$. Give Jeffrey's prior for α.

3. In this question, we consider the special case $\alpha = 0$ so that $\forall t, Y \sim \mathcal{P}(\beta t)$. Give Jeffrey's prior for β.

4. Are the priors defined in the two previous questions proper?

From now on, we consider prior distributions of the form $\pi(\alpha, \beta) = \pi(\alpha)\pi(\beta)$ with $\pi(\alpha) \sim \mathcal{N}(0, \sigma_\alpha^2)$ and $\pi(\beta) \sim \mathcal{N}(0, \sigma_\beta^2)$. You may use $\sigma_\alpha = 10^4$ and $\sigma_\beta = 1$.

2 Simulating from the posterior

The maximum likelihood estimator can be obtained with the command
`glm(y~t, family=poisson(link=identity))$coefficients`
and this value is a useful starting point for your MCMC algorithms.

5. Write a Metropolis-Hastings algorithm to simulate from the posterior distribution of (α, β).

6. Try various values for the variance of the proposal distribution in the Metropolis-Hastings and choose the best.

7. Explain on paper how you checked that the Metropolis-Hastings chain has reached stationarity.

8. Compute the Effective Sample Size of your output.

3 Cut-off point

In 1992, France introduced a penalty point system (*permis à points*). We wish to test whether this had an impact on the number of traffic fatalities. We consider an alternative model with a cut-off point:

$$\begin{aligned} Y_t &\sim \mathcal{P}(\alpha_1 + \beta_1 t) \text{ if } t \leq T_c \\ Y_t &\sim \mathcal{P}(\alpha_2 + \beta_2 t) \text{ if } t > T_c \end{aligned} \tag{2}$$

for $T_c = 1992$.

9. Write a Metropolis-Hastings algorithm to sample from the posterior distribution of $(\alpha_1, \alpha_2, \beta_1, \beta_2)$.

10. Estimate the posterior probability $\mathbb{P}[\beta_2 < \beta_1 | \boldsymbol{Y}]$.

11. Check the impact of the prior on the answer to question 10.

12. Propose a way to test whether the penalty point system had an impact on the number of road fatalities.

13. What are the posterior mean and variance for the parameters of the model?

14. Compute the marginal likelihood of models (1) and (2).
You may wish to use importance sampling.

15. Compute the Bayes' factor to choose between model (1) and model (2). Interpret the answer.

4 Correcting the data

In this section, we do not use the second column of the data (Y_t), as it is only an approximation. We only use the columns for t and Z_t.

The data used until now was actually artificially corrected, due to the following bias: since 2005, the number of fatalities reported is the number of people who die within 30 days of a traffic accident. Up until 2004, the number of fatalities reported is the number of people who die within 6 days of a traffic accident (hence a smaller number). We model this phenomenon as follows.

As previously, let Y_t be the number of people (in year t) who die of their injuries in the 30 days following a traffic accident. As previously, $Y_t \sim \mathcal{P}(\alpha+\beta t)$. We do not observe Y_t; rather, we observe Z_t where

$$\begin{aligned} Z_t &\sim Bin(Y_t, \rho) \quad \text{if } t \leq 2004 \\ Z_t &= Y_t \quad \text{if } t \geq 2005. \end{aligned} \tag{3}$$

The parameter ρ is unknown and represents the probability that a person dies within 6 days of an accident, given that that person died within 30 days of an accident. We put a $U([0,1])$ prior on ρ. We wish to build a Gibbs' sampler for $(\alpha, \beta, \boldsymbol{Y}, \rho)$.

16. If you wish, you may exclude some of the data. If you do, justify your choice.
17. Give the conditional distribution of ρ given $\alpha, \beta, \boldsymbol{Y}, \boldsymbol{Z}$.
18. Give the conditional distribution of Y_t given α, β, ρ, Z_t.
    Hint: Show that conditionally on α, β, ρ, the variable $Y_t - Z_t$ is Poisson distributed.
19. Write the form of the conditional density of α, β given $\rho, \boldsymbol{Y}, \boldsymbol{Z}$.
20. Implement a Gibbs' sampler to simulate from the joint posterior of all the parameters. If necessary, you may wish to use Metropolis-within-Gibbs or any other method.
21. The *Observatoire national interministériel de la sécurité routière* uses the value $\rho = 0.936$. Does your analysis support this choice?