Tutorial Lab 3re – The Reexam – Bayesian Methods ht2019
Michael Höhle, Chun-Biu Li Date: 01 Mar 2020

This exercise sheet constitutes the re-exam homework, which needs to be handed in separately by each student signed-up for the course in order to obtain the ECTS credits for the course. Your solution shall consist of the following:

- A written report as a PDF file containing solutions in the form of results, textual interpretations and graphs for the four homework exercises. Note: plagiarism or other forms of cheating is a serious act – to underline this your report must as cover page contain the signed confirmation that your work is made in accordance with the Rules for Written Exams at Stockholm University. For further information about possible consequences see also the Regulations for Disciplinary Matters at Stockholm University.
- A file `<lastname>.R` containing the R code used to obtain all results and graphics contained in the report. Structured and well-documented code is important, e.g., each function should be preceded by a short text explaining what the input and output parameters are. Further code comments are to be made where needed and indentation should be used – see, e.g., Google's R Style Guide for further guidelines. Results are not to be discussed in the code – this is done in the report. As a trivial quality check: the command `source("<lastname>.R")` should run without errors for your code file.
- Deadline: **Sun 22 Mar 2020 at 18:00 o'clock**. The report has to be handed in as a bundle consisting of a) A scanned copy of your signed `Confirmation.pdf`, b) a PDF file `<lastname>.pdf` containing your report, c) the R file `<lastname.R>` and d) (in case of Sweave/knitr) `lastname.R[nw|md]` before the deadline. If you modified the original data or if your R code relies on external files, your bundle should contain these files as well (optimally as a ZIP file). All files are to be uploaded before the deadline to the Moodle drop-box on the course home page. Please note that there is a 10Mb file limit when uploading files. Delayed hand-ins are not accepted.

A total of 100 points can be reached for the answers in the report. Furthermore, up to 5 additional bonus points can be obtained, should your report and code be written with knitr. In this case please also attach the file `<lastname>.R[nw|md]` to your upload. Your final grade is determined by your sum of regular points and bonus points. Note: A penalty is imposed on reports longer than 30 pages.

Lycka till!

Exercise 1 (25 points)

This exercise is about comparing the results of Bayesian inference for different sampling schemes and priors.

(a) (4 points) Suppose that we want to infer the probability of defection θ ($0 \leq \theta \leq 1$) for a product manufactured from a factory. From a random sample of 30 products, it is found that 3 of them are defected. Assume that the probability for a product to be defected is independent of the others. Use the uniform prior, find the posterior of θ.

(b) (6 points) Instead of fixing the number of samples as above, another sampling scheme is as follows. We keep on sampling the products randomly until 3 defected products are seen. It just happens that the 30th product sampled is the third defected one we found. Again using the uniform prior, find the posterior of θ and compare the results of Parts (a) and (b). *Hint*: The variable corresponds to the "data" in the likelihood function $p(data|\theta)$ is different from that of Part (a).

(c) (4 points) Repeat Part (a) but using Jeffrey's prior, find the prior and posterior as a function of θ.

(d) (6 points) Repeat Part (b) but using Jeffrey's prior, find the prior and posterior as a function of θ.

(e) (5 points) Plot all four posteriors from Part (a)-(d) together and discuss how the design of the sampling scheme and the choice of prior can affect the results.

Exercise 2 (25 points)

To complete this exercise, you need to first download and read the article by Ensign D. I. and Pande V. S., "Bayesian detection of intensity changes in single molecule and molecule dynamics trajectories", J. Phys. Chem. B, 114:280 (2010), available in http://pubs.acs.org/doi/abs/10.1021/jp906786b. Note that knowledge in molecular science is not required in completing this exercise. *IMPORTANT*: The answers of the questions must be written in your own words. Moreover, a clear, concise and logical writing is required to obtain full points of the questions.

(a) (6 points) Clearly explain how the approximation, i.e., the "≈" sign, in Eq. 12 in the article is obtained. In particular, you should tell what the perturbation parameter is and what the order of magnitude of the leading correction term is.

(b) (4 points) Suppose you will give a short presentation about Section 2.5 of the article - Comparing Trajectory Segments, and you only want to demonstrate the general idea instead of showing the technical details. Draw a schematic figure (hand-drawing is ok) to illustrate the workflow how to cluster the change point segments into different states. *Hint*: Imagine what it would look like in a powerpoint slide and be creative!

(c) (4 points) Point out one possible problem when applying the algorithm in Section 2.5 to compare trajectory segments and determine the number of clusters. Please justify your claims.

(d) (5 points) Derive Eq. 40 (i.e., the Bayes factor for the Binomial processes). A step-by-step derivation should be given.

(e) (6 points) Point out two possible problems or criticisms of the proposed change point detection method and discuss how they can be improved/resolved. Justify your claims.

Exercise 3 (30 points)

The file `rain.txt` consist of annual maximum daily rainfall values recorded at the Maiquetia international airport in Venezuela for the years 1961-1998. In December 1999 a daily precipitation of 410 mm caused devastation and an estimated 30.000 deaths.

We will follow a standard approach of modelling annual maxima as independent observations following a Gumbel distribution with cumulative distribution function

$$F(x|\mu,\sigma) = \exp\Big(-\exp\Big(-\frac{x-\mu}{\sigma}\Big)\Big), \quad x \in \mathbb{R}, \mu \in \mathbb{R}, \sigma > 0,$$

hence disregarding from the fact that precipitation is positive.

(a) (4 points) Write a function with header

```
gumbloglik<-function(theta, rain)
```

that computes the log-likelihood function $l(\boldsymbol{\theta} = (\mu,\sigma)'|\boldsymbol{x})$ given data `rain`. Construct a contour plot of the log-likelihood over a suitable region.

(b) (7 points) If we assign an improper flat prior $\pi(\mu,\sigma) \propto 1$ on the parameters, the posterior distribution will be proportional to the likelihood. Write an R function with header `rpost.mh <- function(n,theta0,rain)` which implements a a Metropolis algorithm that samples n values from the joint posterior distribution of μ and σ when using a bivariate Gaussian proposal kernel with covariance matrix $\boldsymbol{\Sigma} = \text{diag}(\tau_\mu^2, \tau_\sigma^2)$. What values of τ_μ and τ_σ do attain an acceptance rate, which ensures a suitable mixing of the chain(s)?

(c) (5 points) Use the sampler from (b) to approximate the posterior mean of the parameters. State also approximate Monte-Carlo standard errors. Do trace-plots, autocorrelation plots and argue whether removal of burn-in is necessary.

(d) (2 points) Make a new contour plot of the log-likelihood, where you add the generated samples as points to the contour plot. *Hint:* Using the argument `cex=0.1` can be helpful in order for points not to become too big.

(e) (5 points) Let X^* denote a future annual maxima from the same distribution. Show that

$$\text{E}(1 - F(410|\mu,\sigma)|\boldsymbol{x}) = P(X^* > 410|\boldsymbol{x}).$$

and use this, together with the above sampler, to approximate the posterior probability of X^* being greater than $410mm$.

(f) (7 points) Use the so called *zero-trick*[1] to write a JAGS model which implements the above sampling situation with n observations from the Gumbel distribution. Since JAGS can not handle improper priors use instead the informative priors $\mu \sim U(0, 100)$ and $\sigma \sim U(10, 100)$. Run the JAGS model for a sufficient number of samples and report posterior median and 95% equal-tailed credibility intervals for the two parameters.

Exercise 4 (20 points)

In this exercise you have to read the publication by Verrall (1990) (available from the course home page), which is about the Bayesian modeling of outstanding claim reserves. You then have to answer a number of questions related to this work.

(a) (2 points) State concisely in 5-6 sentences what the aim of the paper is.

(b) (5 points) Write a 1 page summary motivating and explaining the available data and the mathematical model used in the paper. *Note:* Do not mention any inferential aspects at this point.

(c) (3 points) The file `verell1990.txt` contains the data triangle given on p. 229 of the paper. Read in the data and write R code, which gives you estimates as in Sect. 4.1, i.e. corresponding to the 'no prior' situation. Make a table similar to Table 1 of the paper containing the output of your estimation. Also state your of estimate $\hat{\sigma}^2$. Interpret your results. `Hint:` The R function `lm` might be useful.

(d) (6 points) Write a JAGS model to conduct a Bayesian analysis similar to Sect 4.2 of the paper. As a small extension: you are supposed to use a $\mathcal{G}a(0.001, 0.001)$ prior for $1/\sigma^2$. In your analysis you can set $L = 10^6$. Run the JAGS model for an appropriate number of samples (i.e. perform a convergence assement), remove possibly burn-in and use the output to generate a table similar to Table 3 in the paper. Furthermore, use your output to state numbers corresponding to the 1st column of Table 4, i.e. the number of outstanding claims per year. Finally, state the posterior mean and a 95% highest posterior density interval for the total number of outstanding claims. *Note*: Your results will be slightly different from the numbers in the paper.

(e) (4 points) Describe on approximately 1 page the results of the paper and discuss why a Bayesian inference approach was particularly advantageous for the problem at hand.

References

Verrall, R. (1990). Bayes and empirical Bayes estimation for the chain ladder model. Astin Bulletin, 20(2):217–243.

[1] See https://www.medicine.mcgill.ca/epidemiology/Joseph/courses/common/Tricks.html and note that in JAGS the `zeros` vector has to be provided through the data argument (i.e. instead of setting `zeros[i] <- 0` in the JAGS model).