Bayesian Methods
MT7003 HT21

Take Home Re-Exam

Chun-Biu Li
Felix Gnther

This take home exam consists of 4 exercises. The report of the exam needs to be handed in separately by each student signed-up in Ladok 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 4 exercises. The report must be completed independently. Plagiarism or other forms of cheating is a serious act. To underline this, the *signed confirmation sheet* must be submitted to declare that your work is made in accordance with the rules for written exams at Stockholm University (see course page).

- 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: **Tuesday May 3 2022 at 6:00pm**. 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 (e.g., STAN models in text 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 28 pages.

Lycka till!

Exercise 1 (25 points total)

For parts 1-3 below, let y be the number of heads in n tosses of a coin, whose probability of head is θ.

1. (3 points) If the prior distribution for θ is uniform on the range $[0, 1]$, find the prior predictive distribution for y, $P(y = k) = \int_0^1 P(y = k|\theta)d\theta$.for $k = 0, 1, \cdots, n$.
2. (4 points) Suppose one assigns a $Beta(\alpha, \beta)$ prior distribution for θ, and observes y heads out of n tosses. Show that the posterior mean of θ lies between the prior mean, $\frac{\alpha}{\alpha+\beta}$, and the observed relative frequency of heads, y/n.
3. (4 points) Show that if the prior distribution of θ is uniform, the posterior variance of θ is less than the prior variance.
4. (6 points) From the paper by Tierney and Kadane 1986 on Laplace approximations, derive Eq. (A.1) by showing that the leading correction term has the form a/n with the coefficient a as shown in the paper, where n is the data size. Note: you do not need to work out the higher order terms, i.e., b/n^2 and $O(n^{-3})$.
5. (8 points) Suppose a measurement y is sampled from the normal distribution $N(\theta, \sigma^2)$ with known σ and unknown θ lying in the interval $[0, 1]$. Consider two point estimates of θ: A) the maximum likelihood estimate, restricted to the range $[0, 1]$, and B) the posterior mean based on the assumption of a uniform prior in θ. Show that if σ is large enough, the estimate A) has a higher mean squared error than the estimate B) for any value of θ in $[0, 1]$.

Exercise 2 (25 points total)

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.

1. (6 points) Clearly explain how the approximation, i.e., the "$\approx$" 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.
2. (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!

3. (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.

4. (5 points) Derive Eq. 40 (i.e., the Bayes factor for the Binomial processes). A step-by-step derivation should be given.

5. (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 total)

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.

1. (2 points) State concisely in 5-6 sentences what the aim of the paper is.

2. (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.

3. (5 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.

4. (8 points) Write a STAN model to conduct a Bayesian analysis similar to Sect 4.2 of the paper. As a small extension: you are supposed to use a $Ga(0.001, 0.001)$ prior for $1/\sigma^2$. In your analysis you can set $L = 10^6$. Run the STAN model for an appropriate number of samples and perform a convergence assessment by examining relevant diagnostic statistics and graphical summaries of the MCMC draws. 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% credible interval for the total number of outstanding claims. *Note*: Your results will be slightly different from the numbers in the paper.

5. (5 points) Describe on approximately 1 page the results of the paper and discuss the advantages of using Bayesian inference for the problem at hand.

6. (5 points) A model similar to the chain-ladder model for outstanding claims can be used in infectious disease surveillance for the task of *Nowcasting* new infection counts. Suppose you are interested in examining current trends in infection spread based on the daily case count of new infections for all days up to the current day of analysis. Unfortunately, infection counts for a given day are reported by a health authority with a delay that can vary between individuals, e.g. between one day and up to to three weeks. Therefore not all infections for the most recent days are yet reported. In this case, the triangular data matrix (*reporting triangle*) contains in row i and column j the number of people with an infection on day i reported with a delay of j days. A reasonable assumption is that the total number of new infections on consecutive days is similar. Describe how such an assumption can be incorporated as an extension of the Bayesian chain-ladder/Nowcasting model and provide corresponding formulas.

Exercise 4 (20 points total)

In this exercise we analywe the *copresence* data set using Bayesian generalized linear models. The data set is available from the course home page. In the data Y is a binary variable indicating co-presence of two species in a particular forest at $n = 603$ locations. There is one predictor variable, X, that corresponds to log of the distance of each location to the forest edge.

1. (5 points) Fit a logistic regression model with intercept and slope parameter α and β to the data using vague priors, e.g. $(\alpha, \beta) \sim N(0, \text{diag}(10))$. Plot the posterior distribution of β and discuss whether the proximity to the forest edge is a significant predictor of species co-presence.

2. (5 points) Fit an alternative model using the *complementary log-log* link to the data. That means use the following assumption in your model: $\alpha + \beta * X_i = log(-log(1 - \pi_i))$, with $\pi_i = P(Y = 1|X = x_i)$. Compare the results of both models by plotting the posterior probability of $P(Y = 1|X)$ over the range of observed values of X in the data set. Do the results of both models differ considerably? **Note:** Use a prior distribution of $(\alpha, \beta) \sim N(0, \text{diag}(5))$ for this model.

3. (5 points) Describe a method of your choice to decide which of the two models is more appropriate for analysing the data. Also describe the problems/challenges of your chosen approach (about half a page, including formulas if relevant).

4. (5 points) Focus again on the logistic regression model and implement a Metropolis sampler in R to sample from the posterior distribution using the same prior as in part a). For the proposal distribution you can use, e.g., a multivariate normal distribution,

$$(\alpha_t, \beta_t)|(\alpha_{t-1}, \beta_{t-1}) \sim N((\alpha_{t-1}, \beta_{t-1})^t, \Sigma_p), \;\; \Sigma_t = \begin{bmatrix} \sigma_\alpha^2 & 0 \\ 0 & \sigma_\beta^2 \end{bmatrix},$$

with $\sigma_\alpha^2 = 0.3$ and $\sigma_\beta^2 = 0.2$. Check trace-plots of your samples, discuss and think about burn-in samples, and compare your posterior point estimates with the results from the STAN implementation to check that they are similar.