Tutorial Lab 3 – Exam – Bayesian Methods ht2017
Michael Höhle, Chun-Biu Li Date: 11 Oct 2017

This exercise sheet constitutes the 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: **Wednesday 01 Nov 2017 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 95 points can be reached for the answers in the report. Up to 5 points are given for the quality and documentation level of your R code. 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>.Rnw` 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 32 pages.

Lycka till!

Exercise 1 (22 points)

This exercise is about comparing the results of Bayesian inference for different sampling schemes and priors. *Hint*: All posterior in this Exercise can be expressed in terms of the Beta distribution $Beta(\alpha, \beta)$.

(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) (5 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) (3 points) Repeat Part (a) but using Jeffrey's prior, find the prior and posterior as a function of θ.

(d) (5 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 (23 points)

This exercise gives you an opportunity to see how Bayesian hypothesis test can be applied to a more realistic problem: detecting temporal change points in a time series. In particular, you will learn about the use of Bayes factors, how to set up the hypotheses, how to choose priors, how to perform marginalization, etc. To complete this exercise, you have to 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`. Although the article deals with a special type of biophysical data, namely, the single molecule time series, pre-knowledge of molecular science is not needed. In particular, focus should be put on how the Bayesian framework is formulated when reading the article. You can also skip the Appendices which deals with the Gaussian and binomial processes. *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 why the improper priors are not suitable when evaluating the Bayes factor. *Hint*: See the left column in page 282 of the article.

(b) (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.

(c) (6 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 and scan the result 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!

(d) (5 points) Point out two possible problems or criticisms of the proposed change point detection method and discuss how they can be improved/resolved. Please justify your claims.

Exercise 3 (25 points)

This exercise is about empirical Bayes inference in the Poisson-Gamma model, i.e. consider the following modelling hierarchy

$$\lambda \sim \mathrm{Ga}(a, b)$$
$$y|\lambda \sim \mathrm{Po}(e \cdot \lambda),$$

where $a > 0, b > 0$ are the fixed hyper-parameters and $e > 0$ is a known value.

(a) (4 points) State the posterior distribution of $\lambda|y$. *Hint*: It is a well known distribution, you only need to derive the parameters of this distribution.

(b) (3 Points) State an expression for the posterior mean $E(\lambda|y)$ and show that the posterior mean can be written as a weighted average of prior mean of λ and the maximum likelihood estimator for λ.

(c) (4 points) Show that the marginal distribution for y is negative binomial with size parameter a and probability of success $b/(b + e)$.

(d) (3 points) Implement the marginal likelihood of $\boldsymbol{y}$ as a function of $\boldsymbol{\eta} = (a, b)'$ when the data are $\boldsymbol{y} = (y_1, \ldots, y_n)$' and $\boldsymbol{e} = (e_1, \ldots, e_n)'$. Do this by writing a R function `mloglik(eta, y, e)`, which for data with the vectors `y` and `e` implements the marginal log likelihood.

(e) (3 points) Write a R function `psummary(y, e, a, b, alpha=0.05)` which, given the prior parameter vector $(a, b)'$, for each data pair $(y_i, e_i)'$ computes the posterior mean of λ, an equitailed $(1 - \alpha)\cdot$ 100% credibility region for λ under the above model as well as the posterior probability of the hypothesis $H_0 : \lambda \leq 1$. If `y` and `e` are length n vectors the function should return a matrix of dimension $n \times 4$.

(f) (4 points) As a consistency check, let the data be $\boldsymbol{y} = (13, 5, 36)'$ and $\boldsymbol{e} = (5.7219, 8.9395, 40.8851)'$. In an empirical Bayes setting of the above model determine $\boldsymbol{\eta} = (a, b)'$ by maximizing the marginal likelihood. State the value found for $\boldsymbol{\eta}$. Also use your `psummary` function for the three data pairs. State the results and illustrate the information update by showing prior density, likelihood and posterior density for the first data pair $(y_1, e_1)'$ in appropriate fashion. Comment what you see.

(g) (4 points) Determine an expression for the MSE of the posterior mean $E(\lambda|y, e)$. Implement the MSE as a R function and plot the MSE for both the posterior mean and the MLE for $\lambda \in [0, 2]$ when $e = 10$ and $a = 5, b = 4$. Interpret the plot.

Exercise 4 (25 points)

In this exercise you have to read the publication by Liu et al. (2003) (available from the course home page). Based on your reading you then have to answer a number of questions related to the paper.

(a) (4 points) Write a 1/2 - 1 page summary motivating and explaining using mathematical notation the available data and the model used in the paper. What were the questions of interest and how were these questions were answered.

(b) (1 points) Use the Dialysis Facility Report Data for the fiscal year 2017 available from `https://www.dialysisdata.org/` as the file *FY2017 Dialysis Facility Report Data* (100 MB!). Download the data and restrict to all facilities in the state of New York. Generate a subset dataset `dia`, which for all NY facilities contains the follow-up patient years (`yrs_at_risk`) in 2015, the expected number of deaths (`mu`) for the year 2015, the actual observed number of deaths (`y`) in the year 2015 and the standardized mortality ratio (SMR) (`rho_ml`).
Hint: The 2017 Data Dictionary contains a description of the 3391 variables in the dataset.
Note: While your hand-in exercise shall contain the R code to perform the data pre-processing of the full `DFR_Data_FY2017.csv`, please only include the `dia` data as part of your hand-in, e.g., as CSV file or RDS file.

(c) (2 points) Create a table as in Table 1 of the paper, but using only the 5 provider size classes [0,25), [25-50), [50,100), [100,150), ≥150. Interpret the result.

(d) (4 points) Write a JAGS program implementing the model in Section 5 of Liu et al. (2003). Choose the priors as in the paper and perform an analysis for the 2015 data of all NY facilities in the dataset `dia`. *Hint:* Note the different parameterisation between JAGS and Liu et al. (2003) for the Gamma and the Gaussian distribution.

(e) (2 points) Generate 2,500 samples from your JAGS model and investigate the convergence of your chain. Discuss, if further samples need to be generated and act accordingly. At the end you should have a sample of size 2,500.

(f) (4 points) Compute for each facility the posterior mean $E(\rho_k|\boldsymbol{y},\boldsymbol{\mu})$ using the above samples. Also find an 95% equitailed credibility interval for ρ_k. Use your results to create a plot as in Fig. 3 of the paper. Also create a plot for your data, which is similar to the paper's Fig. 1. Comment on the differences between the two plots you have created in terms of variability and estimation precision.
Hint: The R function `pois.test` might be helpful to create the Fig. 1 like plot.

(g) (3 points) State the MLE, the posterior mean from the above JAGS model as well as the posterior mean from an equivalent empirical Bayes analysis (see Exercise 3) for the facility *Westchester Center for Renal Care*. Explain the differences in the three values.

(h) (2 points) For the facility *Westchester Center for Renal Care*, calculate the posterior probability for the hypothesis that $\rho_k > 1$. Also calculate the prior probability for this hypothesis in the following two situations:

(1) under the model of Liu et al. (2003)

(2) under the empirical Bayes model

Interpret your findings.

(i) (3 points) Extend your JAGS program such that it also determines the rank and the percentile rank for each facility. Determine for each facility the posterior probability that it belongs to the top-3 dialysis facilities in the state of New York (measured in terms of the SMR). For the three facilities having the highest probability to belong to the top-3: show the facility's name, its probability for belonging to the top-3 and its MLE. Which of the three facilities would you go to, if you were a patient? Motivate your choice.

References

Liu, J., Louis, T. A., Pan, W., Ma, J. Z., and Collins, A. J. (2003). Methods for Estimating and Interpreting Provider-Specific Standardized Mortality Ratios. Health Serv Outcomes Res Methodol, 4(3):135–149.