301115 Advanced Statistical Methods

Assignment 2

Spring 2019

In this assignment there are 4 questions. For each question you should draw appropriate plots and summary tables needed to present the results at each stage of the analysis. There is no need to provide commentary on your methods or results beyond that required to fully address the assignment questions.

Your assignment can be submitted as either a PDF or a Word document. It is advisable that your assignment includes all of the code used to perform your analysis in an appendix (not within the body of the report), so that partial credit can be awarded in case of error. Do not use raw R (text) output to present your results, instead present all results using appropriate plots, summary tables and the text of your report. Submission is due by the end of Friday for week 16 (8th Nov 2019). Submission is by the vUWS online system.

Assignment 2

Question 1

a. In an experiment to determine the effectiveness of a new drug, a random sample of 38 patients were treated and 32 showed some benefit from the drug. Assuming a Uniform prior, state and plot the posterior distribution for the proportion of patients that might benefit from the drug.

b. A second drug was also tested on a different random sample of 44 patients and 39 were found to benefit. Compute a posterior mean for the difference in proportions of patients who benefit from the drugs. Also compute a 95% credible interval for the difference in proportions. (Show clear how you obtained this and how many simulations were used (if any)).

Question 2

A survey is conducted to measure alcohol consumption in the population. Each week, participants in the survey record whether they consumed one or more standard drinks that week. For each individual, it is desired to estimate pD, the proportion of weeks in which they consume one or more standard drinks. In modelling pD, we assume that it is fixed for each individual (i.e. does not change over time), but that it varies across the population.

a. The data set `pd.csv` contains estimates of pD from 1000 individuals who have already completed the survey. When attempting to estimate pD for a new individual, a possible choice is to use an "empirical Bayes" approach, in which a prior distribution is formed based on previous estimates. Taking this approach, build a prior distribution from `pd.csv` data by using the `mle` function to fit a Beta distribution (hint: use `dbeta` within `mle`). State the resulting values for the prior distribution parameters, and plot the prior distribution.

b. Using the prior from part a. state and plot the posterior distribution of pD in an individual who reported drinking one or more standard drinks in 13 out of the 20 weeks measured in the survey.

c. A participant drops out of the study before completing 20 weeks, and reports drinking one or more standard drinks in 3 out of the 4 weeks they completed. Using the prior from part a. what is the MAP estimate of pD for this individual. What is the maximum likelihood estimate of pD (i.e. without using a prior?). Compare the two estimates for this individual and comment on their difference.

Question 3

The x and y coordinates of an arrow shot into a target are assumed to be independently Normally distributed with means μ_x and μ_y and standard deviation $s = 2$ for both. A prior for μ_x and μ_y is assumed, that is flat over the circular region of radius 5. i. Compute and plot a grid based prior over the range -10 to 10 in both dimensions using 101 grid points. ii. For a single observation $x = 2.35$ and $y = 0.95$, compute and plot the likelihood on the same grid. iii. Compute and plot the posterior using the prior in a. iv. By simulation or otherwise, find the posterior mean for μ_x and μ_y based on this grid. v. The means of 9 observations are $\bar{x} = 0.43$ and $\bar{y} = 0.09$. By adjusting the standard deviation s, repeat ii.–iv. above for this data.

Question 4

The data set `ftse.csv` contains the log-returns of the FTSE100 stock exchange index for more than 2000 days. It is assumed the these log-returns are Normally (Gaussian) distributed, but that there may be a number of hidden underlying states.

a. Fit a mixture of two Normal distributions to the log-returns
b. Fit a 2-state hidden Markov model (using a Gaussian family) to the log-returns. Compare to the model in a. using a likelihood ratio test and AIC. What are the conclusions?
c. Plot the posterior probability of being in the state with the highest mean for the model in b.
d. What is the stationary distribution (steady states) for a Markov chain with transition matrix as estimated in b. above?