ST2195 Coursework Project

Instructions to candidates

This project contains two questions. Answer **BOTH** questions. All questions will be given equal weight (50%).

Part 1 In this part, you are asked to work with the Markov Chain Monte Carlo algorithm, in particular the Metropolis-Hastings algorithm. The aim is to simulate random numbers for the distribution with probability density function given below

$$f(x) = \frac{1}{2}\exp\left(-|x|\right),$$

where x takes val ue of x. More specifically, you ar them using the following version s random walk Metropolis) that co

Random walk Met

Step 1 Set up an initi e real number s.

Step 2 Repeat the fo

- Simulate mean x_{i-1} and standard
- Compute
- Generate veen 0 and 1.
- If $u < r\,(x$

(a) Apply the ran $s = 1$. Use the generated sa l density plot in the same figure. Note that these provide estimates of $f(x)$.Overlay a graph of $f(x)$ on this figure to visualise the quality of these estimates. Also, report the sample mean and standard deviation of the generated samples (Note: these are also known as the Monte Carlo estimates of the mean and standard deviation respectively).

Practical tip: To avoid numerical errors, it is better to use the equivalent criterion $\log u < \log r\,(x_*, x_{i-1}) = \log f\,(x_*) - \log f\,(x_{i-1})$ instead of $u < r\,(x_*, x_{i-1})$.

(b) The operations in part 1(a) are based on the assumption that the algorithm has converged. One of the most widely used convergence diagnostics is the so-called $\widehat{R}$ value. In order to obtain a valued of this diagnostic, you need to apply the procedure below:

- Generate more than one sequence of $x_0, \ldots, x_N$, potentially using different initial values x_0. Denote each of these sequences, also known as chains, by $(x_0^{(j)}, x_1^{(j)}, \ldots, x_N^{(j)})$ for $j = 1, 2, \ldots, J$.
- Define and compute M_j as the sample mean of chain j as

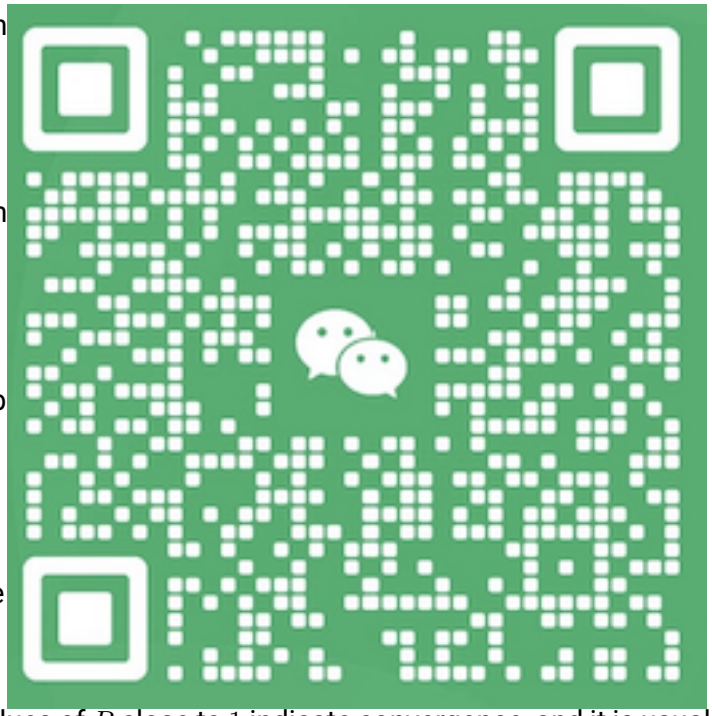
$$M_j = \frac{1}{N} \sum_{i=1}^{N} x_i^{(j)}.$$

  and V_j as the within sample variance of chain j as

$$V_j = \frac{1}{N} \sum_{i=1}^{N} (x_i^{(j)} - M_j)^2.$$

- Define an
- Define an

  and the b
- Compute



In general, values of R close to 1 indicate convergence, and it is usually desired for $\widehat{R}$ to be lower than 1.05. Calculate the $\widehat{R}$ for the random walk Metropolis algorithm with $N = 2000$, $s = 0.001$ and $J = 4$. Keeping N and J fixed, provide a plot of the values of $\widehat{R}$ over a grid of s values in the interval between 0.001 and 1.

Part 2 The 2009 ASA Statistical Computing and Graphics Data Expo consisted of flight arrival and departure details for all commercial flights on major carriers within the USA from October 1987 to April 2008. This is a large dataset; there are nearly 120 million records in total, and it takes up 1.6 gigabytes of space when compressed and 12 gigabytes when uncompressed. The complete dataset, along with supplementary information and variable descriptions, can be downloaded from the *Harvard Dataverse* at
`https://doi.org/10.7910/DVN/HG7NV7`

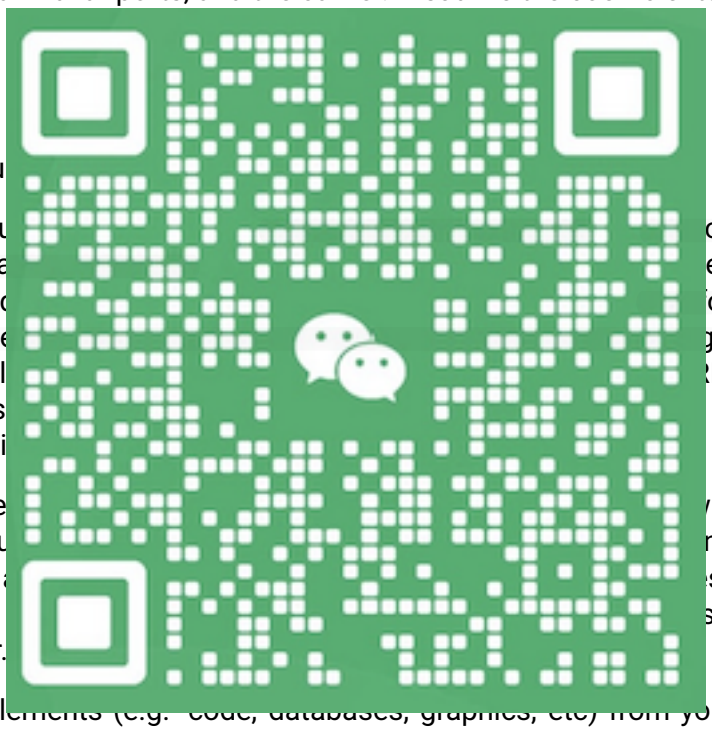
Choose any subset of ten consecutive years and any of the supplementary information provided by the Harvard Dataverse to answer the following questions using the principles and tools you have learned in this course:

(a) What are the best times and days of the week to minimise delays each year?

(b) Evaluate whether older planes suffer more delays on a year-to-year basis.

(c) For each year, fit a logistic regression model for the probability of diverted US flights using as many features as possible from attributes of the departure date, the scheduled departure and arrival times, the coordinates and distance between departure and planned arrival airports, and the carrier. Visualize the coefficients across years.

General Instructions



- All questions shou
- Your answers shou ore than 1 page for part 1 and 6 pa es and table of contents but inclu ormat and also contain adequate e g. In addition to the report, you will Markdown and Jupyter notebooks the designated Atrio or VLE submi
- For part 2, each re data up to the answer for each qu ning operations you carry out, and a scribed in each structured report. s and tables as part of the answer.
- If you are using el ur answer to a previous question to answer the current one, you will need to refer to those elements.
- You should also supply the code you used to answer each question, in a way that can be used by someone else to replicate your analyses. You can do this either as separate scripts or separate RMarkdown/Jupyter notebooks per question, clearly indicating (both with comments and in the filename) which question each script refers to.