ST522/ST816 Computational Statistics: Project 1
Spring 2025

Vaidotas Characiejus

March 20, 2025

The solutions have to be submitted online on itslearning until the deadline of April 23, 2025 at 23:59 by uploading: (a) a PDF file of the solutions; (b) a script file with no special formating containing exactly the same code as used in the solutions for grading purposes. No submissions will be accepted by email or in person. No revisions of the solutions will be allowed. Late submissions will not be accepted.

The report can be written in groups of up to 3 students. Only one report has to be submitted for each group and the students who form a group will receive the same number of points for the report. Every student of the group is expected to be able to explain every part of the report.

You are not permitted to

(i) discuss the solutions with anyone outside your group (this includes your fellow students as well as posting questions about the exercises on the internet);
(ii) use generative models (such as ChatGPT) to write your answers.

Any suspicious similarities of the solutions or any other concerns will be investigated.

You have to provide a suitable amount of explanations and intermediate computations in your solutions. You have to use `R` for your solutions where you perform computations. Any function from any package can be used but a suitable explanation on what the function does needs to be included.

It is strongly recommended to write reports using R Markdown, Quarto, LaTeX, Jupyter Notebook, knitr or some similar tool.

The total number of points is 50.

Good luck!

1 Simulating observations from Student's t-distribuion

Let X be a random variable with Student's t-distribution and parameter $\nu \in (0, \infty)$. The density function of Student's t-distribution with parameter $\nu \in (0, \infty)$ is given by

$$f(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\Big(1 + \frac{x^2}{\nu}\Big)^{-(\nu+1)/2}$$

for $x \in \mathbb{R}$, where Γ is the gamma function.

(a) (2 points) Suppose that $Y = |X|$. Derive the density function of Y.

(b) (5 points) Can we use an exponentially distributed random variable with parameter $\lambda \in (0, \infty)$ as the instrumental random variable for the acceptance-rejection algorithm to simulate observations of Y? Include an explanation. The density function of the exponential distribution with parameter $\lambda \in (0, \infty)$ is given by

$$h(x) = \lambda e^{-\lambda x}$$

for $x > 0$.

(c) (15 points) Propose an instrumental density g for the acceptance-rejection algorithm that simulates observations of X when $\nu \in [1, \infty)$. Explain how the observations of g can be generated and show that g can be used in the acceptance-rejection algorithm to simulate observations of X when $\nu \in [1, \infty)$ by finding the value of

$$\sup_{x \in \mathbb{R}} \frac{f(x)}{g(x)}.$$

Calculate the expected number of simulations until an observation from g is accepted when $\nu = 2$.

2 Density estimation

The objective of this exercise is to fit a density function to time intervals between the starts of successive eruptions of the Old Faithful geyser in Yellowstone National Park in Wyoming, United States. This is a commonly used data set to illustrate the performance of kernel density estimators, mixture distributions, clustering, etc. The data set is one of the built-in data sets in `R` and can be accessed by typing `faithful` in the `R` console. The variable that is the focus of this exercise is `waiting`. Complete the following tasks.

(a) (2 points) Create a histogram of the `waiting` variable. Describe the main characteristic features of the distribution of this variable.

(b) (5 points) Use the kernel density estimator with the standard normal kernel to estimate the density function of the `waiting` variable. Using the oversmoothed bandwidth selector $\hat{h}_{\text{OS}}$, choose the bandwidth and report the chosen value. Plot the estimates of the density function with bandwidths $\hat{h}_{\text{OS}}$, $\hat{h}_{\text{OS}}/2$, $\hat{h}_{\text{OS}}/4$, and $\hat{h}_{\text{OS}}/8$ in a single plot on top of the histogram and add appropriate labels so that it would be easy to tell which density is which. Describe what features of the density function are present for various amount of smoothing. Which of the four bandwidths seems to be the most suitable bandwidth for the `waiting` variable?

(c) (4 points) Does it make sense to use the normal scale bandwidth selector for the `waiting` variable? If we still use it, how will this affect the estimated density? Explain your answer.

(d) (7 points) Calculate the bandwidths using the following selectors for the standard normal kernel and report their values: (a) normal scale

$$h_{\text{NS}} = \left[\frac{8\pi^{1/2}R(K)}{3\sigma_K^4 n}\right]^{1/5} \min\left\{s, \frac{\text{IQR}}{\Phi^{-1}(3/4) - \Phi^{-1}(1/4)}\right\},$$

where s is the sample standard deviation and IQR is the sample interquartile range; (b) pseudo-likelihood; (c) least squares cross-validation (this is also called unbiased cross-validation); (d) biased cross-validation; (e) Sheather-Jones. Plot the estimates of the density function with all of the five different bandwidths in a single plot on top of the histogram and add appropriate labels so that it would be easy to tell which density is which. Which of the five bandwidths seems to be the most suitable bandwidth for the `waiting` variable?

(e) (10 points) Consider the following kernels: (a) standard normal, (b) Epanechnikov, (c) rectangular, (d) triangular, and (e) biweight. Select the bandwidth obtained using the Sheather-Jones bandwidth selector for the standard normal kernel, transform it into equivalent bandwidths that minimise the corresponding asymptotic mean integrated squared error for the remaining four kernels and report their values. Plot the estimates of the density function with all of the five different kernels in a single plot on top of the histogram and add appropriate labels so that it would be easy to tell which density is which. What are the differences between the densities estimated using these five different kernels? Which of the five kernels seems to be the most suitable kernel for the `waiting` variable?