University of Edinburgh
School of Mathematics

Statistical Research Skills

Assignment 3—Simulation Report **Semester 2, 2021–2022**
[**Deadline: 26 April (5PM)**]

Write a 1500 word (excluding references) simulation report with the goal of comparing the performance of the kernel density estimator with that of another density estimator. You should typeset your report using RMarkdown and appropriate use of Tidyverse, as opposed to base R, tools will be rewarded. Please submit a knitted version with the code hidden (use the option echo = false) as well as the Rmd file. However do not rely on us seeing the code, you will be penalised if we need to refer to your code to understand what you did.

Kernel density estimation

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f$. The kernel estimator of f is defined as

$$\widehat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i),$$

where $K_h(\cdot) = K(\cdot)/h$, K is a density (*kernel*), and $h > 0$ is a parameter controlling smoothness of the estimate (*bandwidth*). The estimator is readily available from `R`, using `density`.

Consider dividing your report into two sections:

1 Data Generating Processes and Preliminary Experiments

- Identify the goal of the experiment and what competitors of the kernel density estimator you will be comparing against. The histogram is a natural option, but with some extra research you could find other natural competitors. Remember to include on the bibliography references to methods that you have used.[1]

- Describe formally what the simulation scenarios are (or data generating processes) from which you will be simulating data [say, Scenario 1: $f_1(x) = \phi(x)$; Scenario 2: $f_2(x) = \ldots$].

- Run a one-shot experiment and illustrate in a figure the methods being compared, against the true densities [say, $f_1(x)$, $f_2(x)$, and $f_3(x)$]. Interpret the figure. Are the methods recovering 'well' the true curves? If yes, comment on that. To keep the inquiry simple, you can use `R` default option for the bandwidth parameter. Yet other options exist (e.g. cross-validation). Remember to add references on any methods that you have used.

- Anticipate strengths and weaknesses with the methods, but keeping in mind the disclaimer that this is a one-shot experiment.

[1] While a clear and organized comparison with several methods would be the most interesting one, for the purposes of this report a clear comparison against a single method (say, histogram) will be more appreciated than an unclear one with several methods.

2 Monte Carlo Simulation Study

- Explain in a coherent way what experiment you will be conducting now.
- Conduct the Monte Carlo simulation study for the sample sizes $n = 250$, 500, and 1000.
- For each fixed sample size, report in a figure the ISE (Integrated Squared Error), underlying each simulated data set ISE $= \int \{\widehat{f}(x) - f(x)\}^2 \, \mathrm{d}x$.
- What happens when you increase n?