University of Edinburgh, School of Mathematics

Statistical Research Skills

Assignment 3 - Simulation Report

Johnny Lee

26th Apr 2022

1. Introduction

This report consists of two main parts. In the first part, we aim to compute one-shot experiment on kernel density estimator and compare against its other competitors such as orthogonal series estimator and penalised kernel density estimator. Hence, we will describe the preliminary experiment in the first part. In the second part, we will conduct a Monte Carlo simulation study for different sample sizes on the previously suggested methods. Thus, we will evaluate the integrated squared error (ISE) for different cases of sample sizes. Then we will end this report in the conclusion.

2. Preliminary Experiment

This section consist of the first part of the report where we conduct the preliminary experiments with 3 different density estimators. We will introduce the detailed methodology with its mathematical equation. With random data generation from normal and beta distribution, we will conduct one-shot experiment on these estimators.

2.1 Methodology

2.1.1 Kernel Denisty Estimator

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

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

where $K_h(x) = \frac{K(x)}{h}$, K is a kernel and $h > 0$ is a parameter controlling smoothness of the estimate. Thus, we will perform an univariate density estimate with default kernel and bandwidth settings in this experiment.

2.1.2 Orthogonal Series Estimator

With reference to (Kreyszig, 1991)[5], Kreyszig introduced the orthogonal series estimator using normalised Hermite polynomials. These polynomials form a orthonormal sequence for the univariate case as below,

$$\hat{f}(x) = \frac{1}{(2^i i! \sqrt{x})^{\frac{1}{2}}} \exp(-x^2/2) H_i(x)$$

and

$$H_0(x) = 1, \quad H_i(x) = (-1)^i \exp(x^2) \frac{d^i}{dx^i} \exp(-x^2)$$

where $H_i(x)$ is called the *Hermite polynomial of order i* and orthogonal with respect to the probability density function. Furthermore, we can further simplify the process as Kreyszig showed its properties in a recurrence relation as below,

$$H_{i+1}(x) = 2xH_i(x) - H_i'(x), \quad H_i'(x) = 2iH_{i-1}(x)$$

Thus, we defined a new function to proceed to the iteration above to obtain the series of estimation.

2.1.3 Penalised Kernel Density Estimator

(Kauermann et al, 2009)[4] introduces the penalised likelihood

$$\hat{f}(x) = \sum_{n=-m}^{m} c_n \phi_n(x),$$

where $\phi_n(x)$ is the basis densities. Then the weight c_n is parameterised as follow

$$c_n(y) = \frac{\exp(\beta_n)}{\sum_{n=-m}^{m} \exp(\beta_n)}$$

with $\beta_0 = 0$ and $\boldsymbol{\beta} = \beta_{-m}, \ldots, \beta_{-1}, \beta_1, \ldots \beta_m$ so that $\int f(x)dx = 1$. (Deng et al, 2011)[1] then further suggested the simplified kernel approach that looks similar to kernel density estimator. The following expression shows the penalised approach for nonparametric density estimation for univariate case,

$$\hat{f}(x) = \frac{1}{m} \sum_{n=1}^{m} K\left(\frac{x - \mu_n}{h}\right)$$

where μ_i is a hyperparameter known as knots being placed on an equally spaced locations on the domain of the dataset.

2.2. Data Generating Process

We generated two random distributions, normal distribution and beta distribution. For the first distribution, we randomly generated $X \sim N(0, 1)$ with 1000 samples. Then the second distribution takes $X \sim Beta(2, 4)$ with 1000 samples. These values are chosen on a random basis and we are generating the data in a random process in base `R`.

2.3. One-shot Experiments

Now we proceed to one-shot experiment on two randomly generated distribution in the previous section. First of all, one-shot experiment is the outcome produced by a single simulated data set.

2.3.1. Normal Distribution

Let us look at Figure 1. The figure illustrates the true density with a histogram on the actual data set that we generated. Then we plotted 3 additional line that represent each of the estimators. By looking at the plot we can clearly observe that penalised kernel density estimator resembles the true density line the most compared to others. For the kernel density estimator, the peak is slightly on the right side of the distribution. On the other hand, orthogonal series estimator is shown to have a two hump and being less smoother compared to the rest. Overall, we can conclude that penlaised kernel density estimator is the best among others in the one-shot experiment on normal distribution.
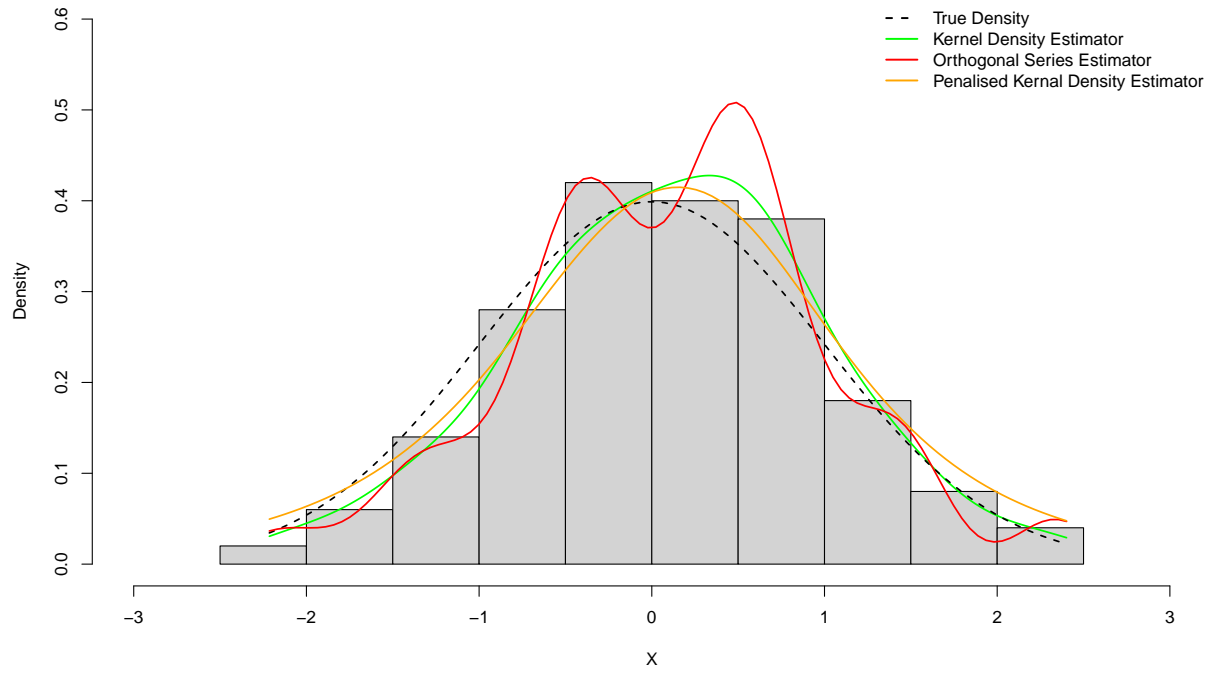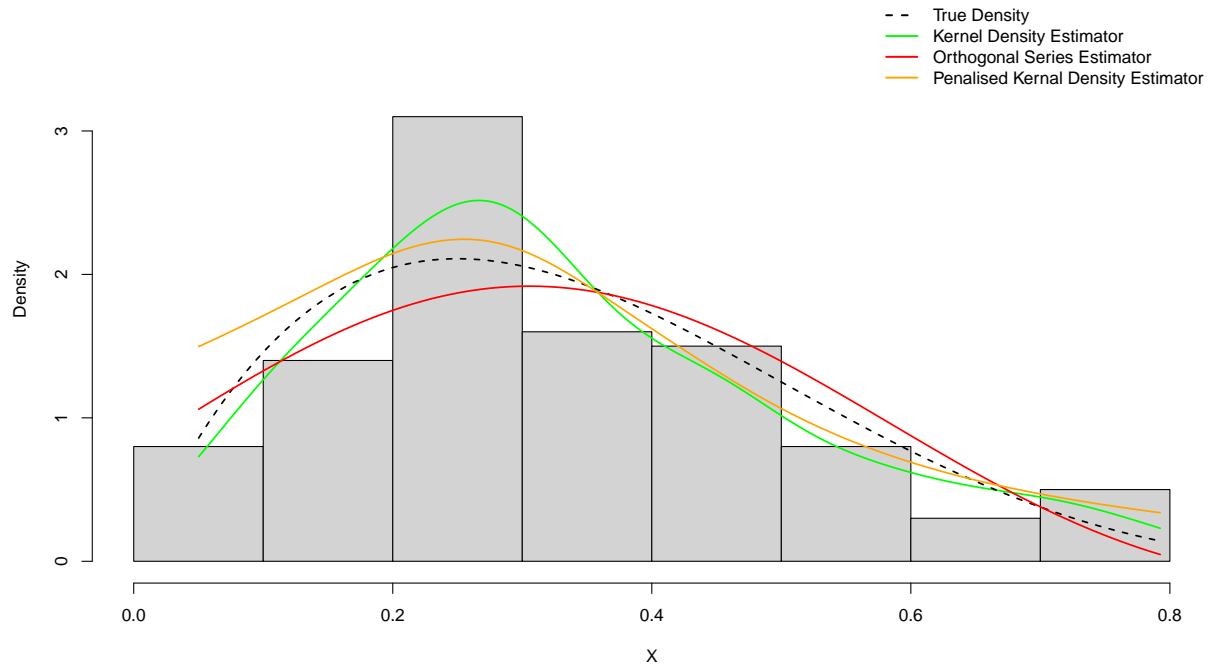
2.3.2. Beta Distribution

Figure 1: One-shot Experiment on Beta distribution

3. Monte Carlo Simulation Study

Now we move onto the Monte Carlo Simulation Study, where we repeat the one-shot experiment R times, for different simulated dataset. This brings advantage in concluding with a solid result and logic. Hence, we repeated this process with 100 times on the randomly generated $X \sim N(0,1)$ on different sample sizes. To that, we computed the integrated squared error,

$$ISE = \int \{\hat{f}(x) - f(x)\}^2$$

to compare the performances of each density estimator on different sample sizes. Note that for the reproduceability, we fixed a seed in R using `set.seed(1)`.
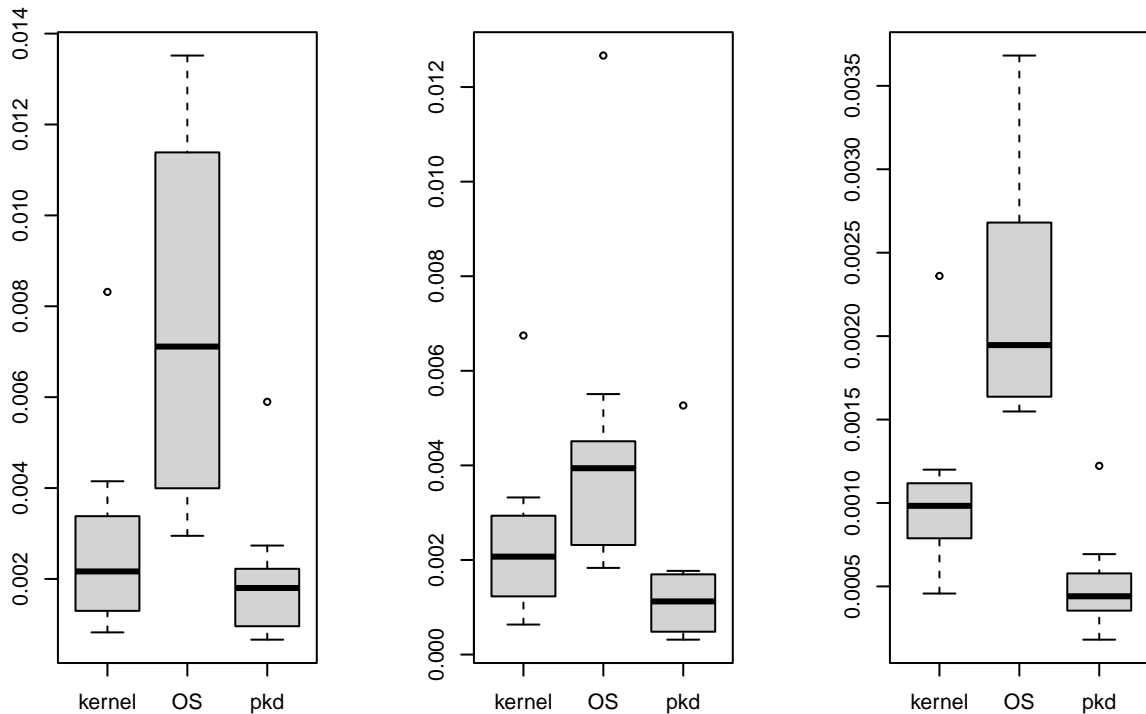

Figure 2: Boxplot of Different Estimators with Different Sample Sizes
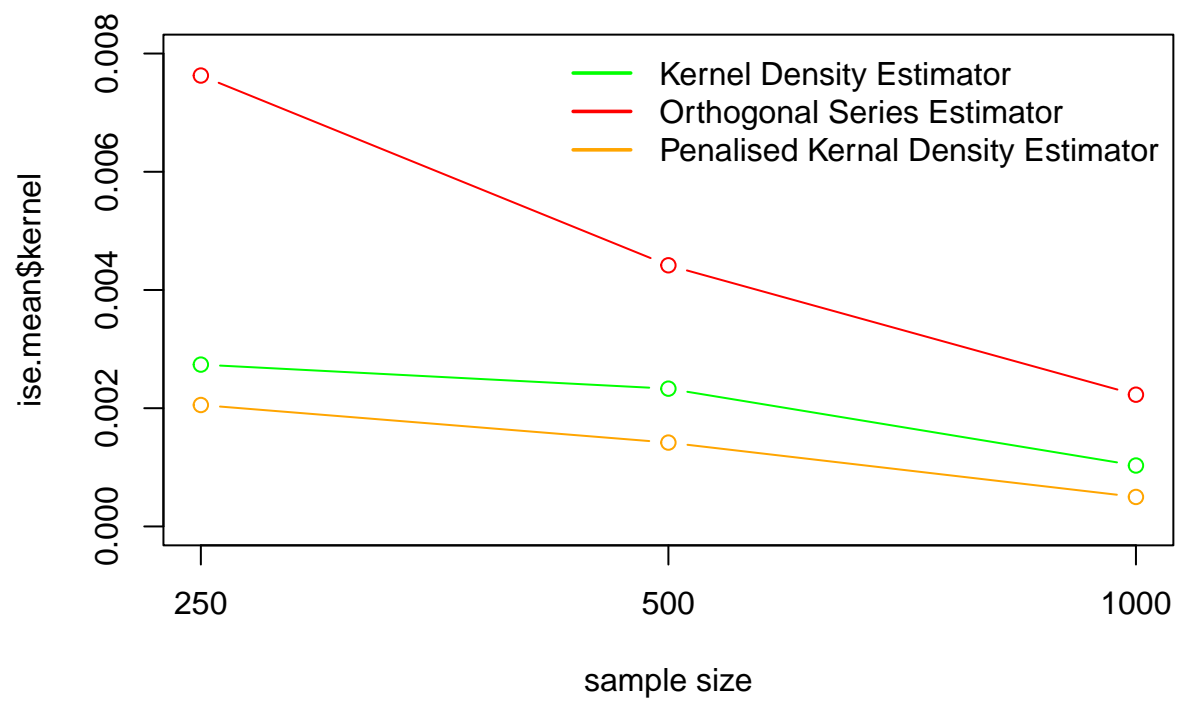
4. Conclusion

Figure 3: Mean Integrated Squared Error with Different Sample Sizes

Reference

1. Deng, H. and Wickham, H., 2011. Density estimation in R. Electronic publication.

2. Girolami, M., 2002. Orthogonal series density estimation and the kernel eigenvalue problem. Neural computation, 14(3), pp.669-688

3. Kauermann, G. and Schellhase, C., 2019. Density Estimation with a Penalized Mixture Approach.

4. Kreyszig, E., 1991. Introductory functional analysis with applications (Vol. 17). John Wiley & Sons.