University of Edinburgh, School of Mathematics

Statistical Research Skills

Assignment 3 - Simulation Report

Johnny Lee, s1687781

26th Apr 2022

1. Introduction

Density estimation is to form an estimate based on the original data of a probability density function which is popular in various fields including statistics and economics. In this report, we aim to investigate the performance of various density estimators and this report consists of two main parts. In the first part, we aim to perform 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 consists 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 univariate 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 by (Silverman, 2018)[4] 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)[3], 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)[2] introduces the penalised likelihood approach

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

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's expression. 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 with a fixed random seed.

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 a pre-experimental design which performs the same algorithm on a single simulated data set. With the outcome, we can design a more sophisticated experiment in the later part and choose suitable hyperparameters for efficient and accurate computation.

2.3.1. Normal Distribution

Let us look at Figure 1. The figure illustrates the true density (dotted line) with a histogram on the actual data set that we generated. Then we plotted 3 additional lines that represent each of the estimators. We can clearly observe that kernel density estimator (green) and penalised kernel density estimator (orange) are having similar shapes to the true density. However, penalised kernel density estimator resembles the true density line the most compared to others. For the kernel density estimator (green), the peak is slightly on the right side of the distribution. On the other hand, orthogonal series estimator (red) is shown to have a two hump and being less smoother compared to the rest. Overall, we can conclude that penalised kernel density estimator is the best among others in the one-shot experiment on normal distribution.

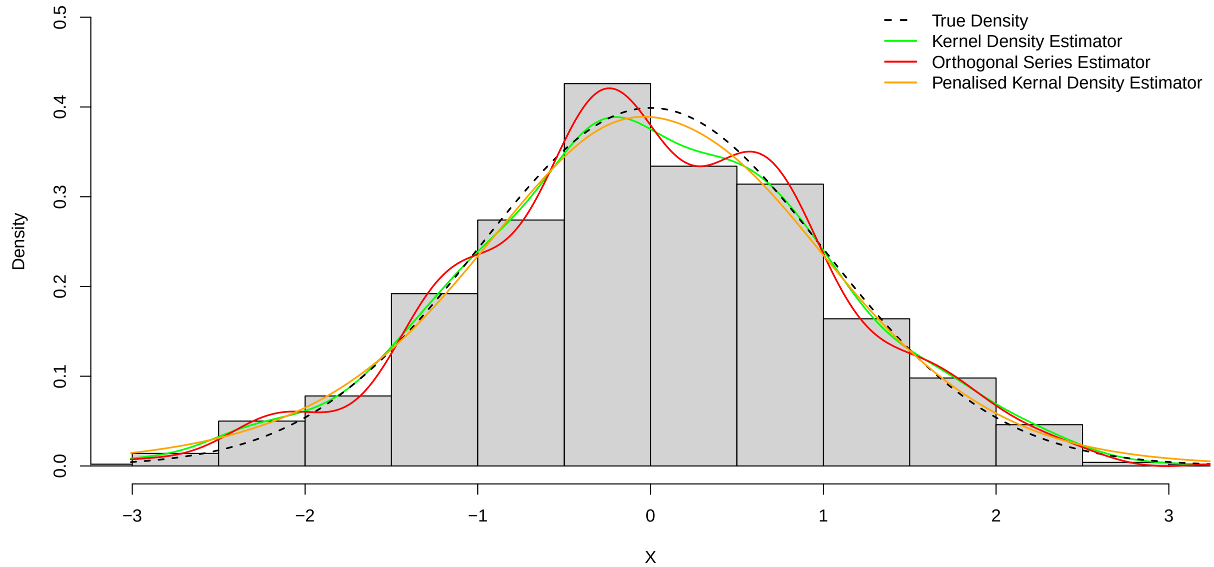

Figure 1: One-shot Experiment on Normal distribution

2.3.2. Beta Distribution

Similarly we performed one-shot experiment on the beta distribution that we generated in the previous part. Looking at Figure 2, we illustrated the true density (dotted line) with a histogram on the actual data set that we generated from. Then we plotted 3 additional lines that represent each of the estimators. By looking at the plot we can clearly observe all curves are smoothly plotted. The kernel density estimator (green) and penalised kernel density estimator (orange) both resembles the true density line. On the other hand, orthogonal series estimator (red) is shown to have the lowest peak with its peak centered on the right side and reasonably apart from the true density. Thus, for beta distribution we can conclude that kernel density estimator and penalised kernel density estimator work well. We will move on to the Monte Carlo simulation to validate our experiment to discuss which estimators have the best performance.

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. Thus, we simulate data $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f_\theta$, for a fixed $\theta \in \Theta$. Our goal in this section is to assess the performance of an estimator,

$$\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$$

at recovering the true θ, over R simulated data sets. Hence, we repeated this process with $R = 100$ times on the randomly generated $X \sim N(0, 1)$ on sample sizes with 250, 500 and 1000. 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. In addition, we computed the mean integrated squared error (MISE), and illustrated the results in figures below. Note that for the reproduceability, we fixed a random seed in all of our experiments.

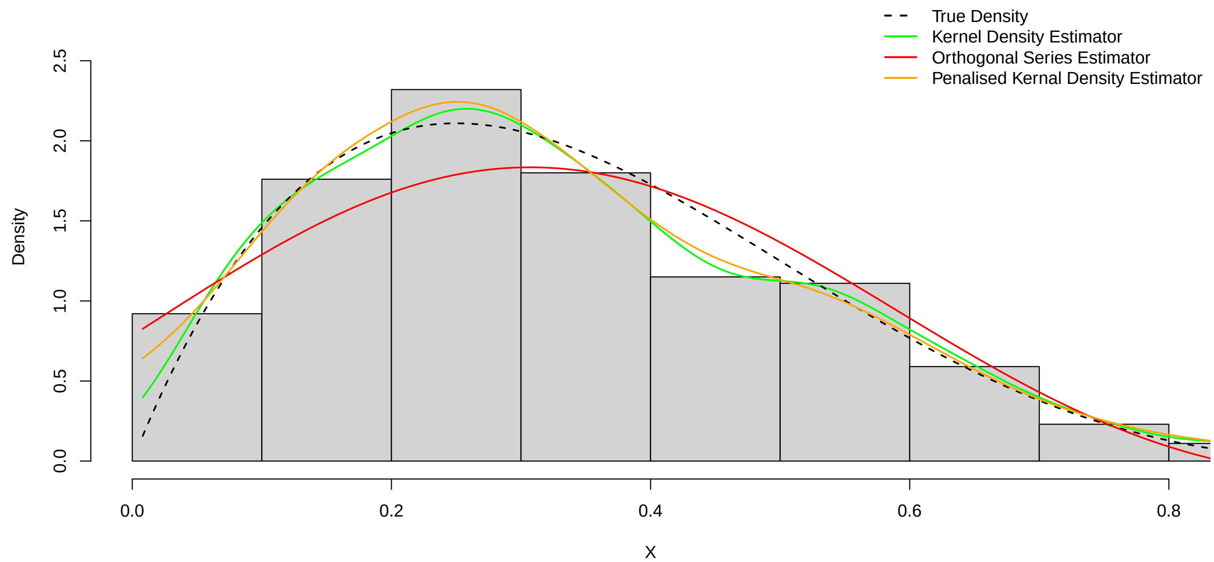

Figure 2: One-shot Experiment on Beta distribution

Referring to Figure 3, we plotted the boxplots of three different estimators with different sample sizes. `kernel` stands for the kernel density estimator, `OS` stands for the orthogonal series estimator and `pkd` stands for the penalised kernel density estimator. Scrutinising its general trend, we can observe that the `OS` has the largest quartile range and highest MISE values. We also observed that as the sample size, n increases, we can see that the ISE values decrease as when $n = 1000$ the ISE values are 10 times less than $n = 250$. Furthermore, penalised kernel density seems to show the shortest interquartile range for the ISE values and the lowest MISE among the others.

Figure 4 further illustrates the MISE values of the three estimators where it solidifies that all estimators have lower MISE values with larger sample size and penalised kernel density estimator having the lowest MISE values. Therefore, if an accurate estimation of densities is required, a large sample size is a better choice.

4. Conclusion

In conclusion, we conducted several experiments on three different density estimators. From one-shot experiment to Monte Carlo simulation study, Penalised kernel density estimator showed the best performance. Among kernel density estimator and orthogonal series estimator, penalised kernel density estimator had the lowest MISE and showed the best fit to the true values.

Reference

1. Deng, H. and Wickham, H., 2011. Density estimation in R. Electronic publication.
2. Kauermann, G. and Schellhase, C., 2019. Density Estimation with a Penalized Mixture Approach.
3. Kreyszig, E., 1991. Introductory functional analysis with applications (Vol. 17). John Wiley & Sons.
4. Silverman, B.W., 2018. Density estimation for statistics and data analysis. Routledge.

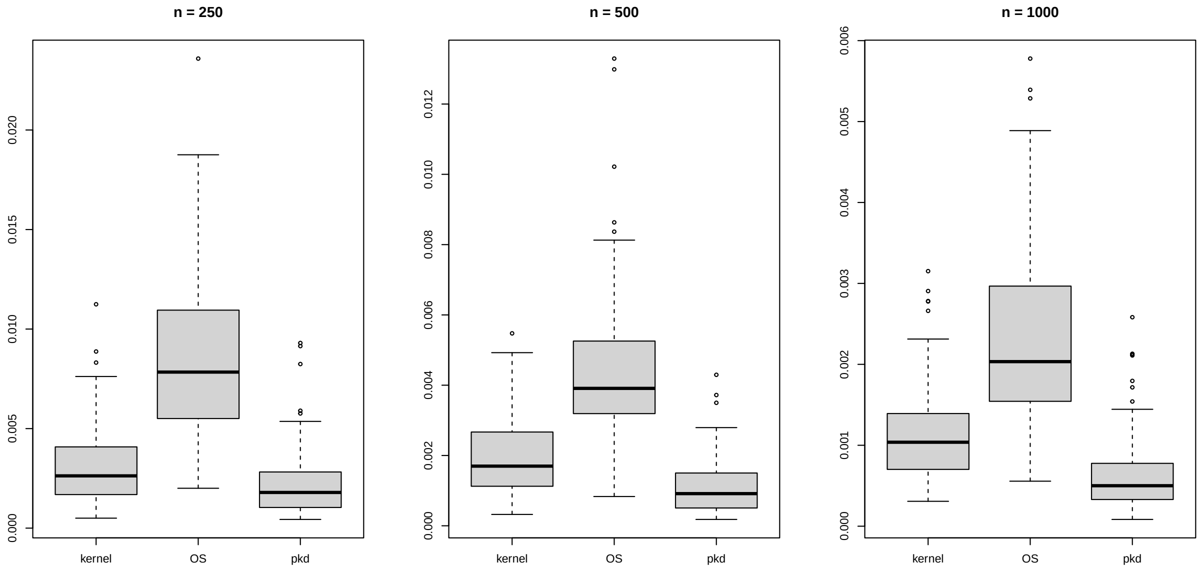

Figure 3: Boxplot of Different Estimators with Different Sample Sizes

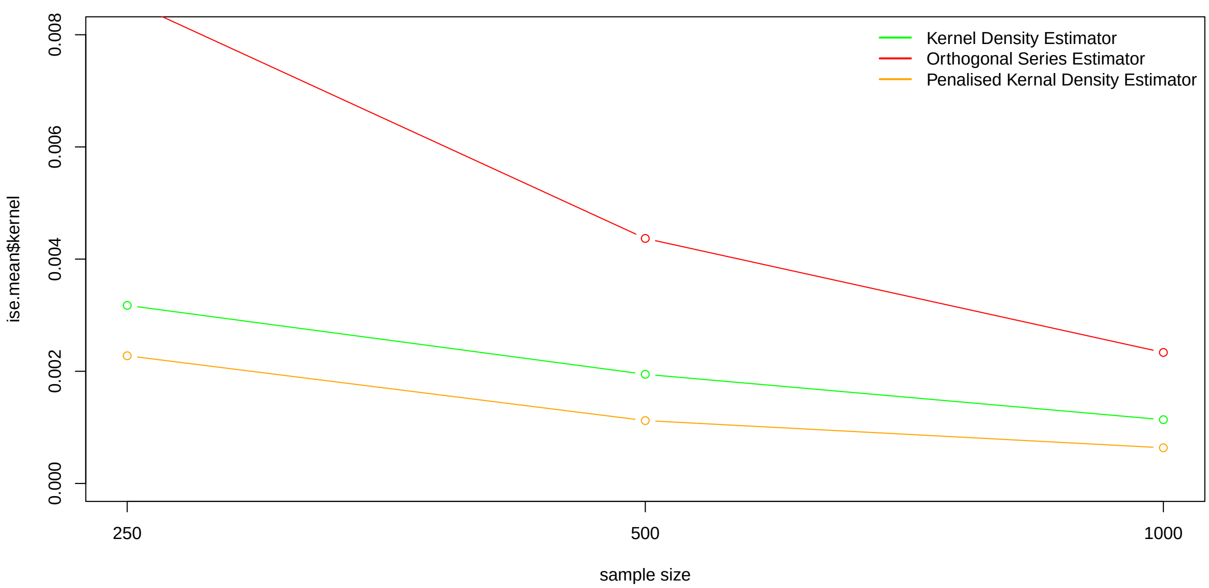

Figure 4: Mean Integrated Squared Error with Different Sample Sizes