TP 3

TP3 : Density estimation

3.1 Preparation

This TP refers to Chapter 3 of the course which was not treated in amphi. Do not hesitate to refer to it.

Consider a realization $x_1, \dots, x_n$ of a vector $X_1, \dots, X_n$ of identical and independent random variables of common density f. The purpose of this TP is to compare the kernels used to estimate the common density f. It must be understood that in practice f is unknown, here, to compare the efficiency of kernel and the size of the h window we will assume in a first part that f is the density of a standard Gaussian, in a second part we will assume that f is the density of a law of larger dimensionality.

1. If $K\colon \mathbb{R} \to \mathbb{R}_+$ is a statistical kernel and $\mu \in \mathbb{R}$ a constant, is the translation of K by the constant a, $\tau_\mu K$ still a statistical core?

2. If $K\colon \mathbb{R} \to \mathbb{R}_+$ is a statistical kernel and $\lambda \in \mathbb{R}*$ a nonzero constant, show that $d_\lambda K$ defined for any $x \in \mathbb{R}$ per $d_\lambda K(x) = \frac{1}{\lambda} K\left(\frac{x}{\lambda}\right)$ is still a statistical core.

3. Show that $K = \frac{1}{2} 1_{[-1.1]}$ is a statistical kernel, it is called the uniform kernel.

4. Show that $K(x) = 1_{[-1.1]}(x)$ is a statistical kernel, called the triangle kernel.

5. Show that $K(x) = \frac{3}{4}(1 - x^2) 1_{[-1.1]}(x)$ is a statistical kernel, called the Epanechnikov kernel.

6. Show that $K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ is a statistical kernel, called the Gaussian kernel.

Let $h > 0$ be a constant called the window. Let K be a statistical core. The function $\widehat{f}_h$, defined for any $x \in \mathbb{R}$, is defined by:

$$\widehat{f}_h(x) = \frac{1}{n} \sum_{k=1}^{n} d_h \tau_{X_i} K(x)$$

This is the density estimate f with the h window and the K kernel.

7. Show that $\widehat{f_h}$ is a probability density.

3.2 Lab session : part 1

The purpose of this part is to define, represent and compare the efficiency of the four kernels of the preparation part for estimating the density of a standard Gaussian f. It is therefore assumed that $X_1, \ldots, X_n$ is a sample of size n of independent variables and identically distributed according to the normal centered law reduced density f.

1. Create four functions $K1, K2, K3, K4$ corresponding respectively to uniform, triangle, Epanechnikov and Gaussian kernels.

2. Represent these four kernels on the same chart (use a different legend and colors). To do so, create a function that you will name **AllplotK** that will enter the parameters of the graph (the step, xmin, xmax, colors etc...) and represent the chart in return.

3. Generate a realization of size n of X according to a standard Gaussian law. (n is currently set at 100 in the script).

4. Set the **fchapeau** function which takes as argument a function K (the kernel), the window h and the realization of the sample X and a value x and returns the image of x by the $\widehat{f_h}$ function.

5. Represent on the same chart the f reference function as well as the four $\widehat{f_h}$ functions obtained with the $K1, K2, K3, K4$ kernels. You will add a different legend and colors to all curves. For this question, set $h = 2$. You will define a function as in question 2 to answer this question. This function will be named **Allplotfchapeauh2**.

6. Repeat the previous question with $h = 1$. Qualitatively, does the estimate differ more when varying the kernel used or the h window used? The new function for this question will be named **Allplotfchapeauh1** .

7. Redo the two previous questions for $n = 10$ and then $n = 1000$. For this question, four graphs must be constructed: the first for $(n, h) = (10, 2)$, the second for $(n, h) = (10, 1)$, the next for $(n, h) = (1000, 2)$ and the last for $(n, h) = (1000, 1)$. You will detail your reasoning in the script and comment on the results. **In the remaining of the Lab session, you will go back to $n = 100$.**

8. We are going to compute the quadratic error of an estimation:

$$SCE(h) = \sum_{i=0}^{500} (\widehat{f_h}(t_i) - f(t_i))^2$$

   the sum of squares of the differences between the image of t_i by the estimate of $\widehat{f_h}$ and the image of t_i per f, where $\{t_0, t_1, t_2, \ldots, t_{500}\}$ is a

discretization of the $[-5.5]$ interval of step $10/500$.
In other words,

$$-5 = t_0 < t_1 = -5 + 10/500 < t_2 = -5 + 20/500 < \cdots < t_{500} = -5 + 5000/500 = 5$$

Create a **SCE** function that takes as parameter a function (the kernel considered), h window, f reference density and returns $SCE(h)$.

9. Create a function **thebesth** that takes a function (the kernel in question) and another function f (the reference) into parameters and returns the index divided by 100 of the minimum list $\{SCE(k/100)\}_{1\leq k\leq 200}$. For each kernel, the best window for estimating the reference function is given by this function.

10. Create a function that graphically represents the four density estimates for these four kernels with the windows obtained via the **thebesth** function. Name it **Allplotfchapeauhoptimal**

3.3 Lab session : part 2

We will now exploit the features of **scikit-learn**. The **estimatedensity** function in the script is used to estimate density by Gaussian kernel (**kernel**$='gaussian'$) with for window h whose reference density is a Gaussian mixture (two Gaussians of medium mu1, mu2 and sigma1, sigma2). This function uses the scikit-learn package.

1. Execute this function with $mu1 = 0$, $mu2 = 5$ and $sigma1 = sigma2 = 1$, $N = 100$ and $h = 0.75$.

2. Compare to any other set parameter as in the previous question, the influence of the h window. We can test values of h between 0.2 and 1.5. Comment.

3. Makes variations to the parameters of the two Gaussian laws that define the Gaussian mixture. Comment.

4. Makes N vary and comment.

5. Other kernels can also be tested for example by replacing '*gaussian*' in the code with '*epanechnikov*'. Make this graph by running the **estimatedensity2** function with the same parameters as in question 1.