UNIVERSITY OF ORLEANS

Non-parametric estimator

Student:

1. Treasure Hunt [16 Points + 4 BONUS Points]

An old pirate is about to retire and wants his treasure, amassed throughout his life, to return to his descendants. Tired of the incessant quarrels between his two sons to find out who among them is the greatest pirate, he cannot decide how he should distribute his inheritance. That's when he had an idea...

He summons each of them separately, and gives them some of the data collected over time during the expedition during which the treasure was hidden:

- the eldest son receives readings of longitude and depth.
- the younger son receives latitude and depth readings.

He finally indicates to each of them, with a small smile, "a cross marks the location of the treasure" and "following my steps you will find the code (7 characters) which will allow you to open the chest. ".

Start by loading the workspace **PIRATES.RData**. It contains 4 variables:

- **t** : measurement instants
- **x** : latitude coordinates (relative).
- **y** : longitude coordinates (relative).
- **p** : seabed depth coordinates (ground elevation for positive values)

Preliminaries We first study the data common to the two threads in order to better understand what they contain.

1. What is the nature of the data? What is the sample size? When do the readings stop? Discuss in a few lines the interest of a non-parametric approach at this time of the study.

2. Give the expression for the kernel density estimator. Show that if the kernel **K** is a density, then this is also true for the estimator.

3. Representing in the same graphic window the kernel estimators of the density of the variables **t** (resp.**p**) by taking the *epanechnikov* kernel and successively using as smoothing parameter the values **0.005**, **0.05**, **5** (resp. **0.07**, **0.7**, **7**) and that determined by cross-validation. Use commandpar ***(mfrow=c(2,2))***.Comment on the graphs obtained. Remember to return the standard graphics window with the commandpar ***(mfrow=c(1,1))***.

4. Propose a model of the law of the instants of measurement from the results of the previous question then test its validity using the Kolmogorov-Smirnov test (it will be admitted that **t** has values in $[0, 4 \times \pi]$).

5. Calculate the values of $\mu_2(K)$ and of $R(K)$ for the triangular kernel. We remember that they are worth respectively $\frac{1}{5}$ and $\frac{3}{5}$ for the epanechnikov nucleus. What percentage of data would be needed in addition to obtain, with the rectangular kernel instead of the epanechnikov kernel, a similar precision in terms of **MISE** when the density is $\mathcal{C}^2$?

6. Represent the cloud of points corresponding to the depth readings (as a function of time). Comment on the result obtained.

7. Estimate the regression function non-parametrically $p = r_P(t) + \epsilon$ using 3 different ways to choose the smoothing parameter. Comment on the results. Which method do you retain (we will use it throughout the rest of the statement)? Why?

8. Give an estimate of the depth at the moment $\mathbf{t} = 3.228297$

1.1. Part A: The eldest son. We will use in this part only the variables **t**, **y** and **p** available to the eldest son. Anxious to reach the treasure before his younger brother, he tries to see what he can get out of the data he has...

1. Represent the cloud of points corresponding to the longitude readings (as a function of time). Comment on the result obtained.

2. Estimate the regression function non-parametrically $y = r_Y(t) + \epsilon$ at each instant of measurement. Save the results obtained under the name **y_e**.

BONUS: Propose a parametric modeling of the function r_Y and represent it on the point cloud of the previous question (Hint: we can do a linear regression of y on $\sin(t)$). Comment on the result obtained

1.2. Part B: The younger son. We will use in this part only the variable **t**, **x** and **p** available to the younger son. Anxious to reach the treasure before his eldest, he tries to see what he can get out of the data he has...

1. Represent the cloud of points corresponding to the latitude readings (as a function of time). Comment on the result obtained.

2. Estimate the regression function non-parametrically $x = r_X(t) + \epsilon$ at each of the instants of measurement and save them under the name **x_e**.

BONUS: Being passionate about trigonometry, the youngest son thinks that $r_X(t)$ can be modeled as a linear combination of $cos(t)$ and $cos(t/2)$. Propose, from this remark, a parametric modeling of the function $r_X(t)$ and represent it on the cloud of points of the previous question (Hint: we can do a linear regression of **x** on $cos(t)$ and $cos(t/2)$). Comment on the result obtained.

1.3. Part C: The two brothers finally reunited. The two brothers failing to find the location of the treasure separately, they finally decide to put aside their rivalry to search for it together by pooling what they have been able to extract from the data in their possession.

1. Represent the cloud of points corresponding to the pairs (latitude, longitude). Comment on the result obtained.

2. Do the same with the data ($\mathbf{x\_e}, \mathbf{y\_e}$)obtained by estimating $r_X(t)$ and $r_Y(t)$ by each of the wires. Comment on the result obtained. Does it give the location of the treasure? What is the safe code?

BONUS: Add to the previous graph the one obtained from parametric modeling proposed in A.3 and B.3. Comment on the result obtained.

3. Explain what the function does find cross and use it to find the coordinates (latitude, longitude) of the treasure location.

BONUS: We can show, using questions A.3 and B.3, that the treasure was hidden at the moment $t = \pi$ Where $t = 3 \times \pi$. Deduce a new estimate of the location of the treasure using the models proposed in A.3 and B.3.

4. In order to figure out how to navigate to the location of the treasure, the two brothers pool their data to try to reconstruct the depth map based on longitude and latitude: $p = r_{XY}(x, y) + \epsilon$. Estimate non-parametrically r_{XY} on the area $[2, 4] \times [1, 1]$ with stitches of 0.03×0.01 (we will consider using the option **structure.2d='common'**).Comment on the result obtained.

5. Give an estimate of the depth to which the treasure is buried. Should the two brothers bring their spacesuits? Did you need to estimate r_{XY} for the knowledge?

2. [8 Points]

We are interested in a problem of vocal dictation in which we seek to find a phoneme

$$(\text{"}sh\text{"}, \text{"}iy\text{"}, \text{"}dcl\text{"}, \text{"}aa\text{"} \text{ and } \text{"}ao\text{"})$$

corresponding to the recording of a sound (described through a curve, called log-periodogram). We have a learning sample on which we have for each recording the corresponding phoneme (variable **PHONEME**) and the log-periodogram (variable **CURVES**). We wish to study the effectiveness of such an approach for automatically assigning the corresponding phoneme to a sound recording.

1. Describe the nature of the data. What type of statistical problem is considered here?

2. Load data (**PHONEMES.RData** file). What is the sample size? How many observations do you have for the different phonemes?

3. Build a training sample containing the first 1000 observations. The rest of the data will constitute the test sample.

4. Give the definition of the Bayes classifier. Can we use it here?

5. Let X be the variable representing the log-periodogram and Y_{SH} the variable which is 1 if the corresponding phoneme is *"sh"* and 0 otherwise. What is the conditional expectation of Y_{SH} knowing X? Define a Y_{SH} vector corresponding to the learning sample, whose components are 1 if the phoneme is "SH" and 0 otherwise. Do the same for the different phonemes.

6. Explain how a supervised classification method can be constructed from these data.

7. Use this method to determine, from the training sample, the phonemes of the sound recordings that are part of the test sample (using the function **funopare.kernel.cv** with options **semimetric='mplsr'**, **5**). Hint: we can estimate the conditional probabilities $p_{SH}, p_{IY}, p_{DCL}, p_{AA}, p_{AO}$ belonging to the different types of phonemes then find the most probable using the commands

$$\begin{aligned} Prob &= cbind(p_{SH}, p_{IY}, p_{DCL}, p_{AA}, p_{AO}) \\ PH &= c("SH", "IY", "DCL", "AA", "AO") \\ Class &= PH[apply(Prob, 1, order)[5,]] \end{aligned}$$

8. Compare the predicted phonemes with the real phonemes and give the classification error rates for each of the phonemes. Comment on the results obtained. Hint: we can use the confusion matrix.