Assigntment 1 - Densitiy Estimation Advanced Multivariate Analysis (AMA) Facultat d'Informàtica de Barcelona (FIB) Universitat Politècnica de Catalunya (UPC) - BarcelonaTech

Bandwidht choice by leave-one-out maximum likelihood

Joan Oliveras Torra, Odysseas Kyparissis, Louis Tichelman

2023-10-26

Contents

1 Histogram **1**
1.1 Exercise 1 . . . 1
1.2 Exercise 2 . . . 1
1.3 Exercise 3 . . . 2
1.4 Exercise 4 . . . 3
1.5 Exercise 5 . . . 3
1.6 Exercise 6 . . . 4
1.7 Exercise 7 . . . 5

2 Kernel Density Estimator **5**
2.1 Exercise 8 . . . 5

1 Histogram

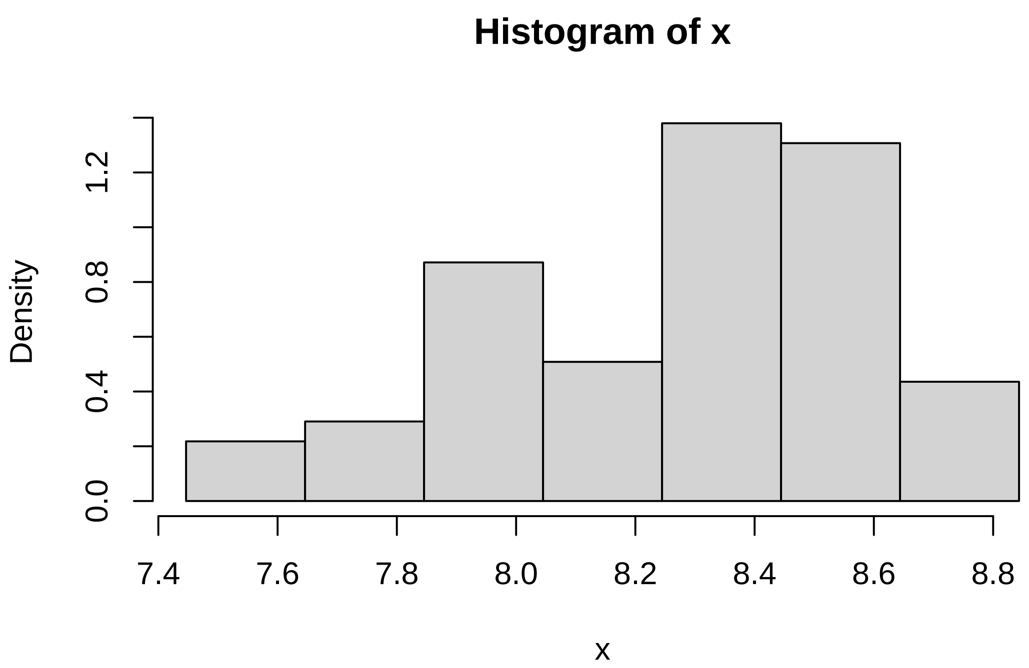
1.1 Exercise 1

A similar relationship to that of the `leave-one-out kernel density estimator`, $\hat{f}_{h,(-i)}(x_i)$, and the `kernel density estimator using all the observations`, $\hat{f}_h(x_i)$, when *both are evaluated at one of the observed data*, is holding between the `histogram estimator of the density function`,$\hat{f}_{hist}(x_i)$, and its `leave-one-out` version, $\hat{f}_{hist,(-i)}(x_i)$, when again *both are evaluated at a single observed point.* The relationship is presented here:

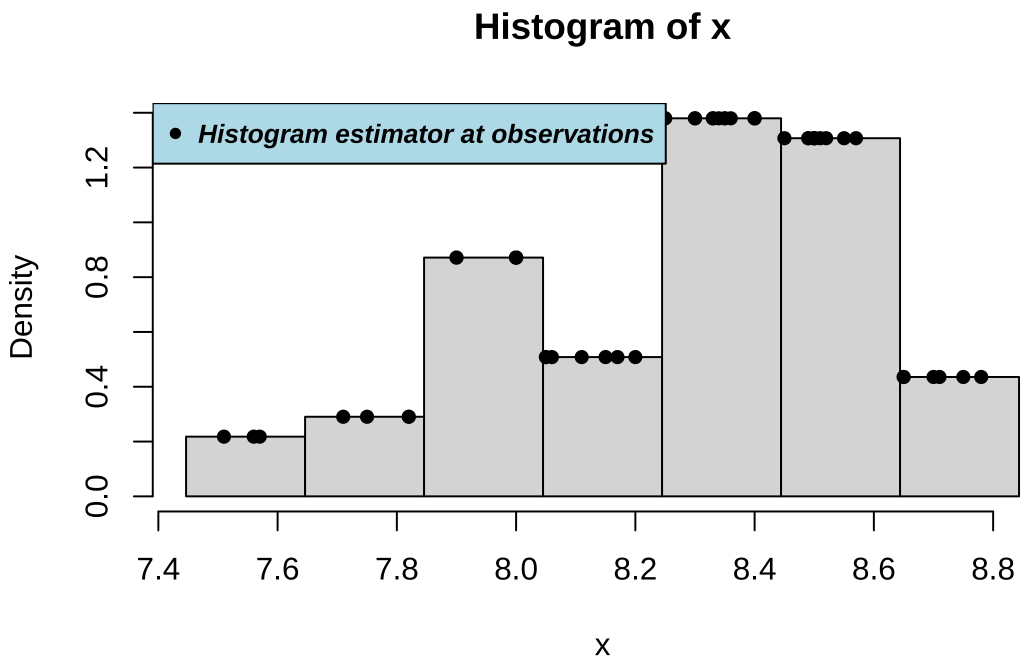
$$\begin{aligned}\hat{f}_{hist,(-i)}(x_i) &= \frac{1}{n-1}\sum_{k\neq i}\frac{1}{b}I_{B_{j(x_i)}}(x_k)\\ &= \frac{1}{(n-1)b}\Big(\Big(\sum_{k=1}^{n}I_{B_{j(x_i)}}(x_k)\Big)-1\Big) = \frac{n}{n-1}\Big(\hat{f}_{hist}(x_i)-\frac{1}{nb}\Big).\end{aligned}$$

1.2 Exercise 2

After reading the ***CDrate dataset***, we call `x` the first column, and the variables ***A***, ***Z*** and ***nbr*** are defined as mentioned in the statement. The histogram of variable ***x*** is presented in the following figure:

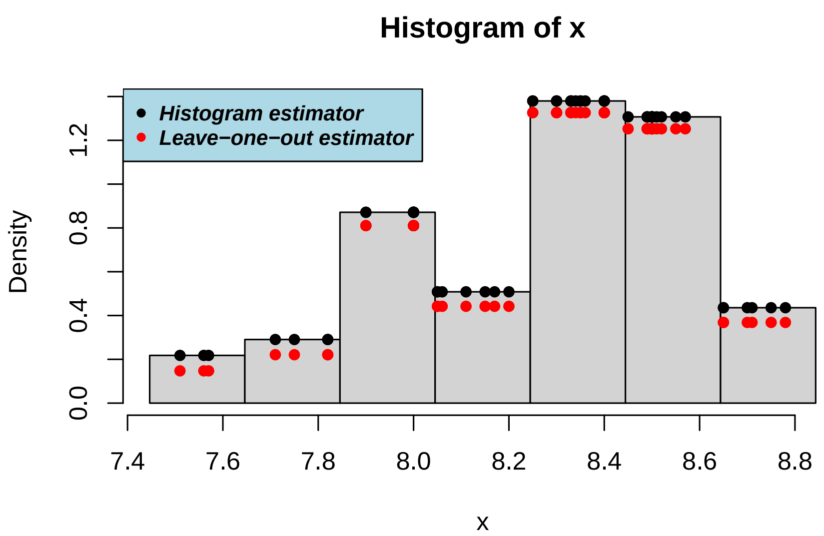


With the usage of the function ***stepfun*** the above histogram is being converted into a function ***hx_f*** that can be evaluated at any point of $\mathbb{R}$, or at a vector of real numbers. By using this function, the histogram generated above is being evaluated at the vector of observed data $\boldsymbol{x}$. Moreover, the histogram accompanied with all the points ($x_i, \hat{f}_{hist}(x_i)$) for $i = 1, ..., n$ are included in the following figure.



1.3 Exercise 3

By using the formula of *exercise 1* and the histogram estimator of *exercise 2*, $\hat{f}_{hist}(x_i)$, the values of $\hat{f}_{hist,(-i)}(x_i)$ are calculated for $i = 1, ..., n$. Moreover the new points are added to the previous plot. We name the function of $\hat{f}_{hist,(-i)}(x_i)$ (*hx_f2*) in the code.



1.4 Exercise 4

In this part of the report, the *leave-one-out log-likelihood function* corresponding to the previous histogram is being calculated with the following formula, and the *number of breaks* that has been used is equal to 7.

```
L7=prod(hx_f2)
looCV_log_lik=sum(log(hx_f2))
```

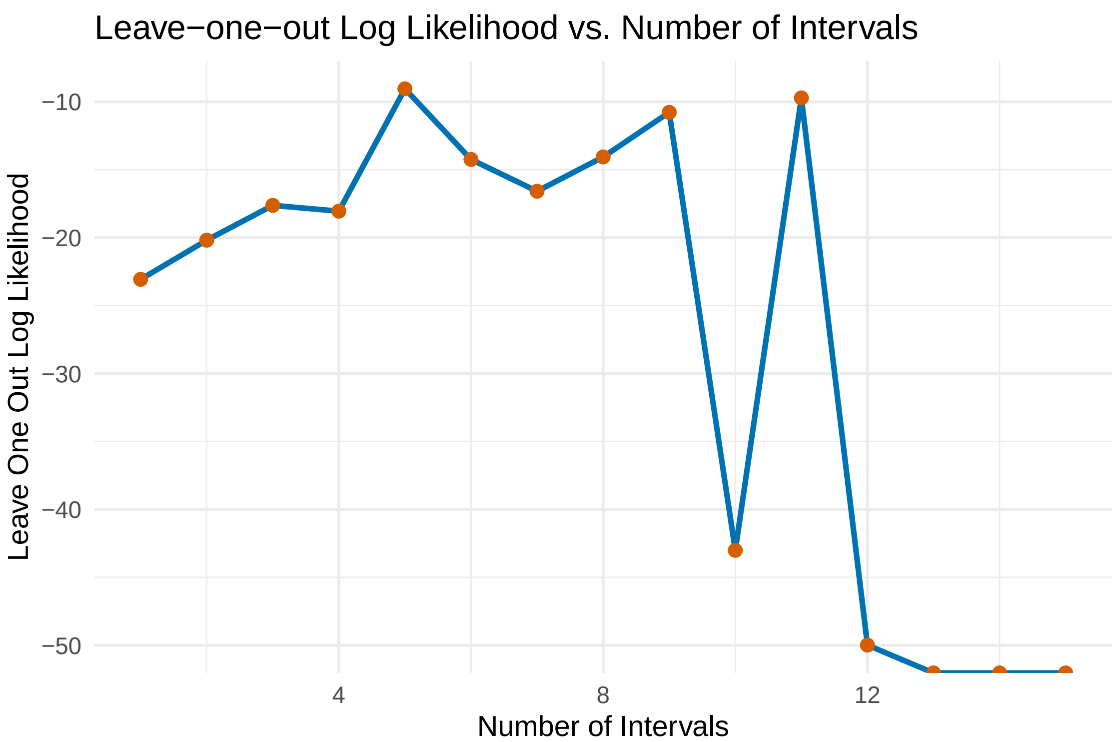
The result of the function is presented here:

```
## [1] "The leave-one-out log-likelihood for nbr=7 is:"
```

```
## [1] -16.58432
```

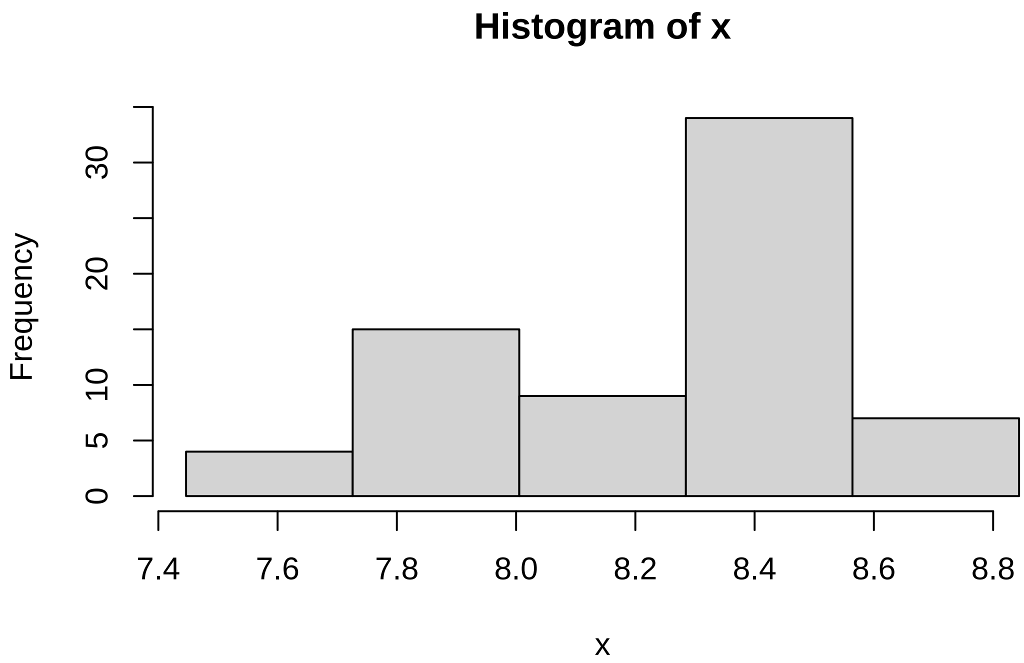
1.5 Exercise 5

In this exercise we are repeating the same process as before, but this time we are iterating through different values for the *number of intervals of the histogram* in the range: $[1, ..., 15]$. For each of them the *leave-one-out log-likelihood function* (*looCV_log_lik*) is computed, once the respected histogram has been firstly generated. Additionally, the values of *looCV_log_lik* are plotted against the values of *number of intervals* and the optimal value of the intervals is selected. For selecting the optimal value, we compare the results of the *looCV_log_lik* function, and the one with the maximum values is considered to be the best result. Finally, the histogram of x is plotted by using the optimal value of *number of intervals*. The mentioned plots are presented in the following figures.



```
## [1] "Number of Intervals with highest leave one out log likelihood: 5"
```

Below the histogram for ***number of intervals*** being equal to ***5*** is presented:

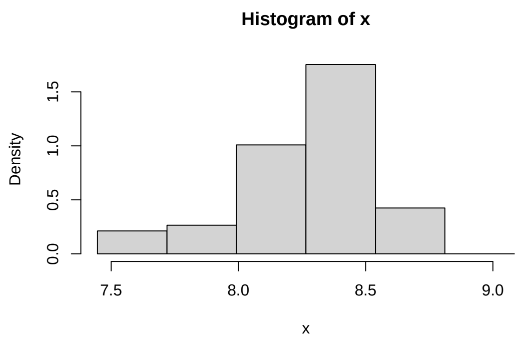


1.6 Exercise 6

In this exercise we let b to be the ***common width of the bins of a histogram***. By considering the set mentioned in the assignment's statement, as the possible values of b, the `leave-one-out log-likelihood function` is calculated for each one of them. Ultimately the selection of the best value for b that maximizes the `leave-one-out log-likelihood function` is taking place and the corresponding histogram is being plotted.

```
## [1] "Bin width (b) with highest leave one out log likelihood:"
```

```
## [1] 0.272977
```

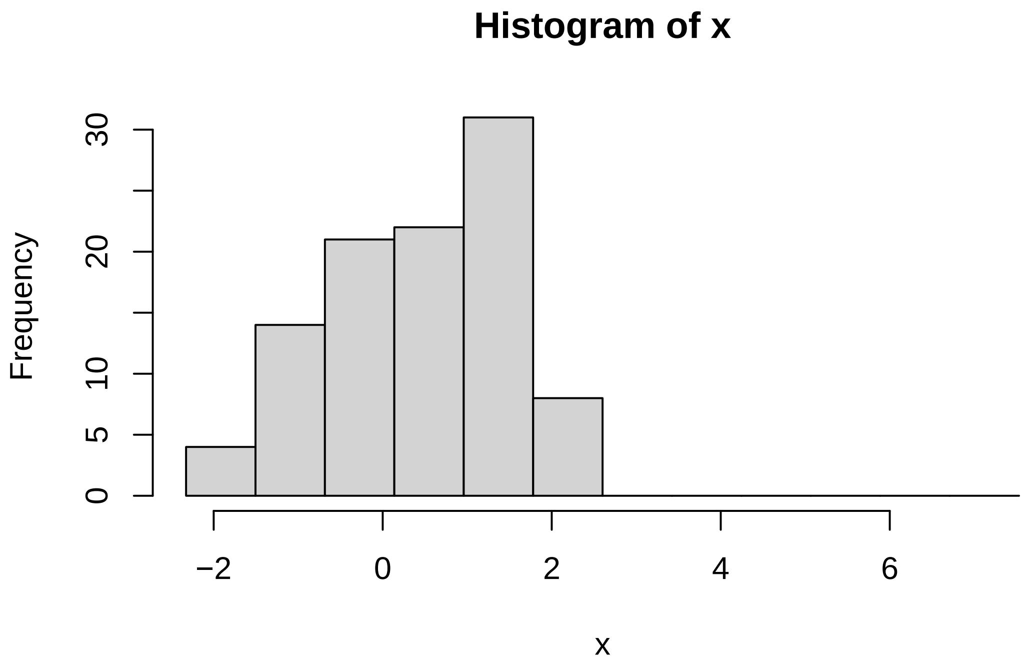


1.7 Exercise 7

By recycling the functions *graph.mixt* and *sim.mixt* defined at *density_estimation.Rmd* which generate a mixture of two Gaussians, we generate $n = 100$ data points from it. Let b be the *bin width* of a histogram estimator of $f(x)$ using the generated data. Below, we are selecting the value of b that maximizes the *leave-one-out log-likelihood function* and, finally, plotting once again the corresponding histogram. A comparison between the results obtained below and the results of *Scott's* formula is taking place.

```
## [1] "bin width with highest leave one out log likelihood (of the ones we conside
```

Below the histogram with that bandwidth



Scotts formula proposes the following bin width:

```
3.49*sd(x)*100^(-1/3)
```

```
## [1] 0.8253815
```

The values are (in this case) pretty similar.

2 Kernel Density Estimator

2.1 Exercise 8

We consider the vector x of data we have generated before from the mixture of two normals. We use the relationship (see exercise sheet) to select the value of h maximizing the leave-one-out log-likelihood function.

```
## [1] "bandwidth with highest leave one out log likelihood (of the ones we conside
```

Below we plot the corresponding kernel density estimator.