

Non-parametrics Statistics: Assignment

Javier Esteban Aragonese and Mauricio Marcos Fajgenbaun

Exercise 1

Part a)

Rubrics: Compute the average gray level, `av_gray_one`, for each image of the digit “1”.

We will work, in this exercise, with a 60.000 x 784 matrix. Each row describes the composition of a particular grayscale image (the images are of digits 0 to 9 and are composed by numbers between 0 and 1). First we are asked to compute the average gray level for each image of the digit “1”.

```
# Let's read the data first from the professor's github:
load(url("https://raw.githubusercontent.com/egarpor/handy/master/datasets/MNIST-tSNE.RData"))

# Create an empty vector and a loop for every one of the 60.000 images.
av_gray_one=c()
for(i in 1:60000){
  # If the label is equal to 1 calculate the mean and append it in the vector
  if(MNIST$labels[i]==1){
    media=mean(MNIST$x[i,])
    av_gray_one=append(av_gray_one,media)
  }
}
```

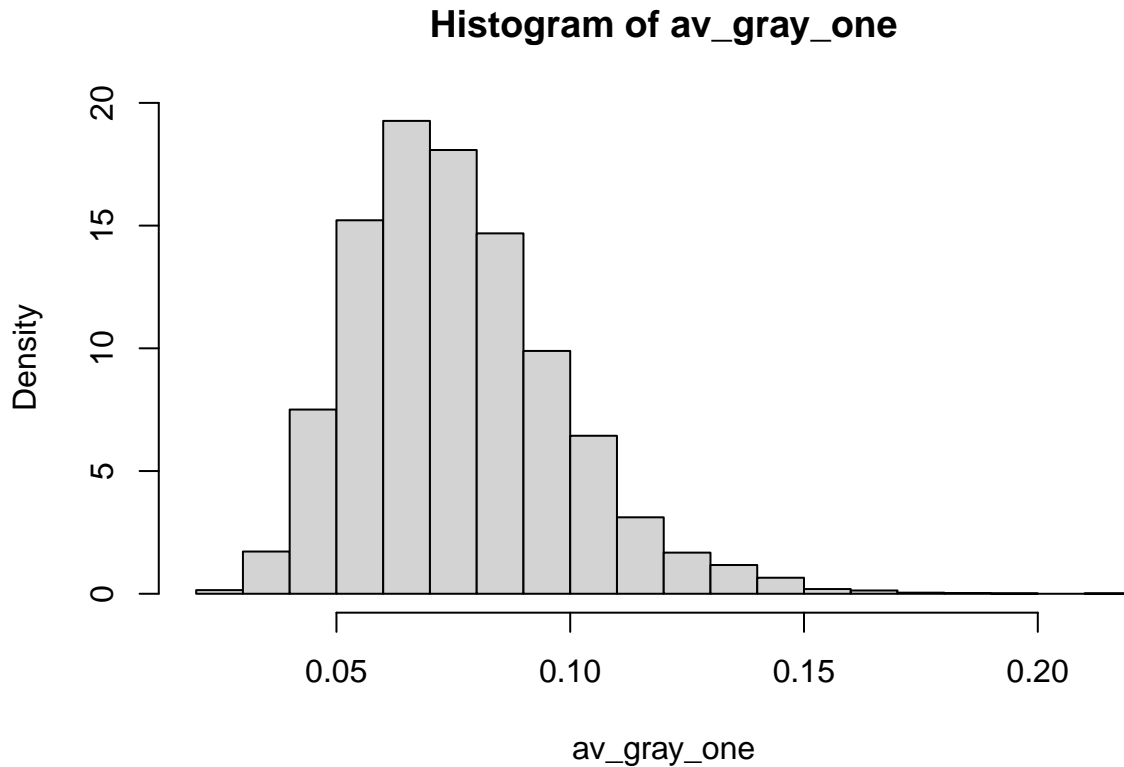
Part b)

Rubrics: Compute and plot the kde of `av_gray_one`, taking into account that it is a positive variable.

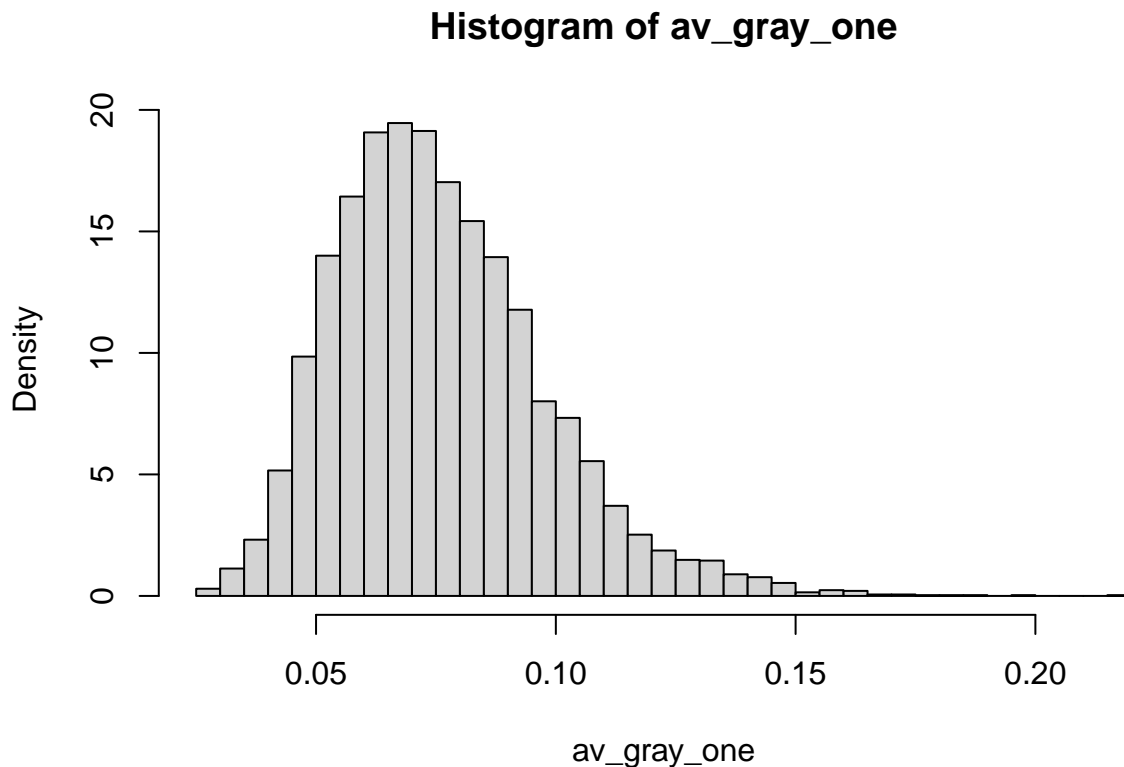
Now, in order to compute and plot the Kernel Density Estimator for the variable `av_gray_one` we first need to find a proper bandwidth. This choice is crucial, as the bandwidth is the one parameter that will make the biggest difference in our density estimation (and not so much the choice of Kernel). In this sense, choosing a big bandwidth (too large) will produce oversmoothing and will tend to miss the features or do not capture totally these. This is, we would have to bear a larger bias. On the contrary, if we choose a bandwidth that is too small for our density estimation, we will finish up interpolating the data (and this is obviously not our intention).

We think that a good first approach to estimate the Kernel Density is to look at the histogram of the variable. This way, we can have a first idea of how this density is supposed to behave in the different parts of the support.

```
hist(av_gray_one, probability=TRUE)
```



```
hist(av_gray_one, probability=TRUE, breaks=35)
```

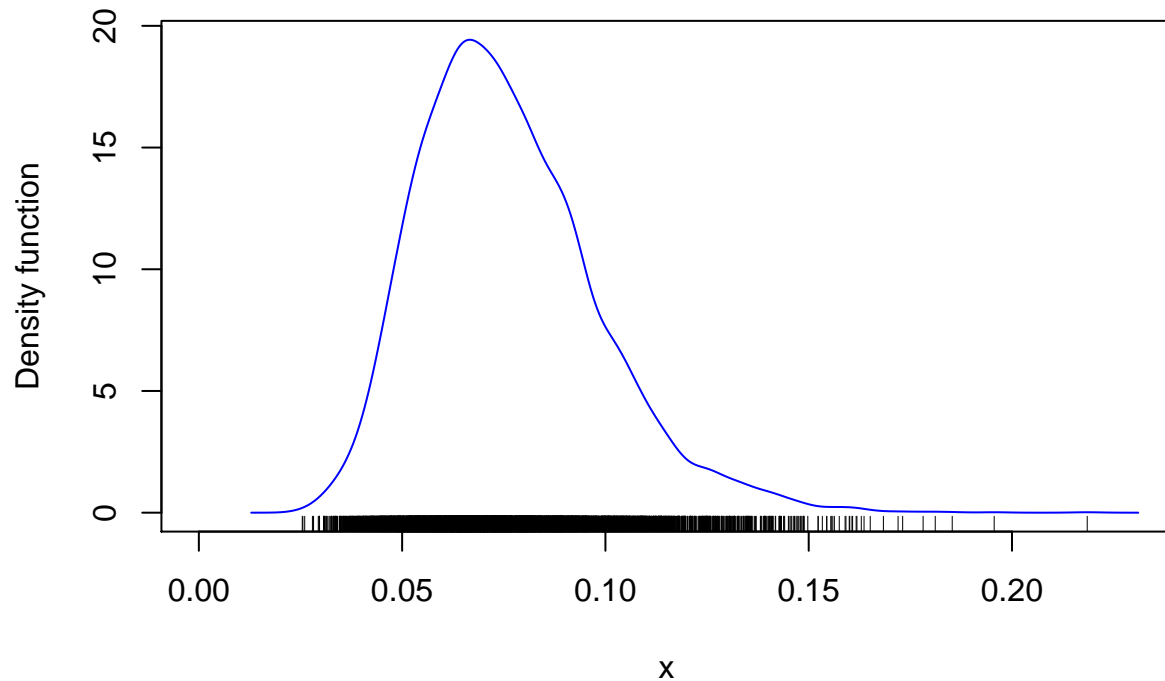


We know that this is a very rudimentary approach to the density estimation, and for that reason we will not pay a lot of attention at selecting the best bandwidth and the best starting point for this histogram. Even when

we change the number of bins the distribution does not change a lot, maybe the skew gets amplified when choosing a larger number of bins. But we do can take some conclusions from this first glance: as we can see, the distribution seems to behave quite symmetrical, maybe with a small skew (longer tail on the right part of the distribution). In this case, we can say that the DPI may be a good bandwidth selector, as it does a relatively good job when estimating distributions that are close to a normal.

Then, we compute the bandwidth that is less computationally expensive: Direct Plug In. Also, this bandwidth selector has a convergence rate much faster than the cross-validation selectors.

```
## [1] 0.003392921
```

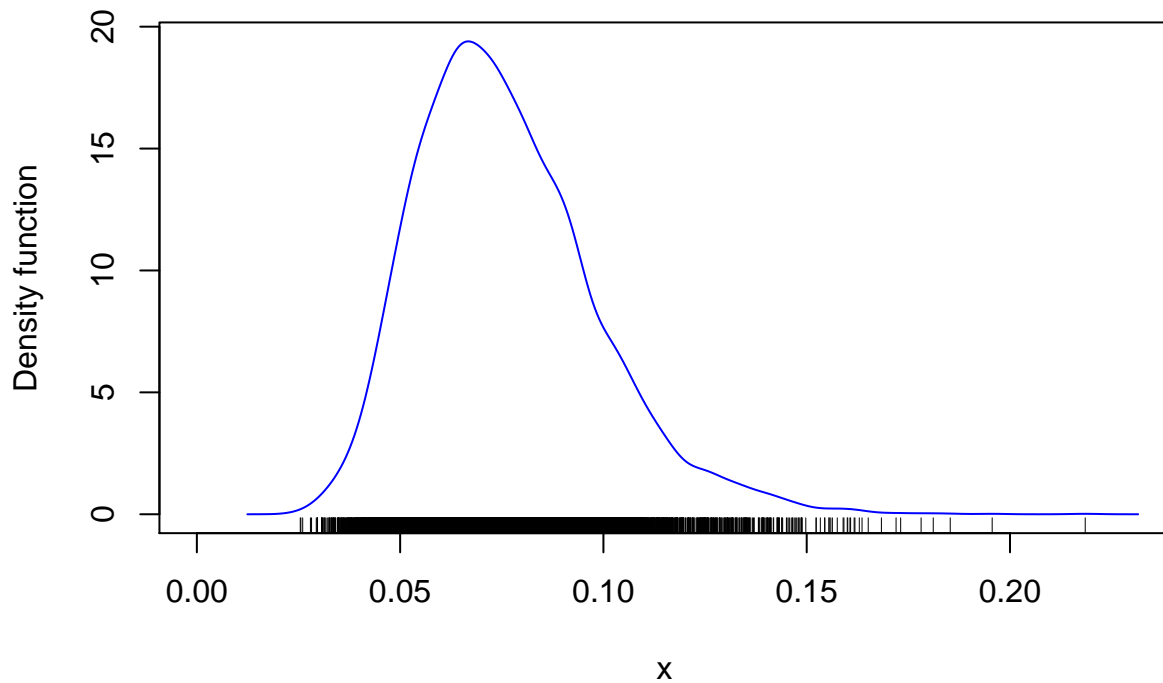


As we know, sometimes the DPI selector tends to have issues when estimating a density that is highly non normal or when facing rough densities. As mentioned above, this is not the case, and we may think this is the right selector. Anyway, let's try and see what happens with cross-validation selectors: Biased Cross Validation and Least Squares Cross Validation.

```
# Computation of Biased Cross Validation Selector
bw2=bw.bcv(x = av_gray_one)
print(bw2)
```

```
## [1] 0.003518536
```

```
plot(ks::kde(x = av_gray_one, h = bw2),xlim=c(0,0.23),col="blue")
rug(av_gray_one)
```

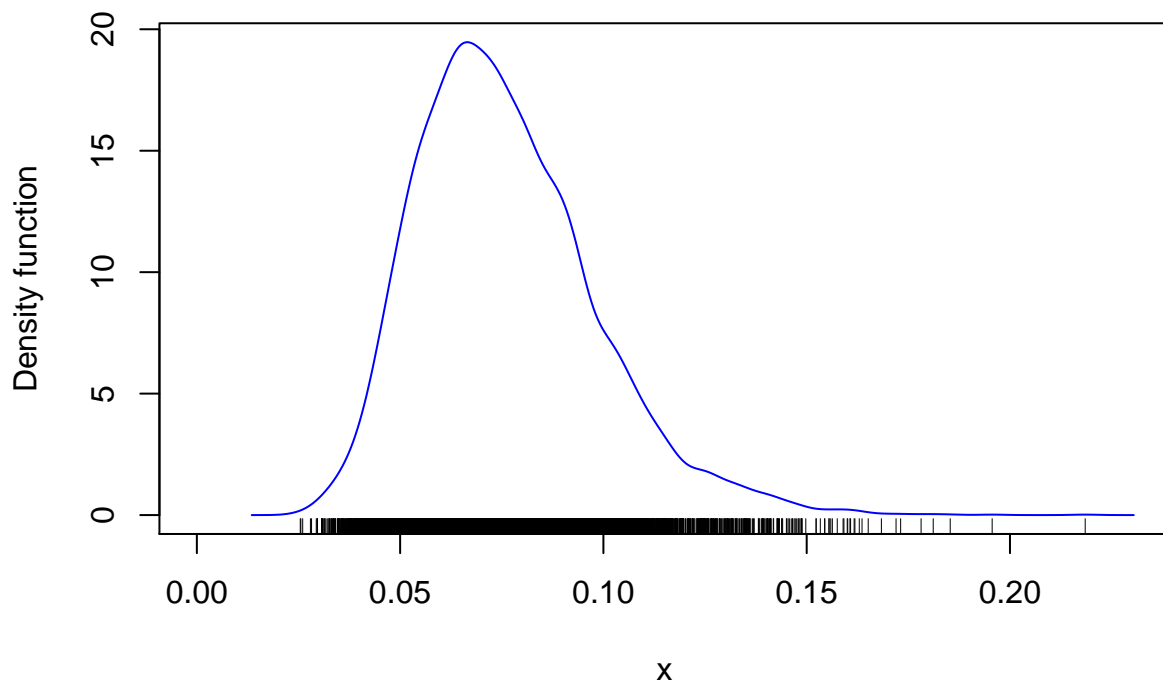


As we can see, the bandwidth that we get when doing BCV is a bit larger than the one obtained with the DPI. Let's see what happens when applying LSCV:

```
# Computation of Biased Cross Validation Selector
bw3=bw.ucv(x = av_gray_one)
print(bw3)
```

```
## [1] 0.003229507
```

```
plot(ks::kde(x = av_gray_one, h = bw3),xlim=c(0,0.23),col="blue")
# Plot the rug to check the amount of points in each part of the support
rug(av_gray_one)
```



This

time we got a smaller bandwidth between the three that we calculated. Nevertheless, we can say that there is not a huge difference between the three bandwidths, and that it should not make a big difference in the case of this distribution. We feel that the best choice we can make is the DPI, for two reasons. First, it seems that taking the bandwidth “in the middle” of the three that we calculated (given that there is not a big difference between them) may be the most “conservative” idea. But the main reason we chose this bandwidth is because our distribution behaves pretty symmetrical and is not very far away from a normal distribution. So this selector should do the job just right.

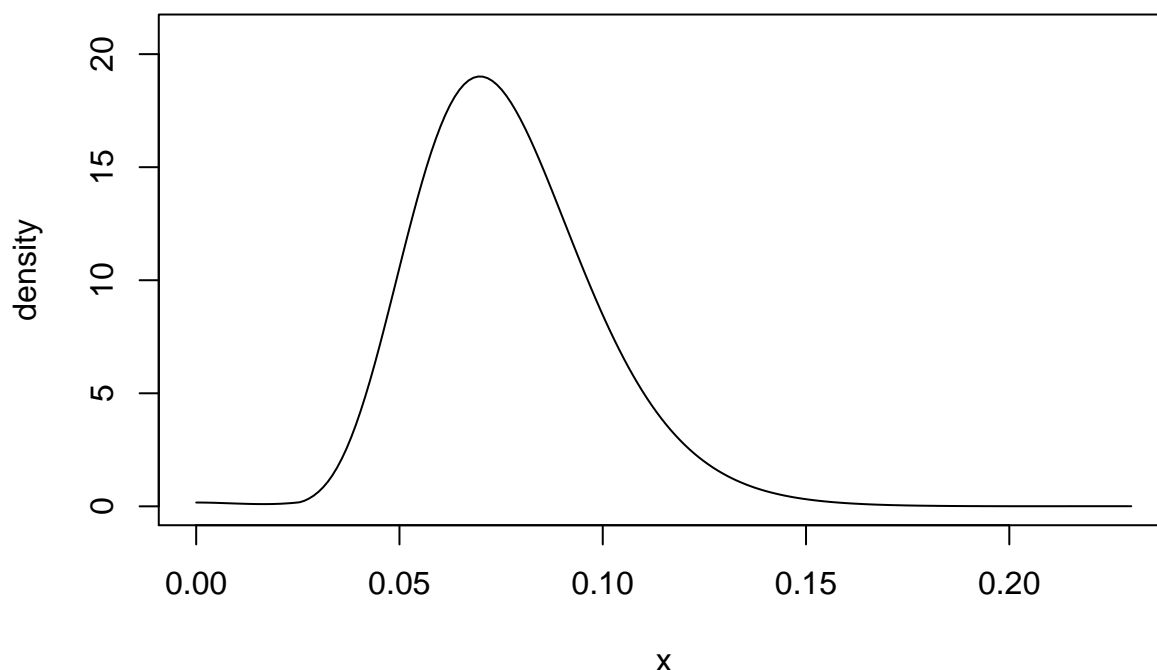
Following the spirit of the previous decision, it seems logical (and right) to continue using a Normal Kernel. Nevertheless, this choice should not be crucial, given our distribution.

Nevertheless, we have to pay attention to the fact that our variable is only positive, as it never takes negative values. This may be of importance, as when we do not declare explicitly that there is no density before zero (given that the kernels are functions defined in the R space) we may encounter some trouble in the boundaries and the estimator will do a bad job there (we may see a big bias close to zero). At the same time, our variable is an average, meaning that we have two boundaries: $[0,1]$. In this sense, we must apply a probit transformation. To do this, we will make use of a package called `kde1d`. This package, by default, calculates a new optimal bandwidth for the transformed data (that may not be optimal for the untransformed data) and when setting two boundaries in the algorithm, the function will perform a probit transformation. In case we would want to do a log-transformation, we would only need to determine one boundary, but this is not the case.

Now, we will estimate and plot two kde, one with a theoretical bound (the upper limit is 1) and another one with a bound equal to our maximum “x” in the sample. This is to get a better grasp on the density estimation plot.

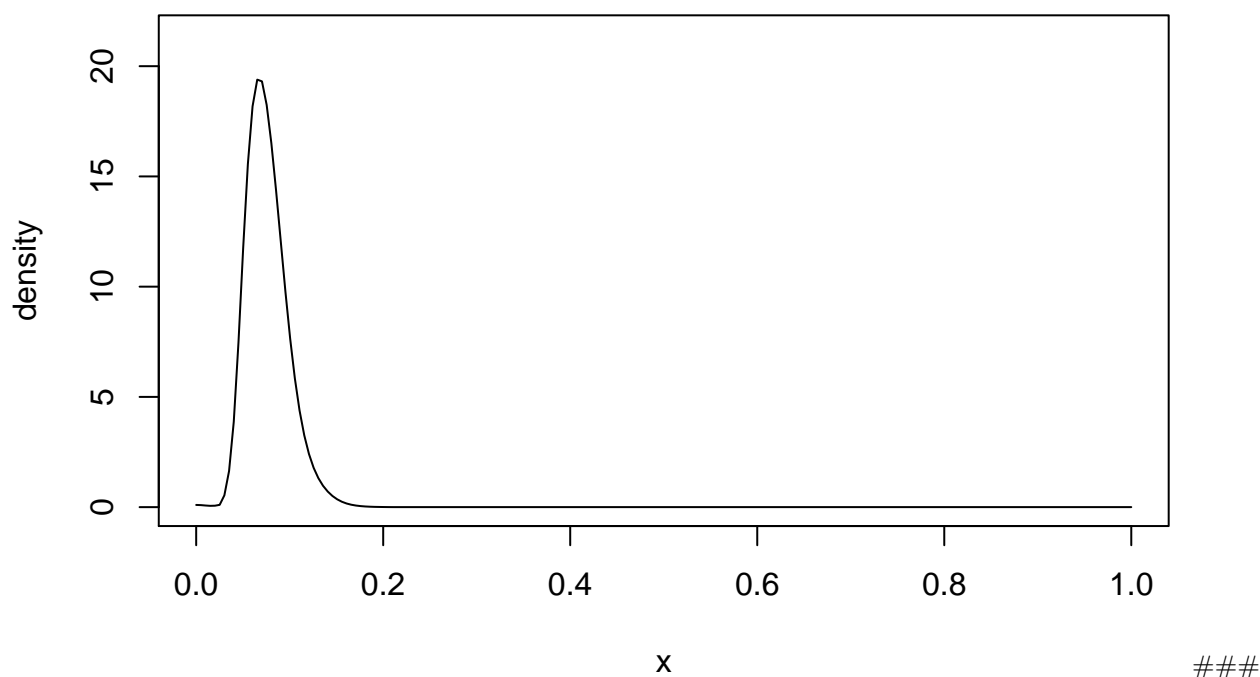
```
# First case, with upper bound equal to the maximum of our sample
probit_1 = kde1d(av_gray_one,xmin=0,xmax=0.23)
summary(probit_1)
```

```
## kernel density estimate ('kde1d'), log-quadratic with bounded support (xmin = 0, xmax = 0.23)
## -----
## nobs = 6742, bw = 0.41, loglik = 16491.97, d.f. = 3.25
plot(probit_1)
```



```
# Second case, with upper bound equal to theoretical bound
probit_2 = kde1d(av_gray_one,xmin=0,xmax=1)
summary(probit_2)

## kernel density estimate ('kde1d'), log-quadratic with bounded support (xmin = 0, xmax = 1)
## -----
## nobs = 6742, bw = 0.09, loglik = 16515.37, d.f. = 7.01
plot(probit_2)
```

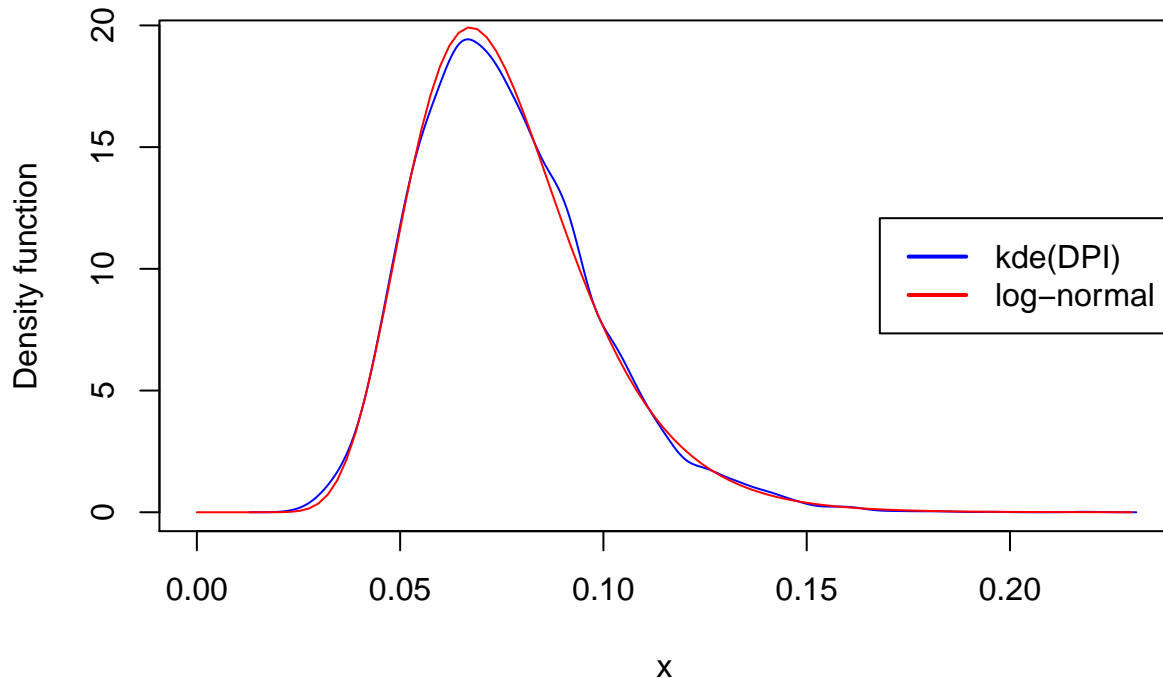


Part c)

Rubric: *Overlay the lognormal distribution density, with parameters estimated by maximum likelihood (use MASS::fitdistr)*

Now it is time to overlay, in the same plot another distribution density: the lognormal. Like any other lognormal, we need to first estimate the two fundamental parameters: mean and standard deviation. In order to achieve this, we will use the package MASS, that will allow us to estimate these parameters through maximum likelihood, and fit the data into this new distribution density so that we can compare.

```
# We fit the vector av_gray_one in a log normal distribution.
fit=fitdistr(av_gray_one,"log-normal")
est=fit$estimate
# We compute the mean and standard deviation, to be able to plot it.
mean=est[1]
sd=est[2]
#We plot the old kde, with bw (obtained through DPI)
plot(ks::kde(x = av_gray_one, h = bw),xlim=c(0,0.23),col="blue")
# We also plot the lognormal distribution
curve(dlnorm(x, meanlog = mean, sdlog = sd),add=TRUE,col="red"
)
legend("right",col=c("blue","red"), legend = c("kde(DPI)","log-normal"),
lwd = 2)
```



As we can see, there is a slight difference in the densities, specially in the peak and slightly in the boundaries.

Part d)

Rubric: Repeat c for the Weibull density.

Now we have to do the same as previously, although with the Weibull density. We can do it again with the same package.

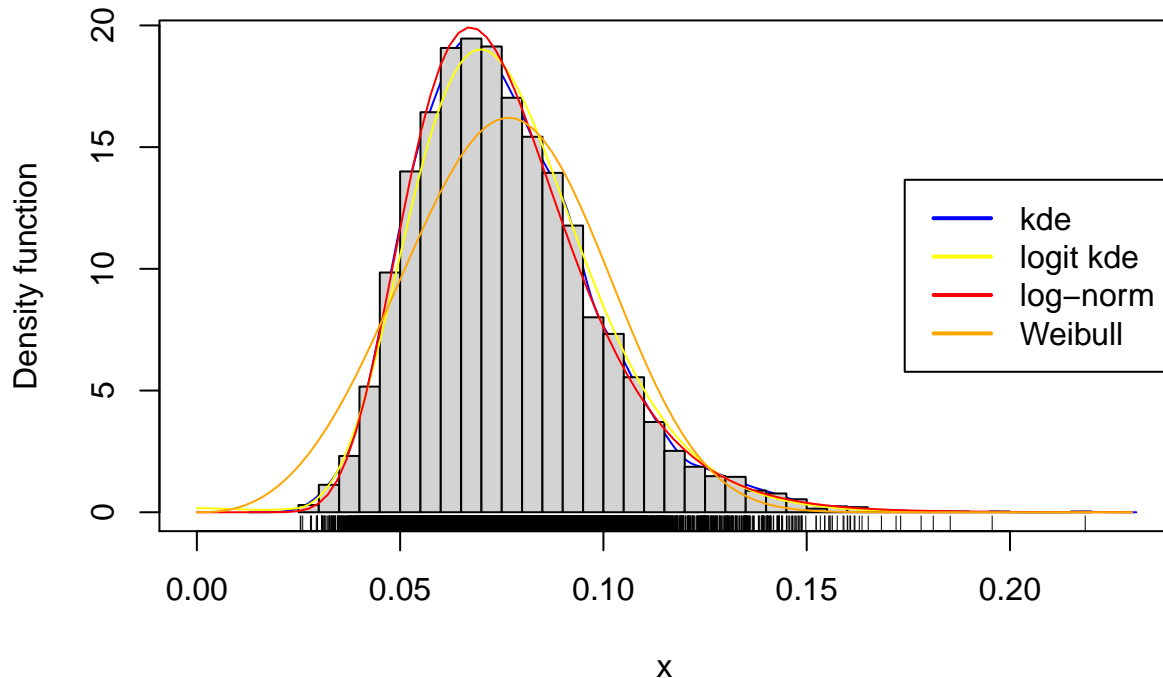
```
# Fit our vector of av_gray_one in a Weibull density.
fit2=fitdistr(av_gray_one, "weibull")
est2=fit2$estimate
# Call the two parameters of this distribution: shape and scale.
shape=est2[1]
scale=est2[2]
```

Part e)

Rubrics: Which parametric model seems more adequate?

Plot the different density estimation:

```
plot(ks::kde(x = av_gray_one, h = bw),xlim=c(0,0.23),col="blue")
hist(av_gray_one,probability = TRUE, add=TRUE,breaks=30)
logit_1 = kde1d(av_gray_one,xmin=0,xmax=0.23)
lines(logit_1,col="yellow")
curve(dlnorm(x, meanlog = mean, sdlog = sd),add=TRUE,col="red")
)
curve(dweibull(x, shape=shape, scale = scale, log = FALSE),add=TRUE,col="orange")
)
rug(av_gray_one)
legend("right",col=c("blue","yellow","red","orange"), legend = c("kde", "logit kde", "log-norm","Weibull"),
lwd = 2)
```



When having a close look at the plot, we can see that the weibull distribution does not fit very well to the data, specially when trying to estimate the picks and the tails (it overestimates the tails and underestimate the pick). Meanwhile, when comparing the other 3 possibilities: the probit transformed kde, the log-normal kde and the non-parametric (and not transformed) kde, we can say that both the log-normal and the nonparametric do a pretty good job, although the log-normal distribution density overestimates a bit the pick and underestimates some parts of the distribution (on the right side).

We can say that maybe the data does not behave so normal, specially because of the skewness. This is why we will finally choose our **not transformed kde**, as the nonparametric option seems a bit better than the parametric ones and in comparison with the transformed one, it seems to capture better the pick. We can also note that the bandwidth we chose for the kernel density estimation is quite small, very close to zero. At the same time, the vector of averages that correspond to the digit “1” is composed of 6742 elements. As we consider that our sample is quite big and that the bandwidth is quite small, we understand this has implications related to the fact that the kde does not have a problem with bias nor the variance really. As $1/n = 1/6742 = 0.00014$ is smaller than the bandwidth (approximately 0.00339) which is equivalent to saying that the bandwidth approximates zero, but at a rate slower than n^{-1} , we can say that asymptotically we have very good properties: bandwidth tends to zero but not so fast, as nh tends to infinity.

Exercise 2

First, we will compute our first own implementation of the Nadaraya-Watson estimator.

Option 1

```
# We create a function with 4 inputs: x (evaluation points), X(vector of size n with predictors), Y (vector of size n with responses)
nadarayaCompleto=function(x,X,Y,h){
  # We create a matrix "kx" of zeros and number of columns as predictors and number of rows as evaluation points
  kx=matrix(0, ncol=length(X), nrow=length(x))
  # We compute the distance from each point to the evaluation points. Please note that we are calculating the squared distance
  for(i in 1:length(X)){
    resta=x-X[i]
    kx[,i]=dnorm((x - X[i]) / h)/h
  }
}
```



```

    }
    # We calculate the weights and they will depend on the evaluation points.
    W <- kx / rowSums( kx)
    drop(W %*% Y)
}

```

Here, we compute the Nadaraya Watson Estimator, proposed by the professor.

```

nw <- function(x, X, Y, h, K = dnorm) {
  Kx <- rbind(sapply(X, function(Xi) K((x - Xi) / h) / h))
  W <- Kx / rowSums(Kx)
  drop(W %*% Y)
}

```

Code taken from the class bookdown

We used the same function as we used in class to compare the results.

```

set.seed(12345)
n <- 10000
eps <- rnorm(n, sd = 2)
m <- function(x) x^2 * cos(x)
X <- rnorm(n, sd = 2)
Y <- m(X) + eps
x_grid <- seq(-10, 10, l = 500)
# Bandwidth
h <- 0.5

```

Code taken from the class bookdown

Now we plot, both the NW estimator proposed by the professor and the one we propose with a twist in the code to see if there is any difference (we should not expect any difference at all).

```

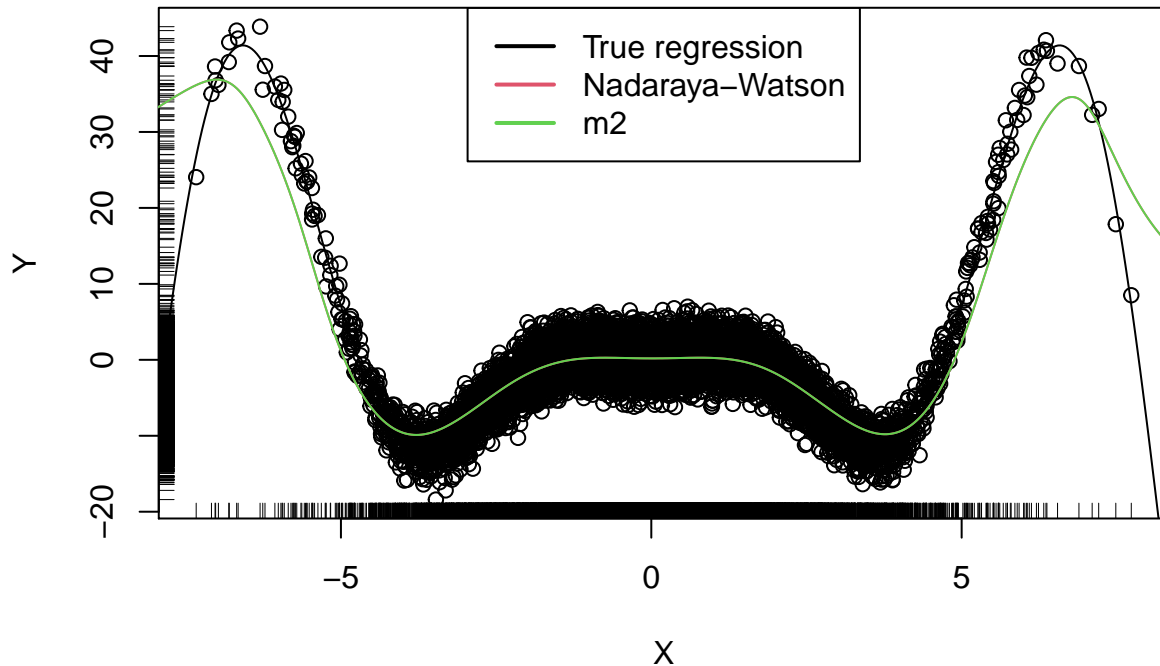
# Plot data
plot(X, Y)
rug(X, side = 1); rug(Y, side = 2)
lines(x_grid, m(x_grid), col = 1)

m1= nw(x = x_grid, X = X, Y = Y, h = h)
#lines(x_grid, m2, col = 3)
lines(x_grid, nw(x = x_grid, X = X, Y = Y, h = h), col = 2)

m2 = nadarayaCompleto(x = x_grid, X = X, Y = Y, h = h)
lines(x_grid, m2, col = 3)

legend("top", legend = c("True regression", "Nadaraya-Watson", "m2"),
      lwd = 2, col = 1:3)

```



As we can see, both the professor and our algorithm work exactly the same: they give us the same regression line as we can not distinguish between them in the plot.

Let's now check the speed of these algorithms, to see if there is any difference. Here we can expect a difference for the simplification in the code.

```
microbenchmark(nw(x = x_grid, X = X, Y = Y, h = h))
```

```
## Unit: milliseconds
##               expr      min       lq      mean  median      uq
##  nw(x = x_grid, X = X, Y = Y, h = h) 419.9829 472.9813 521.9646 503.579 559.175
##      max neval
##  918.375   100
```

```
microbenchmark(nadarayaCompleto(x = x_grid, X = X, Y = Y, h = h))
```

```
## Unit: milliseconds
##               expr      min       lq      mean
##  nadarayaCompleto(x = x_grid, X = X, Y = Y, h = h) 354.4246 421.8664 492.7755
##      median      uq      max neval
##  455.0683 532.4348 870.4986   100
```

We can see that how without changing the maths behind the estimator, our model is a bit faster. The only difference was that we changed the code to a loop, instead of a “sapply” function. As what we are intending in this exercise is to reduce the accuracy to improve efficiency (we want to compute a less precise version than the mathematical definition of the estimator) we will compare the new versions of the estimator to our proposed version (“NadarayaCompleto”) to assess real efficiency gains (and not simple programming tricks).

Option 2

Now, our objective is to reduce gain efficiency by reducing the accuracy of our algorithm. The idea now will be to create a regressor estimator that will only take into account some points of the data and not all of them to calculate the kernels. We only want to take into account the closest points to the point we are trying to estimate and forget about the rest. This way we think we can calculate less kernels and gain efficiency. To do this we create a threshold: when the distance between the point and evaluation points is larger than the threshold, we will assign a zero weight and will not calculate that kernel.

```

# Create a function with the same parameters, plus a threshold.
nadaraya=function(x,X,y,Y,h,thres=1){
  kx=matrix(0, ncol=length(X),nrow=length(x))
  for(i in 1:length(X)){
    resta=x-X[i]
    # Matrix kx will be of zeros and a normal Kernel will be computed only when the distance is inside the
    kx[,i]=ifelse(resta<thres,dnorm((x - X[i]) / h)/h,0)
  }
  W <- kx / rowSums( kx)
  drop(W %*% Y)
}

```

Let's see what happens first in the plot with the regression.

```

# Plot data
plot(X, Y)
rug(X, side = 1); rug(Y, side = 2)
lines(x_grid, m(x_grid), col = 1)

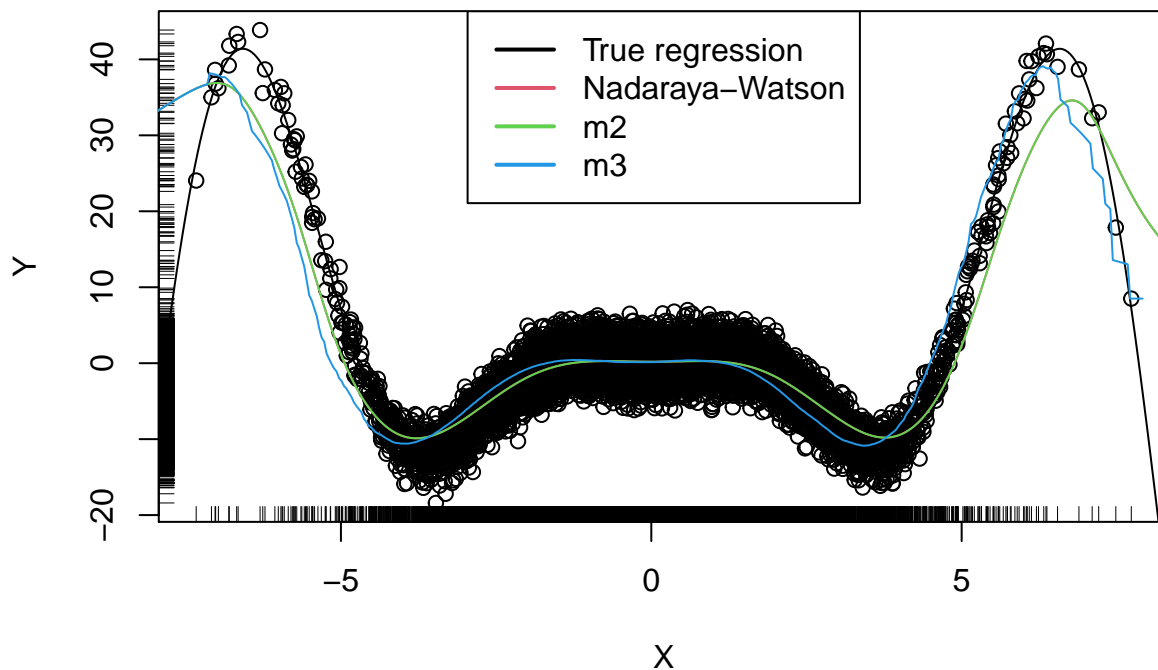
#lines(x_grid, m2, col = 3)
lines(x_grid, nw(x = x_grid, X = X, Y = Y, h = h), col = 2)

m2 = nadarayaCompleto(x = x_grid, X = X, Y = Y, h = h)
lines(x_grid, m2, col = 3)

# We define thresh = 0.2 meaning that if the distance is larger than 0.2 we will not compute the Normal
m3 = nadaraya(x = x_grid, X = X, Y = Y, h = h,thres = 0.2)
lines(x_grid, m3, col = 4)

legend("top", legend = c("True regression", "Nadaraya-Watson", "m2", "m3"),
      lwd = 2, col = 1:4)

```



Let's check the efficiency gain.

```
microbenchmark(nadarayaCompleto(x = x_grid, X = X, Y = Y, h = h))

## Unit: milliseconds
##              expr      min       lq      mean
##  nadarayaCompleto(x = x_grid, X = X, Y = Y, h = h) 383.5492 423.3834 454.6781
##      median        uq      max neval
##  442.4019 464.9745 800.6249   100

microbenchmark(nadaraya(x = x_grid, X = X, Y = Y, h = h, thres = 0.2))

## Unit: milliseconds
##              expr      min       lq
##  nadaraya(x = x_grid, X = X, Y = Y, h = h, thres = 0.2) 740.4385 799.9435
##      mean  median        uq      max neval
##  865.2355 854.4229 905.6829 1173.495   100
```

As we can see, this is a much slower algorithm. We may have gained some accuracy (will check it later), but it took almost twice the time to run. We need another approach.

Option 3

Now, we will try to find another way to reduce the time it takes to run the code for the NW estimator. This time, we will propose another technique: sampling. The idea now will be to take a random sample from the original dataset (this is a new input for the function) and build the estimator only based on this sample dataset.

```
# Write the function with new input "Muestra"
nadarayaSample=function(x,X,y,Y,h,Muestra=10){
# kx will be a matrix of zeros with number of columns equal to the sample size and number of rows equal to the number of samples
  kx=matrix(0, ncol=Muestra,nrow=length(x))
# We define a vector of indexes, to chose x-es and y-es. These indices will have values between 1 and length(X)
  n=sample(1:length(X),Muestra)
# We create a new vector that will be composed of the sample predictors.
  x2=X[n]
# We build a for loop, now for the sample (same as before but now only for the sample).
  for(i in 1:Muestra){
    resta=x-x2[i]
    kx[,i]=dnorm(resta/ h)/h
  }
  W <- kx / rowSums( kx)
  drop(W %*% Y[n])
}
```

Let's check the plots:

```
# Plot data
plot(X, Y)
rug(X, side = 1); rug(Y, side = 2)
lines(x_grid, m(x_grid), col = 1)

#lines(x_grid, m2, col = 3)
lines(x_grid, nw(x = x_grid, X = X, Y = Y, h = h), col = 2)

m2 = nadarayaCompleto(x = x_grid, X = X, Y = Y, h = h)
lines(x_grid, m2, col = 3)
```

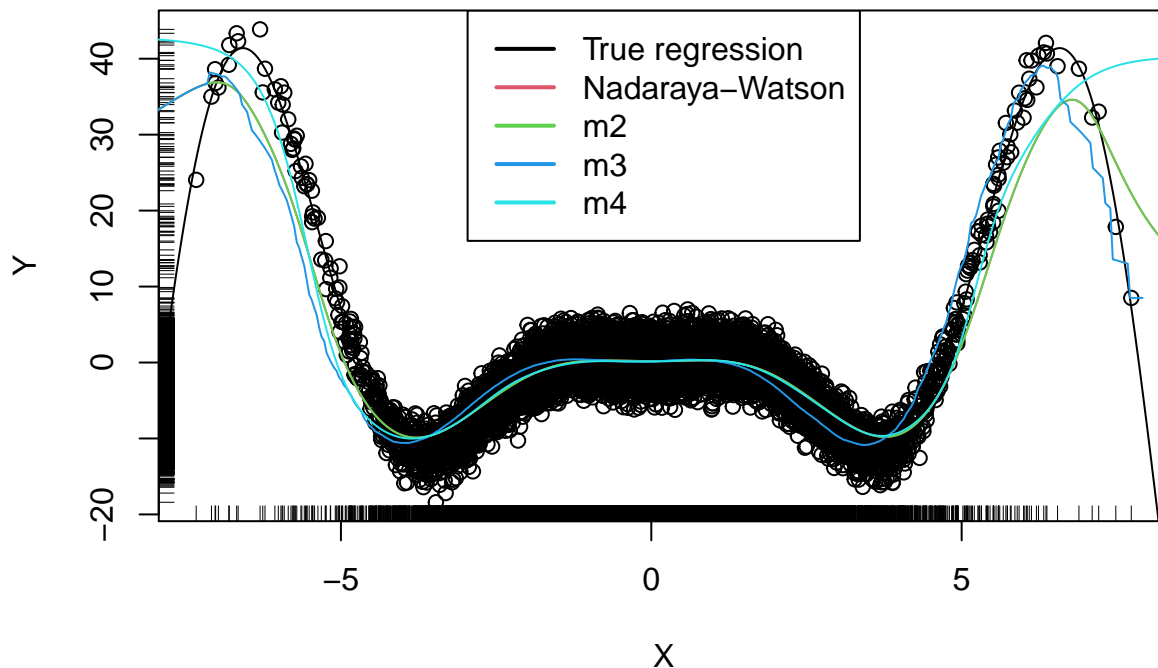
```

# We define thresh = 0.2 meaning that if the distance is larger than 0.2 we will not compute the Normal
m3 = nadaraya(x = x_grid, X = X, Y = Y, h = h, thresh = 0.2)
lines(x_grid, m3, col = 4)

# We choose a sample size equal to 1000 (10% of the dataset)
m4 = nadarayaSample(x = x_grid, X = X, Y = Y, h = h, Muestra=1000)
lines(x_grid, m4, col = 5)

legend("top", legend = c("True regression", "Nadaraya-Watson", "m2", "m3", "m4"),
      lwd = 2, col = 1:5)

```



As we can see in the plot, the new estimator works fairly alright. It has a problem in the boundaries, as a consequence of the lack of points there (lack of information to do a regression).

```
microbenchmark(nadarayaSample(x = x_grid, X = X, Y = Y, h = h, Muestra=1000))
```

```

## Unit: milliseconds
##                                     expr      min
## nadarayaSample(x = x_grid, X = X, Y = Y, h = h, Muestra = 1000) 32.36515
##      lq      mean      median      uq      max neval
## 36.78515 48.62704 40.10617 46.05779 168.9308   100

```

As we can see, we reduced the time of the algorithm more than ten times, in comparison to the “Nadaraya-Completo” that was demanding more than 450 milliseconds. This is a great achievement.

We can see that the fact that we take a random sample gives certain uncertainty to our estimator. This is because we never know what points is the algorithm going to take as a sample and then, we never know in what parts of the density our estimator will fail to estimate with more or less accuracy.

Option 4

Here we will implement a pseudo-cross validation technique to reduce the previous declared uncertainty. We inspired ourselves from both cross-validation and bootstrapping techniques. Now we will take 5 different (the number 5 can be changed, as is a new input in our function and can also be tuned) samples of the same length (“Muestra” is again another parameter to choose in the function) and we compute from each sample

the residual sum of squared between the true regression and this new regression (“nadarayaSampleCV”). The samples are built with replacemtn. Finally, we will end up using the sample that has the lowest Residual Sum of Squares (“RSS”) to perform the regression.

```
# Create a new function with a new parameter "cv".
nadarayaSampleCV=function(x,X,y,Y,h,Muestra=10,cv=5,m){
# I define the RSS as infinite (so that it will work with the first sample)
  RSS=Inf
# We build a for loop for the different "cv" samples
  for(i in 1:cv){
# kx will be the same matrix as before described in other chunks
    kx=matrix(0, ncol=Muestra,nrow=length(x))
    n=sample(1:length(X),Muestra)
    x2=X[n]
    for(i in 1:Muestra){
      resta=x-x2[i]
      kx[,i]=dnorm(resta/ h)/h
    }
    W <- kx / rowSums( kx)

# We store in a new variable the value of the weighted average for the actual sample
    mP=W %*% Y[n]
# We calculate the Residual Sum of Squares for the actual sample in comparison to the "NadarayaCompleto"
    RSSp=sum((m2-mP)^2)
# If the actual sample has a smaller RSS than the previous one in the loop, we keep it.
# Else, we keep with the other one.
    if(RSS> RSSp){
      RSS=RSSp
      mhat=mP
    }
  }
  return (mhat)
}
```

Let's plot now.

```
# Plot data
plot(X, Y)
rug(X, side = 1); rug(Y, side = 2)
lines(x_grid, m(x_grid), col = 1)

#lines(x_grid, m2, col = 3)
lines(x_grid, nw(x = x_grid, X = X, Y = Y, h = h), col = 2)

m2 = nadarayaCompleto(x = x_grid, X = X, Y = Y, h = h)
lines(x_grid, m2, col = 3)

# We define thresh = 0.2 meaning that if the distance is larger than 0.2 we will not compute the Normal
m3 = nadaraya(x = x_grid, X = X, Y = Y, h = h,thres = 0.2)
lines(x_grid, m3, col = 4)

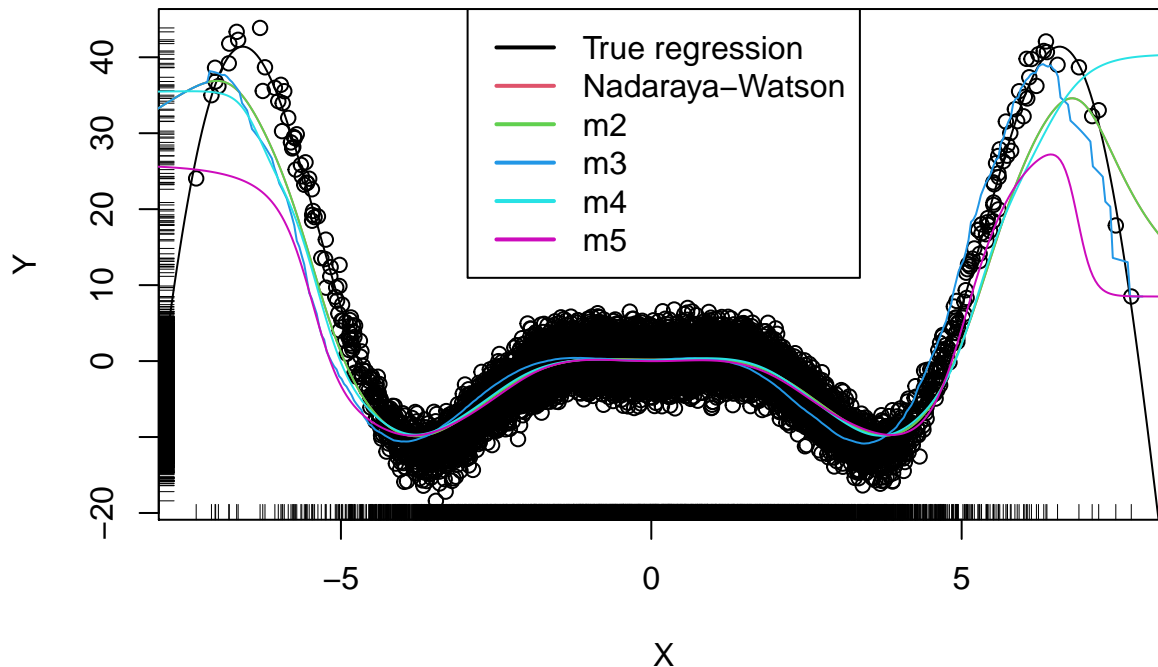
# We choose a sample size equal to 1000 (10% of the dataset)
m4 = nadarayaSample(x = x_grid, X = X, Y = Y, h = h,Muestra=1000)
lines(x_grid, m4, col = 5)
```

```

m5 = nadarayaSampleCV(x = x_grid, X = X, Y = Y, h = h,Muestra=1000,cv=10,m=m)
lines(x_grid, m5, col = 6)

legend("top", legend = c("True regression", "Nadaraya-Watson", "m2", "m3","m4","m5"),
      lwd = 2, col = 1:6)

```



Let's check the speed of "m5".

```
microbenchmark(nadarayaSampleCV(x = x_grid, X = X, Y = Y, h = h,Muestra=1000,cv=9,m=m))
```

```

## Unit: milliseconds
##
## nadarayaSampleCV(x = x_grid, X = X, Y = Y, h = h, Muestra = 1000,      expr
##      min      lq      mean      median      uq      max neval
## 308.2279 363.4432 384.5272 379.7485 402.3654 507.6453   100

```

Let's check what happened with our accuracy. When comparing the first two algorithms: "nw" (the one of our professor) and our "m2" (the one with the loop instead of the "sapply" function) we get:

```
abs(m2-m1)
```

```

## [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [38] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [75] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [112] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [149] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [186] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [223] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [260] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [297] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [334] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [371] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [408] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [445] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

```

```
## [482] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

As we can see, there is no difference between any point in both regression. This is due to the fact that we are not simplifying the mathematical expression of the estimator, but only making it computationally more efficient. When we check other estimators that do a poorer job estimating the regression, but are much faster we get:

```
# Between m3 and m1
abs(m3-m1)
# Between m4 and m1
abs(m4-m1)
# Between m5 and m1
abs(m5-m1)
```

As we can see, we are not able to get another estimator for the regression (another version of the Nadaraya-Watson estimator) that both increases the efficiency and reduce the accuracy by a very short distance ($1e-7$). What we can say is that: if the objective is to make the professor code faster without changing the mathematical expression of the estimator, we can do it by only changing the “sapply” in his code and replace it by a “for” loop. Instead, if we are intended to change this expression, we can do some variants and get to pretty good estimators, that even though they are less accurate than the original one, can get to be more than 10 times faster.

Exercise 3

Part a)

Rubrics: Consider the t-SNE scores y_tsne of the class “3”. Use only the first 2, 000 t-SNE scores for the class “3” for the sake of computational expediency

We first start building a new matrix to use the first 2000 t-SNE scores for the class “3” images.

```
# We build a matrix of two columns whose first row are NAns.
t_SNE=matrix(ncol=2)
# We do a loop in which we check if label of the object is equal to 3.
for(i in 1:60000){
  if(MNIST$labels[i]==3){
    # If it is equal to 3, we add a new row with the values of y_tsne
    t_SNE=rbind(t_SNE,MNIST$y_tsne[i,])
  }
  # When we get to the row number 2001, we stop the loop
  if(length(t_SNE[,1])==2001){
    break
  }
}
# We get rid of the first row, composed by Nans.
t_SNE=t_SNE[2:2001,]
```

Part b)

Rubrics: Compute the normal scale bandwidth matrix that is appropriate for performing kernel mean shift clustering with the first 2, 000 t-SNE scores.

This is a very important part in the procedure of clustering our sample. It is actually the first of the steps to select a suitable bandwidth. In this sense, it is important to note that we need to find a fairly good bandwidth suitable for estimating the gradient of the density and not the density itself. This is due to the fact that not only the target clusters, but also the mean shift algorithm strongly depend on the density gradient (this is proposed by Chacón & Duong (2013)).

This is a crucial point, as a large bandwidth will oversmooth the data and return a lower number of modes. On the opposite, choosing a small bandwidth will undersmooth the data and we will be overestimating the number of modes. Usually, the bandwidth selectors for the density gradient estimation yields larger bandwidths than the ones that are used in estimating the density itself.

It can be understood from what was just mentioned that we do not have to choose the number of clusters before applying this non parametric technique, being this one very nice advantage in comparison to parametric clustering techniques.

Let's compute the Normal Scale Bandwidth with the package `ks`, adding the argument “`deriv.order=1`” to determine that we care about the gradient.

```
Hns <- ks::Hns(x = t_SNE,deriv.order = 1)
print(Hns)
```

```
##           [,1]      [,2]
## [1,] 29.354162 -0.842773
## [2,] -0.842773 27.187406
```

Part c)

Rubrics: Do kernel mean shift clustering on that subset using the previously obtained bandwidth and obtain the modes for $j = 1, \dots, M$, that cluster the t -SNE scores.

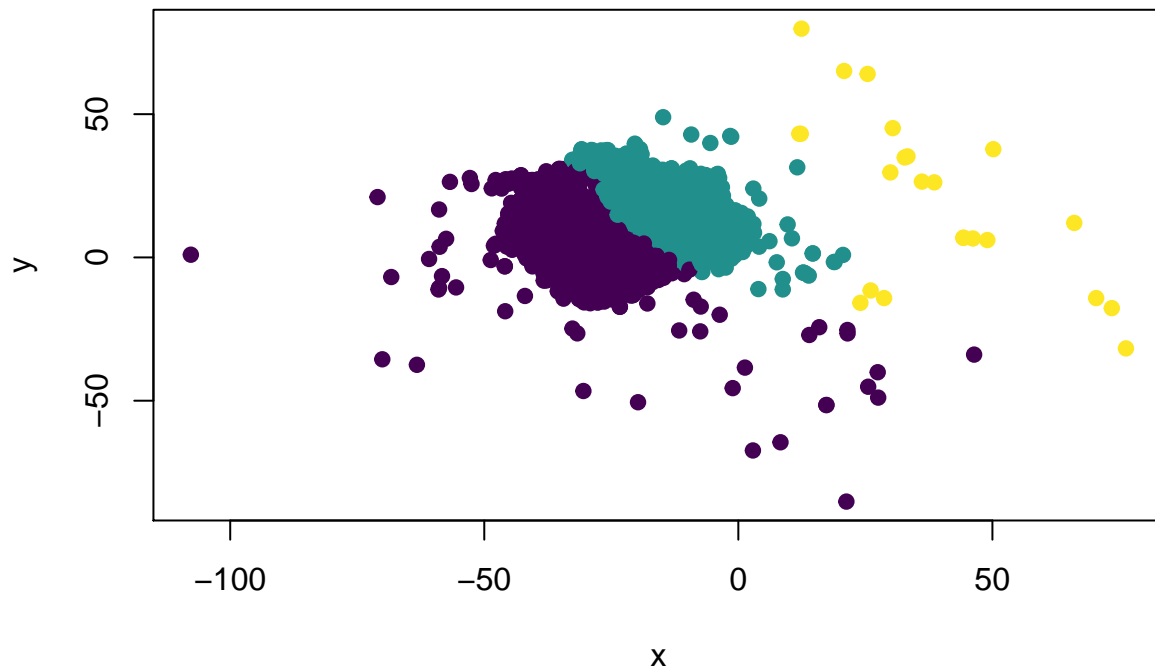
Now, it is time to perform the Kernel Mean Shift Clustering, using the bandwidth obtained in part b). We do this also with the package “`ks`”. It is important to note that both the convergence criteria (of the iteration) and the numerical tolerances will be automatically tested by the algorithm in the package.

```
kms <- ks::kms(x = t_SNE, H = Hns)
summary(kms)
```

```
## Number of clusters = 3
## Cluster label table = 1126 851 23
## Cluster modes =
##           V1          V2
## 1 -27.89169 -1.007476
## 2 -13.52588 14.257348
## 3  45.85105  6.617705
```

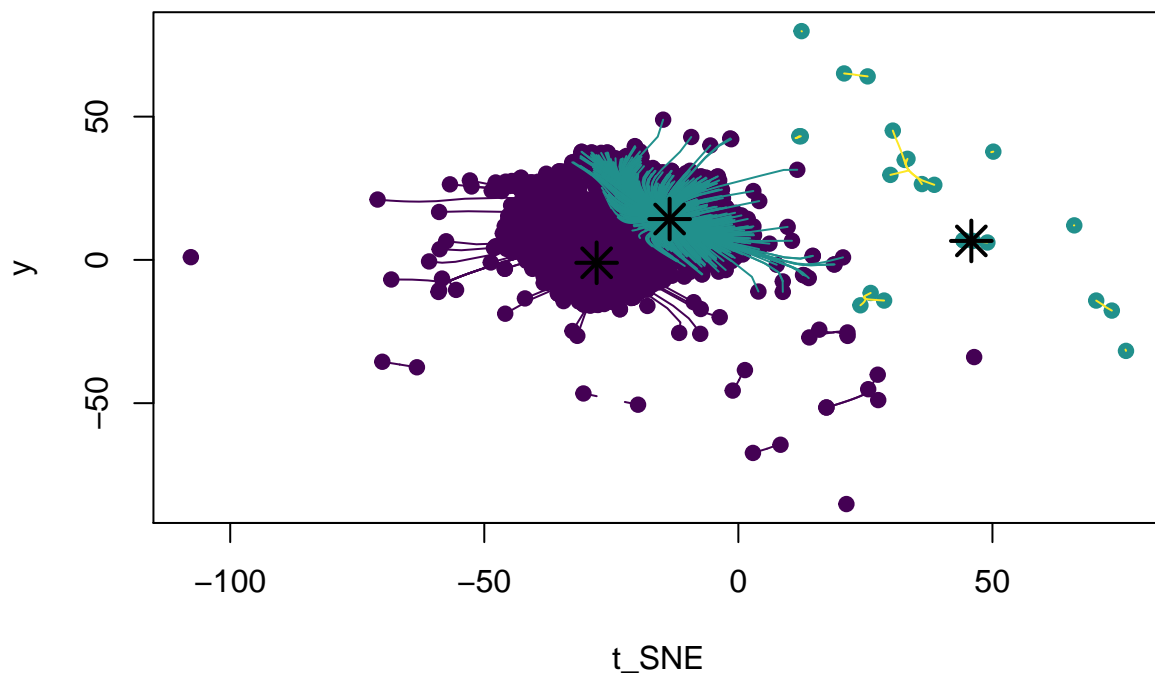
The summary of the Kernel Mean Shift Clustering gives three different outputs: the number of clusters (given the different modes that were found in the density), the sizes of the clusters (how many data points are “attracted” by each mode) and finally the estimated modes (the coordinates of these points).

```
plot(kms, col = viridis::viridis(kms$nclust), pch = 19, xlab = "x", ylab = "y")
```



Another “cool” thing that we can do with this package is to check the ascending paths. This way we can have a very good look at how the points are being “attracted”

```
kms <- ks::kms(x = t_SNE, H = Hns, keep.path = TRUE)
cols <- viridis::viridis(kms$nclust,alpha=1)[kms$label]
plot(kms, col=cols,pch=19,xlab="t_SNE",ylab="y")
for (i in 1:nrow(t_SNE)) lines(kms$path[[i]], col=cols[i])
points(kms$mode,pch=8,cex=2,lwd=2)
```



code has been taken from the notebook of the class

This

Part d)

Rubrics: Determine the M images that have the closest t-SNE scores, using the Euclidean distance, to the modes

As we have found 3 clusters, related to the three modes of the density, now we have to find the three closest points to each of the three modes. So we will repeat a single procedure for the three modes.

```
# First, we create an empty vector and a matrix whose first row is the position of the first mode.
distancia=c()
matriz=matrix(,nrow=2,ncol=2)
# The second row of the matrix will be the position of each of the points.
matriz[1,]=kms$mode[1,]
# We do a for loop and in each iteration we compute the distance between a point and the mode, and we append it to the vector.
for(i in 1:2000){
  matriz[2,]=t_SNE[i,]
  distancia=append(distancia,distance(matriz, method = "euclidean"))
}
```

Now it is time to compute the closest three points and their distances to the first mode.

```
# We build a matrix with dimension 3x4, its first column will be the index of our vector "distancia", the second column will be the minimum distance, the third column will be the index in the vector "distancia" of the minimum distance, and the fourth column will be the index in the vector "distancia" of the second minimum distance.
Mod1_3el=matrix(,ncol=4,nrow=3)
# We make a for loop where we check which is the minimum distance and the index in the vector "distancia" of the minimum distance.
for(i in 1:3){
  Mod1_3el[i,2]=min(distancia)
  Mod1_3el[i,1]=which.min(distancia)
  Mod1_3el[i,3]=t_SNE[which.min(distancia),1]
  Mod1_3el[i,4]=t_SNE[which.min(distancia),2]
  distancia=distancia[-Mod1_3el[i,1]]
}
# We name the columns to get a better visual understanding
colnames(Mod1_3el) <- c("Element","Distance","X","Y")
print(Mod1_3el)
```

```
##      Element Distance      X      Y
## [1,]    1877 0.3364627 -27.69963 -1.283737
## [2,]     263 0.4272029 -27.74574 -1.408975
## [3,]    1011 0.7615577 -52.50696 25.631032
```

We repeat the same procedure for the second mode.

```
distancia=c()
matriz=matrix(,nrow=2,ncol=2)
matriz[1,]=kms$mode[2,]
for(i in 1:2000){
  matriz[2,]=t_SNE[i,]
  distancia=append(distancia,distance(matriz, method = "euclidean"))
}
```

```
## Metric: 'euclidean'; comparing: 2 vectors.
## Metric: 'euclidean'; comparing: 2 vectors.
## Metric: 'euclidean'; comparing: 2 vectors.
## Metric: 'euclidean'; comparing: 2 vectors.
## Metric: 'euclidean'; comparing: 2 vectors.
## Metric: 'euclidean'; comparing: 2 vectors.
## Metric: 'euclidean'; comparing: 2 vectors.
## Metric: 'euclidean'; comparing: 2 vectors.
```

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

```
Mod2_3el=matrix(,ncol=4,nrow=3)
for(i in 1:3){
Mod2_3el[i,2]=min(distancia)
Mod2_3el[i,1]=which.min(distancia)
Mod2_3el[i,3]=t_SNE[which.min(distancia),1]
```


##	Element	Distance	X	Y
## [1,]	915	0.5312555	-13.16573	13.866805
## [2,]	923	0.5975215	-44.40806	7.543359
## [3,]	442	0.9331933	-12.84088	13.623611

```
distancia=c()
matriz=matrix(,nrow=2,ncol=2)
matriz[1,]=kms$mode[3,]
for(i in 1:2000){
  matriz[2,]=t_SNE[i,]
  distancia=append(distancia,distance(matriz, method = "euclidean"))
}
```

57

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

```
## Metric: 'euclidean'; comparing: 2 vectors.
## Metric: 'euclidean'; comparing: 2 vectors.
## Metric: 'euclidean'; comparing: 2 vectors.
## Metric: 'euclidean'; comparing: 2 vectors.
## Metric: 'euclidean'; comparing: 2 vectors.
## Metric: 'euclidean'; comparing: 2 vectors.
## Metric: 'euclidean'; comparing: 2 vectors.
## Metric: 'euclidean'; comparing: 2 vectors.
## Metric: 'euclidean'; comparing: 2 vectors.
## Metric: 'euclidean'; comparing: 2 vectors.
## Metric: 'euclidean'; comparing: 2 vectors.
## Metric: 'euclidean'; comparing: 2 vectors.
## Metric: 'euclidean'; comparing: 2 vectors.
## Metric: 'euclidean'; comparing: 2 vectors.
## Metric: 'euclidean'; comparing: 2 vectors.
## Metric: 'euclidean'; comparing: 2 vectors.
## Metric: 'euclidean'; comparing: 2 vectors.
## Metric: 'euclidean'; comparing: 2 vectors.
## Metric: 'euclidean'; comparing: 2 vectors.
## Metric: 'euclidean'; comparing: 2 vectors.
## Metric: 'euclidean'; comparing: 2 vectors.
## Metric: 'euclidean'; comparing: 2 vectors.
## Metric: 'euclidean'; comparing: 2 vectors.
```

```
Mod3_3el=matrix(,ncol=4,nrow=3)
for(i in 1:3){
  Mod3_3el[i,2]=min(distancia)
  Mod3_3el[i,1]=which.min(distancia)
  Mod3_3el[i,3]=t_SNE[which.min(distancia),1]
  Mod3_3el[i,4]=t_SNE[which.min(distancia),2]
  distancia=distancia[-Mod3_3el[i,1]]
}
colnames(Mod3_3el) <- c("Element","Distance","X","Y")
print(Mod3_3el)
```

```
##      Element Distance      X      Y
## [1,]      118 0.3149848 46.16603 6.615629
## [2,]      423 1.5367648 -23.47996 -4.693805
## [3,]      530 1.5968756 -34.18276 13.325191
```

Now, our objective is to find these points in the original dataset in order to plot them later (in the next part).

We create an empty matrix of dimensions 3x3 in which the columns are the three modes and the rows are
In this matrix, we want to obtain the index of the row in the original matrix of images to know what
n=0

```
posiciones=matrix(nrow=3,ncol=3)
# We do a for loop, where we iterate in the original matrix and check when the label is equal to 3, whe
# Also, we check if this point is one of the closest to any of the modes. If this is the case, we fill
for(i in 1:60000){

  if(MNIST$labels[i]==3){
    n=n+1
    if(n==Mod1_3el[1,1]){
      posiciones[1,1]=i
    }
  }
}
```

```

    if(n==Mod1_3el[2,1]){
      posiciones[2,1]=i
    }
    if(n==Mod1_3el[3,1]){
      posiciones[3,1]=i
    }
    if(n==Mod2_3el[1,1]){
      posiciones[1,2]=i
    }
    if(n==Mod2_3el[2,1]){
      posiciones[2,2]=i
    }
    if(n==Mod2_3el[3,1]){
      posiciones[3,2]=i
    }
    if(n==Mod3_3el[1,1]){
      posiciones[1,3]=i
    }
    if(n==Mod3_3el[2,1]){
      posiciones[2,3]=i
    }
    if(n==Mod3_3el[3,1]){
      posiciones[3,3]=i
    }
  }
}
# In order to reduce computational expenditure, we will stop this loop once we iterated through all o
if(n==2000){
  break
}

}
# We name the columns and rows to get a better understanding of the output.
colnames(posiciones) <- c("Mod1","Mod2","Mod3")
rownames(posiciones) <- c("closest","second closest","third closest")
print(posiciones)

##           Mod1 Mod2 Mod3
## closest      17930 8937 1245
## second closest  2706 9005 4272
## third closest   9779 4491 5359

```

Part e)

Rubrics: Show the closest images associated with the modes. Do they represent different forms of drawing the digit “3”?

Now that we found the three closest points to each one of the three modes, we just need to plot them.

First, let's plot the three images closest to the first mode:

```
show_digit <- function(vec, col = gray(12:1 / 12), ...) {  
  image(matrix(vec, nrow = 28)[, 28:1], col = col, ...)  
}  
for(i in 1:3)  
  show_digit(MNIST$x[posiciones[i,1]  
    , ], axes = FALSE)
```



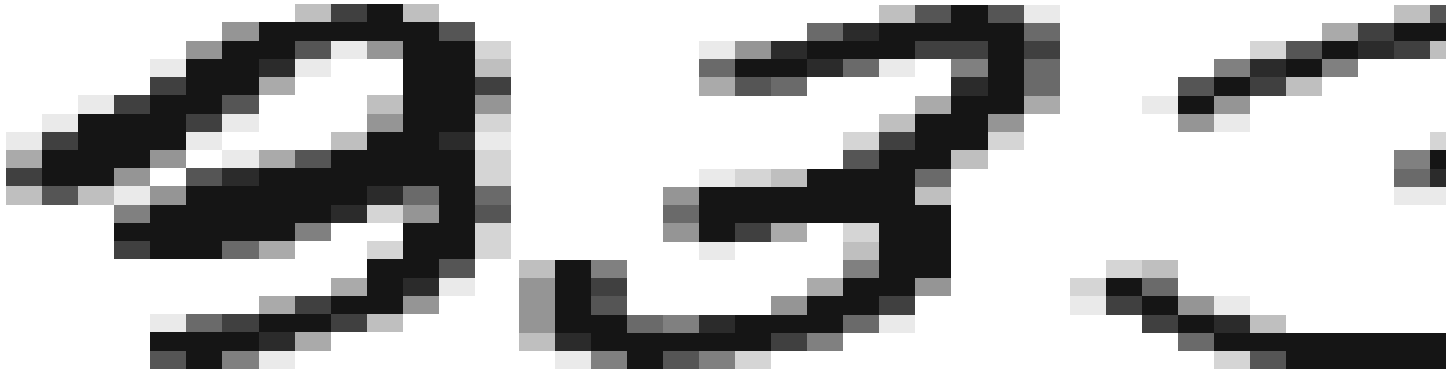
Plotting the three closest images to the second mode:

```
for(i in 1:3)  
  show_digit(MNIST$x[posiciones[i,2]  
    , ], axes = FALSE)
```



Plotting the three closest images to the third mode:


```
for(i in 1:3)
  show_digit(MNIST$x[posiciones[i,3]
, ], axes = FALSE)
```



After inspecting the plotted images we can say that each mode represent a different way of writing the digit three. It is specially easy to note in the second mode, where all the digits (at least the ones that are closest to it) have a very similar shape, they are very rounded. Meanwhile, the first mode shows us another way of writing digit “3”, as they are more inclined and less rounded. Mode number two is a bit more difficult to grasp, as they three images closest to this mode are not so similar, although they do reflect a different way of writing the digit (in comparison with the other two modes).