Report (Group 9): Home Work 2

Course Code: Data Science Lab 3 (MTH312A)

Submitted by, Group 9

Anirban Ghosh(221271), Khyati Singh(221332)

Dasari Charithambika(210302)

Rajdeep Adhya(221385), Rohit Dutta(221396)

Instructor

Dr. Subhra Sankar Dhar

Associate Professor

Department of Mathematics and Statistics,

Indian Institute of Technology, Kanpur

Submission Date: 15 February, 2024

1 Question 1

Download a two sample multivariate data, where dimension of the data is larger than sample size of the data. Check whether the distributions associated with two samples are independent or not.

Let consider

$$\mathbf{X} = \{\underline{X_1}, \underline{X_2}, ..., \underline{X_n}\}$$
$$\mathbf{Y} = \{\underline{Y_1}, \underline{Y_2}, ..., \underline{Y_n}\}$$

are the two sample multivariate data.

- $f(\underline{X_i})$ is the marginal probability density function of $\underline{X_i}$ and $f(\underline{Y_i})$ is the marginal probability density function of $\underline{Y_i}$ $\forall\ i = 1(1)n$.
- $f(\underline{X_i}, \underline{Y_i})$ is the joint probability density function of $\underline{X_i}$ and $\underline{Y_i}$.
- $\varphi_{\underline{X_i}}(t)$ is the marginal characteristic function of $\underline{X_i}$ and $\varphi_{\underline{Y_i}}(t)$ is the marginal characteristic function of $\underline{Y_i}$ $\forall\ i = 1(1)n,\ t$.
- $\varphi_{(\underline{X_i}, \underline{Y_i})}(t)$ is the joint characteristic function of $\underline{X_i}$ and $\underline{Y_i}$.

To check whether the two distributions are independent or not:-

- **By probability density function:**
  - If the joint probability density function can be expressed as the product of their respective marginal probability density functions, then the two distribution are said to independent.

$$f(\underline{X_i}, \underline{Y_i}) = f(\underline{X_i}).f(\underline{Y_i})\ ;\ \forall\ i$$

- **By characteristic function:**
  - If the joint characteristic function can be expressed as the product of their respective

marginal characteristic functions, then the two distribution are said to independent.

$$\varphi_{(\underline{X_i},\underline{Y_i})}(t) = \varphi_{\underline{X_i}}(t).\varphi_{\underline{Y_i}}(t) \ ; \ \forall \ i, t$$

Description of the Data:

We have taken a data named **"Urban Land Cover"** which is intended to assist sustainable urban planning efforts by classification of urban land cover using high resolution aerial imagery. The data can be found in the following link https://archive.ics.uci.edu/dataset/295/urban+land+cover.

This dataset contains a data for classifying a high resolution aerial image into 9 types of urban land cover. Multi-scale spectral, size, shape, and texture information are used for classification. Class is the classification variable. The land cover classes are: trees, grass, soil, concrete, asphalt, buildings, cars, pools, shadows. The data set contains 148 features such as area, round, bright, compact etc. The total number of observations in the data is 168. But we have considered two classes only namely, "shadow" and "asphalt" each class having 45 observations.

So, finally we got a two sample (one for "shadow" and another for "asphalt") multivariate (148 covariates) where each sample has dimension $\mathbf{45 \times 148}$.

Checking Independence:

To reduce the curse of dimensionality, we have used principal component analysis (PCA) and finally out of 148 principal components, we have selected first 10 principal components to explain maximum variability. Then, we subset the whole data according to their class "shadow" and "asphalt".

By probability density function approach:

As the actual joint distribution of the two samples and their respective marginal distributions are not known, we used Kernel density estimation to obtain the joint and the marginal distributions of the two samples.

Then we took the difference between the log of joint density and the log of the product of the two marginal densities of the two samples obtained by KDE method and we plotted the result

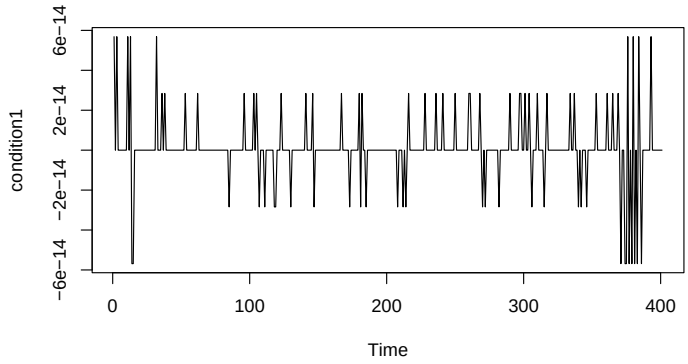

Figure 1: Plot of difference between the log of the joint density and log of the product of the two marginal densities of the two samples obtained by KDE method vs time (refers to the points where KDE has been applied). In the y-axis, the "condition1" refers to the said difference

which is shown in figure 1.

By characteristic function approach:

We have computed the joint characteristics function for the whole data and the respective marginal characteristic functions for the individual two samples.

then we computed the difference between the joint characteristic function and the product of the marginal characteristic functions for 1000 values of t. As, the resultant difference contains some imaginary values also, we have taken the norm of the difference and plotted it, which is shown in figure 2.

Conclusion:

From from figure 1, i.e. the plot of the difference between the log of joint density and the log of the product of the two marginal densities of the two samples obtained by KDE method, we observe that most of the values are close to zero. That possibly indicates that the distribution of the two samples are may be independent.

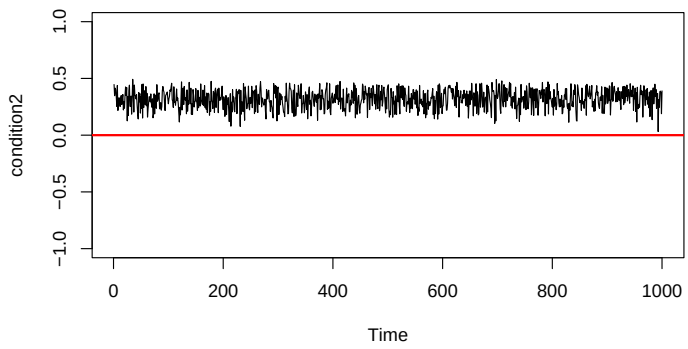

Figure 2: Plot of difference between the joint characteristic function and the product of the marginal characteristic functions for 1000 values of t

Also, from the figure 2, i.e. the plot of difference between the joint characteristic function and the product of the marginal characteristic functions for 1000 values of t, we can observe the norm values are close to zero. This, also, possibly indicates that the underlying distributions of the two samples are may be independent.

2 Question 2

Download a data, which is suitable for non-parametric regression models. For this data, estimate the regression function and its first and second derivatives using local polynomial mean and median approach. Compare the performance of the estimators obtained from both approaches.

Set up:

$$\mathbf{X} = \{x_1, x_2, ..., x_n\}$$
$$\mathbf{Y} = \{y_1, y_2, ..., y_n\}$$

where $\mathbf{X}, \mathbf{Y}$ are independent regressors and responses variables corresponding.

Want to fit the model:

$$y_i = m(x_i) + \epsilon_i \ ; \ i = 1, 2, ..., n$$

If $\{(X_i, Y_i)\}_{i=1(1)n}$ are i.i.d replications of (X, Y), the model can be written as

$$Y_i = m(X_i) + \epsilon_i$$

Want to estimate $m(X)$ i.e the Regression curve:

Suppose that $E(\epsilon) = 0$

For the model $Y = m(X) + \epsilon$, $\mathrm{E}(Y|X = x) = m(X)\ \forall x$, which is Mean regression estimator. Hence, the estimation of $m(x)$ is equivalent to estimate $\mathrm{E}(Y|X = x)$. Hence,

$$\widehat{\mathrm{E}(Y|X = x)} = \int y \left\{ \frac{\widehat{k}_{X,Y}(x, y)}{\widehat{g_X}(x)} \right\} dy$$

on calculating the estimates,

$$\widehat{\mathrm{E}(Y|X = x)} = \frac{\sum_{i=1}^{n} y_i p(\frac{x - X_i}{h_n})}{\sum_{i=1}^{n} p(\frac{x - X_i}{h_n})}$$

This is known as **Nadarya-Watson regression estimator** where,

- x = point of evaluation
- (X_i, Y_i) : given data for the i^{th} individual; $i = 1(1)n$.
- h_n = sequence of band width, such that $n.h_n^2 \to \infty$ and $h_n \to 0$ as $n \to \infty$.
- $p(.)$ = choice of kernel satisfying some conditions

$$\begin{aligned}\widehat{m_n}(x) &= \frac{\sum_{i=1}^{n} y_i p(\frac{x-X_i}{h_n})}{\sum_{i=1}^{n} p(\frac{x-X_i}{h_n})} \\ &= \frac{\sum_{i=1}^{n} y_i w(x_i)}{\sum_{i=1}^{n} w(x_i)} \\ &= \arg\min_\theta \frac{1}{n}\sum_{i=1}^{n}(y_i-\theta)^2 w(x_i) \ldots\ldots (1)\end{aligned}$$

It is called as **local constant mean estimator.**

Instead of square, if we take absolute then

$$\widehat{m_n}(x) = \arg\min_\theta \frac{1}{n}\sum_{i=1}^{n}|y_i-\theta| w(x_i) \ldots\ldots (2)$$

This is called as **local constant median estimator.**

Also, let $(\widehat{m_n^{(0)}}(x_0), \widehat{m_n^{(1)}}(x_0), \ldots, \widehat{m_n^{(p)}}(x_0))$ be a vector of first till p^{th} order derivative of $\widehat{m_n}$ at the point x_0.

Then, it can be shown that, $(\widehat{m_n^{(0)}}(x_0), \widehat{m_n^{(1)}}(x_0), \ldots, \widehat{m_n^{(p)}}(x_0)) = \arg\min_{(\theta_0,\ldots,\theta_p)\in\mathbb{R}^{p+1}} \frac{1}{n}\sum_{i=1}^{n}\{(y_i - \theta_0 - \theta_1(x_i-x_0) - \frac{\theta_2}{2!}(x_i-x_0)^2 - \cdots - \frac{\theta_p}{p!}(x_i-x_0)^p)^2\} p(\frac{x-X_i}{h_n}) \ldots\ldots (3)$

Description of the Data:

We have considered a data on Daily air quality measurements in New York, May to September 1973 which is taken from R software. The data set contains 153 observations on 6 variables.

The variables are:

1. **Ozone:** Mean ozone in parts per billion (ppb) from 1300 to 1500 hours at Roosevelt Island
2. **Solar.R:** Solar radiation in Langleys in the frequency band 4000–7700 Angstroms from 08:00 to 12:00 hours at Central Park

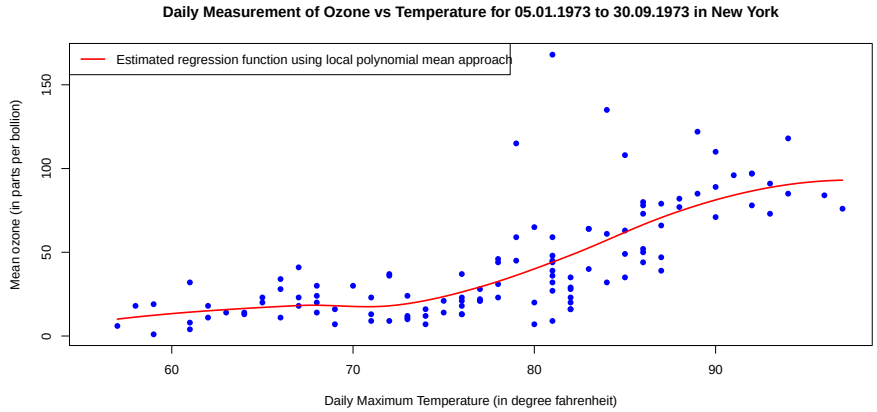

Figure 3: Daily Measurement of Ozone vs Temperature for 05.01.1973 to 30.09.1973 in New York

3. **Wind:** Average wind speed in miles per hour at 0700 and 1000 hours at LaGuardia Airport

4. **Temp:** Maximum daily temperature in degrees Fahrenheit at La Guardia Airport.

Another two variables signifies month and the day of the study.

Source: The data were obtained from the New York State Department of Conservation (ozone data) and the National Weather Service (meteorological data).

Here, in this problem, we have considered Mean ozone (in ppb) as the response and maximum daily temperature (in degree F) as the independent variable or the predictor.

Also, the data contains some blank cells, so we ignored them and ultimately we work with a data of dimension $\mathbf{111} \times \mathbf{2}$.

Estimation of the regression function and its first and second derivatives using local polynomial mean approach:

As described in (1), we estimated the the regression function using local polynomial mean approach and plotted the fitted curve on the original data which is shown in figure 3.

Also, the first order derivative and second order derivatives are obtained according to the formula in (3). Now, as our final data set contains 111 number of observations, we computed the estimated value of local polynomial mean estimator along with the first and second order

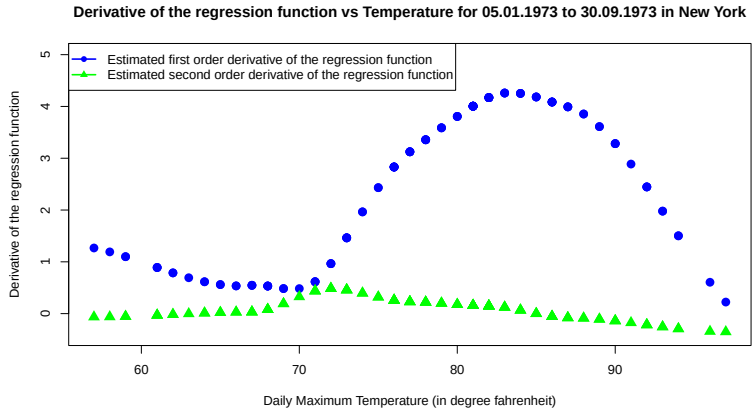

Figure 4: Derivative of the regression function (mean approach) vs Temperature for 05.01.1973 to 30.09.1973 in New York

derivatives at those points and it is displayed in the figure 4. But, here in this report we are showing the results for 5 randomly chosen daily maximum temperature (in ${}^0F$) in table 1.

Table 1: Results related to local mean estimator approach

Temp(in ${}^0F$)	Ozone (in ppb)	Estimated Ozone (in ppb)	$\widehat{m_n^{(1)}}(x)$	$\widehat{m_n^{(2)}}(x)$
80	20	40.07486	3.806933	0.17607920
76	23	26.26380	2.830873	0.25622375
90	110	81.25805	3.282088	-0.13819280
76	18	26.26380	2.830873	0.25622375
74	16	21.17717	1.964496	0.39539802

Estimation of the regression function and its first and second derivatives using local polynomial median approach:

As described in (1), we estimated the the regression function using local polynomial mean approach and plotted the fitted curve on the original data which is shown in figure 5.

Also, the first order derivative and second order derivatives are obtained according to the formula in (3). Now, as our final data set contains 111 number of observations, we computed the estimated value of local polynomial mean estimator along with the first and second order derivatives at those points and it is displayed in the figure 6. But, here in this report we are showing the results for 5 randomly chosen daily maximum temperature (in ${}^0F$) in table 2.

Comparison of the performance of the estimators obtained from both the local

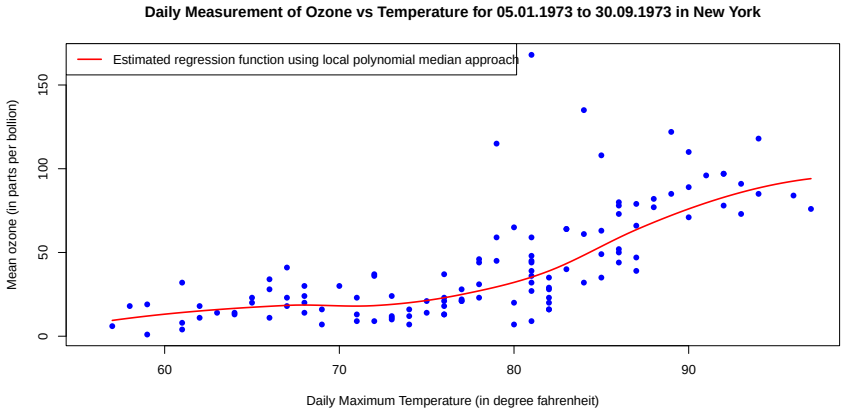

Figure 5: Daily Measurement of Ozone vs Temperature for 05.01.1973 to 30.09.1973 in New York

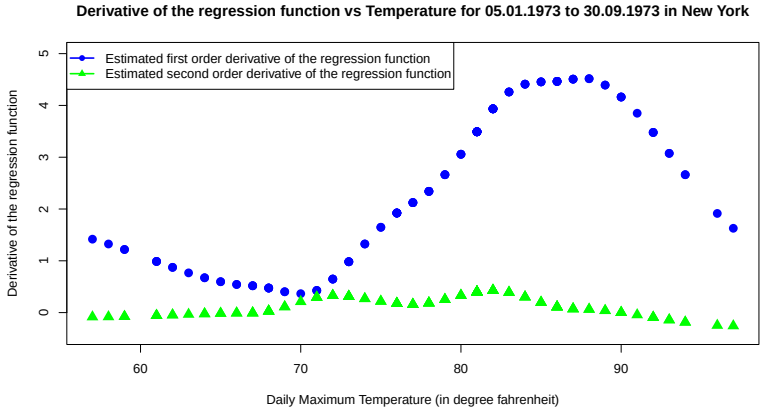

Figure 6: Derivative of the regression function (median approach) vs Temperature for 05.01.1973 to 30.09.1973 in New York

Table 2: Results related to local mean estimator approach

Temp(in ^0F)	Ozone (in ppb)	Estimated Ozone (in ppb)	$\widehat{m_n^{(1)}}(x)$	$\widehat{m_n^{(2)}}(x)$
84	135	48.36216	4.4101384	0.301773704
82	28	38.94930	3.9343560	0.426394373
81	32	35.25979	3.4914622	0.398359577
67	18	18.40020	0.5185993	-0.006913788
69	16	18.43235	0.4025751	0.1116570352

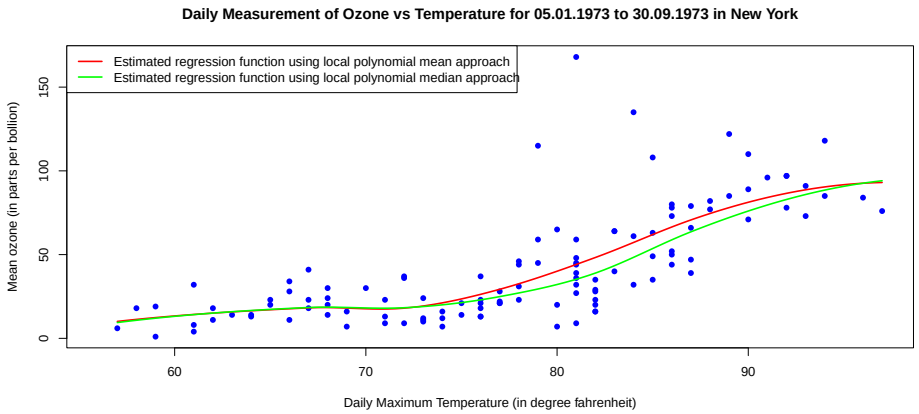

Figure 7: Daily Measurement of Ozone vs Temperature for 05.01.1973 to 30.09.1973 in New York

polynomial mean and median approaches:

The figure 7 shows two regression curves associated with local mean approach as well as local median approach, overlaid on the same graph.

To compare the performance of the local polynomial mean regression and the local median regression approach, we used some criteria such as the root mean squared error (RMSE) and the coefficient of determination (R-squared).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where, n is the total number of observations, y_i is the i^{th} value of the response, $\hat{y}_i$ is the estimated value of the response from the regression function and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

The RMSE and R^2 values for the local polynomial mean estimator approach are found to be 21.74275 and 0.5730593 respectively. So, we can say at around 57% of the total variability of the response is explained by the regression equation obtained by local polynomial mean approach.

The RMSE and R^2 values for the local polynomial median estimator approach are found to be 22.53642 and 0.5413212 respectively. So, we can say at around 54% of the total variability of the response is explained by the regression equation obtained by local polynomial median approach.

Though the RMSE and R^2 values corresponding to mean approach is slightly lesser and greater than that of the median approach; the RMSE values and R^2 values for both the models are quite close to each other.

So, we can possibly say that, the local polynomial mean approach performs slightly better than the local polynomial median approach but there is no significant difference in performance among the two approaches in this case.