Report (Group 9): Home Work 1

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: 24 January, 2024

1 Question 1

Download Iris data and check whether the observations associated with Iris setosa, Iris virginica, and Iris versicolor obtained from the same distribution or not.

- Iris data set gives the measurements in centimeters of the 5 variables for 50 flowers from each of 3 species of iris.
- The variables are
  1. Sepal Length
  2. Sepal Width
  3. Petal Length
  4. Petal Width
- The Species are
  1. Setosa
  2. Versicolor
  3. Virginica
- The dimensions of the Iris data set is **150** x **5**.
- **Definition**: Consider a data cloud $\mathbf{X} = (x_1, ..., x_n)$ with data points $x_i \in \mathbb{R}^d$. Conditioned on it, a statistical depth function assigns to an arbitrary point $z \in \mathbb{R}^d$ its degree of centrality, $z \mapsto D(z|X) \in [0, 1]$. The half-space depth ([1]) is determined as the smallest fraction of data points contained in a closed half-space containing z.
- **Half-space depth [1]:** The half space depth(HD) at x w.r.t F is defined to be

$$HD(F; x) = \inf_H \{P(H) : H \text{ is a closed half-space in } \mathbb{R}^d \text{and } x \in H\}$$

  The sample version of $HD(F; x)$ is $HD(F_n; x)$ Here F_n denotes the empirical distribution of the sample $X_1, ..., X_n$.

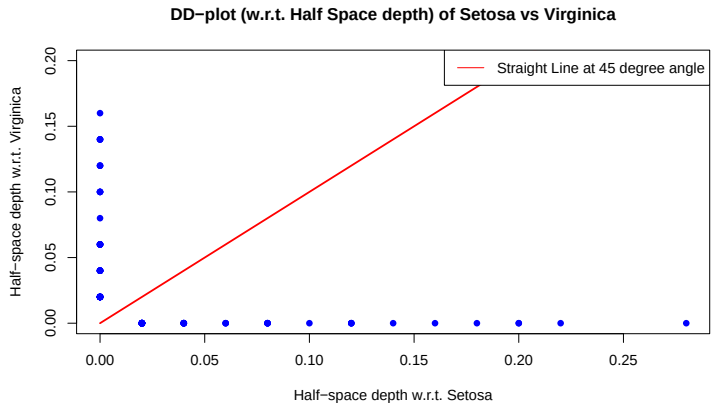

Figure 1: DD-plot (w.r.t. Half Space depth) of Setosa vs Virginica

- In this case, we have to test the hypothesis

$$H_0 : F_1 = F_2 = F_3$$

, where F_i is the distribution associated with species i which are setosa, virginica and versicolor. To evaluate the distributions of the three distinct Iris species, we have employed the notion of data depth, more precisely half-space depth.

- We are focusing on graphical comparisons of two multivariate distributions based on data-depth plots using the half-space depth of their samples. We say the two given distributions are identical, if these plots are segments of the diagonal line from (0, 0) to (1,1) in $\mathbb{R}^2$. Plots that deviate from this line indicate differences between the two distributions.

- We have 50 data points for each of the species setosa, virginica, and versicolor. We analyzed the data by considering two categories at a time. First, we analyzed whether the observations associated with Iris setosa and Iris virginica are obtained from the same distribution or not.

- From the figure 1 we can see that the DD-plot is not concentrated along the diagonal line but exhibits a noticeable perpendicular departure. We can interpret that the locations of the two distributions of species Iris setosa and Iris virginica are not the same.

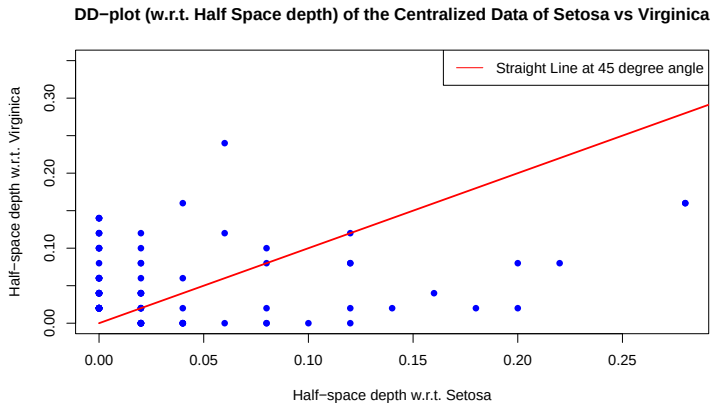

Figure 2: DD-plot (w.r.t. Half Space depth) of the Centralized Data of Setosa vs Virginica

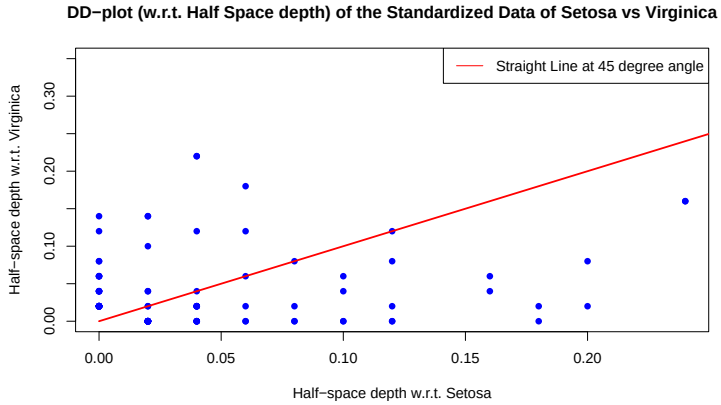

Figure 3: DD-plot (w.r.t. Half Space depth) of the Standardized Data of Setosa vs Virginica

- In order to bring out scale differences, the centers of the samples are equalized by subtracting the data from their respective center. We get that even after centralizing the data and bringing it to its origin the DD-plot exhibits departure from the diagonal line. The following can be seen clearly from the figure 2

- Now we standardize the data both location and scale are equalized by transforming the data to $S_X^{-\frac{1}{2}}X$ and $S_Y^{-\frac{1}{2}}Y$, where S_X and S_Y are dispersion matrices of the centralized setosa data and centralized virginica data. From the figure 3 we can see that there is not much difference from the previous graph and the data is scattered from the diagonal line.

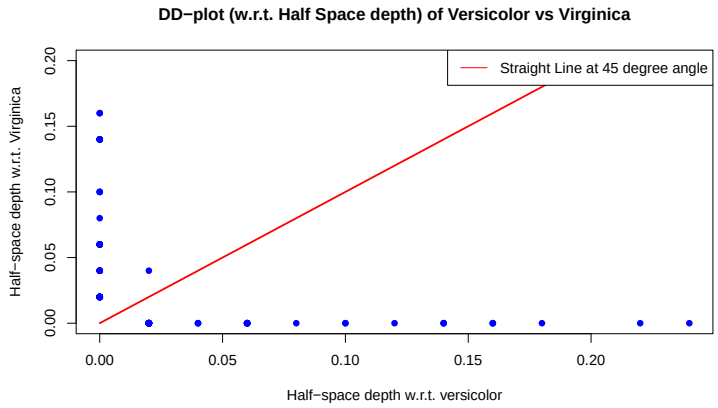

Figure 4: DD-plot (w.r.t. Half Space depth) of Versicolor vs Virginica

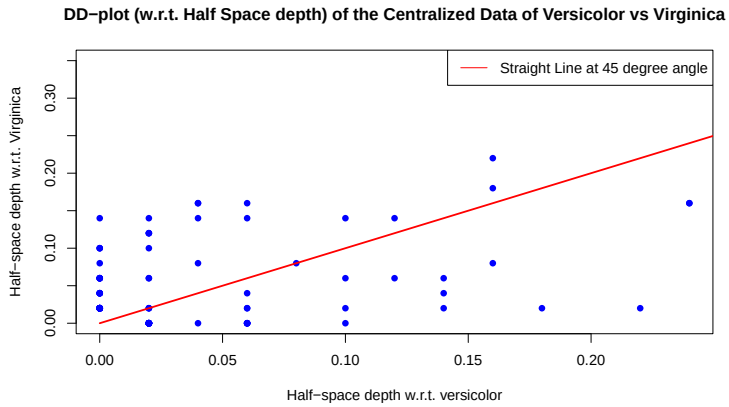

Figure 5: DD-plot (w.r.t. Half Space depth) of the Centralized Data of Versicolor vs Virginica

- Thus, we can conclude that the distributions of two species Iris setosa and Iris virginica are not identical.
- Now, similarly we will analyze the distributions of the observations associated with Iris versicolor and Iris virginica.
- From the figure 4 we can see that the DD-plot is not concentrated along the diagonal line but exhibits a noticeable perpendicular departure. We can interpret that the locations of two distributions of species Iris versicolor and Iris virginica are not the same.
- To bring out scale differences, the centers of the samples are equalized by subtracting the

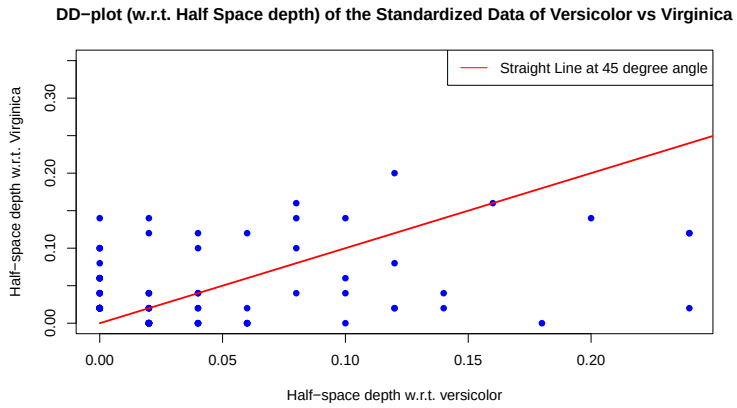

Figure 6: DD-plot (w.r.t. Half Space depth) of the Standardized Data of Versicolor vs Virginica

data from their respective center. We get that even after centralizing the data and bringing it to its origin the DD-plot exhibits departure from the diagonal line. The following can be seen clearly from the figure 5

- Now we standardize the data both location and scale are equalized by transforming the data to $S_X^{-\frac{1}{2}}X$ and $S_Y^{-\frac{1}{2}}Y$, where S_X and S_Y are dispersion matrices of the centralized veriscolor data and virginica data respectively. From the figure 6 we can see that there is not much difference from the previous graph and the data is scattered from the diagonal line.

- Thus, we can conclude that the distributions of two species Iris veriscolor and Iris virginica are not identical.

- From the figure 7 we can see that the DD-plot is not concentrated along the diagonal line but exhibits a noticeable perpendicular departure. We can interpret that the locations of two distributions of species Iris versicolor and Iris setosa are not the same.

- To bring out scale differences, the centers of the samples are equalized by subtracting the data from their respective center. We get that even after centralizing the data and bringing it to its origin the DD-plot exhibits departure from the diagonal line. The following can be seen clearly from the figure 8

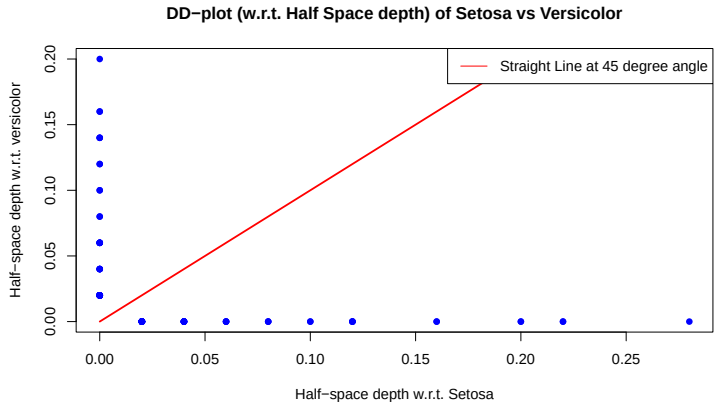

Figure 7: DD-plot (w.r.t. Half Space depth) of Setosa vs Versicolor

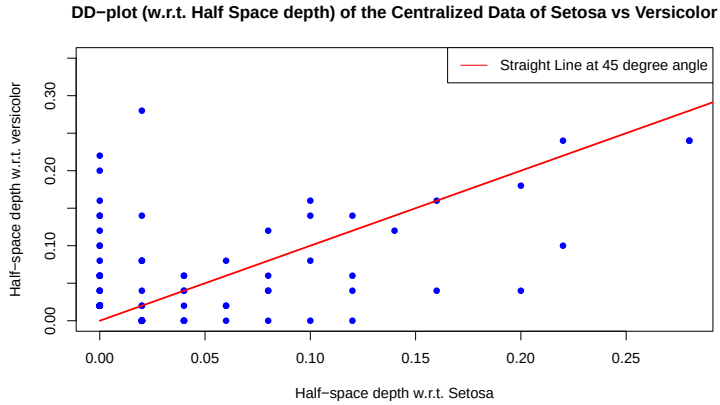

Figure 8: DD-plot (w.r.t. Half Space depth) of the Centralized Data of Setosa vs Versicolor

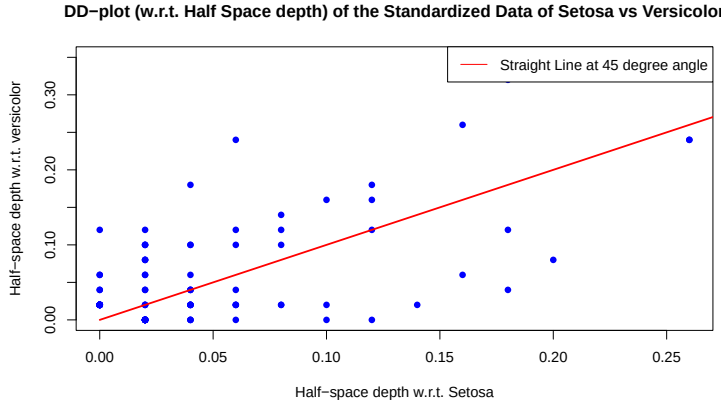

Figure 9: DD-plot (w.r.t. Half Space depth) of the Standardized Data of Setosa vs Versicolor

- Now we standardize the data both location and scale are equalized by transforming the data to $S_X^{-\frac{1}{2}}X$ and $S_Y^{-\frac{1}{2}}Y$, where S_X and S_Y are dispersion matrices of the centralized veriscolor data and setosa data respectively. From the figure 9 we can see that there is not much difference from the previous graph and the data is scattered from the diagonal line.
- Thus, we can conclude that the distributions of the two species Iris veriscolor and Iris setosa are not identical.
- Therefore, we can conclude that the distributions of three species of Iris setosa, Iris virginica, and Iris versicolor are not identical.

2 Question 2

Download a multivariate (i.e, dimension is strictly greater than one) data and compute/draw multivariate quantile contours when $||u|| = i/10$, where $i = 1, \ldots, 9$. Using those contours, describe various features of the data set.

- Iris data set gives the measurements in centimeters of the 5 variables for 50 flowers from each of 3 species of iris.
- For the problem, we have considered the data of Sepal Length and Sepal Width of all the flowers whose dimension is **150** x **2**.
- **Quantile contours:** In multivariate data $\mathbf{X} = \{x_1, x_2, ..., x_n\}$ where $x_i \in \mathbb{R}^d, d \geq 1, i = 1, 2, ..., n$.

$$\hat{Q}_{n,\mathbf{u}} = \arg\min_{Q \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \{||x_i - Q|| + < \mathbf{u}, x_i - Q >\}$$

  where $\mathbf{u} \in \mathbb{B}^d = \{\mathbf{v} \in \mathbb{R}^d : ||\mathbf{v}|| < 1\}$.

- **Population version:**

$$Q_{\mathbf{u}} = \arg\min_{Q \in \mathbb{R}^d} E[||X - Q|| + < \mathbf{u}, X - Q >]$$

First we scaled the data so that all the observations lie around the origin because in the algorithm 1 we considered circles having center at origin $(0, 0)$ so that computations can be done

Algorithm 1 Algorithm for Multivariate Quantile contour

Step 1: For each $i = 1, 2, ..., n$ check whether or not

$$\left\| \sum_{1 \leq j \leq n; j \neq i} \frac{x_i - x_j}{||x_i - x_j||} + (n-1)\mathbf{u} \right\| \leq (1 + ||\mathbf{u}||)$$

If it satisfies for some $i = 1, 2, ..., n$ then $\hat{Q}_{n,\mathbf{u}} = x_i$

Step 2: Otherwise, one needs to solve

$$\sum_{i=1}^{n} \frac{x_i - \hat{Q}_{n,\mathbf{u}}}{||x_i - \hat{Q}_{n,\mathbf{u}}||} + n\mathbf{u} = 0$$

to obtain the desired quantile.

faster. Then to obtain multivariate quantile contours we have implemented the algorithm 1. Here, we have to draw quantile contours for $||\mathbf{u}|| = \frac{i}{10}, i = 1, 2, ..., 9$. For each $||\mathbf{u}||$, we have drawn 10 values of $\mathbf{u}$, corresponding to which we get 10 values of quantiles $||\mathbf{q}||$. They are joined to get the contour for each value of $||\mathbf{u}||$. Here, we have drawn only 10 values of $\mathbf{u}$ for each $||\mathbf{u}||$ because implementation of the algorithm 1 became too time consuming to implement for large number of draws.

2.1 Conclusion

From the figure 10, we observe that the quantile contours corresponding to $||\mathbf{u}|| = \frac{i}{10}$; for $i = 1, 2, 3, 4, 5, 6$ are concentrated around the center (deepest point) and are close to each other signifying high concentration around the origin. This fact can be verified from the KDE (Kernel Density Estimation) plot (figure 11) of the bi-variate distribution of sepal length and sepal width.

Also, the first six quantile contours have less spread towards the right and more spread towards the left signifying a possible region of high concentration area which gives an indication of multi-modal distribution between sepal length and sepal width. The presence of two modes can be verified from figure 11.

The quantile contours exhibit greater concentration towards the center with tails extending outward. This pattern indicates that our data is likely derived from a leptokurtic distribution.

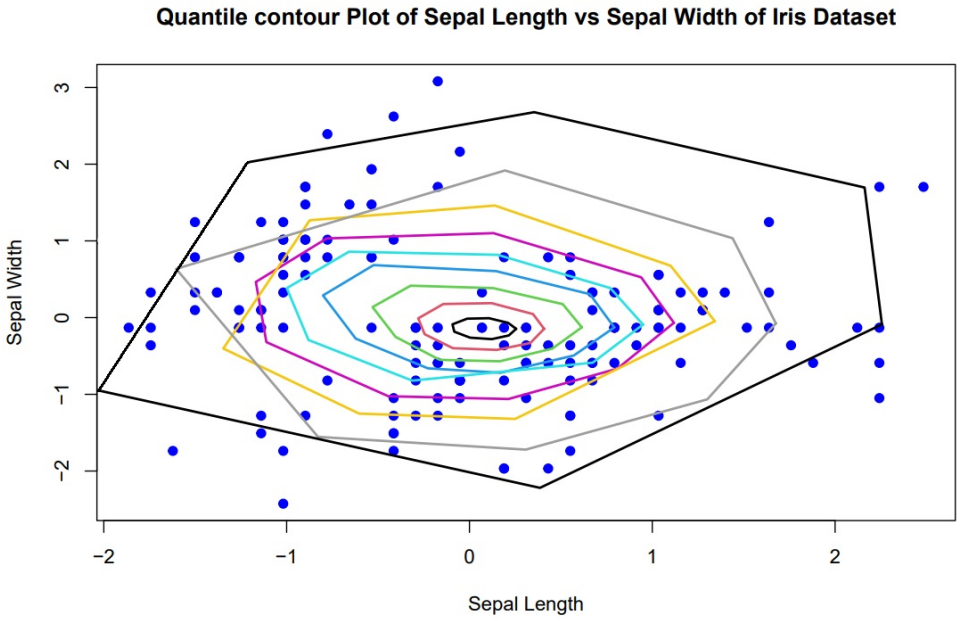

Figure 10: Quantile contour plot of sepal length and sepal width of iris data set. The quantile contours corresponding to $||u|| = \frac{i}{10}; i = 1, 2, 3, \ldots, 9$ are shown in color black, red,green,blue,cyan,magenta,yellow,lightgray, darkgray respectively from inside towards outside.

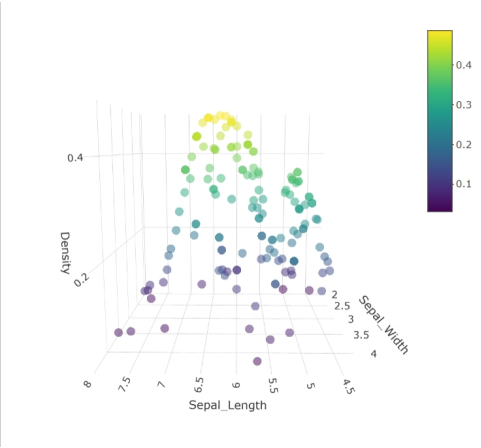

Figure 11: KDE Plot of the joint distribution of sepal length and sepal width of the flowers in iris data set

References

[1] Regina Y Liu, Jesse M Parelius, and Kesar Singh. Multivariate analysis by data depth: descriptive statistics, graphics and inference,(with discussion and a rejoinder by liu and singh). *The annals of statistics*, 27(3):783–858, 1999.