Indian Institute of Technology, Kanpur

Data Science Lab
(MTH312A)
HOME WORK-1
REPORT

Group Members:
Mrinmoy Saha (221352)
Sudesh Kumari (221440)
Vijay Soren (211163)
Manjeet Chaudhary (221346)
Pravin Raman Bharti (221375)

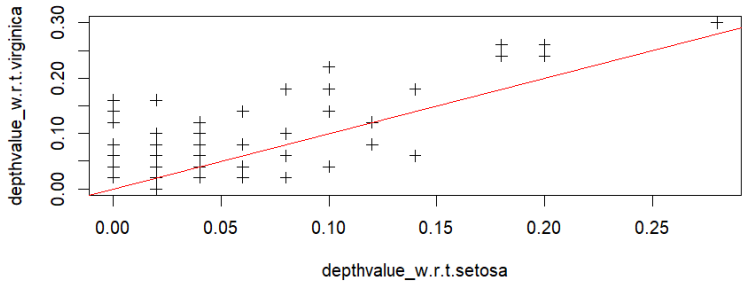
Question 1: Download Iris data and check whether the observations associated with Iris setosa, Iris virginica and Iris versicolor obtained from the same distribution or not.

Answer : Iris dataset is consist of five columns namely Petal Length ,Petal Width, Sepal Length, Sepal Width and lastly Species .First we separate the data based on Species using some standard coding techniques in R. So basically now we have three dataset which are based on setosa , versicolor and virginica. Now we are concerned with checking whether the three multivariate data are from same distribution or not. We take the approach of half space depth to solve this problem.

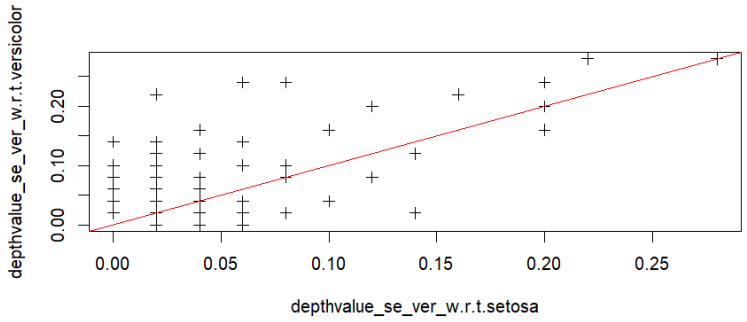
$$\mathrm{HD}(F; x) = \inf_{H}\{P(H) : H \text{ is a closed half-space in } \mathbb{R}^d \text{ and } x \in H\} \tag{1}$$

Lets say we wish to know whether the data based on setosa and versicolor are from same distribution or not.In order to bring out scale differences, the center of the samples should be equalized first by subtracting from the data their respective center. We then combine the data based on setosa and versicolor and lets call it combined data. We now calculate

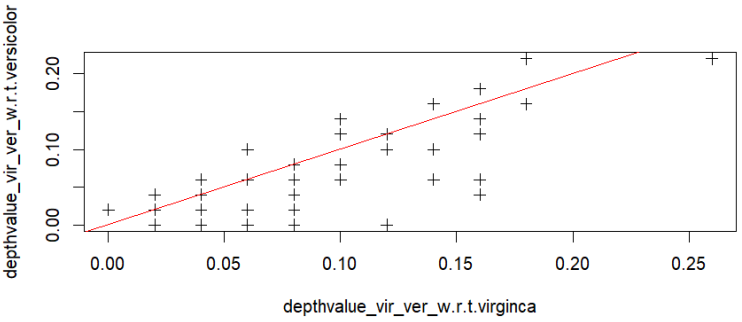
the half space depth of setosa data with respect to the combined data and do the same for versicolor too. If the two half space depths we just calculated are equal then we can say that the distributions are also same. We verify this fact using DD plot and see that they are actually from two different distributions. We perform the same technique for versicolor and virginica and see that they are also from two different distributions. So the three multivariate data are actually not from same distribution.



(a) DD Plot for Setosa and Virginica



(b) DD Plot for Setosa and versicolor



(c) DD Plot for versicolor and Virginica

Figure 1: DD Plot for Setosa and Virginica, Setosa and Versicolor, and Versicolor and Virginica

Interpretation : Roughly speaking, for two given multivariate samples, its DD-plot represents the depth values of those sample points with respect to the two underlying distributions, and thus transforms the samples in any dimension to a simple two-dimensional scatter plot. Let $\{X_1, \ldots, X_m\} \equiv X$ and $\{Y_1, \ldots, Y_n\} \equiv Y$ be two random

samples respectively from distributions F and G which are defined on $\mathbb{R}^d$. The DD-plot is defined as

$$DD(F, G) = \{(DF(x), DG(x)), x \in X \cup Y\}.$$

If both F and G are unknown, the DD-plot is then defined as

$$DD(F_m, G_n) = \{(DF_m(x), DG_n(x)), x \in X \cup Y\}.$$

The DD-plot was first introduced by Liu et al. (1999) for graphical comparisons of two multivariate distributions or samples based on data depth. It is always a two dimensional graph regardless of the dimensions of the samples. Therefore, DD-plots can provide simple diagnostic tools for visual comparisons of two samples of any dimension.

Conclusion : So three multivariate data are not from same distribution.
We can observe DD plot that setosa has higher kurtosis than virginica and versicolor and versicolor has higher kurtosis than virginica.

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, . . . , 9. Using those contours, describe various features of the data set.

Answer : About The Dataset:We choose the iris data set from the R package.We select petal length and petal width and use it as a bivariate data set for further analysis. vspace10pt A contour plot, also known as a contour map or contour chart, is a graphical representation of three-dimensional data in two dimensions. It is commonly used in fields such as mathematics, engineering, geophysics, geography, and the physical sciences to visualize the variation of a function or a dataset over a specific region.

In a contour plot, the data points are represented on a two-dimensional plane, and contours are used to depict regions of constant values. These contours are lines that connect points with the same data value, creating a map of isovalues.

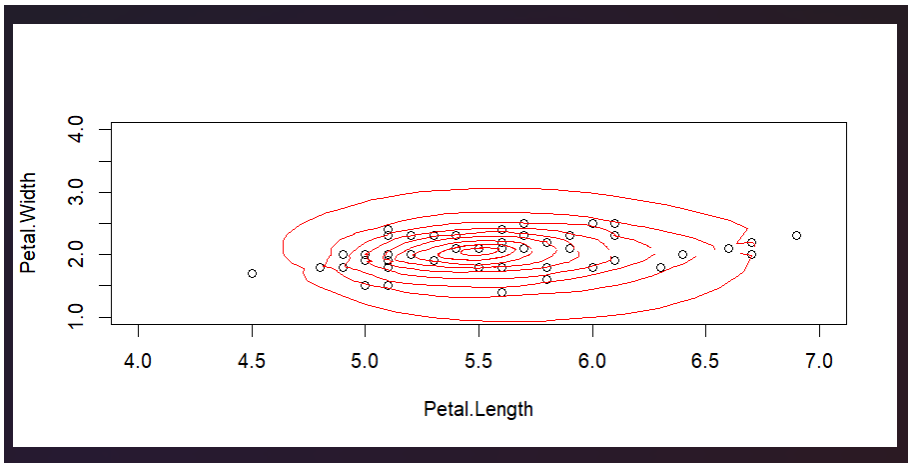


Figure 2: Contour plot of petal length and petal width of Iris data.

Interpretation : A quantile contour plot provides a visual representation of the

distribution of data at different quantiles. In a quantile contour plot, the contours represent regions of equal probability density, and each contour line encloses a specific percentage of the data. The contour lines represent areas of different data densities. Closer contour lines indicate higher density, while wider gaps between lines indicate lower density. So in our plot, we see that at the center contour lines are closer to each other which indicates regions of higher density, and as we go away from the centre gap between the contour lines increases which indicates regions of lower density. If the lines are more spread out in one direction, it indicates higher variability or dispersion in that dimension. So using this fact we see that in the plot there is more spread in the right direction compared to the opposite direction. so the data is positively skewed. We see from the plot there are some outliers in our data set as they fall outside of contour lines. Outliers may appear as isolated points or regions beyond the highest or lowest contour lines. These points can be of interest as they may represent unusual observations in the dataset and they are clearly visible in the plot.Normal distributions are symmetric, so the contour lines will form concentric circles or ellipses around the mean. The center of the plot corresponds to the mean of the distribution.We also find some similar evidence in our data set.

Bagplot: A Bagplot (also known as a Bivariate Boxplot) is a graphical method for visualizing the spread and location of a bivariate dataset. It combines elements of boxplots and scatterplots to provide insights into the distribution and structure of paired observations.The inner polygon, called the bag, is constructed on the basis of Tukey depth, the smallest number of observations that can be contained by a half-plane that also contains a given point.It contains at most 50 percent. of the data points.

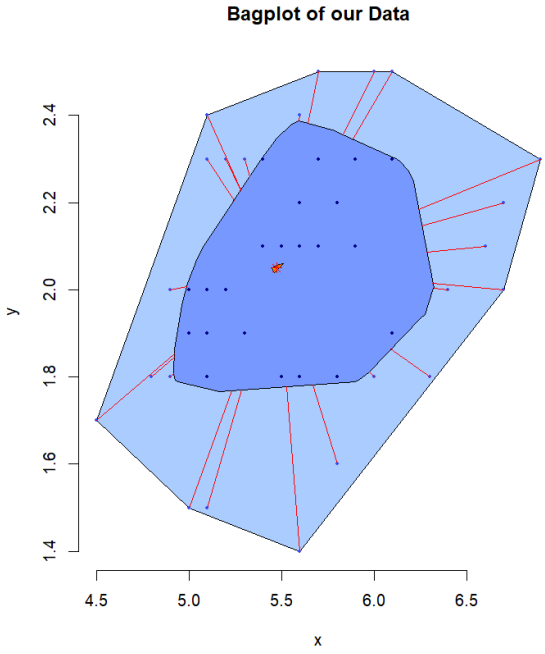


Figure 3: Bag Plot of the data.

Interpretation : Points outside the outer bag are identified as potential outliers. The bagplot provides a visual representation of these outliers, making it easier to identify

extreme observations in the dataset. A larger bag suggests higher variability, while a smaller bag indicates lower variability.So there is higher variability in our data set.The orientation of the bagplot can reveal the correlation between the two variables. A bag that is stretched along a diagonal suggests a correlation. So there is a positive correlation between the two variables.

Conclusion : The plot shows that the petal length and petal width are positively correlated. This means that there is a tendency for flowers with longer petals to also have wider petals, and vice versa. The contours are roughly elliptical, which suggests that the data is approximately bivariate normal.
Overall, the quantile contour plot provides a useful way to visualize the relationship between petal length and petal width in iris flowers. It shows that the two variables are positively correlated and that the data is approximately bivariate normal.