Assignment 1

Lakshika, Mrunal Dhiwar, Ringan Majumdar, Rishiraj Sutar, Sandeep Parmar

2024-01-24

Question 1

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

Data Description

The Iris dataset consists of measurements of four features of three species of Iris flowers, namely:

1. Sepal Length: The length of the iris flower's sepal (the outermost whorl of a flower) in centimeters.
2. Sepal Width: The width of the iris flower's sepal in centimeters.
3. Petal Length: The length of the iris flower's petal (the innermost whorl of a flower) in centimeters.
4. Petal Width: The width of the iris flower's petal in centimeters.

These measurements are taken from 150 different iris flowers, with 50 samples from each of three different species: **Setosa, Virginica** and **Versicolor**, **.**

We need to check whether the observations from the different species come from the same distribution or not.

What is Data Depth and Half Space Depth?

Data depth induces a data-dependent ordering of $\mathbb{R}^d$ in the center-outwards sense — points of high depth form the center of the dataset and low depth points occupy the data cloud's outskirts.

HALFSPACE DEPTH

The halfspace depth of $x \in \mathbb{R}^d$ w.r.t $F \in \mathcal{P}(\mathbb{R}^d)$ is

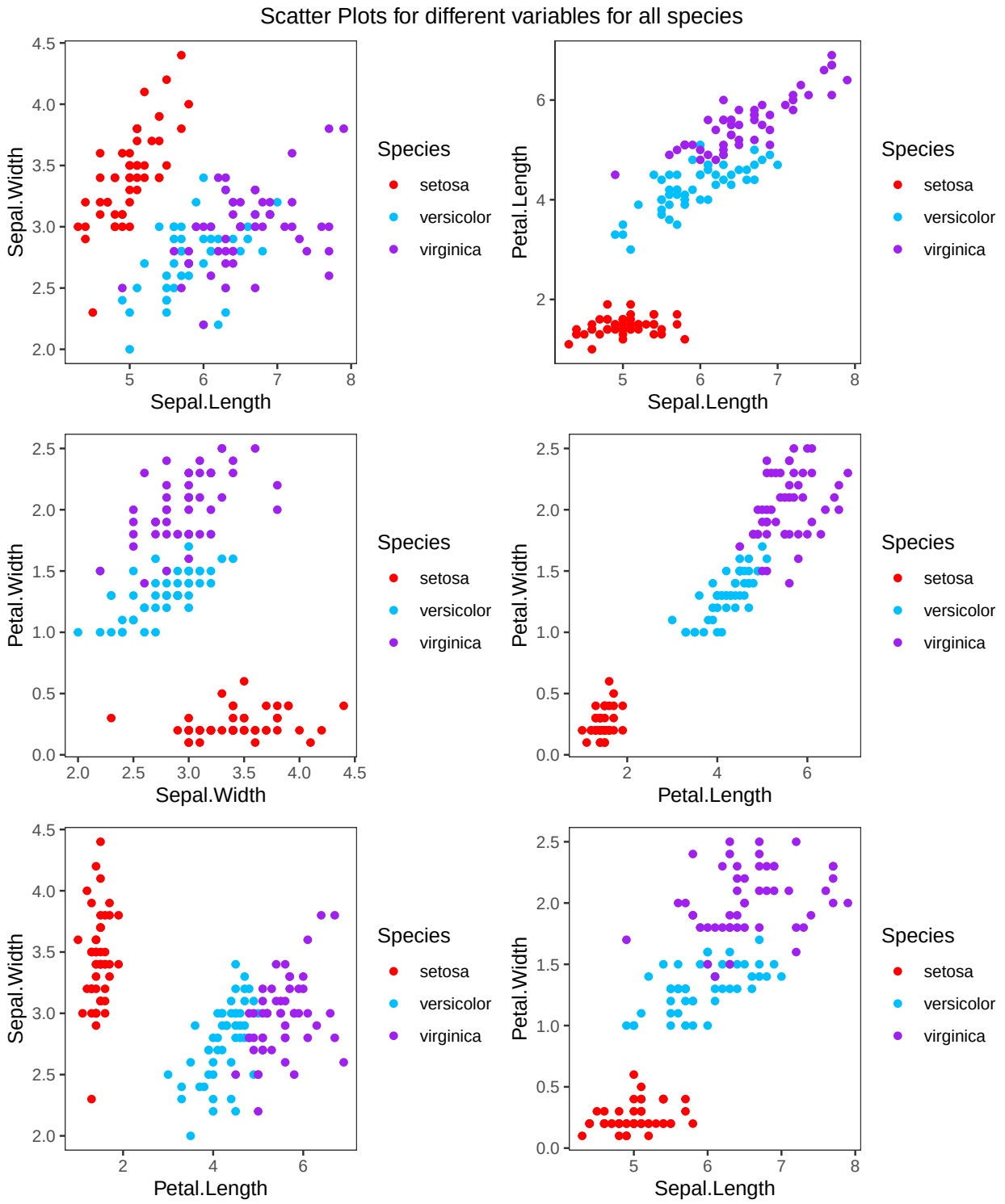
$D(x; F) = \inf_{u \in \mathbb{S}^{d-1}} P(H_{x,u})$,

where $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : ||x|| = 1\}$ is the unit sphere in $\mathbb{R}^d$, and $H_{x,u} = \{y \in \mathbb{R}^d : \langle\ x - y, u\ \rangle \leq 0$ is the closed halfspace whose boundary hyperplane passes through $x \in \mathbb{R}^d$ with inner unit normal $u \in \mathbb{S}^{d-1}$. In other words, the halfspace depth of a point x is the minimum P-mass of a halfspace that contains x.

Methodology

The approach we have taken to solve this problem is using data depth, more specifically half space depth. We have multivariate data for three different species. So, we have focused upon the graphical comparisons . Hence we have made pairwise DD plots for three different species and then commented about whether they are obtained from same populations or not.

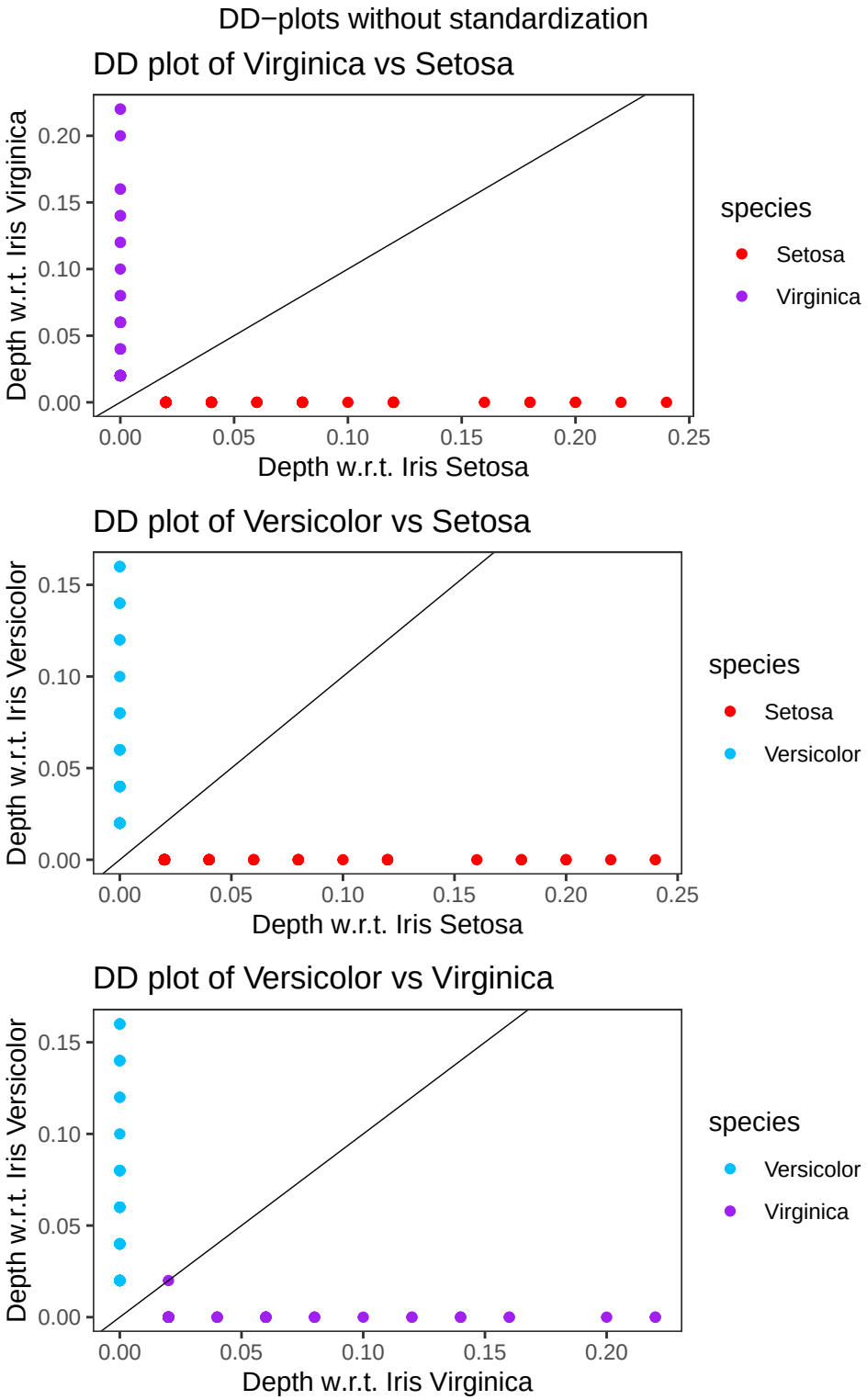
Plots



These plots are all possible scatter plots for the 4 different variables i.e Sepal Length, Sepal Width, Petal Length, Petal Width for the three different species. Clearly observe that the centres of the pairwise observations are far away from each other. Moreover, the nature of the spread of the observations differs in different

species.

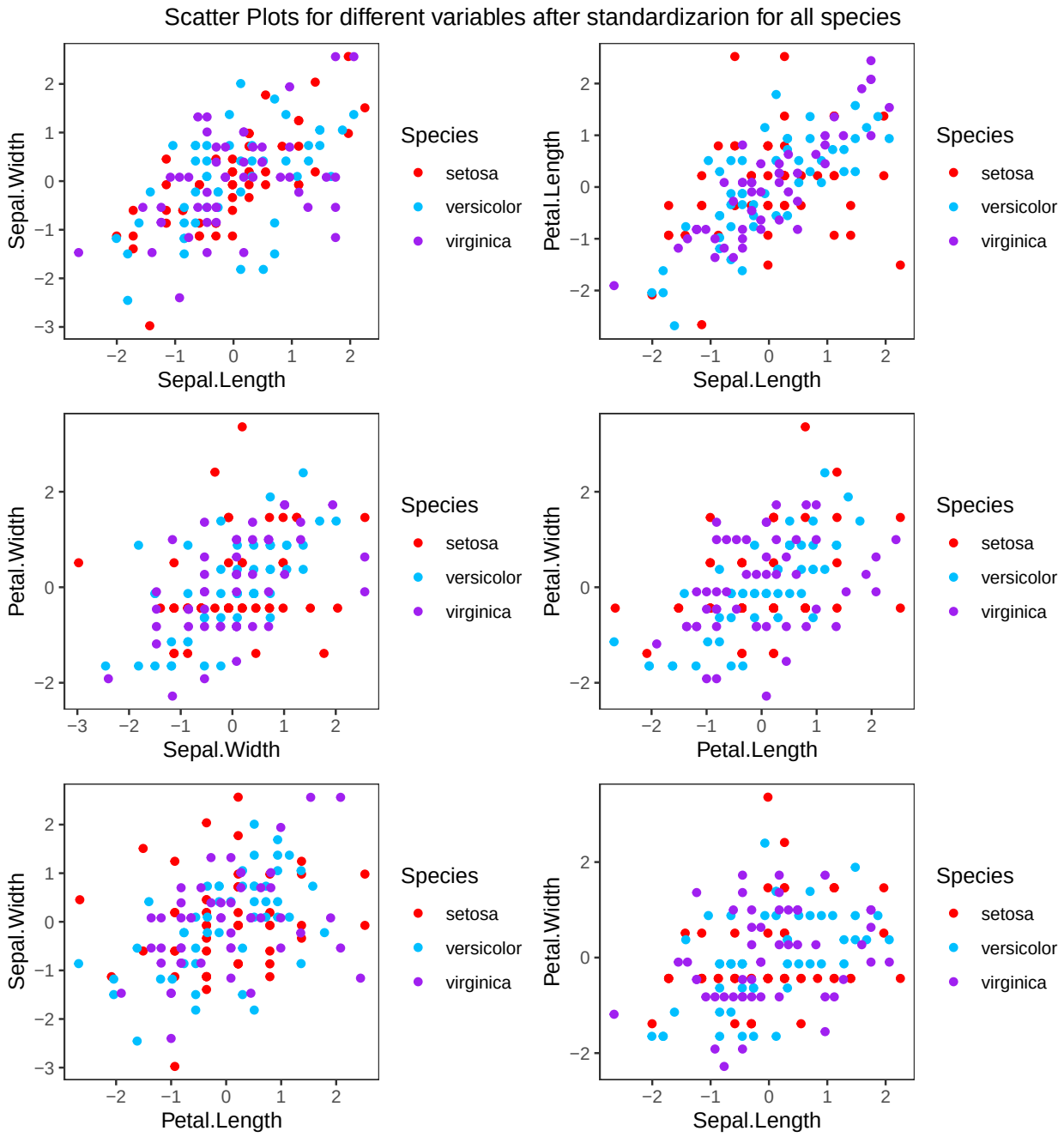
Now, let us make the DD-plots for the raw data.



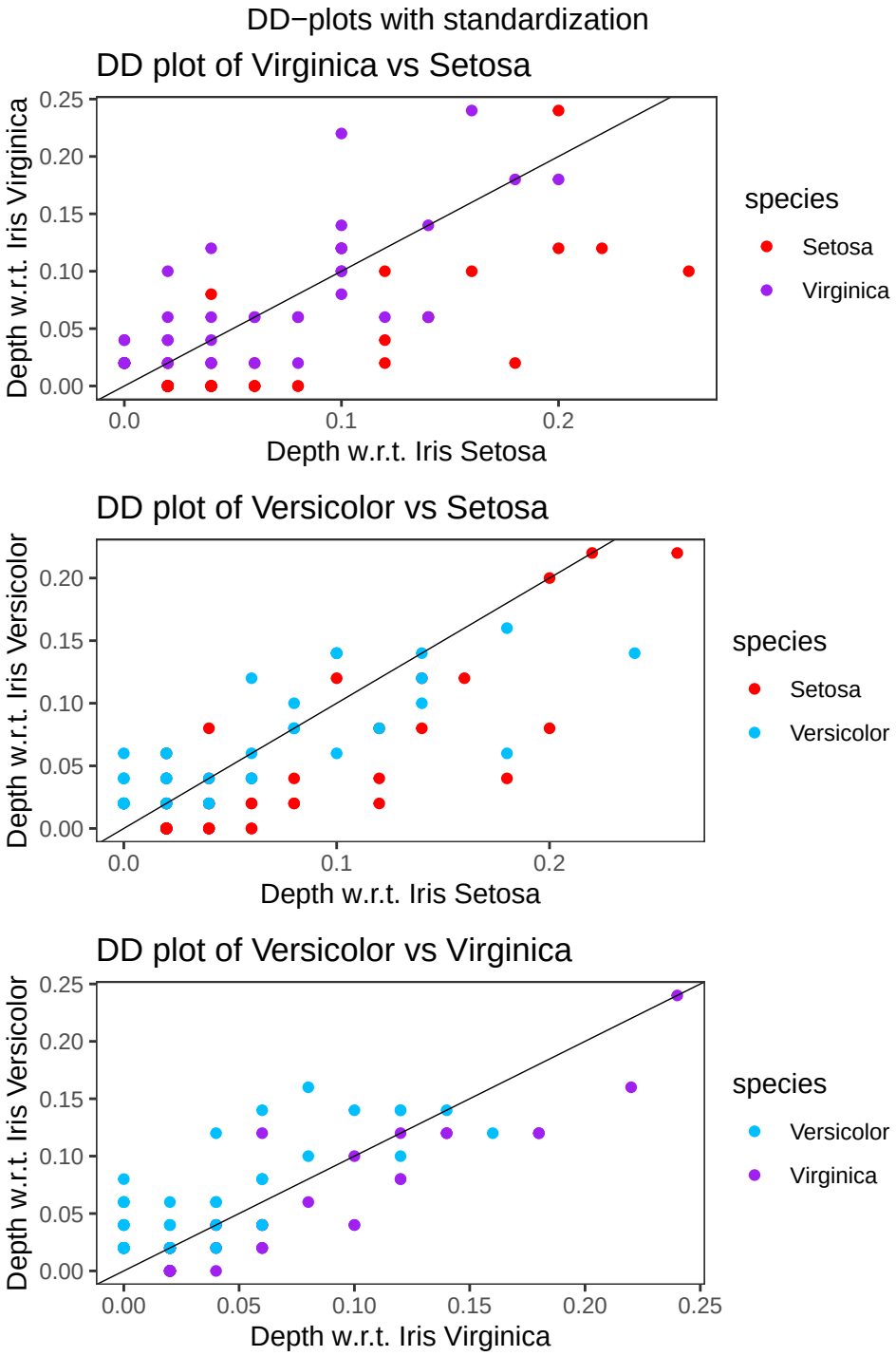
Observe that the plots are L-shaped in nature. On the other hand, for each of the plot the data depth of one

species w.r.t. the other species lies above or below the $y = x$ line. This implies that each of the sample for one species is very far away from the other sample for the other species. In other words, there is a significant difference in location for all the three species.

So, this might be the case that the data is obtained from same distribution but the plots are like this because of the location difference. Now, let us visualise the data after standardization of all the three clusters, namely, setosa, virginica and versicolor.



Clearly, after standardizing the data, all the three species are clustered around the same space. Hence, we have done the DD-plots again after standardizing the data.



There is a significant change in the nature of DD plots after standardization. The points now fall more or less along the $y = x$ line but still the data clouds are not aligned to a satisfactory extent.

Conclusion

Since, the points don't fall along the $y = x$ line even after standardizing, we conclude that the data for different species don't come from the same distribution.

Question 2

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

Data Description

The dataset used in this assignment has been taken from the National Institute of Diabetes and Kidney Disease. It contains information about 200 patients across 4 variables. The different variables are:

1. Age of Person
2. Glucose concentration in Blood
3. Diastolic Blood Pressure
4. Insulin level

For our assignment, we have considered a bivariate data by choosing Glucose concentration and Blood Pressure as our variable of interest.

What are Quantile Contours?

Multivariate Quantiles: Consider the bivariate sample $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, where $\mathbf{x}_i \in \mathbb{R}^2$.

For $\mathbf{u} \in \mathbb{R}^2$ such that $||\mathbf{u}|| < 1$, the quantile $q_{\mathbf{x}}(\mathbf{u})$ is defined as:

$$q_{\mathbf{x}}(\mathbf{u}) = \text{argmin}_{\mathbf{q}} \{E[||(\mathbf{X} - \mathbf{q})|| + \langle \mathbf{u}, \mathbf{X} - \mathbf{q} \rangle - ||(\mathbf{X})|| - \langle \mathbf{u}, \mathbf{X} \rangle]\}$$

Here, $\langle \mathbf{u}, \mathbf{X} - \mathbf{q} \rangle$ denotes the dot product between $\mathbf{u}$ and $\mathbf{X} - \mathbf{q}$ and $\langle \mathbf{u}, \mathbf{q} \rangle$ denotes the dot product between $\mathbf{u}$ and $\mathbf{q}$.

Methodology

To calculate the Quantile points for a fixed u, we have followed the following Steps:

Step 1:

For each $1 < i < n$, check whether the condition

$$\left\| \sum_{j=1, j \neq i}^{n} \frac{\mathbf{X}_j - \mathbf{X}_i}{\|\mathbf{X}_j - \mathbf{X}_i\|} + (n-1)\mathbf{u} \right\| \leq (1 + \|\mathbf{u}\|)$$

satisfies. If it satisfies for some $1 < i < n$, then $Q_{\mathbf{X}}(\mathbf{u}) = \mathbf{x}_i$.

```
step1 <- function(X,mu,n){
  Q <- numeric(length = 2)
  for(i in 1:n){
    if(lhs(X,i,as.matrix(mu),n)<=rhs(mu)){
      Q <- X[i,]
      break
```

```
    }
  }
  if(Q[1]==0 & Q[2]==0){return(c(FALSE,Q))}
  else{return(c(TRUE,Q))}
}
```

Step 2:

To calculate the Quantile points for a fixed $\mathbf{u}$, we need to solve the following equation if the condition in Step 1 is not satisfied:

$$\sum_{i=1}^{n} \frac{\mathbf{X}_i - \hat{Q}_n(\mathbf{u})}{||\mathbf{X}_i - \hat{Q}_n(\mathbf{u})||}\{\mathbf{X}_i - \hat{Q}_n(\mathbf{u})\} + n\mathbf{u} = 0$$

Here, $\hat{Q}_n(\mathbf{u})$ represents the estimated Quantile point for the fixed $\mathbf{u}$.

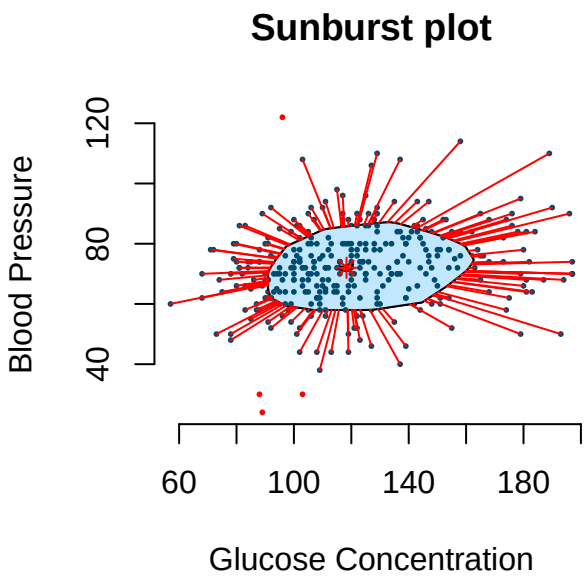
This is an optimization problem, and we have used the **optim** function to do the same.

For ||ui|| = i/10, we have repeated the above algorithm R=25 times using different values of u. Finally, we have connected the corresponding quantile points to obtain the quantile contour.

```
step2 <- function(X,n,u){
  iter <- 1
  Qo <- c(runif(1),runif(1))
  Xnorms <- numeric(length = n)
  for(i in 1:n){
    Xnorms[i] <- norm(X[i,]-c(2,5),type = "2")
  }
  Qmax <- max(Xnorms)
  Qoptim <- optim(par = Qo,fn=eqn,X=X,n=n,u=u,method = "BFGS")
  par <- Qoptim$par
  while( iter<20 & norm(par-c(2,5),type = "2")>Qmax){
    Qo <- c(runif(1),runif(1))
    Qoptim <- optim(par = Qo,fn=eqn,X=X,n=n,u=u,method = "BFGS")
    par <- Qoptim$par
    iter <- iter +1
    #print(iter)
  }
  return(par)
}
```

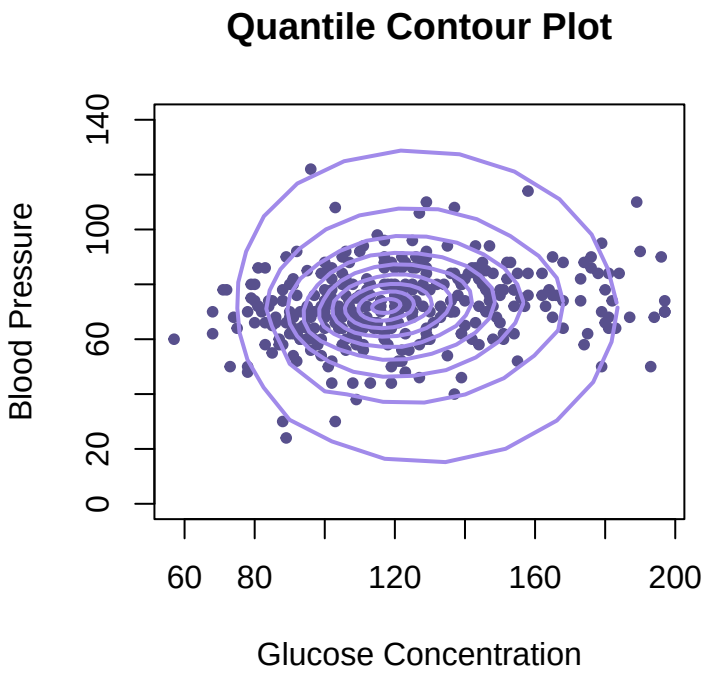
Plots

The centre-outward ordering induced by a data depth immediately gives rise to a simple graphic technique for presenting bivariate data sets. This technique can be viewed as a generalization of the univariate box-plot or box-and-whiskers plot. The plot resembles the sun with its rays radiating in all directions, and is thus named the sunburst plot.



Interpretation: The spikes that extend far from the central ellipse represent outliers in the data. These are values that are not consistent with the majority of the data points and might be due to various factors such as measurement errors or unusual observations that do not follow the general trend.

Then, we have done quantile contour plots. The contours are more closely spaced together around the mean. As we move away from the centre, the contours become more spaced out. The contour plot can also be used to identify outliers, clusters, and trends in the data.



Interpretation: We can see that there are some outliers in the lower left and upper right corners of the plot, where the data points are far away from the contours. The contour plot can also be used to compare the data with a theoretical model or a fitted curve. For example, we can see that the contours are roughly elliptical in shape, which suggests that the data might follow a bivariate normal distribution.