Student ID: 20203001

Statistical Learning with High-Dimensional Data

Exercise 1: General Questions

Question 1: Describe the general setup of resampling techniques and explain how it can be used for parameter tuning.

Answer 1:

Resampling techniques are a range of approach to estimate the precision of a statistic or a method to repeat sampling from a given population. The main techniques are: Bootstrapping, Permutation resampling, Cross Validation and Monte Carlo. Resampling techniques are firstly consisted of repeatedly drawing samples from a training set. For example, the Bootstrapping is a resampling approach in which several smaller samples of the same size.

Question 2: Describe some techniques allowing to select the number of clusters with k-means and the hierarchical clustering.

Answer 2:

Determining the optimal number of clusters is a relevant step in any clustering method and there exist several methods of determination. Two of these methods that will be described are the elbow method and the gap statistic method. These include direct methods and statistical testing method. Moreover, directs methods are based on criterion optimization (criterion like cluster sums of squares). While statistical testing is based on comparing evidence to null hypothesis.

- Elbow method: this method manipulates the total WSS as a function of the number of clusters. This can be applied to the k-means and the optimal number of clusters can be computed as follow:

Elaborate the k-means clustering algorithm for different values of k

Find the total WSS for each k

Plot the curve of WSS against the number of clusters k

Look for a bend on the plot, this point is susceptible to be the optimal number of clusters.

- Gap statistical method: this approach can be applied to the hierarchical clustering and even any other clustering method. In this approach, there is a comparison of the within intra-cluster variation for the chosen k, to their expected values under null reference distribution of the data. Therefore, the optimal number of clusters is the value that maximizes the gap statistic. This approach is elaborated as follow:

Cluster the observed data, vary the number of clusters and calculate the corresponding total within intra-cluster variation

Generate a reference dataset with a random uniform distribution

Compute the standard deviation of the statistics under the null hypothesis

The optimal number of clusters is the smallest value of k for which the gap statistic is within one standard deviation of the gap.

Exercise 2: Hierarchical clustering

Question 1: Explain in a few sentences how can relate the quality of a clustering with the notion of variance

Answer 1:

Variance-based distributed clustering

In a distributed algorithm based on a variance constraint, the data are clustered locally and independently from each other. Although few statistics of the local clustering are carried to the aggregation procedure, they operate the global clustering too, which resume labelling between the sub-clusters by means and perturbation. The algorithm shows a coherent performance about the identification of separated clusters and the real structure of the dataset. In addition, the algorithm finds the number of clusters automatically and this is a relevant solution to the estimation of the number of clusters.

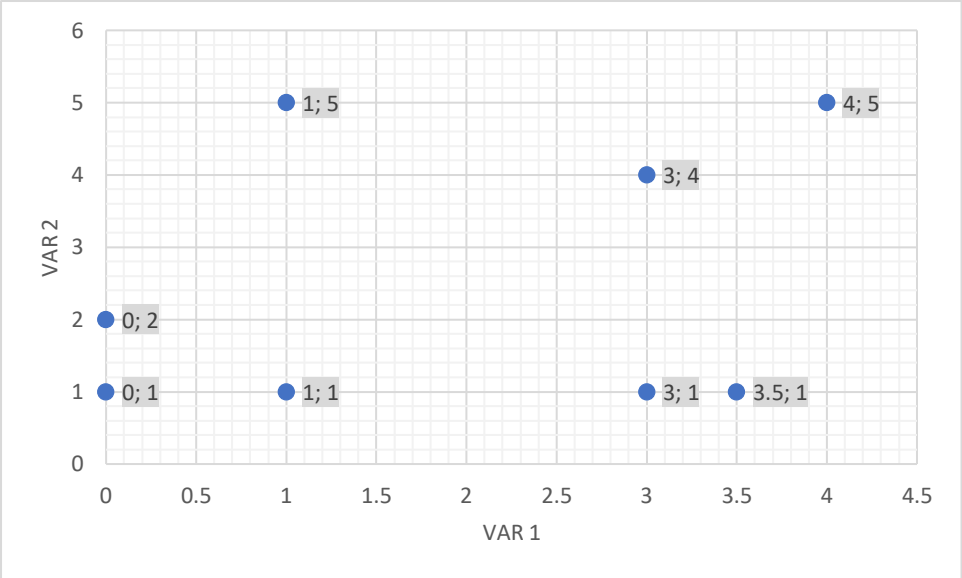
Question 2: We propose to use the hierarchical clustering with the complete linkage (distance of the maximum). Apply this algorithm to the above dataset and draw the associated dendogram (All calculations must be detailed).

Answer 2:

Dataset

Indiv.	Var 1	Var 2
X1	0	1
X2	0	2
X3	1	1
X4	3	1
X5	3,5	1
X6	1	5
X7	3	4
X8	4	5

There is a graphical representation of the dataset on a 2D plot:



Based on the Euclidean distance we will create a distance matrix.

The Euclidean distance is computed as follow:

Distance(X1;X2) = $\sqrt{(0-0)^2+(1-2)^2}$ = 1. (example of distance from X1 to X2)

We will then compute the 27 remaining distances and put them in a table in order to make it more understandable.

	X1	X2	X3	X4	X5	X6	X7	X8
X1								
X2	1	0						
X3	1	1,41421356	0					
X4	3	3,16227766	2	0				
X5	3,5	3,64005494	2,5	0,5	0			
X6	4,12310563	3,16227766	4	4,47213595	4,71699057	0		
X7	4,24264069	3,60555128	3,60555128	3	3,04138127	2,23606798	0	
X8	5,65685425	5	5	4,12310563	4,03112887	3	1,41421356	

First Distance Matrix

From the first distance matrix, we can observed that the minimum distance in between X4 and X5, which means (X4 ; X5) would form our first cluster.

as (X4 ; X5) now forms a single element we will for example consider the Max(distance ((X4 ; X5) ; X1)).

This is: Max(distance ((X4 ; X5) ; X1)) = Max(distance ((X4 ; X1); (X5; X1)) = Max(3;3.5) = 3.5

We are considering the Max function because the clustering is done under a complete linkage. Moreover, if we apply the max function for the remaining elements (X2, X3, X6, X7 and X8), we can thus update the distance matrix to:

	X1	X2	X3	X4,X5	X6	X7	X8
X1	0						
X2	1	0					
X3	1	1,41421356	0				
X4,X5	3,5	3,64005494	2,5	0			
X6	4,12310563	3,16227766	4	4,47213595	0		
X7	4,24264069	3,60555128	3,60555128	3,04138127	2,23606798	0	
X8	5,65685425	5	5	4,12310563	3	1,41421356	0

The minimum distance observed being 1, we can group (X1;X2) to a new cluster.

as (X1 ; X2) now forms a single element we will for example consider the Max(distance ((X1 ; X2) ; (X4;X5)).

This is: Max(distance ((X1 ; X2) ; (X4;X5))= 3,64005494

Moreover, if we apply the max function for the remaining elements, we can thus update the distance matrix to:

	X1,X2	X3	X4,X5	X6	X7	X8
X1,X2	0					
X3	1,41421356	0				
X4,X5	3,64005494	2,5	0			
X6	4,12310563	4	4,47213595	0		
X7	4,24264069	3,60555128	3,04138127	2,23606798	0	
X8	5,65685425	5	4,12310563	3	1,41421356	0

A new cluster is ((X1,X2),X3)

Following an analogic reasoning, we can keep updating the distance matrix step by step:

	(X1,X2),X3	X4,X5	X6	X7	X8
(X1,X2),X3	0				
X4,X5	3,64005494	0			
X6	4,12310563	4,47213595	0		
X7	4,24264069	3,04138127	2,23606798	0	
X8	5,65685425	4,12310563	3	1,41421356	0

A new cluster is (((X1,X2),X3) ; X7)

Following an analogic reasoning, we can keep updating the distance matrix step by step:

	(X1,X2),X3	X4,X5	X6	X7,X8
(X1,X2),X3	0			
X4,X5	3,64005494	0		
X6	4,12310563	4,47213595	0	
X7,X8	5,65685425	4,12310563	3	0

A new cluster is ((X7;X8);X6)

Following an analogic reasoning, we can keep updating the distance matrix step by step:

	(X1,X2),X3	X4,X5	X6,(X7,X8)
(X1,X2),X3	0		
X4,X5	3,64005494	0	
X6,(X7,X8)	5,65685425	4,47213595	0

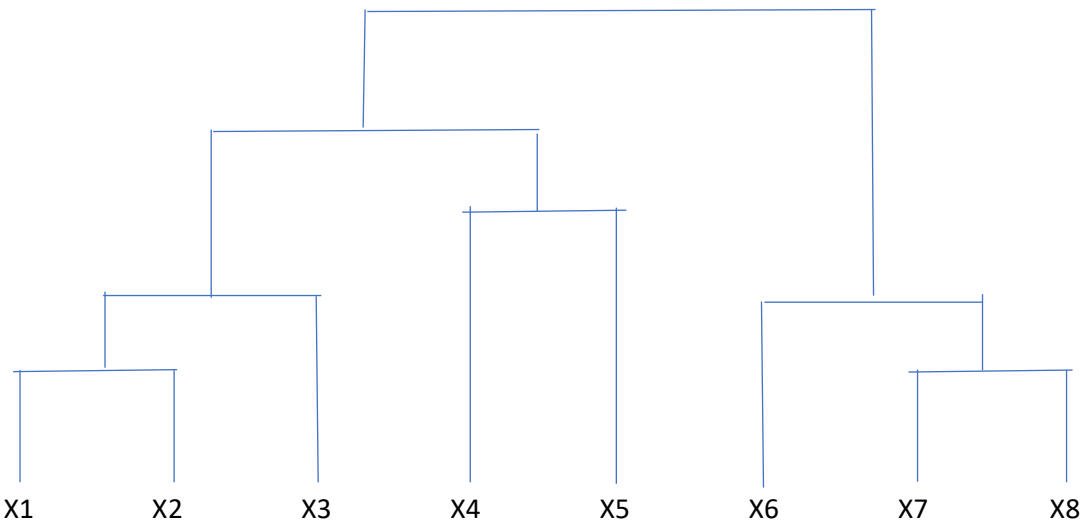
A new cluster is (((X1; X2);X3);(X4;X5))

Following an analogic reasoning, we can keep updating the distance matrix step by step:

	((X1,X2),X3),(X4,X5)	X6, (X7,X8)
((X1,X2),X3),(X4,X5)	0	
X6,(X7,X8)	5,65685425	0

The last cluster is ((((X1;X2);X3);(X4;X5)) ; ((X7;X8);X6))

The dendogram can be drawn as follow:



Exercise 3: The Vélib Data

1. **Loading the data**

See the attached python file

2. **Pretreatment and descriptive analysis**

the summary of the data has been shown and the most useful data were only #Position and #data, these will be processed and analyzed for the remaining part of the exercise.

See the attached python file

3. **Data visualization**

The data will be visualized by the Principal Component Analysis. From which 22 components axes have been chosen based on the rule of 90% or the visualization.

See the attached python file

4. **Clustering**

- **Hierarchical clustering**

See the attached python file

- **K-means**

See the attached python file

5. **Summary**

The two different types of clustering have been undergone on the same data. Taking care of not use some biased methodologies, we have found some similarities in terms of number of clusters (for example). Although the k-means might be a quicker approach technique for this exercise, considering the hierarchical clustering make the cluster look more readable and understandable throughout the dendogram.

See the attached python file