On Clustering Scheme for Kernel K-Means

Lekaa Ali Muhamed[1], Hayder Yahya mohammed[2]

[1] Assist. Prof. Department of Statistics, College of Management and Economics Baghdad University, Baghdad, Iraq
[2] PhD student. Department of Statistics, College of Management and Economics Baghdad University, Baghdad, Iraq

Abstract

Cluster analysis mainly concerned with dividing the number of data elements into clusters observation in the same cluster are homogeneous and are not homogeneous with other clusters، but in the case of nonparametric data it is not possible to deal with classic estimated because of obtaining misleading results This gave rise to adopt efficient estimation methods known as the kernel methods. One of the methods of clustering is Non-Hierarchical clustering aims to divide the dataset into (k) homogeneous cluster groups based on the idea of the central the tendency of the cluster group using (k) averages. There are many methods of non-hierarchical clustering, some depends on the arithmetic mean, and others depend on the mediator or mode.

Keywords: High-dimensional data, k means, kernel k means.

Corresponding Author:
Author Name, Hayder Yahya mohammed
Departement, Statistics
University #1, Baghdad University
Address. Baghdad, Iraq
haiderstatistic@yahoo.com

في مخطط عنقدة اسلوب K من المتوسطات اللبية•

أ.د. لقاء علي محمد[1] حيدر يحيى محمد[2]
[1،2]جامعة بغداد- كلية الادارة والاقتصاد- قسم الاحصاء

المستخلص

التحليل العنقودي هو اسلوب يهدف الى تقسيم بيانات متعددة المتغيرات الى مجاميع او عناقيد ، المشاهدات في العنقود نفسه تكون متجانسة وتكون غير متجانسة مع العناقيد. احد طرق التحليل العنقودي اللاهرمي هي طريقة عنقدة K من المتوسطات التي تهدف إلى تجميع المشاهدات المتماثلة مع بعضها البعض اعتمادا على خصائصها في K من العناقيد ، في حالة البيانات اللاخطية خوارزمية طريقة عنقدة K من المتوسطات تكون نتائجها مضللة وغير دقيقة وبالتالي نستعمل الطرائق اللبية بأسلوب طريقة Kernel K-Means للحصول على افضل الحلول.

الكلمات المفتاحية : بيانات عالية الابعاد ، عنقدة K من المتوسطات ، عنقدة K من المتوسطات اللبية.

• **بحث مستل من اطروحة دكتوراه**

Introduction

Cluster analysis is an un-supervised learning technique which aims to divide a set of data into clusters, or groups. Observations in the same group are similar to each other and the different groups are dissimilar. Clustering methods can be divided into two main basic types: partitional and hierarchical clustering. [elderly people pp579]
clustering algorithms aim to group a set of samples into several clusters such that samples from intra clusters are more similar to each other than samples from inter clusters. the most commonly used clustering methods in practice are k means. [Multiple Kernel k-Means Clustering by Selecting Representative pp1]
k means is an unsupervised learning algorithm partitions the data set into a selected number of clusters under some optimization measures.
Clustering is an important problem which is prevalent in a variety of real world problems , One of the first and widely applied clustering algorithms is k means, which was named by James MacQueen , but was proposed by Hugo Steinhaus even before , Despite being half a century old k means has been widely used and analyzed under various settings. One major drawback of k means is its incapability to separate clusters that are nonlinearly separated. This can be alleviated by mapping the data to a high dimensional feature space and do clustering on top of the feature space which is generally called kernel based methods. [On Robustness of Kernel Clusterinpp1]
kernel functions can be viewed as a nonlinear transformation , that increases the separability of the input data by mapping them to a new high dimensional space , The incorporation of kernel function enables the k means algorithm to explore the inherent data pattern in the new space.[12;pp.1]

2. Materials and methods

2.1. K Means

Suppose the data set has *M* samples $X_1, X_2, \dots X_M$ 1. K Means algorithm aims to partition the *M* samples into *K* clusters, $C_1, C_2, \dots, C_K$, and then returns the centre of each cluster, $m_1, m_2, \dots, m_K$, as the representatintves of the data set. Thus a *M*-point data set is compressed to a *K*point. The batch mode K-Means clustering algorithm using Euclidean distance works as follows [12;pp.2]:

1. choose K centroids at random.
2. make initial partition of objects in to K clusters by assigning objects to closest centroids.
3. calculate the centroids of each of the K clusters.
   I) for objective i calculate its distance to each of the centroids.
   II) allocate object i to cluster with closest centroids.
   III) if object was reallocated , recalculate centroids based on new clusters.
4. repeat step3 for object i=1 to M.
5. repeat step3 and step4 until no reallocation occur.
6. assess cluster structure for fit and stability. [9;pp.180]

2.2. kernel k means

basic idea of a kernel method is to convert an input feature space into ahigher dimensional feature space by a transformation function Clustering algorithms that use sum of squares error criterion perform poorly for data that are distributed in a way that the natural groups of data can be partitioned only nonlinear boundaries , Therefore by transforming the feature space into ahigher dimensional space, the naturally distributed groupings of data can now be partitioned by linear boundaries For this reason, a kernel (transformation) function needs to be defined , As mentioned above , a kernel function

is used to transform data into a higher dimensional feature space , Pattern xi in the input space that is converted into ahigher dimensional feature space by a transformation function can be expressed as $\varphi(xi)$, The transformation function $\varphi(\cdot)$ nonlinearly maps the input feature space of into the higher dimensional feature space. That is, $\varphi: Rp \rightarrow Rq, p < q$, The inner product of two values obtained by the transforming function defined as a kernel function.[5;pp.274]

Kernel K-Means Algorithms:

Step 1: Choose $\delta(xi, Ck)\ (1 \leq i \leq N, 1 \leq k \leq K)$ with the initial value, And form the initial cluster K C1, C2, ..., CK

Step 2: For each clusterCk, calculate $|Ck|$ and $g(Ck)$

Step 3: For each sample of exercise xi and Ck cluster, compute $f(xi, Ck)$ and then Find xi at Cluster nearby.

$$\delta(xi, Ck) = 1 \quad , f(xi, Ck) + g(Ck)$$

$$forall j \neq k$$

$$0, otherwise$$

Step 4: Repeat steps 2 and 3 to find.

Step 5: For each cluster $Ck.$, select the closest sample to the center as a representative of Ck.

$$mk = Arg\,min\, D(\Phi(xi), zk).Xi, where\ \delta(Xi, Ck) = 1$$ [1;pp.2]

3. Results and Discussion

Simulations were performed using the R program according to the following steps:

1. To generate the study data from the natural distribution polluted by using Box-Muller formula, We find observation for the variable (X_i) as follow:

$$X_i = (0.75)\,(N(0,1))\text{^}(2) + (0.25)\exp(N(-2,4))$$

0.75% of the variable is distributed (N (0, 1)) ^ (2) and 0.25% is distributed by exp(N (-2, 4)) and for all sample sizes with the same proportions.[6; pp112], include three experiments that differ in terms of number of variables and sample sizes , These experiments are: The first experiment (N = 15, p = 10), the second experiment (n = 22, p = 15) and finally the third experiment (n = 30, p = 20).

2. Finding the estimation of non-linear methods by the kernel matrix Using Laplacian. The calculation of the bandwidth was based on the Scott formula in calculating the smoothing parameter (h).

3. apply cluster analysis in two ways (K-Means) and (Kernel K-Means) And according to the steps shown in the theoretical side, note that the kernel functions used is Laplace. We will choose the best depending on the percentage of (between-SS / total-SS).

Table 1: Cluster analysis results

size	K	K means percentage	Kernel K means percentage
p=10 ,n=15	3	39.3 %	87.2 %
p=10 ,n=15	4	61.2 %	97.4 %
p=15 ,n=22	3	36.0 %	75.0 %
p=15 ,n=22	5	57.2 %	96.4 %
p=20 ,n=30	5	37.7 %	93.4 %
p=20 ,n=30	7	49.9 %	94.2 %

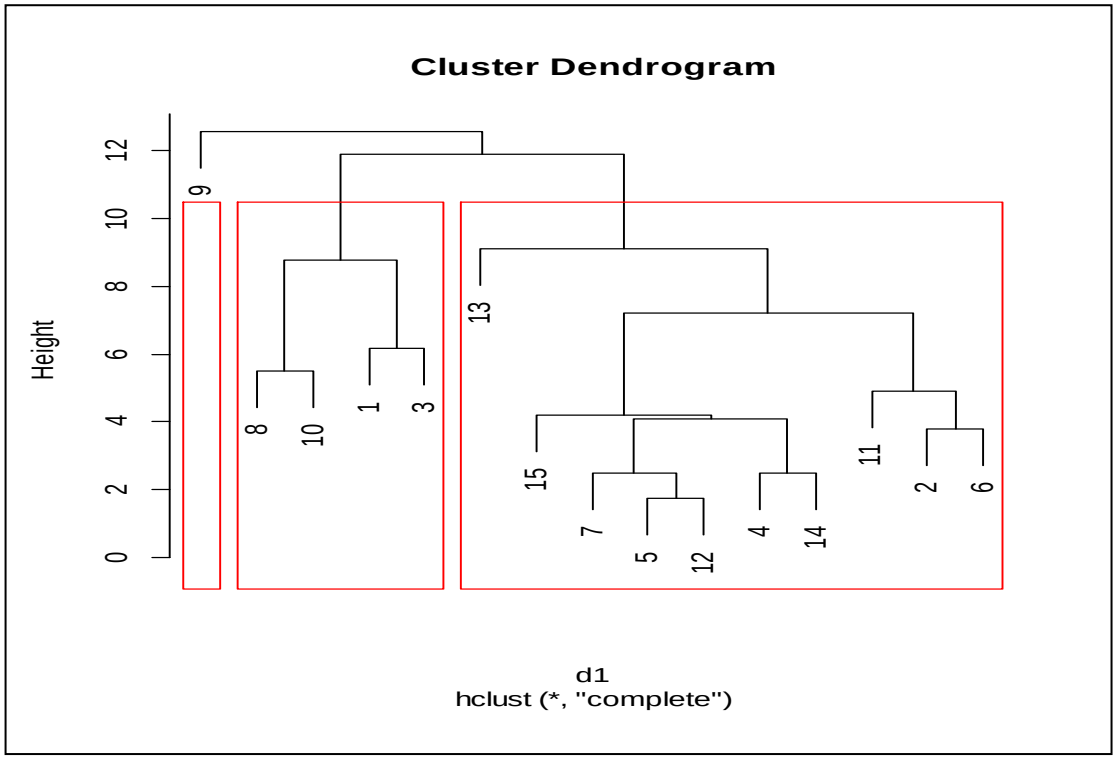

Figure 1: The distribution of observations using K means when n = 15, p = 10, k = 3

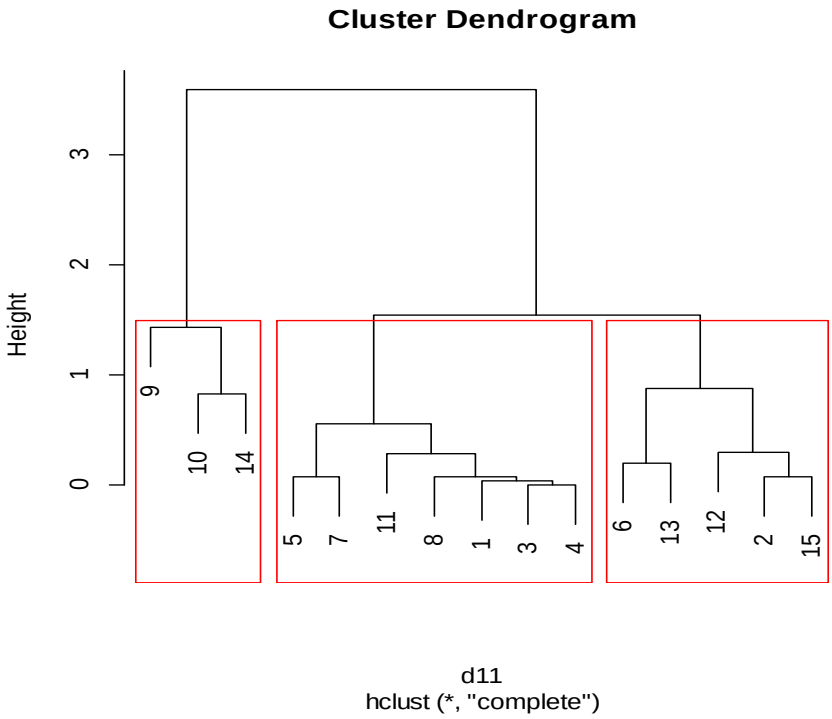

Figure 2: The distribution of observations using kernel K means when n = 15, p = 10, k = 3

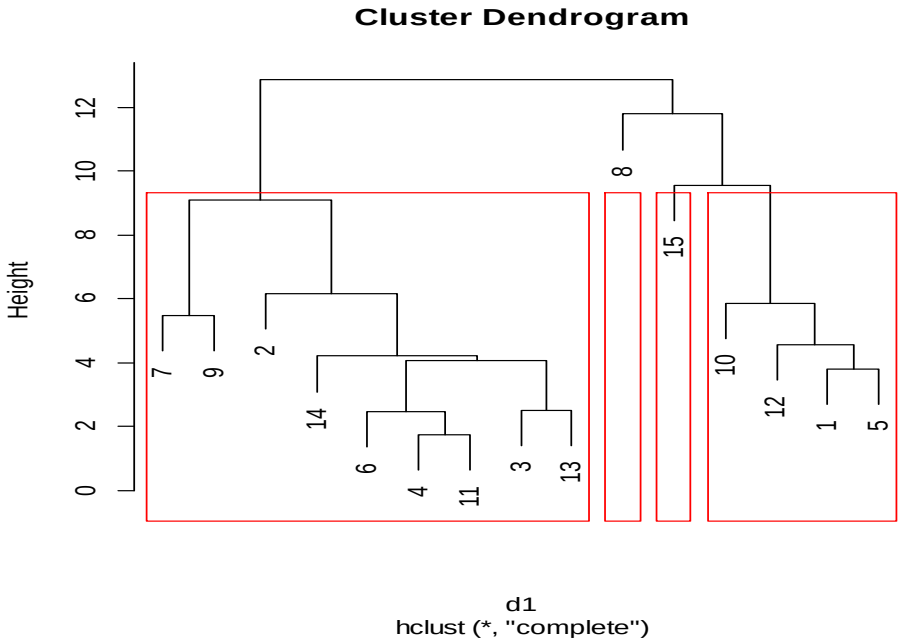

Figure 3: The distribution of observations using K means when n = 15, p = 10, k = 4

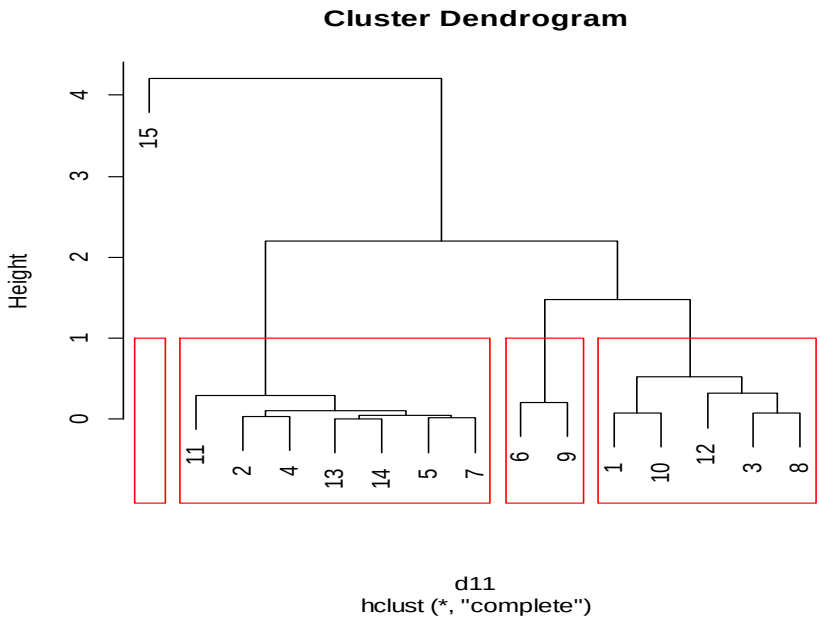

Figure 4: The distribution of observations using kernel K means when n = 15, p = 10, k = 4

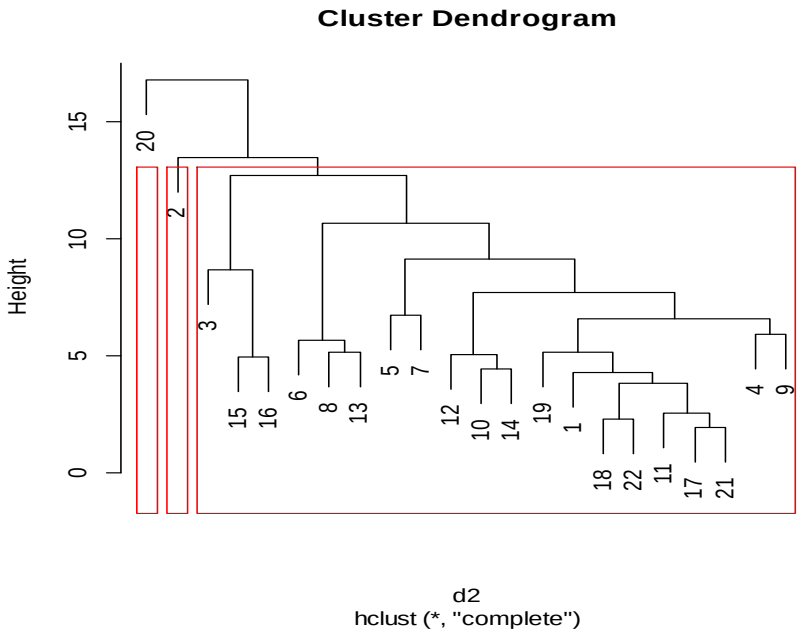

Figure 5: The distribution of observations using K means when n = 22, p = 15, k = 3

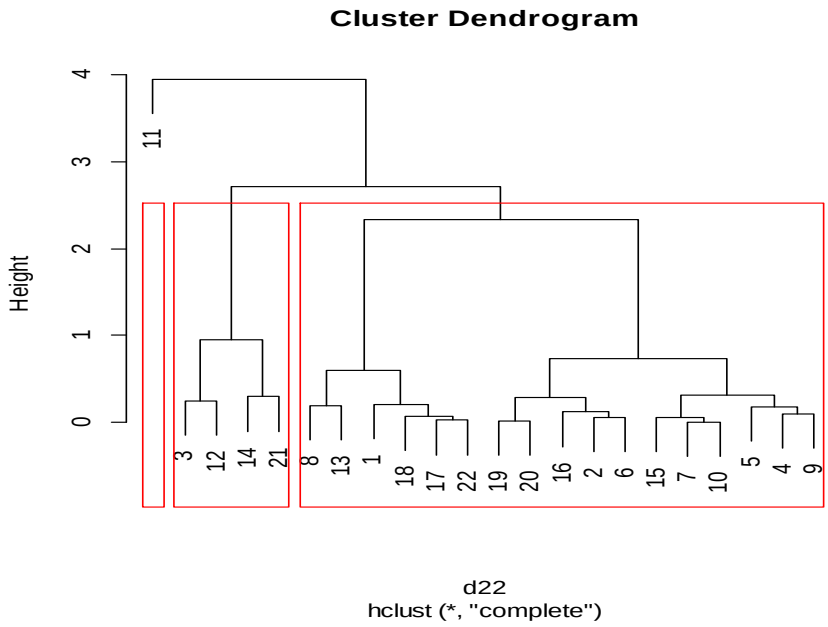

Figure 6: The distribution of observations using kernel K means when n = 22, p = 15, k = 3

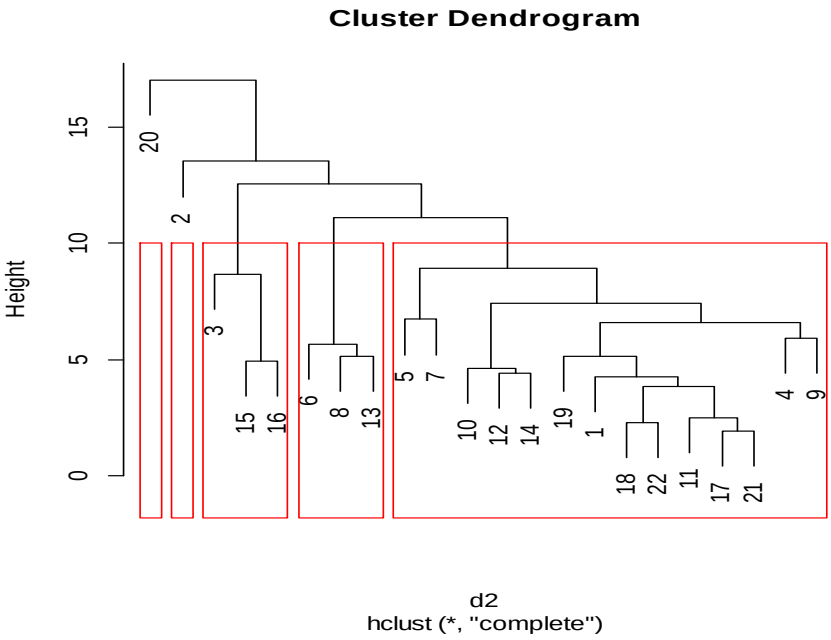

Figure 7: The distribution of observations using K means when n = 22, p = 15, k = 5

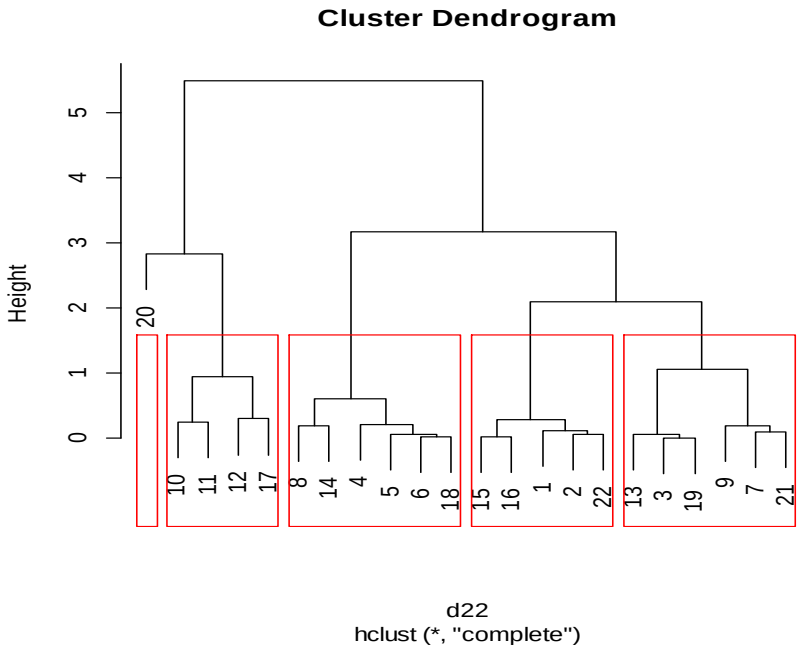

Figure 8: The distribution of observations using kernel K means when n = 22, p = 15, k = 5

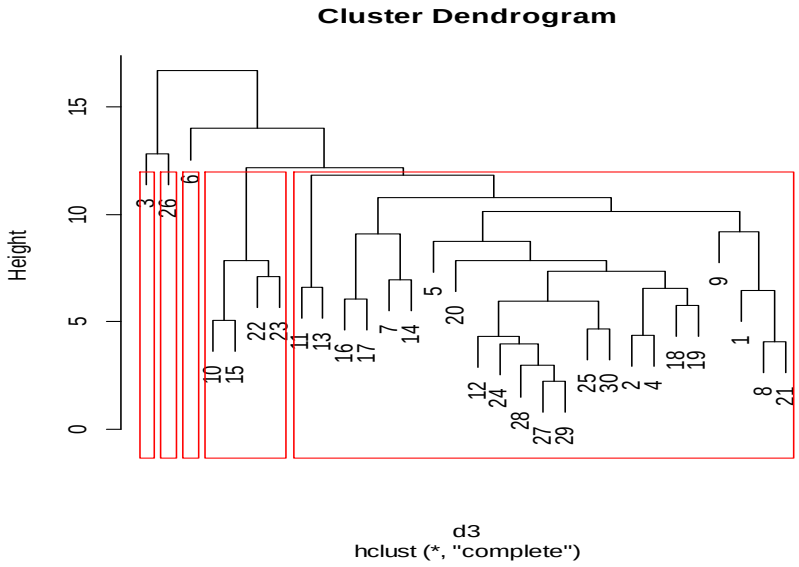

Figure 9: The distribution of observations using K means when n = 30, p = 20, k = 5

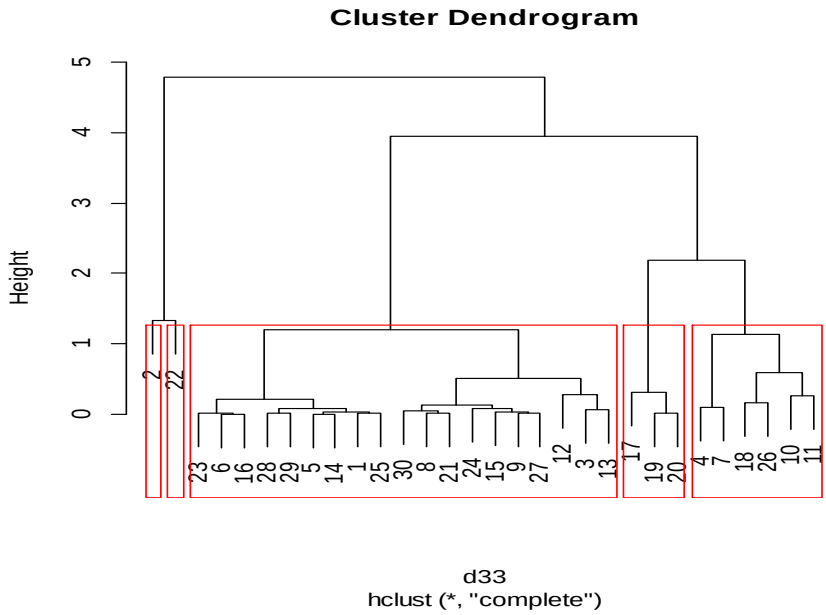

Figure 10: The distribution of observations using kernel K means when n = 30, p = 20, k = 5

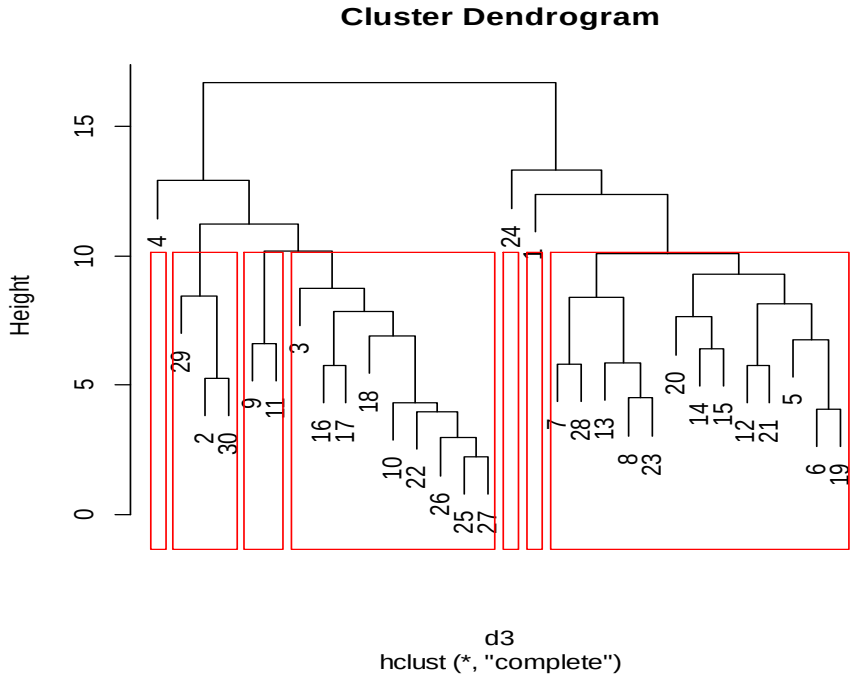

Figure 11: The distribution of observations using K means when n = 30, p = 20, k = 7

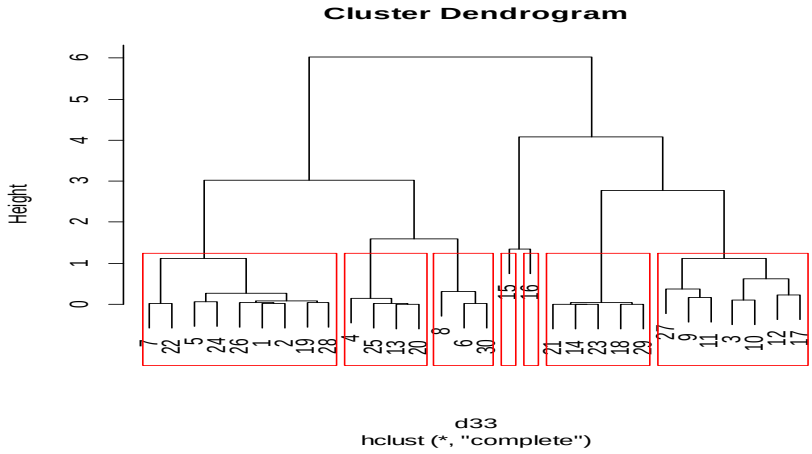

Figure 12: The distribution of observations using kernel K means when n = 30, p = 20, k = 7

Conclusion:

A multivariate data with higher dimensions using of k means will leading to miss understand results, In our simulations we found that the performance of kernel k means is batter than k means and kernel k means Of the percentage of (between-SS / total-SS) in all three sample of simulation. Because kernel k means using Laplace Kernel to solve nonlinear data.

References:

1. M. Alamsyah , A. R. T. H. A., et al. "The Classification of Diabetes Mellitus Using Kernel k-means." Journal of Physics: Conference Series. Vol. 947. No. 1. IOP Publishing, 2018.
2. T. Anderson, "An introduction to multivariate statistical analysis" second edition , john wile & sons , New York. (1984).
3. C. Combes, and J. Azema, "Clustering using principal component analysis applied to autonomy–disability of elderly people." *Decision Support Systems* 55.2 pp.578-586,(2013).
4. J. MacQueen, "Some methods for classification and analysis of multivariate observations". In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Vol. 1, No. 14, pp. 281-297, 1967.
5. F. Rhee, C. H., K. Choi, S., and B. Choi, I. Kernel approach to possibilistic C-means clustering. International Journal of Intelligent Systems, 24(3), pp.272-292, 2009.
6. C. Romesburg, Cluster analysis for researchers. Lulu. com. 2004.
7. B. Schölkopf, B. Smola, and K. Müller, R. "Nonlinear component analysis as a kernel eigenvalue problem". Neural computation, 10(5), pp.1299-1319, 1998.
8. B. Turlach, A. "Bandwidth selection in kernel density estimation: A review". In CORE and Institut de Statistique ,1993.
9. A. Venkatesan, and L. Parthiban,. "Clustering of datasets using PSO-K-Means and PCA-K-means". Int. J. Comput. Intel. Inf, pp.180-184, 2011.
10. K. Wu, L., and M. Yang, S. "Alternative c-means clustering algorithms. Pattern recognition", 35(10), 2267-2278, 2002.
11. S. Yu, L. Tranchevent, Liu X., W. Glanzel, J. Suykens A., B. De Moor, Y.Moreau. "Optimization data fusion for kernel k-means clustering. pattern analysis and machine intelligence" . vol.34,pp.131-139.IEEE. 2012.
12. R. Zhang, and A. Rudnicky, I. "A large scale clustering scheme for kernel k-means". In Object recognition supported by user interaction for service robots, Vol. 4, pp. 289-292. IEEE.7, 2002.