**SOME IMPORTANT NOTES ON CLUSTER ANALYSIS**

***Instruction:*** *Almost all questions are of experimental type. You might be asked to perform simulated experiments, validate concepts and results, etc. Make sure you support your conclusions with valid theory, valid experimental outputs, diagrams, illustrations, graphs and tables, proper codes, etc. You will be marked based on the clarity of your communication and presentation of the experimental results. Its must be like writing proper scientific notes.*

**Question 1**

*Read the context carefully:*

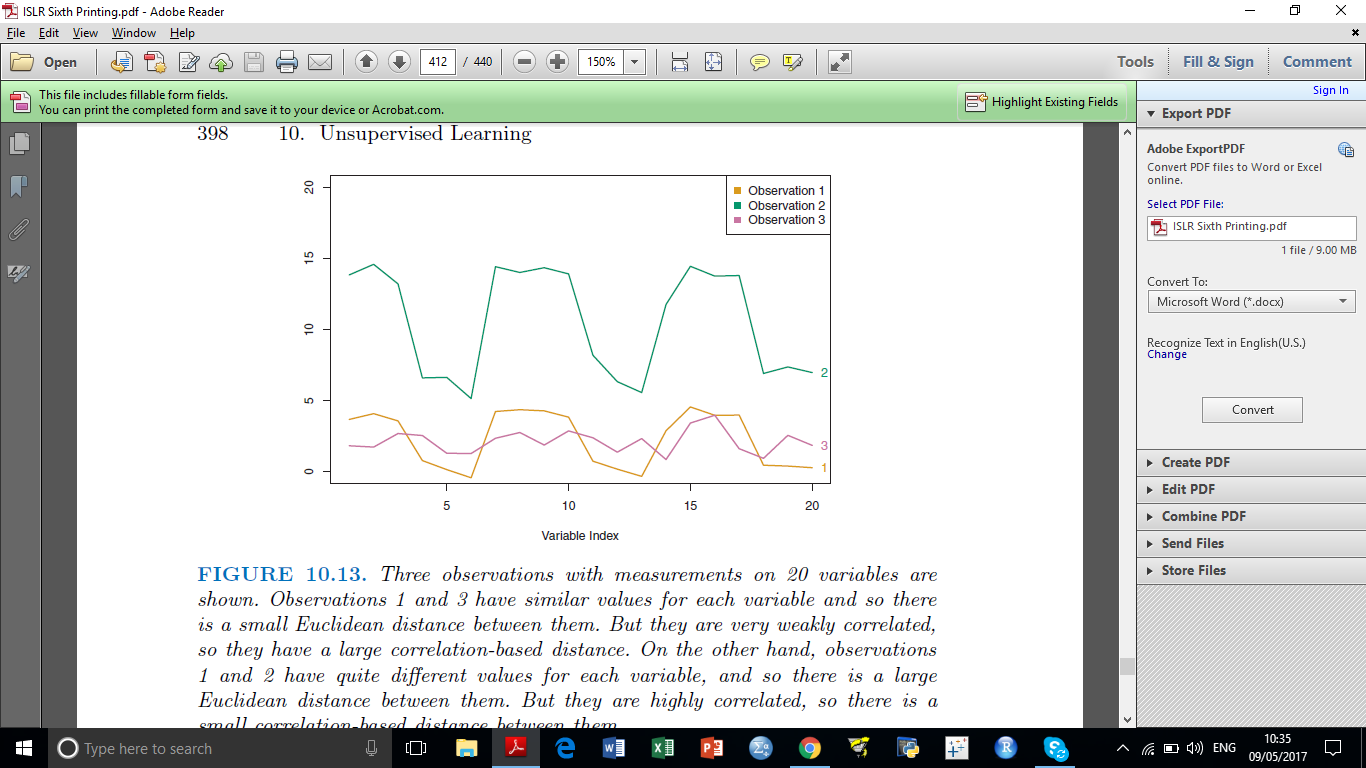
The term hierarchical refers to the fact that clusters obtained by cutting the dendrogram at a given height are necessarily nested within the clusters obtained by cutting the dendrogram at any greater height. However, on an arbitrary data set, this assumption of hierarchical structure might be unrealistic. For instance, suppose that our observations correspond to a group of people with a 50–50 split of males and females, evenly split among Americans, Japanese, and French. We can imagine a scenario in which the best division into two groups might split these people by gender, and the best division into three groups might split them by nationality. In this case, the true clusters are not nested, in the sense that the best division into three groups does not result from taking the best division into two groups and splitting up one of those groups. Consequently, this situation could not be well-represented by hierarchical clustering. Due to situations such as this one, hierarchical clustering can sometimes yield worse (i.e. less accurate) results than K-means clustering for a given number of clusters.

*The above context states a very important limitation of hierarchical clustering algorithm. Perform an experiment with simulated data or data of your selection to validate the above concept. (Note that you will not be judged based on the complexity of the data you are working with. Remember that, simpler the data better the explanation you can give. You will be marked based on the clarity of your communication and presentation). Include all necessary results.*

**Question 2**

*Read the context carefully:*

Thus far, in all cluster analysis examples we have used Euclidean distance as the dissimilarity measure. But sometimes other dissimilarity measures might be preferred. For example, correlation-based distance considers two observations to be similar if their features are highly correlated, even though the observed values may be far apart in terms of Euclidean distance. This is an unusual use of correlation, which is normally computed between variables; here it is computed between the observation profiles for each pair of observations. The following figure illustrates the difference between Euclidean and correlation-based distance.



***Figure:*** *Three observations with measurements on 20 variables are shown. Observations 1 and 3 have similar values for each variable and so there is a small Euclidean distance between them. But they are very weakly correlated, so they have a large correlation-based distance. On the other hand, observations 1 and 2 have quite different values for each variable, and so there is a large Euclidean distance between them. But they are highly correlated, so there is a small correlation-based distance between them.*

Crrelation-based distance focuses on the shapes of observation profiles rather than their magnitudes. The choice of dissimilarity measure is very important, as it has a strong effect on the resulting dendrogram. In general, careful attention should be paid to the type of data being clustered and the scientific question at hand. These considerations should determine what type of dissimilarity measure is used for hierarchical clustering.

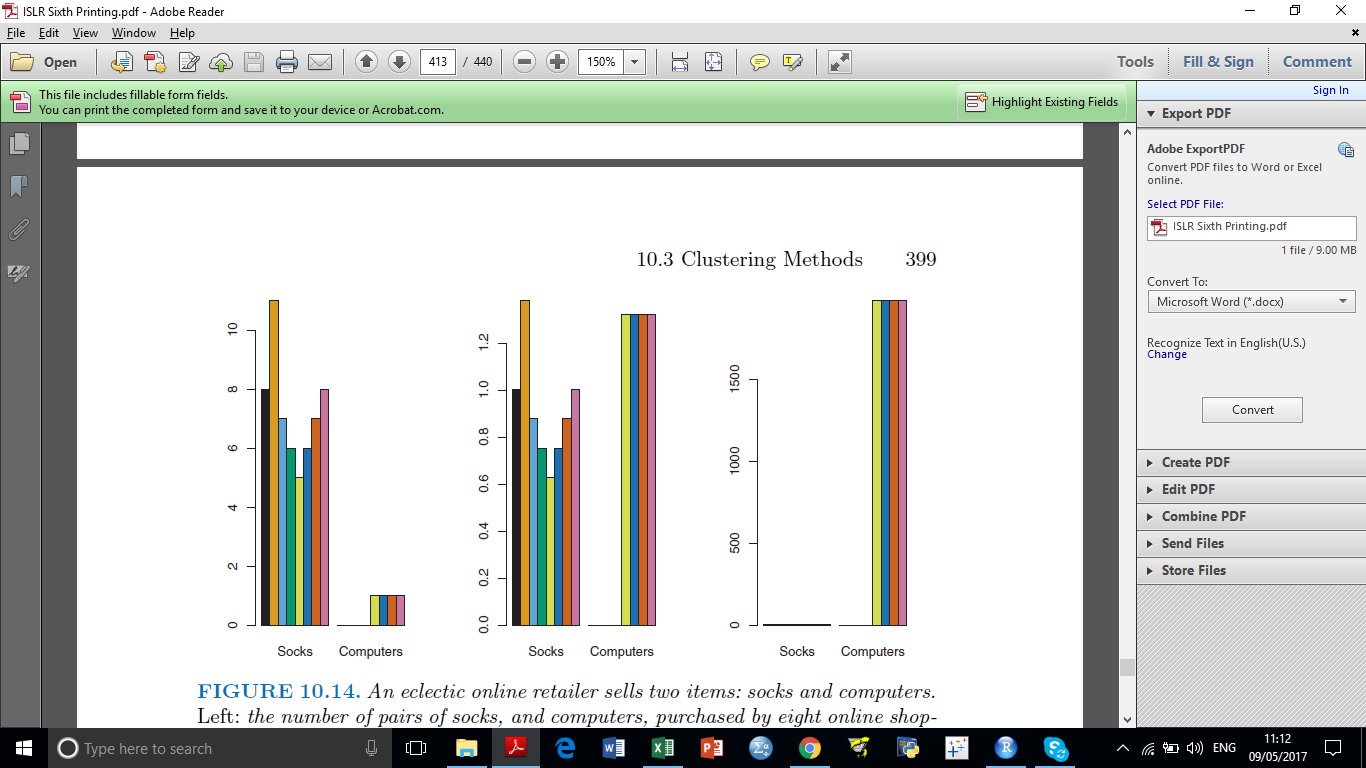
For instance, consider an online retailer interested in clustering shoppers based on their past shopping histories. The goal is to identify subgroups of similar shoppers, so that shoppers within each subgroup can be show items and advertisements that are particularly likely to interest them. Suppose the data takes the form of a matrix where the rows are the shoppers and the columns are the items available for purchase; the elements of the data matrix indicate the number of times a given shopper has purchased a given item (i.e. a 0 if the shopper has never purchased this item, a 1 if the shopper has purchased it once, etc.) What type of dissimilarity measure should be used to cluster the shoppers? If Euclidean distance is used, then shoppers who have bought very few items overall (i.e. infrequent users of the online shopping site) will be clustered together. This may not be desirable. On the other hand, if correlation-based distance is used, then shoppers with similar preferences (e.g. shoppers who have bought items A and B but never items C or D) will be clustered together, even if some shoppers with these preferences are higher-volume shoppers than others. Therefore, for this application, correlation-based distance may be a better choice.

*Based on the idea presented elaborate and simplify the example provided about online retail*

**Question 3**

*Read the context carefully:*

In addition to carefully selecting the dissimilarity measure used, one must also consider whether or not the variables should be scaled to have standard deviation one before the dissimilarity between the observations is computed. To illustrate this point, we continue with the online shopping example just described. Some items may be purchased more frequently than others; for instance, a shopper might buy ten pairs of socks a year, but a computer very rarely. High-frequency purchases like socks therefore tend to have a much larger effect on the inter-shopper dissimilarities, and hence on the clustering ultimately obtained, than rare purchases like computers. This may not be desirable. If the variables are scaled to have standard deviation one before the inter-observation dissimilarities are computed, then each variable will in effect be given equal importance in the hierarchical clustering performed. We might also want to scale the variables to have standard deviation one if they are measured on different scales; otherwise, the choice of units (e.g. centimeters versus kilometers) for a particular variable will greatly affect the dissimilarity measure obtained. It should come as no surprise that whether or not it is a good decision to scale the variables before computing the dissimilarity measure depends on the application at hand. An example is shown in Figure given below. We note that the issue of whether or not to scale the variables before performing clustering applies to K-means clustering as well.



***Figure****. An eclectic online retailer sells two items: socks and computers.*

***Left:*** *the number of pairs of socks, and computers, purchased by eight online shoppers is displayed. Each shopper is shown in a different color. If inter-observation dissimilarities are computed using Euclidean distance on the raw variables, then the number of socks purchased by an individual will drive the dissimilarities obtained, and the number of computers purchased will have little effect. This might be undesirable, since (1) computers are more expensive than socks and so the online retailer may be more interested in encouraging shoppers to buy computers than socks, and (2) a large difference in the number of socks purchased by two shoppers may be less informative about the shoppers’ overall shopping preferences than a small difference in the number of computers purchased. Center: the same data is shown, after scaling each variable by its standard deviation. Now the number of computers purchased will have a much greater effect on the inter-observation dissimilarities obtained. Right: the same data are displayed, but now the y-axis represents the number of dollars spent by each online shopper on socks and on computers. Since computers are much more expensive than socks, now computer purchase history will drive the inter-observation dissimilarities obtained.*

Consider the [USArrests](https://drive.google.com/open?id=0B9Jz-h_XxiRaZ0pJOG5mbHZqbGs) data. We will now perform hierarchical clustering on the states.

1. Using hierarchical clustering with complete linkage and Euclidean distance, cluster the states.
2. Cut the dendrogram at a height that results in three distinct clusters. Which states belong to which clusters?
3. Hierarchically cluster the states using complete linkage and Euclidean distance, after scaling the variables to have standard deviation one.
4. What effect does scaling the variables have on the hierarchical clustering obtained? In your opinion, should the variables be scaled before the inter-observation dissimilarities are computed? Provide a justification for your answer.

**Question 4:**

*Read the context carefully:*

Linkage defines the dissimilarity between two groups of observations. The three most common types of linkage—complete, average and single. Average, complete, and single linkage are most popular among statisticians. Average and complete linkage are generally preferred over single linkage, as they tend to yield more balanced dendrograms.

Unsupervised techniques are often used in the analysis of genomic data. Consider the [NCI60](https://drive.google.com/open?id=0B9Jz-h_XxiRac2h0ZFlrNi1SaG8) cancer cell line microarray data, which consists of 6,830 gene expression measurements on 64 cancer cell lines. You need to hierarchically cluster the cell lines in the NCI60 data, with the goal of finding out whether or not the observations cluster into distinct types of cancer. To begin, we standardize the variables to have mean zero and standard deviation one.

“The choice of linkage certainly does affect the results obtained. Typically, single linkage will tend to yield trailing clusters: very large clusters onto which individual observations attach one-by-one. On the other hand, complete and average linkage tend to yield more balanced, attractive clusters.” Perform cluster analysis on the NC160 data and validate the statement and comment on your observations.

**Question 5:**

*Read the context carefully:*

Both K-means and hierarchical clustering will assign each observation to a cluster. However, sometimes this might not be appropriate. For instance, suppose that most of the observations truly belong to a small number of (unknown) subgroups, and a small subset of the observations are quite different from each other and from all other observations. Then since K Means and hierarchical clustering force every observation into a cluster, the clusters found may be heavily distorted due to the presence of outliers that do not belong to any cluster. Mixture models are an attractive approach for accommodating the presence of such outliers.

*Design an experimental problem to validate the distortion of clusters might be a result due to the presence of outliers. (Designing an experimental data having only 2-dimension will do. This is just of illustration. Use proper visualization and possible use of colour encodings).*

*At the end of the context is mentioned about mixture models. Find out a little more about it, like what is mixture models, what is the intuitive idea behind such models, etc.*

**Question 6:**

1. *List down some of the issues in cluster analysis.*
2. *What are the advantages and limitations of Hierarchical cluster analysis?*
3. *What are the advantages and limitations of K-Means clustering?*

Try to support this with simulated (or real) data as much as possible.