Contents

1 Agenda for Today's Class

- Review of Important Topics
- Review of K-Means Clustering (SPSS Exercise)
- Two-Step Clustering
- Review of Regression (Optional for Math Science Students)

2 Important Topics

- **Multi-collinearity**: Multi-collinearity occurs when two or more predictors in the model are correlated and provide redundant information about the response. Examples of pairs of multi-collinear predictors are years of education and income, height and weight of a person, and assessed value and square footage of a house.
- **Consequences of high multicollinearity**: Multi-collinearity leads to decreased reliability and predictive power of statistical models, and hence, very often, confusing and misleading results.
- Multicollinearity will be dealt with in a future component of this course: Variable Selection Procedures.

3 Two-Step Cluster Analysis

When you have a really large data set or you need a clustering procedure that can rapidly form clusters on the basis of either categorical or continuous data, neither of the previous two procedures are entirely appropriate. Hierarchical clustering requires a matrix of distances between all pairs of cases, and k-means requires shuffling cases in and out of clusters and knowing the number of clusters in advance.

The Two-Step Cluster Analysis procedure was designed for such applications. The name two-step clustering is already an indication that the algorithm is based on a two-stage approach

- In the first stage, the algorithm undertakes a procedure that is very similar to the k-means algorithm.
- Based on these results, the two-step procedure conducts a modified hierarchical agglomerative clustering procedure that combines the objects sequentially to form homogenous clusters.

The Two-Step Cluster Analysis is an exploratory tool designed to reveal natural groupings (or clusters) within a data set that would otherwise not be apparent. The algorithm employed by this procedure has several desirable features that differentiate it from traditional clustering techniques:

- Handling of categorical and continuous variables. By assuming variables to be independent, a joint ***multinomial-normal distribution*** can be placed on categorical and continuous variables. (Interesting, but not examinable).
- Automatic selection of number of clusters. By comparing the values of a ***model-choice criterion*** across different clustering solutions, the procedure can automatically determine the optimal number of clusters.
- Scalability. By constructing a ***cluster features*** (CF) tree that summarizes the records, the Two-Step algorithm allows you to analyze large data files. The Two-Step Cluster Analysis requires only one pass of data (which is important for very large data files).

3.1 Pre-clustering

In two-step clustering, to make large problems tractable, in the first step, cases are assigned to ***preclusters***. In the second step, the preclusters are clustered using the hierarchical clustering algorithm. You can specify the number of clusters you want or let the algorithm decide based on preselected criteria.

In general, the larger the number of sub-clusters produced by the pre-cluster step, the more accurate the final result is. However, too many sub-clusters will slow down the clustering during the second step.

The maximum number of sub-clusters should be carefully chosen so that it is large enough to produce accurate results and small enough not to slow down the second step clustering.

3.1.1 Step 1: Preclustering: Making Little Clusters

The first step of the two-step procedure is formation of preclusters. The goal of preclustering is to reduce the size of the matrix that contains distances between all possible pairs of cases. Preclusters are just clusters of the original cases that are used in place of the raw data in the hierarchical clustering. As a case is read, the algorithm decides, based on a distance measure, if the current case should be merged with a previously formed precluster or start a new precluster. When preclustering is complete, all cases in the same precluster are treated as a single entity. The size of the distance matrix is no longer dependent on the number of cases but on the number of preclusters.

3.1.2 Step 2: Hierarchical Clustering of Preclusters

In the second step, SPSS uses the standard hierarchical clustering algorithm on the preclusters. Forming clusters hierarchically lets you explore a range of solutions with different numbers of clusters.

4 Important Considerations for Two-Step Clustering

4.1 Cluster Features Tree

Two-Step Cluster Analysis is done by building a so-called ***cluster feature tree*** whose ***leaves*** represent distinct objects in the dataset. The procedure can handle categorical and continuous variables simultaneously and offers the user the flexibility to specify the cluster numbers as well as the maximum number of clusters, or to allow the technique to automatically choose the number of clusters on the basis of statistical evaluation criteria.

Additionally, the procedure indicates each variables importance for the construction of a specific cluster. These desirable features make the somewhat less popular two-step clustering a viable alternative to the traditional methods.

4.2 Types of Data

The Two-Step procedure works with both continuous and categorical variables. Cases represent objects to be clustered, and the variables represent attributes upon which the clustering is based.

4.3 Case Order

Note that the cluster features tree and the final solution may depend on the order of objects (or cases). To minimize order effects, randomly order the cases. It is recommended to obtain several different solutions with cases sorted in different random orders to verify the stability of a given solution. In situations where this is difficult due to extremely large file sizes, multiple runs with a sample of cases sorted in different random orders might be substituted.

5 SPSS Implementation

- To implement a Two-Step Cluster Analysis in SPSS, you use the following options: ***Analyze*** > ***Classify*** > ***TwoStep Cluster***.
- **Distance Measure** Log likelihood distance measures are the default; Euclidean distance can be used if all variables are continuous. (Log likelihood distance measures are not part of course).
- **Count of Continuous Variables** Continuous variables are standardized by default. The variables are standardized so that they all contribute equally to the distance or similarity between cases.
- **Number of clusters** You can specify the number of clusters, or you can let the algorithm select the optimal number based on either the Schwarz Bayesian criterion (BIC) or the Akaike information criterion (AIC).
- **Clustering Criterion** BIC and AIC are offered with the default being BIC.

5.1 Graphical Outputs

The lower part of the output indicates the quality of the cluster solution. The silhouette measure of cohesion and separation is a measure of the clustering solutions overall goodness-of-fit. It is essentially based on the average distances between the objects and can vary between -1 and +1. Specifically, a silhouette measure of less than 0.20 indicates a poor solution quality, a measure between 0.20 and 0.50 a fair solution, whereas values of more than 0.50 indicate a good solution . In our case, the measure indicates a satisfactory ("fair") cluster quality. Consequently, you can proceed with the analysis by double-clicking on the output. This will open up the model viewer , an evaluation tool that graphically presents the structure of the revealed clusters.

The model viewer provides us with two windows: the main view, which initially shows a model summary (left-hand side), and an auxiliary view, which initially features the cluster sizes (right-hand side). At the bottom of each window, you can request different information, such as an overview of the cluster structure and the overall variable importance.

6 Clustering Algorithm

To better understand how a clustering algorithm works, lets manually examine some of the single linkage procedures calculation steps. We start off by looking at the initial (Euclidean) distance matrix displayed previously.

Objects	A	B	C	D	E	F	G
A	0						
B	3	0					
C	2.236	1.414	0				
D	2	3.606	2.236	0			
E	3.606	2	1.414	3	0		
F	4.123	4.472	3.162	2.236	2.828	0	
G	5.385	7.071	5.657	3.606	5.831	3.162	0

- In the very first step, the two objects exhibiting the smallest distance in the matrix are merged. Note that we always merge those objects with the smallest distance, regardless of the clustering procedure (e.g., single or complete linkage). (N.B. In the following example, ties will be broken at random.)
- As we can see, this happens to two pairs of objects, namely B and C (d(B, C) = 1.414), as well as C and E (d(C, E) = 1.414). In the next step, we will see that it does not make any difference whether we first merge the one or the other, so lets proceed by forming a new cluster, using objects B and C.

Objects	A	B, C	D	E	F	G
A	0					
B, C	2.236	0				
D	2	2.236	0			
E	3.606	1.414	3	0		
F	4.123	3.162	2.236	2.828	0	
G	5.385	5.657	3.606	5.831	3.162	0

- Having made this decision, we then form a new distance matrix by considering the single linkage decision rule as discussed above. According to this rule, the distance from, for example, object A to the newly formed cluster is the minimum of d(A, B) and d(A, C). As d(A, C) is smaller than d(A, B), the distance from A to the newly formed cluster is equal to d(A, C); that is, 2.236.
- We also compute the distances from cluster [B,C] (clusters are indicated by means of squared brackets) to all other objects (i.e. D, E, F, G) and simply copy the remaining distances such as d(E, F) that the previous clustering has not affected.
- Continuing the clustering procedure, we simply repeat the last step by merging the objects in the new distance matrix that exhibit the smallest distance (in this case, the newly

Objects	A	B, C, E	D	F	G
A	0				
B, C, E	2.236	0			
D	2	2.236	0		
F	4.123	2.828	2.236	0	
G	5.385	5.657	3.606	3.162	0

formed cluster [B, C] and object E) and calculate the distance from this cluster to all other objects.

Objects	A, D	B, C, E	F	G
A, D	0			
B, C, E	2.236	0		
F	2.236	2.828	0	
G	3.606	5.657	3.162	0

- We continue in the same fashion until one cluster is left. By following the single linkage procedure, the last steps involve the merger of cluster [A,B,C,D,E,F] and object G at a distance of 3.162.

Objects	A, B, C, D, E	F	G
A, B, C, D, E	0		
F	2.236	0	
G	3.606	3.162	0

Objects	A, B, C, D, E, F	G
A, B, C, D, E, F	0	
G	3.162	0

7 More on Two-Step Clustering

7.1 Step 1: Pre-clustering: Making Little Clusters

The first step of the two-step procedure is formation of pre-clusters. The goal of pre-clustering is to reduce the size of the Distance matrix (the matrix that contains distances between all possible pairs of cases). Pre-clusters are just clusters of the original cases that are used in place of the raw data in the hierarchical clustering. As a case is read, the algorithm decides, based on a distance measure, if the current case should be merged with a previously formed pre-cluster or start a new precluster.

When preclustering is complete, all cases in the same precluster are treated as a single entity. The size of the distance matrix is no longer dependent on the number of cases but on the number of preclusters.

7.2 Step 2: Hierarchical Clustering of Preclusters

In the second step, SPSS uses the standard hierarchical clustering algorithm on the preclusters. Forming clusters hierarchically lets you explore a range of solutions with different numbers of clusters. Tip: The Options dialog box lets you control the number of preclusters. Large numbers of preclusters give better results because the cases are more similar in a precluster; however, forming many preclusters slows the algorithm.

7.3 Step 1: Pre-clustering: Making Little Clusters

The first step of the two-step procedure is formation of pre-clusters. The goal of pre-clustering is to reduce the size of the Distance matrix (the matrix that contains distances between all possible pairs of cases). Pre-clusters are just clusters of the original cases that are used in place of the raw data in the hierarchical clustering. As a case is read, the algorithm decides, based on a distance measure, if the current case should be merged with a previously formed pre-cluster or start a new precluster.

When preclustering is complete, all cases in the same precluster are treated as a single entity. The size of the distance matrix is no longer dependent on the number of cases but on the number of preclusters.

7.4 Step 2: Hierarchical Clustering of Preclusters

In the second step, SPSS uses the standard hierarchical clustering algorithm on the preclusters. Forming clusters hierarchically lets you explore a range of solutions with different numbers of clusters. Tip: The Options dialog box lets you control the number of preclusters. Large numbers of preclusters give better results because the cases are more similar in a precluster; however, forming many preclusters slows the algorithm.

Some of the options you can specify when using two-step clustering are: **Standardization:** The algorithm will automatically standardize all of the variables unless you override this option.

Distance measures: If your data are a mixture of continuous and categorical variables, you can use only the log-likelihood criterion. The distance between two clusters depends on the decrease in the log-likelihood when they are combined into a single cluster. If the data are only continuous variables, you can use the Euclidean distance between two cluster centers. Depending on the distance measure selected, cases are assigned to the cluster that leads to the largest log-likelihood or to the cluster that has the smallest Euclidean distance.

Number of clusters: You can specify the number of clusters to be formed, or you can let the algorithm select the optimal number based on either the Schwarz Bayesian Criterion or the Akaike information criterion.

Outlier handling: You have the option to create a separate cluster for cases that don't fit well into any other cluster.

Range of solutions: You can specify the range of cluster solutions that you want to see.

8 Two-Step Cluster

When you have a really large data set or you need a clustering procedure that can rapidly form clusters on the basis of either categorical or continuous data, neither of the previous two procedures are entirely appropriate. Hierarchical clustering requires a matrix of distances between all pairs of cases, and k-means requires shuffling cases in and out of clusters and knowing the number of clusters in advance.

The Two-Step Cluster Analysis procedure was designed for such applications. The name two-step clustering is already an indication that the algorithm is based on a two-stage approach: In the first stage, the algorithm undertakes a procedure that is very similar to the k-means algorithm. Based on these results, the two-step procedure conducts a modified hierarchical agglomerative clustering procedure that combines the objects sequentially to form homogenous

clusters. This is done by building a so-called ***cluster feature tree*** whose ***leaves*** represent distinct objects in the dataset. The procedure can handle categorical and continuous variables simultaneously and offers the user the flexibility to specify the cluster numbers as well as the maximum number of clusters, or to allow the technique to automatically choose the number of clusters on the basis of statistical evaluation criteria.

The Two-Step Cluster Analysis requires only one pass of data (which is important for very large data files).

Additionally, the procedure indicates each variables importance for the construction of a specific cluster. These desirable features make the somewhat less popular two-step clustering a viable alternative to the traditional methods.

9 Measures of Fit

Two-Step Cluster Analysis guides the decision of how many clusters to retain from the data by calculating measures-of-fit such as ***Akaikes Information Criterion (AIC)*** or ***Bayes Information Criterion (BIC)***.

These are relative measures of goodness-of-fit and are used to compare different solutions with different numbers of segments.("Relative" means that these criteria are not scaled on a range of, for example, 0 to 1 but can generally take any value.) Compared to an alternative solution with a different number of segments, smaller values in AIC or BIC indicate an increased fit.

SPSS computes solutions for different segment numbers (up to the maximum number of segments specified before) and chooses the appropriate solution by looking for the smallest value in the chosen criterion. However, which criterion should we choose? AIC is well-known for overestimating the correct number of segments, while BIC has a slight tendency to underestimate this number.

Thus, it is worthwhile comparing the clustering outcomes of both criteria and selecting a smaller number of segments than actually indicated by AIC. Nevertheless, when running two separate analyses, one based on AIC and the other based on BIC, SPSS usually renders the same results.

9.1 AIC and BIC in Two-Step Cluster Analysis

(Removed from Last Week's Class due to Version Update)

Two-Step Cluster Analysis guides the decision of how many clusters to retain from the data by calculating measures-of-fit such as ***Akaikes Information Criterion (AIC)*** or ***Bayes Information Criterion (BIC)***.

These are relative measures of goodness-of-fit and are used to compare different solutions with different numbers of segments.("Relative" means that these criteria are not scaled on a range of, for example, 0 to 1 but can generally take any value.)

Important: Compared to an alternative solution with a different number of segments, smaller values in AIC or BIC indicate an increased fit.

SPSS computes solutions for different segment numbers (up to the maximum number of segments specified before) and chooses the appropriate solution by looking for the smallest value in the chosen criterion. However, which criterion should we choose?

- AIC is well-known for overestimating the correct number of segments

- BIC has a slight tendency to underestimate this number.

Thus, it is worthwhile comparing the clustering outcomes of both criteria and selecting a smaller number of segments than actually indicated by AIC. Nevertheless, when running two separate analyses, one based on AIC and the other based on BIC, SPSS usually renders the same results.

Once you make some choices or do nothing and go with the defaults, the clusters are formed. At this point, you can consider whether the number of clusters is "good". If automated cluster selection is used, SPSS prints a table of statistics for different numbers of clusters, an excerpt of which is shown in the figure below. You are interested in finding the number of clusters at which the Schwarz BIC becomes small , but also the change in BIC between adjacent number of clusters is small.

The decision of how much benefit accrued by another cluster is very subjective. In addition to the BIC, a high ratio of distance of measures is desirable. In the figure below, the number of clusters with this highest ratio is three.

Autoclustering statistics

		Schwarz's Bayesian Criterion (BIC)	BIC Change[1]	Ratio of BIC Changes[2]	Ratio of Distance Measures[3]
Number of Clusters	1	6827.387			
	2	5646.855	-1180.532	1.000	1.741
	3	5000.782	-646.073	.547	1.790
	4	4672.859	-327.923	.278	1.047
	5	4362.908	-309.951	.263	1.066
	6	4076.832	-286.076	.242	1.193
	7	3849.057	-227.775	.193	1.130
	8	3656.025	-193.032	.164	1.079
	9	3482.667	-173.358	.147	1.162
	10	3343.916	-138.751	.118	1.240
	11	3246.541	-97.376	.082	1.128
	12	3168.733	-77.808	.066	1.093
	13	3103.950	-64.783	.055	1.022
	14	3042.116	-61.835	.052	1.152
	15	2998.319	-43.796	.037	1.059

1. The changes are from the previous number of clusters in the table.
2. The ratios of changes are relative to the change for the two cluster solution.
3. The ratios of distance measures are based on the current number of clusters against the previous number of clusters.

Figure 1: Schwarz Bayesian Information Criterion

Information Criterions We define two types of information criterion: the Akaike Information Criterion (AIC) and the Schwarz?s Bayesian Information Criterion (BIC). The Akaike information criterion is a measure of the relative goodness of fit of a statistical model. AIC = 2p ? 2 ln(L)

- p is the number of predictor variables in the model.
- L is the value of the Likelihood function for the model in question.
- For AIC to be optimal, n must be large compared to p.

An alternative to the AIC is the Schwarz BIC, which additionally takes into account the sample size n. BIC = p ln n ? 2 ln(L) When using the AIC (or BIC) for selecting the optimal model, we choose the model for which the AIC (or BIC) value is lowest.

Akaike Information Criterion

- Akaike?s information criterionis a measure of the goodness of fit of an estimated statistical model. The AIC was developed by Hirotsugu Akaike under the name of ?an information criterion? in 1971.
- The AIC is a model selection tool i.e. a method of comparing two or more candidate regression models. The AIC methodology attempts to find the model that best explains the data with a minimum of parameters. (i.e. in keeping with the law of parsimony)
- The AIC is calculated using the ?likelihood function? and the number of parameters . The likelihood value is generally given in code output, as a complement to the AIC. (Likelihood function is not on our course)
- Given a data set, several competing models may be ranked according to their AIC, with the one having the lowest AIC being the best. (Although, a difference in AIC values of less than two is considered negligible).