5

Cluster Analysis

This primer focuses on cluster analysis. We assume you have read the section on cluster analysis in Chapter 11, but we repeat parts of it here to set context. Cluster analysis refers to a class of statistical methods that identifies subgroups of individuals who show common profiles across a set of variables within a subgroup but whose profiles are distinct when contrasted with individuals in other subgroups. The subgroups are referred to as *clusters.* There are many different types of cluster analysis. We discuss in this primer three types, (1) *hierarchical clustering* (also known as *connectivity-based clustering*), (2) *partitioning clustering* (also called *centroid clustering*), and (3) *mixture modeling*. There are other forms of cluster analysis besides these three, such as *distribution based clustering models* and *density based clustering models*. Discussion of them, however, is beyond the scope of this primer. Interested readers are referred to Everitt et al. (2011) and Hennig, Meila, Murtagh and Rocci (2015).

HIERARCHICAL CLUSTERING AND DISTANCE SCORES

Hierarchical clustering calculates a matrix of distance scores between all possible pairs of individuals with the distance scores reflecting how dissimilar a given pair of individuals is across the X variables in the analysis. Individuals with small distance scores are grouped into the same cluster; individuals with large distance scores are grouped into different clusters. Distance scores are at the heart of hierarchical clustering, so we elaborate them here. We consider three such indices.

One common index of distance is the *squared Euclidean distance score*. It is defined as the sum of the squared differences between scores for two individuals across the variables. For example, suppose adolescents rate statements about different discipline strategies their mothers might use if they broke a serious family rule. The ratings are made on a 0 to 10 scale, where 0 indicates strong disagreement with the statement, 5 indicates neither agreement nor disagreement, and 10 indicates strong agreement with the statement, with higher numbers reflecting increasing degrees of agreement. Here are the statements:

X1 = She would take things away from me (like my computer, cell phone or TV).
X2 = She would ground me.

X3 = She would hit me.
X4 = She would cut me off from my friends - make me stop seeing them.
X5 = She would yell at me.
X6 = She would let me know how disappointed she is.
X7 = She would be cold towards me.
X8 = She would try to explain to me why I should not do it again.

Here are scores that two adolescents might provide for the statements (in the first two columns):

	Adolescent 1	Adolescent 2	Difference	Squared Difference
X1	8	9	-1	1
X2	2	0	2	4
X3	0	2	-2	4
X4	2	1	1	1
X5	2	1	1	1
X6	8	9	-1	1
X7	0	0	0	0
X8	2	2	0	0
			Sum:	12

The third column is the difference between the ratings for the two adolescents and the fourth column is the square of these differences. The sum of this column is the squared Euclidean distance score, which for these two individuals is 12. If two individuals have identical profiles, the Euclidean distance score is zero. If they have maximally discrepant profiles, then for this example, the Euclidean distance score would be 800. A score of 12 indicates profiles that are fairly similar. Sometimes instead of working with raw scores, analysts will first standardize scores on each variable and use standard scores instead. This is often done when the variables have different metrics (e.g., one variable is scored on a 1 to 5 scale and another variable is scored on a 1 to 100 scale). In the present example, all variables have a common metric (0 to 10), so standardization is not used.

A second type of distance score is called the *Euclidean distance score* and is simply the square root of the squared Euclidean difference score. For the above example, it equals the square root of 12, which is 3.46. A third index is called the *Manhattan distance score* and it is the sum of the absolute differences between profiles across

variables. For the above two individuals, we calculate the absolute value of the entries in column 3 and then sum the scores, yielding a value of 8. If we divide it by the number of variables, 8, it reflects the average disparity between variables, in this case, 8/8 = 1.0. The Euclidean distance score gives more weight to larger disparities than smaller disparities when forming the aggregate. The Manhattan distance scores give equal weight to variable disparities when forming the aggregate, whether those disparities are large or small. To transform the Euclidean distance score to the "average" disparity on the original variable metrics, you can divide it by the square root of the number of variables (in this case the square root of 8 = 2.83). For our example 3.46/2.83 = 1.22, which is slightly larger than the value of 1.00 computed using the average of the Manhattan index because of the greater weight given to larger disparities.

Once distance scores are computed for all possible pairs of individuals, an algorithm is applied to group together into clusters individuals with small distance scores. This is usually done in steps. Suppose for the discipline style example, we use Manhattan distance scores. At step 0, we consider everyone as being in a distinct cluster. If we have 100 individuals, we have 100 clusters. At step 1, we merge into a cluster the two individuals with the smallest distance score. We now have 99 clusters, one with two individuals in it and everyone else being their own cluster. We will refer to the former cluster as cluster A. For the next step (also called an *iteration*), we want to merge two of the 99 clusters into a larger cluster to give us 98 clusters. However, we do not have numerical distance scores between Cluster A (the one with two people in it) and each of the remaining clusters, which we need to decide what "units" to merge next.

We illustrate the method used to accomplish this by selecting one of the other clusters, which we will refer to as cluster B (which, at this point, has but one individual in it). Suppose individual 1 in cluster A has a distance score of 8 with the individual in cluster B and individual 2 in cluster A has a distance score of 12 with the individual in cluster B. One strategy to index the distance between cluster A and cluster B is to use the largest of these two distance scores, which is 12. This is known as a *complete linkage* algorithm. Another possibility is to use the smallest of the two distance scores, in this case, 8, which is known as a *single linkage* algorithm. A third possibility is to use the average of the two distance scores ((8+12)/2 = 10), which is known as an *average linkage* algorithm. Yet another possibility is to use what is called *Ward's method*. This is a more complex algorithm that reflects how much the sum of squares for scores in one cluster increases when we merge into it the scores in the other cluster (see Everitt, Landau, Leese & Stahl, 2011, for details). Whichever strategy is used, one ends up with a revised distance score reflecting the distance between the newly formed cluster A and cluster B. This process is performed for cluster A with each of the other 99 single-person clusters.

We now have a new set of distance scores between all possible pairs of the 99 clusters.

At Step 2, we merge together the two clusters (of the 99) that have the smallest distance score, forming 98 clusters. We recalculate the distance scores between the 98 clusters using the above principles and, at the next iteration, merge together the two clusters with the smallest distance score, thereby creating 97 clusters. The process repeats over and over until at the final step, we have merged everyone into one big cluster.

At each step in this process, we make note of what the value of the distance score was that led us to merge the 2 clusters with the smallest distance score. For example, at step 1, the merging may have happened for the two clusters that had a distance score of 1. At step 2, the merging may have happened for the two clusters that had a distance score of 3. At step 3, the merging may have happened for the two clusters that had a distance score of 4. And so on. Prior to the analysis, we define a theoretical "cut-point" where we designate a distance score as being too large to justify merging clusters together. For example, we might decide that a distance score of 20 or greater is too large to justify merging two individuals or two clusters into a larger cluster. We identify the step that merged two clusters at that value or just under it and define the final cluster solution as being the one that occurred at that step. Note that at this step, there still may be individuals that are in their own, one person cluster.

The approach described is called hierarchical clustering because the clustering occurs in a hierarchical or sequential fashion. Hierarchical clustering actually refers to a whole family of methods that vary in how the initial distance scores are defined and how distance scores are re-calculated between clusters. Hierarchical clustering can be either *agglomerative* by starting with N clusters for N individuals and then aggregating them into larger clusters, per the above example, or it can be *divisive* by starting with one large cluster that contains everyone and then dividing it up at successive steps into increasingly more fine grained clusters. Hierarchical clustering is computationally challenging for large N and some researchers are dissatisfied with the prospect of ending up with many "outlier" clusters consisting of single individuals. Nevertheless, the approach has been used for many interesting substantive applications. The worked examples for this primer contain an example of hierarchical clustering.

K-MEANS AND CENTROID CLUSTER ANALYSIS

Another popular approach to cluster analysis is called *centroid clustering*, with the most common instantiation being *k-means cluster analysis*. In this approach, each cluster is conceptualized as having a centroid (e.g., a mean value) on each target variable and the focus is on defining clusters so as to minimize within-cluster variation on a given X relative to this centroid. The general strategy is to divide the data into groups such that

the within-cluster variability within each group is as small as possible. Using the terminology of analysis of variance, the strategy seeks to minimize the sum of squares within-groups with the consequent side effect of maximizing the sum of squares between-groups. Different algorithms have been suggested for accomplishing this, including the Hartigan-Wong method, the Lloyd method, the Forgy method, and the MacQueen method (see Everitt et al., 2011, and Hennig, Meila, Murtagh & Rocci, 2015), with most simulation studies favoring the Hartigan-Wong method.

In k-means cluster analysis, researchers must specify *a priori* the number of clusters to extract. With this information in hand, the computer algorithm then sets about the task of assigning individuals to clusters in a way that minimizes the sum of squares of the X variables, considered multivariately, within each cluster. One way of thinking about this task is to think of each possible cluster number as a different model that should adequately represent the population data. For example, there is a two cluster model, a three cluster model, a four cluster model, and so on and we want to choose the "best" of these models, i.e., the model that best reflects the data dynamics. There are a variety of diagnostics that researchers use to accomplish this task.

We consider in the remainder of this section three issues (1) choosing the number of clusters, (2) interpreting the clusters, and (3) relating cluster membership to other variables.

Choosing the Number of Clusters

When conducting a k-means cluster analysis, each individual is assigned to a subgroup/cluster. We can conceptualize the clusters as a qualitative variable (the cluster the individual is in) that can be used in later statistical modeling. For example, in a two group/cluster model, the variable of cluster membership has two levels; in a three group/cluster model, the variable has three levels; and so on. One way to evaluate the different cluster models is to examine the overall percent of variation in scores that each model accounts for. Given a set of variables to cluster analyze (such as the eight discipline items discussed above), we first calculate an overall index of variability across all of the measures by listing the scores for every individual and each variable in a long, single column. We then calculate a sum of squares of this vector of scores using the standard statistical formula for a sum of squares. The result is referred to as the *sum of squares total*. We then use standard analysis of variance methods to calculate the percent of this variation that can be accounted for (in the form of eta squared multiplied by 100) by a two cluster model, by a three cluster model, by a four cluster model, and so on. Obviously, the percent of variance accounted for will improve as we increase the number of clusters. However, what we look for is when there are large changes in the percent of

variance accounted for by each successive model and when the changes become trivial, much like a scree test in factor analysis. Here are the percentages of variance accounted for by models for up to 10 clusters when applied to the discipline example:

Model (Number of Clusters)	Percent of Variance Accounted For
2	23.4%
3	36.3%
4	41.8%
5	46.4%
6	50.8%
7	53.7%
8	56.8%
9	58.9%
10	60.3%

If we use as a rough cut-off of increments of at least 5% explained variance, a model with about four clusters seems viable, but there is no clear cut-point.

The algorithms of K-means cluster analysis cluster individuals into groups with the intent to minimize the sum of squares within clusters, i.e., they minimize the sum of the squared distance between a person's score on X and the mean of X for the group/cluster to which s/he belongs for all X considered multivariately. For this reason, many methodologists like to apply a scree test not just to the percent of variance accounted for by each model but also to the sum of squares within for each model. These values can be quite large and non-intuitive, so it is common to examine a plot of them in the spirit of the scree test. The plot that results for the discipline example is shown in Figure 5.1. There is not a clear, definitive "elbow" in the plot where the sum of squares within flattens out, but the trend again seems to favor a 3 or 4 cluster models.

A third approach to evaluating the models is to calculate information fit indices for each model in the form of the classic Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC). One then determines with these statistics which model has the largest likelihood of producing the data (Pelleg & Moore, 2000). An explication of the AIC and BIC indices is presented in the regression mixture modeling primer. The section from that primer with minor edits is appended to this primer for easier access. We do not describe in detail the AIC or BIC results for the discipline example, but they favored a four cluster solution.

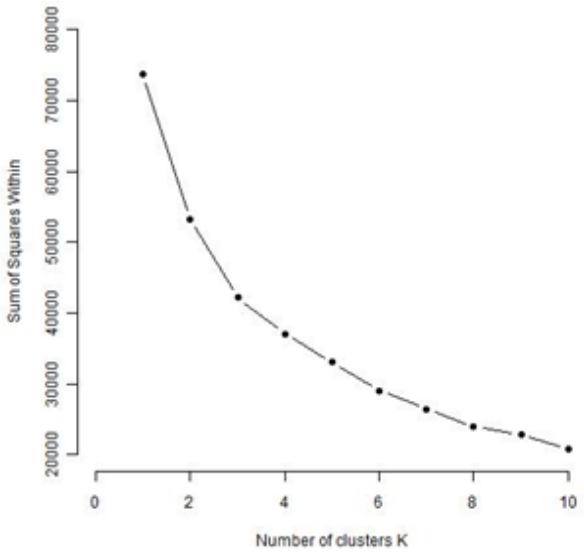

FIGURE 5.1. Plot of Sum of Squares Within

Yet another approach sometimes used to choose the number of clusters is to compare the *average silhouette width* for each model. The silhouette width is an index that reflects the compactness and separation of clusters, considered simultaneously (Rousseeuw, 1987). Consider, as, an example, a three cluster model. For each individual in a given cluster, we calculate the average distance that a person is from all other people in the same cluster. We average these values across all individuals in the cluster and refer to this quantity as A. Next, for each individual in that cluster, we calculate the average distance the person is from all other people in a different cluster and average these values. We do this for each of the other clusters and then use the value that is lowest across the clusters. We refer to this value as B. We then subtract A from B to yield an index of separation relative to homogeneity. We divide this difference by the larger of A or B, yielding a value between 0 and 1.00. The larger the value, the better differentiated the clusters are and the more we prefer the model. Silhouette values near 0.50 or larger are generally considered to reflect reasonable cluster structuring, although there is controversy about the most appropriate rule-of-thumb. Here are the silhouette width values for our example:

Model (Number of Clusters)	Average Silhouette Width
2	0.22
3	0.35
4	0.42
5	0.26
6	0.30
7	0.32
8	0.34
9	0.35
10	0.36

The results tend to favor a four cluster model.

The different methods for choosing the number of clusters sometimes converge with one another and sometimes not. If there was always strong convergence between them, we would not bother computing and examining different indices because once you have examined one, you have your answer. Coupled with the interpretability of the clusters (e.g., the substantive sense that the patterning of the X mean scores across clusters makes) and the size of the clusters (we often are not interested in clusters that are extremely small in terms of the number of individual in them), you make your final choice given performance of the different models on the different indices. See the worked example associated with this primer for an illustration.

Interpreting the Solution

To gain perspectives on the chosen cluster model, we usually examine the mean scores for each variable across clusters, the standard deviations for each variable within each cluster (hoping that they are small, which implies within-cluster homogeneity), and the relative sample size for each cluster, to gain a sense of how large the cluster is in the population. As an example, here are the cluster means for a four cluster solution for the maternal discipline styles:

Discipline Style	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Would take things away (like cell phone, TV)	1.52	9.45	9.44	9.37
Would ground me	1.90	9.17	9.11	9.42

Would hit me	2.04	3.40	1.31	8.86
Would cut me off from my friends	1.72	7.90	7.68	8.27
Would yell at me	3.91	4.19	3.81	9.53
Would let me know how disappointed she is	3.48	3.36	9.22	9.45
Would be cold towards me	1.95	1.32	8,30	9.12
Would explain why I should not do it again	3.67	9.71	9.25	8.78

Recall that the metric for each variable is from 0 to 10 with higher scores indicating greater levels of agreement. Cluster 1 is characterized by mothers who are likely lax because they do not exhibit any of the discipline strategies. Cluster 4 is characterized by mothers who are likely harsh because they exhibit use of virtually all of the discipline strategies. Cluster 2 is characterized by mothers who exert power assertion by depriving their adolescent child of desired objects (e.g., cell phones, access to friends and events) but who also try to explain the reasons why the transgression was bad. Cluster 4 is mothers similar to those in Cluster 3 but who also use guilt induction and rejection. The percent of mothers in the four clusters were 12%, 32%, 25%, and 31% for clusters 1 through 4, respectively. The within-group standard deviations for each variable (not shown here) tended to be around 2.0

Some researchers conduct one way analyses of variance and associated pairwise mean comparisons for each input variable as a function of cluster membership to document cluster differences in means. Statisticians argue that p values for these tests are dubious because they fail to take into account the uncertainty associated with the assignment of individuals to clusters, thereby underestimating standard errors. Also, the analyses inherently embrace a two-step process for statistical inference, namely (1) the accurate recovery of a population cluster structure through the analysis of sample data followed by (2) the accurate recovery of differences in population cluster means through traditional F tests following Step 1. The performance of significance tests in the context of this two-step approach are not well understood. As such, any significance tests must be treated with caution.

Relating Cluster Membership to Other Variables

As noted, one can create a new qualitative variable in the data to represent cluster membership, i.e., which of the four clusters the person is classified into. This variable can then be used in statistical modeling with other variables. For example, how do the different discipline style clusters relate to engagement in future problem behaviors on the part of adolescents? Is ethnicity of the mother related to the multivariate pattern of discipline strategies as reflected by the clusters? And so on.

A difficulty with such modeling is that membership in a given cluster is best conceptualized as being probabilistic rather than absolute. For example, a given mother might have a certain probability of being in Cluster 1, a probability of being in Cluster 2, a probability of being in Cluster 3, and a probability of being in Cluster 4. If a particular mother is assigned to Cluster 2 in a k-means cluster analysis, then this is analogous to treating the four probabilities as having the values 0.0, 1.0, 0.0, and 0.0, respectively. This might be unrealistic. Perhaps the classification is not so clear cut and the probabilities of being in the four classes are more akin to 0.0, 0.55, 0.45, and 0.0 for the mother. In the cluster analytic literature, this view of classification is called *fuzzy clustering*. Ideally, we would take the uncertainty associated with fuzzy clustering into account when estimating parameters relating membership to other variables. A limitation of traditional applications of k-means clustering is that it does not. To be sure, the matter may not be problematic if the true assignment probabilities are well-differentiated. Nevertheless, recognition of the probabilistic nature of classification is important to keep in mind.

Strategies related to k-means based clustering that estimate fuzzy cluster probabilities have been developed (Kaufman & Rousseeuw, 1990). The methods can be used to calculate a *confusion matrix* to evaluate classification clarity. Consider a three cluster solution. For each individual, we assign the individual to the cluster that the individual has the highest probability of being in. For individuals classified into Cluster 1, we calculate the mean probability they received of being in Cluster 1, the mean probability they received of being in Cluster 2, and the mean probability they received of being in Cluster 3. Suppose the three mean probabilities are 0.92, 0.05, and 0.03, respectively. This suggests a well-differentiated cluster categorization. The process is repeated for each cluster, yielding the confusion matrix, an example of which is shown in Table 1. Larger values in the diagonal are better as are lower values in the off-diagonal.

Table 1: Confusion Matrix

	Mean Probability		
Cluster Classified Into	Cluster 1	Cluster 2	Cluster 3
Cluster 1	0.92	0.05	0.03
Cluster 2	0.05	0.90	0.05
Cluster 3	0.04	0.04	0.92

When the confusion matrix is well-differentiated, analyses that relate cluster memberships to other variables (ignoring the probabilistic nature of assignments) may not be problematic. If the confusion matrix is less favorable, then one could select only prototypical individuals for each cluster (e.g., those with a cluster probability of at least 0.90 in one of the clusters) and then explore membership correlates for that subset. Inferential tests, however, are still somewhat ad hoc and approximate. Alternatively, one can invoke a statistical theory that explicitly takes the probabilities into account (see below).

Additional Methods for Centroid Clustering

K means clustering performs best when the number of individuals in the population clusters is about the same (the proportion of individuals in a cluster is often referred to as *cluster density*) and the variables within a cluster are multivariately normally distributed. To be sure, it can handle some deviations from these properties, but depending on additional facets of the population structure, k-means clustering can mislead.

There are other forms of centroid clustering distinct from k-means. A method that is less sensitive to outliers than k-means analysis (and hence is a form of robust clustering) is called *partitioning around medoids* (PAM; Kaufman & Rousseuw, 1990). Rather than focusing on cluster means, this approach identifies an exemplar individual in the data for each cluster who is nearest the center of the cluster in the sense that the distance between the medoid (the prototypical individual) and all other individuals in the cluster is minimized. Another robust method is called *trimmed k-means cluster analysis* (Cuesta-Albertos et al., 1997; Garcia-Escudero et al., (2010), which trims an *a priori* specified amount of the data (e.g., 10%) prior to conducting k-means analysis. The method uses empirical criteria designed to clarify the cluster structure for determining which cases to trim, a process called *self-trimming* in the cluster analytic literature. A more general version of this algorithm is described in Fritz, Garcıa-Escudero and Mayo-Iscar (2012), which has the advantages of more readily accommodating unequal population cluster sample sizes as well as differing within-cluster covariance matrices (i.e., scatter) for the target variables. If one sets the trimming factor to zero, the approach yields a more general class of centroid clustering than traditional k-means modeling, albeit without robustification.

We think of k-means clustering as identifying well-separated and roughly equal-sized, spherical-shaped “blobs” of individuals in the population. More nuanced approaches use model-based perspectives that take into account distribution shapes and cluster intersection in different ways. For a description of these approaches, see Fraley and Raftery (2002) and Russell, Murphy and Raftery (2015).

MATTERS OF METRIC

When the variables being studied are on the same quantitative metric, application of cluster analytic methods is reasonably straightforward. However if the variables are on different metrics, the situation is more complex. For example, suppose we want to conduct a cluster analysis on 10 different personality traits, but the traits are measured on different metrics, with some scales ranging from 0 to 10, others from 10 to 50, and so on. In general, variables with larger variability due to differing metrics will dominate the cluster analysis. In many cases, this will produce artifactual results.

Some researchers deal with such scenarios by standardizing variables before conducting the cluster analysis. The analysis is then undertaken on the standard scores. Care must be taken when using this strategy because when we standardize variables, it has the effect of equating the variances of each variable in the set, i.e., all the X will have a variance of 1.0. Does a standard deviation carry the same meaning for each X? If the use of physical punishment has a small standard deviation, then does a standard score of 1.0 on it have the same meaning as a standard score of 1.0 for the use of guilt for which there is greater variability? For elaboration of this dilemma, see the primer on dominance analysis.

An alternative transformation that preserves variance differences is to re-score each variable so that all variable metrics range from 0 to 10, where 0 is the lowest possible response on the scale, 10 is the highest possible response, and 5 is the scale midpoint. Consider a variable whose response metric is from 10 to 30. First subtract the lowest possible score (in this case, 10) from each person's score so the metric now ranges from 0 to 20 rather than 10 to 30. Next, divide each person's transformed score by the highest score on the new metric (in this case, 20). Now the metric ranges from 0 to 1.0. Then multiply this result by 10. The new metric will be from 0 to 10. If you use this process for each variable, they all will be on a 0 to 10 metric but with unequal variances that may be more meaningful than the equal variance case.

As an aside, k-means cluster analytic strategies generally are not appropriate for variables with binary metrics (see, for example, the discussion by SPSS, 2015). An alternative is to use hierarchical clustering but with distance scores that are appropriate for binary variables, such as the *Jaccard coefficient* or the *matching coefficient* (see Everitt et al., 2011).

MIXTURE MODELING

A third approach to cluster analysis is called *mixture modeling*, of which latent class analysis (LCA) and latent profile analysis (LPA) are members. *Latent class analysis* is

the term used when the approach is applied to exclusively binary metrics. *Latent profile analysis* is the term used when the approach is applied to exclusively continuous metrics. When both binary and continuous variables are used, the generic term *mixture model* is used. These methods seek to group individuals into clusters based on the ability of the clusters to account for the correlations/covariances between the various X.

Mixture modeling is best explained using the framework of factor analysis (see the primer on factor analysis if you are unfamiliar with it). In traditional factor analysis, we specify a set of observed variables whose correlational pattern we are interested in explaining. If we fit a one factor model to the data, we are hypothesizing that the correlations between the variables are due to an unknown common cause that impacts each of them, per Figure 5.2. The unknown "factor" is assumed to be a continuous variable. According to the model in Figure 5.2, the correlation between, say, X1 and X2, can be explained by the fact that the latent factor influences both X1 and X2. If we were to somehow identify the factor, measure it, and partial it out or hold it constant, the correlation between X1 and X2 would vanish. As we seek to discern what the factor might be, we examine the magnitude and pattern of the factor loadings and, based on those loadings, deduce what the substantive content of the factor. With mixture modeling, we engage in the same process, but instead of the underlying factor being continuous, it is conceptualized as being categorical with an unknown number of levels.[1] For example, the Xs might represent the disciplinary styles identified earlier and our claim as theorists is that the correlations between them can be accounted for by an unknown categorical variable with an unknown number of levels that serves as a common cause to each of them. Each level of the underlying factor in a mixture model represents a subgroup or a "cluster." Just as we had to determine the number of clusters in k-means clustering, we also must do so in mixture modeling. Thus, we can fit a model where the underlying factor has 2 levels/clusters, another model where it has 3 levels/clusters, another model where it has 4 levels/clusters, and so on. For each model, we obtain an index of model fit that reflects how well the model reproduced the observed correlations between the variables (X1 to X8). We then compare the fits of the different models and choose the best fitting model. This defines the number of levels/clusters for the latent factor.

The indices for evaluating model fit are not the same as those described for k-means clustering because the fit function in mixture modeling is focused on reproducing the correlational structure among the variables, which is not the focus of k-means clustering. K-means clustering seeks to minimize within-cluster variance while maximizing between-cluster variance. The two methods are decidedly distinct.

[1] Technically, mixture modeling focuses on covariances rather than correlations, but we will refer to correlations because they are more familiar to readers and the basic ideas are the same.

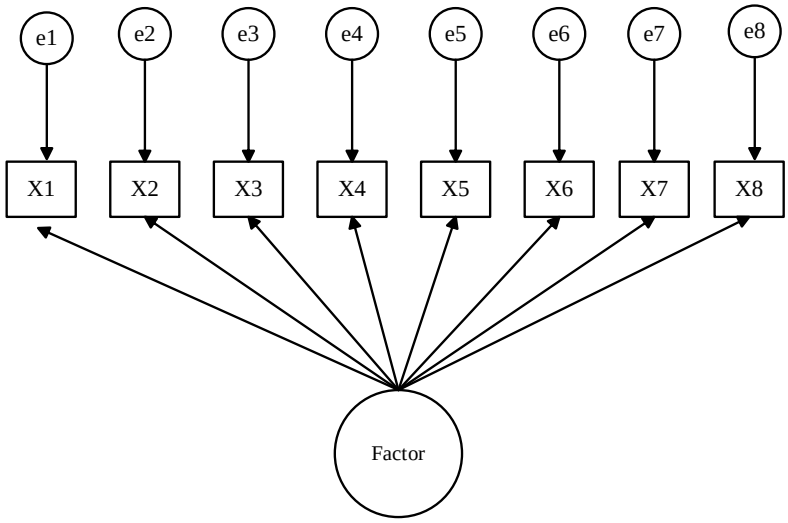

FIGURE 5.2. Traditional One Factor Model

Once the number of levels is determined, the mathematics of mixture modeling allow us to estimate the mean score of each X for each level of the factor. Also, like k-means analysis, each individual in the sample is classified into a given cluster. However mixture modeling uses the logic of fuzzy clustering, so individuals are assigned a distinct probability of being in each cluster, with the probabilities summing to 1.00. The individual is "assigned" to the category that has the highest probability of the individual being in.

Two indices of model fit used in mixture modeling that map onto those used in mixture modeling are the AIC and BIC. Mixture modeling also yields a confusion matrix that can be used to evaluate the differentiation of cluster probabilities. Finally, mixture modeling allows one to use the *Vuong-Lo-Mendell-Rubin test*. This is a significance test that compares the fit of a model with the fit of a model with one less cluster. If the p value is statistically significant, then this means that the model with more clusters fits the data better than the model with one less cluster. The idea is not to add clusters that do not show significant improvement in fit based on this test (see Nylund, Asparouhov & Muthén, 2007).

. As with k means cluster analysis, another consideration in determining the number of levels of the underlying factor is the substantive meaningfulness of the results. For example, adding another level might make the model fit better but if the new level/cluster does not make substantive sense, one might be hesitant to add it.

Once a final model is settled upon, mean values for each X variable for each cluster

are reported by the mixture model output and the clusters are interpreted accordingly. Mixture models also provide estimates of the proportion of the population that is in each subgroup.

An advantage of the mixture modeling approach is that it can be used with binary or continuous measures or any combination of them. For continuous measures, the metrics do not have to be comparable. When predictors and outcomes of cluster membership are modeled, a well-developed statistical theory for taking into account the uncertainty associated with cluster assignment is available. For more background on mixture modeling as applied to cluster analysis, see Vermunt and Magidson (2002) and Finch and Bronk (2011).

CLUSTER REPLICATION

Because of its exploratory nature, it generally is useful to replicate one's results from both k-means and mixture modeling approaches with one or more independent samples.

SUMMARY AND CONCLUDING COMMENTS

Cluster analysis is a useful method for conducting profile analyses across sets of variables and identifying meaningful subgroups relative to those profiles. There are many different types of cluster analysis, with the most well-known ones being hierarchical clustering, centroid clustering, and mixture modeling (latent class analysis and latent profile analysis). Hierarchical clustering relies on the use of distance scores (of which Euclidean distance scores are among the more popular), k-means clustering relies on minimizing within-cluster variability relative to cluster means, and mixture modeling posits underlying categorical latent factors that are thought to serve as common causes of the variables in the profile analysis. The former two methods are usually applied to variables with a common metric, although standardization and transformation strategies can be used to adjust for metric differences. Mixture modeling can be applied to variables with varying metrics.

A key issue in all cluster analyses is how to choose the number of subgroups/clusters that account well for the data. The different clustering methods use different criteria for doing so. K-means cluster analysis relies on indices like the percent of variance account for by the clusters, the sum of squares within, the AIC, the BIC, and the average silhouette width, among others. Mixture models rely on the AIC and BIC, confusions matrices, and the Vuong-Lo-Mendell-Rubin test, among others. All methods also take into account the substantive meaning of the clusters when making decisions about the number of clusters to select.

When classifying individuals into subgroups or clusters, the traditional k-means approach uses an absolute assignment method to clusters whereas mixture models and other forms of centroid clustering use fuzzy clustering concepts. Given the probabilistic nature of classifications, it is important to keep in mind the uncertainty associated with classifications. K-means modeling works best when the population clusters are well-separated, about equally-sized, have spherical shapes, and outliers are absent. More recent approaches using trimmed data (Fritz et al., 2012) are generally more flexible.

Because of its exploratory nature, it generally is important to replicate one's results from both k-means and mixture modeling approaches with one or more independent samples.

Cluster analysis has been applied to a wide range of contexts using both cross-sectional and longitudinal data. For the latter, various forms of transition analysis are of growing interest.

REFERENCES

Burnham, K. & Anderson, D. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods and Research*, 33, 261-304.

Cuesta-Albertos, J.A., Gordaliza, A. & Matran, C, (1997). Trimmed k-Means: An attempt to robustify quantizers. *The Annals of Statistics*, 25, 553–576.

Everitt, B., Landau, S., Leese, M. & Stahl, D. (2011). *Cluster analysis*. New York: Wiley.

Finch, W. & Bronk, K. (2011). Conducting confirmatory latent class analysis using Mplus. *Structural Equation Modeling*, 18, 132–151.

Fraley, C. & Raftery, A.E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97, 611-631.

Fritz, H., Garcıa-Escudero, L. & Mayo-Iscar, A. (2012). tclust: An R package for a trimming approach to cluster analysis. *Journal of Statistical Software*, 47, 1-26.

Garcia-Escudero, L.A., Gordaliza, A, Matran, C, & Mayo-Iscar, A, (2010). A review of robust clustering methods. *Advances in Data Analysis and Classification*, 4, 89–109.

Hennig, C., Meila, M., Murtagh, F. & Rocci, R. (2015). *Handbook of cluster analysis*. New York: Chapman & Hall.

Kaufman, L. & Rousseeuw, P.J. (1990) *Finding groups in data: An introduction to cluster analysis*. Wiley, New York.

Kuha, J. (2004). AIC and BIC : Comparisons of assumptions and performance. *Sociological Methods & Research*, 33, 188-208.

Nylund, K., Asparouhov, T. and Muthén, B. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling*, 14, 535–569

Pelleg, D. & Moore, A. (2000). X means: Extending K-means with efficient estimation of the number of clusters. At www.aladdin.cs.cmu.edu/papers/pdfs/y2000/xmeans.pdf.

Raftery, A.E. (1995). Bayesian model selection in social research (with discussion). *Sociological Methodology*, 25, 111-195.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational & Applied Mathematics*, 20, 53-65.

Sclove, L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, 2, 333-343

SPSS (2015). Clustering binary data with k-means (should be avoided). Retrieved on August 30, 2015 from http://www-01.ibm.com/support/docview.wss?uid=swg21477401.

Russell, N., Murphy, T.B. & Raftery, A.E. (2015). Bayesian model averaging in model-based clustering and density estimation. Technical Report no. 635, Department of Statistics, University of Washington.

Tein, J., Coxe, S. & Cham, H. (2013). Statistical power to detect the correct number of classes in latent profile analysis. *Structural Equation Modeling*, 20, 640–657.

Vermunt, J. & Magidson, J. (2002). Latent class cluster analysis. In Hagennars, J. and McCutcheon, A. (Eds.) *Applied latent class analysis*. Cambridge, UK: Cambridge University Press.

Wasserman, L. (1997). Bayesian model selection and model averaging (Working Paper No. 666). Carnegie Mellon University, Department of Statistics.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? *Biometrika*, 92, 937-950.

APPENDIX: INFORMATION INDICES FOR MODEL CHOICE

When choosing between the different models to determine the number of classes, a commonly used set of comparative fit indices is based in a statistical theory known as *information theory*. Two such indices are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). In general, researchers calculate an AIC index and/or a BIC index for the different models and then choose the model that has the best BIC or AIC value. In this appendix, we develop the logic of these indices, taking a few liberties in the interest of pedagogy. We first develop the concept of a log likelihood, a concept that is central to both the AIC and BIC. We then describe the model comparison process for the AIC, followed by consideration of that process for the BIC.

Log Likelihoods

Suppose we have a very large population and half the population is male and half the population is female. The probability of a randomly selected case being a male is 0.50 and this also is true for being a female. Stated more formally:

p(male) = 0.50 p(female) = 0.50

If we randomly select two cases, the probability of a given joint result across the two selections or "trials" is the product of their probabilities. As such, the probability of observing two males is

p(male)*p(male) = (0.50)(0.50) = 0.25

This is known as the multiplication rule for independent trials. Stated more formally, let p(A) = the probability of event A on a trial and p(B) = the probability of event B on a second (independent) trial. The joint probability of both events A and B is the product of the individual probabilities p(A) p(B). To be more concrete, there are four combinations that can result, each with a probability of 0.25:

Probability of a male on the first trial followed by a male on the second trial: 0.25
Probability of a male on the first trial followed by a female on the second trial: 0.25
Probability of a female on the first trial followed by a male on the second trial: 0.25
Probability of a female on the first trial followed by a female on the second trial: 0.25

and if we do not care about the order of appearance in the trials,

Probability of two males: 0.25
Probability of a male and a female: 0.50
Probability of two females: 0.25

We now shift gears review another facet of statistical theory that we will make use of. If we know that a very large set of scores is normally distributed with a certain mean and standard deviation, then we can use knowledge of the probability density function for a normal distribution to compute the probability of obtaining any given value when we randomly select a case from that distribution. The density formula is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \; e^{\frac{-.5(x-\mu)^2}{\sigma^2}}$$

where x is the score value in question, μ is the mean of the distribution, σ is the standard deviation of the distribution, π is the mathematical constant pi, *e* is the constant associated with the Naperian logarithm, and the density describes the height of the normal curve at the value of x. We can use this density in conjunction with calculus to calculate the probability of observing the score in question. As an example, if scores are normally distributed with a mean of 100 and a standard deviation of 13.77, then, using the above formula, we find that the likelihood of a score of 99 is 0.0289. For a score of 87, it is 0.0186.[2]

Suppose we randomly select two scores from an extremely large population where scores are normally distributed with a mean of 100 and a standard deviation of 13.77. The probability that the scores will be 87 and 99, using the joint probability theorem described above, is (0.0289)(0.0186) = 0.00053754. Stated another way, the probability of observing these two data points given that the mean is 100 and the standard deviation is 13.77 (and assuming a normal distribution) is 0.00053754, with further adjustments to account for disinterest in the order of selection.

Suppose we randomly sample 100 data points from the population and calculate the likelihood of those 100 data points occurring using a strategy similar to the above method. The strategy would involve multiplying each probability by one another, with the result being a very, very small number. To make things more manageable and so as not to work with such small numbers, statisticians transform the final result by calculating the log of it, yielding what is called a *log likelihood*. The log likelihood is

[2] Technically, the probability of observing an exact value for a continuous variable is zero. We compute the likelihoods here by focusing on the interval defined by the real limits of the number (e.g., 98.5 to 99.5) in conjunction with the integral that scales the area under the curve to 1.00.

indicative of (but not equal to) the probability of obtaining the sample data given a "model" that states (a) the scores are normally distributed, (b) the mean is 100, and (c) the standard deviation is 13.77.

Log likelihoods are negative because the log of numbers less than 1.00 is always negative. For example, the natural log of 1.00 is zero, the natural log of 0.50 is -0.69, the natural log of 0.25 is -1.39, and the natural log of .01 is -4.61.[3]

Now, let's turn the above situation on its head. Suppose we have a set of 100 data points but we do not know the mean and standard deviation of the (assumed normal) distribution from which they come. We might, based on theory or logic, decide to "test" a model that states the mean is 95 and the standard deviation is 15. Using the probability density function from above and the strategies described, we can calculate the log likelihood for this model. The closer the log likelihood value is to zero (i.e., the less negative it is), the more likely the data came from the postulated model. We might formulate a second (competing) model that the mean is 100 and the standard deviation is 13.75 and calculate the log likelihood for it. Again, the closer the value of the log likelihood for this model is to zero, the more likely it is the data came from the model positing a mean of 100 and a standard deviation of 13.75.

We can compare the log likelihood values for the two models and we might find that one model results in a log likelihood closer to 0 than the other model. The model with the log likelihood closer to zero is more likely to have produced the data, hence we would prefer it to the model with the more negative log likelihood. Such is the fundamental logic of choosing between models based on their relative log likelihoods: We calculate the log likelihood of competing models and then choose the model with the log likelihood that is closest to zero. To be sure, the above explanation is simplistic and glosses over technicalities, but hopefully it conveys the general idea of comparing log likelihoods for two models.

As an aside, the above logic also is central to the well-known method of estimation called *maximum likelihood estimation*. In this approach, to estimate the mean of a distribution, one conceptually posits different models each representing a possible population mean value, calculates the likelihood of observing the data given the "model," and then selects the value/model that has the maximum likelihood.

[3] Actually, some operationalizations of log likelihoods can yield positive numbers, but discussion of this point is beyond the scope of this primer.

Model Comparisons using the AIC

The AIC is an index of model likelihood or "model fit" based on a log likelihood. A common representation of it is

$$AIC = (-2)\ (LL) + 2k \qquad [1]$$

where LL is the log likelihood associated with the model in question and k is the number of estimable parameters in the model (such as when we estimate an intercept and the various regression coefficients). By multiplying the log likelihood by -2, the AIC essentially becomes a positive number, with larger numbers indicating lower likelihoods of the model. The AIC also includes what is often referred to as a penalty function for lack of parsimony, namely 2k. If the model has many parameters in it that must be estimated, then the AIC will be larger, everything else being equal. With the AIC, model parsimony is rewarded.[4] In general, the smaller the value of AIC, the better the "fit" of the model to the data. To make this intuitive, if the probability of the data given the model is 0.25, the log likelihood will be -1.39 and multiplying this by -2 yields 2.78. If the probability of the data given the model is much higher, say 0.50, the log likelihood is -0.69 and multiplying this by -2 yields 1.38. So, the smaller the value, the better the model. To this term, a penalty function is added that inflates the value of AIC for models that estimate more parameters.

There are many variations of the AIC. For example, some researchers use the above formula but with a small sample bias correction incorporated into it. This is sometimes referred to as AIC_c. The nuances of the different versions of the AIC are described in Burnham and Anderson (2004). Do not be surprised if for some software you observe AIC indices that are quite different in magnitude from other software. The important idea for all them is that we can compare different models using their respective AICs and then choose models that have "better" AICs when compared to other models.

Sometimes we compare more than two models, i.e., we might compare three, four or five models. When comparing more than two models, it is common to first identify the model with the lowest AIC value (which is the best fitting model of all the models being considered). One then calculates the difference in AIC values between each of the models and this best fitting model (subtracting the latter from the former). For the best fitting model, the difference will be zero and for all other models, it will be positive in value,

[4] Technically, the 2k term is part of the mathematical theory underlying the derivation of AIC. Also, choosing the value of -2 to multiply the LL by is not arbitrary. This value has a clear rationale. See Burnham and Anderson (2004).

with the larger the disparity, the worse the fit of the target model relative to the best fitting model.

General rules of thumb have been proposed to contextualize the magnitude of the difference in AICs between models (see Burnham & Anderson, 2004). The most common rules of thumb are as follows:

1. If the disparity in AICs is < 2, then the two models have about the same support

2. If the disparity in AICs is > 2 and < 4, then the better fitting model has positive support relative to the model it is compared with

3. If the disparity in AICs is > 4 and < 10, then the better fitting model has strong support relative to the model it is compared with

4. If the disparity in AICs is > 10, then the better fitting model has very strong support relative to the model it is compared with.

Of course, one must be careful when applying rules of thumb like this because they may not apply in all contexts. Indeed, some analysts object to their specification, arguing that they can result in the same rigid and counterproductive use of a criterion like "$p < 0.05$" that plagues null hypothesis testing frameworks.

Another standard for comparing two models vis-a-vis the AIC is to examine what is called the *evidence ratio*. Let D = the AIC for the worse fitting model of the two models minus the AIC for the better fitting model of the two models (and let *e* be the traditional Naperian constant). The evidence ratio is defined as

$$ER = 1 / e^{(-D / 2)}$$

where ER stands for "evidence ratio." It indicates how much more likely the better fitting model is (given the data) than the worse fitting model (given the data). For example, if the AIC for the better fitting model is 100 and for the worse fitting model it is 102, then the evidence ratio is

$$1 / e^{-(102-100) / 2)} = 2.63$$

The better fitting model is 2.63 times more likely to have yielded the data than the model it is being compared with.

Finally, some researchers normalize AIC differences relative to all models being compared so that they sum to 1. These are called *Akaike weights* and indicate the "weight

of evidence" in favor of a model relative to *all* models in the comparison set. Akaike weights are distinct from evidence ratios because Akaike weights are impacted by the particular set of models being compared when the number of models is greater than two. Let us first describe how Akaike weights are calculated and then we will make them more concrete with an example.

To calculate the Akaike weight, each model is assigned an index of its likelihood relative to that of the best fitting model using the value from the denominator of the evidence ratio, $e^{(-D/2)}$, as the index. Let T = the sum of the $e^{(-D/2)}$ values across all the models being considered. Then the Akaike weight for a given model is defined as

$e^{(-D/2)} / T$

The weight ranges from 0 to 1.00, with higher values favoring the model in question.

To make this concrete, suppose we fit five different models to a set of data. Here is a table with the AICs, the differences between the model AIC versus the model with the lowest AIC, and the Akaike weights (w):

Model	AIC	D	$e^{(-D/2)}$	$w = e^{(-D/2)}/T$
1	204	2	0.3678	0.2242
2	202	0	1.0000	0.6094
3	206	4	0.1353	0.0824
4	206	4	0.1353	0.0824
5	214	12	0.0024	0.0015
Sum			T = 1.6408	1.0000

The sum of the weights across all five models is 1.00. The weights represent a continuous measure of relative strength of evidence for each model. Each weight can be crudely interpreted as the probability that the model is the best model among the set. In the present case, the data support Model 2.

The basic idea when evaluating models is to examine multiple criteria, including the magnitude of the difference in AICs, the evidence ratios, the Akaike weights, and the substantive meaning/logical coherence of the models, in order to choose the best one.

Model Comparisons using the BIC

We describe the logic of the BIC using the Schwartz BIC, which is formally defined as

$$BIC = -2\ LL + \ln(N)\ k \qquad [2]$$

where k = the number of estimable parameters in the model, N = the sample size, and LL = the model log likelihood. Like the AIC, the smaller the BIC, the better the model fit, everything else being equal. Like the AIC, there is a penalty function for lack of parsimony, but the penalty is different than the AIC. The penalty is somewhat harsher for the BIC as opposed to the AIC. There are other instantiations of the BIC, and we discuss these below. For current purposes, we use the Schwartz formulation.

Like the AIC, it is not uncommon for the model with the smallest BIC to be used as a reference point for comparing models, with a common practice being to calculate the difference between each model in the model set and the model with the best BIC, like we did for the AIC. For the best fitting model, this difference will be zero.

To evaluate models in terms of BIC differences, general rules of thumb are (see Raftery, 1995):

1. If the BIC disparity < 2.2, then the better fitting model and the model it is compared with have about the same support

2. If the BIC disparity > 2.2 and < 6, then the better fitting model has positive support relative to the model it is compared with

3. If the BIC disparity > 6 and < 10, then the better fitting model has strong support relative to the model it is compared with

4. If the BIC disparity > 10 then the better fitting model has very strong support relative to the model it is compared with

For similar but slightly different standards, see Wasserman (1997).

One also can calculate what is called a *Bayes Factor* (BF) for each model relative to the best fitting model. It is defined as

$$BF = e^{(D'/2)}$$

where D' is the BIC difference between the target model and the best fitting model. The Bayes factor is the probability that the model with the lower BIC produced the data divided by the probability the model in question produced the data. For example, a BF = 10 means it is 10 times more likely the model with the minimum BIC produced the data than the model in question.

Finally, a relative model weight, analogous to the Akaike weight, can be computed by normalizing model likelihoods relative to *all* models in the comparison set so that they sum to 1. Let D = the difference in the BIC for the model in question minus the value of the BIC for the best fitting model, T = the sum of the index $e^{(-D/2)}$ across each model. The relative weight for a model is

$e^{(-D/2)} / T$

The weight ranges from 0 to 1.00, with higher values favoring the model. Again, the sum of the weights across models is 1.00.

As with the AIC, the basic idea when evaluating models is to examine multiple criteria, including the magnitude of the difference in BICs, the Bayes factors, the relative weights, and the substantive meaning/logical coherence of the models, in order to choose the best one.

You will encounter variants of the BIC, but the basic logic in applying them is the same. For example, like the AIC_c, there is a sample size adjusted BIC that is similar to Schwartz' BIC, but it applies a somewhat milder penalty function (Sclove, 1987). There also are variants of both the AIC and BIC to deal with dispersion issues in count regression models (called QAIC and QBIC).

Which Method is Better, AIC or BIC?

A debated topic in statistics is which approach to model comparison is better, one based on AICs or one based on BICs. There are advocates on both sides of the matter and we dare not venture into this controversy here. The BIC tends to favor simpler models more so than the AIC. This can be both a strength and a weakness. Interested readers are referred to Burnham and Anderson (2004), Yang (2005), and Kuha (2004). Kuha argues for the use of both indices.

An issue with both approaches is that researchers can be lulled into thinking that the best fitting model within a set of models is the true model. This is not necessarily the case. Researchers can choose the best of a set of wrong models, which is not our goal.

In mixture modeling, the choice of the number of latent levels/clusters for a factor is often guided by the AIC and BIC values of the models with differing numbers of levels/clusters.