

Chapter 2

Statistical and Methodological Considerations When Using Cluster Analysis in Neuropsychological Research

Chad L. Cross

Introduction

Multivariate data are inherently complex. Owing to that complexity, it is often desirable to find relationships among a suite of variables from which patterns or structures can be determined either to gain a more thorough understanding of outcome variables or to develop groups that can be subjected to further analyses. For example, for disorders like traumatic brain injury or schizophrenia where there is no prototypical neuropsychological profile, one might desire to determine whether there were homogeneous subgroups of patients that would be identified by their neuropsychological test performance. If identified, these neuropsychological subgroups might evince different outcomes, responses to treatment, and underlying neuropathology. In this example, developing subgroupings of homogenous entities (patients) involves determining similarities and differences among neuropsychological variables in multivariate space. One type of analysis, cluster analysis, is well suited for this task and has been used widely among disparate disciplines such as biology (Eisen, Spellman, Brown, & Botstein, 1998; Jiang, Tang, & Zhang, 2004), computer science (Wallace, Keil, & Rai, 2004), education (Myers & Fouts, 1992), marketing (Punj & Stewart, 1983), and neuropsychology (Allen, Goldstein, & Warnick, 2003; Allen et al., 2010; Goldstein, 1990; Thaler et al., 2010).

The views presented in this chapter are those of the author(s) and do not necessarily represent the views of the US Department of Veterans Affairs.

C.L. Cross, Ph.D., P.Stat®, L.C.A.D.C., M.F.T. (✉)

Veterans Health Administration, Office of Informatics and Analytics, Las Vegas, NV, USA

School of Community Health Sciences, University of Nevada, Las Vegas,

4505 Maryland Parkway, Las Vegas, NV 89154, USA

e-mail: chad.cross@va.gov

Cluster analysis is generally considered a grouping technique as opposed to a classification technique—the latter of which is classically reserved for those methods used for developing a model based on data in which group membership is already known, and the task of the researcher is to find a way to reliably classify new objects into one of the known groups (Johnson & Wichern, 2007). A practitioner, for example, may wish to determine a reliable way to classify patients with right and left hemisphere lesions using neuropsychological tests, and to do so might use test data collected on patients with right and left hemisphere lesions to develop a prediction formula for such purposes using classification techniques. Two familiar examples of classification techniques are discriminant function analysis (DFA) (Betz, 1987; Johnson & Wichern, 2007) and polytomous logistic regression (Cross & Petersen, 2001; Hosmer & Lemeshow, 2001). Clustering, or more generally, grouping techniques, is used to develop subclasses of relatively homogeneous entities based on a set of variables and a given set of rules and algorithms for developing groups.

Cluster analysis is also typically concerned with grouping objects (patients in the above examples), as opposed to focusing on finding groupings of variables, the latter of which is the focus of correlation-type analyses (Rodgers & Nicewander, 1988) or various multivariate partitioning methods such as factor analysis (Tabachnick & Fidell, 2007).

The aim of cluster analysis, then, is to find the best possible grouping of objects such that the degree of association between objects within a given cluster is maximized and for which any other grouping with the same set of objects would decrease this degree of association. Other definitions and extensions of this definition can be found in books specializing in cluster analysis techniques (e.g., Everitt, Landau, Leese, & Stahl, 2011), but this broadly describes the idea behind cluster analysis methodology.

As a simple example of how cluster analysis may work in a given circumstance, consider Fig. 2.1, a neuropsychological example similar to the playing card example commonly used to illustrate clustering (Johnson & Wichern, 2007, p. 672). In the example, the two brain hemispheres and four brain lobes are used to develop various subgroups of patients. What is immediately apparent when examining this figure is that there are multifarious ways in which to cluster a given set of patients. In fact, finding the number of ways in which to parse a set of n objects into r nonempty subsets with $i=1, \dots, r$ groups of varying sizes is a common problem in combinatorial mathematics and is easily calculated as a Stirling number of the second kind using the following formula (Sharp, 1968):

$$\left(\frac{n}{r} \right) = \frac{1}{r!} \sum_{i=1}^r (-1)^i \binom{r}{i} (r-i)^n \quad (2.1)$$

Using this formula or one of the many online applets for calculating Stirling numbers, one finds that there are 127 ways to partition the brain hemisphere/lobe example into two groups, 966 ways to partition into three groups, 1,701 ways to partition into four groups, and onward. The purpose of this illustration is to point out both the potential complexity involved in finding cluster solutions and the

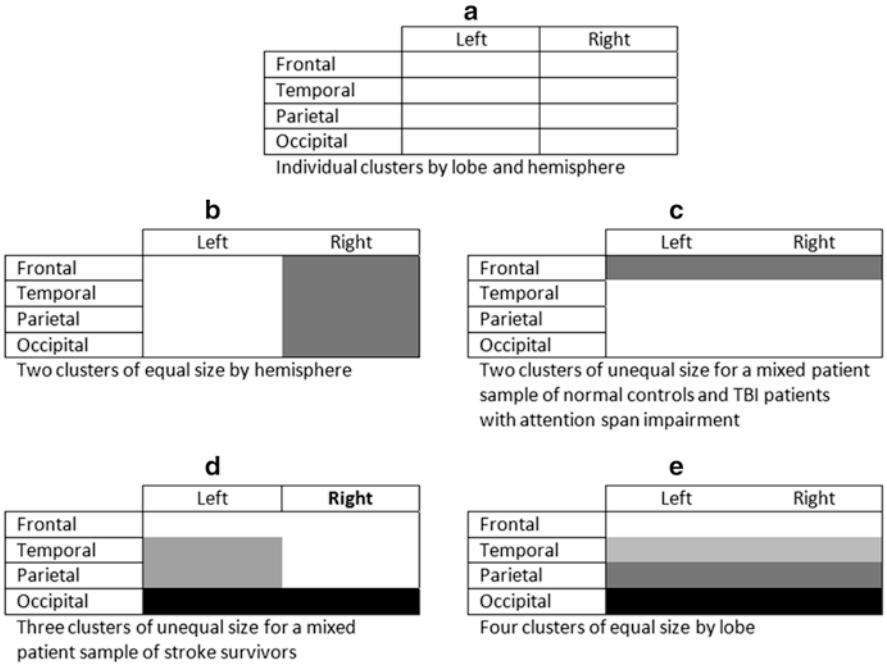


Fig. 2.1 Illustrative example of potential clustering solutions under a variety of scenarios. This example considers the two hemispheres and four lobes of the brain. Clusters of equal size are seen in scenarios A, B, and E, and clusters of unequal size are seen in scenarios C and D. The number and relative size of clusters in these examples depends on the unique set of circumstances surrounding the magnitude and type of input variables used to parse the data into various solutions. See Johnson and Wichern (2007) for an analogous example using playing cards

extensive calculations that must be employed in order to find the most parsimonious solution to a clustering problem.

The goal of this chapter is to provide a review of clustering methods, including hierarchical agglomerative methods and iterative partitioning methods. Recommendations for determining the appropriate number of clusters and for comparing clustering methods also will be discussed. Further, validation techniques will be addressed. The chapter will conclude with a discussion of data issues commonly encountered in neuropsychological research, such as non-normality of data and incomplete data records from patients, and techniques for handling these situations.

This chapter is not meant to be an exhaustive treatment of cluster analysis, but rather is intended to cover common topics and issues in cluster analysis, much as the extensive neuropsychology application article by Morris, Blashfield, and Satz (1981) did 30 years ago. For the interested reader, there is an excellent reference work by Everitt et al. (2011). As a final note, the purpose of this chapter is to provide an application overview of a very broad area of statistics. In order to do this, some equations are provided throughout the text to illustrate particular points; most readers

will be using computer software packages to perform cluster analysis and likely will read over most of the equations and rely on the textual explanation. For the more mathematically curious, several references are provided throughout the chapter.

Association and Similarity

The idea of finding associations or some degree of similarity among objects is not a new idea by any means. Finding co-occurring relationships (i.e., co-relations or correlations) among variables was first approached as an interesting problem by Sir Francis Galton in 1885 and was later formalized into common usage and formulation by Pearson some years later. An interesting review of the various formulations and usages of correlation as a measure of association can be found in Rodgers and Nicewander (1988); interestingly, Pearson himself published a lengthy review of the concept in 1920 (Pearson, 1920) noting among other things that there are a vast number of ways to measure correlation. Regardless of the various formulations for measures of association, there would be no such measures if there were not an innate sense that some things simply share a degree of association that is greater than would be expected by chance.

Visualization

Because of the ability of humans to naturally discriminate, clustering objects together seems a natural task (Everitt et al., 2011). The idea of developing a visual way of displaying information is often one of the first steps to accomplish when presented with novel data—graphing data before doing anything else is, in a casual sense, a very useful way to prevent us from doing something silly (e.g., reporting a positive relationship when a scatter plot clearly depicts a negative one). Indeed, if simple graphical analysis does not reveal interesting patterns, there is often little reason to pursue more formal analyses. However, this is not always the case, as aptly stated by Huff (1954), “Averages and relationships and trends and graphs are not always what they seem. There may be more in them than meets the eye, and there may be a good deal less” (p. 8).

Discerning relationships in multiple dimensions, as is often necessary in data analysis, can be difficult, however. Consider Fig. 2.2, where a two-dimensional graph is shown that relates a measure of attention span and age. In this figure there seems to be a general relationship between attention span and age. Additionally, there appear to be three distinct groupings of subjects of similar ages. Further analysis of these data may in fact provide reason to believe that there are distinct clusters based on these two variables, or it may be that gaps in the measurement of certain ages has led to the resulting graph. Higher dimensions—even three dimensions—can be more difficult to visualize. Consider Fig. 2.3, where a three-dimensional

Fig. 2.2 An illustration of cluster visualization in two dimensions. Clusters of objects are easy to discern in this figure, and hence one would likely investigate this relationship further—particularly in terms of including more subjects to occupy the age gaps in the data

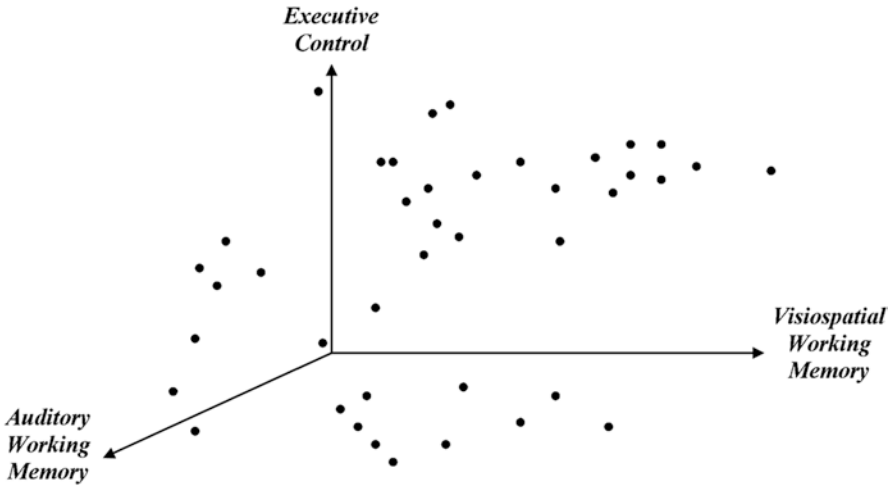
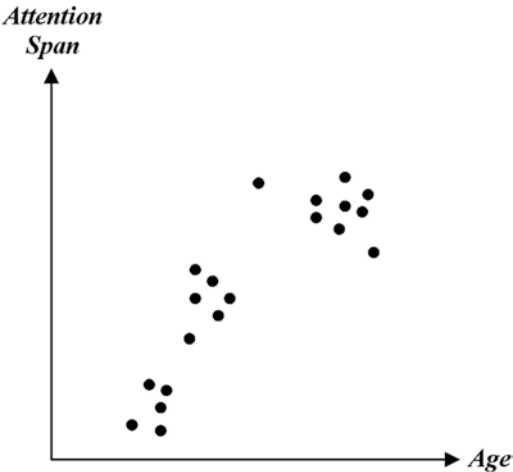


Fig. 2.3 An illustration of data visualization in three dimensions. Though some patterns are discernible (e.g., a positive relationship between executive control and visuospatial working memory), it is difficult to assess if true groupings exist without the ability to rotate the graph along its axes

figure shows subjects evaluated for executive control, auditory working memory, and visuospatial memory. The ability to discern objects in space is much more difficult unless one can rotate the figure on its axes.

The task of discerning relationships among sometimes dozens of variables for hundreds or thousands of subjects is clearly beyond the scope of traditional data visualization. The use of multivariate techniques such as multidimensional scaling (Johnson & Wichern, 2007) can be useful, as can developing a series of 2-D or 3-D graphs. An interesting development in high-dimensional visualization, Orca, was

Table 2.1 An illustration of possible binary outcomes when comparing two patients using a 2 × 2 table

| | | Patient <i>j</i> | | Total |
|------------------|---|------------------|---------------|------------------|
| | | 1 | 0 | |
| Patient <i>i</i> | 1 | Y_{11} | Y_{10} | $Y_{1\cdot}$ |
| | 0 | Y_{01} | Y_{00} | $Y_{0\cdot}$ |
| Total | | $Y_{\cdot 1}$ | $Y_{\cdot 0}$ | $Y_{\cdot\cdot}$ |

Where Y_{ij} represents the frequency of (i, j) matches across a set of variables and dot notation is used to represent the sum of frequencies across either rows (i) or columns (j)

developed by Lumley (2001) and is available for the R statistical software environment. Sarkar (2008) also developed lattice graphics for R for visualization. Some have also used principal components analysis (Tabachnick & Fidell, 2007) to reduce dimensionality for graphical representation, an idea termed “projection pursuit” by Everitt et al. (2011).

Quantitative Measures of Association

As mentioned previously, most often for cluster analysis in neuropsychological research, the interest lies in finding homogeneous groupings among objects (e.g., patients or research subjects) as opposed to focusing on finding groupings of variables, the latter of which is the focus of correlation-type analyses (Rodgers & Nicewander, 1988) or various multivariate partitioning methods such as factor analysis (Tabachnick & Fidell, 2007). The literature is replete with various methods of measuring association, and it is not uncommon in the literature and in various textbooks, to see references to measures as “similarities,” “dissimilarities,” “association,” or “proximity.” Regardless of the name, the idea is that there is a mathematical method to measure spatial relationships in a way that can be used as a criterion for developing groupings. Also, various measures of association have been developed to address different types of data including binary, categorical, and continuous.

Measures for Binary Data

Binary data are often collected on research subjects. Subjects may be those with or without a given condition, those meeting a certain diagnostic suite and those that do not, and so on. Consider Table 2.1. In this table, patient i and patient j are measured on a set of binary variables, and the frequency of matches is tallied; the statistic of interest is then based on various ways of using these tallied results; for example, counting patients along the diagonal (i.e., those that match) or those along the off-diagonal (e.g., those that differ). Many authors have developed measures of binary association, many of which can be found in Johnson and Wichern

(2007) or Everitt et al. (2011). A very extensive review and documentation of these measures is provided in Gower and Legendre (1986).

A common measure of similarity is to consider examining total matches for each patient as a function of the overall total:

$$\text{Similarity}(i, j) = \frac{Y_{11} + Y_{00}}{Y_{..}} \quad (2.2)$$

Equation 2.2 has intuitive appeal, as it uses all of the available data from Table 2.1 and is easily interpretable. Here, the similarity between patient i and patient j is represented as the sum of all matches proportionate to the total count. One can immediately see that a measure of similarity is functionally equivalent to a measure of dissimilarity by subtracting it from unity:

$$\text{Dissimilarity}(i, j) = 1 - \frac{Y_{11} + Y_{00}}{Y_{..}} = 1 - \text{Similarity}(i, j) \quad (2.3)$$

Other common measures ignore all cases where both patient i and patient j have (0, 0) matches such that the focus is only on what is relevant to either patient as opposed to what is not.

Measures for Categorical Data

Categorical variables can be classified in a similar fashion as binary ones, with summation occurring two categories at a time. However, as Everitt et al. (2011) points out, this may be problematic in some cases, particularly when there are many non-matches for the patients being compared. For that reason, it is suggested by that author, and we agree, that it makes more intuitive sense simply to allocate a binary (0, 1) score to each variable and then average them. For k variables:

$$\text{Similarity}(i, j) = \frac{\sum_{k=1}^p Y_{ijk}}{p} \quad (2.4)$$

Similarity is then the sum of all 0 and 1 scores for each item, k , averaged over all p variables. Measures for categorical data of this type can be found in the genetics literature, where dissimilarity measures are commonly used for lineage analysis (e.g., Tajima, 1993; Tamura et al., 2011).

Measures for Continuous Data

As with the measures for binary data, there are many options available for continuous data (see Everitt et al., 2011 for a fairly extensive list with formulations). These metrics generally rely on some measure of spatial distance or some measure of correlation among variables, with subjects assigned to groups in which their

measures are most similar. By far the most common measure of distance is the Euclidian distance, which most everyone learns as an extension of the Pythagorean theorem in basic algebra (i.e., $a^2 + b^2 = c^2$) to find the distance between points in a Cartesian plane. This is easily extended to points in multivariate space, where objects nearer in space are more similar than those farther apart in space. For k variables, a generalized distance between two individuals (i, j) is:

$$\text{Distance}(i, j) = \left[\sum_{k=1}^p |x_{ik} - x_{jk}|^r \right]^{1/r} \quad (2.5)$$

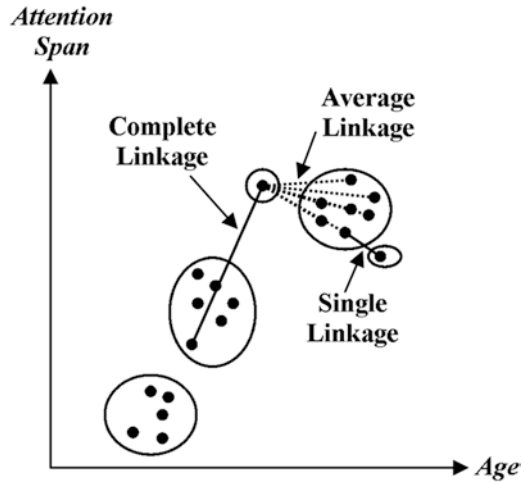
If we make $r=2$, we obtain the standard Euclidean distance as the square root of the sum of squared differences. Interestingly, if we make $r=1$, we obtain a measure termed the “city block” or “Manhattan” distance, as it represents distances measured in a street-like configuration, and making $r=3$ results in a measure termed the “Minkowski” distance (Everitt et al., 2011; Johnson & Wichern, 2007). By far the most intuitive and most often used (in fact it is often the default measure in many statistical packages) is the Euclidean distance. It should also be noted that one could calculate similar measures for binary data by assigning a score of “0” if subjects match on given variable and “1” otherwise such that the distance measure provides a mismatch score (Johnson & Wichern, 2007).

Measures for Combined Data Types

In some cases, data of a mixed type are collected for research subjects. This may happen, for example, if you collect a demographic questionnaire on your subjects (e.g., gender, age category, or income category) along with standardized neuropsychological tests that produce a continuous-type score, either raw or standardized. Can both data types be used simultaneously in a single run of a cluster analysis? The answer is yes but with mixed results. This demonstrates the idea of commensurability in data analysis, wherein it is often desirable to have the same measurement scale among variables.

A simulation study by Bacher, Wenzig, and Vogler (2004) for the SPSS-based TwoStep cluster method (see [TwoStep Clustering](#)), which allows both continuous and categorical variables in a model simultaneously, showed poor results for models of mixed-type data owing to categorical variables being assigned more weight as a virtue of the formulation for combining data types whenever possible. There are several suggested ways to calculate similarity for data of mixed type, and the interested reader is directed to work by Gower (1971) and Ichino and Yaguchi (1994). Based on these issues, it is our recommendation that users develop clusters with data of similar types. For example, if clusters are developed for a set of cognitive measures, these clusters can later be summarized by categorical variables. That is, if three clusters are found from the data, one could, after the fact, summarize gender ratios, age categories, and so on among the clusters or even test (e.g., the binomial or chi-square test) to see if there are significant differences in the frequencies of categorical observations among the clusters.

Fig. 2.4 An illustration of cluster proximity measures. Single linkage examines the distance to the nearest neighbor, complete linkage uses the distance to the farthest neighbor, and the average linkage uses the average distance to neighbors, thereby ameliorating the impact of extreme observations. One would use the same linkage for all measures; multiple linkages are shown here for illustration only



Measures of Cluster Proximity

We last introduce sets of measures that are useful for finding distances between clusters or groups of individuals as opposed to distances between individuals. These measures are referred to as “linkage” or “Interclass” distance measures (Everitt et al., 2011; Johnson & Wichern, 2007). These types of measures are most generally used for data of the continuous type, but there are some less well-known measures designed for categorical data, and we leave the interested reader with a reference summarizing those methods (Everitt et al., 2011, p. 61).

When one has groups of subjects and wishes to investigate if they are proximate enough in space to combine, as in hierarchical clustering methods (see [Hierarchical Clustering](#)), there are several methods one can use. These are the “single linkage” or “nearest-neighbor distance,” the “complete linkage” or “furthest-neighbor distance,” the “average linkage,” and Ward’s method. The first three of these are illustrated in Fig. 2.4. The single linkage method uses the nearest distance between single objects in space, the complete linkage method uses the farthest neighbor in space, and the average linkage method uses the average of the distances in space. A general suggestion is to consider using average linkage for hierarchical clustering, as this measure tends to reduce the impact of extreme individuals (i.e., subjects lying furthest away from a cluster centroid) in a cluster.

As a final proximity measure, Ward’s method (Ward, 1963) is very commonly used in hierarchical clustering methods, as it has intuitive statistical appeal. This statistical appeal originates from the fact that Ward’s method is based on a sum-of-squares approach, an approach familiar to anyone that has used a linear model such as simple linear regression or ANOVA. Ward’s method works by attempting to minimize increases in cluster sum of squares when clusters are combined into potential new clusters. That is, clusters begin as independent subjects occupying

their own cluster, and then proximate subjects are combined stepwise such that at each step the increase in overall intergroup sum of squares is kept at a minimum. For example, if there are three subjects A, B, and C, A can join with B or A can join with C, and that combination which has the lowest sum of squares will be the two that are combined.

Ward's method is widely used and is the default in many programs. Because the sum-of-squares increase in Ward's method is mathematically proportional to the squared Euclidean distance (see [Measures for Continuous Data](#)), the combination of these two methods—that is calculating similarity using the squared Euclidean distance and then clustering objects using Ward's method—is often used in combination as default methods. We do not disagree with this approach and suggest that this combination be considered a null model against which other potential combinations of similarity and linkage measures be tested for parsimony.

Clustering Methods

There are a great many clustering methods available, many of which have been included as ad hoc methods in proprietary computer programs. Clearly not all available methods are in wide use, and many have likely never been examined by the casual user as possibilities because they are not included in the most popular computer statistical analysis packages used by practitioners. Inasmuch as the intention of this chapter is to introduce cluster analysis methods and ideas in an application-oriented approach, we cover in this section only the most common clustering algorithms, as they are likely to be the most commonly used among the reading audience; as it happens, then, they tend to be the best-documented approaches as well, and include [hierarchical clustering](#), [optimization clustering](#), and [model-based clustering](#).

Hierarchical Clustering

These methods either begin with each individual as its own cluster and then continue until all individuals are in a single cluster (agglomerative methods), or they begin with all individuals in a single cluster and then iteratively partition them into clusters of varying size until each individual is its own cluster (divisive methods). The hierarchical methods use measures of similarity and proximity discussed in the previous section. A very informative review of various methods is provided in Everitt et al. (2011, Table 4.1, p. 79) for the interested reader. One of the most visually useful diagrams that can be generated for hierarchical clusters is the dendrogram, which is illustrated in the figures referenced below.

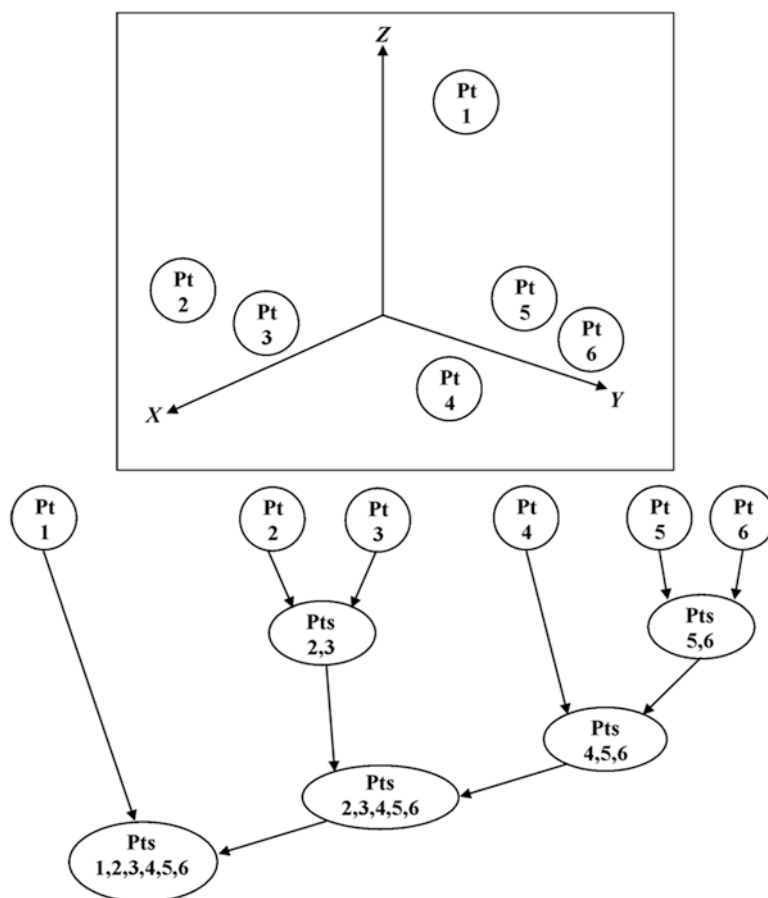


Fig. 2.5 An illustration of a hierarchical agglomerative approach to clustering. In this approach, patients (“Pt”) in closer proximity to one another are clustered together, and then clusters of patient groups are agglomerated iteratively until a single cluster is formed

Hierarchical Agglomerative Methods

Agglomerative methods are by far the most often used, and this is owed to the general recursive formula proposed by Lance and Williams (1967) in which a multitude of different models easily can be generated based on changing the coefficients of the recurrence formula. A useful illustration of an agglomerative technique is illustrated in Fig. 2.5. In this illustration, patients (“Pt”) shown in 3-D space are agglomerated stepwise based on their generalized spatial proximity. These techniques have been used widely in neuropsychological investigations (Allen et al., 2010; Goldstein, 1990; Heinrichs & Awad, 1993; Hill, Ragland, Gur, & Gur, 2002; Rogers et al., 2004; Seaton, Goldstein, & Allen, 2001; Thaler et al., 2010; and chapters within this volume).

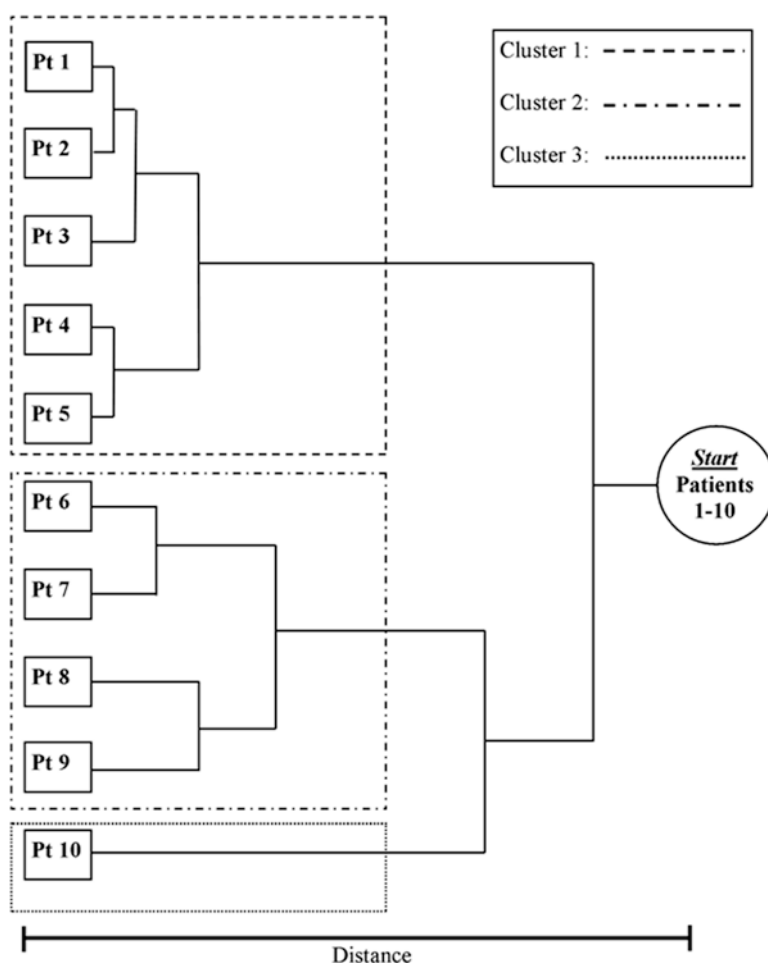


Fig. 2.6 An illustration of a dendrogram from an iterative divisive clustering strategy. All 10 patients (“Pt”) begin in a single cluster, and patients are iteratively divided until each patient represents an individual cluster. Clusters of varying size can be decided upon based on a distance rule

Hierarchical Divisive Methods

Divisive methods are very similar to agglomerative methods and can be conceptualized as an agglomerative approach working in reverse. Divisive methods, per se, have not been used broadly in the neuropsychology literature, though they have been used in other pattern recognition studies (Chavent et al., 2006). As an illustration of a dendrogram (a branching “tree” diagram from the Greek root) using a divisive technique, consider Fig. 2.6. In this figure, patients are divided based on some measure of distance from one another, with potential clusters highlighted. In this

context, dendrograms can be an important tool for visualizing potential clusters, albeit subjectively. In the figure, a measure of distance (see above for examples of these functions) is plotted on the horizontal line at the bottom of the graphic. Branches of varying lengths connect patients, with the length of branches representing distance in variable space (e.g., shorter branches depict more similarity as defined by distance). For example, in the dendrogram, patients 1 and 2 are more similar to each other than patients 8 and 9, and groupings of patients into clusters are the result of examination of the various branching patterns.

Optimization Clustering

A second group of clustering techniques is optimization techniques. Optimization techniques for finding clusters of homogeneous entities are a popular option in the literature and are often done as a follow-up technique after an initial partitioning of the data using hierarchical techniques has been completed (Allen et al., 2010; Thaler et al., 2010; and chapters within this volume). In these methods, optimization is loosely defined as a method that either minimizes or maximizes some numerical criterion defined by the user in an effort to find the “best” or most parsimonious set of clusters. Owing to the vast number of potential ways to develop optimization schemes, we shall restrict our discussion to the k-means techniques commonly used, simply as a result of their ready availability in canned statistical packages. The interested reader is referred to the literature for a full review of optimization techniques and their various qualities (Everitt et al., 2011).

k-Means Clustering

k-Means clustering techniques are designed to group items or subjects into a specified number of clusters. The number of clusters specified by the user may come from previous studies, from an initial run of a hierarchical technique, examination of a dendrogram, and so on. Often several potential cluster sets are formed and then the user, based on theoretical reasoning or a statistical criterion (discussed in [Validation and Choosing Solutions](#)), decides the cluster solution that is best supported by the data.

Nearly all computer packages implement some form of Euclidean distance measure to define a dispersion matrix. This matrix can be thought of, in the sense of an ANOVA, as a matrix that represents the total sum of squares between each object and a mean of all objects in a cluster. One then can mathematically partition this total dispersion into its “within-group” and “between-group” components. Optimization is then simply a matter of developing a cluster solution that either minimizes the within-group variability or maximizes the between-group variability. This is accomplished by minimizing the trace of the within-group matrix—that is, minimizing the sum of the diagonal matrix elements of the within-group dispersion

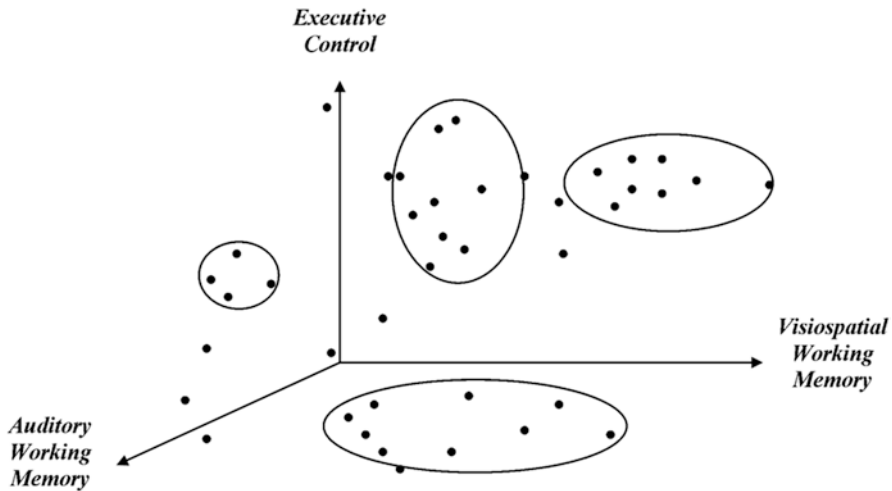


Fig. 2.7 Illustration of a four-cluster, k-means solution. Those subjects (represented as points in the 3-D graphic) not currently assigned to a given cluster will be recursively assigned to clusters until the within-group dispersion is minimized and the between-group dispersion is maximized. This will result in the optimal four-cluster solution

matrix (as it happens, this also maximizes the trace of the between-group dispersion matrix) (Everitt et al., 2011). Hence, the optimal clusters are formed by assigning cases to cluster groups such that the addition of cases to a given group results in the minimum Euclidean distance between the object and the centroid of the existing cluster. The process continues until each object is “optimally” placed into a given cluster. The end product is then a cluster solution, with a predefined number of clusters, which statistically minimizes the within-group variability and maximizes the between-group variability. As you can imagine, this may require an incredible amount of computer time and resources. Rather than start with random starting points, then, one can use seed values defined by, for example, centroids derived from a hierarchical clustering scheme or suggested by an examination of a dendrogram.

As an example, consider Fig. 2.7. In this example, the clinical researcher has chosen to examine a four-cluster solution. It is easy to see that objects closest in space have been assigned to the same clusters. As the algorithm continues, all subjects will be assigned to one of the four existing clusters. One can immediately see a potential shortfall of this technique in that subjects that appear to be equidistant between two clusters will be assigned to a given cluster, and can result in overlap of clusters. This could mean that the number of clusters chosen by the practitioner was incorrect, that the variables used to define distances were not appropriate or were incomplete, or that some individuals in the data set simply cannot be nicely grouped together. Nonetheless, k-means techniques remain a popular choice among researchers.

Model-Based Clustering

Model-based approaches are a third type of clustering techniques. They are non-heuristic methods of clustering that use common mathematical or statistical reasoning to develop solutions. The general idea behind these models is that one can use statistical algorithms and information criteria to find the best solution given the data (e.g., TwoStep clustering) or that the collected data belong to a mixture of subpopulations, each with their own probability density functions (finite mixture models), and hence this information can be used to determine cluster membership.

TwoStep Clustering

When deciding to use cluster analysis as a way to group objects, one would often like a method that allows the combination of categorical and continuous data, that allows for automatic noise handling for outliers, and that provides an automated way to find the optimal number of clusters. To that end, the popular and oft-used statistical package by psychologists, SPSS (IBM SPSS, Armonk, New York), has implemented the TwoStep clustering method. This method allows users to use non-commensurate data in a single model, allows the use of log-likelihood or Euclidean distance measures, and provides an information-theoretic approach (Burnham & Anderson, 2002) for establishing the number of clusters in the final solution.

The TwoStep clustering algorithm obtains its name from the fact that two steps are involved in the progress of the algorithm. In the first step, pre-clusters are formed based on an initial pass through the data, and then the pre-cluster solutions are used in step 2 to find the final cluster solutions. The final cluster solution is found by optimizing a dispersion measure much like that used in k-means clustering, with the final solution selected as that solution which provides the most information with the least number of clusters—that is, the most parsimonious solution. Bacher et al. (2004) provide an extensive overview of this method using simulations and comparisons to various other software packages. In general they found that the TwoStep algorithm did not perform as well as expected, particularly when using different data types or when clusters greatly overlap.

Our experience mirrors that of Bacher et al. (2004) in that the use of TwoStep clustering sometimes does not provide informative solutions. This could be a result of generally small sample sizes or the fact that clusters often overlap when examining neuropsychological data. TwoStep clustering techniques have not been used as broadly in the neuropsychological literature as k-means, though some have reported using this algorithm (Libon et al., 2010; Peters Graf, Hayden, & Feldman, 2005), and it is sometimes employed in related disciplines, for example: education (Halsell, 2007), general psychology (Glasø, Matthiesen, Nielsen, & Ståle, 2007), and health psychology (Bulger, Matthews, & Hoffman, 2007). Certainly more investigative work needs to be done with this algorithm to fully understand its potential for use with standard neuropsychological data sets.

Finite Mixture Models

Models of this type use a formal statistical approach to determine cluster membership. In its basic form, one is attempting to determine how measurements from a set of objects were formed. That is, objects in the data set are assumed to have arisen from a set of underlying statistical distributions, and the task is to estimate the parameters of these mixed distributions and then use these parameters to find the most probable set of subpopulations (i.e., clusters) in the data set.

If each object is assumed to come from a distribution, then one can use this information to generate a model using standard statistical techniques. For example, if objects are assumed to come from a mixture of subpopulations that are represented as a mixture of normal density functions, the clustering problem is reduced to a problem of finding parameter estimates for this multivariate normal distribution and then using an information-theoretic approach (Burnham & Anderson, 2002) to find the most parsimonious set of subgroups leading to the mixture distribution. Many estimation techniques are available, with the most common being finding maximum likelihood estimates for model parameters—which makes intuitive sense in that the solution to these estimates will maximize the probability of having obtained the data under investigation obtaining the data in hand (Johnson & Wichern, 2007). Other estimation techniques, for example, using Bayesian inference, are being used more often now owing to the availability of rapid computer processing (see Everitt et al., 2011).

Just as with any statistical technique, the value of a solution is only as valuable as the information used to generate the model. With that in mind, mixture models should be used with relatively large data sets, with sample size determined via a power analysis. Mixture models have found their way into the neuropsychological literature, though not to a large degree (Donoghue, 1995; Palmer, Dawes, & Heaton, 2009; Wessman et al., 2009). It is anticipated that use of these models will increase with the availability of large data sets related to, for example, genetic studies of certain neurological conditions, or data mining of existing databases.

Validation and Choosing Solutions

The ultimate goal of cluster analysis is to find a way to determine if certain entities share commonality when measured across a given set of variables and, if so, what might explain these commonalities. Hence, it is of interest to determine if a solution is valid and how many clusters there may be in a given data set. Outside of statistically based criteria, this task is somewhat subjective, and therefore solutions are open to criticism. This is certainly expected and warranted, as a solution is only as good as the data used to generate it, and multiple interpretations may be possible with any data set. It is therefore useful to consider ways to evaluate the potential validity of cluster solutions. We have already covered a common method used for

model-based algorithms—namely, the use of information criteria to assist in finding a parsimonious solution. In addition to this method, we have included a brief discussion of other techniques.

External Criteria

It is very common, and certainly makes intuitive sense, to use an external criterion to assist in determining clusters. For example, one may wish to validate a three- or four-cluster solution for a particular group of inpatient subjects because such a solution has been postulated from theory to exist, and thence one will use this theoretically derived postulate to test the underlying theory. As another example, one may wish to see if the data they collected follow what is known about developmental stages of children or what would be expected based on trauma to certain regions of the brain owing to a TBI. Whatever the case, the use of external criteria can be quite useful but should be well documented and explained so that one is not accused of tinkering with potential solutions until the most parsimonious one happens to be what one has theorized should be present in the data, that is, avoid a “fishing expedition.”

Explained Variance

A paper by Milligan and Cooper (1985) provides an extensive list and evaluation of potential ways to evaluate and compare different cluster solutions. There are numerous techniques for evaluating explained variance as a means to determining the “best” number of clusters. This has appeal in that many statistical decisions are made based on maximizing explained variance and minimizing unexplained variance. One of the better performers in the review paper was proposed by Beale (1969), and we have used it extensively owing to its ease of calculation and relatively straightforward interpretation. Beale’s approach relies on finding the F -ratio between competing cluster solutions; the explicit formula is most easily found in Everitt et al. (2011). Essentially the formula takes the ratio of the difference in the sum of squares for two solutions and divides it by the degrees of freedom, which is a function of the sample size, number of groups in the two competing solutions, and the number of variables used to derive the solution. Its usage is simple: find various solutions to a cluster problem, say a four-cluster and a five-cluster solution. The algorithm proceeds thus: (1) find the total sum of squares for each solution by measuring distance between each point and the centroid of the cluster to which it belongs; (2) using Beale’s formula, find the F -ratio comparing the sums of squares for each solution; (3) find the p -value for the F -ratio; and (4) if the p -value is below a designated criteria, then statistically more variability is being explained by one

solution compared to the other. One can easily use this idea to test multiple solutions and correct the p -value using a Bonferroni adjustment. It is a straightforward task to write code to do the calculations or to simply write a spreadsheet macro or series of cell formulas.

A similar approach was taken by Calinski and Harabasz (1974) in the development of their estimator, often referred to as “pseudo F ” (pseudo distributions generally do not satisfy strict independence and multivariate normality assumptions; Timm, 2002). This estimator takes the ratio of between-group sum of squares to within-group sum of squares. One can then determine if clusters of varying cluster number have significantly different F -ratios. Sarle (1983) utilized a slightly different approach, called the cubic cluster criterion, which compares cluster solutions to a uniform distribution of clusters and then uses a function of the sum of squares to calculate the test statistic. Both of these techniques have been implemented in SAS software (SAS Institute Inc., Cary, North Carolina).

Discriminant Function Analysis

The purpose of DFA is to find a series of functions that can be used to reliably parse groups where membership is already known (Tabachnick & Fidell, 2007). Clearly membership is not known a priori in cluster analysis. However, DFA is an interesting ad hoc method for examining cluster solutions. First, it allows one to easily generate a visualization of a given solution by plotting clusters on canonical axes that are independent. Second, it provides a means to evaluate, after the fact, how many individuals would be correctly or incorrectly reclassified into particular groups based on the DFA algorithm. This is useful because one can select a jackknife approach to reclassification such that the data used to generate the model and the data used to test the model are the same; this saves valuable resources in terms of using part of a data set to build a model and the remaining to test the model in traditional validation. Last, one can generate a set of discriminant functions that can be used to determine what cluster group to which a new subject most likely belongs. This can be very valuable when one is using cluster analysis to develop groups for screening purposes. In this approach, the researcher has identified the number of clusters using one of the clustering approaches previously described, and the variables used to develop the clusters are then entered into the DFA to see how well they classify the individuals back into the clusters; a jackknife reclassification approach (Efron, 1982) is suggested, as it reduces the bias of cross-validating a model using the same data as was used to develop the model. The correspondence between the clusters identified with the cluster analysis and those predicted by the DFA provide an indication of the relative stability (i.e., reproducibility) of the cluster solution, with the greater the correspondence, the higher the stability of the cluster solution.

Assumptions and Considerations

Cluster analysis is an intriguing technique to consider when faced with the issue of developing groups where subjects may share commonality of measurements on a given set of neuropsychological instruments. Unfortunately, cluster analysis is often best used as an exploratory technique for validation of existing theory as opposed to a theory-generating procedure because the number of clusters which one is attempting to find is a nuisance parameter in the cluster model because it is not known *a priori*. The goal, then, is to find the number of clusters that provide the most explanatory ability for a given set of data and that are definable in a useful way. Validation of a given set of clusters through replication of experiments, as well as using external criteria for validating solutions, can be a useful way to develop convincing clusters that can be generalized beyond a single study.

Cluster analysis in general does not suffer from a set of assumptions that is often unrealistic, as compared to some other multivariable techniques. There are some issues of concern to consider, however. Distributional assumptions are not necessarily required, though normality of data is of concern for mixture models that assume underlying multivariate normality. One should consider, then, the necessity of applying transformations prior to data analysis such that outliers do not overly influence the shape, size, and overlap of potential clusters. Standardizing data prior to analysis is often recommended, and we concur that this can be valuable because unstandardized variables can have demonstrable influence on distances depending on their scale of measurement. For example, the potential range of scores for a test like Trail Making (time in seconds) is much larger than the range for a test like Grooved Pegboard (number of pegs placed), and so entering raw scores rather than standardize scores may cause Trail Making scores to have a greater influence on the cluster solution than Grooved Pegboard.

Another important consideration is to ensure that variables are independent, particularly for mixture models where collinearity can be problematic. Redundant information is uninformative in cluster solutions and should be avoided if possible. Sample size issues should be considered as well. A power analysis to determine sample size is a routine consideration for experimental design; however, power analyses for cluster algorithms are not generally available. The best approach, of course, is to have as many subjects as possible and to consider having at least ten subjects for each variable measured. For optimization algorithms, sample size necessarily must be large and devoid of outliers.

Missing data are problematic for virtually any analysis, particularly when patterns of missingness are not random (Tabachnick & Fidell, 2007). As always, missing data should be avoided as a standard practice, by ensuring that subjects are included in a study by virtue of the completeness of their data. With small data sets, however, one hates to lose the information from any individual, even if incomplete and so a data replacement technique may be the preferred option. As with any replacement technique for missing data, the use of a simple measure such as a mean

can distort the underlying data distribution because overall variability in a given measure will be reduced. Hence, one should replace data using an algorithm (e.g., expectation-maximization [EM] algorithm or regression with random error) that will incorporate random error among the replaced values so that overall variability for a given measure is unaffected by replaced values. Inasmuch as distances among variables are often used in clustering techniques, measures with random error are unlikely to cause major issues with final cluster solutions. Final cluster solutions obtained when using different replacement algorithms to replace missing values should be the same if, in fact, missing values are not problematic for the solution. Hence, it is recommended that one consider different replacement algorithms to check for stability of final cluster solutions.

Concluding Remarks

As a final set of considerations, we leave the reader with four points to consider. The first is that many cluster analytic methods are relatively simple algorithms that may not have extensive mathematical or statistical theory supporting them. This is generally not an issue, as cluster analysis is used to find parsimonious solutions with some measure of explanatory validity. Second, cluster analysis methods exist among many, disparate disciplines. As such, there are vocabulary differences when examining the literature, and the researcher is cautioned to consider this when reading. Third, different cluster techniques result in different solutions when using the same data. We therefore recommend that one strategically plan and document their approach and that some sort of external validation is considered. Last, cluster analyses *impose* structure on a data set, but we use cluster analyses to *find* structure. This is a puzzling conundrum but should always be considered when examining a solution. As humans, we often are drawn to solutions we believe to be true, and the imposition of structure fits a lens through which we see the experimental world in which we work. Therefore, when using and publishing results derived from cluster analyses, we should be open to questions and criticisms and be prepared to provide convincing arguments for why certain solutions seem more reasonable than others.

Software Considerations

As a final section, we have provided a table highlighting statistical packages that can be used for cluster analysis. The complexity of algorithms needed to generate potential solutions is vast and nearly impossible without the help of a computer. Table 2.2 is designed to provide some insight into available software, but is not meant to be exhaustive or to persuade the user to use one package compared to another.

Table 2.2 Highlights of popular statistical packages for cluster analyses with contact information

| Package | Comments | Reference |
|------------|--|--|
| MINITAB | <ul style="list-style-type: none">• MINITAB contains algorithms for basic cluster analysis, include hierarchical and k-means methods• This software is largely menu driven and does not require extensive programming knowledge | MINTAB Inc. Quality plaza 1829 pine hall road State College, PA 16801-3008 http://www.minitab.com |
| R | <ul style="list-style-type: none">• R contains by far the largest number of procedures for cluster analysis, with nearly 100 packages available for hierarchical, k-means, and model-based clustering• R is a programming language and therefore users must be knowledgeable about programming before using this software• R is completely free and available for download on many platforms | The R project for statistical computing http://www.r-project.org/ |
| SAS | <ul style="list-style-type: none">• SAS contains a complete suite of cluster analytic methods. Procedures include CLUSTER (hierarchical models), FASTCLUS (k-means models), MODECLUS (nonparametric density estimates), VARCLUS (hierarchical and disjoint methods), and TREE (dendrogram generation)• Use of SAS requires extensive programming knowledge | SAS Institute Inc. 100 SAS campus drive Cary, NC 27513-2414 919.644.8000 http://www.sas.com/ |
| SPSS | <ul style="list-style-type: none">• SPSS contains algorithms for hierarchical cluster analysis, k-means cluster analysis, and TwoStep cluster analysis• SPSS is largely menu driven, and most methods can be used without extensive programming knowledge | IBM SPSS 1 new orchard road Armonk, NY 10504-1722 914.499.1900 http://www-01.ibm.com/software/analytics/spss/ |
| STATISTICA | <ul style="list-style-type: none">• STATISTICA contains methods for hierarchical clustering, k-means clustering, EM clustering, and cross-validation for finding cluster solutions• This program is largely menu driven and does not require extensive programming knowledge | StatSoft, Inc. 2300 East 14th street Tulsa, OK 74104 918.749.1119 http://www.statsoft.com/ |

References

Allen, D. N., Goldstein, G., & Warnick, E. (2003). A consideration of neuropsychologically normal schizophrenia. *Journal of the International Neuropsychological Society*, 9, 56–63.

Allen, D. N., Leany, B. D., Thaler, N. S., Cross, C., Sutton, G. P., & Mayfield, J. (2010). Memory and attention profiles in pediatric traumatic brain injury. *Archives of Clinical Neuropsychology*, 25, 618–633.

Bacher, J., Wenzig, K., & Vogler, M. (2004). *SPSS TwoStep cluster—a first evaluation*. Retrieved February 15, 2008, from <http://www.statisticalinnovations.com/products/twostep.pdf>

- Beale, E. M. L. (1969). Euclidean cluster analysis. *Bulletin of the International Statistical Institute: Proceedings of the 37th Session (London), Book 2* (pp. 92–94). Voorburg, The Netherlands: ISI.
- Betz, N. E. (1987). Use of discriminant analysis in counseling psychology research. *Journal of Counseling Psychology, 34*, 393–403.
- Bulger, D. A., Matthews, R. A., & Hoffman, M. E. (2007). Work and personal life boundary management: Boundary strength, work/personal life balance, and the segmentation-integration continuum. *Journal of Occupational Health Psychology, 12*, 365–375.
- Burnham, K. P., & Anderson, D. (2002). *Model selection and multi-model inference: A practical information-theoretic approach* (2nd ed.). New York: Springer.
- Calinski, R. B., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics, 3*, 1–27.
- Chavent, M., Ding, Y., Fu, L., Stolowy, H., & Wang, H. (2006). Disclosure and determinants studies; an extension using the division clustering method (DIV). *European Accounting Review, 15*, 181–218.
- Cross, C. L., & Petersen, C. E. (2001). Modeling snake microhabitat from radiotelemetry studies using polytomous logistic regression. *Journal of Herpetology, 35*, 590–597.
- Donoghue, J. R. (1995). Univariate screening measures for cluster analysis. *Multivariate Behavioral Research, 30*, 385–427.
- Efron, B. (1982). The jackknife, the bootstrap and other resampling plans. *SIAM CBMS-NSF Monographs, 28*.
- Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences, 95*, 14863–14868.
- Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis* (5th ed.). New York: Wiley.
- Glasø, L., Matthiesen, S. B., Nielsen, M. B., & Ståle, E. (2007). Do targets of workplace bullying portray a general victim personality profile? *Scandinavian Journal of Psychology, 48*, 313–319.
- Goldstein, G. (1990). Neuropsychological heterogeneity in schizophrenia: A consideration of abstraction and problem-solving abilities. *Archives of Clinical Neuropsychology, 5*, 251–264.
- Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics, 27*, 857–872.
- Gower, J. C., & Legendre, P. (1986). Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification, 5*, 5–48.
- Halsell, J. N. (2007). *Using cluster analysis to evaluate the academic performance of demographic homogeneous subsets*. Unpublished doctoral dissertation, University of Nevada, Las Vegas, Nevada.
- Heinrichs, R. W., & Awad, A. G. (1993). Neurocognitive subtypes of chronic schizophrenia. *Schizophrenia Research, 9*, 49–58.
- Hill, S. K., Ragland, J. D., Gur, R. C., & Gur, R. E. (2002). Neuropsychological profiles delineate distinct profiles of schizophrenia, an interaction between memory and executive function, and uneven distribution of clinical subtypes. *Journal of Clinical and Experimental Neuropsychology, 24*, 2002.
- Hosmer, D. W., & Lemeshow, S. (2001). *Applied logistic regression* (2nd ed.). New York: Wiley.
- Huff, D. (1954). *How to lie with statistics*. New York: W. W. Norton.
- Ichino, M., & Yaguchi, H. (1994). Generalized Minkowski metrics for mixed feature-type data analysis. *IEEE Transactions on Systems, Man and Cybernetics, 24*, 698–708.
- Jiang, D., Tang, C., & Zhang, A. (2004). Cluster analysis for gene expression data: A survey. *IEEE Transactions on Knowledge and Data Engineering, 16*, 1370–1386.
- Johnson, R. A., & Wichern, D. W. (2007). *Applied multivariate statistical analysis* (6th ed.). Upper Saddle River, NJ: Pearson.
- Lance, G. N., & Williams, W. T. (1967). A general theory of classification sorting strategies: 1. Hierarchical systems. *Computer Journal, 9*, 373–380.

- Libon, D. J., Schwartzman, R. J., Eppig, J., Wambach, D., Brahin, E., Peterlin, B. L., et al. (2010). Neuropsychological deficits associated with complex regional pain syndrome. *Journal of the International Neuropsychological Society*, 16, 566–573.
- Lumley, T. (2001). Orca [R [RJava]]. *Proceedings of the 2nd International Workshop on Distributed Statistical Computing, Vienna, Austria*. Available online at <http://www.ci.tuwien.ac.at/Conferences/DSC-2001/Proceedings/Lumley.pdf>
- Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50, 159–179.
- Morris, R., Blashfield, R., & Satz, P. (1981). Neuropsychology and cluster analysis: Potentials and problems. *Journal of Clinical Neuropsychology*, 3, 79–99.
- Myers, R. E., III, & Fouts, J. T. (1992). A cluster analysis of high school science classroom environment and attitude toward science. *Journal of Research in Science Teaching*, 29, 929–937.
- Palmer, B. W., Dawes, S. W., & Heaton, R. K. (2009). What do we know about neuropsychological aspects of schizophrenia? *Neuropsychology Review*, 19, 365–384.
- Pearson, K. (1920). Notes on the history of correlation. *Biometrika*, 13, 25–45.
- Peters, K. R., Graf, P., Hayden, S., & Feldman, H. (2005). Neuropsychological subgroups of cognitively-impaired-not-demented (CIND) individuals: Delineation, reliability, and predictive validity. *Journal of Clinical and Experimental Neuropsychology*, 27, 164–188.
- Punj, G., & Stewart, D. W. (1983). Cluster analysis in marketing research: Review and suggestions for application. *Journal of Marketing Research*, 20, 134–148.
- Rodgers, J. L., & Nicewander, A. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42, 59–66.
- Rogers, T. T., Ralph, M. A. L., Garrard, P., Bozeat, S., McClelland, J. L., Hodges, J. R., et al. (2004). Structure and deterioration of semantic memory: A neuropsychological and computational investigation. *Psychological Review*, 111, 205–235.
- Sarkar, D. (2008). *Lattice: Multivariate visualization with R*. New York: Springer.
- Sarle, W. S. (1983). *The cubic cluster criterion*. SAS Technical Report A-108. Cary, NC: SAS Institute.
- Seaton, B. E., Goldstein, G., & Allen, D. (2001). Sources of heterogeneity in schizophrenia: The role of neuropsychological functioning. *Neuropsychological Review*, 11, 45–67.
- Sharp, H. (1968). Cardinality of finite topologies. *Journal of Combinatorial Theory*, 5, 82–86.
- Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics*, 5th ed. Boston, MA: Allyn and Bacon.
- Tajima, F. (1993). Unbiased estimation of evolutionary distance between nucleotide sequences. *Molecular Biology and Evolution*, 10, 677–688.
- Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., & Kumar, S. (2011). MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distances, and maximum parsimony methods. *Molecular Biology and Evolution*, 28, 2731–2739.
- Thaler, N. S., Bellow, D. T., Randall, C., Goldstein, G., Mayfield, J., & Allen, D. N. (2010). IQ profiles are associated with differences in behavioral functioning following pediatric traumatic brain injury. *Archives of Clinical Neuropsychology*, 25, 781–790.
- Timm, N. H. (2002). *Applied multivariate statistics*. New York: Springer.
- Wallace, L., Keil, M., & Rai, A. (2004). Understanding software project risk: A cluster analysis. *Information and Management*, 42, 115–125.
- Ward, J. H. (1963). Hierarchical groupings to optimize an objective function. *Journal of the American Statistical Association*, 58, 236–244.
- Wessman, J., Paunio, T., Tuulio-Henriksson, A., Koivisto, M., Partonen, T., Suvisaari, J., et al. (2009). Mixture model clustering of phenotype features reveals evidence for association of DTNBP1 to a specific subtype of schizophrenia. *Biological Psychiatry*, 66, 990–996.

Cluster Analysis in Neuropsychological Research

Recent Applications

Allen, D.N.; Goldstein, G. (Eds.)

2013, VII, 136 p., Hardcover

ISBN: 978-1-4614-6743-4