1

Cluster Plots with More than Two Target Variables

In this document I discuss methods used to construct cluster plots when there are three or more target variables in the cluster analysis. I focus on two approaches, one based on principal components analysis (PCA, used in my programs on consensus and medoid cluster analyses) and the other on discriminant function analysis (DFA, used in my program on trimmed k-means cluster analysis). I assume you are somewhat familiar with both PCA and DFA. When there are only two target variables, the visual display of the cluster plot is straightforward and typically uses a two-dimensional scatterplot with one variable on each axis (sometimes transformed, sometimes not). With three or more target variables, representing the clusters in a two dimensional visual is challenging. PCA and DFA are used to address these challenges.

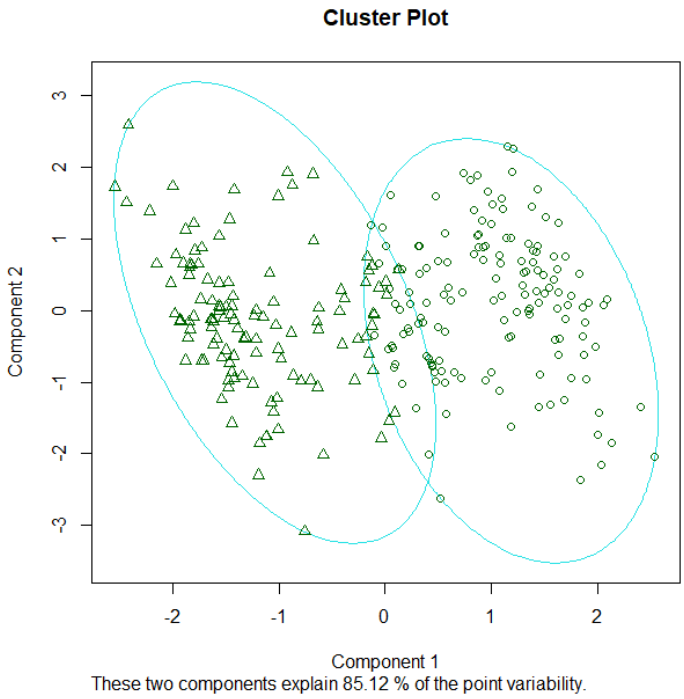
PRINCIPAL COMPONENTS ANALYSIS

PCA is often characterized as a form of factor analysis but it is not that. To be sure, it can be adapted for use in factor analysis but in its original form, it was intended as a data reduction method to characterize the variation in a set of variables using fewer, more succinct linear combinations of the target variables. These linear combinations are called **principal components**. In general if the number of target variables is k, then there are k possible principal components. The first principal component that is extracted from the data describes the most variation in the target variables considered as a whole (per the covariance matrix of the variables). The second principal component describes the second most amount of variation that is not accounted for by the first component. Because of this property, only the first and second principal components are extracted from the target variable to construct a cluster plot to visualize cluster dynamics in the data. Each individual is assigned a score on each of the two principal components (called **component scores**) and these pairs of scores are then used to form a traditional scatterplot of them. The scores typically are standardized in form. The idea is that the first two principle components capture variation in the target variables that can then be mapped onto the cluster structure. For details and an example, see (Wang et al., 2018).

On a cluster plot that uses PCA, it is not atypical for the plot to report a percentage on each axis. This indicates how much of the variation that is explained by each component. Some plots only report the sum of the total percent explained variance across the two components. In general, when you sum these percents, the larger the result the better.

The data points that constitute the different clusters on the cluster plot usually are

assigned different shapes or colors. Cluster centroids are included with 95% ellipsoid bands drawn around them. These bands represent the cluster circles that separate one cluster from another. The clusters can overlap on the plot. Small overlaps indicate that the clusters are somewhat similar; if the overlap is too great, then one cannot distinguish between the overlapping clusters. Clearly separated clusters suggest partial empirical support for the clusters. Here is an example PCA based plot for multiple target variables in conjunction with a two cluster solution:

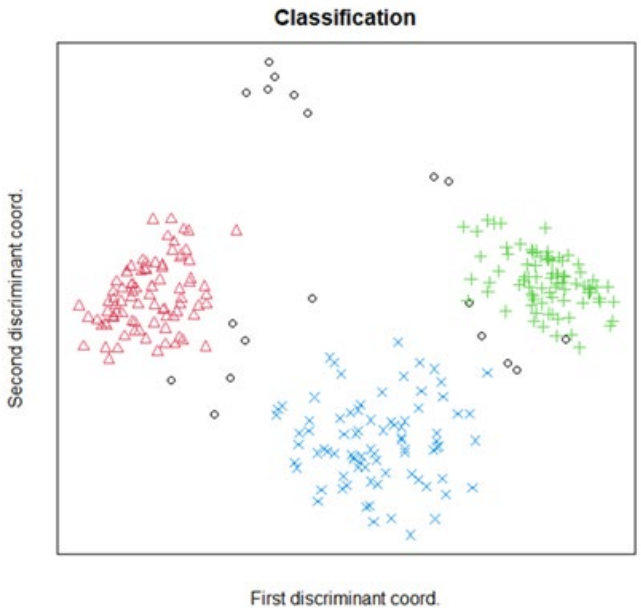


Use of PCA as a strategy for visualizing clusters works well in many situations but there also are scenarios where it is suboptimal. If the first two components when considered together account only for a small amount of variation, then focusing on those dimensions does not provide much information. PCA plots also can be problematic when the data has non-linear relationships between variables, when there are highly skewed distributions, and in the presence of outliers. It turns out that PCA focuses on somewhat different features of the data than clustering algorithms do and these differences can distort the visual. Even if the clusters appear well-separated on a plot that uses PCA, this does not necessarily mean they are distinct in the original data space. That is why we do not rely solely on plots when evaluating cluster structure. They are useful but only one piece of information we take into account. For PCA tutorials, see Dunteman (1989) and Tabachnick and Fidell (2013).

DISCRIMINANT FUNCTION ANALYSIS

A second approach eschews PCA and instead uses DFA. This is the case for the trimmed k-means program on my website. Both PCA and DFA are data reduction techniques but they use different statistical frameworks with somewhat different goals. Like PCA, DFA defines linear "components" underlying the observed data but the components are called **discriminant functions**. Whereas PCA seeks to explain variation in the target variables, FDA seeks to define the underlying discriminant functions so as to maximize the separation (e.g., the mean differences) of known groups, in this case, the clusters from the cluster analysis. Like PCA, there are multiple underlying discriminant functions, with the first extracted discriminant function best separating the groups, the second extracted discriminant function being next best independent of the first discriminant function, and so on. Each individual can be assigned a score on each discriminant function. A higher score on a given discriminant function indicates that the data point for the individual is closer to the group or cluster that the discriminant function best separates.

Traditional DFA assumes linear relationships between variables. When applied to trimmed k-means analysis, outliers are not as problematic as with traditional k-means analysis because outliers are eliminated vis-à-vis the trimming process. Ellipsoid bands usually do not accompany DFA plots. Here is an example plot from the trimmed k-means program on my website in which the uncolored data points represent trimmed cases:



We again seek well separated clusters that are relatively homogeneous within a cluster. Note that the plot can change depending on the amount of trimming you do. I often construct multiple plots with different levels of trimming to literally visualize how trimming may be affecting matters. For introductions to DFA, see Klecka (1980) and Tabachnick and Fidell, (2013).

REFERENCES

Dunteman, G. (1989). *Principal components analysis*. SAGE

Klecka, W. (1980). *Discriminant analysis*. SAGE

Tabachnick, B. & Fidell, L. (2013). *Using multivariate statistics* (6th edition). Pearson Education Limited

Wang, K. et al. (2018). Principal component analysis of early alcohol, drug and tobacco use with major depressive disorder in US adults." Journal of Psychiatric Research, 100, 113-120.