Paper 2868-2018

Factors vs. Clusters

Diana Suhr, SIR Consulting

ABSTRACT

Factor analysis is an exploratory statistical technique to investigate dimensions and the factor structure underlying a set of variables (items) while cluster analysis is an exploratory statistical technique to group observations (people, things, events) into clusters or groups so that the degree of association is strong between members of the same cluster and weak between members of different clusters. Factor and cluster analysis guidelines and SAS® code will be discussed as well as illustrating and discussing results for sample data analysis. Procedures shown will be PROC FACTOR, PROC CORR alpha, PROC STANDARDIZE, PROC CLUSTER, and PROC FASTCLUS.

INTRODUCTION

Exploratory factor analysis (EFA) investigates the possible underlying factor structure (dimensions) of a set of interrelated variables without imposing a preconceived structure on the outcome (Child, 1990). The analysis groups similar items to identify dimensions (also called factors or latent constructs).

Exploratory cluster analysis (ECA) is a technique for dividing a multivariate dataset into “natural” clusters or groups. The technique involves identifying groups of individuals or objects that are similar to each other but different from individuals or objects in other groups.

Cluster analysis, like factor analysis, makes no distinction between independent and dependent variables. Factor analysis reduces the number of variables by grouping them into a smaller set of factors. Cluster analysis reduces the number of observations by grouping them into a smaller set of clusters.

There is no right or wrong answer to “how many factors or clusters should I keep?”. The answer depends on what you’re going to do with the factors or clusters. To find a good factor solution, you must look at how the items group together and what dimensions they measure. To find a good cluster solution, you must look at the characteristics of the clusters at successive steps and decide when you have an interpretable solution or a solution that has a reasonable number of homogeneous clusters.

DATA

Data for both factor analysis and cluster analysis examples are free datasets found online. The factor analysis example uses data (n=1418) collected online during 2015 using the **Nerdy Personality Attributes Scale (NPAS),** https://openpsychometrics.org/_rawdata/ . NPAS was developed as a project to quantify what "nerdiness" is and provides an estimate of how much a respondent's personality is similar to the average for those who identify themselves as nerds versus those who do not. The NPAS has 26 questions with a five point scale, 1=Disagree, 3=Neutral and 5=Agree. The NPAS is for educational use only and is not to be used as a diagnostic or screening tool.

Data for the cluster analysis example (n=406) is a set of car characteristics including acceleration, horsepower, MPG, cylinders, weight, displacement, model, origin, and car name. The dataset can be found at https://perso.telecom-paristech.fr/eagan/class/igr204/datasets.

EXPLORATORY FACTOR ANALYSIS (EFA)

Psychologists searching for a neat and tidy description of human intellectual abilities lead to the development of factor analytic methods. Galton, a scientist during the 19th and 20th centuries, laid the foundations for factor analytic methods by developing quantitative methods to determine the interdependence between 2 variables. Karl Pearson was the first to explicitly define factor analysis. In

1902, Macdonnell was the first to publish an application of factor analysis, a comparison of physical characteristics between 3000 criminals and 1000 Cambridge undergraduates (Child, 1990).

EFA

- is a variable reduction technique which identifies the number of latent constructs (dimensions) and the underlying factor structure of a set of variables
- hypothesizes underlying latent constructs (a dimension not measured directly)
- estimates factors which influence responses on observed variables
- allows you to describe and identify the number of factors (dimensions or latent constructs)
- includes unique error terms due to unreliability in measurement
- traditionally is used to explore the possible underlying factor structure of a set of measured variables without imposing any preconceived structure on the outcome (Child, 1990).

EFA decomposes an adjusted correlation matrix. Variables are standardized in EFA, e.g., mean=0, standard deviation=1, diagonals are adjusted for unique factors, 1-u. Squared multiple correlations (SMC) are used as communality estimates on the diagonals.

The amount of variance explained is equal to the trace of the decomposed adjusted correlation matrix, the sum of the adjusted diagonals or communalities. Observed variables are a linear combination of the underlying and unique factors. Factors are estimated, (X1 = b1F1 + b2F2 + . . . e1 where e1 is a unique error term).

Eigenvalues indicate the amount of variance explained by each factor. Eigenvectors are the weights that could be used to calculate factor scores. In common practice, factor scores are calculated with a mean or sum of measured variables that "load" on a factor.

The EFA Model is $Y = X\beta + E$

where Y is a matrix of measured variables
X is a matrix of common factors
β is a matrix of weights (factor loadings)
E is a matrix of unique error terms, error variation

Communality is the variance of observed variables accounted for by a common factor. A large communality value indicates a strong influence by an underlying construct. Community is computed by summing squares of factor loadings

d_1^2 = 1 – communality = % variance accounted for by the unique factor
d_1 = square root (1-community) = unique factor weight (parameter estimate)

Figure 1 below shows 4 factors (circles) each measured by 3 observed variables (rectangles) with unique factors (error). Since measurement is not perfect, error or unreliability is estimated and specified explicitly in the diagram. Factor loadings (parameter estimates) help interpret factors. Loadings are the correlation between observed variables and factors, are standardized regression weights if variables are standardized (weights used to predict variables from factor), and are path coefficients in path analysis. Standardized linear weights represent the effect size of the factor on variability of observed variables.

Goals of factor analysis are

1) to help an investigator determine the number of latent constructs underlying a set of items (variables)
2) to provide a means of explaining variation among variables (items) using a few newly created variables (factors or dimensions)
3) to define the content or meaning of factors or dimensions, e.g., latent constructs

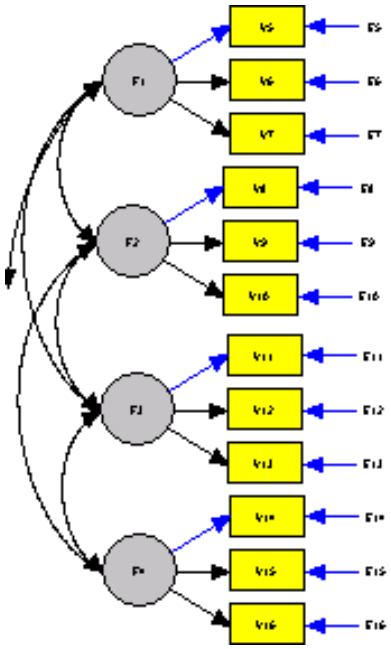


Figure 1 Factor Analytic Model

Assumptions underlying EFA are
- interval or ratio level measurement
- random sampling
- relationship between observed variables is linear
- a normal distribution (each observed variable)
- a bivariate normal distribution (each pair of observed variables)
- multivariate normality

Limitations of EFA are
- reliability of the measurement instrument. Avoid an instrument with low reliability.
- sample size (larger sample → larger correlation)
  - ◊ a minimal number of cases for reliable results, more than 100 observations and 5 times the number of items.
  - ◊ since some subjects may not answer every item, a larger sample is desirable. For example, 30 items would require at least 150 cases (5*30), a sample of 200 subjects would allow for missing data.
- sample selection
  - ◊ representative of population
  - ◊ do not pool populations.
- variables could be sample specific, e.g., a unique quality possessed by a group does not generalize to the population
- nonnormal distribution of data
- correlations describe relationships and are the basis of factor analysis. No causal inferences can be made from correlations alone.

Factor Extraction
Factor analysis seeks to discover common factors. The technique for extracting factors attempts to take out as much common variance as possible in the first factor. Subsequent factors are, in turn, intended to account for the maximum amount of the remaining common variance until, hopefully, no common variance remains.

Direct extraction methods obtain the factor matrix directly from the correlation matrix by application of specified mathematical models. Most factor analysts agree that direct solutions are not sufficient. Adjustment to the frames of reference by rotation methods improves the interpretation of factor loadings by reducing some of the ambiguities which accompany the preliminary analysis (Child, 1990). The process of manipulating the reference axes is known as rotation.

Rotation applied to the reference axes means the axes are turned about the origin until some alternative position has been reached. The simplest case is when the axes are held at 90^{o} to each other, orthogonal rotation. Rotating the axes through different angles gives an oblique rotation (not at 90^{o} to each other).

Criteria for Extracting Factors

Determining the number of factors to extract in a factor analytic procedure means keeping the factors that account for the most variance in the data. Criteria for determining the number of factors are:

1) Kaiser's criterion, suggested by Guttman and adapted by Kaiser, considers factors with an eigenvalue greater than one as common factors (Nunnally, 1978)
2) Cattell's (1966) scree test. The name is based on an analogy between the debris, called scree, that collects at the bottom of a hill after a landslide, and the relatively meaningless factors that result from overextraction. On a scree plot, because each factor explains less variance than the preceding factors, an imaginary line divides the markers for successive factors and generally runs from top left of the graph to the bottom right. If there is a point below which factors explain relatively little variance and above which they explain substantially more, this usually appears as an "elbow" in the plot. This plot bears some physical resemblance to the profile of a hillside. The portion beyond the elbow corresponds to the rubble, or scree, that gathers. Cattell's guidelines call for retaining factors above the elbow and rejecting those below it.
3) Proportion of variance keeps a factor if it accounts for a predetermined amount of the variance (e.g., 5%, 10%).
4) Interpretability criteria
    a. Are there at least 3 items with loading values set a priori, e.g. > 0.45.
    b. Do the variables that load on a factor share some conceptual meaning?
    c. Do the variables that load on different factors seem to measure different constructs?
    d. Does the rotated factor pattern demonstrate simple structure? Are there relatively
        i. high loadings on one factor?
        ii. low loadings on other factors?

Reliability Analysis Prior to EFA Analysis

Reliability refers to the accuracy and precision of a measurement procedure (Thorndike, Cunningham, Thorndike, & Hagen, 1991). Reliability may be viewed as an instrument's relative lack of error. In addition, reliability is a function of properties of the underlying construct being measured, the measurement instrument itself, the groups being assessed, the testing environment, and the purpose of assessment. Reliability answers the question, "How well does the instrument measure what it purports to measure?".

Some degree of inconsistency is present in all measurement procedures. The variability in a set of item scores is due to the actual variation across individuals in the phenomenon that the scale measures, made up of true score and error. Therefore, each observation of a measurement (X) is equal to true score (T) plus measurement error (e), or $X = T + e$.

Reliability can be assessed by internal consistency – measured with Cronbach's coefficient alpha. Internal consistency is a procedure to estimate the reliability of a measurement instrument from a single administration of a single form. Internal consistency depends on the individual's performance from item to item based on the standard deviation of the test and the standard deviations of the items. Cronbach's coefficient alpha is measured on a scale from zero to one with a value closer to one being a more reliable measurement instrument and showing higher internal consistency.

Reversing Items

Examination of reliability analysis and item-to-total correlations reveal which, if any, item scales should be reversed. Items could be worded so that most responses will be in the same "direction". However, if

appropriate responses are in the opposite "direction" of the scale, item responses are reserved. Negative item-to-total correlations determine which item response scales to reverse. For example, with a scale of 1=strongly agree, 2=agree, 3=disagree, 4=strongly disagree, responses are reversed to 1=strongly disagree, 2=disagree, 3=agree, 4=strongly agree.
To reverse an item on a 4-point scale the SAS code is `q3r = 5 - q3;`
Substitute the values 1, 2, 3, 4 to verify responses are reversed.
For a 5-point scale the SAS code is `q12r = 6 - q12;`
Substitute the values 1, 2, 3, 4, 5 to verify responses will be reversed.

EFA Steps

1) Reliability analysis, reverse items if needed
2) Set variance levels a priori, proportion of variance and cumulative variance
3) Initial extraction
   - each factor accounts for a maximum amount of variance that has not previously been accounted for by any of the other factors
   - factors are uncorrelated
   - eigenvalues represent the amount of variance accounted for by each factor
4) Determine the number of factors to retain
   - scree test, look for elbow (optional)
   - proportion of variance
   - cumulative variance
   - prior communality estimates are not perfectly accurate, cumulative proportion must equal 100% so some eigenvalues will be negative after factors are extracted, e.g., if 5 factors are extracted, cumulative proportion equals 100% and factors 6 and above have negative eigenvalues
   - interpretability
     - at least 3 observed (measured) variables per factor
     - common conceptual meaning
     - each factor measures a different construct or dimension
     - rotated factor pattern has simple structure (no cross loadings)
5) Rotation – a transformation
6) Interpret solution
7) Calculate factor scores
8) Results in a table
9) Prepare results, paper

SAS Code

PROC FACTOR and options for EFA

DATA =	specifies dataset to be analyzed
PRIORS =SMC	squared multiple correlations used as adjusted diagonals of the correlation matrix
METHOD =ML,ULS	specifies maximum likelihood and unweighted least squares methods
ROTATE =	VARIMAX(orthogonal at a 90 degee angle)
	PROMAX (oblique, not at a 90 degree angle)
SCREE	requests a scree plot of the eigenvalues
N =	specifies number of factors
MINEIGEN=1	specifies select factors with eigenvalues greater than 1
OUT =	data and estimated factor scores, use raw data and N=
FLAG =	include a flag (*) for factor loadings above a specified value
REORDER =	sort the loadings from largest to smallest values for each factor

PROC CORR and options for reliability

DATA =	specifies dataset to be analyzed
ALPHA	computes Cronbach's coefficient alpha
NOCORR	suppresses Pearson correlations
NOMISS	excludes observations with missing values from the analysis

EFA EXAMPLE and Results

Syntax for preliminary steps is not shown. The syntax below illustrates code for the final analysis:

```
proc factor data=rawsub1 method=ml priors=smc rotate=varimax
            n=4 reorder flag=.45; *scree;
     var Q1-Q26;
title2 'factor analysis - 4 factors';
run;

proc corr data=rawsub1 nomiss nocorr alpha;
  var q7 q12 q25 q5 q26 q24;
title2 'reliability - 4 factors with loadings ge .45';
proc corr data=rawsub1 nomiss nocorr alpha;
  var q23 q6 q9;
proc corr data=rawsub1 nomiss nocorr alpha;
  var q17 q16 q3 q13;
proc corr data=rawsub1 nomiss nocorr alpha;
  var q11 q19 q15 q2;
run;
```

A rotated four factor solution with factor loadings greater than or equal to 0.45 found 6 items for factor one, 3 items for factor two, 4 items for factor three and 4 items for factor four. Interpretability guidelines specify at least 3 observed (measured) variables or items per factor. Items loading on each factor are shown in Table 1. Each factor measures a dimension of a how respondents identified with having attributes of a nerdy personality. Seventeen items with factor loadings greater than or equal to 0.45 were retained for a four factor solution (from the twenty-six item scale).

Table 1. NPAS Rotated Factor Pattern Loadings and Reliability – Four Factor Solution

Factor	Item Number	Loading	Item
1	7	0.704	I watch science related shows.
1	12	0.623	I spend more time at the library than any other public place.
1	25	0.586	I care about super heroes.
1	5	0.531	I collect books.
1	26	0.472	I can be socially awkward at times.
1	24	0.465	I am a strange person.
			Factor 1 reliability (Cronbach alpha) = 0.78 with 6 items
2	23	0.650	I get excited about my ideas and research.
2	6	0.578	I prefer academic success to social success.
2	9	0.540	I like science fiction.
			Factor 2 reliability (Cronbach alpha) = 0.67 with 3 items
3	17	0.764	I am more comfortable interacting online than in person.
3	16	0.583	I gravitate towards introspection (Note: introspection is the examination or observation of one's own mental and emotional processes)
3	3	0.559	I like to play RPFs. (ex D&D)
3	13	0.520	I would describe my smarts as bookish.
			Factor 3 reliability (Cronbach alpha) = 0.71 with 4 items
4	11	0.681	I am more comfortable with my hobbies than I am with other people
4	19	0.639	I have played a lot of video games.
4	15	0.523	I have started writing a novel.
4	2	0.468	I was in advanced classes.
			Factor 4 reliability (Cronbach alpha) = 0.69 with 4 items

EXPLORATORY CLUSTER ANALYSIS (ECA)

Cluster analysis originated in anthropology through studies by Driver and Kroeber in 1932 and was introduced to psychology by Zubin in 1938 and Tryon in 1939. Cluster analysis techniques were used by Cattell beginning in 1943 for trait theory classification in personality psychology.

Exploratory Cluster Analysis can be used in

- Marketing to form clusters of customers who have similar buying habits or demographics to take advantage of similarities, develop marketing strategies to target offers to subgroups.
- Management as a requirement for predicting the sales of a product to different cities and is achieved by clustering demographically similar cities.
- IT to cluster load balancing in application servers.
- Software development to restructure and improve software functionality.
- Analyzing social networks to identify communities.
- Internet search engines to search more intelligently.
- Data mining for the process of discovering meaningful correlations, patterns and trends from large amounts of data, using pattern recognition technologies as well as statistical and mathematical techniques. Clustering is often one of the first steps in data mining analysis.
- Diagnosis to use symptom checklists and psychological inventories to cluster patients into subgroups that have similar characteristics for diagnosis and appropriate treatment.
- Archaeology to cluster skulls excavated from archaeological digs into civilizations from which they originated.
- Geology to analyze mineral contents of excavated materials to study their origins and spread.
- Biology to organize and group different species of animals before identifying differences between animals.
- Astronomy to classify stars into a main sequence, white giants, and red dwarfs, according to temperature and luminosity.
- Food industry to make recommendations for new items based on customer preferences
- The military for body measurement data to reduce the number of different uniform sizes kept in inventory.
- Law enforcement to manage resources more effectively by identifying areas of higher and similar types of crime.
- Education to identify and examine groups of students or schools with similar properties and needs.

Method

Cluster analysis does not specify a particular statistical method or model. There are no assumptions about the underlying distribution of the data for cluster analysis. Regression, factor analysis, and discriminant analysis have specified methods or models. Cluster analysis groups observations that are similar while factor analysis groups variables that are similar. The choice of a cluster analytic method depends on the goal of the analysis, the size of the data set, and types of variables. Methods commonly used for small data sets are impractical for data files with thousands of cases.

The appropriate clustering algorithm and parameter settings (including values such as the distance function, density threshold or the number of expected clusters) depend on the individual data set and intended use of the results. Cluster analysis is an iterative process of knowledge discovery and optimization modifying data processing and model parameters until the result achieves preferred as well as appropriate properties.

Warning for Cluster Analysis

Cluster analysis has no mechanism for differentiating between relevant and irrelevant variables. Therefore the choice of variables included in a cluster analysis must be related to conceptual considerations. The clusters formed can be very dependent on the variables included. Data used in cluster analysis can be interval, ordinal or categorical. However, having a mixture of different types of variable will make the analysis more complicated. With cluster analysis you need to have some way of

measuring the distance between observations. The type of measure used will depend on what type of data you have.

How Alike (or Different) Are the Cases?

Distance is a measure of how far apart two objects are, while similarity measures how similar two objects are. For cases that are alike, distance measures are small and similarity measures are large. There are many different definitions of distance and similarity. Some, like the Euclidean distance, are suitable for only continuous variables, while others are suitable for only categorical variables.

Standardizing the Variables

If variables are measured on different scales, variables with large values contribute more to the distance measure than variables with small values. Variables can be standardized using z-scores. When looking at the distance between two people based on their IQs and incomes in dollars, differences in incomes would dominate any distance measures. A difference of $100 when squared becomes 10,000, while a difference of 30 IQ points squared would be 900. Variables measured in large numbers will contribute more to the distance than variables recorded with smaller numbers. One strategy is to run the cluster analysis twice, once without standardizing and once with standardizing to see how much difference, if any, is found in the resulting clusters.

PROC STANDARD syntax

```
PROC STANDARD DATA=<input dataset> OUT=<output dataset>
  MEAN = <mean value>  (specifies mean value)
  STD = <std value>    (specifies standard deviation value)
  <options>:
  VAR <variables>;
  FREQ <variable>;
  WEIGHT <variable>;
  BY <variables>;
```

Hierarchical Clustering

There are numerous ways in which clusters can be formed. Hierarchical clustering is one of the most straightforward methods. It can be either agglomerative or divisive.
To form clusters using a hierarchical cluster analysis, you must select:

- a criterion for determining similarity or distance between cases,
- a criterion for determining which clusters are merged at successive steps,
- the number of clusters you need to represent your data.

Agglomerative hierarchical clustering begins with every case being a cluster. At successive steps, similar clusters are merged. The algorithm ends with everybody in one cluster. In agglomerative clustering, once a cluster is formed, it cannot be split; it can only be combined with other clusters. Agglomerative hierarchical clustering doesn't let cases separate from clusters that they've joined. Once in a cluster, always in that cluster. Divisive clustering starts with all cases in one cluster and ends up with individuals as clusters. Neither the first step nor the last step is the best solution.

Plotting Cluster Distances: The Dendrogram

A visual representation of the distance at which clusters are combined can be shown with a dendrogram. The dendrogram is read from left to right. Vertical lines show joined clusters. The position of the line on the scale indicates the distance at which clusters are joined. The observed distances are rescaled to fall into the range of 1 to 25, the actual distances are not seen; however, the ratio of the rescaled distances within the dendrogram is the same as the ratio of the original distances.
Tip: When you read a dendrogram, you want to determine at what stage the distances between clusters that are combined is large. Look for large distances between sequential vertical lines.

Syntax for PROC TREE - dendrogram

```
PROC TREE <options>;
  NAME <variables>;
```

```
  HEIGHT <variables>;
  PARENT <variables>;
  BY <variables>;
  COPY <variables>;
  FREQ <variables>;
  ID <variables>;
run;
```

K-Means Clustering

Hierarchical clustering requires a distance or similarity matrix between all pairs of cases. That's a huge matrix if there are tens of thousands of cases in the data file. A clustering method that doesn't require computation of all possible distances is *k*-means clustering. It differs from hierarchical clustering in several ways:

- determine the number of clusters a priori.
- run an analysis for each different number of clusters, e.g. 3 clusters, 4 clusters, etc..
- the algorithm repeatedly reassigns cases to clusters, so the same case can move from cluster to cluster during the analysis. In agglomerative hierarchical clustering, cases are added only to existing clusters.
- the k-means algorithm, where *k* is the number of clusters you want, assigns cases to the cluster for which its distance to the cluster mean is the smallest .
- the algorithm centers around finding *k*-means.
- start out with an initial set of means and classify cases based on their distances to the centers,
- compute the cluster means again, using the cases that are assigned to the cluster,
- reclassify all cases based on the new set of means,
- keep repeating this step until cluster means don't change much between successive steps.
- finally, calculate the means of the clusters once again and assign the cases to their permanent clusters.

Two-Step Cluster

With a large data set or to use a procedure that can rapidly form clusters on the basis of either categorical or continuous data, neither of the previous two procedures works as well as two-step clustering. Hierarchical clustering requires a matrix of distances between all pairs of cases, and *k*-means requires shuffling cases in and out of clusters and knowing the number of clusters in advance.

Step 1: Preclustering: Making Little Clusters

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

Step 2: Hierarchical Clustering of Preclusters

- in the second step, use the standard hierarchical clustering algorithm on the preclusters.
- forming clusters hierarchically lets you explore a range of solutions with different numbers of clusters.

Tip: A large number of preclusters gives better results because the cases are more similar in a precluster; however, forming many preclusters slows the algorithm.

Cluster Analysis Steps

Cluster analysis methods are not clearly established. There are many options one may select when performing cluster analysis. For example, each case could start as a separate cluster and continue to combine similar clusters until there is one cluster left. This method uses distances between data points. The most common and generally accepted way to calculate distances between objects in a multi-dimensional space is to compute Euclidean distances (an extension of Pythagoras' theorem).

The choice of methods used for cluster analysis could depend on the size of the data set as well as the types of variables. Hierarchical clustering is appropriate for a small data set. K-means clustering where the number of clusters is known works for a moderately large data set. A large data set or a mixture of continuous and categorical variables requires a two-step procedure.

Note: While the theoretical foundation of these methods is excellent, there is one key problem known as overfitting (unless constraints are put on the model complexity). A more complex model will usually be able to explain the data better, which makes choosing the appropriate number and type of variables a key factor in model development.

Steps

- start with a number of cases to subdivide into homogeneous groups or clusters,
- choose the variables,
- decide whether to standardize the variables in some way so that they all contribute equally to the distance or similarity between cases; the analysis can be run with both standardized and unstandardized variables and then compared for the number of clusters,
- decide which clustering procedure to use, based on the number of cases and types of variables that you want to use for forming clusters.

For hierarchical clustering,

- choose a statistic that quantifies how far apart (or similar) two cases are,
- select a method for forming the groups; could have as many clusters as you do cases (not a useful solution),
- determine how many clusters are needed to represent the data; look at how similar clusters are (descriptive statistics for variables for each cluster).

For *k*-means clustering,

- select the number of clusters,
- the algorithm iteratively estimates the cluster means and assigns each case to the cluster for which its distance to the cluster mean is the smallest,
- assigns cases to the cluster where the smallest distance to the cluster mean.

For two-step clustering,

- cases are assigned to “preclusters”,
- the preclusters are clustered using the hierarchical clustering algorithm,
- specify the number of clusters you want or let the algorithm decide based on preselected criteria.

PROC CLUSTER

The CLUSTER procedure hierarchically clusters the observations in a SAS data set by using one of eleven methods. The data can be coordinates or distances. If the data are coordinates, PROC CLUSTER computes squared Euclidean distances.

There are numerous clustering methods. All methods are based on the usual agglomerative hierarchical clustering procedure. Each observation begins in a cluster by itself. The two closest clusters are merged to form a new cluster that replaces the two old clusters. Merging of the two closest clusters is repeated until only one cluster is left. The various clustering methods differ in how the distance between two clusters is computed.

The CLUSTER procedure is not practical for very large data sets because the CPU time is roughly proportional to the square or cube of the number of observations. The FASTCLUS procedure requires time proportional to the number of observations and can be used with much larger data sets than PROC CLUSTER. If you want to cluster a very large data set hierarchically, use PROC CLUSTER for a preliminary cluster analysis. Then use PROC FASTCLUS to determine cluster membership.

Plots of the pseudo *F* statistic and cubic clustering criterion (CCC) are options in PROC CLUSTER. Values of CCC, pseudo F, and observed overall r-squared, and approximate expected overall r-squared are shown with PROC FASTCLUS. Values of cubic clustering criterion greater than 2 or 3 indicate good clusters. Values between 0 and 2 indicate potential clusters, but they should be taken with caution; large

negative values can indicate outliers. Note: approximate expected R square and CCC values are not valid for correlated variables. For correlated variables, use cluster principal component scores from the PROC PRINCOMP procedure.

PROC CLUSTER displays a history of the clustering process, showing statistics useful for estimating the number of clusters in the population from which the data are sampled. PROC CLUSTER also creates an output data set that can be used by PROC TREE to draw a tree diagram of the cluster hierarchy or to output the cluster membership at any desired level. For example, to obtain the five-cluster solution, you could first use PROC CLUSTER with the OUTTREE= option, and then use this output data set as the input data set to the TREE procedure. With PROC TREE, specify NCLUSTERS=5 and the OUT= options to obtain the five-cluster solution and draw a tree diagram.

Syntax PROC CLUSTER

```
PROC CLUSTER DATA=<input dataset>
    METHOD=<name>
    OUTTREE=<dataset>
    std                 (standardize, mean=0, std=1)
    ccc                 (cubic clustering criterion plot)
    pseudo              (pseudo F plot)
    <options>;
  BY <variables>;       (separate analysis)
  COPY <variables>;     (copied from input dataset to OUTTREE=dataset)
  FREQ <variable>;      (frequency of occurrence of variables)
  ID <variables>;       (to identify observations)
  RMSSTD <variables>;   (variable containing root mean squared deviations)
  VAR <variables>;      (variables in the analysis)
```

PROC FASTCLUS

PROC FASTCLUS is intended for use with large data sets, 100 or more observations. With small data sets, the results can be highly sensitive to the order of the observations in the dataset. PROC FASTCLUS uses algorithms that place a larger influence on variables with larger variance. Standardizing the variables before performing the cluster analysis is recommended. PROC FASTCLUS produces brief summaries of the clusters it finds. For more extensive examination of the clusters, you can request an output data set containing a cluster membership variable.

PROC FASTCLUS performs a disjoint cluster analysis on the basis of distances computed from one or more quantitative variables. The observations are divided into clusters such that every observation belongs to one and only one cluster; the clusters do not form a tree structure as they do in the CLUSTER procedure. For an analysis for different numbers of clusters, run PROC FASTCLUS several times changing the number of clusters requested in each analysis.

By default, PROC FASTCLUS uses Euclidean distances, so the cluster centers are based on least squares estimation. The cluster centers are the means of the observations assigned to each cluster when the algorithm is run to complete convergence. PROC FASTCLUS is designed to find good clusters, not the best possible clusters, with only two or three passes through the data set. PROC FASTCLUS can be an effective procedure for detecting outliers which appear as clusters with only one member.

Syntax PROC FASTCLUS

```
PROC FASTCLUS DATA=     <input data>
   INSTAT =             <dataset created with OUTSTAT=>
   OUT =                <output SAS dataset>
   OUTSTAT =            <output SAS dataset containing statistics>
   CLUSTER =            <specifies name for cluster membership variable>
   MEAN =               <output SAS dataset containing cluster centers>
   MAXCLUSTERS=n;       (if MAXCLUSTERS is not specified, n=100, by default)
 VAR <variables>;       (variables in the analysis)
 ID <variables>;        (to identify observations)
```

```
 FREQ <variable>;        (frequency of occurrence of variables)
 WEIGHT <variable>;      (compete weighted cluster means)
 BY <variables>;         (separate analysis)
```

Graphical Representation
A plot or scatter plot provides a graphical representation of the clusters (see Display 5 for an example). Canonical variables are created using the PROC CANDISC procedure and plotted with PROC SGPLOT.

Syntax for PROC SGPLOT;
```
proc candisc data=<dataset name> out=<out dataset name> noprint;
  var <variables>;
  class <variables>;
proc sgplot data=<out dataset name rom proc candisc>7;
  scatter y=can2 x=can1 / group=class variable name from proc candisc>;
run;
```

Example – Cluster Analysis using car dataset

Data for the cluster analysis example (n=406) is a set of car characteristics including acceleration, horsepower, MPG, cylinders, weight, displacement, model, origin, and car name. For marketing purposes, a cluster analysis could identify groups of vehicles by characteristics.

Data was standardized due to different measurement scales. Analysis was completed with PROC CLUSTER and PROC FASTCLUS. Eigenvalues from PROC CLUSTER were examined for a three cluster solution (see Display 2).

SAS Code
```
proc standard data=rawsub1 mean=0 std=1 out=stan1;
 var MPG Cylinders Displacement Horsepower Weight Acceleration;
run;

proc means data=stan1;
 var MPG Cylinders Displacement Horsepower Weight Acceleration;
run;

proc cluster data=stan1 outtree=tree1 std
      method=average ccc pseudo;
var MPG Cylinders Displacement Horsepower Weight Acceleration;
title2 'standardized data';
run;

**** 3 clusters;
proc fastclus data=stan1 out=clus3 maxclusters=3 summary maxiter=100;
 var MPG Cylinders Displacement Horsepower Weight Acceleration;
  id car;
title2'standardized data - 3 clusters';
run;

proc means data=clus3 maxdec=2;
class cluster;
var MPG Cylinders Displacement Horsepower Weight Acceleration;
run;

proc candisc data=clus3 out=can3 noprint;
  var MPG Cylinders Displacement Horsepower Weight Acceleration;
  class cluster;

proc sgplot data=can3;
  scatter y=can2 x=can1 / group=cluster;
run;
```

```
data clus3x; set clus3; keep carid cluster;
proc sort data=rawsub1; by carid;
proc sort data=clus3x; by carid;
data new3;
  merge rawsub1 clus3x; by carid;
proc means data = new3  maxdec=2;
 var MPG Cylinders Displacement Horsepower Weight Acceleration;
  class cluster;
run;
```

Results

Cluster analysis using a k-means cluster procedure found three groups of cars using characteristics of MPG, number of cylinders, displacement, horsepower, weight, and acceleration to differentiate the groups. Display 1 shows mean values for the total group. Display 2 illustrates eigenvalues of the correlation matrix from PROC CLUSTER. Three clusters were identified to explain 96.26% of the variance in the data.

Variable	N	Mean	Std Dev	Minimum	Maximum
MPG	392	23.45	7.81	9.00	46.60
Cylinders	392	5.47	1.71	3.00	8.00
Displacement	392	194.41	104.64	68.00	455.00
Horsepower	392	104.47	38.49	46.00	230.00
Weight	392	2977.58	849.40	1613.00	5140.00
Acceleration	392	15.54	2.76	8.00	24.80

Display 1. Means for Car Data (PROC MEANS)

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	4.78826616	4.05963505	0.7980	0.7980
2	0.72863111	0.47016380	0.1214	0.9195
3	0.25846732	0.13329030	0.0431	0.9626
4	0.12517701	0.06200061	0.0209	0.9834
5	0.06317640	0.02689440	0.0105	0.9940
6	0.03628200		0.0060	1.0000

Display 2. Eigenvalues for Standardized Data (PROC CLUSTER)

Cluster analysis using car data found three groups with distinct characteristics. Cluster 2 has lower average values for cylinders (4.00), horsepower (75.66), and weight (2239.85) than Clusters 1 (5.54, 100.32, 3108.12) or Cluster 3 (7.98, 161.09, 4125.14) but a larger value for MPG (miles per gallon) (30.31 (Cluster 2), 20.52 (Cluster 1), 14.66 (Cluster 3)).

The Cluster Summery shown in Display 3 includes the distance between cluster centers, the nearest cluster, and frequency in each cluster. Display 4 shows the means for each cluster. A graphical representation of the cluster locations is shown in Display 5.

Cluster Summary						
Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Radius Exceeded	Nearest Cluster	Distance Between Cluster Centroids
1	120	0.5254	2.7700		2	2.1394
2	182	0.5058	3.3298		1	2.1394
3	104	0.4918	2.5177		1	3.2168

Display 3. PROC FASTCLUS Cluster Summary

Cluster	N Obs	Variable	N	Mean	Std Dev	Minimum	Maximum
1	120	MPG	118	20.52	3.23	15.00	32.70
		Cylinders	120	5.54	1.03	3.00	8.00
		Displacement	120	196.36	51.06	80.00	350.00
		Horsepower	117	100.32	12.04	72.00	133.00
		Weight	120	3108.12	365.36	2310.00	4060.00
		Acceleration	120	16.35	2.18	11.30	22.20
2	182	MPG	181	30.31	5.43	18.00	46.60
		Cylinders	182	4.00	0.21	3.00	6.00
		Displacement	182	106.57	23.69	68.00	262.00
		Horsepower	179	75.66	13.16	46.00	113.00
		Weight	182	2239.85	306.18	1613.00	3250.00
		Acceleration	182	16.64	2.41	11.60	24.80
3	104	MPG	99	14.66	2.37	9.00	23.00
		Cylinders	104	7.98	0.20	6.00	8.00
		Displacement	104	347.32	45.22	231.00	455.00
		Horsepower	104	161.09	25.75	120.00	230.00
		Weight	104	4125.14	436.75	3086.00	5140.00
		Acceleration	104	12.60	1.89	8.00	18.50

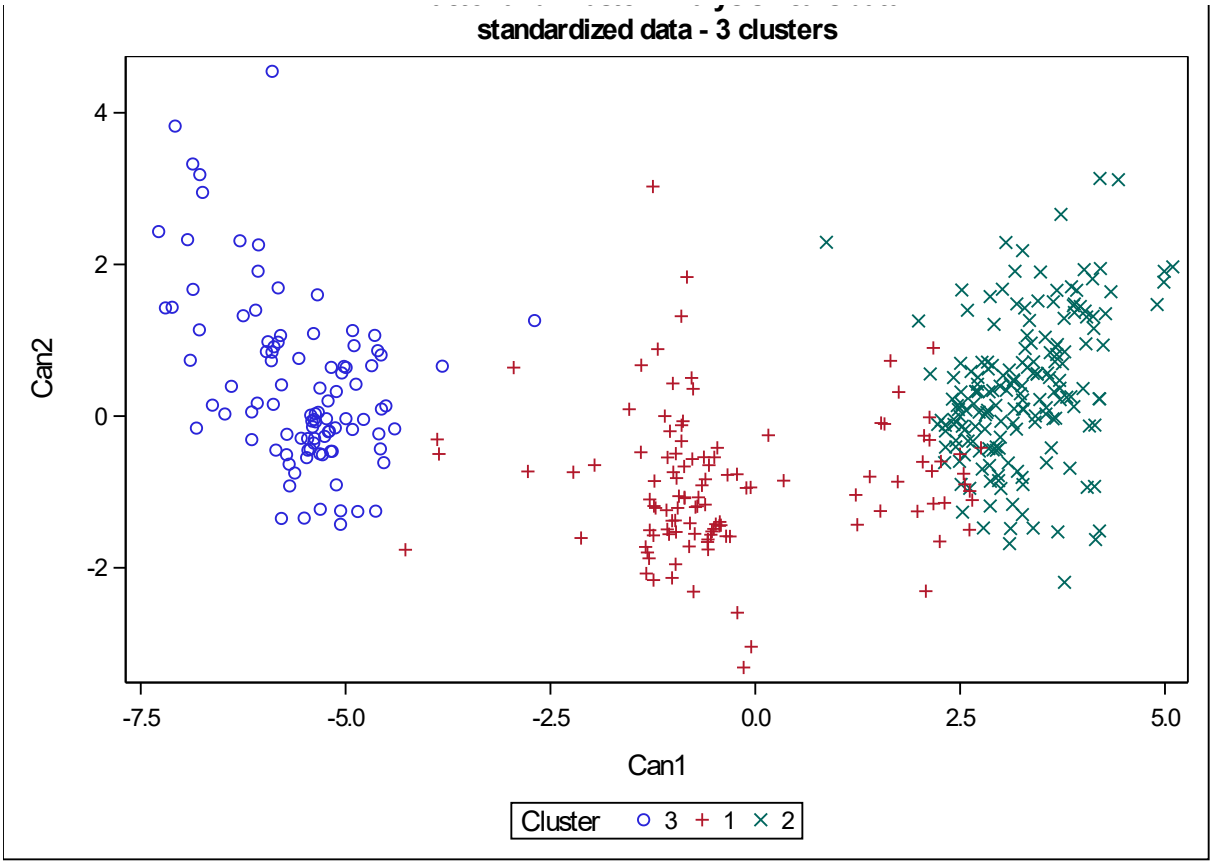
Display 4. Cluster Means (PROC MEANS)

Discussion

Exploratory factor analysis techniques investigate the underlying factor structure of a set of variables while exploratory cluster analysis groups observations using variables to determine similarities in group membership. Cluster analysis groups observations that are similar while factor analysis groups variables that are similar.

Factor analysis explicitly defines a model including measured variables, error terms, and latent constructs (factors or dimensions). Factor analysis specifies underlying assumptions: a linear relationship between variables, normal distribution of measured variables, bivariate normality, and multivariate normality.

Cluster analysis does not define a model, has no underlying assumptions about the distribution of the data. Selecting the appropriate clustering method depends on the data set and the intended use for the results.



Display 5. Three Cluster SGPLOT

CONCLUSION

Exploratory factor analysis identifies the underlying structure of a set of variables without imposing a preconceived structure on the outcome. Exploratory cluster analysis classifies and organizes data into meaningful clusters or groups by maximizing the similarity between observations in each cluster without any preconceived notion. Results from SAS® procedures PROC CORR, PROC FACTOR, PROC CLUSTER, PROC FASTCLUS give you the power to maximize your knowledge, answer your research questions, and determine business strategies.

REFERENCES

Bailey, Ken (1994). "Numerical Taxonomy and Cluster Analysis". *Typologies and Taxonomies*. p. 34. ISBN9780803952591.

Cattell, R. B. (1943). The description of personality: Basic traits resolved into clusters. Journal of Abnormal and Social Psychology, 38, 476-506.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, 1, 245-276.

Child, D. (1990). The essentials of factor analysis, second edition. London: Cassel Educational Limited.

Cluster analysis. (2006). In Collins Dictionary of Sociology. Retrieved from http://0-www.credoreference.com.source.unco.edu/entry/collinssoc/cluster_analysis.

DeVellis, R. F. (1991). Scale Development: Theory and Applications. Newbury Park, California: Sage Publications.

Estivill-Castro, Vladimir (June 2002 2002). "Why so many clustering algorithms —A PositionPaper" (http://cs.nju.edu.cn/zhouzh/zhouzh.files/course/dm/reading/reading05/estivill- castro_sigkddexp02.pdf). *ACM SIGKDD Explorations Newsletter* 4 (1): 65–75. doi:10.1145/568574.568575 (http://dx.doi.org/10.1145%2F568574.568575).

Everitt, B.S., Landau, S. and Leese, M. (2001), Cluster Analysis, Fourth edition, Arnold.

Hans-Peter Kriegel, Peer Kröger, Jörg Sander, Arthur Zimek (2011). "Density-based Clustering" (http://wires.wiley.com/WileyCDA/WiresArticle/wisId-WIDM30.html).*WIREs Data Mining and Knowledge Discovery* 1 (3): 231–240. doi:10.1002/widm.30 (http://dx.doi.org/10.1002%2Fwidm.30).

Hatcher, L. (1994). A step-by-step approach to using the SAS® System for factor analysis and structural equation modeling. Cary, NC: SAS Institute Inc.

Kline, P. (1994). An Easy Guide to Factor Analysis. London: Routledge.

Lloyd, S. (1982). "Least squares quantization in PCM". *IEEE Transactions on Information Theory*28 (2): 129–137. doi:10.1109/TIT.1982.1056489.edit

Manly, B.F.J. (2005), Multivariate Statistical Methods: A primer, Third edition, Chapman and Hall.

Nunnally, J. C. (1978). Psychometric theory, 2nd edition. New York: McGraw-Hill

Penn State, https://onlinecourses.science.psu.edu/stat505/node/138

Rencher, A.C. (2002), Methods of Multivariate Analysis, Second edition, Wiley.

SAS® Language and Procedures, Version 6, First Edition. Cary, N.C.: SAS Institute, 1989.

SAS® Online Doc 9. Cary, N.C.: SAS Institute.

SAS® Procedures, Version 6, Third Edition. Cary, N.C.: SAS Institute, 1990.

SAS/STAT® User's Guide, Version 6, Fourth Edition, Volume 1. Cary, N.C.: SAS Institute, 1990.

SAS/STAT® User's Guide, Version 6, Fourth Edition, Volume 2. Cary, N.C.: SAS Institute, 1990.

Retrieved from support.sas.com
Cluster_PROCFASTCLUSOverview_support.sas.com_documentation_cdl_en_statug_63347_HTML_d.
ClusterAnalysisOverview_support.sas.com_documentation_cdl_en_statug_63347_HTML_d.

Thorndike, R. M., Cunningham, G. K., Thorndike, R. L., & Hagen E. P. (1991). Measurement and evaluation in psychology and education. New York: Macmillan Publishing Company.

Truxillo, Catherine. (2003). Multivariate Statistical Methods: Practical Research Applications Course Notes. Cary, N.C.: SAS Institute.

CONTACT INFORMATION

Diana Suhr, PhD
SIR Consulting
Greeley CO 80634
SuhrSirs@msn.com