

Partitive Clustering Methods

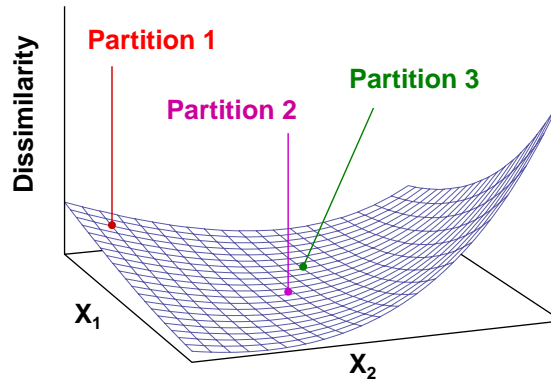
Partitive Clustering Techniques

- *k*-Means Clustering
- Nonparametric Clustering
- Fuzzy Clustering
- Supervised Clustering

Partitive (Optimization) Clustering

The goal of partitive clustering is to minimize or maximize some specified criterion. Two popular criteria are

- within-cluster similarity
- separation.



Partitive Clustering

- *Partitive clustering* (also known as optimization clustering) divides a data set into a number of clusters by trying to minimize some error function.
- In theory, one could consider all possible partitions; that is, you could calculate the value of the optimization criteria (error function) for each possible partition and choose the partition that yields an optimal value for the criteria.
- In practice, however, this is impractical.

Possible Number of Partitions

The number of possible partitions of n objects into g groups is given by:

$$N(n, g) = \frac{1}{g!} \sum_{m=1}^g (-1)^{g-m} \binom{g}{m} m^n$$

For example, the number of partitions of 50 observations into 4 clusters, $N(50,4)$, is equal to 5.3×10^{28} . $N(100, 4)$ generates 6.7×10^{58} partitions.

Complete enumeration of every possible partition, therefore, is generally impossible.

Solution to the Combinatorial Explosion

- This combinatorial explosion has led to the development of clustering algorithms that, by rearranging existing partitions and only keeping those partitions that result in an improvement in a specified numerical criterion, are designed to search this vast space of potential partitions.
- These methods are called *hill-climbing* algorithms.

Hill Climbing

1. Find some initial partition of the n observations into g groups.
2. Calculate the change in the clustering criterion produced by moving each observation from its own to another group.
3. Make the change that leads to the greatest improvement in the clustering criterion.
4. Repeat the previous two steps until no move causes the criterion to improve.



Problems with Heuristic Algorithms

- Partitive clustering methods make explicit assumptions about the shape of the clusters.
- They require that you take an initial guess at the number of clusters that will eventually be found.
- Partitive clustering methods are influenced by
 - the choice of the initial seeds,
 - by the presence of outliers, and
 - by the order in which the seeds are read.

Separation

The proposed cluster partition should produce groups (clusters) that are as well-separated from each other as possible.

Separation measures dissimilarity between observations in the **different** groups.

Example Measures:

- Sum of dissimilarities between observations in different clusters (between-cluster variation)
- Minimum dissimilarity between observations in different clusters.

Heterogeneity

The proposed cluster partition should produce groups that have a relatively cohesive structure.

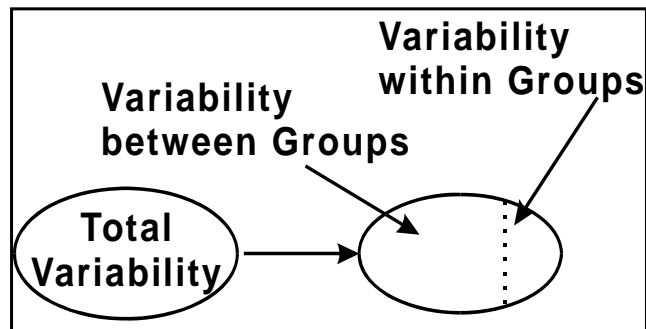
Lack of homogeneity (that is, heterogeneity) measures dissimilarity between observations in the same group.

Example Measures:

- Sum of dissimilarities between observations from the same cluster (within-cluster variation)
- Maximum distance between two objects from the same cluster (cluster diameter).

Partitioning Variability

Total variability can be partitioned into within-cluster and between-cluster variability.



A Natural Grouping Criterion

Partitioning variability suggests a natural grouping criterion.

Choose the partition corresponding to either

- the minimum value of the within-group sum of squares,
or (equivalently)
- the maximum value of the between-group sum of squares.

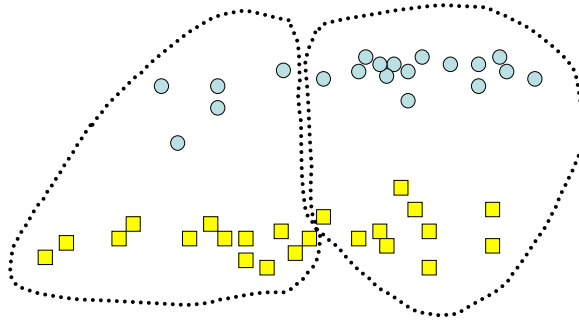
The Trace(\mathbf{W}) Criterion

- The within-cluster sum of squares is really contained in the within-cluster sums of squares and cross products (SSCP) matrix.
- Ideally, you want to summarize the within-cluster sum of squares as a single number.
- Perhaps the most popular criterion minimizes $\text{trace}(\mathbf{W})$, the sum of the diagonal elements.
- Minimizing $\text{trace}(\mathbf{W})$ is equivalent to minimizing the sum of the squared Euclidean distances between a set of observations and their group mean.

The Trace(\mathbf{W}) Criterion

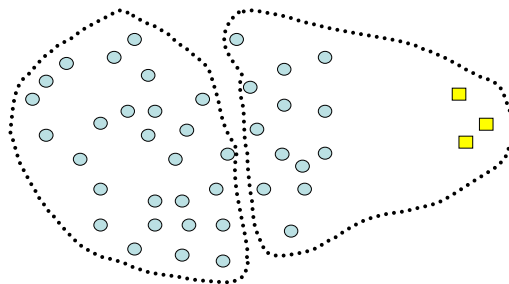
- Thus, a commonly used clustering criterion is minimization of $\text{trace}(\mathbf{W})$, a measure of the within-cluster variability.
- $\text{Trace}(\mathbf{W})$ is scale dependent, which means that different standardization methods generate different solutions.
- $\text{Trace}(\mathbf{W})$ imposes spherical structure on the clusters and tends to produce equal-sized clusters.

The Spherical Structure Problem



The actual clusters are specified by shape, and the discovered clusters are circled (dotted line).

The Similar Size Problem



The actual clusters are specified by shape, and the discovered clusters are circled (dotted line).

An Alternative to Trace(\mathbf{W})

- Another way to summarize the within-cluster variability is the determinant of \mathbf{W} , $\det(\mathbf{W})$.
- It can recover nonspherical clusters, thus eliminating the spherical structure problem.
- However, it still succumbs to the similar size problem.

K-Means Clustering

1. Specify the number of clusters k .
2. Partition the items into k initial clusters (or specify k initial centroids [or seed points]).
3. Proceed through the complete list of N items, assigning an item to the cluster value whose centroid is nearest. Recalculate the centroid for the cluster receiving the new item and for the cluster losing the item.
4. Repeat (3.) until no more reassignments take place.

The FASTCLUS Procedure

```
PROC FASTCLUS DATA=<input-data-set><options>;  
  VAR variables;  
  ID variable;  
  FREQ variable;  
  WEIGHT variable;  
  BY variables;  
RUN;
```

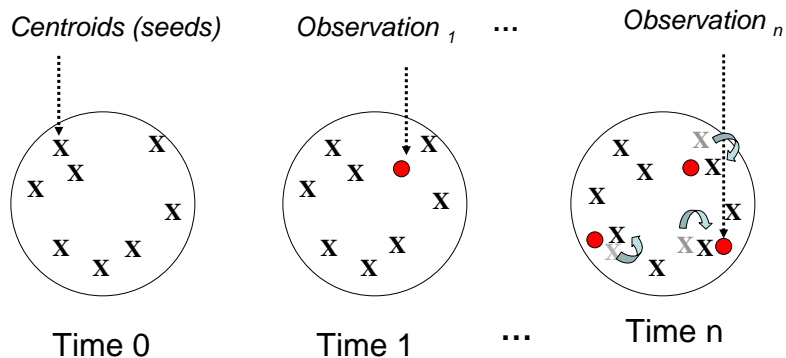
Initial Seed Selection

- PROC FASTCLUS always selects the first complete (no missing values) observation as the first seed. Two tests are performed on subsequent observations to see if they qualify as a new seed:
 - First, an old seed is replaced if the distance between the observation and the closest seed is greater than the minimum distance between seeds.
 - If the observation fails the first test for seed replacement, a second test is made. The observation replaces the nearest seed if the smallest distance from the observation to all seeds other than the nearest one is greater than the shortest distance from the nearest seed to all other seeds. If the observation fails this test, PROC FASTCLUS goes on to the next observation.

The MAXITER Option

- If the (default) MAXITER= option is specified (in the PROC FASTCLUS statement), the algorithm first assigns all observations in their temporary clusters before adjusting the centroids.
- That is, all centroids are adjusted at the same time.

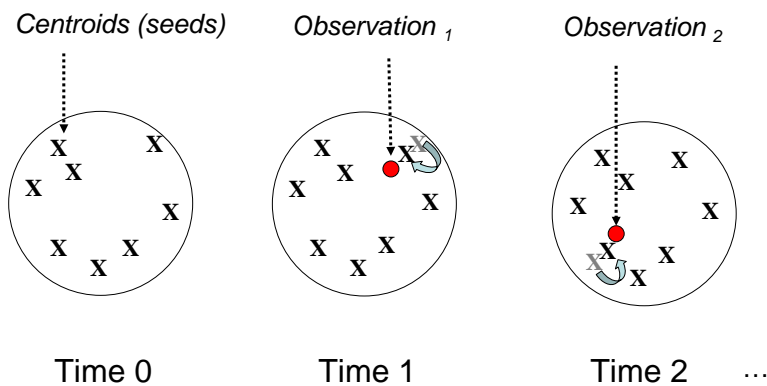
The MAXITER Option



The DRIFT Option

- If the DRIFT= option is specified, the nearest cluster centroid (mean) moves as each observation arrives.
- This method is sometimes referred to as incremental, on-line, or adaptive training.

The DRIFT Option



The L_p Clustering Criterion

- The FASTCLUS procedure uses an L_p (least p^{th} powers) clustering criterion instead of the least-squares (L_2) criterion used in most k -means clustering algorithms.
- The LEAST= p option specifies the power p to be used.
- For general values of p , the FASTCLUS procedure tries to minimize the p^{th} root of the mean of the p^{th} powers of the absolute differences between the observation and the corresponding cluster seeds.

The LEAST Option

- When LEAST=1 (City Block Distance)
 - minimize the **mean absolute difference** between the data and the corresponding cluster **medians**.
- When LEAST=2 (Euclidean Distance)
 - minimize the **root mean square difference** between the data and the corresponding cluster **means**.
- When LEAST=MAX
 - minimize the **maximum absolute difference** between the data and the corresponding cluster **midranges**.

Selection of p in the LEAST option

- Values of p less than 2 reduce the effect of outliers on the cluster centers compared with least-squares methods; values of p greater than 2 increase the effect of outliers.
- If you do not specify the LEAST= option, PROC FASTCLUS uses the least-squares (LEAST=2) criterion. Also, optimization of the criterion is generally not completed.
- If, however, you specify the LEAST= option, the maximum number of iterations is increased to allow the optimization process a chance to converge.

Characteristics of Hierarchical Clustering and k-Mean Clustering

- Most clustering methods are biased toward finding clusters that possess certain characteristics related to size (number of members), shape, or dispersion.
- For example, methods based on the least-squares criterion, such as k -means clustering and Ward's minimum-variance method, tend to find clusters with roughly the same number of observations in each cluster.
- Average linkage is somewhat biased toward finding clusters of equal variance.
- In fact, many clustering methods are incapable of detecting clusters with highly elongated or irregular shapes.

Nonparametric Clustering

- The methods with the least bias are those based on nonparametric density estimation (Silverman 1986; Scott 1992).
- The density estimate at a point is calculated by dividing the number of observations within a (hyper)sphere centered at the point by the product of the sample size and the volume of the (hyper)sphere.

Nonparametric Clustering

- Capable of detecting clusters of unequal size and dispersion, even if they have highly irregular shapes.
- Obtain good results for compact clusters of equal size and dispersion, but they require larger sample size.
- Less sensitive to changes in scale than are most other commonly used clustering methods.

Nonparametric Clustering

- Nonparametric clustering methods are less sensitive than most other methods to changes in the scale or, in fact, any transformation in which all points initially lying on a line still lie on a line after the transformation.
- This is known as an affine transformation. *Affine transformations* preserve the ratio of distances. For example, the midpoint of a line remains the midpoint.

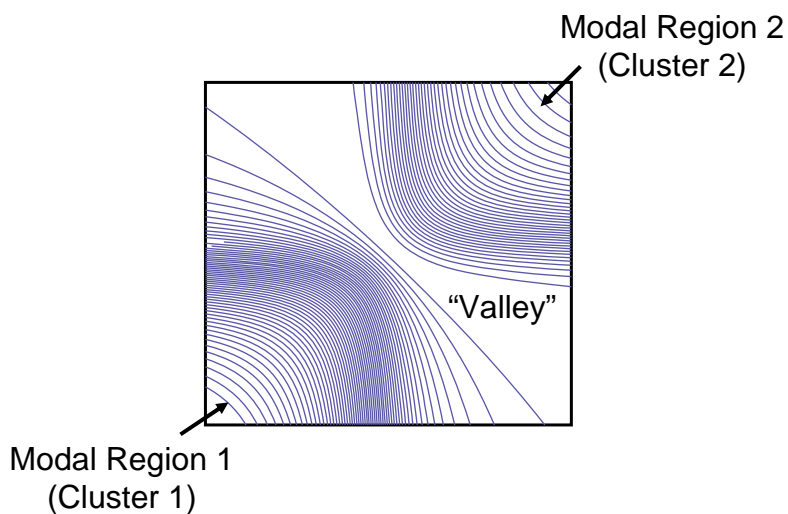
Nonparametric Cluster Definition

- In nonparametric clustering, a cluster is defined as the region that surrounds a local maximum (mode) of the probability density function.
- A cluster, therefore, is a connected set of observations that form exactly one modal region.

Modal Region

- A *modal region* is defined by the probability that a randomly selected point will lie within a hypersphere of radius r from point x .
- If a population has two clusters, by definition it must have two modal regions with a valley between them.
- Intuitively, it is desirable that the boundary between the two clusters should follow the bottom of this valley.

Searching for Boundaries



The MODECLUS Procedure

- General form of the MODECLUS procedure:

```
PROC MODECLUS <options>;  
  BY variables;  
  FREQ variable;  
  ID variable;  
  VAR variables;  
RUN;
```

The MODECLUS Procedure

- All the clustering methods in PROC MODECLUS are designed to locate the estimated cluster boundaries as modal regions with a valley between them.
- You do not need to tell PROC MODECLUS how many clusters you want.
- Instead, you specify a smoothing parameter and, optionally, a significance level. PROC MODECLUS then determines the number of clusters.

The MODECLUS Procedure

- There is no simple answer to the question of which smoothing parameter to use. It is usually necessary to try several different smoothing parameters.
- PROC MODECLUS accepts either distance data or coordinate data.

The MODECLUS Procedure

- For fixed-radius kernels, specify the radius as a Euclidean distance by using the R = option in the PROC MODECLUS statement.
- For variable-radius kernels, specify the number of neighbors desired within the sphere using the K= option.
- The larger the value of K, the larger the resulting clusters tend to be.

Sphere of Support

It is convenient to refer to the sphere of support of the kernel at observation x_i as the neighborhood of x_i .

The observations within the neighborhood of x_i are the neighbors of x_i .

Sphere of Support

The estimated density is given by

$$\hat{f}_i = \frac{n_i}{nv_i}$$

where n is the total number of observations in the sample,

n_i is the number of neighbors within the neighborhood of observation x_i , and
 v_i is volume of the neighborhood of x_i .

Nonparametric Clustering Methods

- Method 0
 - Begin with each observation in a separate cluster. For each observation and each of its neighbors, join the cluster to which the observation belongs with the cluster to which the neighbor belongs.
 - This method does **not** use density estimates. With a fixed clustering radius, the clusters are obtained by cutting the single linkage tree (hierarchical clustering) at a specified radius.

Nonparametric Clustering Methods

- Method 1
 - Begin with each observation in a separate cluster. For each observation, find the nearest neighbor with a greater estimated density. If such a neighbor exists, join the cluster to which the observation belongs with the cluster to which the specified neighbor belongs.
 - **For most purposes, METHOD=1 is recommended.**

Nonparametric Clustering Methods

- Method 2
 - Begin with each observation in a separate cluster. For each observation, find the neighbor with the greatest estimated density that exceeds the estimated density of the observation. If such a neighbor exists, join the cluster to which the observation belongs with the cluster to which the specified neighbor belongs.

Nonparametric Clustering Methods

- Method 3
 - Begin with each observation in a separate cluster. For each observation, find the neighbor with greater estimated density such that the slope of the line that connects the point on the estimated density surface at the observation with the point on the estimated density surface at the neighbor is a maximum.

Nonparametric Clustering Methods

- Method 4
 - This method is equivalent to the first stage of two-stage density linkage **without** the use of the MODE= option.

Nonparametric Clustering Methods

- Method 5
 - This method is equivalent to the first stage of two-stage density linkage (see hierarchical clustering) with the use of the MODE= option.

Nonparametric Clustering Methods

- Method 6
 - Begin with all observations unassigned. PROC MODECLUS first forms a list of seeds, each seed being a single observation. It adds to the cluster any unassigned seed that is a neighbor of a member of the cluster or that shares a neighbor with a member of the cluster. If the THRESHOLD= option is less than 0.5, PROC MODECLUS also forms a list of unassigned observations in decreasing order of estimated density.
 - Method 6 is the most computationally intensive; **it should be used with caution.**

Scaling the Input Variables

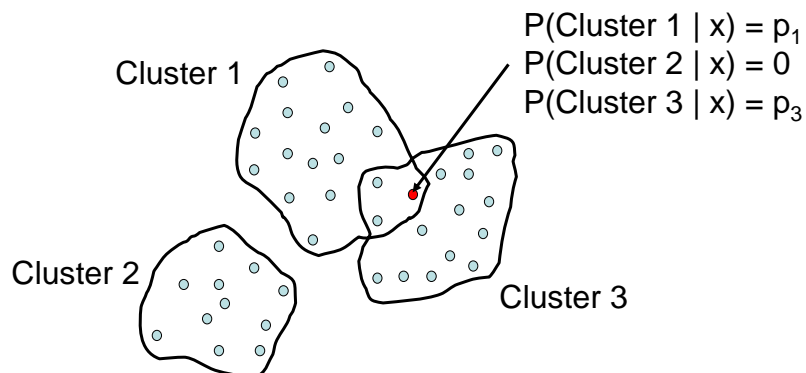
- Variables with large variances tend to have disproportionate influence on the clusters.
- The STD option in PROC MODECLUS scales variables to have equal variances.
- Outliers should be removed before using the STD option.

Standardization in PROC MODECLUS

- Even if the variables use comparable units of measurement, standardization or scaling may still be desirable, particularly if the scale estimates of the variables are not related to their expected importance in defining clusters.
- If you want two variables to have equal importance in the analysis, they should have roughly equal scale estimates. Conversely, if you want one variable to have more effect than another, then assign it a larger scale estimate (a larger range).

Fuzzy Clustering

- In fuzzy clustering, each observation belongs to all clusters with a certain probability, which could be zero.



Fuzzy Clustering

- When you group people or objects into clusters, discrete clusters are often reasonable.
- For example, a customer can either be a responder or a nonresponder. An animal is either a cat, or a dog, or a bird, and so on.
- However, sometimes the boundaries between clusters are not so clear.

Fuzzy Clustering

- Consider a stock market example. You suspect that certain types of companies will perform well under some economic conditions, while others perform well under different economic conditions.
- Sometimes a specific company performs similarly to the companies in group 1, while at other times it performs more like the companies in group 2.
- The boundaries that divide the clusters are unclear, or fuzzy.

Q-Technique Factor Analysis

- In Q-technique factor analysis, observations can be scored on their similarity to a cluster.
- These scores can be converted to a percentage of variance explained; that is, they represent the proportion of variability in each observation explained by each cluster.
- These scores are not, strictly speaking, probabilities of group membership, but they do answer the question “What is the degree of membership in each cluster exhibited by each observation (case)?”

R-Technique Factor Analysis

- In R-technique factor analysis, the most common form of factor analysis, the data table has variables in the columns and observations in the rows.
- The implied similarity measure is the correlation between variables.

Not Your Usual Factor Analysis!

		Observations				
		Lori	Joe	Mary	Tina	Dan
Variables	9	15	11	11	9	
	25	12	11	18	22	
	35	22	33	28	24	
	15	10	11	22	16	
	24	15	26	19	28	
	11	9	11	9	7	
	... more variables					
	12	20	14	12	11	
	26	14	25	18	22	
	9	17	27	11	20	

Q-Technique Clustering

- Q-technique clustering is simply factor analysis on a transposed data matrix; that is, in Q-technique clustering, columns represent the observations and rows represent the variables.
- The implied similarity measure is the correlation between observations.

Q-Technique Clustering

- It is important to have more rows than columns in the transposed data table (more variables than observations).
- This constraint either severely limits the number of observations that can be considered or it forces consideration of a large number of variables.

Q-Technique Clustering

- Because Q-technique factor analysis is performed on a matrix of observations and not variables, the assumptions of factor analysis must apply to the variables rather than the cases.
- This is typically not reasonable, and for that reason, any hypothesis tests regarding the resulting clusters are probably invalid.
- However, Q-technique clustering may still be extremely useful as a descriptive tool.

Factor Analytic Model for Clustering

- $Y = X\beta + E$
- where
 - Y observations to be clustered
 - X factor scores
 - β factor pattern
 - E residuals (errors).

Factor Analytic Model for Clustering

- A frequent source of confusion in the field of factor analysis is the term factor. Here it refers to a hypothetical, unobservable variable.
- *Factors* are estimates of the latent commonalities that you hypothesize exist in the population.
- In this case, the factors are (uncorrelated) clusters.

Randomly Sampling the Variables

- Factor analysis assumes independent rows.
- Normally this means randomly sampling observations.
- In Q-technique clustering, the rows are the variables that you measured.
- Select a wide variety of variables that allow you to differentiate among your subjects or units.

Carefully Selected Observations

- Just as you carefully select variables for R-factor analysis, in Q-factor analysis, carefully select representative observations to be clustered.
- These observations should be good representatives of the groups that you think exist in the population.



What Is the Similarity Measure Used?

- In fuzzy clustering, correlation is the similarity measure.
- Correlation tends to ignore mean differences.
- For example, the following subjects have very different magnitudes, yet they yield a correlation value of 1.0:

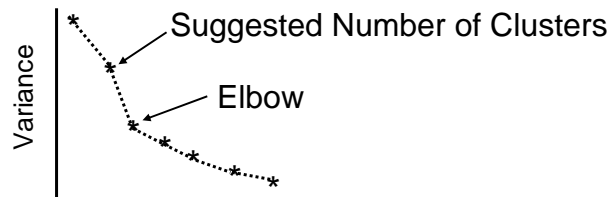
Subject 1	Subject 2
2	15
3	16
1	14
6	19

Determining the Number of Clusters

- When you perform Q-technique clustering, you initially extract as many clusters as there are columns in the analysis (or in the correlation matrix).
- However, you do not typically want to retain all the clusters.
- One of the most useful methods for determining how many clusters to retain is to plot the eigenvalues of the factors by the factor numbers (1, 2, 3, ...).

Scree Plots

- Scree plot of eigenvalues:



- Proportion of variance explained by each cluster
- Cumulative variance explained by clusters
- Account for 100% of the variance (default).

Understanding Factor Loadings

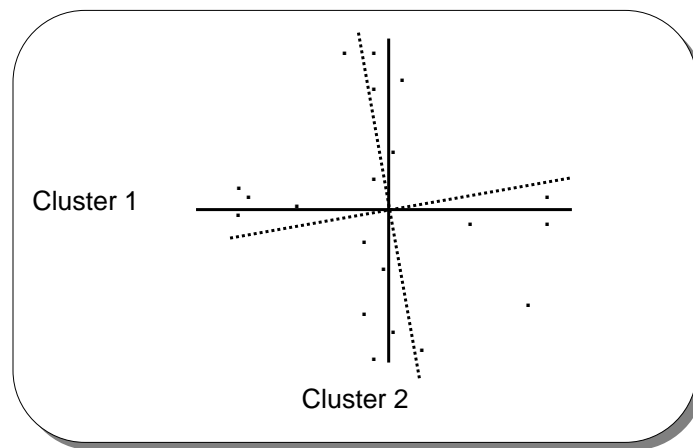
Factor1 Factor2

Lori	.92	.22
Joe	-.87	.12
Mary	.89	-.08
Tina	-.15	.80
Dan	.02	.95

Understanding Factor Loadings

- The loadings can be interpreted as the correlation between each column in the data matrix and the corresponding cluster.
- By squaring the factor (cluster) loadings, you can obtain an R^2 statistic that expresses the proportion of variability in each observation that is explained by a cluster.

Interpreting Cluster Loadings: Rotation



Complexity of Interpretation

- All cases are involved in defining each cluster:
 - Sometimes your cases do not clearly load on any one cluster, but load moderately on more than one.
 - Sometimes they do not load on any clusters at all.
- Factor analysis uses an over-parameterized model:
 - Rotating different numbers of clusters can produce dramatically different cluster structures.
 - Using different rotation methods can produce different cluster structures.