

Introduction to Cluster Analysis

Cluster Analysis

- Cluster Analysis is a collection of techniques for aggregating objects into groups based on similarity measures or distances (dissimilarity measures).
- Cluster Analysis is a set of methods for constructing a (hopefully) sensible and informative classification of an initially unclassified set of data, using the variable values observed on each individual. (Everitt, 1998)

Examples of Clustering Applications

- Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs.
- Land use: Identification of areas of similar land use in an earth observation database.
- Insurance: Identifying groups of motor insurance policy holders with a high average claim cost.
- City-planning: Identifying groups of houses according to their house type, value, and geographical location.
- Earth-quake studies: Observed earth quake epicenters should be clustered along continent faults.

Various Terms used for Cluster Analysis

- botryology
- classification
- clumping
- competitive learning
- morphometrics
- nosography
- nosology
- numerical taxonomy
- partitioning
- Q-analysis
- systematics
- taximetrics
- taxonorics
- typology
- unsupervised pattern recognition
- vector quantization
- winner-take-all learning
- aciniformics
- agminatics

Unsupervised Learning

- *Unsupervised learning* is learning without *a priori* knowledge about the classification of samples; learning without a teacher. (Kohonen, 1995)

Data Structures

Data matrix

$$\begin{bmatrix} x_{11} & \dots & x_{1j} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{ij} & \dots & x_{ip} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nj} & \dots & x_{np} \end{bmatrix}$$

Dissimilarity matrix

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix}$$

Distance and Similarity

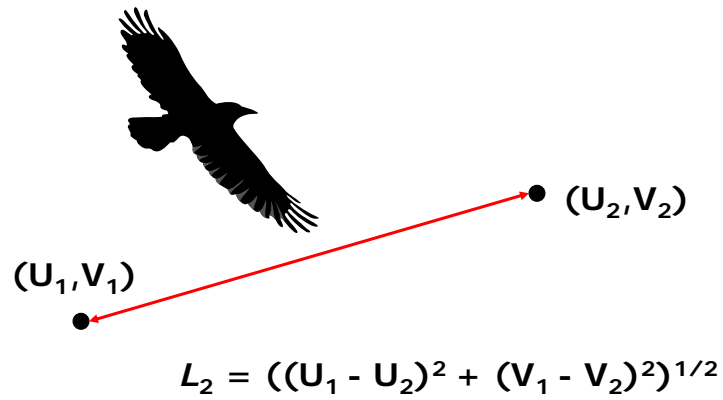
- Items within a cluster are similar, and/or the distance between them is small.
- Items in different clusters are dissimilar, and/or the distance between them is large.

Distance Measures for Metric Variables

1. Euclidean (Ruler) Distance

$$\begin{aligned}d_2(\mathbf{x}_r, \mathbf{x}_s) &= \left[(\mathbf{x}_r - \mathbf{x}_s)' (\mathbf{x}_r - \mathbf{x}_s) \right]^{\frac{1}{2}} \\ &= \left[\sum_{j=1}^p (x_{rj} - x_{sj})^2 \right]^{\frac{1}{2}}\end{aligned}$$

Euclidean Distance



Distance Measures for Metric Variables

2. Standardized Euclidean Distance

$$d_{2s}(\mathbf{x}_r, \mathbf{x}_s) = \left[(\mathbf{z}_r - \mathbf{z}_s)' (\mathbf{z}_r - \mathbf{z}_s) \right]^{1/2}$$

Distance Measures for Metric Variables

3. Mahalanobis Distance

$$d_M(\mathbf{x}_r, \mathbf{x}_s) = \left[(\mathbf{x}_r - \mathbf{x}_s)' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_r - \mathbf{x}_s) \right]^{\frac{1}{2}}$$

where $\boldsymbol{\Sigma}$ is estimated by \mathbf{S}

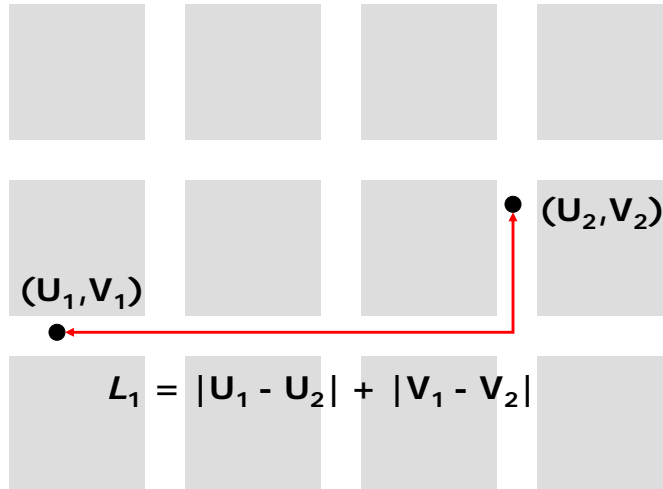
Distance Measures for Metric Variables

4. General Minkowski Metrics

$$d_m(\mathbf{x}_r, \mathbf{x}_s) = \left[\sum_{j=1}^p |x_{rj} - x_{sj}|^m \right]^{\frac{1}{m}}$$

For $m=1$, d_1 measures city block distance
For $m=2$, d_2 measures Euclidean distance

Manhattan Distance



Other metric scales

- Ordinal variables
 - Treat like an interval-scaled variable
- Ratio-scaled variables
 - A positive measurement on a nonlinear scale, approximately at an exponential scale
 - Apply logarithmic transformation
 - Alternatively, rank the variable and treat the ranks as an interval-scaled variable.

Similarity Metrics

- Similarity between people, units, or objects can be quantified using either correlation metrics or distance metrics.
- The metric chosen to operationalize similarity can have substantial impact on the kinds of clusters recovered.

Similarity Measures for Categorical Variables

- Pairs of items can be compared based on the presence or absence of characteristics. Similar items have more characteristics in common than dissimilar items.

Similarity Measures for Categorical Variables

- Let \mathbf{x} be a dichotomous variable that assumes the value 1 if the characteristic is present, and 0 otherwise. Consider a pair of items compared on a set of 5 characteristics, with these results:

	Characteristic				
	\mathbf{x}_1	\mathbf{x}_2	\mathbf{x}_3	\mathbf{x}_4	\mathbf{x}_5
Item i	1	0	0	1	1
Item j	1	1	0	1	0

Similarity Measures for Categorical Variables

Let x_{rj} = score of the r^{th} item on the j^{th} variable

$$\begin{aligned} \text{Then } (x_{rj} - x_{sj})^2 &= 0 && \text{if } x_{rj} = x_{sj} \\ &= 1 && \text{if } x_{rj} \neq x_{sj} \end{aligned}$$

Also, the squared Euclidean distance is

$$d_2^2 = \sum_{j=1}^p (x_{rj} - x_{sj})^2 = \text{number of mismatches}$$

Similarity Measures for Categorical Variables

This measure of similarity weights the 1-1 matches and the 0-0 matches equally. In some cases, a 1-1 match is clearly a stronger indication of similarity than a 0-0 match (for example, the ability to read ancient Greek).

Similarity Measures for Categorical Variables

Consider the following table summarizing the number of matches and mismatches for 2 items on a set of p characteristics:

		Item j		Total
		1	0	
Item i	1	a	b	a+b
	0	c	d	c+d
Total		a+c	b+d	p

Similarity Coefficients

Coefficient	Rationale
$\frac{a+d}{p}$	Equal weights for 1-1 and 0-0 matches
$\frac{2(a+d)}{2(a+d)+b+c}$	Double weights for 1-1 and 0-0 matches
$\frac{a+d}{a+d+2(b+c)}$	Double weights for unmatched pairs

Similarity Coefficients

Coefficient	Rationale
$\frac{a}{a+b+c}$	0-0 matches not counted at all
$\frac{2a}{2a+b+c}$	No 0-0 matches; double weights for 1-1 matches
$\frac{a}{a+2(b+c)}$	No 0-0 matches; double weights for unmatched pairs

Example on calculating the values of a similarity coefficient

Individual	Height	Weight	Eye Color	Hair Color	Handedness	Gender
1	68"	140	Green	Blond	Right	Female
2	73"	185	Brown	Brown	Right	Male
3	67"	165	Blue	Blond	Right	Male
4	64"	120	Brown	Brown	Right	Female
5	76"	210	Brown	Brown	Left	Male

Example on calculating the values of a similarity coefficient

Define six dichotomous variables:

- $X_1 = 1$ if Height > 72", 0 otherwise
- $X_2 = 1$ if Weight > 150, 0 otherwise
- $X_3 = 1$ if brown eyes, 0 otherwise
- $X_4 = 1$ if blond hair, 0 otherwise
- $X_5 = 1$ if right handed, 0 otherwise
- $X_6 = 1$ if female, 0 otherwise

Example on calculating the values of a similarity coefficient

Individual	Height	Weight	Eye Color	Hair Color	Handedness	Gender
1	68"	140	Green	Blond	Right	Female
2	73"	185	Brown	Brown	Right	Male
3	67"	165	Blue	Blond	Right	Male
4	64"	120	Brown	Brown	Right	Female
5	76"	210	Brown	Brown	Left	Male

The scores for individuals 1 and 2 on the $p=6$ dichotomous variables are:

Individual	X_1	X_2	X_3	X_4	X_5	X_6
1	0	0	0	1	1	1
2	1	1	1	0	1	0

Example on calculating the values of a similarity coefficient

Individual	X_1	X_2	X_3	X_4	X_5	X_6
1	0	0	0	1	1	1
2	1	1	1	0	1	0

The number of matches and mismatches can be displayed as:

		Individual 2		Total
		1	0	
Individual 1	1	1	2	3
	0	3	0	3
Total		4	2	6

Example on calculating the values of a similarity coefficient

Define a similarity coefficient as:

$$s = \frac{a + d}{p} = \frac{\# \text{ matches}}{\# \text{ variables}}$$

Example on calculating the values of a similarity coefficient

Then the matrix displaying the similarities between all individuals is:

		Individual				
		1	2	3	4	5
Individual	1	1				
	2	1/6	1			
	3	4/6	3/6	1		
	4	4/6	3/6	2/6	1	
	5	0	5/6	2/6	2/6	1

What Makes a Good Similarity Metric?

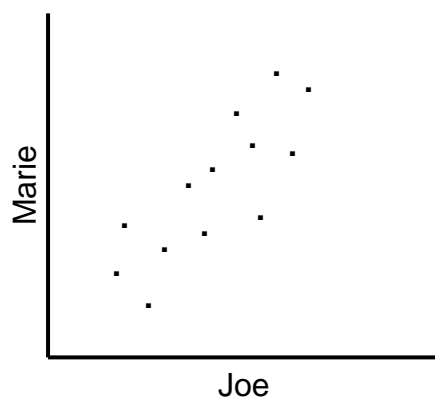
A good similarity metric should exhibit the following:

1. Symmetry: $d(x,y) = d(y,x)$
2. Triangular inequality: $d(x,y) \leq d(x,z) + d(y,z)$
3. Nonidentical distinguishability: if $d(x,y) \neq 0$ then $x \neq y$
4. Identical nondistinguishability: if $x = y$, then $d(x,y) = 0$.

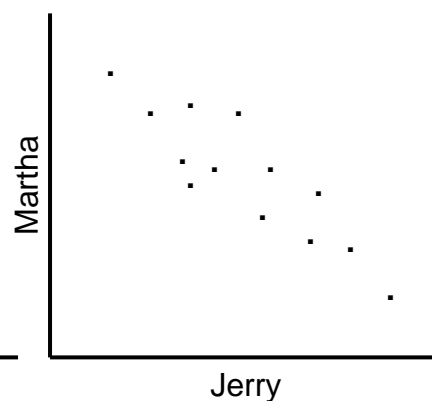
Sometimes the easiest measures to use, for example, Pearson's correlation, fail to meet one or more of these criteria.

Correlation-Type Metrics

Similar:



Dissimilar:



The Problem with Correlation-Type Metrics

<u>Variable</u>	<u>Observation 1</u>	<u>Observation 2</u>
1	5	51
2	4	42
3	3	33
4	2	24
5	1	15

The (Pearson) correlation between observations 1 and 2 is 1.0, but are the observations **really** similar?

Relationship between distance and similarity

Let d_{ij} = distance between items i & j
 s_{ij} = similarity between items i & j

If distances are calculated, but instead need to determine similarities, one possible choice might be:

$$s_{ij} = \frac{1}{1 + d_{ij}}$$

which enjoys the constraint: $0 \leq s_{ij} \leq 1$

The DISTANCE Procedure

PROC DISTANCE computes various measures of distance, dissimilarity, or similarity between the observations (rows) of a SAS data set.

PROC DISTANCE also provides various nonparametric and parametric methods for standardizing variables. Different variables can be standardized with different methods.

Relationship between distance and similarity

However, distances cannot always be constructed from similarities. The construction is possible only if the matrix of similarities is nonnegative definite.

If $\mathbf{S} = (s_{ij})$ is nonnegative definite with $s_{ii} = 1$, then

$$d_{ij} = \sqrt{2(1 - s_{ij})}$$

has the properties of distance.

Types of Clusters Possible

- Disjoint clusters place each object in one and only one cluster.
- Hierarchical clusters are organized so that one cluster may be entirely contained within another cluster, but no other kind of overlap between clusters is allowed.
- Overlapping clusters can be constrained to limit the number of objects that belong simultaneously to two clusters, or they can be unconstrained, allowing any degree of overlap in cluster membership.
- Fuzzy clusters are defined by a probability or grade of membership of each object in each cluster. Fuzzy clusters can be disjoint, hierarchical, or overlapping.

Common Data Representations of Objects to be Clustered

- a square distance or similarity matrix, in which both rows and columns correspond to the objects to be clustered. A correlation matrix is an example of a similarity matrix.
- a coordinate matrix, in which the rows are observations and the columns are variables, as in the usual SAS multivariate data set. The observations, the variables, or both may be clustered.

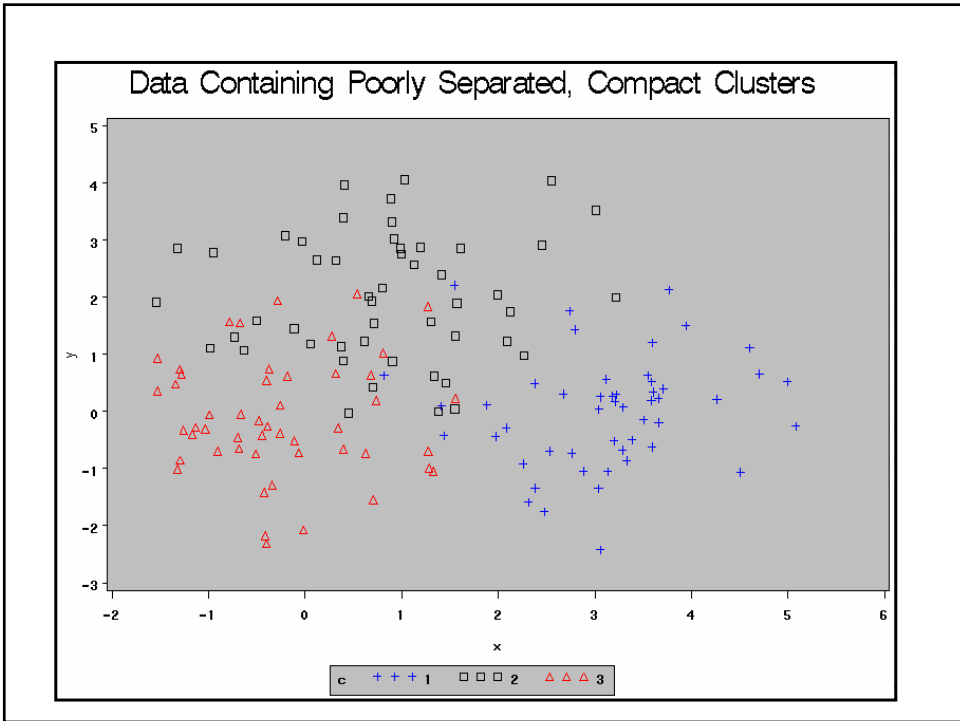
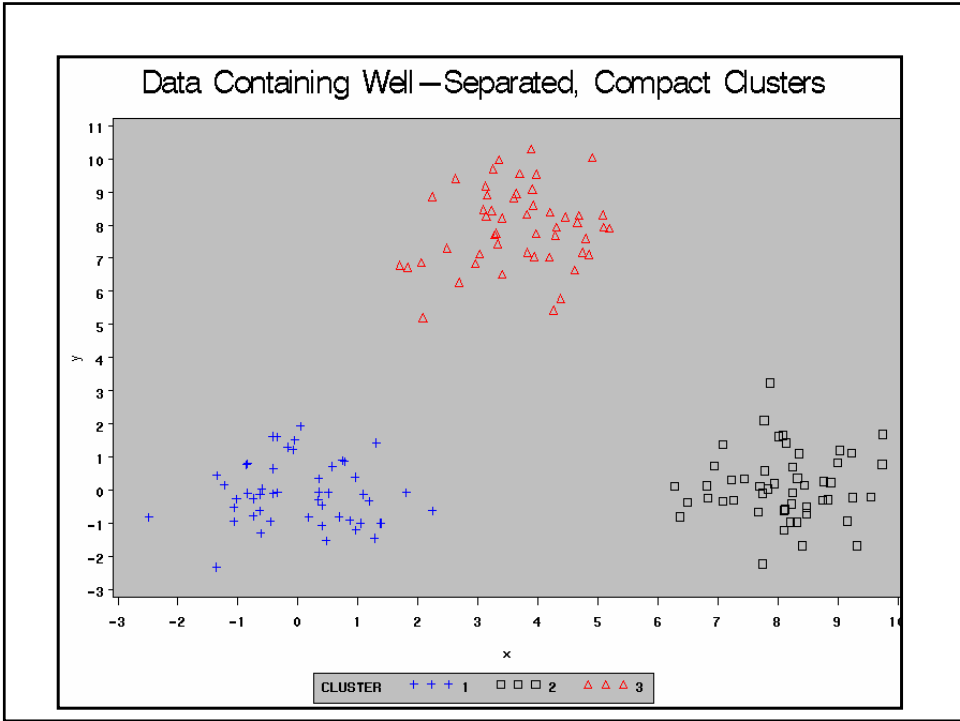
Example of a Distance Matrix

	Beaver	Raccoon	Fox	Dog	Wolf
Beaver	0	0.4817	0.8851	0.8580	0.9987
Raccoon	0.4817	0	0.8765	0.9876	0.8865
Fox	0.8851	0.8765	0	0.3956	0.4408
Dog	0.8580	0.9876	0.3956	0	0.4146
Wolf	0.9987	0.8865	0.4408	0.4146	0

Graphical Aids for Clustering

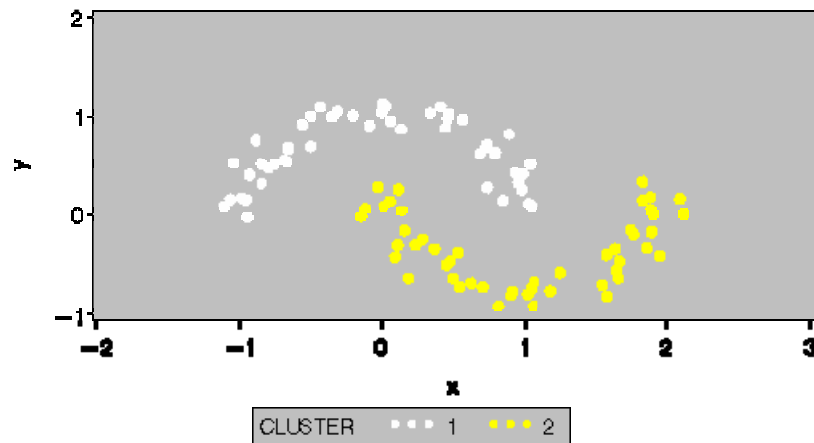
Scatter Plots

- 2-D & 3-D plots of the original variables
- 2-D & 3-D plots of the principal components



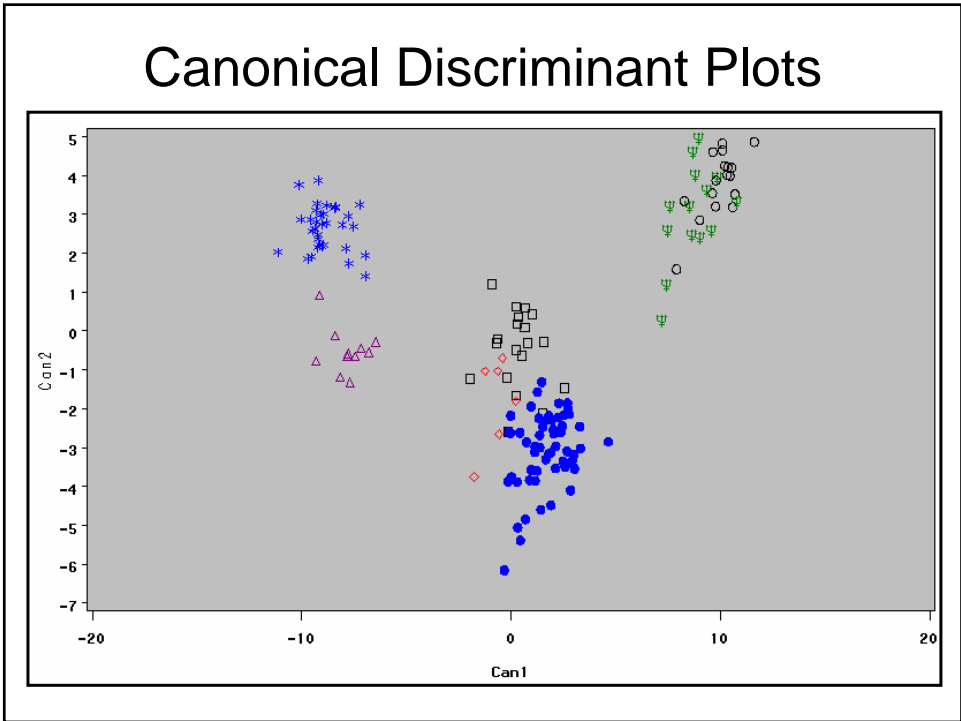
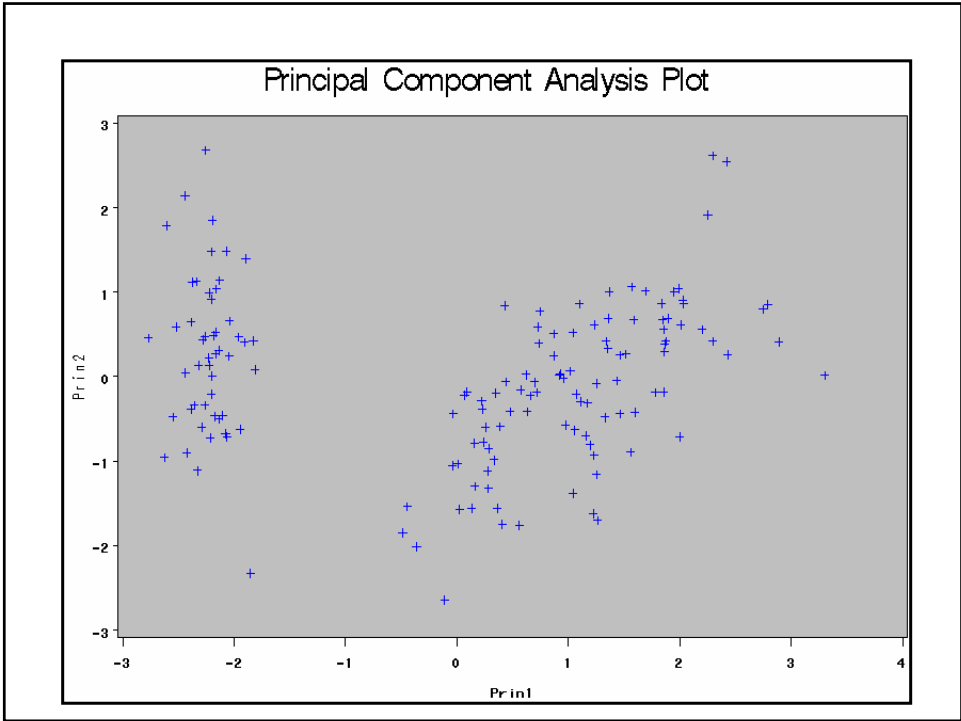
Nonconvex Clusters

Two-Stage Density Linkage Cluster Analysis
of Data Containing Nonconvex Clusters

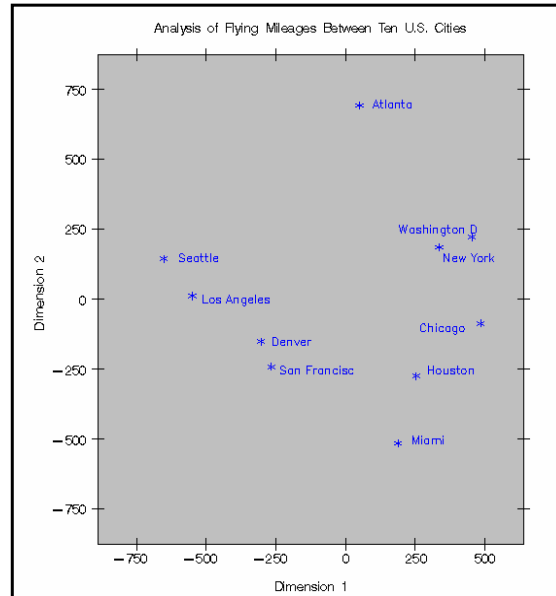


Plotting More Than Two Dimensions

- How do you plot more than two dimensions?
 - The G3D procedure can be used to plot three dimensions at a time.
 - For more than three dimensions you can use
 - principal component plots
 - canonical discriminant plots
 - multidimensional scaling (MDS) plots.



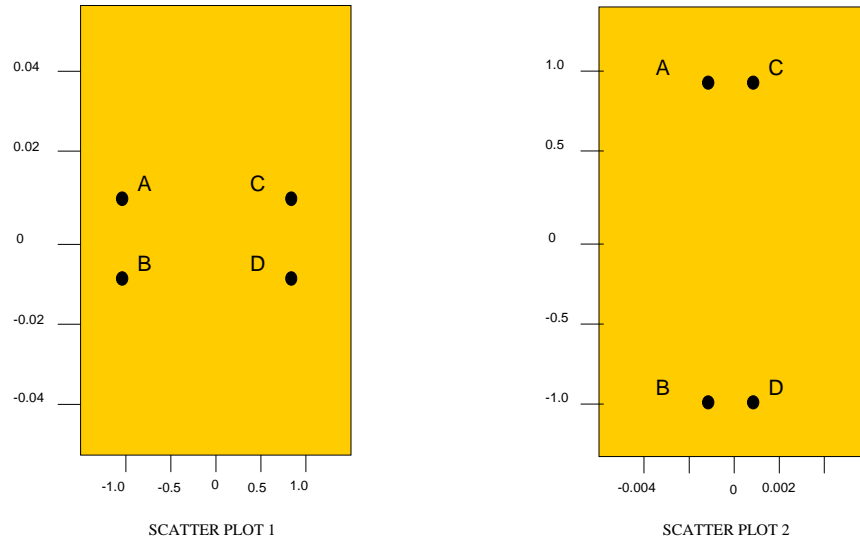
Multidimensional Scaling Plots



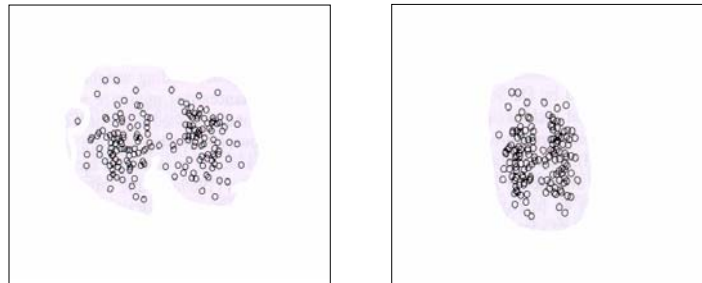
The Order Effect

- In some cases, the order in which the observations are presented can make a difference.
- A researcher could apply the same analysis to the same data (except for input order) and find entirely different clusters!

The Importance of Scale



The Standardization Problem



Before

After

Standardization using the reciprocal of the standard deviation may actually dilute the differences between groups. (Everitt, 2001)

The STDIZE Procedure

- General form of the STDIZE procedure:

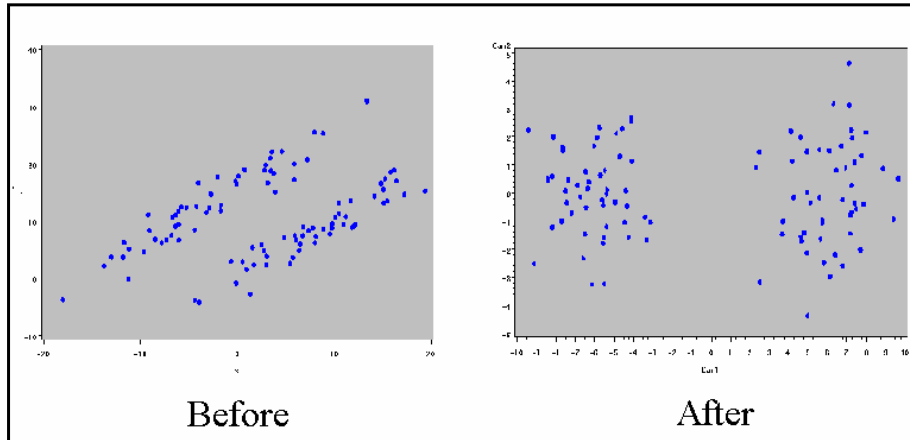
```

PROC STDIZE <options>;
      VAR variables;
      BY variables;
      LOCATION variables;
      SCALE variables;
RUN;
  
```

Standardization Methods in PROC STDIZE

METHOD	LOCATION	SCALE
MEAN	mean	1
MEDIAN	median	1
SUM	0	sum
EUCLEN	0	Euclidean Length
USTD	0	standard deviation about origin
STD	mean	standard deviation
RANGE	minimum	range
MIDRANGE	midrange	range/2
MAXABS	0	maximum absolute value
IQR	median	interquartile range
MAD	median	median absolute deviation from median
ABW(c)	biweight 1-step M-estimate	biweight A-estimate
AHUBER(c)	Huber 1-step M-estimate	Huber A-estimate
AWAVE(c)	Wave 1-step M-estimate	Wave A-estimate
AGK(p)	mean	AGK estimate (ACECLUS)
SPACING(p)	mid minimum-spacing	minimum spacing
L(p)	L(p)	L(p) (Minkowski distances)
IN(ds)	read from data set	read from data set

Cluster Preprocessing



The ACECLUS Procedure

(Approximate Covariance Estimation for CLUStering)

PROC ACECLUS attempts to estimate the pooled within-cluster covariance matrix from coordinate data without knowledge of the number or the membership of the clusters (Art, Gnanadesikan, and Kettenring 1982).

PROC ACECLUS outputs a data set containing canonical variable scores to be used in the cluster analysis proper.

The ACECLUS Procedure Some Background

Let $\mathbf{W} = (w_{ij})$ = within-cluster covariance matrix.

Let $\mathbf{A} = (a_{jk})$ be defined as

$$a_{jk} = \frac{\sum_{i=2}^n \sum_{h=1}^{i-1} d_{ih} (x_{ij} - x_{hj})(x_{ik} - x_{hk})}{2 \sum_{i=2}^n \sum_{h=1}^{i-1} d_{ih}}$$

The ACECLUS Procedure Some Background

where

$$d_{ih} = \begin{cases} 1 & \text{if } \sum_{j=1}^p \sum_{k=1}^p m_{jk} (x_{ij} - x_{hj})(x_{ik} - x_{hk}) \leq u^2 \\ 0 & \text{otherwise} \end{cases}$$

where u is an appropriately chosen value and $\mathbf{M} = (m_{jk})$ is an appropriate metric.

The ACECLUS Procedure Some Background

If all of the following conditions hold, **A** equals **W**:

- all within-cluster distances in the metric **M** are less than or equal to u
- all between-cluster distances in the metric **M** are greater than u
- all clusters have the same number of members n_c

The ACECLUS Procedure The Choice of **M**

1. Obtain an initial estimate of matrix **A**. The identity matrix or the total-sample covariance matrix is often used.
2. Set matrix **M** to the inverse of matrix **A**, that is, $\mathbf{M}=\mathbf{A}^{-1}$.
3. Recompute matrix **A** using the formula

$$a_{jk} = \frac{\sum_{i=2}^n \sum_{h=1}^{i-1} d_{ih} (x_{ij} - x_{hj})(x_{ik} - x_{hk})}{2 \sum_{i=2}^n \sum_{h=1}^{i-1} d_{ih}}$$

4. Repeat steps 2 and 3 until the estimate stabilizes.

The ACECLUS Procedure

The Choice of u

Specify a value p , $0 < p < 1$, to be transformed into a value t that is then multiplied by $1/\sqrt{(2p)}$ times the root mean square distance between observations in the current metric on each iteration to yield u .

The value of u changes from iteration to iteration.

This method can be used with a poor initial estimate of \mathbf{A} . (See the PROPORTION= option in the PROC ACECLUS statement.)

The ACECLUS Procedure

Recommended Choice of \mathbf{M} & p

In most cases, the analysis should begin using values of p between 0.5 and 0.01 and using the full total-sample covariance matrix as the initial estimate of \mathbf{A} .