Multivariate Analysis

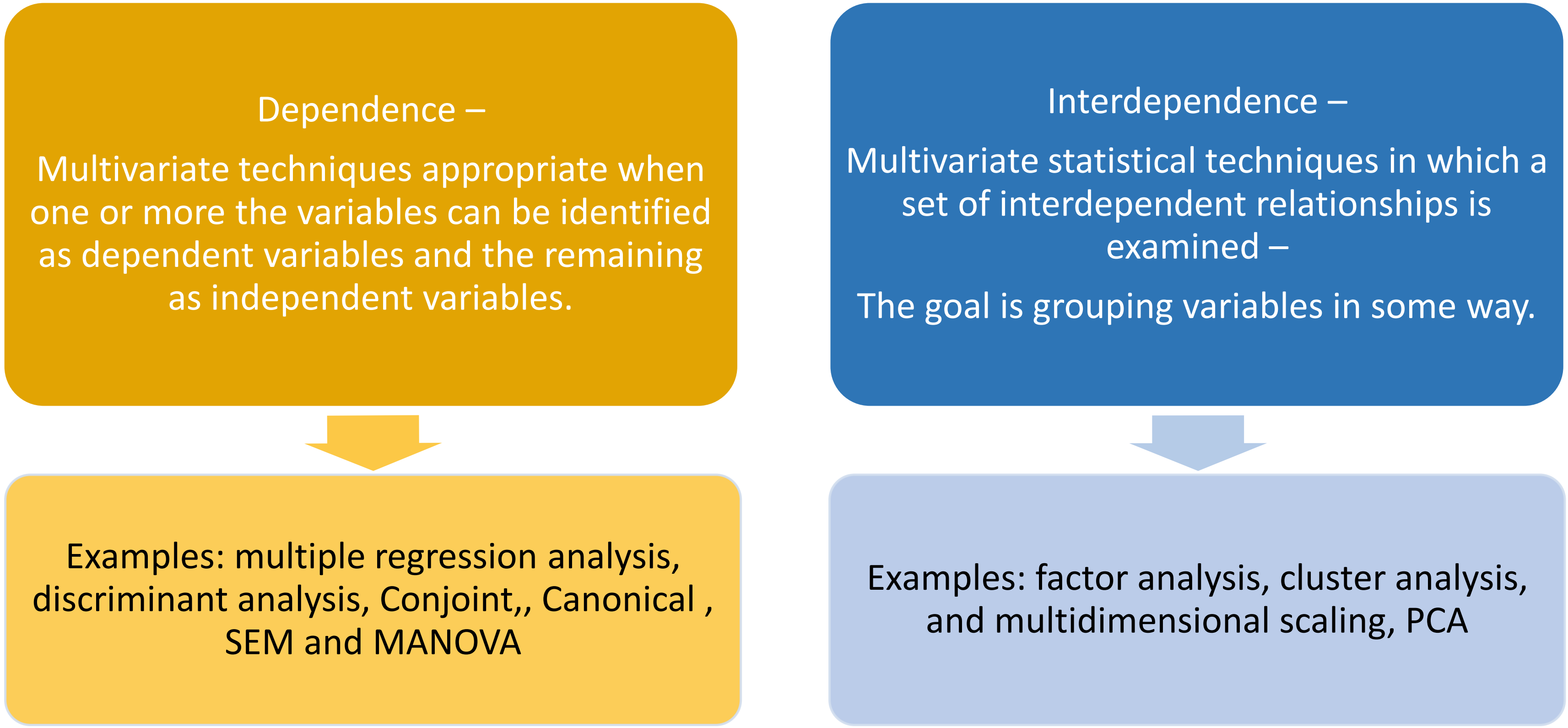
Telkom University

TÜVRheinland®
Precisely Right.

Agenda

- A Classification of Multivariate Technique
- Multidimensional Scaling Analysis
- Factor Analysis
- Cluster Analysis
- Principal Components Analysis
- Diskriminan Analysis
- Multiple Regression Definition
- Multiple Discriminant Analysis
- MANOVA

Dependence vs Interdependence Methods



Dependence –

Multivariate techniques appropriate when one or more the variables can be identified as dependent variables and the remaining as independent variables.

Examples: multiple regression analysis, discriminant analysis, Conjoint,, Canonical , SEM and MANOVA

Interdependence –

Multivariate statistical techniques in which a set of interdependent relationships is examined –

The goal is grouping variables in some way.

Examples: factor analysis, cluster analysis, and multidimensional scaling, PCA

Multivariate Analysis in Statistical Terms

Structured approach to multivariate model building :

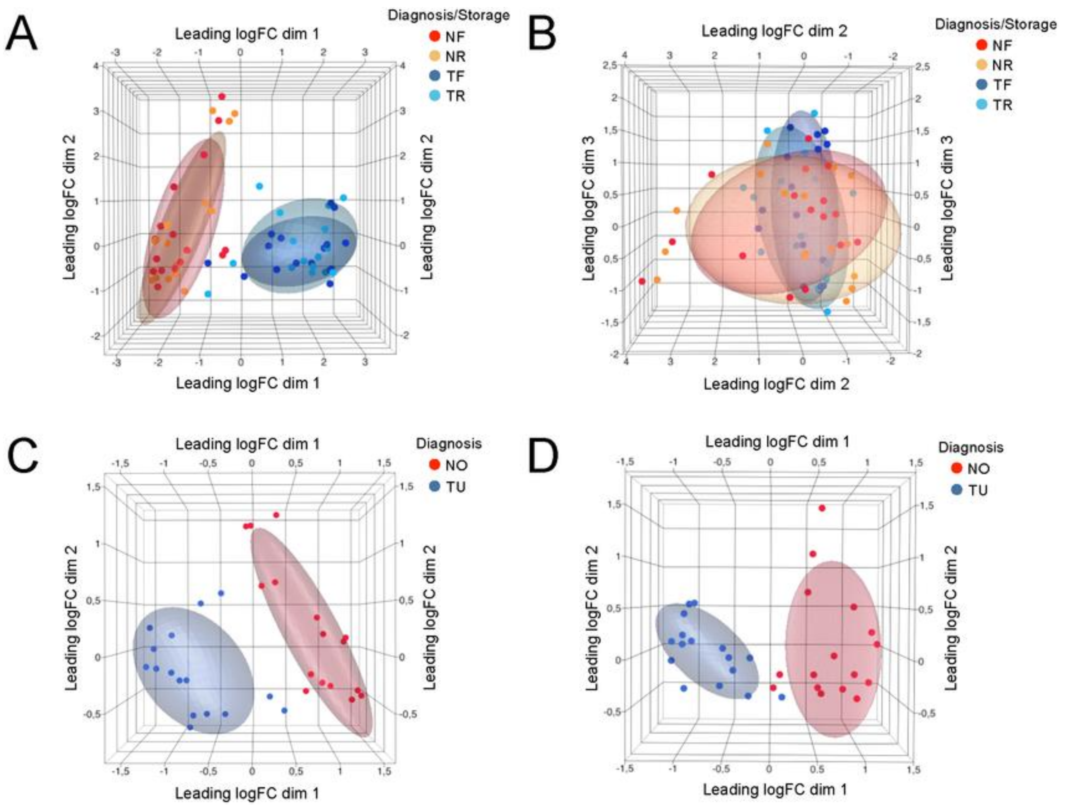
- **Stage 1**
  - Define the research problem, objective, and multivariate technique to be used.
- **Stage 2**
  - Develop the analysis plan.
- **Stage 3**
  - Evaluate the assumptions
- **Stage 4**
  - Estimate the multivariate model and assess overall model fit.
- **Stage 5**
  - Interpret the variate(s)
- **Stage 6**
  - Validate the multivariate model.

Agenda

- A Classification of Multivariate Technique
- **Multidimensional Scaling Analysis**
- Factor Analysis
- Cluster Analysis
- Principal Components Analysis
- Multiple Regression Definition
- Multiple Discriminant Analysis
- MANOVA

Multidimensional
Scaling Analysis

- Generally regarded as exploratory data analysis (Ding, 2006).
- Reduces large amounts of data into easy-to-visualize structures.
- Attempts to find structure (visual representation) in a set of distance measures, e.g. dis/similarities, between objects/cases.
  - Shows how variables/objects are related perceptually.
- How? By assigning cases to specific locations in space.
- Distances between points in space match dis/similarities as closely as possible:
  - Similar objects: Close points
  - Dissimilar objects: Far apart points



Agenda

- A Classification of Multivariate Technique
- Multidimensional Scaling Analysis
- **Factor Analysis**
- Cluster Analysis
- Principal Components Analysis
- Multiple Regression Definition
- Multiple Discriminant Analysis
- MANOVA

Factor Analysis

- Factor analysis (FA) is an interdependence technique whose primary purpose is to define the underlying structure among the variables in the analysis.
- The purpose of FA is to condense the information contained in a number of original variables into a smaller set of new composite dimensions or variates (factors) with a minimum loss of information.

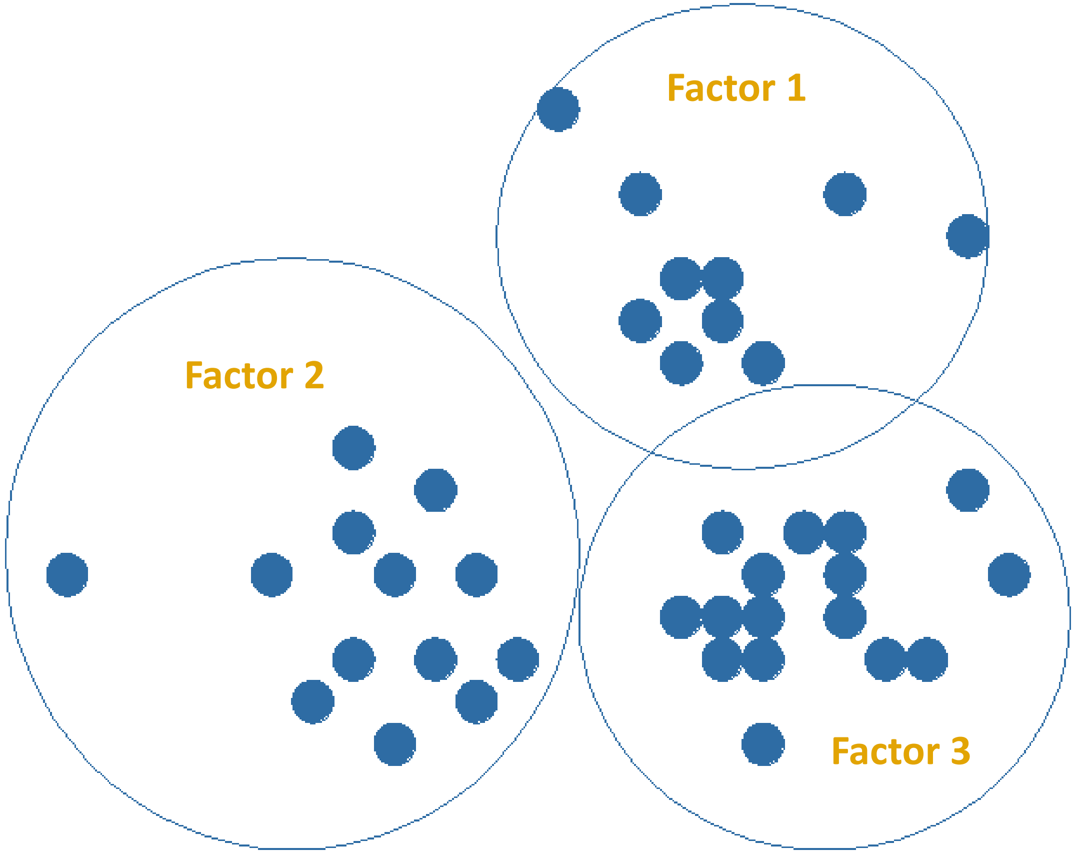
When to use
Factor Analysis

- Data Reduction
- Identification of underlying latent structures
  - Clusters of correlated variables are termed factors
  - Example: factor analysis could potentially be used to identify the characteristics) out of a large number of characteristics) that make a person popular.
- Candidate characteristics: Level of social skills, selfishness, how interesting a person is to others, the amount of time they spend talking about themselves (Talk 2) versus the other person (Talk 1), their propensity to lie about themselves

Conceptual Model of Factor Analysis

- The variance of several variables may be largely explained by a smaller number of underlying clusters (factors), with each factor consisting of several variables

- FA uses correlations among many variables to sort related variables into clusters called "factors"



Basic Concept of Factor Analysis

- The fundamental idea underlying the factor analysis is that some but not all variables can be directly observed.
- Those unobserved variables are referred to as either latent variables or factors.
- Information about latent variables can be gained by observing their influence on observed variables.
- Factor analysis examines covariation among a set of observed variables trying to generate a smaller number of latent variables.

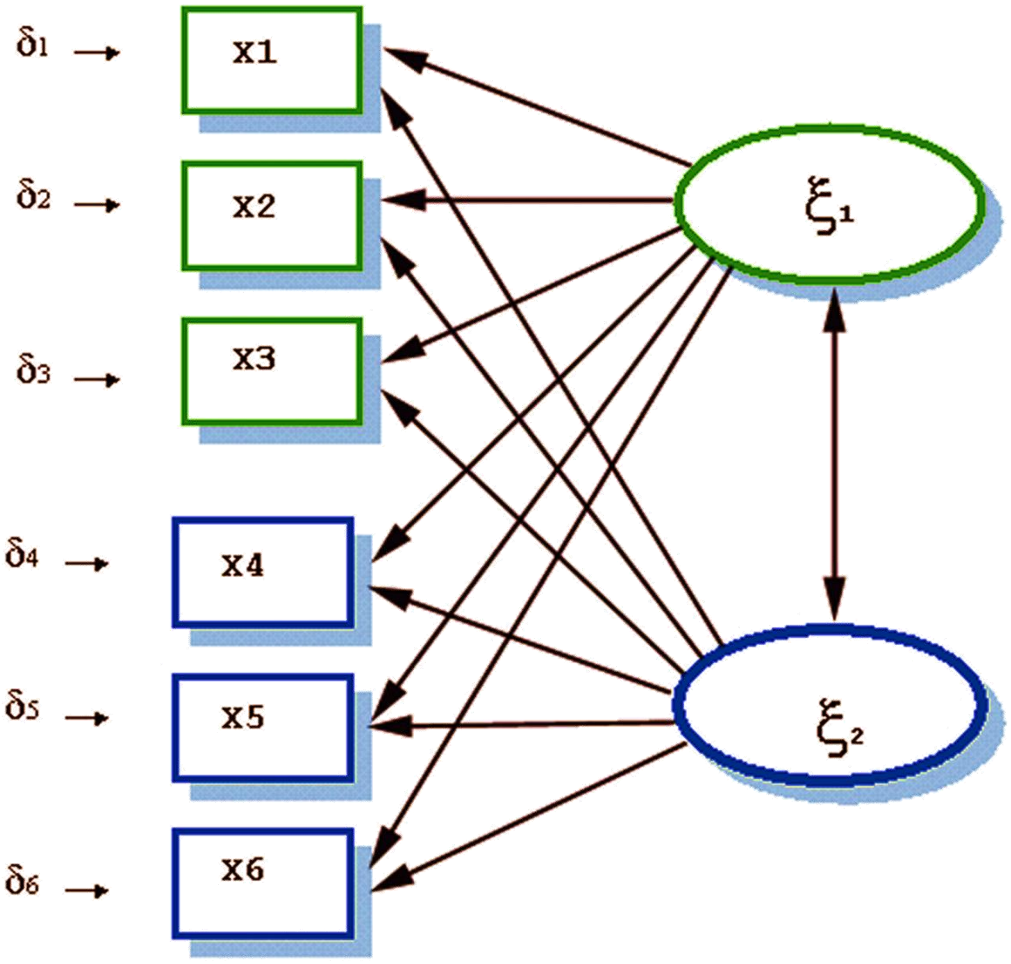
Basic Concept of Factor Analysis

Exploratory Factor Analysis

- In exploratory factor analysis (EFA), observed variables are represented by squares and circles represent latent variables.
- Causal effect of the latent variable on the observed variable is presented with straight line with arrowhead.
- The latent factors (ellipses) labeled with ξ's (Xi) are called common factors and the δ's (delta) (usually in circles) are called errors in variables or residual variables.
- Errors in variables have unique effects to one and only one observed variable - unlike the common factors that share their effects in common with more than one of the observed variables.

Basic Concept of
Factor Analysis

Exploratory Factor Model
(Nokelainen, 1999.)

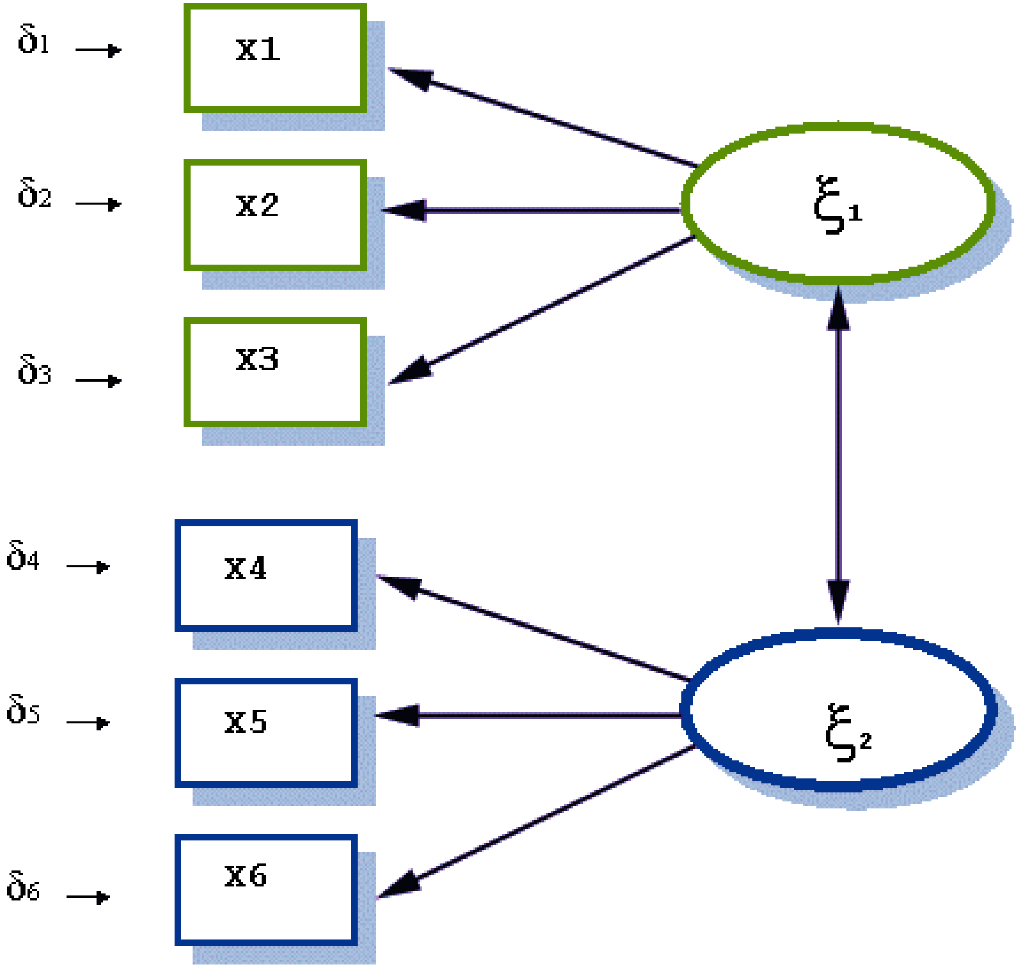


Basic Concept of
Factor Analysis

Confirmatory Factor Analysis

- One of the biggest problems in EFA is its inability to incorporate substantively meaningful constraints.
- That is due to fact that algebraic mathematical solution to solve estimates is not trivial, instead one has to seek for other solutions.
- That problem was partly solved by the development of the confirmatory factor model, which was based on an iterative algorithm (Jöreskog, 1969).
- In confirmatory factor analysis (CFA), which is a special case of SEM, the correlations between the factors are an explicit part of the analysis because they are collected in a matrix of factor correlations.
- With CFA, researcher is able to decide a priori whether the factors would correlate or not. (Tacq, 1997.) Moreover, researcher is able to impose substantively motivated constraints,
  - which common factor pairs that are correlated,
  - which observed variables are affected by which common factors,
  - which observed variables are affected by a unique factor and
  - which pairs of unique factors are correlated. (Long, 1983.)

Basic Concept of Factor Analysis



Confirmatory Factor Model
(Nokelainen, 1999.)

Structural Equation Models

- Structural equation modeling (SEM), as a concept, is a combination of statistical techniques such as exploratory factor analysis and multiple regression.
- The purpose of SEM is to examine a set of relationships between one or more Independent Variables (IV) and one or more Dependent Variables (DV).
- Both IV's and DV's can be continuous or discrete.
- Independent variables are usually considered either predictor or causal variables because they predict or cause the dependent variables (the response or outcome variables).
- Structural equation modeling is also known as 'causal modeling' or 'analysis of covariance structures'.
- Path analysis and confirmatory factor analysis (CFA) are special types of SEM

Model Constructing

- One of the most well known covariance structure models is called LISREL (LInear Structural RELationships) or Jöreskog-Keesling-Wiley –model.
- LISREL is also a name of the software (Jöreskog et al., 1979), which is later demonstrated in this presentation to analyze a latent variable model.
- The other approach in this study field is Bentler-Weeks -model (Bentler et al., 1980) and EQS –software (Bentler, 1995). The latest software release attempting to implement SEM is graphical and intuitive AMOS (Arbuckle, 1997).

All the previously mentioned approaches to SEM use the same pattern for constructing the model:

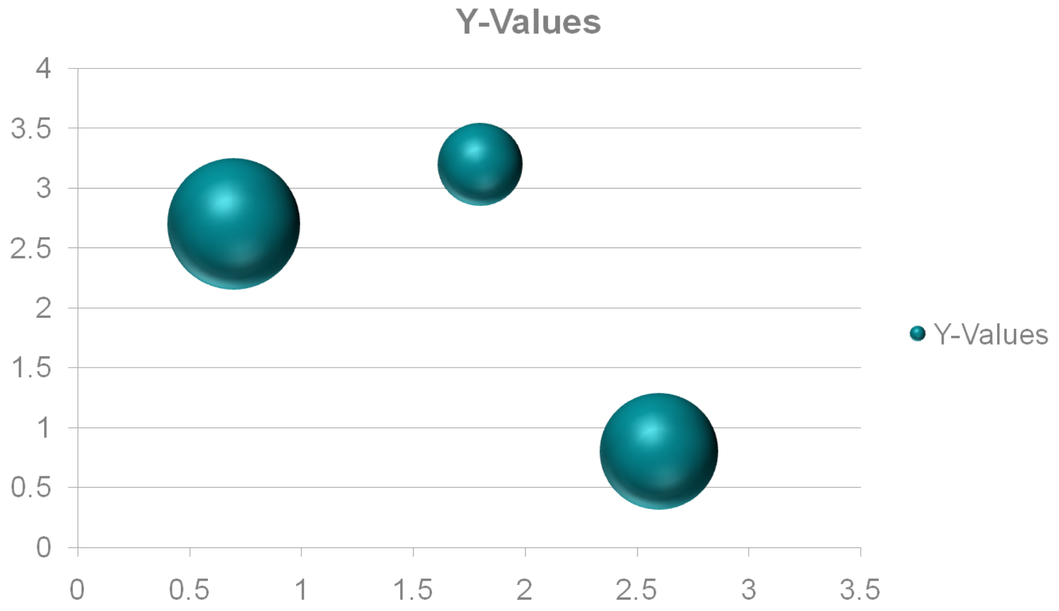
- model hypotheses,
- model specification,
- model identification and
- model estimation.

Agenda

- A Classification of Multivariate Technique
- Multidimensional Scaling Analysis
- Factor Analysis
- **Cluster Analysis**
- Principal Components Analysis
- Multiple Regression Definition
- Multiple Discriminant Analysis
- MANOVA

Cluster Analysis

- Basically cluster analysis aims to determine a 'natural group' of individual groups (objects, points, units, or others). This group of individuals can form a complete population or a sample of a larger population.
- More generally, cluster analysis aims to allocate a group of individuals to groups that are mutually independent so that individuals within the group are similar to each other, while individuals in different groups are not similar. This grouping is usually called a partition.



Agenda

- A Classification of Multivariate Technique
- Multidimensional Scaling Analysis
- Factor Analysis
- Cluster Analysis
- **Principal Components Analysis**
- Multiple Regression Definition
- Multiple Discriminant Analysis
- MANOVA

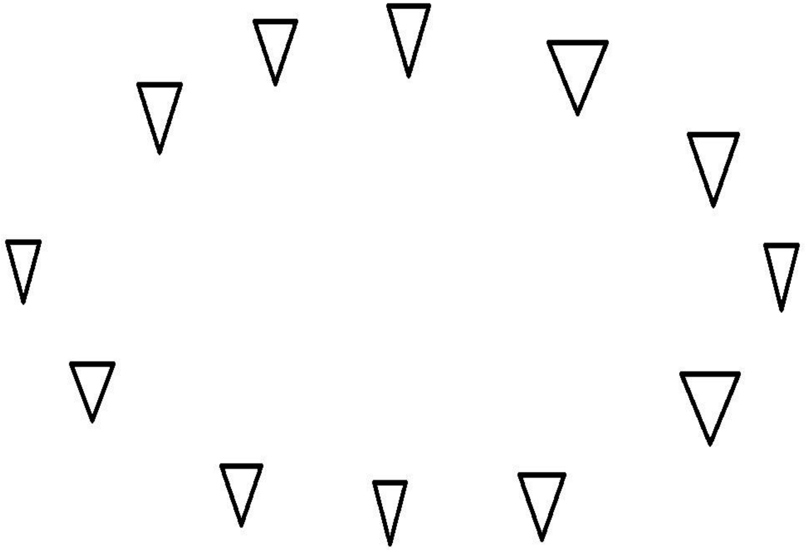
Principal Component Analysis

In the present big data era, there is a need to process large amounts of unlabeled data and find some patterns in the data to use it further.

- Need to discard features that are unimportant and discover only the representations that are needed.
- It is possible to convert high-dimensional data to low-dimensional data using different techniques, this dimension reduction is important and makes tasks such as classification, visualization, communication and storage much easier.
- The loss of information should be less while mapping data from high-dimensional space to low-dimensional space.
- Does the data set ‘span’ the whole of d dimensional space?
- For a matrix of m samples x n genes, create a new covariance matrix of size n x n.
- Transform some large number of variables into a smaller number of uncorrelated variables called principal components (PCs).
- developed to capture as much of the variation in data as possible

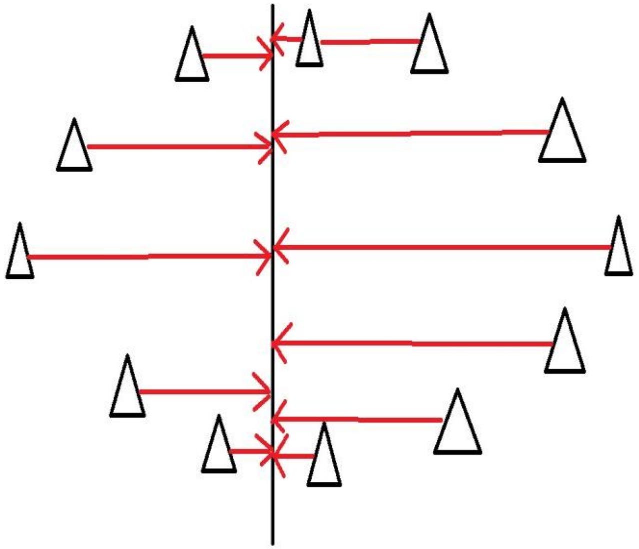
What is
Principal Component Analysis

- They are the directions where there is the most variance, the directions where the data is most spread out.

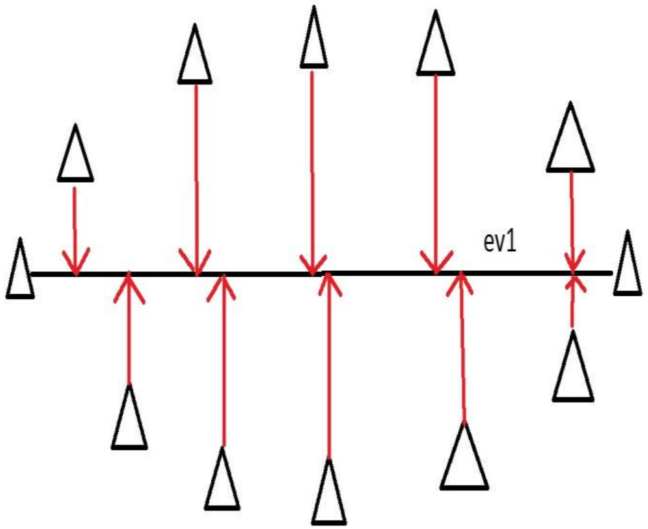


What is
Principal Component Analysis

- To find the direction where there is most variance, find the straight line where the data is most spread out when projected onto it. A vertical straight line with the points projected on to it will look like this:



- On this line the data is way more spread out, it has a large variance. In fact there isn't a straight line you can draw that has a larger variance than a horizontal one. A horizontal line is therefore the principal component in this example



Agenda

- A Classification of Multivariate Technique
- Multidimensional Scaling Analysis
- Factor Analysis
- Cluster Analysis
- Principal Components Analysis
- **Multiple Regression Definition**
- Multiple Discriminant Analysis
- MANOVA

Multiple Regression

Equation and Line

Multiple regression equation – given a collection of sample data with several (*k-many*) explanatory variables, the regression equation that algebraically describes the relationship between the response variable y and two or more explanatory variables *X1, X2, ….Xk* and is:

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k$$

- We are using more than one exploratory variable to predict a response variable now
- In practice, you need large amounts date to use several predictor / exploratory variables
- Guideline: Your sample size should be 10 times larger than the number of *x* variables

Multiple regression line – the graph of the multiple regression equation

- This multiple regression line still fits the sample points best according to the least squares property

Multivariate Analysis

Multiple Regression

Equation and Line

- Multiple Regression Analysis
  - Design requirements
  - Multiple regression model
  - R2
  - Testing R2 and b’s
  - Comparing models
  - Comparing standardized regression coefficients

- Method for studying the relationship between a dependent variable and two or more independent variables.

- Purposes:
  - Prediction
  - Explanation
  - Theory building

Agenda

- A Classification of Multivariate Technique
- Multidimensional Scaling Analysis
- Factor Analysis
- Cluster Analysis
- Principal Components Analysis
- Multiple Regression Definition
- **Multiple Discriminant Analysis**
- MANOVA

What is MDA,

and How Can It be Used to Predict Bankruptcy?

- Multiple discriminant analysis (MDA) is a statistical technique similar to multiple regression
- It identifies the characteristics of firms that went bankrupt in the past
- Then, data from any firm can be entered into the model to assess the likelihood of future bankruptcy

Discriminant Analysis
Defined

- Multiple discriminant analysis is an appropriate technique when the dependent variable is categorical (nominal or nonmetric) and the independent variables are metric.
- The single dependent variable can have two, three, or more categories.
- Examples
  - Gender – Male vs. Female
  - Heavy Users vs. Light Users
  - Purchasers vs. Non-Purchasers
  - Good Credit Risk vs. Poor Credit Risk
  - Member vs. Non-Member
  - Attorney, Physician, or Professior

Agenda

- A Classification of Multivariate Technique
- Multidimensional Scaling Analysis
- Factor Analysis
- Cluster Analysis
- Principal Components Analysis
- Multiple Regression Definition
- Multiple Discriminant Analysis
- **MANOVA**

Multiple Analysis of Variance (MANOVA)

Definition:

- Analysis involving the investigation of the main and interaction effects of categorical (independent) variables on multiple dependent interval variables.

Purpose:

- To determine if individual categorical independent variables have an effect on a group, or related set of interval dependent variables.

For example:

- We may conduct a study where we try two different textbooks (independent variables), and we are interested in the students' improvements in math and physics. In that case, we have two dependent variables, and our hypothesis is that both together are affected by the difference in textbooks.

Assumptions:

- The independent variables are categorical
- There are multiple dependent variables that are continuous and interval.
- There is a relationship between the dependent variables
- The number of observations for each combination of the factor are the same (balanced experiment)