ASM1: multivariate analysis

Chiara Seghieri,
c.seghieri@santannapisa.it

What is Multivariate Analysis?

Starting from a large data matrix:

Multivariate analysis ***helps to****:*

• Classify the **observations** on the basis of the collected variables. Identify the possible existence of groups of observations
→ Data Classification techniques (i.e. Cluster Analysis)

• Reduce the number of **variables**. Study the possible relationships among variables.
→ Data Reduction techniques (i.e. principal component analysis, factor analysis, more advanced like IRT)

Data Classification

Data classification techniques involve a way of **grouping together cases** that are **similar** in a dataset on the basis of the selected variables. The most common way of doing this is through a cluster analysis.

Cluster Analysis

- Term first used by Tryon, 1939
- also called segmentation analysis or taxonomy analysis;
- It covers several different algorithms and methods for grouping objects of similar kind into categories;

- A general question facing researchers in many areas of inquiry is **how to *organize* observed data** into meaningful structures, that is, to develop taxonomies. Cluster Analysis is a means to achieve this.

Cluster Analysis

Cluster analysis is an exploratory data analysis tool which aims to sort different objects into groups in a way that:

- the degree of association between two objects is **maximal** if they belong to the **same group** and
- **minimal** if they belong to **different groups**

Examples of Clustering Applications

Marketing: help marketers discover distinct groups in their customer and then use this knowledge to develop targeted marketing programs.

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

Types of Cluster Analysis

- Two main types of cluster analysis are:

✓ **Hierarchical** cluster analysis (e.g. single linkage, group linkage, average linkage, Ward's method)

✓ **Non hierarchical** cluster analysis (e.g. *k*-means)

✓ Other advanced methods (not covered here)

Hierarchical Clustering

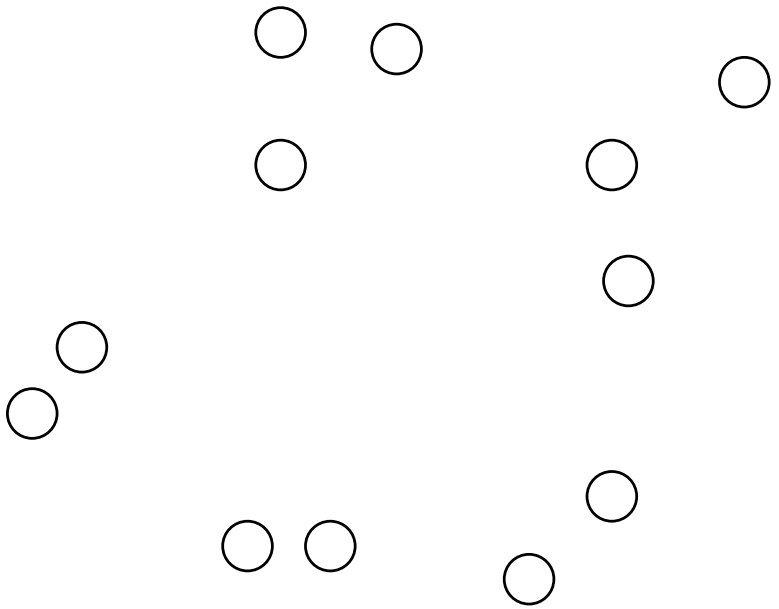
- Two main types of hierarchical clustering
  - Agglomerative:
    - Start with the points as individual clusters
    - At each step, merge the closest pair of clusters until only one cluster left
  - Divisive:
    - Start with one, all-inclusive cluster
    - At each step, split a cluster until each cluster contains a point
- Traditional hierarchical algorithms use a similarity or distance matrix for deciding which clusters to merge/split

Agglomerative clustering algorithm

- Most popular hierarchical clustering technique
- Basic algorithm
  1. Compute the distance matrix between the data points
  2. Let each data point be a cluster
  3. **Repeat**
  4. Merge the two closest clusters
  5. Update the distance matrix
  6. **Until** only a single cluster remains
- Key operation is the computation of the distance between two clusters
  - Different definitions of the distance between clusters lead to different algorithms

Input/ Initial setting

Start with clusters of individual points and a distance/proximity matrix



samples	A	B	C	D	E	F	G
A	0	0.5000	0.4286	1.0000	0.2500	0.6250	0.3750
B	0.5000	0	0.7143	0.8333	0.6667	0.2000	0.7778
C	0.4286	0.7143	0	1.0000	0.4286	0.6667	0.3333
D	1.0000	0.8333	1.0000	0	1.0000	0.8000	0.8571
E	0.2500	0.6667	0.4286	1.0000	0	0.7778	0.3750
F	0.6250	0.2000	0.6667	0.8000	0.7778	0	0.7500
G	0.3750	0.7778	0.3333	0.8571	0.3750	0.7500	0

Distance/Proximity Matrix

$$d(x_{i.}, x_{j.}) = \sqrt{\sum_{k=1}^{p} (x_{i\,k} - x_{j\,k})^2}$$

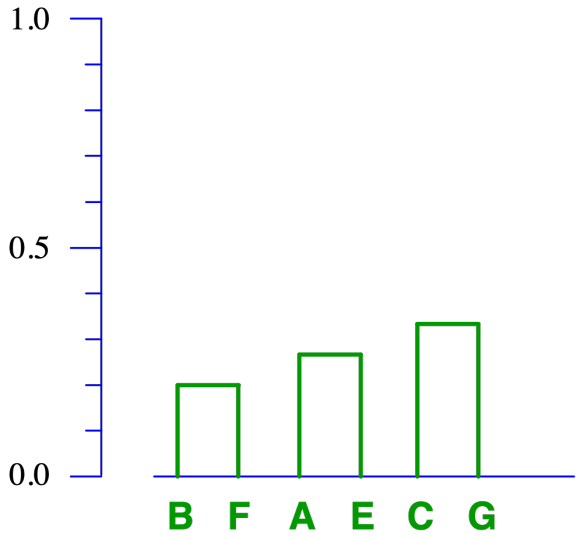
Euclidean distance

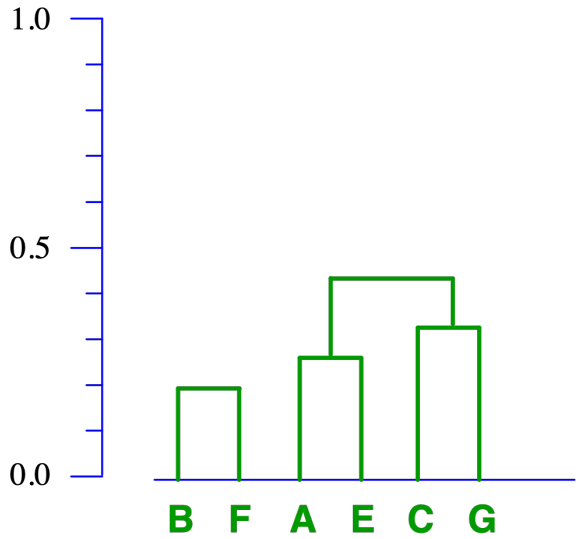
Intermediate State

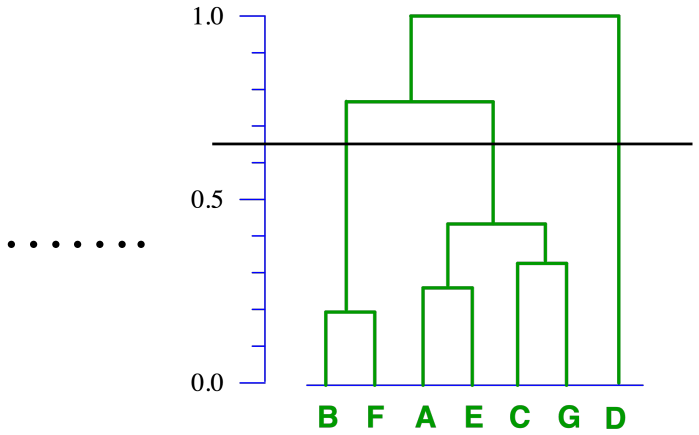
samples	A	(B,F)	C	D	E	G
A	0	0.6250	0.4286	1.0000	0.2500	0.3750
(B,F)	0.6250	0	0.7143	0.8333	0.7778	0.7778
C	0.4286	0.7143	0	1.0000	0.4286	0.3333
D	1.0000	0.8333	1.0000	0	1.0000	0.8571
E	0.2500	0.7778	0.4286	1.0000	0	0.3750
G	0.3750	0.7778	0.3333	0.8571	0.3750	0

As a next step we keep repeating this step, but the only problem is how to calculated the dissimilarity between the merged pair (B,F) and the other units. There are several choices, one of the most popular ones is the *maximum*, or *complete linkage*, method: the dissimilarity between the merged pair and the others will be the maximum of the pair of dissimilarities in each case. For example, the dissimilarity between B and A is 0.500, while the dissimilarity between F and A is 0.6250 hence we choose the maximum of the two, 0.6250, to quantify the dissimilarity between (B,F) and A

Determining the number of clusters: Graphically: the dendrogram

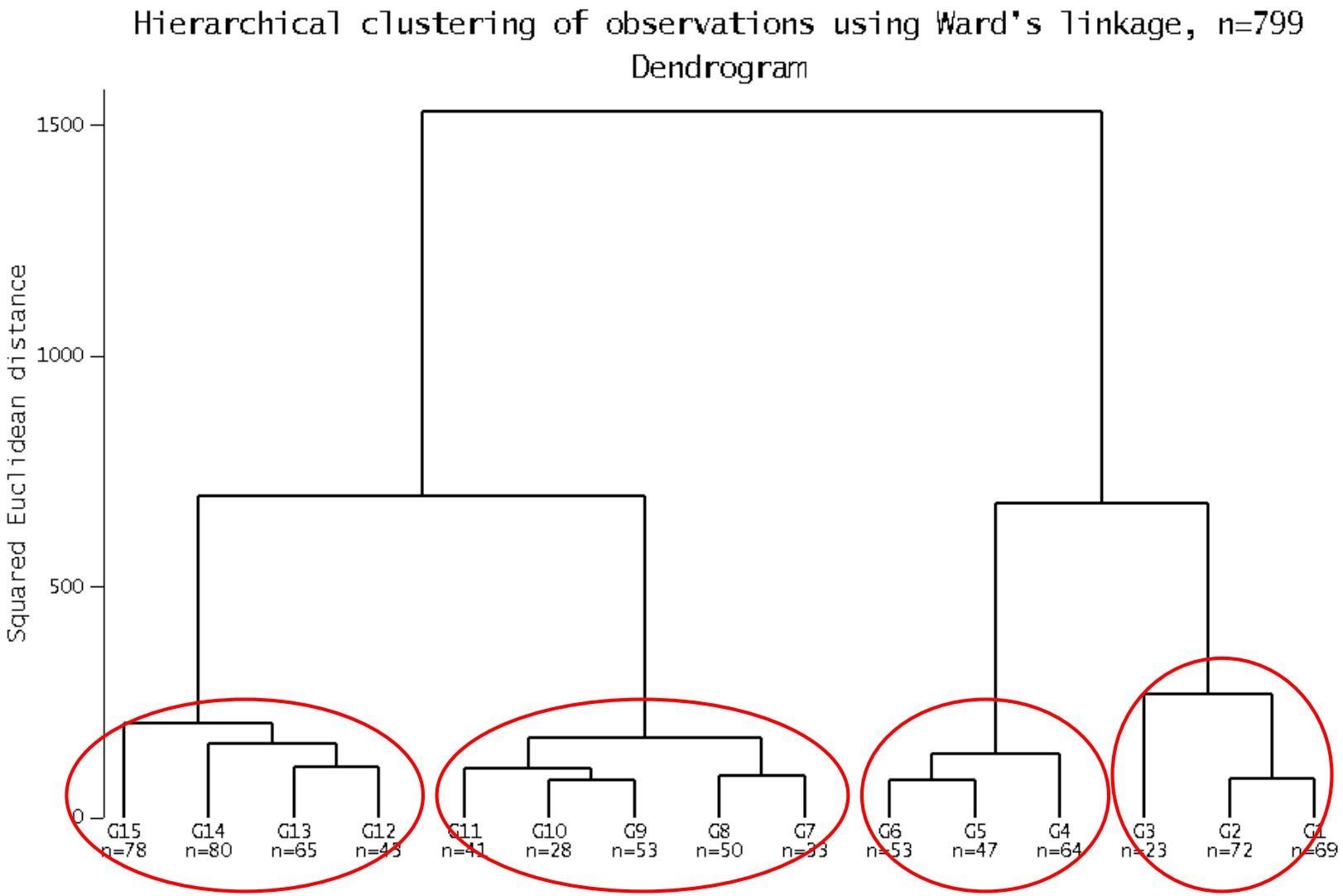






Final step: how many clusters?
It is based on where whe decide to “cut” the dendrogram.
In this example: 3 clusters

Determining the number of clusters: dendrogram



Determining the number of clusters: stopping rules

- ✓ Two common indices: Calinski-Harabasz (Pseudo F) and Duda-Hart.

The Calinski-Harabasz index is given by the ratio of the between-cluster variance and within-cluster variance. Each corrected by the number of cluster at any step of the hierarchical clustering.

Large values of Pseudo F indicate close-knit and separated clusters. In particular, peaks in the pseudo F statistic are indicators of greater cluster separation

Hierarchical Clustering Techniques

- **Distance** (or similarity measures) are often used to determine whether two observations are from a particular cluster.

- The most **common measures** that are used are:

1. The *minimum* distance between observations in two clusters (single linkage)
2. The *maximum* distance between observations in two clusters (complete linkage)
3. The *average* distance between observations in two clusters (group average)

Hierarchical Clustering techniques

- **Ward's method**: an alternative approach, wards method, does not use distance measure to determine clusters. Instead, clusters are formed with this method by maximizing the within cluster homogeneity.

Hierarchical Clustering Techniques

- **Pro :** no need to decide how many clusters you need
- **Con:** imposes a structure on the data that may be inappropriate.
- **Con:** once a case has been allocated to a particular cluster it cannot then be re-allocated to another cluster. **However,** these techniques may still be useful in an exploratory sense to get some idea how many clusters there are before doing a non hierarchical analysis (which does require the required number of clusters to be specified).

Non hierarchical clustering methods

Non-hierarchical methods may also be used. The most common one is the **'*k* means cluster analysis'**

Non hierarchical clustering methods

Non-hierarchical clustering techniques basically follow these steps (see Sharma 1997):

- Select *k* initial cluster centroids, where *k* is the desired number of clusters
- Assign each case to the cluster that is the closest to it
- Re-assign each case to one of the *k* clusters according to a pre-determined stopping rule

Not hierarchical Clustering

K-means clustering:

- based on general multivariate data
- yields K mutually exclusive and exhaustive clusters
- method alternates defining K cluster centroids and assigning each object to (possibly new) cluster to whose centroid it is closest
- solution usually depends on particular initial set of "seed" clusters (and/or centroids) with which algorithm begins
  - most solutions are suboptimal (obtaining local but not necessarily global optimum for loss function being optimized), so serious "local optimum" problem exists
  - necessary to start from more than one (usually several) starting points to guarantee globally optimal K-means solution

Phases of the clustering procedure

1. Choose the variables
2. Choose the measure of distance
3. Choose the clustering algorithm
4. Interpretate the results

Data reduction methods (PCA and Factor)

In many disciplines we study phenomena or constructs that cannot be directly measured

(self-esteem, personality, intelligence)

It often is required to take multiple observations for each case, and in the end we may have more data than can be readily interpreted

Items are representations of underlying or latent factors.

- We want to know what these factors are

We have an idea of the phenomena that a set of items represent (construct validity).

Because of this, we want to "reduce" them to a smaller set of factors

Data reduction methods

- Useful to deal with large data-sets *(many variables)*.

- The idea behind these methods is to avoid *double counting* of the same information by distinguishing between the individual information content of each variable and the amount of information shared across several variables.

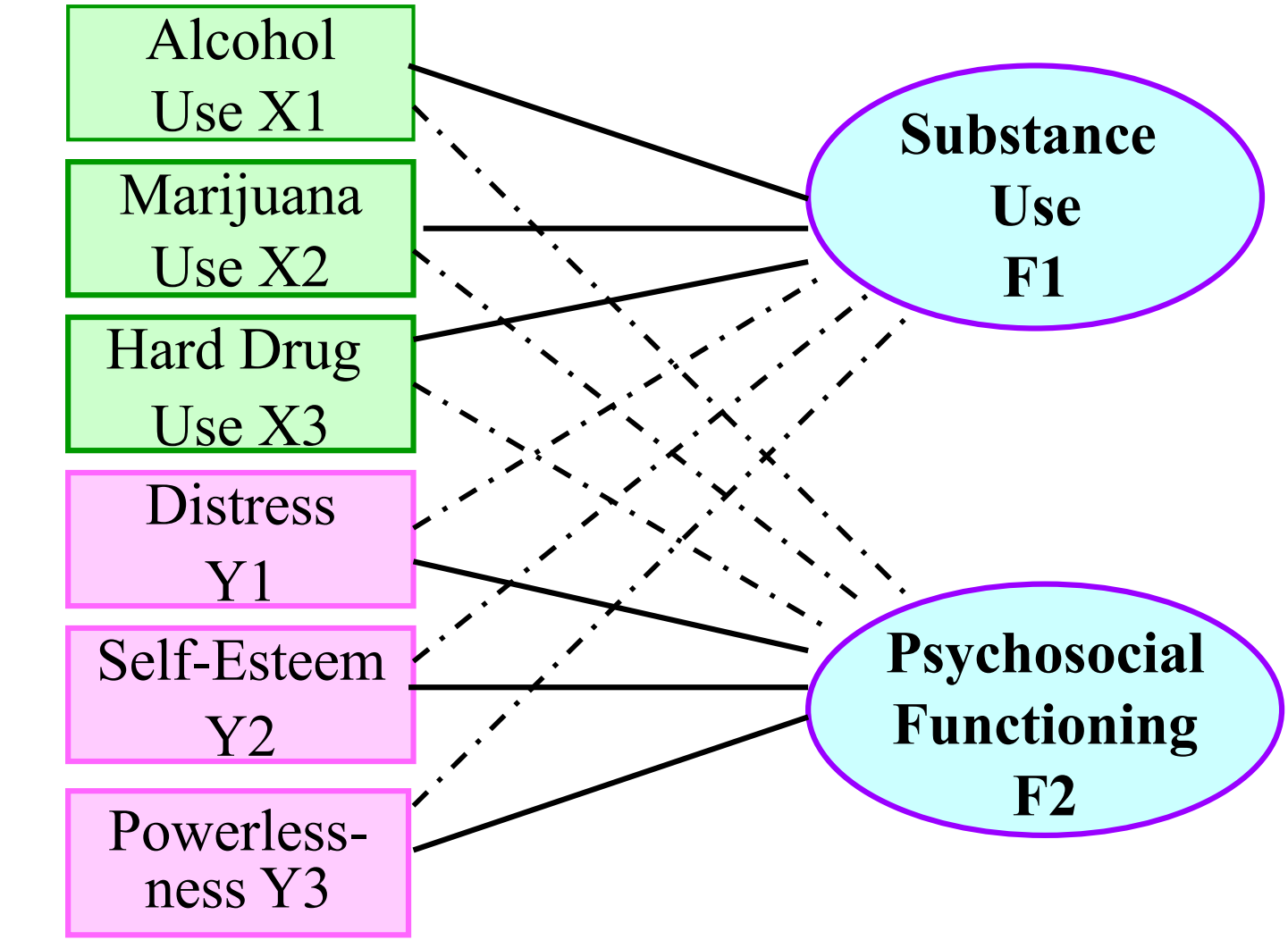
- A data-set containing *m* correlated variables is condensed into a reduced set of **$p \leq m$ new variables** (respectively factors or principal components according to the method used) that are ***uncorrelated* between each other and *explain a large proportion of the original variability.***

Factor and Principal Component Analysis

These are multivariate statistical methods designed to:

- exploit the correlations between the original variables
- create a smaller set of new artificial (or latent) variables that can be expressed as a combination of the original ones

- The higher the correlation across the original variables, the smaller the number of artificial variables (factors or components) needed to adequately describe the phenomenon



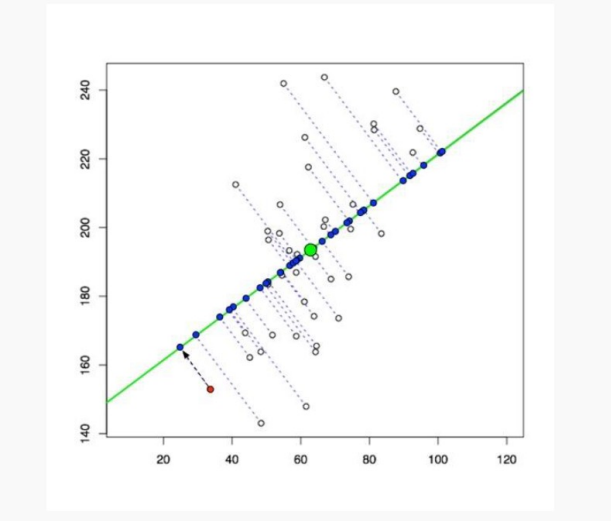
Example of PCA-FA

PCA- Principal Component Analysis

process of forming a new set of variable PCs as linear combinations of the original ones that captures the maximum amount of variance in the observed data.

- $PC_1 = b_{11}X_1 + b_{21}X_2 + \ldots + b_{k1}X_k$
- $PC_2 = b_{12}X_1 + b_{22}X_2 + \ldots + b_{k2}X_k$
- .....
- $PC_f = b_{1f}X_1 + b_{2f}X_2 + \ldots + b_{kf}X_k$

From a geometric point of view, this is equivalent to determining a one-dimensional subspace (a line in the plane) onto which the points are projected in such a way as to represent, with the highest possible resolution and within certain constraints, the differences that exist among the units. Orthogonal projection of the database onto a space defined by the top f PCs



Factor Analysis

The aim of factor analysis is to explain the interdependence that exists within a large set of variables through a small number of underlying, unobservable factors that are uncorrelated with one another.

In this sense, factor analysis goes beyond principal component analysis because, rather than merely producing a synthetic transformation of the observed variables, it involves estimating a model that reproduces their covariance structure.

Factor Analysis

Exploratory FA

- Summarizing data by grouping correlated variables
- is a data-driven approach which is generally used as an investigative technique to identify relationships among variables.
- does not have independent and dependent variables, but is an interdependence technique in which all variables are considered simultaneously

Confirmatory FA (theory driven)

- More advanced technique
- When factor structure is known or at least theorized
- Testing generalization of factor structure to new data, through SEM (structural equation modeling) methods. Most often used to determine the extent to which an already established theory about relationships among variables is supported by empirical data.

How does it work?

Correlation Matrix

- Analyses the pattern of correlations between variables in the correlation matrix
- Which variables tend to correlate highly together?
- If variables are highly correlated, likely that they represent the same underlying dimension

Example: Frailty

- We have a concept of what "frailty" is, but we can't measure it directly.
- We think it combines strength, weight, speed, agility, balance, and perhaps other dimensions
- We would like to be able to describe the components of frailty with a summary of strength, weight, etc.

Frailty Variables

Speed of fast walk (+)

Speed of usual walk (+)

Time to do chair stands (-)

Arm circumference (+)

Body mass index (+)

Skinfold thickness (+)

Shoulder rotation (+)

Upper extremity strength (+)

Pinch strength (+)

Grip strength (+)

Knee extension (+)

Hip extension (+)

Time to do Pegboard test (-)

EFA Research Design

- Factor analysis is performed most often only on **metric variables**
  - Specialized methods exist for the use of dummy variables, but a small number of "dummy variables" can be included in a set of metric variables that are factor analyzed.
  - If a study is being designed to reveal factor structure, strive to have at least 3/5 variables for each proposed factor.

- For **sample size**:
  - the sample must have more observations than variables.

Key Concepts

In factor analysis, each variable is expressed as a linear function of a certain number m of common factors—responsible for its correlation with the other variables—and of a single specific factor, which accounts for the variability unique to that variable itself.

F is the latent (i.e.unobserved, underlying) variable.

X's are observed (i.e. manifest) variables.

Three things influence observed variables. Two are types of factors: **common factors**, which give rise to more than one of the observed variables; **specific factors**, which give rise to only one of the observed variables.

The third thing that influences observed variables is **measurement error and basically anything unexplained by common or specific variance (randomness)**.

Key Concepts

Each of the elements that influence observed variables also contribute to those variables' variance.

The variance of a given observed variable is due in part to factors that influence other observed variables, factors that influence only the given observed variable, and measurement error.

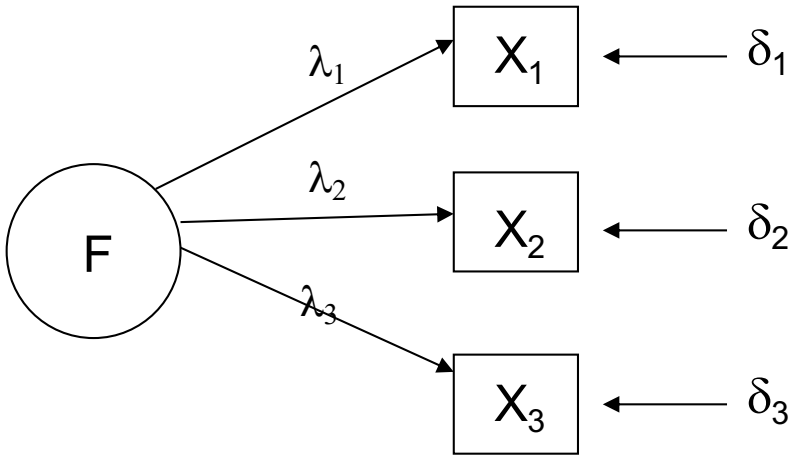
Common variance is sometimes referred to as "communality," amount of variance that is shared among a set of items. Items that are highly correlated will share a lot of variance. It ranges between 0 and 1. Values closer to 1 suggest that extracted factors explain more of the variance of an individual item.

the specific variance (due to the factor that influence that item only) and the error variance are often combined and referred to as "uniqueness."

Key Concepts

- F is latent (i.e.unobserved, underlying) variable
- X's are observed (i.e. manifest) variables
- λ_j is the "loading" for X_j.
- δ_j variability in the X_j NOT explained by F

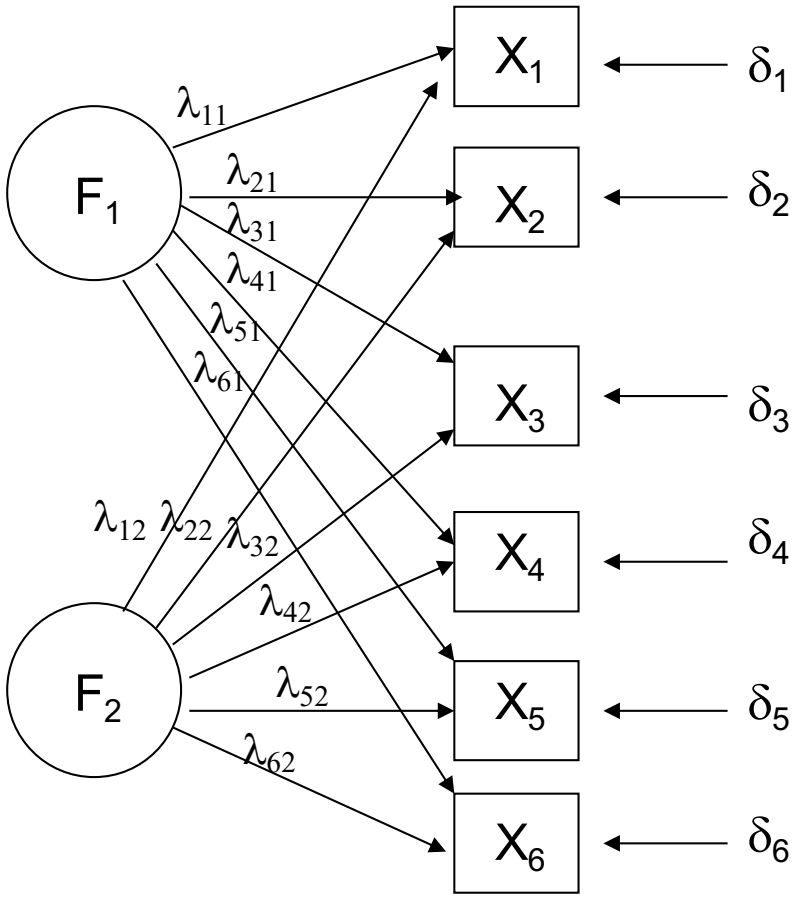
One Common Factor Model: Model Specification



- The factor F is <u>not observed</u>; only X_1, X_2, X_3 are observed
- δ_i represent variability in the X_i NOT explained by F
- X_i is a <u>linear</u> function of F and δ_i

λ_j is the "loading" for X_j. Relationship between the factor and the variable. "Saturazione" in Italian

Two Factor Model



$$X_1=\lambda_{11}F_1+\lambda_{12}F_2+\delta_1$$
$$X_2=\lambda_{21}F_1+\lambda_{22}F_2+\delta_2$$

Factor loadings: λ_{ij}

$\lambda_{ij} = corr(X_i,F_j)$

Communality of X_i: h_i^2

$h_i^2 = \lambda_{i1}^2 + \lambda_{i2}^2$=% variance of X_i explained by F_1 AND F_2

Uniqueness: $1- h_i^2$

Each variable is expressed as a linear function of a m common factors, which account for its correlation with the other variables, and a single specific factor, which accounts for the variability unique to the variable itself.

$$X_j = \mu_j + \lambda_{j1}F_1 + \cdots + \lambda_{jm}F_m + U_j$$

X is a k-dimensional random variable with mean vector μ and variance–covariance matrix **Σ**:

X = μ + **Λ**F + U

– **Λ**: k × m matrix of factor weights (factor loadings)
– F: m-dimensional random vector of common factors
– U: k-dimensional vector of specific factors

(i) E(F) = 0
(ii) Cov(F) = E(FF') = I_m
(iii) E(U) = 0
(iv) Cov(U) = E(UU') = Ψ= diag(ψ_1, …, ψ_k)
(v) E(FU') = 0

The common factors have zero mean, unit variance, and are uncorrelated with one another, the specific factors have zero mean, variance **ψ**j with j=1,…,k, and are uncorrelated with each other and with the common factors.

Usually we assume that variables are centered ($\mu=0$), the model is then:

$$X = \mathbf{\Lambda} F + U$$

$$X_j = \sum_{h=1}^{m} \lambda_{jh} F_h + U_j$$

Since U and F are not correlated:

$$Var(X_j) = \sum_{h=1}^{m} \lambda_{jh}^2 + \psi_i$$

$c_i = \mathbf{\Sigma}_{h=1} \mathbf{\lambda}_{jh}^2$ is the communality of the j-th variable; it corresponds to the portion of the variance of X_j explained by the common factors.

– $\mathbf{\psi}_j$ is the specificity of X_j, the residual part of the variance of X_j not explained by the common factors (specific variance or uniqueness).

If the factor model is valid:

- The covariance matrix of Xi is the sum of two matrices with **communalities** and **specificities** on the diagonal.
- Covariances between Xi and Xj are fully explained by common factors.
- Covariances between factors and variables equal the factor loadings.

FA is **invariant to changes in the scale** of observations.

$$\mathbf{\Sigma} = E(XX') = E[(\mathbf{\Lambda}F + U)(\mathbf{\Lambda}F + U)'] = \ldots = \mathbf{\Lambda\Lambda}' + \mathbf{\Psi}$$

$$Cov(X,F) = E(XF') = E[(\mathbf{\Lambda}F + U)F'] = \mathbf{\Lambda}$$

The matrix $\mathbf{\Lambda}$ of factor loadings is not identifiable: there is no unique solution for determining the factor loadings.
If an orthogonal rotation of the factors is performed using the orthonormal matrix Q of order m, we obtain:

$$X = \mathbf{\Lambda} Q Q' F + U$$

- $QQ' = I_m$
- $Q'F \rightarrow$ orthogonally rotated factors
- $\mathbf{\Lambda} Q \rightarrow$ factor loadings of the rotated factors

We there therefore impose constrains to the rotation.

Methods for Estimating Model (extracting the factor)

- Principal Factor Method
- Maximum Likelihood (ml)

The methods differ mainly from the way the correlation matrix is analyzed.

✸ ml: the data are multivariate normal distributed and we can calculate the loglikelihood. It is an iterative procedure which might have problems of convergence

Model estimation – principal factors

- A non-parametric estimation method that does not require distributional assumptions on the k-dimensional variable X.

- Unknown parameters: **Λ**, F, **Ψ**, and UU.

- Iterative procedure: the initial estimates of the communalities c_j are calculated (using the square of the multiple correlation coefficient of the regression of each item on the others). The estimated communalities are then substituted for the elements on the diagonal of R.

Estimation of factor scores

- **Factor scores:** values taken by the common factors corresponding to the sample observations.

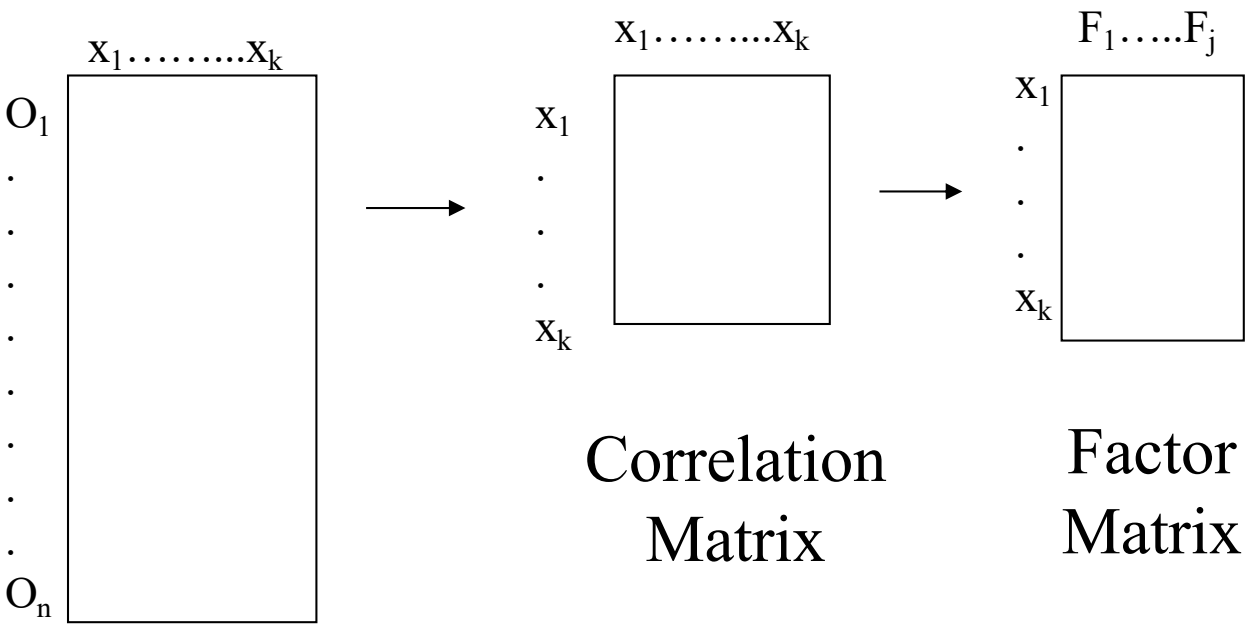
Bartlett's estimator:

Normality of X is assumed, use ML estimation methods.
This is the **unbiased estimator of the factor scores**.

Thompson's estimator – Bayesian method

As in all Bayesian approaches, inferences on the parameters (considered as random variables) are conducted **conditionally on the observed data**.
By assuming a normal prior distribution for Fi, it is possible to calculate the **posterior distribution** Fi|Xi
This estimator is **more accurate than Bartlett's estimator**.

Data Matrix


x1……...xk
O1
On
x1……...xk
x1
xk
F1…..Fj
x1
xk
Data Matrix
Correlation Matrix
Factor Matrix

Factor matrix:

1. Columns represent derived factors
2. Rows represent input variables
3. Loadings represent degree to which each of the variables "correlates"with each of the factors
4. Loadings range from -1 to 1
5. Inspection of factor loadings reveals extent to which each of the variables contributes to the meaning of each of the factors.
6. High loadings provide meaning and interpretation of factors (~ regression coefficients)

Frailty Example

We believe there is a latent construct –frailty- that defines the interrelationship among items. We acknowledge however that frailty cannot explain all the variance among items in the frailty questionnaire, so we model the unique variance as well.

Variable	F1:size	F2: speed	F3: Hand strength	F4: Leg strength
arm_circ	**0.97**	-0.01	0.16	0.01
skinfld	**0.71**	0.10	0.09	0.26
fastwalk	-0.01	**0.94**	0.08	0.12
gripstr	0.19	0.10	**0.93**	0.10
pinchstr	0.26	0.09	**0.57**	0.19
upextstr	0.08	0.25	0.27	0.14
kneeext	0.13	0.26	0.16	**0.72**
hipext	0.09	0.09	0.14	**0.68**
shldrrot	0.01	0.22	0.14	0.26
pegbrd	-0.07	-0.33	-0.22	-0.06
bmi	**0.89**	-0.09	0.09	0.04
uslwalk	-0.03	**0.92**	0.07	0.07
chrstand	0.02	-0.43	-0.07	-0.18

The factor loading scores show the correlation between factors and individual variables on a scale from -1 to 1. A negative value indicates an inverse impact on the factor.

Communality for arm_circ= 0.97^2+(-0.01^2)+0.16^2+0.01^2= 0.96. 96% of the variation in arm_circ is explained by the factor model. Sum of all com= the total common variance shared among all items for the 4 factor solution.

Total var explained by F1 (sum of squared loadings): (0.97^2)+(0.71^2)+…+(0.02^)=0.29

In principal components, each communality represents the total variance across all 13 items!.

Unique Solution?

- The factor analysis solution is NOT unique!
- More than one solution will yield the same “result.”

Rotation

The factor loadings could be plot in a scatterplot, with each variable represented as a point.

The axis of this plot could be rotated in any direction without changing the relative locations of the points to each other; however, the actual coordinates of the points, that is, the factor loadings would change.

Sometimes such rotations allow a clearer view of the factors

Rotational Strategies

The goal of all of rotational strategies is to obtain a clear pattern of loadings, that is, factors that are somehow clearly marked by high loadings for some variables and low loadings for others.

"rotate" factors:

- redefine factors such that loadings on various factors tend to be very high (-1 or 1) or very low (0)
- intuitively, it makes sharper distinctions in the meanings of the factors

Typical rotational strategies are varimax (variance maximizing), quartimax, and equamax.

Variable Selection and Use with Other Techniques

- Three Elements in **variable selection**
  1. Variable specification – researcher must specifically designate variables to be analyzed.
  2. Factors are always produced – EFA always generates factors, researcher has the responsibility to evaluate the usefulness and validity of the factors.
  3. Factors require multiple variables – EFA must have at least two correlated variables to form a factor.
  4. Using Factor Analysis with **Other Multivariate Techniques**
  - Factors may identify concepts more useful than individual variables.
  - Factors help mitigate the impact of multicollinearity on the interpretation of correlated variables.

Criticisms of Factor Analysis

- Labels of factors can be arbitrary or lack scientific basis

- "Garbage in, garbage out"
  - really a criticism of input variables
  - factor analysis reorganizes input matrix
- Too many steps that could affect results
- Too complicated
- Correlation matrix is often poor measure of association of input variables.