Factor Analysis

Prof. GPanda, FNAE,FNASc,FIET(UK)
IIT Bhubaneswar

Outline

- Introduction
- Basic concepts
- Factor Analysis Model
- Statistics Associated with Factor Analysis
- Example

Introduction

Factor Analysis:

A data reduction technique designed to represent a wide range of attributes on a smaller number of dimensions.

The basic assumption of factor analysis is that for a collection of observed variables there are a set of *underlying* variables (unobserved) called **factors** (smaller than the observed variables), that can explain the interrelationships among those variables.

Introduction

Example:

Unlike variables directly measured such as speed, height, weight, etc., some variables such as creativity, happiness,love,hate, comfort, etc., are **not measurable entities and are also unobserved**.

They are **constructs** that are derived from the measurement of others like directly observable variables .

Introduction

- Factor is a construct that is not directly observed but that needs to be inferred from the input variables.
- The identification of such underlying dimensions (factors) simplifies the understanding and description of complex constructs.

Introduction

- Generally, the number of factors is much smaller than the number of measures.
- Therefore, the expectation is that a **factor** represents a **set of measures.**
- From this angle, factor analysis is viewed as a data-reduction technique as it reduces a large number of overlapping variables to a smaller set of factors that reflect construct(s)

Understanding Factor Analysis

- Observed correlations between variables result from their sharing of factors.
- Example: Correlations between a person’s test scores might be linked to shared factors such as general intelligence, critical thinking and reasoning skills, reading comprehension etc.

Understanding Factor Analysis

- A major goal of factor analysis is to represent relationships among sets of variables as simple as possible yet keeping factors meaningful(explain the correlations among a set of variables).

- A good factor solution is both simple and interpretable.

Understanding Factor Analysis

Finally it is,

A technique that analyses data on a relatively large set of variables and produces a smaller set of factors which are used to represent the **original variables as linear combinations of the smaller set of factors**; so that the set of factor captures as much information as possible from the original set of data

Types of Factor Analysis

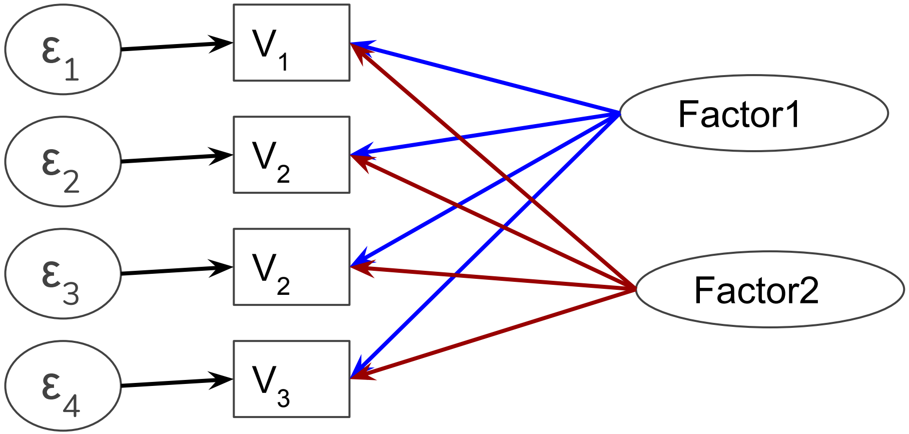
Exploratory Factor Analysis(EFA):

EFA is used when it is not known that how many factors are necessary to explain the interrelationships among a set of observed variables, to determine if those variables can be grouped into a smaller number of underlying factors.

It means that EFA is basically used to explore the underlying dimensions(factors) of the construct of interest.

Path diagram for Explorative FA

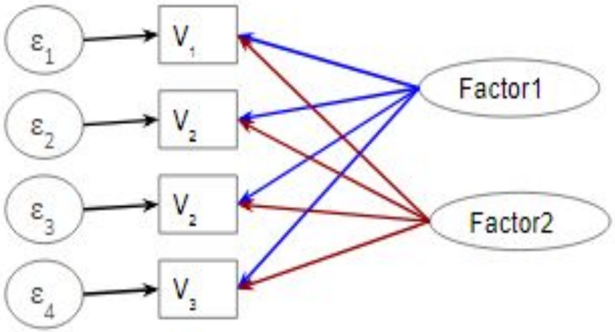
In this analysis, all items are assumed to be related to all factors.



Path diagram for Explorative FA

Measured Variables: v_1,v_2,v_3,v_4
These variables are those that the researcher has observed or measured.

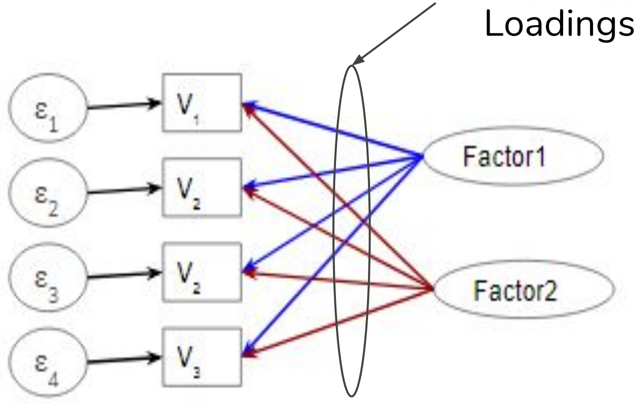
Unmeasured or Latents variables:
Factor1, Factor2
These variables are not directly measurable, rather the researcher only has indicators of these measures.



Path diagram for Explorative FA

Factor Loadings:

- Measure the relationship between the items and the factors.
- Factor loadings can be interpreted like correlation coefficients; ranging between -1.0 and +1.0.
- The closer the value is to 1.0, positive or negative, the stronger the relationship between the factor and the item.

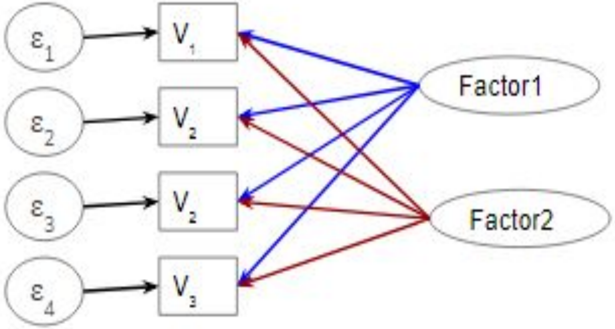


Factor Loadings

Path diagram for Explorative FA

Unique variance in measurement:

- $\varepsilon_1,\varepsilon_2,\varepsilon_3,\varepsilon_4,\varepsilon_5$
- Each of the indicator variables has some unique variance, apart from the common variance which is explained by the factors.



Types of Factor Analysis

Confirmatory Factor Analysis(CFA):

Researchers use CFA when they want to assess the extent to which the hypothesized organization of a set of identified factors fit the data.

It means that CFA us used when the researcher has some knowledge about the underlying structure of the construct under investigation.

Factor Analysis Model

Each variable is expressed as a linear combination of factors. The factors are some common factors plus a unique factor. The factor model is represented as:

$$X_i = \mu_i + \lambda_{i1}F_1 + \lambda_{i2}F_2 + \lambda_{i3}F_3 + \ldots + \lambda_{im}F_m + U_i$$

where

X_i = i^{th} observed variable and $\mu_i = E(X_i)$ {E is expectation}

λ_{ij} = Factor Loading of j^{th} factor on i^{th} example

F_j = common factor *j*

U_i = the unique variance for variable *i(can consider as error term)*

m = number of common factors

Factor Analysis Model

In Matrix form:

$$\underset{p\times1}{\boldsymbol{x}} = \underset{p\times1}{\boldsymbol{\mu}} + \underset{p\times m}{\boldsymbol{L}}\;\underset{m\times1}{\boldsymbol{F}} + \underset{p\times1}{\boldsymbol{\varepsilon}}$$

F and ε are independent

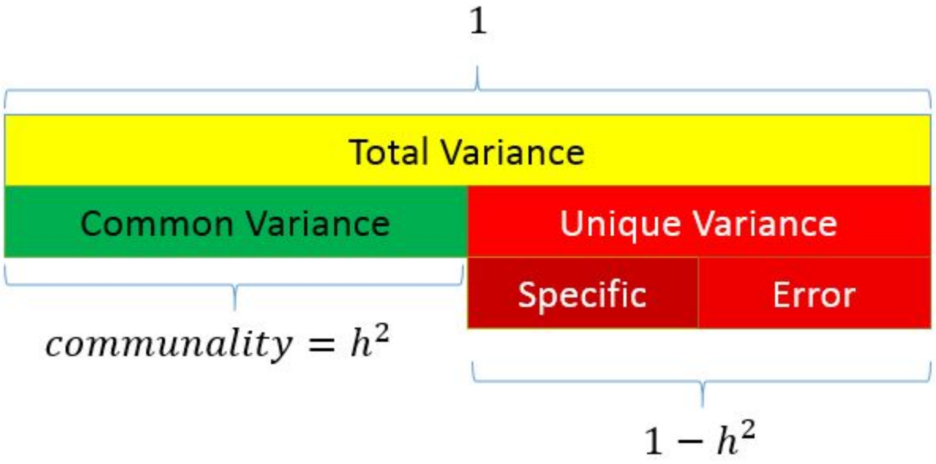
Where, p is the number of observed variables , m is number of factors, L represent Loadings, F represent factors and ε is unique variance.

$$E(\boldsymbol{F}) = 0, \mathrm{Cov}(\boldsymbol{F}) = \boldsymbol{I}$$

$$E(\boldsymbol{\varepsilon}) = 0, \mathrm{Cov}(\boldsymbol{\varepsilon}) = \boldsymbol{\psi}$$, Ψ is a diagonal matrix

Factor Analysis Model

Factor analysis assumes that variance can be partitioned into two types of variance, common and unique.



Factor Analysis Model

Common Variance: It is the amount of variance that is shared among a set of items. Items that are highly correlated will share a lot of variance.

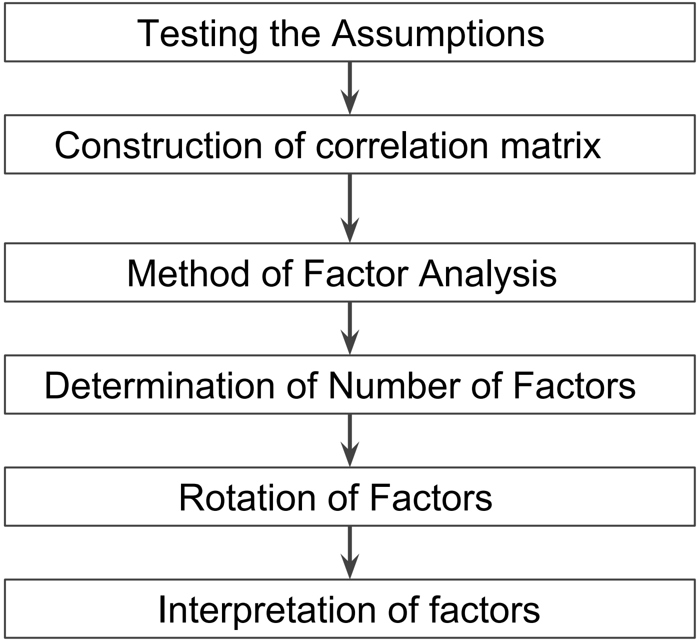
Communality: (also called **h^2**) is a definition of common variance that ranges between 0 and 1. Values closer to 1 suggest that extracted factors explain more of the variance of an individual item.

Factor Analysis Model

Unique variance(also called **1- h^2)** is any portion of variance that's not common. There are two types:

- **Specific variance**: is variance that is specific to a particular item.
- **Error variance:** comes from errors of measurement and basically anything unexplained by common or specific variance .

Conducting Factor Analysis



Testing the Assumptions

Construction of correlation matrix

Method of Factor Analysis

Determination of Number of Factors

Rotation of Factors

Interpretation of factors

Factor Analysis Assumptions

- Variables are correlated and some are having high degree of co-variance.
- The factor analysis model assumes that variables are determined by common factors and unique factors. All unique factors are assumed to be uncorrelated with each other and with the common factors.
- Observed variables can be expressed as some linear combinations of the underlying factors.
- Adequate sample size

Bartlett test of sphericity

- It test the null hypothesis that all the correlation between the variables is Zero, i.e., it tests whether the correlation matrix is a identity matrix or not.
- If it is an identity matrix then factor analysis becomes inappropriate.
- If the **value** of the test statistic for **sphericity is large** and the associated **significance level is small**, it is **unlikely** that the population correlation matrix is an identity.

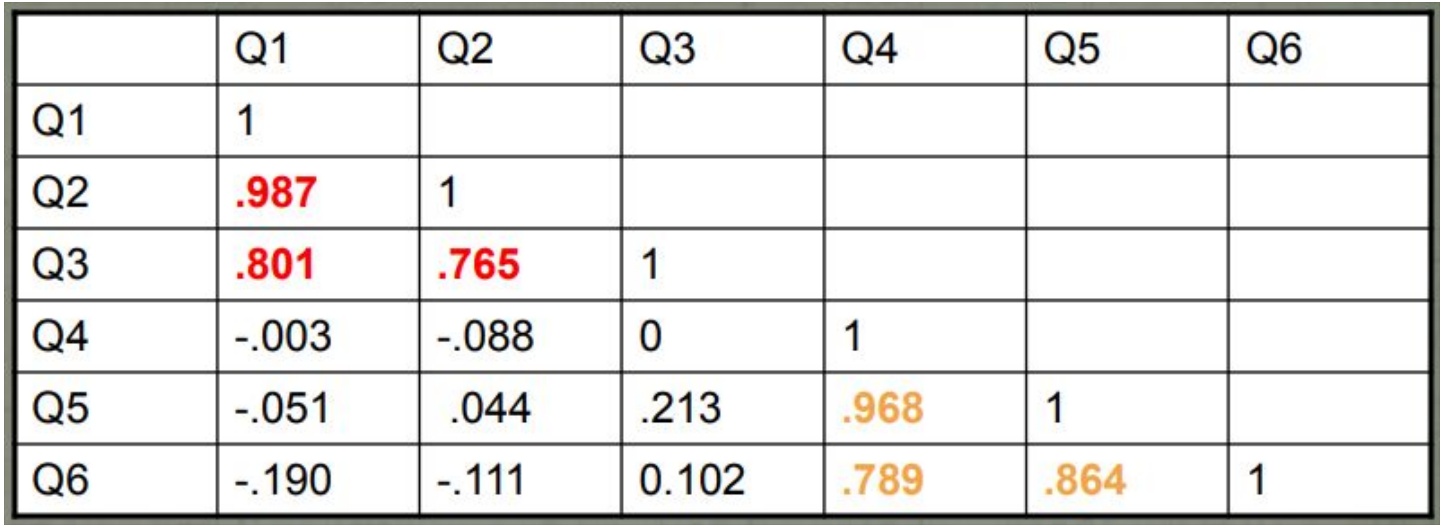
Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy

- This test checks the adequacy of data for running the factor analysis. The statistic is a measure of the proportion of variance among variables that might be common variance.
- The value of KMO ranges from 0 to 1.
- The larger the value of KMO more adequate is the sample for running the factor analysis.
- Kaiser recommends accepting values greater than 0.5 as acceptable.

Construction of 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.
- Correlation coefficients greater than 0.3 in absolute value are indicative of acceptable correlations.
- Factor analysis pinpoints the clusters of high correlations between variables and for each cluster, it will assign a factor

Correlation matrix



	Q1	Q2	Q3	Q4	Q5	Q6
Q1	1					
Q2	.987	1				
Q3	.801	.765	1			
Q4	-.003	-.088	0	1		
Q5	-.051	.044	.213	.968	1	
Q6	-.190	-.111	0.102	.789	.864	1

- Q1-3 correlate strongly with each other and hardly at all with 4-6
- Q4-6 correlate strongly with each other and hardly at all with 1-3

Method of Factor Analysis

Principal Components Method:

- Estimates of initial factors are obtained using Principal components analysis.
- It looks at the total variance among the variables as the common variance removing unique variance from the model..
- In this method, the factor explaining the maximum variance is extracted first.

Method of Estimation

To estimate factor loadings find the eigen values and corresponding eigenvectors of the correleation matrix ‘S’

Let $\lambda_1,\lambda_2,\lambda_3 \ldots\ldots\lambda_p$ be the eigen values of ‘p’ variables and

$e_1,e_2,e_3, \ldots\ldots\ldots e_p$ be the corresponding eigen vectors.

From principal component Analysis

Correlation S = $\sum_{j=1}^{p} \lambda_j e_i e_j^T$

Method of Estimation

$$S = \left[\sqrt{\lambda_1}\, e_1 \;\; \sqrt{\lambda_2}\, e_2 \;\; \ldots \;\; \sqrt{\lambda_p}\, e_p\right] \begin{bmatrix} \sqrt{\lambda_1}\, e_1 \\ \sqrt{\lambda_2}\, e_2 \\ \\ \\ \sqrt{\lambda_p}\, e_p \end{bmatrix}$$

$$= \left[\sqrt{\lambda_1}\, e_1 \;\; \sqrt{\lambda_2}\, e_2 \;\; \ldots \;\; \sqrt{\lambda_m}\, e_m\right] \begin{bmatrix} \sqrt{\lambda_1}\, e_1 \\ \sqrt{\lambda_2}\, e_2 \\ \\ \\ \sqrt{\lambda_m}\, e_m \end{bmatrix} + \left[\sqrt{\lambda_{m+1}}\, e_{m+1} \;\; \ldots \;\; \sqrt{\lambda_p}\, e_p\right] \begin{bmatrix} \sqrt{\lambda_{m+1}}\, e_{m+1} \\ \\ \\ \sqrt{\lambda_p}\, e_p \end{bmatrix}$$

$$= \underset{\Lambda\Lambda^T}{} + \Psi$$

Method of Estimation

Principal components method Approximating this with the model equation $\sum = L\;F + \varepsilon$

Therefore,

Factor Loadings L = Λ

$$= [\sqrt{\lambda_1}\, e_1 \;\; \sqrt{\lambda_2}\, e_2 \; \dots \; \sqrt{\lambda_m}\, e_m]$$

Communality(h_{ij}^{2}) = $\sum_{j=1}^{m} l_{ij}^{2}$ Where l_{ij} is factor loading of j^{th} factor on i^{th} variable.

Determination of number of factors

Eigenvalue:

The Eigenvalue for a given factor measures the variance in all the variables which is accounted for by that factor.

It is the amount of variance explained by a factor. It is also called as characteristic root

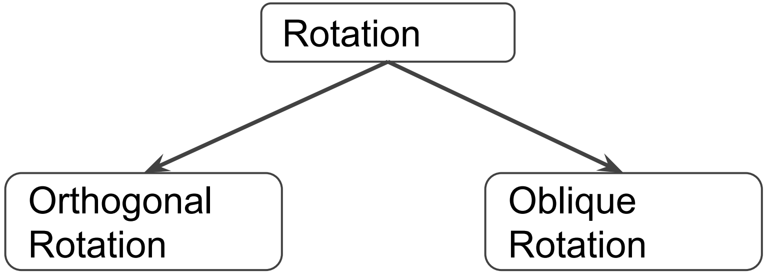
Kaiser Guttmann Criterion:

This method states that the number of factors to be extracted should be equal to the number of factors having an Eigenvalue of 1 or greater than 1.

Rotation of Factors

Maximizes high item loadings and minimizes low item loadings, thereby producing a more interpretable and simplified solution.

- Two common rotation techniques orthogonal rotation and oblique rotation



Rotation of Factors

Orthogonal rotation: Yields uncorrelated factors. Ex: Varimax

Varimax attempts to minimize the number of variables that have high loadings on a factor. This enhances the interpretability of the factors.

Oblique rotation: Yields correlated factors

Oblique rotations are less frequently used because their results are more difficult to summarize.

Examples: Quartimax ,Equamax,Promax

Interpretation of Factors

- The final decision about the number of factors to choose is the number of factors for the rotated solution that is most interpretable.
- To identify factors, group variables that have large loadings for the same factor.
- Plots of loadings provide a visual for variable clusters.
- Interpret factors according to the meaning of the variables

Example

Given correlation matrix of consumer-Preference data of food items.

Taste	1	0.02	0.96	0.42	0.01
Good buy for Money	0.02	1	0.13	0.71	0.85
Flavour	0.96	0.13	1	0.50	0.11
Suitable for Snack	0.42	0.71	0.5	1	0.79
Provides lots of energy	0.01	0.85	0.11	0.79	1

Example

Step -1: Need to estimate factor loadings.

Let S be the correlation matrix, Consider the eigen value equation

$SX = \lambda X$,

where, λ is eigen value corresponding to eigenvector X.

Need to solve $|S-\lambda\mathbf{I}| = 0$

Find the values of λ, then for each λ solve $(S-\lambda\mathbf{I}).X = 0$ to get eigenvector X

Example

Eigenvalues :

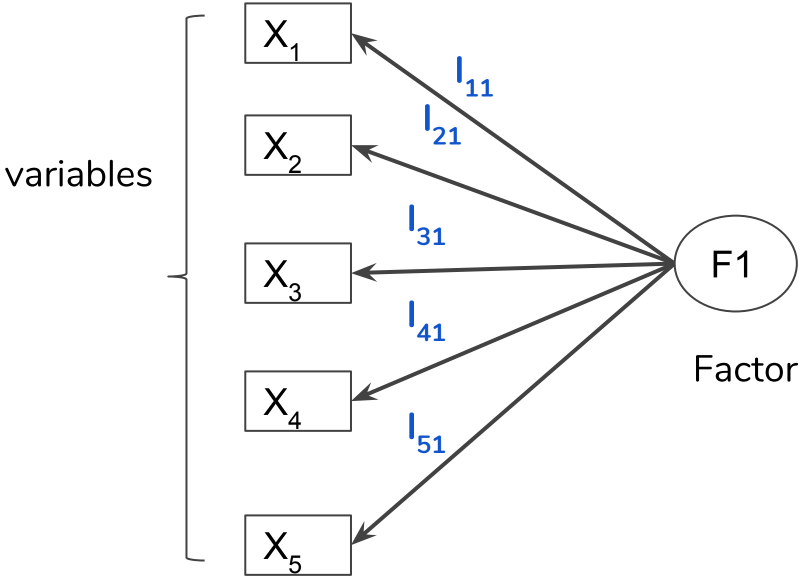
$\lambda_1 = 2.85,\ \lambda_2 = 1.81,\ \lambda_3 = 0.03,\ \lambda_4 = 0.20,\ \lambda_5 = 0.10$

Eigenvectors:

$$E_{1,} = \begin{pmatrix} 0.33 \\ 0.46 \\ 0.38 \\ 0.56 \\ 0.47 \end{pmatrix} \quad E_{2} = \begin{pmatrix} 0.61 \\ -0.39 \\ 0.56 \\ -0.08 \\ -0.40 \end{pmatrix} \quad E_{3,} = \begin{pmatrix} -0.10 \\ -0.07 \\ 0.71 \\ 0.00 \\ -0.01 \end{pmatrix} \quad E_{4,} = \begin{pmatrix} -0.70 \\ -0.74 \\ -0.17 \\ 0.60 \\ 0.22 \end{pmatrix} \quad E_{5,} = \begin{pmatrix} -0.14 \\ 0.28 \\ -0.12 \\ 0.57 \\ -0.75 \end{pmatrix}$$

Example

Estimate of Factor Loading:



Factor loading of jth factor on ith variable

$$l_{ij} = e_{ij}\sqrt{\lambda_j}$$

In vector form :

$$L_j = E_j\sqrt{\lambda_j}$$

Example

$$L_1 = E_1\sqrt{\lambda_1} = \begin{pmatrix} 0.33 \\ 0.46 \\ 0.38 \\ 0.56 \\ 0.47 \end{pmatrix} \times \sqrt{2.85} = \begin{pmatrix} 0.56 \\ 0.78 \\ 0.65 \\ 0.94 \\ 0.80 \end{pmatrix}$$

$$L_2 = E_2\sqrt{\lambda_2} = \begin{pmatrix} 0.61 \\ -0.39 \\ 0.56 \\ -0.08 \\ -0.40 \end{pmatrix} \times \sqrt{1.81} = \begin{pmatrix} 0.82 \\ -0.52 \\ 0.75 \\ -0.11 \\ -0.54 \end{pmatrix}$$

Example

Similarly,

$$L_3 = \begin{pmatrix} -0.13 \\ -0.01 \\ 0.13 \\ -0.00 \\ -0.00 \end{pmatrix} \quad L_4 = \begin{pmatrix} -0.04 \\ -0.34 \\ -0.08 \\ 0.27 \\ 0.10 \end{pmatrix} \quad L_5 = \begin{pmatrix} -0.04 \\ 0.09 \\ -0.04 \\ 0.18 \\ -0.24 \end{pmatrix}$$

Example

Step 2 : Factor Extraction(Factors should be less than variables).

Considering Factors with Eigen value ≥ 1

	Factor 1	Factor 2
Taste	0.56	0.82
Good buy for Money	0.78	-0.52
Flavour	0.65	0.75
Suitable for Snack	0.94	-0.11
Provides lots of energy	0.80	-0.54

Example

Communality(h^2) of the i^{th} variable $= \sum_{j=1}^{m} l_{ij}^{2}$, with m factors

	Factor 1	Factor 2	Communality($h^{2)}$
Taste	0.56	0.82	$(0.56)^2 + (0.82)^2 = 0.986$
Good buy for Money	0.78	-0.52	$(0.78)^2 + (-0.52)^2 = 0.878$
Flavour	0.65	0.75	$(0.65)^2 + (0.75)^2 = 0.985$
Suitable for Snack	0.94	-0.11	$(0.94)^2 + (-0.11)^2 = 0.896$
Provides lots of energy	0.80	-0.54	$(0.80)^2 + (-0.54)^2 = 0.932$

For variable ‘Taste’, Communality 0.986 implies 98.6% of variance of Taste is explained by two factors

Example

Step-3: Factor Rotation

Orthogonal rotation using Transformation matrix T,

$$T = \begin{pmatrix} Cos\theta & -Sin\theta \\ Sin\theta & Cos\theta \end{pmatrix} \quad => \quad T^TT = T\,T^T = T\,T^{-1} = \mathbf{I}$$

$$F_1^* = F_1\,Cos\theta - F_2Sin\theta$$

$$F_2^* = F_1Sin\theta + F_2Cos\theta$$

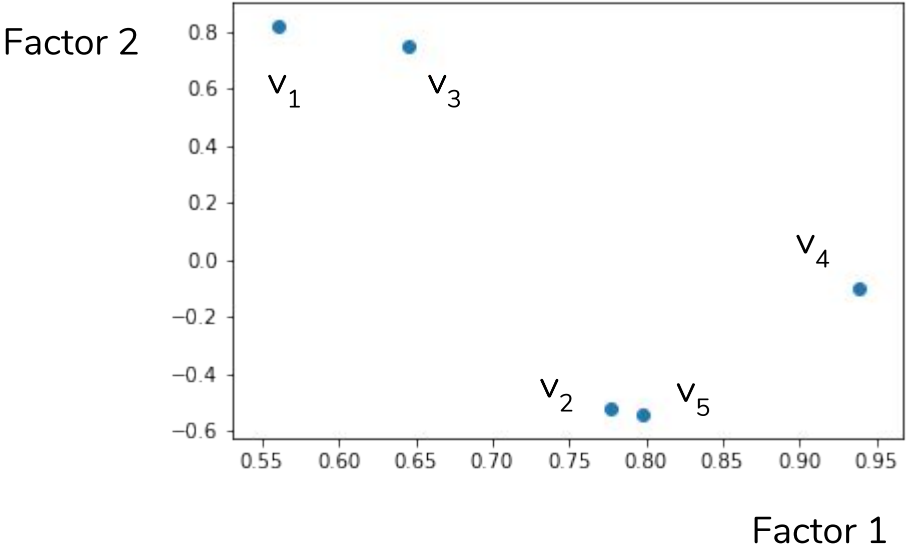
Example

Factors after orthogonal transformation with θ = 45 degrees

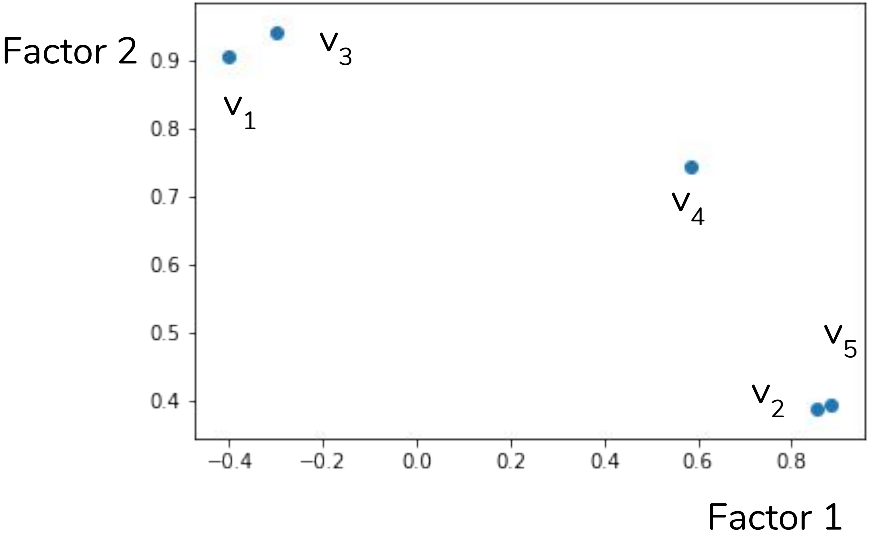
	Factor 1	Factor 2
Taste	-0.40	0.91
Good buy for Money	0.85	0.38
Flavour	-0.30	0.94
Suitable for Snack	0.58	0.74
Provides lots of energy	0.88	0.39

🟥 Factors contributing to majority of variance of the variables

Example



Before Rotation



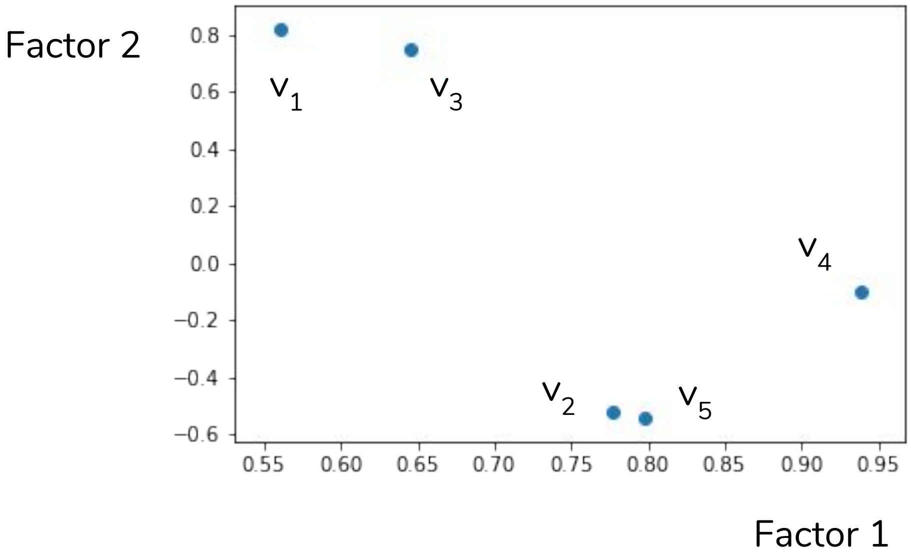
After Rotation

Example

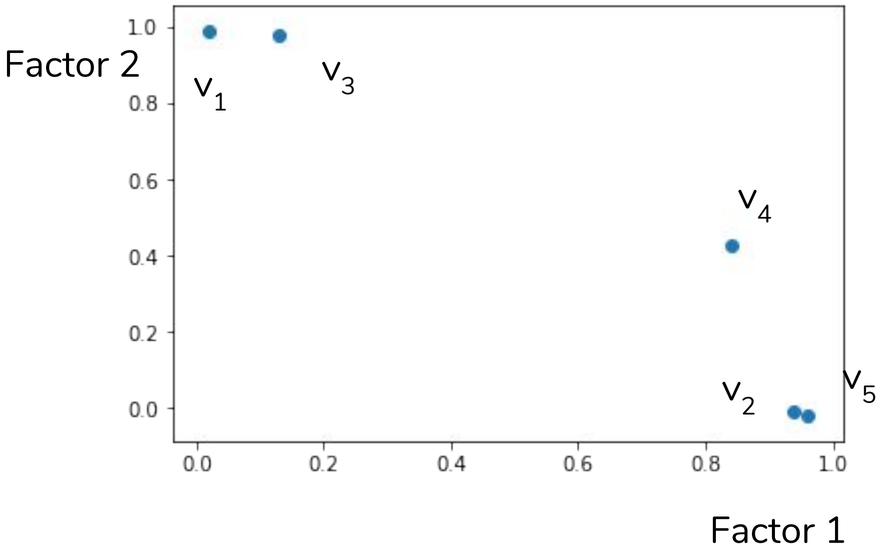
Applying **Varimax Rotation** on Factors

	Factor 1	Factor 2
Taste	0.02	0.99
Good buy for Money	0.94	-0.01
Flavour	0.13	0.98
Suitable for Snack	0.84	0.43
Provides lots of energy	0.96	-0.02

Example

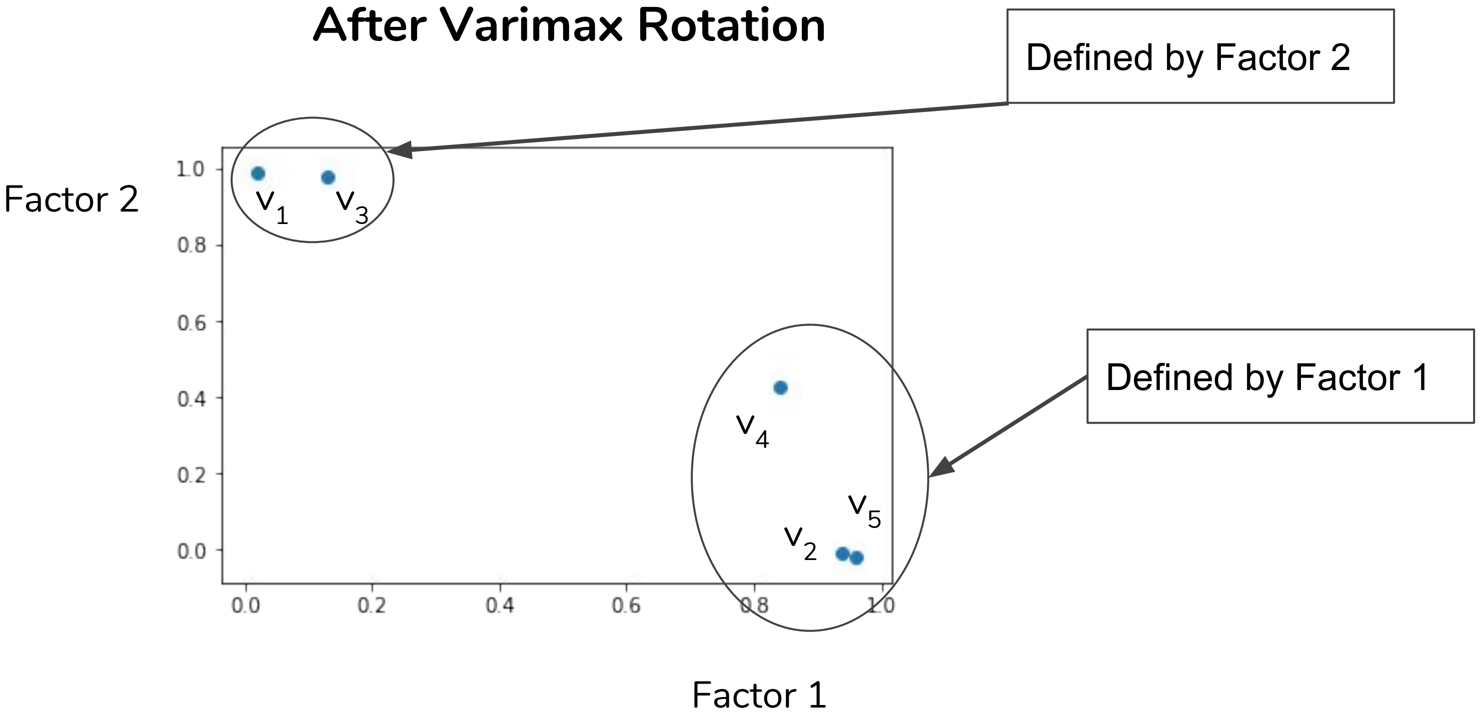


Before Rotation



After Varimax Rotation

Example



Defined by Factor 2

Defined by Factor 1

Example

Step-4: Interpretation

- Since we can categorize whole variable into two factors
- Factor1 explains variance of Good buy for Money, Suitable for Snack and Provides lots of energy.
- Factor2 explains variance of Taste and flavour.
- **We can name Factor1 as Assertion on food and Factor2 as Property of food**.
- Now instead of 5 we have only 2 variables.

Summary

- Makes analysis simpler.
- Applicable when there is significant covariance between variables.
- Uses Principal components methods(can also use others) to estimate factor loadings and does factor extraction.
- Rotation is done for good interpretation of factors.
- Factors can be used instead of Variables for further analysis(done by factor scores which can be obtained from Regression or Least Squares.)

References

- Applied Multivariate Statistical Analysis, 5th Edition, authors:Richard A.Johnson, Dean W. Wichern.
- An easy guide to factor analysis Book by Paul Kline.
- Exploratory Factor Analysis,Book by Duane T. Wegener and Leandre R Fabrigar

Thank You