Factor Analysis (PCR)

Extracting Hidden Factors in Multivariate Data

Contents

Factor Analysis

- Factor Analysis is primarily used for,

Data Reduction	or	Structure Detection
The purpose of data reduction is to replace original variables with a smaller number of uncorrelated variables		The purpose of structure detection is **to examine the underlying (or latent) relationships between the variables**

- It is an interdependence technique: **No distinction between dependent and independent variables** (Also called as Unsupervised Method)

Statistical 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:

$$X_i = l_{i1}F_1 + l_{i2}F_2 + l_{i3}F_3 + \cdots + l_{im}F_m + u_i + e_i$$

where,

l_{ij}	:	Loading of variable i on common factor j
F_j	:	Common factor j
u_i	:	Mean of variable i
m	:	Number of common factors
e_i	:	Specific factor

The model looks similar to linear regression but **factors are unobservable**

Statistical Model – Five Variables and Two Common Factors

- Each variable is expressed as a linear combination of two factors

The factor model is represented as:

$$X_1 - u_1 = l_{11}F_1 + l_{12}F_2 + e_1$$
$$X_2 - u_2 = l_{21}F_1 + l_{22}F_2 + e_2$$
$$X_3 - u_3 = l_{31}F_1 + l_{32}F_2 + e_3$$
$$X_4 - u_4 = l_{41}F_1 + l_{42}F_2 + e_4$$
$$X_5 - u_5 = l_{51}F_1 + l_{52}F_2 + e_5$$

- F_1 and F_2 are common factors
- e's are specific factors
- Each variable loads on two common factors

Statistical Model – Five Variables and Two Common Factors

- The common factors themselves can be expressed as linear combinations of the observed variables.

$$F_i = W_{i1}X_1 + W_{i2}X_2 + W_{i3}X_3 + \ldots + W_{ik}X_k$$

where,

F_i : Estimated score of i^{th} factor

W_i : Weight or factor score coefficient

k : Number of variables

Common Factor and Unique Factor

Common Factor

- It is an abstraction, a **hypothetical construct that affects some or all observed variables**
- We want to estimate the common factors that contribute to the variance in our variables
- Example: Athletics performance can be thought of as combination of 'speed' and ‘stamina'

Unique Factor

- It is a **factor that contributes to the variance of only one variable**
- We want to exclude these unique factors from our solution
- The unique factors are unrelated to one another and unrelated to the common factors

Factor Analysis – Assumptions

The unobservable random variables F_j and e_i are assumed to satisfy following conditions:

1. F_j and e_i are independent

1. $E(F_j)=0$, $Cov(F_j, F_k)=0$ for $j \neq k$ and $V(F_j)=1$

1. $E(e_i)=0$ and $Cov(e_j, e_k)=0$ for $j \neq k$

When these assumptions are satisfied, the model is called Orthogonal Factor Model

Estimation of Factor Loadings

Typically 2 methods are used to estimate the factor loadings:

1. Principal Component Method

2. Maximum Likelihood Method

   (Assuming common Factors and specific factors follow Normal distributions)

- The number of factors to be included in the model is determined by considering the proportion of variance explained by the *m* factors model
- The portion of the variance of i^{th} variable contributed by m common factors is called as "**Communality**"
- The portion due to specific factor is called as "**Uniqueness**" or "**Specific Variance**"

Interpretation of Factor Loadings

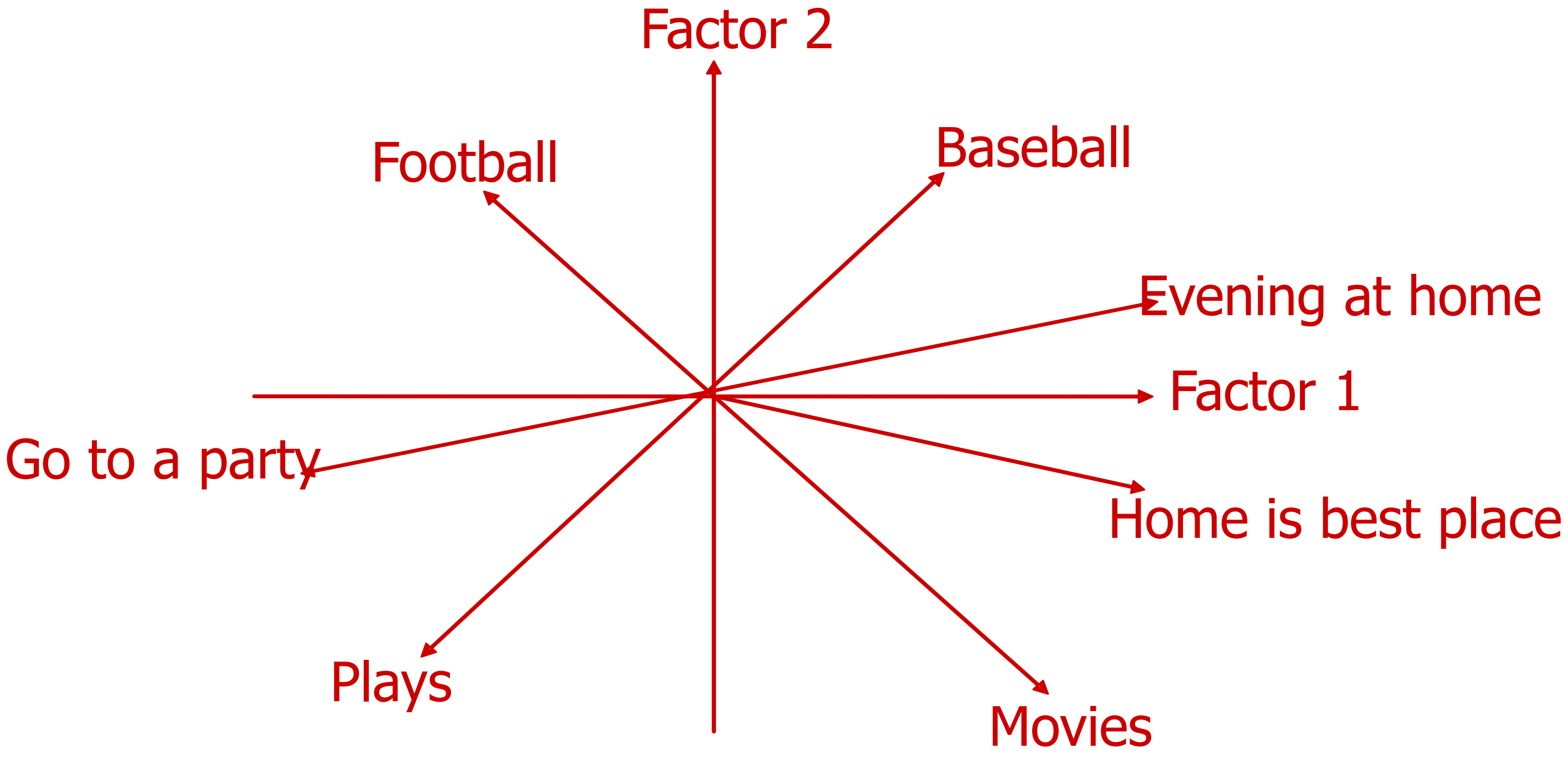
A factor can be interpreted in terms of the variables that load high on it

- **Factor loadings are correlations between the variables and the factors**. So, when factor loadings are high on a variable or a group of variable, it means that the factor represents the group of variables.
- Another useful aid in interpretation is to **plot the variables, using the factor loadings as coordinates**. Variables at the end of an axis are those that have high loadings on only that factor, and hence describe the factor.

Ideally we should like to see a pattern of loading such that each variable loads highly on a single factor and has small to moderate loading on the remaining factors.

Interpretation of Factor Loadings

Factors Underlying Selected Psychographics and Lifestyles



Rotation of Factors

- Factor rotation is simply **changing the "viewing angle" of the factor space**
- Through rotation the **loading matrix is transformed into a simpler one that is easier to interpret**
- **After rotation:**

Factors can be interpreted if ⇨	**Each variable has high loading on only one factor and moderate/low loadings on other factors.**

- The rotation is called **Orthogonal Rotation if the axes are maintained at right angles**

Rotation of Factors

Varimax Procedure:

- Axes maintained at right angles
- An orthogonal method of rotation that minimizes the number of variables with high loadings on a factor
- Orthogonal rotation results in uncorrelated factors

Factor Scores

The factor scores for the i^{th} factor may be estimated as follows:

$$F_i = W_{i1}X_1 + W_{i2}X_2 + W_{i3}X_3 + \ldots + W_{ik}X_k$$

- Intuitively, **factors are latent variables that underlie the scores in observed variables**. Factor scores would therefore represent the score of each person on the underlying latent variable.
- The **interpretation of each of these factors is based on the content of the original variables** so that each factor is interpreted as whatever the attributes with high loadings for this particular factor have in common.

Case Study – Athletics Records

Background

- Data for various countries' national athletics records is available.

Objective

- To discover hidden factors which can be interpreted as different skills required by athletes.

Available Information

- Data Source: Applied Multivariate Statistical Analysis by Richard A. Johnson , Dean W. Wichern
- Sample size is 55 countries athletics.
- Records for 8 different categories (activity – running) are available. Categories differ on the basis of length of tracks.

Data

Athleticsdata

Variables

Observations

Country	100m_s	200m_s	400m_s	800m_min	1500m_min	5000m_min	10000m_min	Marathon_min
Argentina	10.39	20.81	46.84	1.81	3.7	14.04	29.36	137.72
Australia	10.31	20.06	44.84	1.74	3.57	13.28	27.66	128.3

Column	Description	Type	Measurement	Possible Values
Country	Country Name	Categorical		
100m_s	Time for 100 meter running	Continuous	Seconds	Positive Values
200m_s	Time for 200 meter running	Continuous	Seconds	Positive Values
400m_s	Time for 400 meter running	Continuous	Seconds	Positive Values
800m_min	Time for 800 meter running	Continuous	Minutes	Positive Values
1500m_min	Time for 1500 meter running	Continuous	Minutes	Positive Values
5000m_min	Time for 5000 meter running	Continuous	Minutes	Positive Values
10000m_min	Time for 10000 meter running	Continuous	Minutes	Positive Values
Marathon_min	Time for Marathon running	Continuous	Minutes	Positive Values

Factor Analysis in R

```
# Import the data
data<-read.csv("Athleticsdata.csv", header=TRUE)
```

```
# Perform factor analysis
athletics<-subset(data,select=c(-Country))

fact<-factanal(athletics,2,rotation="varimax",scores="regression")

fact
```

- **subset()** is used to remove the column named “Country” from the data.
- **factanal()** from base R performs factor analysis on the given numeric data matrix.
- **rotation=“varimax”** performs varimax rotation of loading matrix
- **scores=**“regression” is used to specify method for factor scores

Factor Analysis in R

Output

```
Call:
factanal(x = athletics2, factors = 2, scores = "regression",     rotation = "vari
max")

Uniquenesses:
     X100m_s       X200m_s       X400m_s    X800m_min   X1500m_min   X5000m_min
       0.079         0.077         0.151        0.135        0.082        0.034
 X10000m_min Marathon_min
       0.018        0.086

Loadings:
             Factor1 Factor2
X100m_s      0.287   0.916
X200m_s      0.376   0.885
X400m_s      0.537   0.749
X800m_min    0.686   0.628
X1500m_min   0.795   0.535
X5000m_min   0.898   0.400
X10000m_min  0.904   0.406
Marathon_min 0.913   0.284

               Factor1 Factor2
SS loadings      4.071   3.268
Proportion Var   0.509   0.408
Cumulative Var   0.509   0.917

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 16.65 on 13 degrees of freedom.
The p-value is 0.216
> |
```

Interpretation:

- Two factors explain 92% of common variance.
- Factor 1 can be termed as ‘Stamina’ and factor 2 as ‘Speed’

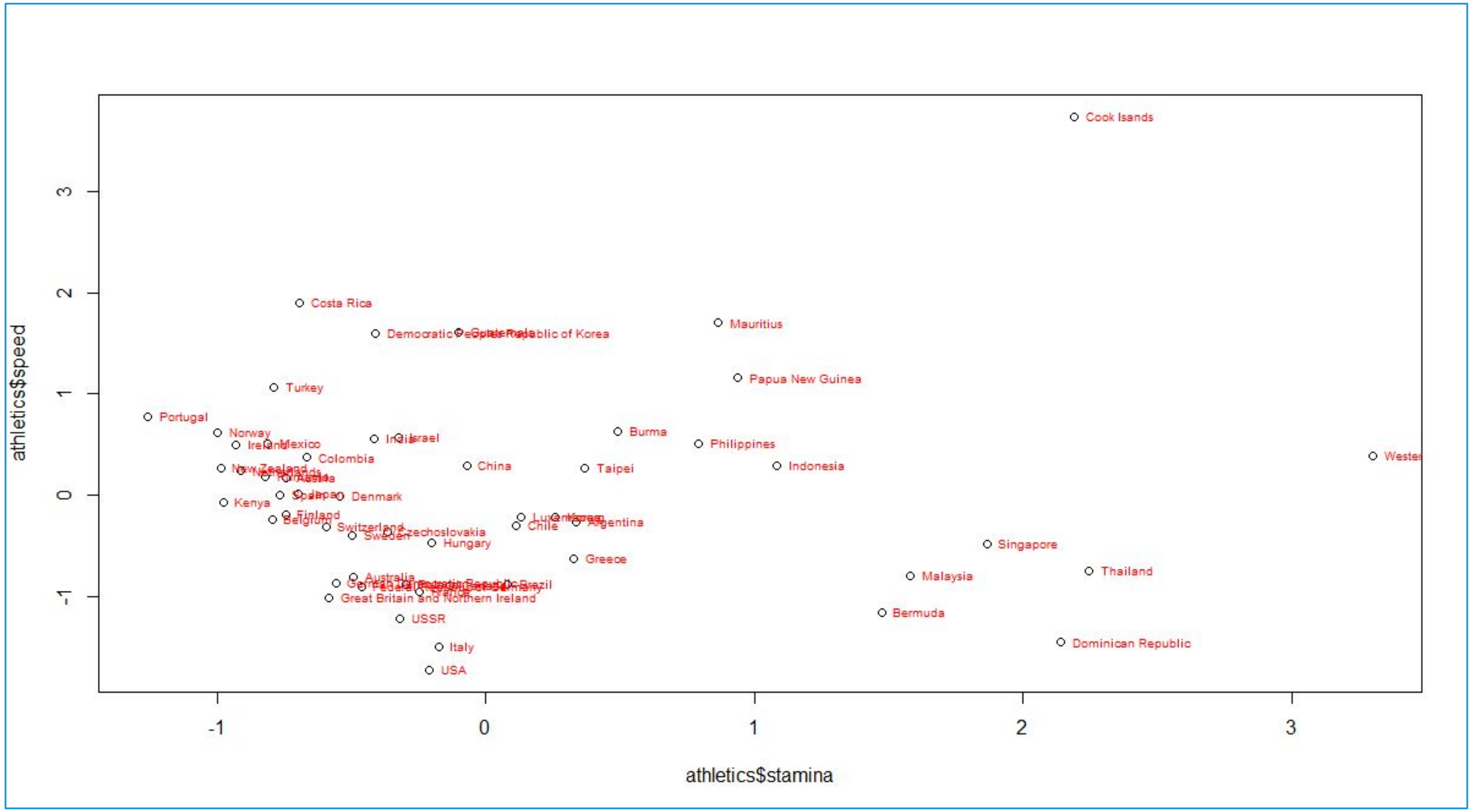
Factor Analysis in R – Factor Scores

```
# Adding Factor Scores to main data & plotting it
data$stamina<-fact$score[,1]
data$speed<-fact$score[,2]
plot(data$stamina,data$speed)
text(data$stamina,data$speed,data$Country,cex=0.6, pos=4,col="red")
```

- Factor scores are stored in the data frame ‘athletics’
- **text()** is used to assign names to each points in the scatter diagram.

Factor Analysis in R

```
# Output
```



Interpretation:
The points represent the scores on factor 1 and factor 2 for each country.

Quick Recap

In this session, we learnt about **Factor Analysis:**

Factor Analysis – Basics and Assumptions	• Each variable is expressed as a linear combination of factors. Some are common factors and one is unique.
	• The unobserved random variables
	    • Fj and ei are independent
	    • E(Fj)=0, Cov(Fj, Fk)=0 for j≠k and V(Fj)=1
	    • E(ei)=0 and Cov(e j, e k)=0 for j≠k

Estimation and Interpretation and Rotation of Factors	• Factors can be estimated by Principal Component Method or Maximum Likelihood Method • Factor loadings are correlations between the variables and the factors, high loadings on a group of variables suggests that the factor represents those variables • Rotation means changing the “viewing angle” of the factor space. The Varimax Procedure is the most common method of factor rotations
Factor Analysis in R	• **`factanal()`** function performs maximum likelihood factor analysis