UNIT IV

PRINCIPAL COMPONENT ANALYSIS, FACTOR ANALYSIS

Principal components, Algorithm for conducting PCA, deciding on how many principal components to retain, H-plot.

Factor analysis model, Extracting common factors, determining number of factors, Transformation of factor analysis solutions, Factor scores.

Introduction:

An important machine learning method for dimensionality reduction is called Principal component analysis. It is a method that uses simple matrix operations from linear algebra and statistics to calculate a projection of the original data into the same number or fewer dimensions.

It is originally introduced by Pearson in 1901 and independently by Hotelling in 1933.

Definition: PCA

Principal Component Analysis is a Statistical procedure that uses an orthogonal transformation which converts a set of correlated variables to a set of uncorrelated variables.

Principal Components: (P.C)

The first P.C of the observations is that linear combination Y_i of the original variables given by $y_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p$ —— ①
whose sample variance is greatest for all coefficient $a_{11}, \dots a_{1p}$ (which is represented by vector a_1).

The second principal components y_2 is that l.c $y_2 = a_{21}x_1 + a_{22}x_2 + \cdots + a_{2p}x_p$ —— ②
whose sample variance is greatest for all co-eff. & it is represented as vector a_2.
Clearly $a_2' a_1 = 0$ so that y_1 & y_2 are uncorrelated.
∴ j^{th} P.C is that l.c. $y_j = a_j' x$, which has greatest variance subject $a_j' a_j = 1$; $a_j' a_i = 0$ $(i<j)$ —— ③

To find the co-efficients the first P.C's we need to Choose the elts a_1 so as to maximize the variance of y_1 sub. to the constraint $a_1' a_1 = 1$. The variance of y_1 is

$Var(y_1) = Var(a_1' x) = a_1' S' a_1$ where S is the variance covariance matrix of the original variable x.

Points to remember :

1. Total variation $T = \sum_{i=1}^{p} \lambda_i \, |trace(s)|$

2. $Cov(x, y_j) = Cov(x, x'a_j) = E(x, x')\, a_j = S a_j = \lambda_j a_j$

3. $\gamma_{x_i y_j} = \frac{Cov(x_i, y_j)}{\sigma_{x_i} \sigma_{y_j}} = \frac{\lambda_j a_{ji}}{\sqrt{S_{ii}}\sqrt{\lambda_j}} = \frac{\sqrt{\lambda_j}\, a_{ji}}{\sqrt{S_{ii}}}$

1. Compute the principal component to the follo. 3x3 Variance covariance matrix for $n = 20$

$$S = \frac{1}{20-1}\begin{bmatrix} 54.8895 & 188.0405 & -34.4255 \\ 188.0405 & 3819.3495 & -107.4514 \\ -34.4285 & -107.4515 & 68.925 \end{bmatrix}$$

Solution:

$$S = \begin{bmatrix} 2.8889 & 9.8968 & -1.8120 \\ 9.8968 & 201.0183 & -5.65553 \\ -1.8210 & -5.6553 & 3.6276 \end{bmatrix}$$

To obtain the eigen values, consider $|S - \lambda I| = 0$

$\Rightarrow \lambda^3 - 207.5349\lambda^2 + 1219.1842\lambda - 1294.1323 = 0$

$\Rightarrow \lambda = 201.5167,\ 1.3865,\ 4.6316.$

To get the eigen vectors we have to solve $(S - \lambda I)x = 0$.

Case (i) When $\lambda = 201.5167$

$$(S - \lambda_1 I)x = \begin{bmatrix} -198.6272 & 9.8968 & -1.812 \\ 9.8968 & -0.4984 & -5.6553 \\ -1.812 & -5.6553 & -197.8891 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = 0$$

$\Rightarrow 198.6272\, x_1 + 9.8968\, x_2 - 1.812\, x_3 = 0$

$9.8968\, x_1 - 0.4984\, x_2 - 5.6553 x_3 = 0$

$-1.812\, x_1 - 5.6553\, x_2 - 197.8891 x_3 = 0$

Solving these eqn by Cross multiplication rule,

$x_1 = -1.1718 \quad x_2 = -34.6163, \quad x_3 = 1$

$$\Rightarrow X_1 = \begin{bmatrix} -1.1718 \\ -34.6163 \\ 1 \end{bmatrix}$$

The normal vector is $e_1 = \begin{bmatrix} -0.0338 \\ -0.9969 \\ 0.0288 \end{bmatrix}$ ($\because$ if $X = \begin{pmatrix} a \\ b \end{pmatrix}$ then normal vector $X = \begin{pmatrix} \frac{a}{\sqrt{a^2+b^2}} \\ \frac{b}{\sqrt{a^2+b^2}} \end{pmatrix}$)

Case (ii)

When $\lambda_2 = 4.6316$

The $(S_2 - \lambda_2 I) X = \begin{bmatrix} -1.7427 & 9.8968 & -1.812 \\ 9.8968 & 196.3867 & -5.6553 \\ -1.812 & -5.6553 & -1.004 \end{bmatrix} X = 0$

Solving these we get, $x_1 = -0.6812$, $x_2 = 0.0631$, $x_3 = 1$

The normal vector is $e_2 = \begin{bmatrix} -0.5622 \\ 0.05207 \\ 0.8253 \end{bmatrix}$

case (iii)

when $\lambda_3 = 1.3865$

we have to solve $(S - \lambda_3 I) X = 0$

$$\begin{bmatrix} 1.5024 & 9.8968 & -1.812 \\ 9.8968 & 199.6318 & -5.6553 \\ -1.812 & -5.6553 & 2.2411 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = 0$$

$\Rightarrow x_1 = 1.5138, \quad x_2 = -0.0467, \quad x_3 = 1$

Normal vector $e_3 = \begin{bmatrix} 0.8341 \\ -0.0257 \\ 0.551 \end{bmatrix}$

The population variance explained by the three principal components is given by

$$P_1 = \frac{\lambda_1}{\lambda_1+\lambda_2+\lambda_3} = \frac{201.5167}{207.5348} = 0.971$$

$$P_2 = \frac{\lambda_2}{\lambda_1+\lambda_2+\lambda_3} = \frac{4.6316}{207.5348} = 0.0223$$

$$P_3 = \frac{\lambda_3}{\lambda_1+\lambda_2+\lambda_3} = \frac{1.3865}{207.5348} = 0.0067$$

$\therefore P_1 + P_2 + P_3 = 0.999 \simeq 1$

The i^{th} P.C $= \frac{1}{i} = e_i' = X$

$$Y_1 = e_1' X = -0.0338\, x_1 - 0.999\, x_2 + 0.0288\, x_3$$

$$Y_2 = e_2' X = -0.5622\, x_1 + 0.05207\, x_2 + 0.8253\, x_3$$

$$Y_3 = e_3' X = 0.8341\, x_1 - 0.0257\, x_2 + 0.551\, x_3.$$

$\therefore$ The first principal component covers the maximum & the other two principal components covers least.

Note:

1. The results of principal component analysis depend on the measurement scales.

2. Variables with the highest sample variances tend to be emphasized in the first few principal components.

3. PCA using the Covariance function should only be considered if all the variables have the same units of measurements.

If the variables have different unit of measurement, (ie; pounds, feet, gallons etc) or if we wish each variable to receive equal weight in the analysis then the variables should be standardized before conducting a principal component analysis. To standardize a variable subtract the mean & divide by the Std. deviation

$$Z_{ij} = \frac{X_{ij} - \overline{x}_j}{S_j}$$ where

X_{ij} = data for variable j in sample unit i

$\overline{x}_j$ = Sample mean for variable j

S_j = sample S.D for variable j.

Note: The variance - covariance matrix of the std. data is equal to the correlation matrix for the unstandardized data.

∴ PCA using the Std. data = PCA using the correlation matrix.

Factor analysis:

It is a statistical method used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors. For example, it is possible that variations in six observed variables mainly reflect the variations in two unobserved variables.

Objectives:

1. To understand the terminology of Factor analysis, including interpretation of factor loadings, specific variances and communalities.
2. Understand how to apply both PCA & maximum likelihood methods for estimating the parameters of a factor model.
3. Understand factor rotation & interpret rotated factor loadings.

Notations:

Let X_i denote Observable trait i

$$X_{p\times 1} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix} = \text{vector of traits}$$

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} = \text{population mean vector}$$

$E(X_i) = \mu_i$ denotes the population mean of variable i

Consider m unobservable common factor $f_1, f_2, \ldots, f_m$

$$\underset{m\times 1}{f} = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_m \end{pmatrix} = \text{vector of common factors}$$

Let
$$X_1 = \mu_1 + l_{11} f_1 + l_{12} f_2 + \cdots + l_{1m} f_m + \varepsilon_1$$
$$X_2 = \mu_2 + l_{21} f_2 + l_{22} f_2 + \cdots + l_{2m} f_m + \varepsilon_2 .$$
$$\vdots$$
$$X_p = \mu_p + l_{p1} f_1 + l_{p2} f_2 + \cdots + l_{pm} f_m + \varepsilon_p$$

where
$$L = \begin{pmatrix} l_{11} & l_{12} & \cdots & l_{1m} \\ l_{21} & l_{22} & \cdots & l_{2m} \\ \vdots & & & \\ l_{p1} & l_{p2} & \cdots & l_{pm} \end{pmatrix} = \text{matrix of factor loadings}$$

(instead of l_{11} we can use λ_{11})

$$\delta \text{ or } \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{pmatrix} = \text{vector of specific factors}$$

Generally, we get $\underset{p\times 1}{X} = \underset{p\times 1}{\mu} + \underset{(p\times m)(m\times 1)}{Lf} + \varepsilon .$ (Note: $m < p$)

or $\boxed{X = \mu + \Lambda F + \delta}$

$$\Rightarrow X - \mu = \Lambda F + \delta \quad \text{——— (1)}.$$

Factor model: Assumptions

$E(x) = \mu$, $E(F) = 0$, $E(\delta) = 0$

$Cov(x) = \Sigma$, $Cov(F) = E(FF^T) = I$

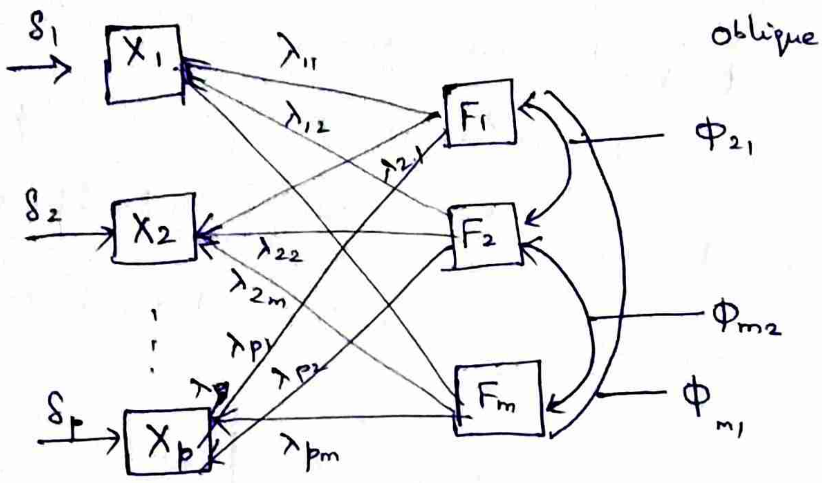
$$Cov(\delta) = \Psi = \begin{bmatrix} \psi_{11} & 0 & \cdots & 0 \\ 0 & \psi_{22} & \cdots & 0 \\ \vdots & & & \\ 0 & 0 & \cdots & \psi_p \end{bmatrix}$$

$Cov(F\delta) = 0$

Purpose of Factor analysis:

The purpose of Factor analysis is to describe, if possible, the covariance relationships among many variables in terms of a few underlying but unobservable, random quantities called "Factors".

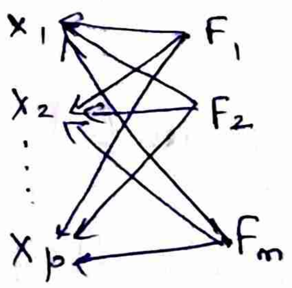
- The determination of a small no. of factors based on a particular no. of inter-related quantitative variables.
- Unlike variables directly measured such as speed, height, weight etc, some variables such as egoism, creativity, happiness, religiosity, comfort are not a single measurable entity.
- They are constructs that are derived from the measurement of other, directly observable variables.



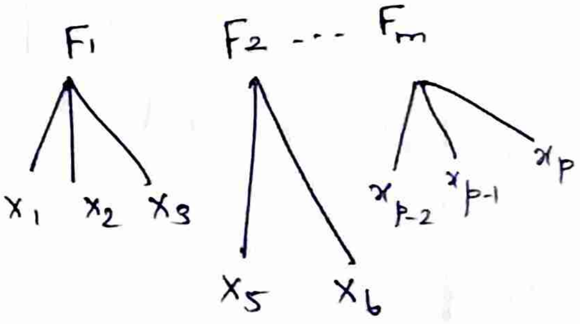
Types of Factor analysis

1. Exploratory Factor model
2. Confirmatory Factor model.

Exploratory :



Confirmatory



- Factor analysis quantifies those constructs (factors) with the help of the manifest variables.
- Factor analysis also reduces information (dimension)

Assumptions:

1. Variables must be interrelated
   - 20 unrelated variables = 20 factors
   - matrix must have sufficient no. of correlations

2. Sample must be homogeneous

3. Metric variables assumed

4. MV normality not required

5. Sample size
   - min 50, prefer 100
   - min 5 observations/item, prefer 10 observations/item

Types of Factor analysis:

1. Exploratory Factor analysis (EFA)
   - Used to discover underlying structure
   - Principal component analysis (PCA)(Thurstone)
     - Considers the total variance & derive factors that contain little amount of unique & error variance
     - Unity inserted on diagonal of matrix
     - Often used in physical science
   - Factor analysis (common factors analysis)(Spearman)
     - Considers only the common or shared variance & ignores the unique & error variance. It is complicated thus less used.
     - In SPSS known as principal axis factoring.

- Both PCA and FA gave similar answers most of the time & especially when the no. of variables are > 30 or the communalities > 0.6 for most variables.

2. Confirmatory factor analysis (CFA)

- Used to test whether data fit a prior expectations for data structure
- Structural equations modelling.

Basic Logic of EFA

- Items you want to reduce
- Creates mathematical combination of variables that maximizes variance you can predict in all variables → Principal component or factor.
- New combination of items from residual variance that maximizes variances you can predict in what is left → second principal component or factor. Continue until all variance is accounted for.
- Select the minimal no. of factors that captures the most amount of variance
- Interpret the factors
- Rotated matrix & loadings are more interpretable.

Concepts & Terms:

1. Factor - linear composite. A way of turning multiple measures into one thing.
2. Factor Score - Measure of one person's score on a given factor.
3. Factor loadings - Correlation of a factor score with an item. Variables with high loadings are the distinguishable features of the factor.
4. Communality - (h^2) - Variance in a given item accounted for by all factors. Sum of squared factor loadings in a row from factor analysis results. These are presented in the diagonal in common factor analysis.
5. Factorally pure - A test that only loads on one factor.
6. Scale score - Score for individual obtained by adding together items making up a factor.
7. Eigen value - Column sum of squared loadings & indicates the relative importance of each factor in accounting for the variance associated with the set of variables.

How many factors?

- Because we are trying to reduce the data, we don't want as many factors as items.
- Because each new component or factor is the best linear combination of residual variance, data can be explained relatively well in many less factors than original number of items.
- Stop taking additional factors is a difficult decision. Primary methods:
- Scree plot – Not a test
  - — Look for bend in plot
  - — Include factor located right at bend point.
- Kaiser (or Latent root) Criterion
  - — Eigen values greater than 1
  - — Also, 1 is the amount of variance accounted for by a single item ($r^2 = 1.00$).

If eigen value < 1.00 then factor accounts for less variance than a single item.

Rotation of factors:

- After rotation, variance accounted for by a factor is spread out. First factor no longer accounts for max variance possible; other factors get more variance. Though the total variance accounted for remains the same.
- Two types of rotation
  - Orthogonal (factors uncorrelated)
  - Oblique (factors correlated).
- Orthogonal rotation (rigid 90 degrees). Factors remain uncorrelated after transformation.

— Varimax — Simplifying column weights to 1s and 0s. Factor has items loading highly, others don't load. Not appropriate if you expect a single factor. Maximizes the sum of variances of required loadings of the factor matrix. (use it with Kaiser's normalisation)

— Quartimax — simplify to 1s to 0s in a row. Items load high on 1 factor, almost 0 on others. Appropriate if you expect single general factor.

— Equimax — Mix of Varimax & Quartimax.

- Oblique or correlated components (less or more than 90 degrees). Accounts for same variance but factor correlated.

- Not meaningful with PCA
- Many factors are theoritically related, so rotation method not force orthogonality.
  - Let loadings are more closer match simple structure
  - Correlated solutions will get you close to simple structure
  - Oblimin and Promax (requires a large data set < 150) are good.

Factor analysis example :

Academic Ability of a student

Quantitative Ability

1. Math score
2. Computer programming score
3. Physics score

Verbal Ability

1. English
2. Verbal reasoning score

Finding out factor:

- Independent variables
  - Math score
  - Verbal reasoning score
  - Physical score
  - Communication skill score
  - Computer programming score
  - Statistics score