FACTOR ANALYSIS

Akshai James (194102301)

8 JUN 2020

1 Theory

Factor analysis is a statistical method used to describe variability among observed,correlated variables in terms of a potentially lower number of unobserved variables called factors.Factor analysis searches for joint variations in response to unobserved latent variables.The observed variables are modeled as linear combinations of potential factors,plus "error" terms.Factor analysis aims to find independent latent variables.

The theory behind factor analytic methods is that the information gained about the interdependencies between observed variables can be used later to reduce the set of variables in a dataset.Factor analysis commonly used in biology,personality theories,marketing,finance.It may help to deal with datasets where there are large numbers of observed variables that are thought to reflect a smaller number of underlying latent variables.It is one of the most commonly used inter-dependency techniques and is used when relevant set of variables shows a systematic inter-dependence and the objective is to find out the latent factors that create a commonality.

Principle component analysis(PCA) can be considered as a more basic version of exploratory factor analysis(Figure.1),that was developed in early days prior to advent of high speed computers.Both PCA and Factor analysis aim to reduce the dimensionality of a set of data.Factor analysis is clearly designed with the objective to identify certain unobservable factors from the observaed variables.

1.1 Definition

Suppose we have a set of P observable random variables, $\{x_i\}$;i=1,2,....,P with means $\{\mu_j\}$;j=1,2,...P.

For some unknown constants l_{i_j} and K, unobserved random variables F_j (called "common factors"),where i varies from 1 to P and j varies from 1 to K,where P$\geq K$

We have that the terms in each random variable should be writeable as a linear combination of the common factors.

$$x_i - \mu_i = l_{i1} \ F_1 + ...+= l_{iK} \ F_K + e_i$$

Here e_i are unobserved stochastic error terms with zero mean and finite variance ,which may not be same for all i.
In matrix form,

$$X - \mu = LF + e$$

If we have N observations,then we will have dimensions X_{P*N} , L_{P*K} , F_{K*N} .Each column of x and F denotes values for one particular observation and matrix L does not vary across observations.

We will impose the following assumptions on F.

- F and e are independent
- E(F) =0 (E is expectation)
- Cov(F) = I (Cov is cross-covariance matrix,to make sure that the factors are uncorrelated and I is identity matrix)

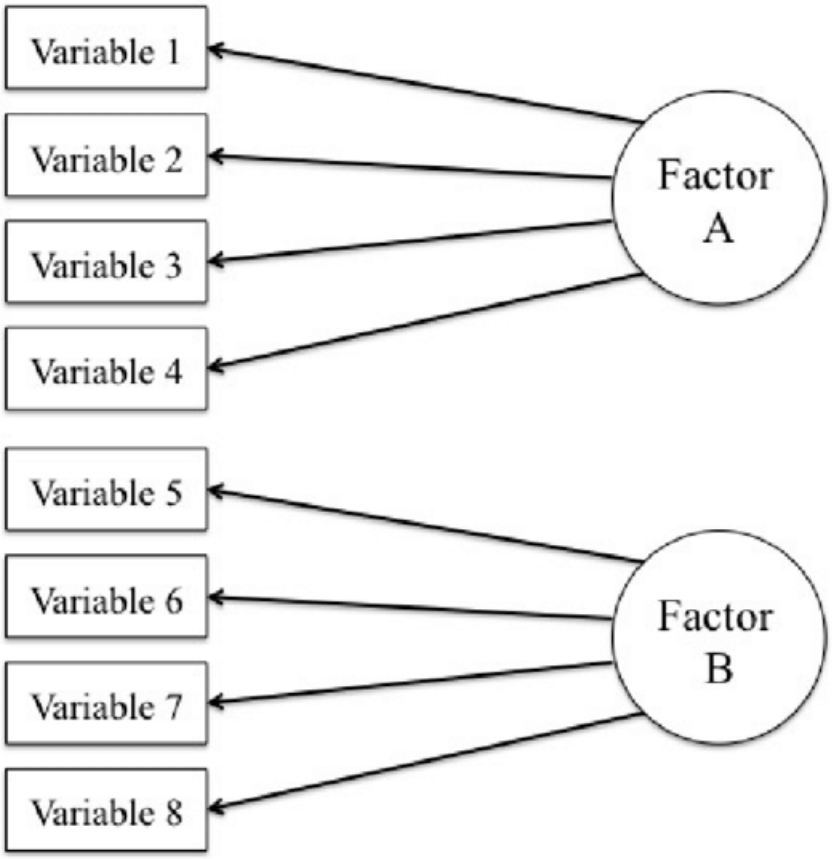

Figure 1: Exploratory-Factor-Analysis

Any solution of the above set of equations following the constraints for F is defined as the factors and L as the loading matrix.
Suppose,

$$\text{Cov(X- } \mu \text{)} = \sum$$

Then,

$$\sum = Cov(X - \mu \ ) = \text{Cov(LF+e)}$$

From the conditions imposed on F above;

$$\sum = L * Cov(F) * L^T + Cov(e)$$

Suppose, Cov(e) = H

$$\sum = L * L^T + H$$

1.2 Algorithm

1. Read the input X such that $X \epsilon R^d$
2. Find the mean of the input data $\mu \epsilon R^d$
3. Find $\widetilde{X} = X - \mu$
4. Obtain the matrix A for N data

   $A = \begin{bmatrix} \tilde{X_1} & \tilde{X_2} & \cdots & \tilde{X_N} \end{bmatrix}_{d \times N}$
5. Find the covariance matrix of A

   $\hat{C} = \frac{AA^T}{N}$
6. Find top three eigenvalues of A and corresponding eigenvectors
7. Find the Loading matrix using equaton

   $\hat{V} = \widetilde{Q} \cdot \widetilde{D}$

   $\widetilde{Q} = \begin{bmatrix} \hat{e_1} & \hat{e_2} & \hat{e_3} \end{bmatrix}$

   $\widetilde{D} = \begin{bmatrix} \sqrt{T}_1 & 0 & 0 \\ 0 & \sqrt{T}_2 & 0 \\ 0 & 0 & \sqrt{T}_3 \end{bmatrix}$
8. Find factor matrix using equation

   $Z = \left(\hat{C}^{-1} \cdot \hat{V}\right)^T A$