Principal component Analysis

Friday weekly Meeting

ACIPH

November 18

Learning Objectives

At the end of the session, participants will be able to:

- Meaning of PCA
- Define some important terms
- Describe common assumptions for PCA
- Understand Steps in PCA
- Practice Wealth Index in SPSS

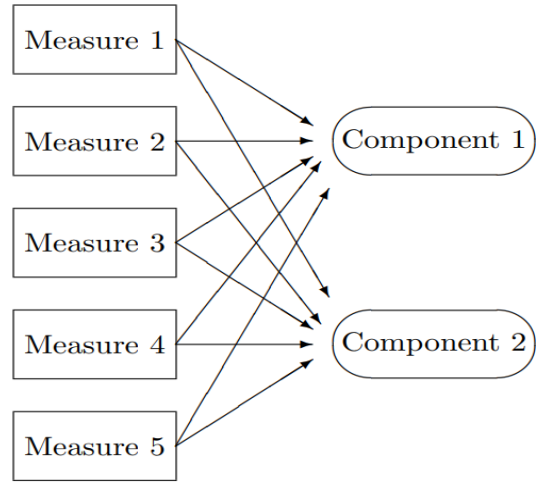
Introduction

Principal Component Analysis (PCA)

- It is a data reduction approach to create **one or more index variables** from a larger set of measured variables
- Reduces the number of observed variables to a smaller number of principal components which account for most of the variance of the observed variables
- Using a linear combination (basically a weighted average) of a set of variables.

$$C = w_1(Y_1) + w_2(Y_2) + w_3(Y_3) + w_4(Y_4)$$

- The created index variables are called components.



Definitions

- **An observed/indicator/measured variable:** is a variable directly measured from an observation.
- **A principal component:** is a linear combination of weighted observed variables.
- **A latent variable :** a variable produced from construct indirectly by determining its influence to responses on measured variables.
- **Eigenvalues:** indicate the amount of variance explained by each principal component or each factor.
- **Orthogonal:** means at a 90 degree angle, perpendicular.
- **Obilque:** means other than a 90 degree angle

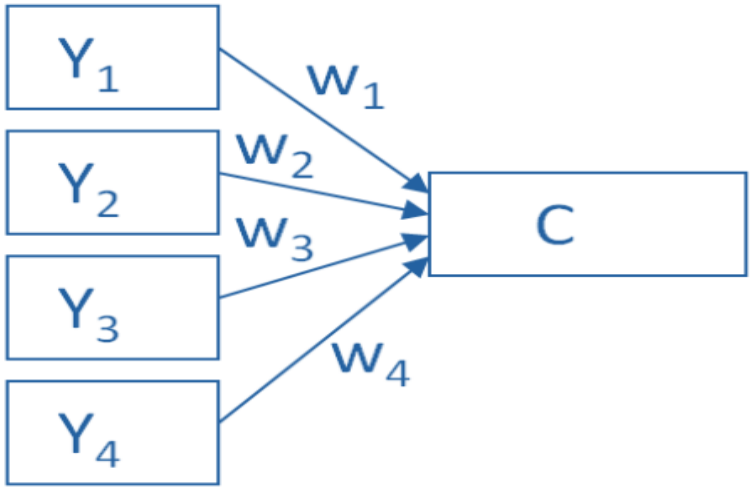
Assumptions of PCA

- Measurement scale is interval or ratio level
- Random sample - at least 5 observations per observed variable and at least 100 observations.
- Larger sample sizes recommended for more stable estimates,

  10-20 observations per observed variable
- Linear relationship between observed variables
- Normal distribution for each observed variable
- It is variable reduction techniques.
- It assumes the absence of outliers in the data.

Difference Between PCA and EFA

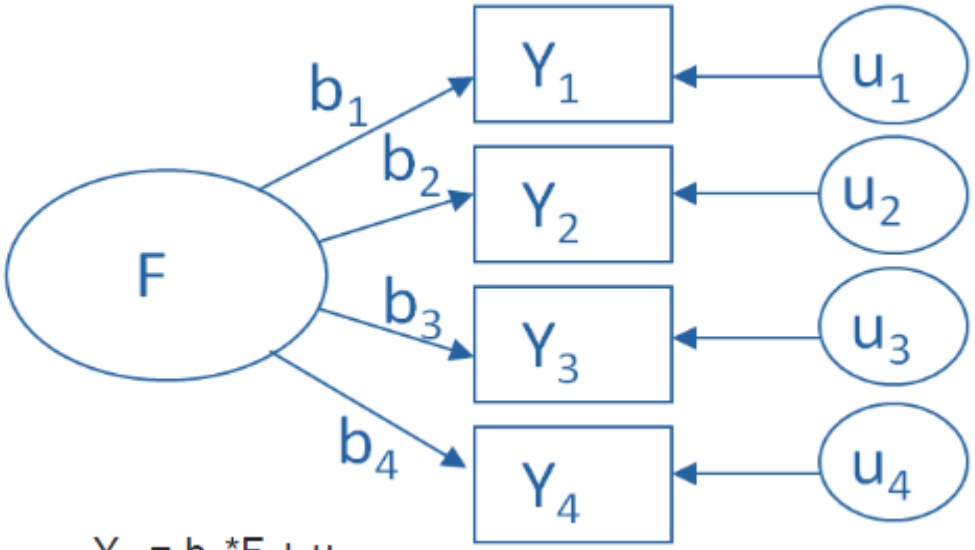
Principal Component Analysis



$C = w_1(Y_1) + w_2(Y_2) + w_3(Y_3) + w_4(Y_4)$

Principal Components retained account for a maximal amount of variance of observed variables

Explanatory Factor Analysis



$Y_1 = b_1*F + u_1$
$Y_2 = b_2*F + u_2$
$Y_3 = b_3*F + u_3$
$Y_4 = b_4*F + u_4$

Steps Involved in PCA

There are three main steps in conducting PCA analysis

Step 1: Assessment of the suitability of the data for PCA analysis

- sample size, and the strength of the relationship among the variables (or items)

Sample size

- Little agreement among authors concerning how large a sample should be, the recommendation generally is:

- ✓ The larger is the better.
- ✓ At least 300 cases for factor analysis /or 150

  (Tabachnick and Fidell , 2007)
- ✓ The ratio of participants to items recommends a 10 to 1 ratio

  (Nunnally , 1978)

Addis Continental Institute of Public Health

Strength of Association

- The strength of the inter correlations among the items;
  - Correlation matrix for evidence of coefficients greater than 0.3.

    (Tabachnick and Fidell)
  - Two statistical measures are generated by SPSS

i). Bartlett's test of sphericity (Bartlett 1954)

- BTS should be significant ($p < .05$) to be considered appropriate

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

(Kaiser 1970, 1974).

- The KMO index ranges from 0 to 1, with 0.6 suggested as the minimum value for a good factor analysis

Addis Continental Institute of Public Health

Step 2: Factor extraction

- There are a number of techniques that can be used to assist in the decision concerning the number of factors to retain:

  i). Kaiser's criterion

  ii). Scree test and

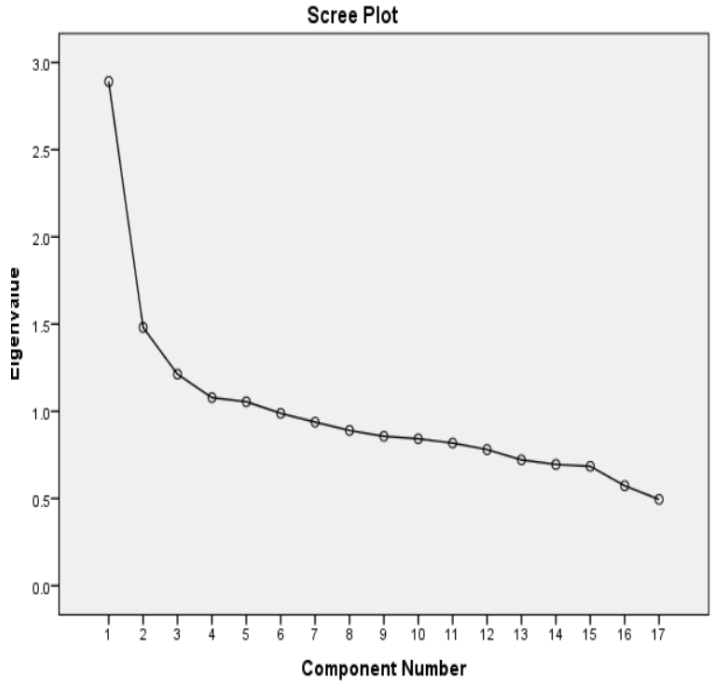
  iii). Parallel analysis.

Kaiser's criterion

- known as Kaiser's criterion, or the eigen value rule
- Only factors with an eigen value of 1.0 or more are retained for further investigation
- The eigen value of a factor represents the amount of the total variance explained by that factor.

Scree test

- An approach that can be used is Catell's scree test (Catell 1966).
- This involves plotting each of the eigenvalues of the factors
- Inspecting the plot to find a point at which the shape of the curve changes direction and becomes horizontal.
- Retaining all factors above the elbow, or break in the plot,
- Factors contribute the most to the explanation of the variance in the data set.



Parallel analysis

- Parallel analysis involves comparing the size of the eigenvalues with those obtained from a randomly generated data set of the same size.

- MonteCarlo: - Sample size, Variables , 100 iteration

  - generate random data matrices

- Only those eigenvalues that exceed the corresponding values from the random data set are retained

Addis Continental Institute of Public Health

Step 3: Factor rotation and interpretation

- Once the number of factors has been determined, the next step is to try to interpret them.
- To assist in this process, the factors are ‘rotated’.
- This does not change the underlying solution—rather, it presents the pattern of loadings in a manner that is easier to interpret.
- SPSS does not label or interpret each of the factors for you.

- There are two main approaches to rotation, resulting in either
  - orthogonal (uncorrelated)
  - oblique (correlated) factor solutions.
- Orthogonal rotation results in solutions that are easier to interpret and to report
- However, they do require the researcher to assume (usually incorrectly) that the underlying constructs are independent (not correlated).
- Oblique approaches allow for the factors to be correlated, but they are more difficult to interpret, describe and report

Example: Wealth index

- Household characteristics of a survey

- Variables of :

- b03a b03b b03c b03d b03e b03f b03g b03h b03i b03j b03k b03l b03m b03n b03o b03p b03q.

Interpretation 1/2

Step 1: suitable for factor analysis

- Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO)>0.6.
- Bartlett's Test of Sphericity value is significant P<0.05
- Correlation Matrix >0.3

Step 2: Determine how many components are included

- Initial eagle values greater than 1
- Cumulative index % of explained.

Step 3: Kaiser criterion for shape of the plot.

- Screeplot

Interpretation 2/2

Step 4: Eigenvalue criterion

- By comparing results from **Monte Carlo** Generated number.
- If your value is larger than the criterion value from parallel analysis

Step 5: Component Matrix.

- Kaiser criterion >0.4 will be included to the components

Step 6: Pattern Matrix criteria

- Ideally, we can decide items loading on each component(3 or more)
- You can decide the number of components by :
- Changing eigen value from one to -*Fixed number of factors to extract* .
- Removing items with low communality values tends to increase the total variance explained- (Good to be >0.3 for communality value)

Thank you