Contents

1 Review of Important Definitions

- An observed variable can be measured directly, is sometimes called a measured variable or an indicator or a manifest variable.
- A principal component is a linear combination of weighted observed variables. Principal components are uncorrelated and orthogonal.
- A latent construct can be measured indirectly by determining its influence to responses on measured variables. A latent construct could is also referred to as a factor, underlying construct, or unobserved variable.
- Factor scores are estimates of underlying latent constructs.
- Unique factors refer to unreliability due to measurement error and variation in the data.
- Principal component analysis minimizes the sum of the squared perpendicular distances to the axis of the principal component while least squares regression minimizes the sum of the squared distances perpendicular to the x axis (not perpendicular to the fitted line).
- Principal component scores are actual scores.
- Eigenvectors are the weights in a linear transformation when computing principal component scores. 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.
- An observed variable ***loads*** on a factors if it is highly correlated with the factor, has an eigenvector of greater magnitude on that factor.

- Communality is the variance in observed variables accounted for by a common factors. Communality is more relevant to EFA than PCA.

2 Exploratory Factor Analysis

Principal Component Analysis (PCA) and Exploratory Factor Analysis (EFA) are both variable reduction techniques and sometimes mistaken as the same statistical method. However, there are distinct differences between PCA and EFA. Similarities and differences between PCA and EFA will be examined.

2.1 Principal Component Analysis (PCA)

- Is a variable reduction technique
- Is used when variables are highly correlated
- Reduces the number of observed variables to a smaller number of principal components which account for most of the variance of the observed variables
- Is a large sample procedure

The number of components extracted is equal to the number of observed variables in the analysis. The first principal component identified accounts for most of the variance in the data. The second component identified accounts for the second largest amount of variance in the data and is uncorrelated with the first principal component and so on.

The total amount of variance in PCA is equal to the number of observed variables being analyzed. In PCA, observed variables are standardized, (e.g., mean=0, standard deviation=1).

Components accounting for maximal variance are retained while other components accounting for a trivial amount of variance are not retained. Eigenvalues indicate the amount of variance explained by each component. Eigenvectors are the weights used to calculate components scores.

2.2 Factor Analysis

Factor analysis is a statistical procedure to identify interrelationships that exist among a large number of variables, i.e., to identify how suites of variables are related.

Factor analysis can be used for exploratory or confirmatory purposes. As an exploratory procedure, factor analysis is used to search for a possible underlying structure in the variables. In confirmatory research, the researcher evaluates how similar the actual structure of the data, as indicated by factor analysis, is to the expected structure.

The major difference between exploratory and confirmatory factor analysis is that researcher has formulated hypotheses about the underlying structure of the variables when using factor analysis for confirmatory purposes.

As an exploratory tool, factor analysis doesn't have many statistical assumptions. The only real assumption is presence of relatedness between the variables as represented by the correlation coefficient. If there are no correlations, then there is no underlying structure.

2.3 Exploratory Factor Analysis (EFA)

- Is a variable reduction technique which identifies the number of ***latent constructs*** and the underlying factor structure of a set of variables

- Hypothesizes an underlying construct, a variable not measured directly
- Estimates factors which influence responses on observed variables
- Allows you to describe and identify the number of latent constructs (also known as factors)
- Includes unique factors, error due to unreliability in measurement
- Traditionally has been used to explore the possible underlying factor structure of a set of measured variables without imposing any preconceived structure on the outcome.

The figure below shows four ***factors*** (ovals) each measured by 3 observed variables (rectangles) with unique factors. Since measurement is not perfect, error or unreliability is estimated and specified explicitly in the diagram.

Factor loadings (parameter estimates) help interpret factors. Loadings are the correlation between observed variables and factors, are standardized regression weights if variables are standardized (weights used to predict variables from factor). Standardized linear weights represent the effect size of the factor on variability of observed variables.

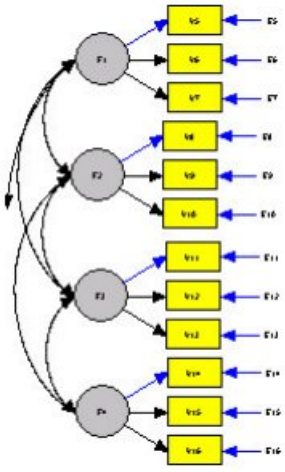

Figure 1: Factor Analysis

In Exploratory Factor Analysis, observed variables are a linear combination of the underlying factors (estimated factor and a unique factor).

Communality is the variance of observed variables accounted for by a common factor. Large communality is strongly influenced by an underlying construct.

2.4 Similarities between PCA and EFA

- PCA and EFA have these assumptions in common:
  - 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

- 'Over-sample' to compensate for missing values
- Linear relationship between observed variables
- Normal distribution for each observed variable
- Each pair of observed variables has a bivariate normal distribution
- PCA and EFA are both variable reduction techniques. (If communalities are large, close to 1.00, results could be similar.)

2.5 Differences between PCA and FA

These techniques are typically used to analyze groups of correlated variables representing one or more common domains; for example, indicators of socioeconomic status, job satisfaction, health, self-esteem, political attitudes or family values. Principal components analysis is used to find optimal ways of combining variables into a small number of subsets, while factor analysis may be used to identify the structure underlying such variables and to estimate scores to measure latent factors themselves. The main applications of these techniques can be found in the analysis of multiple indicators, measurement and validation of complex constructs, index and scale construction, and data reduction. These approaches are particularly useful in situations where the dimensionality of data and its structural composition are not well known.

When an investigator has a set of hypotheses that form the conceptual basis for her/his factor analysis, the investigator performs a confirmatory, or hypothesis testing, factor analysis. In contrast, when there are no guiding hypotheses, when the question is simply what are the underlying factors the investigator conducts an exploratory factor analysis.

The factors in factor analysis are conceptualized as "real world" entities such as depression, anxiety, and disturbed thought. This is in contrast to principal components analysis (PCA), where the components are simply geometrical abstractions that may not map easily onto real world phenomena.

2.5.1 Treatment of Variance

Another difference between the two approaches has to do with the variance that is analyzed. In PCA, all of the observed variance is analyzed, while in factor analysis it is only the shared variances that is analyzed.

- Principal Components retained account for a maximal amount of variance of observed variables. Factors account for common variance in the data.
- PCA Analysis decomposes correlation matrix. EFA decomposes adjusted correlation matrix.
- PCA: Ones on the diagonals of the correlation matrix. EFA Diagonals of correlation matrix adjusted with unique factors.

- PCA: Minimizes sum of squared perpendicular distance to the component axis. EFA: Estimates factors which influence responses on observed variables.
- PCA: Component scores are a linear combination of the observed variables weighted by eigenvectors. EFA: Observed variables are linear combinations of the underlying and unique factors.

2.6 PCA Terminology

- PC loadings are correlation coefficients between the PC scores and the original variables.
- PC loadings measure the importance of each variable in accounting for the variability in the PC. It is possible to interpret the first few PCs in terms of 'overall' effect or a 'contrast' between groups of variables based on the structures of PC loadings.
- high correlation between PC1 and a variable indicates that the variable is associated with the direction of the maximum amount of variation in the dataset.
- More than one variable might have a high correlation with PC1. A strong correlation between a variable and PC2 indicates that the variable is responsible for the next largest variation in the data perpendicular to PC1, and so on.
- if a variable does not correlate to any PC, or correlates only with the last PC, or one before the last PC, this usually suggests that the variable has little or no contribution to the variation in the dataset. Therefore, PCA may often indicate which variables in a dataset are important and which ones may be of little consequence. Some of these low-performance variables might therefore be removed from consideration in order to simplify the overall analyses.

2.7 Bi-plot Display of PCA

- Bi-plot display is a visualization technique for investigating the inter-relationships between the observations and variables in multivariate data.
- To display a bi-plot, the data should be considered as a matrix, in which the column represents the variable space while the row represents the observational space.
- The term bi-plot means it is a plot of two dimensions with the observation and variable spaces plotted simultaneously.
- In PCA, relationships between PC scores and PCA loadings associated with any two PCs can be illustrated in a bi-plot display

2.8 Communality

Communality refers to the total amount of variance an original variable shares with all other variables included in the analysis.This is the proportion of each variable's variance that can be explained by the principal components . (It is denoted as h^2 and can be defined as the sum of squared factor loadings).

Initial - By definition, the initial value of the communality in a principal components analysis is 1.

Extraction - The values in this column indicate the proportion of each variable's variance that can be explained by the principal components. Variables with high values are well represented in the common factor space, while variables with low values are not well represented. They are the reproduced variances from the number of components that you have saved. You can find these values on the diagonal of the reproduced correlation matrix.

2.9 What is a Rotation

Ideally, you would like to review the correlations between the variables and the components and use this information to interpret the components; that is, to determine what construct seems to be measured by component 1, what construct seems to be measured by component 2, and so forth. Unfortunately, when more than one component has been retained in an analysis, the interpretation of an unrotated factor pattern is usually quite difficult. To make interpretation easier, you will normally perform an operation called a rotation. A rotation is a linear transformation that is performed on the factor solution for the purpose of making the solution easier to interpret.

2.10 Varimax Rotation

A varimax rotation is an orthogonal rotation, meaning that it results in uncorrelated components. Compared to some other types of rotations, a varimax rotation tends to maximize the variance of a column of the factor pattern matrix (as opposed to a row of the matrix). This rotation is probably the most commonly used orthogonal rotation in the social sciences.

2.11 Interpreting the Rotated Solution

Interpreting a rotated solution means determining just what is measured by each of the retained components. Briefly, this involves identifying the variables that demonstrate high loadings for a given component, and determining what these variables have in common. Usually, a brief name is assigned to each retained component that describes its content. The first decision to be made at this stage is to decide how large a factor loading must be to be considered "large."

Guidelines are provided in statistical literature for testing the statistical significance of factor loadings. Given that this is an introductory treatment of principal component analysis, however, simply consider a loading to be large if its absolute value exceeds 0.40.