A SURVEY PAPER ON COMPARATIVE STUDY BETWEEN PRINCIPAL COMPONENT ANALYSIS (PCA) AND EXPLORATORY FACTOR ANALYSIS (EFA)

Parul M.Jain[*]

V.K.Shandliya[**]

Abstract

Principal Component Analysis (PCA) and Exploratory Factor Analysis (EFA) are both variable reduction techniques. There are distinct differences between PCA and EFA. Similarities and differences between PCA and EFA are studied in this paper. Principal Components retained account for a maximal amount of variance of observed variables while Factors account for common variance in the data. PCA decomposes correlation matrix while EFA decomposes adjusted correlation matrix. Exploring basic theory of multivariate analysis, which involves a mathematical procedure to transform a number of correlated variables into a number of uncorrelated variables have been studied, compared and analyzed for better performance.

Keywords: Principal Component Analysis, Exploratory Factor Analysis, Principal Components, Correlated variables, uncorrelated variables, eigenvalues, eigenvectors.

[*] M.E. (C.S.E.) First Year, Sipna College of Engineering & Technology, Amravati

[**] Professor, Computer Science & Engineering Department, Sipna College of Engineering & Technology, Amravati

1. Introduction

Principal component analysis (PCA) is a mathematical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components is less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has the largest possible variance and each succeeding component in turn has the highest variance possible under the constraint that it be orthogonal to the preceding components. Principal components are guaranteed to be independent only if the data set is jointly normally distributed. PCA is sensitive to the relative scaling of the original variables while 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. In other words, it is possible, for example, that variations in three or four observed variables mainly reflect the variations in fewer unobserved variables. Factor analysis searches for such joint variations in response to unobserved latent variables. The observed variables are modeled as linear combinations of the potential factors, plus "error" terms. The information gained about the interdependencies between observed variables can be used later to reduce the set of variables in a dataset. Computationally this technique is equivalent to low rank approximation of the matrix of observed variables. Factor analysis is related to principal component analysis (PCA), but the two are not identical. Latent variable models, including factor analysis, use regression modeling techniques to test hypotheses producing error terms, while PCA is a descriptive statistical technique.

2. Definitions

2.1 Principal Component Analysis (PCA)

PCA is a variable reduction technique. It is used when variables are highly correlated. It reduces the number of observed variables to a smaller number of principal components which account for most of the variance of the observed variables. It is a large sample procedure. 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, diagonals of the matrix are equal to 1. The amount of variance explained is equal to the trace of the matrix (sum of the diagonals of the decomposed correlation matrix).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. Components accounting for maximal variance are retained while other components accounting for a trivial amount of variance are not retained. Eigen values indicate the amount of variance explained by each component. Eigen vectors are the weights used to calculate components scores.

2.2 Exploratory Factor Analysis (EFA)

EFA is a variable reduction technique which identifies the number of latent constructs and the underlying factor structure of a set of variables. It hypothesizes an underlying construct, a variable not measured directly. It estimates factors which influence responses on observed variables. It allows to describe and identify the number of latent constructs .It includes unique factors, error due to unreliability in measurement. It 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.

3. The PCA and EFA models

3.1 PCA MODEL

PCA is mathematically defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on. A data matrix is defined, $\mathbf{X}^T$, with zero empirical mean where each of the n rows represents a different repetition of the experiment, and each of the m columns gives a particular kind of datum. The singular value decomposition of $\mathbf{X}$ is $\mathbf{X} = \mathbf{W\Sigma V}^\mathbf{T}$, where the m × m matrix $\mathbf{W}$ is the matrix of eigenvectors of the covariance matrix $\mathbf{XX}^T$, the matrix $\mathbf{\Sigma}$ is an m × n rectangular diagonal matrix with nonnegative real numbers on the diagonal, and the n × n matrix $\mathbf{V}$ is the matrix of eigenvectors of $\mathbf{X}^T\mathbf{X}$. The PCA transformation that preserves dimensionality is then given by:

$$\mathbf{Y}^{\mathrm{T}} = \mathbf{X}^{\mathrm{T}}\mathbf{W}$$
$$= \mathbf{V}\boldsymbol{\Sigma}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W}$$
$$= \mathbf{V}\boldsymbol{\Sigma}^{\mathrm{T}}$$

V is not uniquely defined in the usual case when m < n − 1, but **Y** will usually still be uniquely defined. Since **W** is an orthogonal matrix, each row of $\mathbf{Y}^T$ is simply a linear transformation of the corresponding row of $\mathbf{X}^T$. The first column of $\mathbf{Y}^T$ is made up of the "scores" of the cases with respect to the "principal" component; the next column has the scores with respect to the "second principal" component, and so on. If we want a reduced-dimensionality representation, we can project **X** down into the reduced space defined by only the first L singular vectors, $\mathbf{W}_L$:

$$\mathbf{Y} = \mathbf{W}_L^{\mathrm{T}}\mathbf{X} = \boldsymbol{\Sigma}_L\mathbf{V}^{\mathrm{T}}$$

where $\boldsymbol{\Sigma}_L = \mathbf{I}_{L\times m}\boldsymbol{\Sigma}$ with $\mathbf{I}_{L\times m}$ the $L \times m$ rectangular identity matrix.

The matrix **W** of singular vectors of **X** is equivalently the matrix **W** of eigenvectors of the matrix of observed co variances **C** = **X** $\mathbf{X}^T$,

$$\mathbf{X}\mathbf{X}^{\mathrm{T}} = \mathbf{W}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}$$

Given a set of points in Euclidean space, the first principal component corresponds to a line that passes through the multidimensional mean and minimizes the sum of squares of the distances of the points from the line. The second principal component corresponds to the same concept after all correlation with the first principal component has been subtracted from the points. The singular values (in **Σ**) are the square roots of the eigenvalues of the matrix $\mathbf{XX}^T$. Each eigenvalue is proportional to the portion of the "variance" that is correlated with each eigenvector. The sum of all the eigenvalues is equal to the sum of the squared distances of the points from their multidimensional mean. PCA essentially rotates the set of points around their mean in order to align with the principal components. This moves as much of the variance as possible using an orthogonal transformation into the first few dimensions.PCA is often used in this manner for dimensionality reduction. PCA is sensitive to the scaling of the variables. If we have just two variables and they have the same sample variance and are positively correlated, then the PCA will entail a rotation by 45° and the "loadings" for the two variables with respect to the principal component will be equal. But if we multiply all values of the first variable by 100, then the principal component will be almost the same as that variable, with a small contribution from the other variable, whereas the second component will be almost aligned with the second original variable.

3.2 The EFA model

In multivariate statistics, exploratory factor analysis (EFA) is a statistical method used to uncover the underlying structure of a relatively large set of variables. EFA is a technique within factor analysis whose overarching goal is to identify the underlying relationships between measured variables. It is commonly used by researchers when developing a scale which is a collection of questions used to measure a particular research topic and serves to identify a set of latent constructs underlying a battery of measured variables. It is used when the researcher has no a priori hypothesis about factors or patterns of measured variables. Measured variables are any one of several attributes of people that may be observed and measured. An example of a measured variable is the physical height of a human being. The numbers of measured variables to include in the analysis are carefully considered. EFA procedures are more accurate when each factor is represented by multiple measured variables in the analysis. EFA is based on the common factor model. Within the common factor model, measured variables are expressed as a function of common factors, unique factors, and errors of measurement. Common factors influence two or more measured variables, while each unique factor influences only one measured variable and does not explain correlations among measured variables.

4. Difference between Principal Component Analysis and Exploratory Factor Analysis

Principal Components retained account for a maximal amount of variance of observed variables. Exploratory factor analysis account for common variance in the data.PCA decomposes correlation matrix while EFA decomposes adjusted correlation matrix. In PCA ones are on the diagonals of the correlation matrix while in EFA diagonals of correlation matrix are adjusted with unique factors.PCA minimizes sum of squared perpendicular distance to the component axis while EFA estimates factors which influence responses on observed variables. In PCA component scores are a linear combination of the observed variables weighted by Eigen vectors while in EFA observed variables are linear combinations of the underlying and unique factors.PCA decomposes a correlation matrix with ones on the diagonals. The amount of variance is equal to the trace of the matrix, the sum of the diagonals, or the number of observed variables in the analysis. PCA minimizes the sum of the squared perpendicular distance to the component axis. Principal components retained account for a maximal amount of variance. The component score is a linear combination of observed variables weighted by eigenvectors. Component scores are a transformation of observed variables,

(C1 = b11x1 + b12x2 + b13x3 + . . .)

The PCA Model is Y = XB

Where Y is a matrix of observed variables

X is a matrix of scores on components

B is a matrix of eigenvectors (weights)

EFA decomposes an adjusted correlation matrix. The diagonals have been adjusted for the unique factors. The amount of variance explained is equal to the trace of the matrix, the sum of the adjusted diagonals or communalities. Factors account for common variance in a data set. Squared multiple correlations (SMC) are used as communality estimates on the diagonals. Observed variables are a linear combination of the underlying and unique factors. Factors are estimated, (X1 = b1F1 + b2F2 + . . . e1 where e1 is a unique factor).

The EFA Model is Y = Xb+ E

Where Y is a matrix of measured variables

X is a matrix of common factors

b is a matrix of weights (factor loadings)

E is a matrix of unique factors, error variation

5. Principal Component Analysis Methods

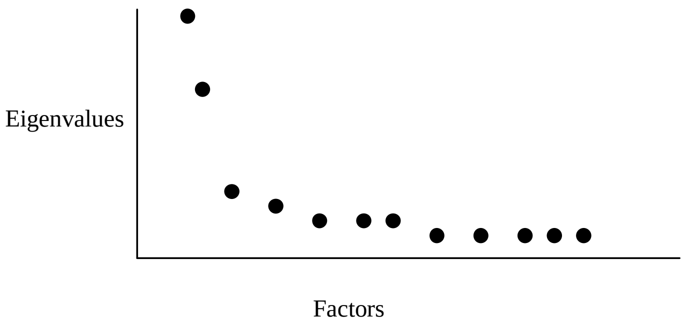
Two types of methods have been used for PCA. Firstly, there are the more conventional matrix methods, in which all the data are used to estimate the Variance-covariance structure and express it in a matrix. In practice this usually means that the matrix is diagonalized using some numerical technique such as singular value decomposition (SVD). The second type is data method since they work directly with the data. This approach is suitable for real-time applications or for very high dimensional problems where the computational expense is an important consideration. Neural networks with Hebbian learning have been proposed for adaptive PCA. Simple PCA which is a faster method that does not require learning parameters has also been developed. In matrix method the goal is to find the eigenvectors of the covariance matrix. These eigenvectors correspond to the directions of the principal components of the original data; their statistical significance is given by their corresponding eigenvalues.

6. Exploratory Factor Analysis Methods

Exploratory factor analysis (EFA) is generally used to discover the factor structure of a measure and to examine its internal reliability. EFA is often recommended when researchers have no hypotheses about the nature of the underlying factor structure of their measure. Exploratory factor analysis has three methods: (1) decide the number of factors, (2) choosing an extraction method, (3) choosing a rotation method.

6.1. Deciding the number of factors

The most common approach to deciding the number of factors is to generate a scree plot. The scree plot is a two dimensional graph with factors on the x-axis and eigenvalues on the y-axis. Eigenvalues are produced by a process called principal components analysis (PCA) and represent the variance accounted for by each underlying factor. They are not represented by percentages but scores that total to the number of items. A 12-item scale will theoretically have 12 possible underlying factors; each factor will have an eigenvalue that indicates the amount of variation in the items accounted for by each factor. If a first factor has an eigenvalue of 3.0, it accounts for 25% of the variance (3/12=.25). The total of all the eigenvalues will be 12 if there are 12 items, so some factors will have smaller eigenvalues. They are typically arranged in a scree plot in descending order like the following:



From the scree plot you can see that the first couple of factors account for most of the variance, then the remaining factors all have small eigenvalues. The term “scree” is taken from the word for the rubble at the bottom of a mountain. A researcher might examine this plot and decide there are 2 underlying factors and the remainder of factors is just “scree” or error variation. So, this approach to selecting the number of factors involves a certain amount of subjective judgment. Another approach is called the Kaiser-Guttmann rule and simply states that the number of factors is equal to the number of factors with eigenvalues greater than 1.0. We tend to recommend the scree plot approach because the Kaiser-Guttmann approach seems to produce many factors.

6.2. Factor Extraction

Once the number of factors is decided the researcher runs another factor analysis to get the loadings for each of the factors. To do this, one has to decide which mathematical solution to use to find the loadings. There are about five basic extraction methods (1) PCA, which is the default in most packages. PCA assumes there is no measurement error and is considered not to be a true exploratory factor analysis; (2) maximum likelihood (a.k.a. canonical factoring); (3) alpha factoring, (4) image factoring, (5) principal axis factoring with iterated communalities (a.k.a. least squares).The extraction method will produce factor loadings for every item on every extracted factor.

6.3. Rotation

Once an initial solution is obtained, the loadings are rotated. Rotation is a way of maximizing high loadings and minimizing low loadings so that the simplest possible structure is achieved. There are two basic types of rotation: orthogonal and oblique. Orthogonal means the factors are assumed to be uncorrelated with one another. This is the default setting in all statistical packages but is rarely a logical assumption about factors in the social sciences. Not all researchers using EFA realize that orthogonal rotations imply the assumption that they probably would not consciously make. Oblique rotation derives factor loadings based on the assumption that the factors are correlated, and this is probably most likely the case for most measures. So, oblique rotation gives the correlation between the factors in addition to the loadings.

7. Conclusion

Principal Component Analysis and Exploratory Factor Analysis are powerful statistical techniques. The techniques have similarities and differences. Principal components analysis is used to find optimal ways of combining variables into a small number of subsets, while factor analysis is 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. The difference between the two approaches is that in PCA, all of the observed variance is analyzed, while in exploratory factor analysis it is only the shared variance that is analyzed. Principal components are weighted composites of the observed variable, which is why they are properly referred to as components not factors. Factor analysis estimates the proportion of common factor variance and attempts to factor this common variance, ignoring the specific and error variance. Principal components are likely to combine specific factor variance and error variance into the components. Principal components are useful as data reduction but not for understanding the structure of the data.

8. References

[1] Baker, P. C., Keck, C. K., Mott, F. L. & Quinlan, S. V. (1993)

[2] Cattell, R. B. (1966). The scree test for the number of factors. Multivariate Behavioral Research, 1, 245-276.

[3] Child, D. (1990). The essentials of factor analysis, second edition. London: Cassel Educational Limited.

[4] Nunnally, J. C. (1978). Psychometric theory, 2nd edition. New York: McGraw-Hill.

[5] Kim, J.-O., & Mueller, C.W. (1978). Factor analysis: Statistical methods and practical Issues. Newbury Park:Sage.

[6] Gorsuch, R.L. (1990). Common factor analysis versus component analysis: Some well and little known facts. Multivariate Behavioral Research, 25, 33-39.

[7] Pearson, K. (1901). "On Lines and Planes of Closest Fit to Systems of Points in Space"(PDF). Philosophical Magazine **2** (11): 559–572.

[8] Abdi. H.& Williams, L.J. (2010). "Principal component analysis."Wiley Interdisciplinary Reviews: Computational Statistics, 2: 433–459.

[9] Shaw P.J.A. (2003) Multivariate statistics for the Environmental Sciences, Hodder-Arnold. ISBN 0-340-80763-6.

[10] Barnett, T. P., and R. Preisendorfer. (1987). "Origins and levels of monthly and seasonal Forecast skill for United States surface air temperatures determined by canonical correlation Analysis". Monthly Weather Review 115.