Principal Components Analysis

Principal components analysis (PCA) was originally a data reduction strategy to obtain a smaller set of meaningful "components" from a set of related variables. Formulated by Harold Hotelling (1933) in part to solve the problem of multiple indicators of educational ability (e.g., reading speed, arithmetic speed) to discover a single general construct (known as "general ability" or g).[1] Principal components may be used as a data reduction tool to explore the dimensionality of a set of items in a scale, and it is the initial step in exploratory factor analysis. PCA also underlies the weighted composite process of many classic multivariate methods, including MANOVA, discriminant analysis, cluster analysis, and canonical correlation (Jolliffe, 2002; Takemura, 1985).

Let's consider a hypothetical example with a correlation matrix of a set of measures of elementary school abilities.

	Reading Comprehension	Reading Fluency	Vocabulary	Math Conceptual Understanding	Math Procedural Knowledge	Math Problem Solving
Reading Comprehension	1					
Reading Fluency	.54	1				
Vocabulary	.44	.80	1			
Math Conceptual Understanding	.04	.21	.18	1		
Math Procedural Knowledge	.17	.13	.09	.49	1	
Math Problem Solving	.09	.10	.11	.72	.68	1

Notice the clusters of high correlations among the reading-related measures and high correlations among the math-related measures. This pattern suggests there may be two meaningful components of ability underlying the correlations. Principal components uses an algebraic process to identify these potentially meaningful clusters of correlations.

Eigenvalues

To quantify the degree to which a correlation matrix of a set of variables can be represented by one or more clusters of highly correlated variables, PCA derives *eigenvalues* that represent the amount of variance accounted for by each component. There are as many eigenvalues as variables, and each represents the variance of each component. If there are a few major constructs underlying the set of correlations (e.g., reading and math ability), there will be a few large eigenvalues and several small eigenvalues. In the abilities example above, we should expect two large eigenvalues and four very small eigenvalues (a total of six, because there are six variables). The eigenvalues are in the metric of the variance/covariance values of the variables involved and sum to the total of the variances of the variables analyzed, so the values do not have a very meaningful interpretation by themselves. We interpret them in terms of their size relative to the other eigenvalues. The proportion of variance accounted for by each component can be obtained, however, by dividing the eigenvalue by the total variance (i.e., the sum of the eigenvalues).

Mathematically, eigenvalues are derived by a decomposition of the variance-covariance matrix by solving a set of simultaneous equations. The matrix equation, called the characteristic equation, is solved to obtain the eigenvalues, so eigenvalues are sometimes referred to as characteristic roots.

$$(\mathbf{S} - \boldsymbol{\lambda}\mathbf{I})\,\mathbf{v} = 0$$

[1] Karl Pearson proposed the general mathematical concepts now associated with eigenvalues (characteristics roots derived from associations among a set of variables). Hotelling applied the concepts to educational measurement in correspondence with Thurstone, Spearman, Thorndike and others who developed factor analysis (Tatsuoka, 1988).

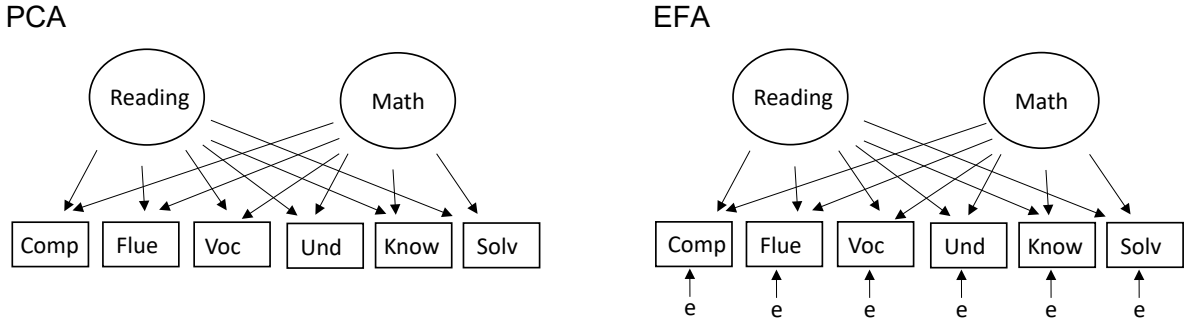
$\mathbf{S}$ is the variance-covariance matrix, $\boldsymbol{\lambda}$ contains the unknown eigenvalues, $\mathbf{I}$ is a special matrix of 1s and 0s, called an identify matrix, and $\mathbf{v}$ is the eigenvector (which are weighting values). The eigenvalues are first found by setting the quantity in parentheses equal to 0, and then those values are used to solve for the eigenvectors. In a very simple case, say with two variables, the two eigenvalues are computed from solving the quadratic equation below, which is the regular algebraic equivalent to the matrix expression in parentheses above, where $(\mathbf{S} - \boldsymbol{\lambda}\mathbf{I}) = 0$. If the known values (elements) of this simple 2 × 2 variance-covariance matrix were labeled a, b, c, and d,[2] the matrix equation could be solved for values of the unknown λ (see Tatsuoka, 1988, pp. 135-137):

$$\lambda^2 - (a+d)\lambda + ad - bc = 0$$

The values a, b, c, and d, are known, so are replaced by constant values. Assuming some standard conditions, quadratic equations can be solved to obtain two valid solutions—the two desired eigenvalues, λ_1 and λ_2. In general, we will have a polynomial of the order equal to the number of variables and that number of solutions to the polynomial equation. In other words, we will have the same number of eigenvalues as variables.

Principal Components and Factor Analysis

Although PCA is the typical first step when conducting an exploratory factor analysis (EFA) as well as the default method whenever factor analysis is requested from a statistical software program, it is not really a true factor analysis method. The loadings from PCA, which describe the strength of the relationship between the item and the component, are linear weights that account for the full variance of the variables. Another way to state this is that PCA assumes no measurement error in the relationship between the component and the items. This is the primary difference from factor analysis which assumes that the factors account for only some of the variance in the items and that there is some remaining error variance left over. Graphically we can represent the difference with the following graphic:[3]



The circles represent "components" for PCA or "factors" for EFA (many authors like to maintain this distinction, although in practice, the term "factor" is used for PCA too), and the arrows represent regression estimates of the path between the component or factor and the item, referred to as loadings. The EFA model is more in line with classical test theory ideas that any observed variable is a function of a true score and error ($\mathrm{X} = \mathrm{T} + \mathrm{E}$), so for investigating the structure of a construct and estimating the loadings, many researchers (including myself) prefer one of the "true" factor analytic methods to PCA. Although PCA and EFA may show similar results in many samples, PCA can give poor estimates of loadings in small samples (Snook & Gorsuch, 1989; see Preacher & MacCallum, 2003 for an illustrative comparison). EFA is really a family of different estimation methods, all with the same general goals,

[2] Here a, b, c, and d are used to label the variance-covariance matrix values in the fairly standard system that we have used before for labeling 2 × 2 tables, where $\mathbf{S} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$. The positive diagonal values (a and d) are the variances for the two variables and the negative diagonal (c and b) are the covariances.

[3] I have omitted a double headed arrow between the components and factors, which would represent the correlation between them. PCA derives weights based on the assumption that the components are independent, or orthogonal. Whether the factors are correlated is another issue that we will have to deal with in another course, although, in short, it is an unlikely assumption in most settings.

decision steps, and format of the results. Confirmatory factor analysis (CFA) generally has similar goals in that it is concerned with understanding the factors structure of the items and estimating the loadings, but it involves a different process. The basic assumptions about error variance for EFA are parallel to those of CFA, however, so these two methods are more related to one another than either is to PCA. For more information and references on factor analysis see the handout "A Quick Primer on Exploratory Factor Analysis" for my structural equation modeling course, http://web.pdx.edu/~newsomj/semclass/.

References

Jolliffe, I. T. (2002). *Principal component analysis and factor analysis, second edition*. New York: Springer.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology, 24*(6), 417-441.

Preacher, K. J., & MacCallum, R. C. (2003). Repairing Tom Swift's electric factor analysis machine. *Understanding statistics: Statistical issues in psychology, education, and the social sciences, 2*(1), 13-43.

Snook, S.C., & Gorsuch, R.L. (1989). Principal component analysis versus common factor analysis: A Monte Carlo study. *Psychological Bulletin, 106*, 148-154.

Tabachnick, B.G. and Fidell, L.S. (2013). *Using multivariate statistics, sixth edition*. Boston: Pearson.

Takemura, A. (1985). A principal decomposition of Hotelling's T^2 statistic. In P. R. Krishnaiah (Ed.), *Multivariate analysis VI* (pp. 538–597). New York, NY: Elsevier.

Tatsuoka, M. M. (1988). *Multivariate analysis: Techniques for educational and psychological research*. New York: Macmillan.