Principal Components Analysis vs. Factor Analysis…and Appropriate Alternatives
By Dr. Jon Starkweather, Research and Statistical Support consultant

During my academic childhood; which is a label I apply to the time when I was earning my Bachelor's degree, I was introduced to the use of Principal Components Analysis (PCA) and Factor Analysis (FA) while taking an undergraduate tests and measurement class. Later, during my academic adolescence (earning my Master's degree), I was introduced to the confusion surrounding the appropriate choice of PCA or FA under specific circumstances. During my academic young adulthood (earning my PhD), I was introduced to the use of Structural Equation Modeling for confirmatory factor analysis. Since that time, I have learned there are alternative analyses which can be used in place of traditional PCA or FA when attempting to reduce the number of variables or identify underlying structure. The purpose of this article was to hopefully clarify what are the key elements which should be considered prior to choosing between these alternatives and PCA or FA. The article also provides a brief review of two of those alternative analyses and links are provided to tutorial pages for conducting the various analyses in SPSS and the R programming language.

Principles of Measurement

In order to discuss PCA and FA, and their alternatives, we must first discuss some principles of measurement. PCA and FA grew out of early measurement and intelligence research, such as Binet and Henri (1895), Pearson (1901), and Spearman (1904) among others. The latent factor idea grew out of the combination of intelligence research, and the classical test theory of measurement (Thurston, 1947; Novak, 1966). Essentially, under classical test theory, observed score equals true score plus error. Here, true score refers to the amount of the characteristic or ability a case actually has at the time of measure. Latent simply refers to something we can not directly observe or measure. Observed variables can be measured directly, unobserved or latent variables can only be measured indirectly. Measuring a person's height is fairly direct, whether we use inches or centimeters the error associated with a measurement of a person's height is relatively low. But, when we measure characteristics or abilities which can only be measured indirectly, for instance sadness, the error associated with that measurement is likely greater than the error associated with the height measurement. Confounding the issue of direct versus indirect measurement is the issue of measurement scale.

Measurement scales were made popular by Stevens (1946, 1951, 1957). There are four measurement scales; nominal, ordinal, interval, and ratio. Nominal scale is essentially naming things with numbers; football jersey numbers – the numbers simply identify objects. Ordinal scale has the added characteristic of sequence; finishing positions of a race – the numbers identify objects and convey sequential order. Interval scale has the additional characteristic of equal intervals between units of measure; time of day or clock time – the numbers identify objects or points, convey sequence, and there are equal intervals between the units (e.g. the interval between 1 o'clock and 2 o'clock is the same as the interval between 4 o'clock and 5 o'clock [accept on Fridays]). Ratio scale has the additional characteristic of a true zero point; pounds of weight or Kelvin temperature – an object cannot weigh negative 120 pounds. The 'accept on Fridays' comment above is an important one because it highlights another issue in measurement; the issue of objective versus subjective measurement. The clock is an example of objective measurement while our perception tends to be much more subjective. We tend to perceive the interval between 4 o'clock and 5 o'clock as greater than the clock measures it. The clock has no variance (if it is working properly) when measuring the interval between 4 and 5 p.m. on multiple days. Our judgments of the interval between 4 and 5 p.m. on Fridays tends to be different (or vary) when compared with our judgment of the interval between 4 and

5 p.m. on other days. It is generally considered best if one measures directly, objectively, and with instruments that provide interval or ratio scaled variables. Often, practical considerations prevent this best case scenario and the data then constrains which analysis should be run. These principles of measurement should be carefully considered prior to choosing any analysis and especially when considering the choice between PCA, FA, and their alternatives.

PCA and FA

PCA is a variable reduction technique which maximizes the amount of variance accounted for in the observed variables by a smaller group of variables called *components*. As an example, consider the following situation. Let's say, we have 500 questions on a survey we designed to measure persistence. We want to reduce the number of questions so that it does not take several hours to complete the survey. It would be appropriate to use PCA to reduce the number of questions by identifying and removing redundant questions. For instance, if question 122 and question 356 are virtually identical (i.e. they ask the exact same thing but in different ways), then one of them is not necessary. The PCA process allows us to reduce the number of questions or variables down to their principal components while maximizing the amount of variance in those variables accounted for by the principal components. A key to understanding PCA is recognizing the components as groups of variables (questions, items, etc.) which were the inputs of the PCA. The components are not latent factors. PCA is not a model based technique and involves no hypothesis about the substantive meaning of or relationships between latent factors. Rotation strategies, which focus on the relationship between components, can be applied to components to aid interpretation, but these components are not the same as latent factors. PCA is occasionally, but very confusingly, called exploratory factor analysis (EFA). The use of the word *factor* in EFA is inappropriate and confusing in this context because, we are really interested in components and variable reduction, not factors. This issue is made more confusing by some software packages (e.g. PASW/SPSS & SAS) which list or use PCA under the heading factor analysis.

FA is typically used to identify or confirm the latent factor structure for a group of measured variables. Latent factors are unobserved variables which typically can not be directly measured; but, they are assumed to *cause* the scores we observe on the measured or indicator variables. FA is also used to reduce the number of variables which can reasonably measure or convey the latent factor structure. FA is a model based technique. It is concerned with modeling the relationships between measured variables, latent factors, and error. Therefore, because of the recognition of error; FA is typically more consistent across samples (i.e. the results tend to be more generalizable and replicable than PCA). The ability of FA to recognize unique item variance (sometimes referred to as item error variance) is a key in distinguishing it from PCA – which considers all variance equally and attempts to account for as much of it as possible without regard to types of variance. FA relies on assumptions of linearity, multivariate normality, and homoscedasticity.

Both PCA and FA take as input a correlation or covariance matrix. Both PCA and FA can be more easily interpreted with the application of a rotation strategy (e.g. varimax or oblimin). PCA and FA tend to show similar results when performed on a single data set, but they are not interchangeable. As stated in O'Rourke, Hatcher, and Stepanski (2005):

> "Both (PCA & FA) are methods that can be used to identify groups of observed variables that tend to hang together empirically. Both procedures can also be performed with the SAS FACTOR procedure and they generally tend to provide similar results. Nonetheless, there are some important conceptual differences between principal

> component analysis and factor analysis that should be understood at the outset. Perhaps the most important deals with the assumption of an underlying causal structure. Factor analysis assumes that the covariation in the observed variables is due to the presence of one or more latent variables (factors) that exert causal influence on these observed variables" (p. 436).

Both PCA and FA can be used as exploratory analysis. But; PCA is predominantly used in an exploratory fashion and almost never used in a confirmatory fashion; because it is primarily suited for data reduction. FA can be used in an exploratory fashion or a confirmatory fashion because; it is primarily suited for identifying and/or confirming factor structure. In both scenarios, the focus is on identifying the variables/items which load on the factors well. The choice of which is used (PCA or FA) should be driven by the goals of the analyst. If you are interested in reducing the observed variables down to their principal components while maximizing the variance accounted for in the variables by the components, then you should be using PCA. If you are concerned with modeling the latent factors which cause the scores on your observed variables, then you should be using FA.

Some Alternatives

Categorical principal components analysis (CATPCA) is appropriate for data reduction when variables are categorical (e.g. nominal or ordinal) and the researcher is concerned with identifying the underlying components of a set of variables (items, survey questions, etc.) while maximizing the amount of variance accounted for in those items (by the principal components). The primary benefit of using CATPCA is that it derives weights from the input data that produce optimal linear relationships in the output data. CATPCA does not assume linear relationships among numeric (interval or ratio) data nor does it require assuming multivariate normal data. Furthermore, the optimal scaling used in SPSS during the CATPCA analysis allows the researcher to specify which level of measurement he or she wants to maintain (e.g. nominal, ordinal, interval/ratio, spline-nominal, & spline-ordinal) in the optimally scaled variables. The R programming language also has various packages (e.g. polycor) and functions (e.g. hetcor) which can be used to create a matrix of different types of correlations for different measurement scales for each pair of variables in a data set. Some, such as those in the aspect package, will also do optimal scaling. The resulting correlation matrix can be passed to a PCA function which will result in less biased results than simply using Pearson correlations for all types of variables. For example, the hetcor function “computes a heterogenous correlation matrix, consisting of Pearson product-moment correlations between numeric variables, polyserial correlations between numeric and ordinal variables, and polychoric correlations between ordinal variables” (Fox, 2010, p. 2). The resulting matrix from a hetcor function can then be passed to a PCA function. But, the hetcor function is not the only function which can be used to create specific types of correlations and the hetcor function does not offer an alternative for strictly nominal variables.

Correspondence Analysis is appropriate when attempting to determine the proximal relationships among two or more categorical variables. Correspondence analysis is also available in the R programming language using a variety of packages and functions (e.g. ca package contains the ca function – for correspondence analysis). Using correspondence analysis with categorical variables is analogous to using correlation analysis and principal components analysis for continuous or nearly continuous variables. They provide the researcher with insight as to the relationships among variables and the dimensions or eigenvectors underlying them. A key part of correspondence analysis is the multi-dimensional map produced as part of the output. The correspondence map allows researchers to visualize the relationships among categories by plotting them in a spatially accurate way on dimensional axes; in other words, which categories are close to other categories on empirically derived

dimensions. Correspondence analysis is nonparametric and does not offer a statistical significance test because; it is not based on a distribution or distributional assumption (Garson, 2010). Comparison of different models (e.g. different variables entered/removed) should be done with categorical or logistic regression. Again, correspondence analysis requires categorical variables only. Correspondence analysis accepts nominal variables, ordinal variables, and/or discretized interval - ratio variables (e.g. quartiles), although creating discrete categories from a continuous variable is generally discouraged.

Recommendations

The choice of which analysis to use should be evaluated by the researcher with strong emphasis on what the a-priori goals of the study were, what type of data has been collected, and what properties the data displays during initial data analysis. Close attention should be paid to scatter plot matrices of all the variables (or items) with a keen eye for linearity.

If the goal was to reduce the number of variables while maximizing the total variance accounted for, then traditional PCA is appropriate. If the data is nominal or ordinal, then CATPCA is appropriate. CATPCA can be implemented using SPSS CATPCA with optimal scaling. Or, if one uses the R programming language, there are several packages available (e.g. aspect, homals, polycor) which contain functions for optimally scaling and/or correlating differently scaled variables. Each can produce an appropriate correlation matrix on which to conduct the PCA in R.

If the goal was to document or confirm latent factor structure, while accounting for measurement error when the data are interval or ratio, homoscedastic, and multivariate normal; then FA is appropriate. If those assumptions are not met, then a variety of alternatives are available. One can use the hetcor function (or some other function) in R to do the FA. Or, one could conduct multiple correspondence analysis in SPSS or in the R programming language (using the ca package). Or, one could conduct joint correspondence analysis in R (using the ca package). There are likely other options available, especially in the R programming language. But the general message of this article is that the researcher should not feel constrained to one analysis without considering alternatives which may be better suited to the data.

Until next time, *I get by with a little help from my friends*.

References & Resources

Bartholomew, D. J. (2007). Three faces of factor analysis. In R. Cudec & R. MacCallum (Eds.), *Factor analysis at 100: Historical developments and future directions* (pp. 9 – 22). Mahwah, NJ: Lawrence Erlbaum Associates.

Bentler, P. M., & Kano, Y. (1990). On the equivalence of factors and components. *Multivariate Behavioral Research, 25,* 67 – 74.

Binet, A., & Henri, V. (1895). La psychologie individuelle. *Annee Psychologique, 2,* 411 – 463.

Clausen, S.-E. (1998). Applied correspondence analysis: An introduction. Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-121. Thousand Oaks, CA: Sage.

Fox, J. (2010). “Polychoric and polyserial correlations” from *Package ‘polycor’*. Retrieved July, 14, 2010 from http://cran.r-project.org/web/packages/polycor/polycor.pdf

Garson, D. (2010). "Correspondence Analysis", from *Statnotes: Topics in Multivariate Analysis.* Retrieved July 14, 2010 from http://faculty.chass.ncsu.edu/garson/pa765/statnote.htm

Greenacre, M. (2007). Interdisciplinary statistics: Correspondence analysis in practice (2nd ed.). Boca Raton, FL: Taylor & Francis Group, LLC.

Le Roux, B., & Rouanet, H. (2010). Multiple correspondence analysis. Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-163. Thousand Oaks, CA: Sage.

Meulman, J. J., & Heiser, W. J. (2009). PASW Categories 18. Chicago, IL: SPSS, Inc.

Meulman, J. J., Van Der Kooij, A. J., & Heiser, W. J. (2004). Principal components analysis with nonlinear optimal scaling transformations for ordinal and nominal data. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 43 – 72). Thousand Oaks, CA: Sage Publications, Inc.

Novick, M. R. (1966) The axioms and principal results of classical test theory. *Journal of Mathematical Psychology, 3,* 1 – 18.

O'Rourke, N., Hatcher, L., & Stepanski, E.J. (2005). *A step-by-step approach to using SAS for univariate and multivariate statistics,* Second Edition. Cary, NC: SAS Institute Inc.

Pearson, K. (1901). On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine, 2,* 559–572.

Spearman, C. (1904). General intelligence objectively determined and measured. *American Journal of Psychology, 15,* 201–293.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103,* 677 – 680.

Stevens, S. S. (1951). Mathematics, measurement and psychophysics. In S.S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1 – 49). New York: Wiley.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review, 64,* 153 – 181.

Thurstone, L. L. (1947). *Multiple factor analysis.* Chicago, IL: University of Chicago Press.