

# Exploratory Factor Analysis

Brian Habing - University of South Carolina – Last Updated: October 4, 2005

*FA is not worth the time necessary to understand it and carry it out.*  
-Hills, 1977

*Factor analysis should not be used in most practical situations.*  
-Chatfield and Collins, 1980, pg. 89.

*At the present time, factor analysis still maintains the flavor of an art,  
and no single strategy should yet be "chiseled into stone".*  
-Johnson and Wichern, 2002, pg. 517.

38 - The number of articles *ISI Web of Science* shows with the topic  
"Factor Analysis" for the week ending October 1, 2005.

1213 - The number of FA articles in 2005 through October 1<sup>st</sup>.

## **References:**

- Chatfield, C. & Collins, A. J. (1980). *Introduction to Multivariate Analysis*. London: Chapman and Hall.
- Hair, J.F. Jr., Anderson, R.E., Tatham, R.L., & Black, W.C. (1998). *Multivariate Data Analysis, (5<sup>th</sup> Edition)*. Upper Saddle River, NJ: Prentice Hall.
- Hatcher, L. (1994). *A Step-by-Step Approach to Using the SAS System for Factor Analysis and Structural Equation Modeling*. Cary, NC: SAS Institute Inc.
- Hills, M. (1977). Book Review, *Applied Statistics*, 26, 339-340.
- Johnson, D.E. (1998). *Applied Multivariate Methods for Data Analysis*. Pacific Grove, CA: Brooks/Cole Publishing.
- Johnson, R.A. & Wichern, D.W. (2002). *Applied Multivariate Statistical Analyses (5<sup>th</sup> Edition)*. Upper Saddle River, NJ: Prentice Hall.
- Mardia, K.V., Kent, J.T., & Bibby, J.M. (1979). *Multivariate Analysis*. New York: Academic Press.
- Sharma, S. (1996). *Applied Multivariate Techniques*. United States: John Wiley & Sons.
- Stevens, J. (2002). *Applied Multivariate Statistics for the Social Sciences (4<sup>th</sup> Edition)*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Thompson, B. (2000). Q-Technique Factor Analysis: One Variation on the Two-Mode Factor Analysis of Variables. In Grimm, L.G. & Yarnold, P., *Reading and Understanding More Multivariate Statistics*. Washington, D.C.: American Psychological Association.

**Goal:** To explain the variance in the observed variables in terms of underlying latent factors.

**The Model:** The following assumes that the  $p$  observed variables (the  $X_i$ ) that have been measured for each of the  $n$  subjects have been standardized.

$$\begin{aligned} X_1 &= \lambda_{11}F_1 + \dots + \lambda_{1k}F_k + u_1 \\ X_2 &= \lambda_{21}F_1 + \dots + \lambda_{2k}F_k + u_2 \\ &\vdots \\ X_q &= \lambda_{q1}F_1 + \dots + \lambda_{qk}F_k + u_q \end{aligned}$$

The  $F_j$  are the  $k$  common factors, the  $u_i$  are the  $q$  unique (or specific) errors, and the  $\lambda_{ij}$  are the  $q \times k$  factor loadings. The  $F_j$  have mean zero and standard deviation one, and are generally assumed to be independent. (We will assume this orthogonality below, but it is not true for oblique rotations.) The  $u_i$  are also independent and the  $F_j$  and  $u_i$  are mutually independent of each other.

In matrix form this can be written as:

$$X_{q \times 1} = \Lambda_{q \times k} F_{k \times 1} + u_{q \times 1}$$

which is equivalent to

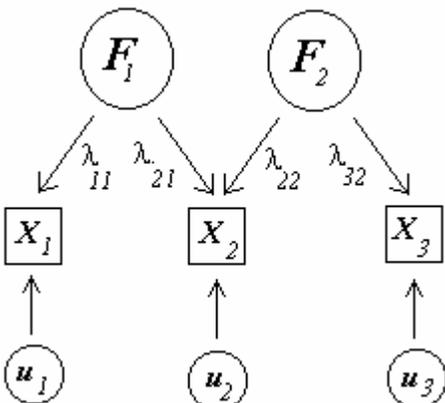
$$\Sigma = \Lambda \Lambda^T + \text{cov}(u)$$

where  $\Sigma_{q \times q}$  is the correlation matrix of  $X_{q \times 1}$ . Since the errors are assumed to be independent,  $\text{cov}(u)$  should be a  $q \times q$  diagonal matrix. This implies that:

$$\text{Var}(X_i) = \sum_{j=1}^k \lambda_{ij}^2 + \text{Var}(u_i)$$

The sum of  $X_i$ 's squared factor loadings is called its communality (the variance it has in common with the other variables through the common factors). The  $i^{\text{th}}$  error variance is called the uniqueness or specificity of  $X_i$  (the variance that is specific to variable  $i$ ).

**Path Diagram:** Factor Analytic models are often represented by path diagrams. Each latent variable is represented by a circle, and each manifest variable is represented by a square. An arrow indicates causality... which can get to be a pretty complicated subject in some models!



The diagram at the left is for a 2-factor model of 3 variables ( $k=2, q=3$ ), where the first variable loads only on factor 1 ( $\lambda_{12}=0$ ), the second variable loads on both factors, and the third variable loads only on factor 2 ( $\lambda_{31}=0$ ).

**Types of Factor Analysis:** Some authors refer to several different types of factor analysis, such as R-Factor Analysis or Q-Factor Analysis. These simply refer to what is serving as the variables (the columns of the data set) and what is serving as the observations (the rows).

	Columns (what the factors explain)	Rows (measured by the columns)
R	variables	participants
Q	participants	variables
O	occasions	variables
P	variables	occasions
T	occasions	participants
S	participants	occasions

Thompson (2000, pg. 211)

What we have been doing so far has been R factor analysis, we have been looking for latent factors that lie behind the variables. This would, for example, let us group different test questions that seem to be measuring the same underlying construct. In Q factor analysis we would be looking for factors that seem to be underlying the examinees and be asking how many types of people are there. The math is the same, but the terminology and goals can be different. Everything we have below refers to R factor analysis.

### **Is the data appropriate?**

**Correlation** - It doesn't make sense to use factor analysis if the different variables are unrelated (why model common factors if they have nothing in common?)

Rule of thumb: A substantial number of correlations are  $> 0.3$ .

Some other methods also include: Bartlett's test of sphericity, the measuring of sampling adequacy (MSA), the anti-image correlation matrix, and the Kaiser-Meyer-Olkin measure (KMO). I haven't seen that much is gained by using these others.

**Multivariate Normality** – In order to use maximum likelihood estimation or to perform any of the tests of hypotheses, it is necessary for the data to be multivariate normal. This is not required for fitting the model using principal components or principal factor methods.

**Sample Size:** A general rule of thumb is that you should have at least 50 observations and at least 5 times as many observations as variables.

Stevens (2002, pg. 395) summarizes some results that are a bit more specific and backed by simulations. The number of observations required for factors to be reliable depend on the data. In particular on how well the variables load on the different factors.

A factor is reliable if it has:

3 or more variables with loadings of 0.8 and any  $n$

4 or more variables with loadings of 0.6 and any  $n$

10 or more variables with loadings of 0.4 and  $n \geq 150$

Factors with only a few loadings require  $n \geq 300$

Obviously this doesn't cover every case, but it does give some guidance.

**How many factors?** There are several rules for determining how many factors are appropriate for your data.

Mardia, Kent, and Bibby (1979, pg. 258) point out that there is a limit to how many factors you can have and actually end up with a model that is simpler than what your raw data. The quantity  $s$  is the difference between the number of unique values in your data's  $q \times q$  correlation matrix and the number of parameters in the  $k$  factor model:

$$s = \frac{1}{2}(q - k)^2 - \frac{1}{2}(q + k)$$

It only makes sense to perform a factor analysis if  $s > 0$ , and some programs will not let you estimate the factor analytic model if it is not true. Even though you could always exactly fit a 5 factor model to 5 variables using principal components and no specificity,  $s > 0$  only for one or two factors. The minimum number of variables required for different numbers of factors are:

# Factors	2	3	4	5	6
Variables Required	5	7	8	9	11

In general this is not something we worry about too much since we usually want to have a much smaller number of factors than variables. Also, recall that we want several variables loading on each factor before we can actually trust that factor to be meaningful anyway.

***Kaiser's Criterion / Eigen Value > 1*** – Take as many factors as there are eigenvalues  $> 1$  for the correlation matrix. Hair, et.al. (1998, pg. 103) reports that this rule is good if there are 20 to 50 variables, but it tends to take too few if there are  $< 20$  variables, and to many if there are  $> 50$ . Stevens (2002, pg. 389) reports that it tends to take too many if there are  $> 40$  variables and their communalities are around 0.4. It tends to be accurate with 10-30 variables and their communalities are around 0.7.

***Scree Plot*** – Take the number of factors corresponding to the last eigenvalue before they start to level off. Hair, et.al. (1998, pg. 104) reports that it tends to keep one or more factors more than Kaiser's criterion. Stevens (2002, pg. 390) reports that both Kaiser and Scree are accurate if  $n > 250$  and communalities  $\geq 0.6$ .

***Fixed % of Variance Explained*** – Keep as many factors as are required to explain 60%, 70%, 80-85%, or 95%. There is no general consensus and one should check what is common in your field. It seems reasonable that any decent model should have at least 50% of the variance in the variables explained by the common factors.

***A priori*** – If you have a hypothesis about the number of factors that should underlie the data, then that is probably a good (at least minimum) number to use.

In practice there is no single best rule to use and a combination of them is often used, so if you have no a priori hypothesis check all three and use the closest thing to a majority decision.

There are also a variety of other methods out there that are very popular with some authors: Minimum average partial correlation, parallel analysis, and modified parallel analysis are three of them. Three that require multivariate normality are: likelihood ratio test, AIC (Akaike's information criterion), and SBC (Schwartz's Bayesian criterion).

## **Methods for Fitting the Factor Analysis Model:**

**Principal Components Factor Analysis** – Just take the first  $m$  loadings from the principal components solution and multiply by the square root of the corresponding eigen value. This is not really appropriate since it attempts to explain all of the variance in the variables and not just the common variance. It therefore will often have highly correlated errors. However, Hair, et.al. (1998, pg. 103) reports that it often gives similar results to other methods if there are  $\geq 30$  variables or if most variables have communalities  $> 0.6$ .

**Principal Factor Factor Analysis** – (a.k.a. Principal Axis Factoring and sometimes even Principal Components Factoring!) Come up with initial estimates of the communality for each variable and replace the diagonals in the correlation matrix with those. Then do principal components and take the first  $m$  loadings. Because you have taken out the specificity the error matrix should be much closer to a diagonal matrix. There are various initial estimates used for the initial communalities: the absolute value of the maximum correlation of that variable with any of the others, the squared multiple correlation coefficient for predicting that variable from the others in multiple regression, and the corresponding diagonal element from the inverse of the correlation matrix. There seems to be no agreement on which is best... but the first is a slight bit easier to program.

**Iterated Principal Factor Factor Analysis** - This begins the same way as the principal factor method, but you use the fitted model to get better estimates of the communalities, and then you keep repeating the process. This method will often fail to fit because of the Heywood case (estimated communalities  $> 1$ , negative error variances) though.

**Maximum Likelihood Factor Analysis** – Requires the assumption of multivariate normality and is difficult to program. It does allow for various tests of hypotheses however.

**Other Methods** – Alpha Factoring, Image Factoring, Harris Factoring, Rao's Canonical Factoring, Unweighted Least Squares

I would recommend using the iterated principal factor solution if it does not have impossible communalities or error variances. If the iterated method fails then I would use the non-iterated principal factor method if the error covariances do not seem unreasonably large. If that too fails you either need more factors, or the factor analytic model is possibly inappropriate for your data set.

**Rotation:** It is often considered best to use an orthogonal rotation because then all of the mathematical representations we've used so far will work. Two of the most popular are:

**Varimax** – Maximizes the sum of the variances of the squared factor loadings within the columns. This tends to force each variable to load highly on as few factors as possible. Ideally it will cause each variable to load on only one factor and thus point out good summed scores that could be made. If there is an overall underlying general factor that is working on all of the variables this rotation will not find it. (e.g. The math test has geometry, trigonometry, and algebra... but does it also have an overall general math ability too?)

**Quartimax** – Whereas varimax focuses on the columns, quartimax focuses on the rows. Sharma (1996, pg. 120) reports that: it tends to have all of the variables load highly on one factor, and then each variable will tend to load highly on one other factor and near zero on the rest; unlike varimax this method will tend to keep the general factor. It sometimes doesn't seem to differ much from the initial solution.

There are also oblique rotations that don't keep the independence between the factors. Popular ones include Promax, Oblimin, and Orthoblique. Interpret these with caution!

**Interpretation:** Once you have your factor loadings matrix it is necessary to try and interpret the factors. It is common to indicate which of the loadings are actually significant by underlining or circling them (and possibly erasing the non-significant ones). Significant is measured in two ways.

**Practical Significance** - Are the factor loadings large enough so that the factors actually have a meaningful effect on the variables. (e.g. Roughly speaking, if two things only have a correlation of 0.2 it means that the one only explains 4% of the variation in the other!) Hair et.al. (1998, pg. 111) recommends the following guidelines for practical significance:

- ±0.3 Minimal
- ±0.4 More Important
- ±0.5 Practically Significant

**Statistical Significance** - We also want the loading to be statistically significantly different from zero. (e.g. If the loading is 0.3, but the confidence interval for it is -0.2 to 0.8 do we care?)

Stevens (2003, pg. 294) reports the following rules of thumb based on sample size

<i>n</i>	50	100	200	300	600	1000
loading	0.722	0.512	0.384	0.298	0.210	0.162

**Further Guidance** – Johnson (1996, pg. 156) recommends not including factors with only one significant loading. Hatcher (1994, pg. 73) reports that we want at least 3 variables loading on each factor and preferably more. This is also related to a result discussed earlier in the section on sample size. Finally, it seems reasonable if you are using varimax to remove variables which load heavily on more than one factor.

**Factor Scores:** Exploratory factor analysis helps us to gain an understanding of the processes underlying our variables. We might also want to come up with estimates for how each of the observations rates on these unobservable (latent) factors.

**Surrogate Variable** – Choose a single variable that loads highly on that factor and use its value. It is very simple to calculate and simple to explain... too simple...

**Estimate Factor Scores** – Use a statistical method to actually estimate the values of the  $F_j$  for each observation. Mardia, et.al. (1979, pg. 274) reports that if multivariate normality holds then Bartlett's weighted least squares method is unbiased, and Thompson's regression method will be biased but have a smaller overall error. Factor scores however are difficult to interpret (what does it mean to take a complicated weighted average of the observed values?) and the exact values estimated depend on the sample.

**Summed Scores** – If each variable loads on only a single factor, then make subscores for each factor by summing up all of the variables that load on that factor. This is the compromise solution and is often the one used in practice when designing and analyzing questionnaires.

**Some Final Steps:** Johnson and Wichern (2002, pg. 517) suggest the following to see if your solution seems reasonable: plot the factor scores against each other to look for suspicious observations and for large data sets, split them in half and perform factor analysis on each half to see if the solution is stable.