A Survey in Factor Analysis

Molly Folkert

Biola University

Readings in Mathematics

May 12, 2014

Abstract

This paper contains information on *concepts* and *considerations* in applying factor analysis to data. It gives a *theoretical model* for it and a list of the steps needed in factor analysis, giving a basic outline for each decision in the steps. It discusses *confirmatory* versus *exploratory* factor analysis, gives three methods for *determining the number of factors*, and two methods for *factoring* the data. It then discusses the basic categories for *rotational methods* in factor analysis: *orthogonal* and *oblique* rotations. Finally, it applies factor analysis to a survey conducted in 2013 at Biola University with a discussion of the results and implications.

1. Introduction

Factor analysis is a statistical method of analyzing data collected by a priori means through intercorrelations to categorize the variations into a smaller number of categories (Fruchter, 1954). In essence, this process looks at trends in data, analyzes the strength, also known as the correlations of the patterns, and groups certain areas together based on their strengths. Once the groups have been finalized, they are inspected to see if each group contains some common characteristics, which could explain the differences in responses or results in the data. The explanations, called factors, are what can be concluded as influencing the variation seen in the data.

Many fields use factor analysis, adding possible explanations to phenomena in specific areas of their particular field. In this paper, I seek to provide fundamental concepts and cautions to factor analysis in section two, and then explain the theoretical concepts in section three. I discuss the steps in applying factor analysis in section four, and then use this method on a set of data recently collected at Biola University in section five.

2. A Brief History and Important Knowledge

Factor analysis surfaced when Charles Spearman (1863-1945) presented the idea in 1904. Spearman observed that tests of abilities tend to have positive intercorrelations, so he continued to analyze with this in mind. He discovered that in a group of four tests, each test had multiple factors influencing results, not just one. This led to practice with intelligence tests in psychology, one of the main fields that use factor analysis. There are many possible factors that influence the measured level of intelligence, such as arithmetic ability and reading comprehension. The social sciences employ this analysis often, seeking to explain behaviors, personality characteristics, and opinions as results of varying levels of influence from separate areas. For instance, in sociology, survey responses can be affected by socioeconomic status or political affiliation.

The results of factor analysis should only be interpreted with caution. First, factor analysis should not be used as a definitive set of explanations. Each set of data comes from

different circumstances and therefore it cannot be said that one result from a single factor analysis is applicable under all conditions. Secondly, before factors can be applied to many situations, there should be multiple studies to support the claim. A single example is not enough to draw concrete conclusions. Lastly, though factors from the analysis can be seen as correlated, they cannot be said to cause results to vary one way or another. A common misconception is that correlation implies that one factor causes another factor to behave a certain way – this is not the case! While correlation gives strong evidence for a relationship, each of the factors could be the result of another cause not identified in the particular study.

There are terms that arise in factor analysis that anyone interpreting results should know. The main concept to understand is *correlation*. Correlation is a "measure of the strength of a certain type of relationship between two measurement variables" (Utts, 2005). Correlation essentially refers to the tendency of two numerical variables to relate with each other. The correlation coefficient can range from -1.0 to 1.0, with any number greater than 0 implying that as one variable increases, the other increases with it, and a negative number implying that as one variable increases, the other decreases.

Another crucial term to understand is *factor*, which in this context refers to a possible explanation for variation in data. The emphasis of factor analysis is to group related variables together under a common factor. *Variance* refers to deviation from one point to another under the same variable of observation, such as different responses to a survey question. The more deviation there is within the data, the larger the variance will be. An *eigenvalue of a factor* is a "measure of how much of the variance of the observed values a factor explains" (Rahn, 2013). An eigenvalue greater than 1.0 means that it explains more variance than just a single variable would. Eigenvalues come from $det(A - \lambda I_n) = 0$, where λ is the eigenvalue that solves the equation, A is the matrix of data, and I_n is the identity matrix.

3. Theoretical Model for Factor Analysis

The theoretical model for factor analysis is as follows:

$$x = \Lambda f + e$$

where x is a column vector of *n* variables, *f* is a vector of *k* common factors, *e* is a vector of *p* residuals, and Λ is a $p \times k$ matrix of factor loadings. This model can also be summarized as the following:

$$Y_1 = \alpha_{1,1}F_1 + \alpha_{1,2}F_2 + \cdots + \alpha_{1,m}F_m + e_1$$
$$Y_2 = \alpha_{2,1}F_1 + \alpha_{2,2}F_2 + \cdots + \alpha_{2,m}F_m + e_2$$
$$\vdots$$
$$Y_n = \alpha_{n,1}F_1 + \alpha_{n,2}F_2 + \cdots + \alpha_{n,m}F_m + e_n$$

where Y_n is the variable being explained by the factors, which is the individual components *x* in the first model. $\alpha_{i,j}$ for i=1,…,n and j=1,…,m is the factor loading for the factor *m* and variable *n*, given by Λ in the first model. Next, F_m is a function of the factor, where the collection of factors is stored in *f* from the first model. Finally, e_n is the error, or the residual for that variable, given by *e* in the first model. It is important to understand that F_m is a function, and not a constant. Each individual subject or person giving unique data has different characteristics that influence each factor, so factors need to be able to vary between subjects.

Each variable is explained by a linear combination of the factor loadings multiplied by the pull, or influence of that factor based on each individual subject. For example, consider a survey that asked eight questions and used the Likert Scale to measure responses. Each question gave a statement for the subject to think about, then decide whether they strongly disagree, disagree, are neutral, agree, or strongly agree with it. Depending on the subject's opinions and experiences, different weights for each factor would impact their answer to the question.

4. Steps in Performing Factor Analysis

In factor analysis, there are multiple steps to reaching the final product of the analysis. One component of the work is to decide how to explore the data. The next step is determining the number of factors. Then the choice between the types of factoring depends on the goal of the analysis and the intended usage of the results. The final major aspect of factor analysis to consider is the method of rotation to fit the data to its most meaningful position. Each of these steps has a different process depending on the chosen method of execution.

4.1. Exploratory vs. Confirmatory Factor Analysis

Factor analysis can be used in a two main ways – more commonly, exploring possible factors and situations in which the variation can be explained, called "exploratory factor analysis", or confirming or repudiating proposed factors and the way they group, called "confirmatory factor analysis". Sometimes, after much research in a particular area is done, scientists are able to hypothesize about relationships that might be expected which allows for confirmatory factor analysis, but this is not usually the case since most areas do not have enough prior research.

Exploratory factor analysis follows similar steps as confirmatory, but does not require hypothesis testing to validate hypotheses. Both methods require calculating correlations between data, choosing the number of factors to describe the variations, applying a specific method of factoring to the data, and rotating the data to fit a set of reference or principal axes to accommodate as much variation as possible.

4.2. Determining the Number of Factors

The next step in factor analysis requires the researcher to decide the number of factors to extract. There are multiple ways of deciding how many factors to use, both subjective and objective. One method is to make an educated guess on the number of factors to use. In confirmatory analysis, this would be more useful since there will be certain factors expected, but exploratory does not leave much room for this method.

The Kaiser criterion is a rule for choosing the number of factors. It first calculates

eigenvalues for a large number of factors, usually more than 15 if possible. The eigenvalue of a factor is found by the following:

$$\sum_{i=1}^{n}(\alpha_{i,1})^2$$

which squares all the factor loadings for that particular factor and adding them up. Then, it lists the factors in descending order of eigenvalue. All eigenvalues under 1.0 are dropped, since the factors do not explain more variation than a single variable would. Since all measurements are subject to error, factors with eigenvalues less than 1.0 can be kept or ones with eigenvalues greater than 1.0 can be dropped, at the researcher's discretion.

Variance-explained criteria take into account proportion of variation explained. While eigenvalues and the Kaiser criterion are valuable for finding heavily influential factors, sometimes the number of factors may only account for 60% of the data's variation, or even less. The variance-explained criterion bases its decision on the proportion of variance explained by factors rather than the degree of influence each factor has. With each factor, the cumulative proportion of variance explained is calculated and once that proportion reaches .90, the researcher chooses that many factors.

Scree plots are more subjective than the previous two methods for determining the number of factors to hold. A scree plot uses eigenvalues like the Kaiser criterion does, but does not have a default cutoff point. Instead, each eigenvalue is plotted on the y-axis of a graph with its rank on the x-axis, and a straight line is drawn connecting each consecutive point. Then, the researcher examines the graph for an "elbow shape", where the lines stop decreasing drastically and become more negatively linear. Wherever the researcher decides the points begin straighten is the cutoff point for factors.

4.3. Factoring Methods

Once the number of factors is chosen, the next step is to choose the method of factoring. There are multiple means of factoring, such as Principal Axes factoring and Maximum-Likelihood analysis. Each factoring method has a separate set of criteria for choosing the factors

and depend heavily on the analyst's intention for accounting for the variation within the data.

The most common form of factoring is Principal Axes Factoring, which seeks to weight the factors so that they extract the maximum variance and give the smallest possible residuals (Fruchter, 1954). This process usually starts out with a similar matrix from a method called principal component factoring, which has 1.0's down the main diagonal, since in theory, a factor perfectly correlates with itself. In principal axes factoring, the communality coefficients,

$$h^2 = \sum_{j=1}^{m} (\alpha_{1,j})^2$$

which is are the sum of the squares of all the factor loadings on a variable, are calculated. They then replace the 1.0's on the diagonal, which allows room for measurement error. The process is repeated with the replaced numbers, recalculating the communality coefficients and substituting them in the diagonal until they stabilize. Then, the factor loadings in the stabilized communality coefficient matrix are used for grouping variables.

Another method for factoring is the Maximum-Likelihood Method. This method comes from the maximum-likelihood estimator in statistics. The maximum-likelihood estimator is used in statistics to calculate a guess for a parameter of a set of data. The Maximum-Likelihood Method executes a similar process, estimating the most likely correlation for a factor against the variables. This provides a stronger estimation for factors in the population rather than just the sample. The Maximum Likelihood Method has the same desirable traits as the maximum-likelihood estimator, such as consistency, sufficiency, and efficiency (Heeler, 1977).

4.4. Rotational Methods

Rotation methods are another important concept to understand in factor analysis. Rotation methods, either orthogonal or oblique, involve "moving the factor axes measuring the locations of the measured variables in the factor space so that the nature of the underlying constructs become more obvious to the researcher" (Thompson, 2004). Rotation comes from the notion that the reference axes are chosen in an arbitrary manner. The axes may be rotated in any way, as

long as the data points, or the correlations of variables, are not moved. This would allow axes to be set around clusters of correlation points to find the least residuals and easily identify which factors influence certain variables.

Orthogonal axes are one of the two categories of methods of rotation. Orthogonal vectors u and v have an angle of 90° between them, where their dot product, $u \cdot v = 0$. In factor analysis, axes with an angle of 90° between them are *independent* of each other. This means that the factors have no obvious relation to each other with their influence. For example, analysis could result in two factors, political affiliation and opinion on the importance of family, which have no consistent interaction between them. In other words, in this example, certain political parties do not necessarily hold certain views on the importance of family in any regular pattern. Orthogonal rotation assumes that factors are indeed independent.

Oblique rotation, on the other hand, takes into consideration that some factors will often relate to each other somehow. This means that factors that are positively related will have a smaller angle between them, while negatively related factors will have a larger angle. This allows variables and the strength of their correlations to certain factors to be accurately portrayed. A pair of factors that would be best fit to oblique axes is political affiliation and social class. There are numerous studies and examples where members of a certain social class tend to belong to a particular political party and vice versa. Since these factors would lend themselves to influencing each other, they would not be independent of one another, and therefore more appropriately set to be on oblique axes.

5. An Application of Factor Analysis

In fall of 2013, two professors at Biola University in the Math and Computer Science Department, Dr. Jason Wilson and Dr. William McCarty, oversaw the administration of a survey to a sample of 110 students on campus. Dr. Wilson, professor of an introductory statistics class and a course in biostatistics, asked students to complete this for a class project. While the students analyzed the survey data, a factorial analysis was not attempted until January of 2014.

Dr. McCarty put together the survey using questions from a previously administered survey, done with a group of 5,000 Lutherans between the ages of 15 and 65. Altogether, 23 different questions were asked along with six demographic questions. For each non-demographic question, each participant was asked to read the statement and check a box indicating that they agreed with the statement, disagreed with the statement, or were unsure about their thoughts. Once the surveys were collected, each response was recorded as -1, 0, or 1.

The nature of the questions ranged from Gospel teachings to how the individual viewed God and their relationship with Him. All of the borrowed questions came from the sections titled "Salvation by Works" and " Service Without Proclamation" in the book *A Study in Generations* (Strommen, 1972). Dr. McCarty picked each question, adding in a few of his own questions as well.

Though not directly related to factor analysis, the preliminary analysis of the survey showed good news – Biola students answered questions correctly more frequently than the original sampled group for every single question except for one. The question which did not produce a statistically significant result for a than the original study was the statement that said, "There is nothing which science cannot eventually understand." More analysis included a linear regression model based on major, denominational affiliation, gender, and age of conversion to Christianity. The results of the preliminary analysis led to the decision to carry out factor analysis. The main focus of this section is the factor analysis on the differences within responses to questions, so the preliminary analysis has been omitted.

Analysis was completed through an open-source statistical program called R (www.r-project.org) The function factanal() was employed, which makes use of the Maximum-Likelihood Method of factor analysis. The chosen rotation method was an orthogonal method called *varimax*, which is the most widely used method of orthogonal factoring. The following table displays the factor loadings of three factors on the 23 questions, modified for readability into Table 1. The original output from R is can be found in the Appendix.

Question	Factor 1	Factor 2	Factor 3	Question	Factor 1	Factor 2	Factor 3
17	**0.378**	0.195	--	54	0.177	0.190	**0.257**
18	**0.140**	--	0.103	15E1	**0.650**	--	0.202
26	**0.285**	--	0.262	15E2	**0.398**	0.205	--
28	**0.244**	--	--	15E3	**0.642**	--	--
30	--	**0.819**	-0.106	15E4	**0.414**	--	-0.297
31	0.201	**0.637**	0.171	320	0.145	**0.294**	0.268
42	--	0.254	**0.297**	531	--	--	**0.273**
43	**0.372**	--	--	534	0.153	**0.661**	--
44	**0.671**	0.240	--	539	--	**0.653**	--
45	0.259	0.173	**0.302**	62E1	0.356	0.311	**0.383**
51	0.171	**0.241**	0.194	62E2	**0.521**	0.207	0.305
53	-0.131	0.131	**0.465**	Total	11	6	6

Table 1: Factor loadings for 23 variables against 3 factors

As seen above in Table 1, there are some variables that did not have any factor loading with certain factors, labeled with "--" in the cell. Bolded are the highest factor loadings for each variable. This groups eleven questions on factor 1, six questions on factor 2, and six questions on factor 3. After consideration of the content of the questions in each group, a few common characteristics surfaced with each factor. The questions grouped with their factors can be found in the Appendix.

These are the results of confirmatory factor analysis with three factors hypothesized by Dr. Wilson. The results contained one correctly hypothesized factor with no perfect match of which questions were grouped together. Factor 1 contained questions whose main topic was the teachings of the Gospel. Questions included doctrinal content and proper theology in orthodox

Christianity. For this reason, factor 1 can be understood as "Proper Understanding of the Gospel". Factor 2 dealt with controversial issues within society and problems that society seems to have with Christianity. Therefore, factor 2 can be understood as "Society's Influence on Christian Thinking". Thirdly, factor 3 had questions that asked about an individual and their part in Christian life, such as how someone's actions affect God's view of them and what they can do in their life. As such, factor 3 has been set as "Individual Abilities and Work".

With these factors in mind, students' responses see to be influenced by three distinct groups of thinking – doctrine, society, and the individual. Each of these areas seem to influence how a student at Biola forms opinions on specific issues within Christianity, so it could be beneficial for Biola University to make sure each of these topics are covered properly in the education. The good news of the analysis is that it shows that education in Biblical Studies does impact thinking patterns, but there is always room for improvement on methodology.

6. Conclusion Statements

The results of this study are not necessarily applicable in all situations but they do offer insight into an aspect of Biola University that is crucial to its purpose. The factors found here through the Maximum-Likelihood Method of factoring and varimax rotation are simple, yet important factors to take into account. This analysis provides room for discussion and gives a starting place for other analysis to occur, possibly leading to different methods of factoring or more factors to take into account. The possibilities for more work are endless.

Factor analysis as a whole offers an additional perspective into the underlying influences of complex situations and is an incredibly useful tool in the world. Though it can be rigorous and computationally complex, the software available today allows for analysis of data to be accomplished by many people without extensive statistical training. Factor analysis is an incredibly helpful method in statistics and its applications, and in future years, it will most likely be used more often than it is now in various situations and for countless applications with data.

Appendix A: R Scripts

The following includes the scripts used in RStudio:

```
part1 = survey[, 8:23] #first part of the survey data

Q15E4 = -1*survey[, 24] #inverting all the answers here for this
column

part3 = survey[, 25:30] #third part of the survey data

x = cbind(part1, Q15E4 ,part3) #binding the data together for analysis of
proper input

x1 = na.omit(x) #omitting any unanswered questions from the data

fact3 = factanal(x1, 3 ,rotation="varimax") #factor analysis with 3 factors

fact3

Call:

factanal(x = x1, factors = 3, scores = c("regression"), rotation = "varimax")

Uniquenesses:

  Q17   Q18   Q26   Q28   Q30   Q31   Q42   Q43   Q44   Q45   Q51   Q53   Q54

Q15E1

0.813 0.967 0.847 0.938 0.318 0.525 0.847 0.852 0.483 0.812 0.875 0.750 0.867

0.534

Q15E2 Q15E3 Q15E4  Q320  Q531  Q534  Q539 Q62E1 Q62E2

0.797 0.587 0.740 0.820 0.918 0.538 0.572 0.630 0.592


Loadings:

      Factor1 Factor2 Factor3

Q17    0.378   0.195

Q18    0.140           0.103

Q26    0.285           0.262

Q28    0.244

Q30            0.819  -0.106
```

Q31	0.201	0.637	0.171
Q42		0.254	0.297
Q43	0.372		
Q44	0.671	0.240	
Q45	0.259	0.173	0.302
Q51	0.171	0.241	0.194
Q53	-0.131	0.132	0.465
Q54	0.177	0.190	0.257
Q15E1	0.650		0.202
Q15E2	0.398	0.205	
Q15E3	0.642		
Q15E4	0.414		-0.297
Q320	0.145	0.294	0.268
Q531			0.273
Q534	0.153		0.661
Q539		0.653	
Q62E1	0.356	0.311	0.383
Q62E2	0.521	0.207	0.305

	Factor1	Factor2	Factor3
SS loadings	2.686	2.093	1.598
Proportion Var	0.117	0.091	0.069
Cumulative Var	0.117	0.208	0.277

Test of the hypothesis that 3 factors are sufficient.

The chi square statistic is 256.95 on 187 degrees of freedom.

The p-value is 0.000522

Factor 1 includes the following questions:

- The main emphasis of the Gospel is on God's rules for right living.
- Although there are many religions in the world, most of them lead to the same God.
- Hard work will always pay off if you have faith in yourself and stick to it.
- Being tolerant means that one accepts all religions--including Christianity—as equally important before God.
- God is satisfied if a person lives the best life he can.
- A person at birth is neither good nor bad.
- We are saved by grace through faith, after having done all that we can do.
- Although salvation comes as a gift we must work earnestly to show that we are deserving of that gift.
- God's grace makes it possible for us to do the good works necessary to deserve salvation. Grace is necessary but not sufficient.
- God's grace is sufficient in itself to save and no good works need be added to it. Grace is both necessary and sufficient.
- Warning people of hell will only antagonize them. People should be drawn to Christ by a message of pure love.

Factor 2 includes the following questions:

- Sin is whatever people (society) think is wrong behavior.
- The Bible teaches that God is like a friendly neighbor living upstairs.
- There is nothing which science cannot eventually understand.
- Christians should leave other people alone and not try to change their religion.
- The Church's task to help eliminate physical sufferings of people is more important than proclaiming the Gospel by preaching and teaching.
- A worship service must be beautiful to be really meaningful to me.

Factor 3 includes the following questions:

- Hard work keeps people from getting into trouble.

- Salvation depends upon being sincere in whatever you believe.
- A man should stand on his own two feet and not depend on others for help or favors.
- If I say I believe in God and do right, I will get to Heaven.
- Missionaries should not proclaim God's Law [e.g., the Ten Commandments] too often to people suffering from poverty and sickness.
- Christians should not proclaim the Gospel to others until they are invited to do so.

References

Biola University. (1908). Mission, Vision & Values. *Biola University Webpage*. Retrieved May 7, 2014 from the World Wide Web: http://www.biola.edu/about/mission/

Fruchter, B. (1954). Introduction to Factor Analysis. Princeton, New Jersey: D. Van Nostrand Company, Inc.

Heeler, R.M., Whipple, T. W., and Hustad, T.P. (1977, February). Maximum Likelihood Factor Analysis of Attitude Data. Journal of Marketing Research 42-51. Retrieved May 3, 2014 from the World Wide Web: http://www.jstor.org/stable/3151053

Rahn, M. (2013). Factor Analysis: A Short Introduction, Part 1. Retrieved April 19, 2014 from the World Wide Web: http://www.theanalysisfactor.com/factor-analysis-1-introduction/

Rummel, R.J. (1967, December). Understanding Factor Analysis. The Journal of Conflict Resolution 444-480. Retrieved April 23, 2014 from the World Wide Web: http://www.jstor.org/stable/173151

Scott, J. T. (1966, July). Factor Analysis and Regression. Econometrica 552-562. Retrieved April 21, 2014 from the World Wide Web: http://www.jstor.org/stable/1909769

Strommen, M. P., Brekke, M. L., Underwager, R. C., and Johnson, A. L. (1972). A Study of Generations. Minneapolis, Minnesota: Augsburg Publishing House.

Thompson, B. (2004). Exploratory and Confirmatory Factor Analysis. Washington, DC: American Psychological Society.

Utts, J. M. (2005). Seeing Through Statistics. Belmont, California: Thomson Brooks/Cole.