Exploratory Factor Analysis and Principal Components Analysis

Nicholas Hopewell – 0496633

AMOD 5240

Fall 2017

Put simply, factor analysis reduces a complex amount of information down to a simpler description of the important aspects of said information. If we were interested in investigating some construct and we had many variables which could be used for this investigation, exploratory factor analysis (EFA) would inform us of how we could reduce the number of variables while retaining crucial context.

To begin to understand exploratory factor analysis, it is helpful to know when one might want to use factor analysis. Quite frequently, researchers are interested in measuring variables which cannot be observed directly but rather must be inferred from additional variables which the researcher can measure directly. These variables are referred to as latent (or latent variables), a term derived from a Latin word meaning "to lie hidden." For example, an economist might be interested in researching quality of life of certain individuals, societies, or countries. While quality of life itself cannot be directly measured, as it has many facets, variables such as life expectancy, education, and standard of living *can* be measured in an attempt to quantify quality of life. If we measured differences in these variables for multiple societies, for example, it would be useful to determine in some way whether these differences truly reflect one variable or not. More specifically, we would like to know whether all of these variables are driven by quality of life. Any situation where we are looking for clusters of related variables, it is useful to conduct EFA and/or principal components analysis (PCA).

EFA has three important applications:

1. To understand the underlying structure of a given set of variables. These sets of variables are often relatively large because the amount of information derived from a large set of variables is a primary motivating factor for doing EFA in the first place.
2. To design a valid questionnaire to measure a latent variable such as business confidence or business moral.
3. To reduce a large data set to a smaller size while retaining as much of the original information as possible and especially the most important information.

To expand on the final application, it might be helpful to consider an example. In the context of regression, one assumption of multiple linear regression is that the predictor variables included in the model are not too highly related to each other. An issue arises when the bivariate correlation between two or multiple predictors in a regression model are strong (this is called Multicollinearity). Such a circumstance would make it much more difficult to determine the importance of each individual predictor variable as variance in the outcome accounted for by the separate variables would be relatively similar, or at least not different enough to meaningful. Factor analysis could be used to combine variables which are collinear, effectively reducing the size of the data set without losing any important details.

These important goals of EFA fall under what is known as the principle of parsimony (similarly, Occam's Razor or MDL: Minimum Description Length for comparing multiple data-derived models in data mining contexts). These concepts try to express the same message: when explaining a relationship or describing a phenomenon, the best method is the simplest method. In other words, assuming the same amount of information can be communicated, a simpler model is always better than a more complicated one. To gain an appreciation of why the simplest explanation is the most acceptable explanation, it is best to think about the purpose of an analysis in an even more broad sense. An analysis does not stop at the derivation of a model or the selection of the 'best' model within a given context. In the words of Haldey Wickham, the end product of an analysis is not a model, it is rhetoric. An analysis has no value unless the key components of whatever is being explored are determined, learned from, and communicated in a way that resembles some sort of call to action. Thus, from a statistical perspective as well as a practical perspective, factor analysis has many benefits.

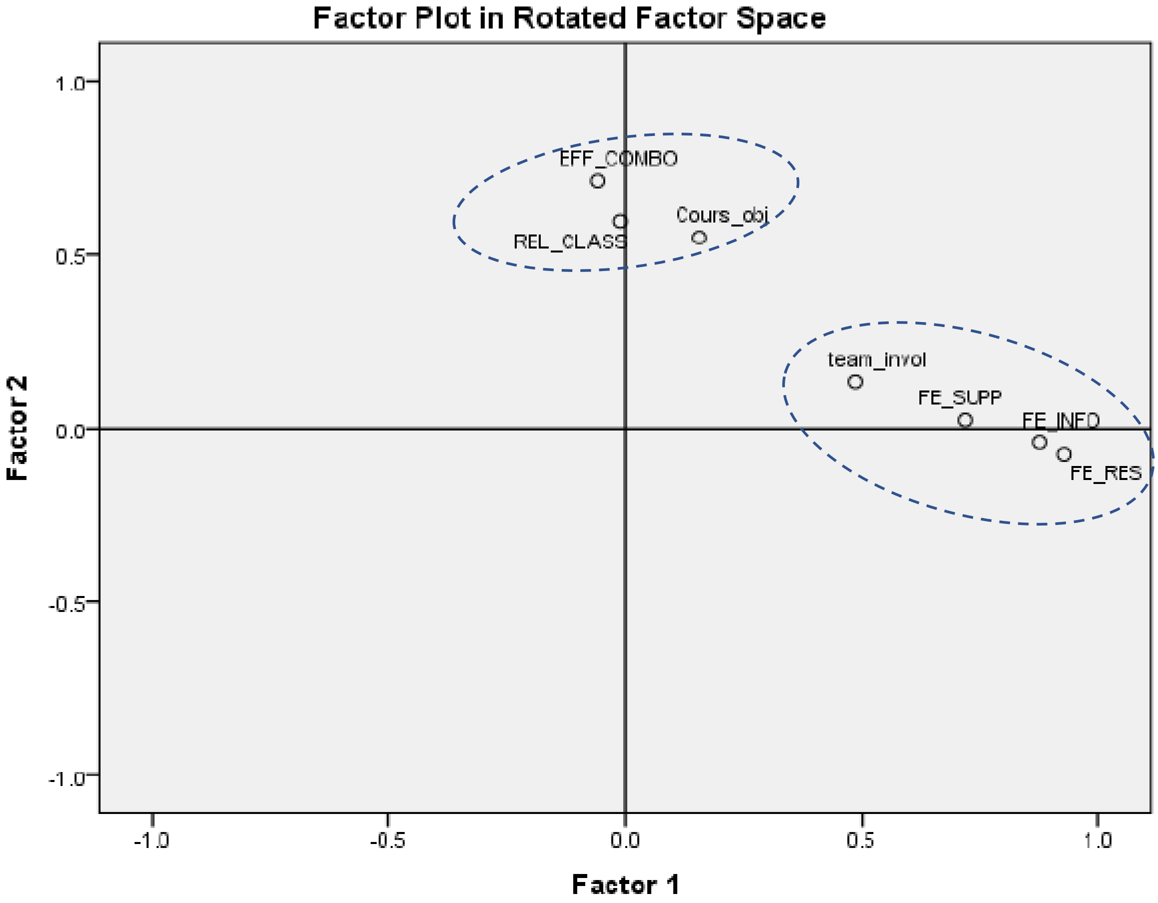
But what exactly are these 'factors' we try to analyse? Consider this: if we looked at a correlation matrix (*R*-matrix) of multiple measured variables and found clusterings of large correlation coefficients between subsets of variables, this might hint that these related variables are all measuring characteristics of some other underlying construct. These underlying constructs are referred to as factors and are simply another name for latent variables. It is important to note that these clusters of related variables are only likely to be measuring the same underlying construct if they are not only interrelated but also do not correlate with other variables outside of the cluster (or at least to a much lower degree). EFA would allow us to reduce these meaningful clusters of related variables into a much smaller number of factors. By doing this, EFA establishes parsimony by deriving the simplest model which can explain as much as is needed to explain about the variance in the matrix. In other words, the most interesting relationships between variables have been retained while the less interesting relationships have not.

Factor Plots and Factor Loadings

When visualizing a factor plot, the factors are represented as classification axes. Simply put, if the variance in our data was adequately explained by two factors, factor 1 would be represented as the x-axis of the plot while factor 2 would be represented as the y-axis of the plot. The scales of both axes range from -1 to 1 which is simply the range of a Pearson's correlation. If the measurement variables were then plotted along these axes, the position of these variables in 2-dimensional space (their coordinates along each axis) would indicate their strength of relationship with the two factors. This is because the position of the variables along the axes would depend entirely on their correlation with each of the factors. If it were determined to retain three factors, we would simply need to design a 3-dimensional plot (one dimension for each factor). Retaining more than three factors cannot be visualized on paper as a factor plot but would require a multi-dimensional visualization tool.

In practice, if a clustering of variables truly did measure different facets of the same underlying factor, we would expect that this cluster would have larger coordinates for the factor it is related to and smaller coordinates for any additional factors. In a two-factor scenario, this would mean that a cluster would fall around 0 (or a low/negative value) on the axis of the factor it was not related to and around .5 to 1 on the axis of the factor it was related to. The exact coordinate a variable lies along each axis of a factor plot is known as the variable's factor loading for each factor. Given what we know about the axes of a factor plot, a variable's factor loading is simply its Pearson's correlation with each factor. Because of this, it is possible to square a variable's factor loading value to get an indication of the substantive importance of a variable to a given factor. The result is an R^2 value (also known as the coefficient of determination) which, within the context of EFA, is the amount of variability within given factor which is shared by a particular variable. R^2 values are often reported in a way that implies causality such as the percentage of variance in one variable *explained* or a*ccounted for* another variable. Because R^2 is derived from a correlation coefficient, it should not be used to infer a causal relationship. For instance, if the variable ‘education level’ shared 21.3% of the variance in the factor ‘quality of life,’ it is not necessarily the case that education level causes this variation.

Below is an example of a factor plot. The actual variables do not matter in this example. The importance of the plot is to show factor loadings for each variable to each factor and that two clusters, each separately measuring one of the two underlying factors, seem to apparent.



Representing Factors Mathematically

Although the above example was taken from a quick google image search, I will use it as well as the approximate factor loadings of these variables to create a factor matrix (explained below).

Because we represent factors as axes in a factor plot and these axes are straight lines, factors may be represented as a modified equation for a straight line. The equation for a factor is very similar to the equation for a linear regression model without the inclusion of an intercept value. This is because in the case of EFA the intercept does not matter; since the axes always intercept at 0.00, the intercept is simply 0 and thus does not need to be included in the equation.

Also, instead of including a regression coefficient, each variable's factor loading is included in the formula. An equation for each line is created which includes every variable of interest. The equations differ for each factor in the sense that the factor loadings for the variables will be different for each factor.

Factor 1 = 0.5team_invol + 0.71FE_SUPP + 0.84FE_INFO + 0.92FE_RES
– 0.01REL_CLASS – 0.08EFF_COMBO + 0.16Cours_obj + error term

Factor 2 = 0.18team_invol + 0.02FE_SUPP – 0.04 FE_INFO – 0.08FE_RES
+ 0.53REL_CLASS + 0.68EFF_COMBO + 0.51Cours_obj + error term

Notice the above factor loadings are high for the variables which are important to each factor and low for the variables which are not important to each factor. These factor loadings tell the same story as seen in the factor plot: two clusters of variables are apparent which may be measuring two different underlying factors. These factor loadings can be displayed more simply in a factor matrix (called a components matrix in principal components analysis) where the number of columns are equal to the number of factors and reach row contains one variable's factor loadings for each factor.

Interpretation of Factors: A Warning

Once we have concluded that subsets of our measured variables are being driven by a number of factors, a substantial leap must be made by the researcher to conclude that these factors truly represent some construct in the real world. Previously, quality of life was discussed as an example of a latent variable (or a factor). Quality of life is well-researched, and this has led to the development of numerous quantitative measurement scales. Furthermore, the variables which are measured to determine quality of life make sense in the real world. For instance, it is seen time and time again that areas with higher standards of living, better quality and accessibility to education, and higher life expectancy rates consistently score higher in every measure of quality of life we have developed. We can see the impacts that these things (variables) have on someone's overall quality of life (factor).

When we want to quantify factors which are very abstract, not fully understood, and much less tangible, things become complicated and conclusions can be debatable. If instead of trying to understand quality of life we wanted to find a cluster of measurable variables which could represent personality types (which cannot be directly measured), it would require a much larger stretch of inference (or imagination) from the researcher. This is because although factors might exist in a statistical sense, what we *believe* these factors represent in the real world may be little more than a belief. When all is said and done, exploratory factor analysis leads to an assumption that what we found through statistical analysis has meaning beyond simply numbers on a page or a screen. Ultimately, the nature of this real-world meaning must be guessed at based off quantitative measures of things we think should be related to a broader concept. Can even the wisest researcher take a matrix of factor loadings and determine the existence of universal personality types which we can use to describe the behaviours and outcomes of all individuals? Some might say that this is very unlikely. In these cases, the scientific process must have time to develop a large body of research to support the claim that these factors have meaning in the real world. Even when such a body of research has been developed, it is up to the critical consumer of the research to realize that the individuals who came to these overarching conclusions are humans who make decisions based off of past and present judgement and whose findings often confirm personal beliefs and uphold conventions of the research area (good and bad).

Factor Scores

Now that we have an equation for our factors, we can estimate an individual's score on a factor by looking at how they scored on the related measured variables. This estimate takes the form of one number and is referred to as a factor score. The ability to determine someone's score on a factor by inputting values into the factor equation is similar to the process of inputting values into a linear regression equation to determine a predicted value of an outcome variable. Although conceptually similar, to determine a factor score we use a weighted sum based off factor loadings and individual scores on each measurement. Imagine these arbitrary variables used in the factor equation example were each measured on a scale from 1 to 10 and an individual scored: team_invol(7), FE_SUPP(4), FE_INFO(9), FE_RES(6), REL_CLASS(5), EFF_COMBO(8), Cours_obj(6). This individuals factor score for each factor would be the following:

Factor 1 = 0.5team_invol + 0.71FE_SUPP + 0.84FE_INFO + 0.92FE_RES
– 0.01REL_CLASS – 0.08EFF_COMBO + 0.16Cours_obj
= (0.5 x 7) + (0.71 x 4) + (0.84 x 9) + (0.92 x 6)
– (0.01 x 5) – (0.08 x 8) + (0.16 x 6)

= **19.61**

Factor 2 = 0.18team_invol + 0.02FE_SUPP – 0.04 FE_INFO – 0.08FE_RES
+ 0.53REL_CLASS + 0.68EFF_COMBO + 0.51Cours_obj
= (0.18 x 7) + (0.02 x 4) – (0.04 x 9) – (0.08 x 6)
+ (0.53 x 5) + (0.68 x 8) + (0.51 x 6)

= **11.65**

These factor scores simply represent a way to quantify the degree to which an individual associates with whichever latent variables we are trying to measure. If the measurement scales of the individual variables are not equal (scales do not have the same range of possible values) the resulting factor score cannot be interpreted. In this case, scales must be normalized or a different approach to calculating factor scores must be taken. A common method is to replace the factor loading values in the factor equation with factor coefficients determined by applying a regression technique. An explanation about how this technique is implemented would be lengthy and would not add much to the understanding of EFA so I will not cover the technique in detail. Essentially, the technique corrects for multiple different measurement scales by stabilizing differences in units of measurements and variances of each measured variable.

In the end, what started as a large amount of information across many variables is now a small number of values for each factor. Factor scores can be used to continue an analysis in place of the original data. For example, since interrelated variables were combined, using factor scores to develop a regression model would overcome the issue of multicollinearity of predictors. To determine how identifiable groups of individuals differ in terms these factors, we could do a simple *t*-test using factor scores. Overall, the process of EFA makes a complex analysis more manageable and interpretable if the context is appropriate.

How Else are Factors Found?

Factors are found through multiple different methods. Discovering factors using factor analysis has already been discussed but additional methods can be used. The choice of these methods comes down to whether the researcher wants to generalize their findings to a population, whether a specific hypothesis is being tested, or if the nature of the data is simply being explored in order to generate a hypothesis. When a researcher wants to test a specific hypothesis about the structure of latent variables, he or she would be interested in conducting a confirmatory factor analysis (something outside of the scope of this paper). Because factor analysis was developed with the intention of generating a hypothesis, it is assumed that factor analysis will be applied to a population of interest. Principal components analysis (very closely related to principal axis factoring) is one example of a very widely used method that treats the sample as the population. Because of this, any insights and conclusions found using this technique cannot be generalized beyond the current sample until repeated analyses using different samples are done and the same or similar factor structure is found. Other common techniques which will not be discussed include the maximum-likelihood method and Kaiser’s alpha factoring.

Factor Analysis vs. Principle Components Analysis

These two techniques differ in two main ways: the estimates of communality used as well as related assumptions which must be made. Assumptions which must be made about the real-world importance of factors as result of EFA have already been discussed. As a simple overview, communality refers to the proportion of variance present in a variable which is shared by other variables or measures (also called common variance). This is opposed to unique variance which is the proportion of variance which is attributed to only one variable or measure (or unique to that variable). Random variance, which is not reliably attributed to any specific variable included in a model, is referred to as error or error variance. In the context of EFA, communality specifically refers to the proportion of variance which is accounted for by the factors extracted from the data. To appreciate why common variance is of interest during EFA, it is helpful to remind oneself of the goal of EFA: to discover ‘common’ or underlying factors of which subsets of variables are collectively measuring and thus being driven by. An issue arises here because although we are interested in the common variance of our measures enough to proceed with EFA, we only get a sense of the proportions of common variance after undertaking EFA. This can be overcome with multiple methods to estimate communality, such as completing multiple regressions for each variable separately as the outcome variable and each of the other variables as predictor variables and taking the multiple R^2 values as the estimate of communality for the corresponding outcome variable. This method is precisely how researchers often first estimate communality when doing EFA.

Principle components analysis, on the other hand, is not interested in estimating communality, but instead treats all variance as common variance and simply reduces the correlated variables into a smaller number of orthogonal (uncorrelated) variables called principle components (which can be thought of as factors). PCA finds strong patterns in data where the first component accounts for the most variance in the data and each following component accounts for less than the one preceding it, and is interested in how a variable may contribute to a component. The term ‘factor’ and ‘component’ are often used interchangeably because these two methods are sometimes seen as two different ways to tell the same story, with major differences residing in how the two are calculated.

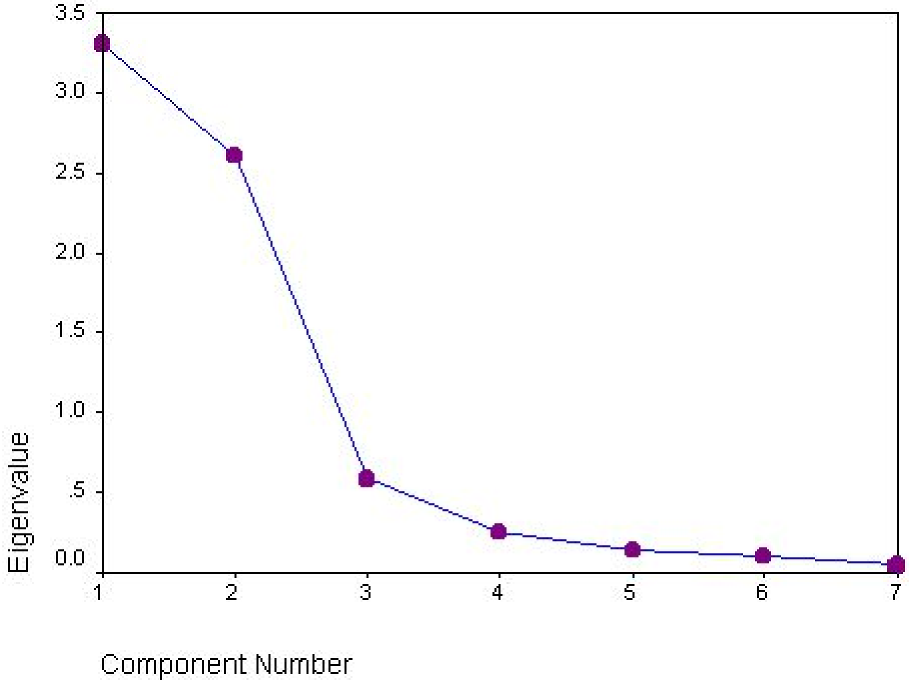
A very brief overview of principle components analysis is as follows: the researcher begins with an *R*-matrix of the measured variables; eigenvalues of the matrix are then used to calculate linear components (also known as variates or factors) of the matrix; finally, the ‘eigenvalues’ are used to calculated ‘eigenvectors’ whose elements are simply the weights of each variable on the component (just like factor loadings used in factor equations). Recall that a variable’s factor loading is simply its Pearson’s correlation with each factor, and if those values were squared we would get an indication of the substantive importance of a variable to a given factor. Similarly, an eigenvalue is a measure of the substantive importance of an eigenvector it is associated with. Without going into detail about eigenvalues, what really matters is that the largest eigenvalue for each of the eigenvectors gives the researcher an idea of how important a component is and whether or not it should be retained after retaining more important factors. It is best to retain components which have large eigenvalues relative to all other components and only retain an adequate number of components to capture important variability in our measured variables. This is because the higher a component’s loading values (factor loadings), the greater the proportion of variance in the variables that component explains. It is useful to know that an eigenvalue for a particular component can be calculated by first squaring its factor loadings and then summing those squared values.

Using the Scree Plot for Factor Extraction

The Y-axis of a scree plot is a range of eigenvalues with a lower limit of 0 and an upper limit that depends on the data, while the X-axis is a range of components. How many components are determined and how they fall on the scree plot is related to the number of variables and how much total variability there is to explain in the first place. It was mentioned that it is beneficial to retain factors with relatively high eigenvalues and ignore factors with relatively low eigenvalues. This is because the eigenvalue of a component gives an idea of the component’s importance, and when these values are plotted it is easy to compare the relative importance of each component.

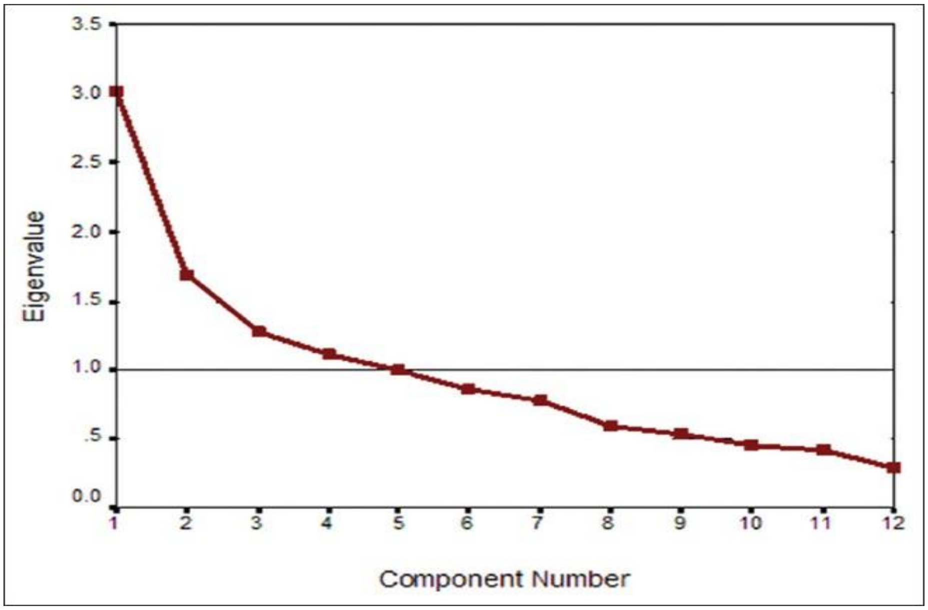
While there is no agreed-upon ‘perfect’ method of retaining factors, almost universally researchers look for what is called a point of inflexion in the scree plot when making such decisions. Essentially, in almost all cases, the first 2-3 components will have large eigenvalues followed by a noticeable change in the slope of the line connecting to next component and then a long trailing off of many components with relatively small eigenvalues. If the first two components capture most of the variance in the variables, the slope of the line connecting the components would change noticeably following those components. The point of inflexion is simply the point (representing a component) on the scree plot that connects components with large eigenvalues to the following components with smaller eigenvalues. It is convention to retain the components to the left of this point (first two components in this example) and not retain the component at the point of inflexion nor the components to its right. Some have argued that the component at the point of inflexion should be retained, although this is rarely done. In this case, we would have extracted 2 components using PCA with our original variables. Extracting components in this manner is quite useful when the sample size that you are taking your data from includes more than 200 individual records.

Below is a scree plot showing a point of inflexion at components three, thus, the first two components would be retained.



It is recommended to consider other methods when retaining components rather than solely making a qualitative judgement based on a scree plot. It is common for researchers to retain components with eigenvalues greater than one because it has been argued that this represents the fact that a component is explaining a significant amount of variance in the measured variables. Others argue that components with eigenvalues greater then 0.7 should be retained. It is important to remember that the retained components do not perfectly represent the original variables but instead reflect common variance in the data previously discussed. These methods are fundamentally flawed because an eigenvalue of 1.0 does not mean the same thing for every analysis. If an analysis included 100 variables, then an eigenvalue of 1.0 would mean that a component accounts for about 1% of the variability in the measured variables. If an had only 10 variables, then this eigenvalue would mean that the component accounts for 10% of the variability in the data. In addition, following these guidelines often leads to retaining more components than is necessary to capture the required proportion of common variance while achieving parsimony, which defeats the purpose of a factor analysis or PCA in the first place.

Below is a scree plot using eigenvalues greater than 1.0 as a cut-off to extract components. In the below example, five components would be retained.



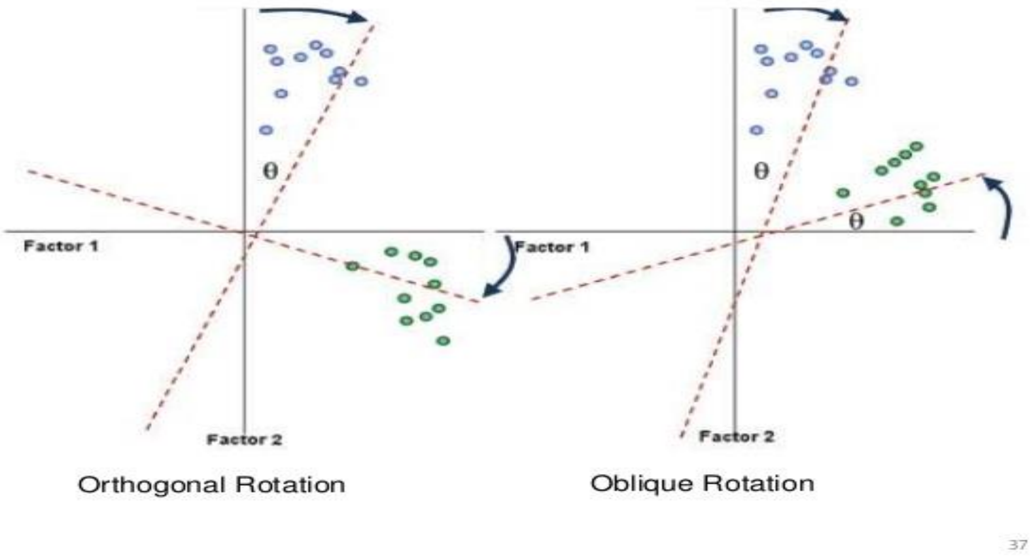
The process of determining whether not enough components have been extracted is outside the scope of this paper. I will mention that, depending on the research problem, one might want to be stricter when determining the cut-off for extracting components while others may want to extract more than the only the most statistically important components.

Factor Rotations

Due to the nature of how principle components are extracted, it is common that when variable loadings are calculated for the extracted factors, most variables will have high loadings for the first principle component and much smaller loadings on the rest of the components. To discriminate between components, the axes of a factor plot are rotated in such a way that variables load highest for only one component. Recall that a factor plot had the first factor as the X-axis and the second factor as the Y-axis and the axes scales ranged from -1 to 1 to allow the Pearson's correlation of each variable to each factor to be plotted. Also recall that the terms 'component' and 'factor' represent the same thing and can are often used interchangeably. Essentially, the axes can be rotated such that the cluster of variables which load highest to each component are intersected by that components axis line. This ensures that the factor loadings for those variables are maximized to that component and minimized to all other components. Recall from the factor equations that variables have loading values for all factors, and just because a certain variable loads highly on one factor does not mean it has a loading value of 0 on another factor. Clusters of variables having a factor loading close to zero for other components only is the case after the axis rotation is performed.

The axes of a factor plot are rotated in two ways. The fist is an orthogonal rotation where the axes remain perfectly perpendicular once rotated. This simply means that before and after rotation the factors remain independent of each other (the rotation does not allow them to correlate). An orthogonal rotation is the only option when the factors are determined to be independent. The second is an oblique rotation where the axes do not remain perpendicular after rotation and thus the factors do not remain independent (factors are allowed to correlate). If the researcher has a good reason (usually based on past research findings) to believe that the factors are not independent but instead share some sort of relationship, then an oblique rotation might be the best choice of rotation. The issue arises when related variables are clustered such that an orthogonal rotation does not sufficiently maximize the factor loadings of these clustered variables to their primary factor. It is common to simply compare the results of both types of rotations and retain the rotated factor space which yields an appropriate resolution only for cases where the factors might be not independent. Factor transformation matrices are beyond the scope of this paper.

Below is a display of these two types of rotations.



Notable orthogonal rotation techniques include quartimax and varimax. Notable oblique rotation techniques include promax and oblimin. These techniques will not be described but it should be known that in many circumstances orthogonal rotations do not make sense because it cannot be justified that certain variables are in no way related to each other. This is often the case in any data taken from human samples because variables measured are likely to share some relationship. A couple of R packages which can be used for rotations are “corpcor” and “GPAtotation”.

Reference:

Textbook: Field, A. P., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. London: Sage.