Chapter 3: Performing a factor analysis

For an overview of key terms, refer to page 90-92

What is factor analysis?

Factor analysis = An interdependence technique, with as main goal to determine the underlying structure among the variables in the analysis. Variables are the building blocks of relationships. As we add more variables, more and more overlap will exist between those variables. As the variables become more correlated, the researcher now needs ways in which to manage these variables. Factor analysis provides the tools for analyzing the structure of the interrelationships among a large number of variables. By defining sets of variables that are highly interrelated, called factors. If we have a conceptual basis for understanding the relationships between variables, these dimensions (factors) may have meaning for what they collectively represent.
Many researchers call it only exploratory, useful for searching structure among a set of variables or as a variable reduction method. However, sometimes it can also be used to test hypotheses, as a confirmatory method.

Factor analysis decision process

Stage 1- objectives

the general purpose is to find a way to condense the information contained in a number of original variables into a smaller set of new, composite dimensions or variates with a minimum loss of information. In meeting its objectives there are 4 issues:

- Specifying the units of analysis. If the objective is to summarize the characteristics, a correlation matrix would be created. This is done by conducting R factor analysis. Factor analysis may also be applied to a correlation matrix of the individual respondents based on their characteristics. This is called a Q factor analysis and combines a large number of people into distinctly different groups based on their characteristics. A cluster analysis can also be used to group people.
- Achieving data summarization and /or data reduction. The fundamental concept in data summarization is structure. Through structure the researcher can view the set of variables at different levels of generalization, where groups of individuals are viewed for what they represent collectively. In factor analysis all variables are simultaneously considered no matter if they are dependent or independent. It still employs the concept of the variate, but not in predicting dependent variables, but in explaining the entire variable set. Structure is defined by the interrelatedness among variables allowing for the specification of a smaller number of dimensions(factors) representing the original set of variables. Factor analysis can also be used to achieve data reduction by either identifying representative variables from a much larger set of variables for use in subsequent multivariate analysis or by creating an entirely new set of variables to replace the original set of variables. Estimates of the factors and the contributions of each variable to the factors(loadings) are all that's required to the analysis with data summarization. In data reduction, these loadings are also important, but it uses them for either identifying variables for subsequent analysis with other techniques or making estimates for the factors themselves.
- Variable selection. The researcher implicitly specifies the dimensions that can be identified through the character and nature of the variables submitted to factor analysis. The researcher also must remember that factor analysis always creates factors. It is a potential candidate for 'garbage in, garbage out'.

- Using factor analysis results with other multivariate techniques. Variables determined to be highly correlated and members of the same factor would be expected to have similar profiles across groups in multivariate analysis. Knowledge of the structure of the variables itself would give the researcher a better understanding of reasoning behind the entry of variables in this technique.

Stage 2- Designing a factor analysis

The design of a factor analysis involves 3 basic decisions:

- Calculation of the input data to meet the specified objectives of grouping variables or respondents. With R type analysis, the researcher uses the traditional correlation matrix as input (correlations among variables). In the Q type analysis this correlation matrix would be based on the correlation between individual respondents. What is the difference between Q type analysis and cluster analysis? Q type analysis is based on the intercorrelation among correspondence, whereas cluster analysis forms groupings based on a distance-based similarity measure between respondents.
- Design of the study in terms of number of variables, measurement properties of variables and the types of allowable variables. Two specific questions must be answered at this point. What type of variables can be used in factor analysis? Metric variables are the best option, nonmetric variables can cause problems. In case of nonmetric variables, dummy variables should be created.
  How many variables should be included? The researcher should attempt to minimize the number of variables included, but still maintain a reasonable number of variables per factor.
- The sample size necessary. The researcher would generally not analyze samples of 50 or smaller, and preferably over 100. The researcher should try to obtain the highest cases-per-variable ratio to minimize chances of over fitting.

Stage 3- assumptions in factor analysis

The assumptions underlying factor analysis are more conceptual than statistical.

- Conceptual issues. A basic assumption is that some structure does exist in the set of selected variables. the presence of correlated variables and the subsequent definition of factors do not guarantee relevance, even if they meet the statistical requirements. The researcher must also make sure that the sample is homogeneous with regard to the underlying factor structure.
- Statistical issues. Some degree of multicollinearity is desirable, because the objective is to identify interrelated sets of variables. The next step is measuring the degree of intercorrelatedness and to check if this is sufficient to produce factors.
  Overall measures of correlation. The researcher must ensure that the data matrix has sufficient correlations to justify the application of factor analysis. If the correlations are low, or if all correlations are equal (which implies no structure exists) the researcher should question the application of factor analysis. Several approaches are available.
  If no correlations are larger than .3, factor analysis is inappropriate. The correlations can also be analyzed by computing partial correlations. A partial correlation is the correlation that is unexplained when the effects of other variables are taken into account. If partial correlations are high, factor analysis is inappropriate. Partial correlations above 0.7 are high.
  Another method is performing the Bartlett test of sphericity, a statistical test for the presence of correlations.

A third measure is the measure of sampling adequacy. It ranges from 0-1 with 1 when each variable is perfectly explained without error by the other variables.
Variable specific measures of intercorrelation. The researcher should examine the MSA values for each variable and exclude those falling in the unacceptable range. In deleting the variables, the researcher should first delete the variable with the lowest MSA value and then recalculate the factor analysis.

Stage 4- deriving factors and assessing overall fit

Once the variables are specifies the researcher is ready to apply factor analysis to identify the underlying structure. In doing so, decisions should be made concerning the method of extracting the values and the number of factors selected to represent the underlying structure in the data.

Selecting the factor extraction method.

- Partitioning the variance of a variable in order to select a method the researcher should first have an understanding of the variance for a variable and how it is divided. When a variable is correlated with another variable, we say it shares variance with the other variable. The amount of sharing is the squared correlation. The total variance of any variable can be divided into 3 types:
  1. **Common variance** = The variance that is shared with all other variables in the analysis. The communality is the estimate of the shared variance.
  2. **Specific variance** = The variance only associated with a specific variable.
  3. **Error variance** = Variance due to unreliability in the data-collection process, measurement error, or a random component in the measured phenomenon.

Common factor analysis versus component analysis

The selection of the method is based on two criteria: The objectives of the factor analysis and the amount of prior knowledge about the variance in the variables.

1. **Component analysis** is used when the objective is to summarize most of the original information in a minimum number of factors for prediction purposes. It considers the total variance and derives factors that contain small portions of unique variance and error variance. This method is the most appropriate when data reduction is a primary concern and when prior knowledge suggests that specific and error variance represent a relatively small proportion of the total variance.
2. **Common factor analysis** is used primarily to identify underlying factors or dimensions that reflect what the variables share in common. It considers only the common variance assuming that both the unique and error variance are not of interest in defining the structure of the variables. This method is most appropriate when the primary objective is to identify the latent dimensions or constructs represented in the original variables and the researcher has little knowledge about the amount of specific and error variance and wishes to eliminate variance. Common factor analysis is often viewed as more theoretically based. However, there are a view problems.

- Common factor analysis suffers from factor indeterminacy: for any individual respondent, several different factor scores can be calculated from a single factor model result.
- The communalities are sometimes not estimable or may be invalid

Criteria for the number of factors to extract.

Although considerable debate remains over which factor model is the most appropriate, empirical research demonstrates similar results in many cases. How do we decide on the number of factors

to extract? The first factor may be viewed as the single best summary of linear relationships exhibited in data. The second factor is defined as the second-best linear combination of the variables, subject to the constraint that is **orthogonal** to the first factor. To be orthogonal to the first factor, the second factor must be derived from the variance remaining after the first factor has been extracted. The process continues extracting factors accounting for smaller and smaller amounts of variance until all of the variance is explained. In deciding when to stop factoring, the researcher must combine a conceptual foundation with some empirical evidence. He generally begins with some predetermined criteria, such as the general number of factors plus some general thresholds of practical relevance. These criteria are combines with empirical measures of the factor structure. The following stopping criteria for the number of factors have been developed:

- **Latent root criterion**. Any individual factor should account for the variance of at least a single variable if it is to be retained for interpretation. Only factors having latent roots or eigenvalues of more than 1 are considered significant. This method is most reliable with a number of variables between 20 and 50.
- **A priori criterion**. When applying this method, the researcher already knows how many factors to extract before the factor analysis. The researcher instructs the computer to stop the analysis when the desired number of factors has been extracted.
- **Percentage of variance criterion**. This approach is based on achieving a specified cumulative percentage of total variance extracted by successive factors.
- **Scree test criterion**. The scree test is used to identify the optimum number of factors that can be extracted before the amount of unique variance begins to dominate the common variance structure.
- **Heterogeneity of the respondents**. Shared variance among variables is the baseis for both common and component analysis. If the sample is heterogeneous with regard to at least one subset of the variables, then the first factors will represent those variables that are more homogeneous across the entire sample.

  Researchers usually use more than one method in determining how many factors to extract. After the factors are interpreted, the practicality is assessed. Negative consequences can arise from selecting either too many, or too few factors to represent the data. With the use of too few factors, the structure is not revealed. With the use of too many factors, the interpretation becomes too complex when the results are rotated.

Stage 5- interpreting the factors

A strong conceptual foundation is of great importance in interpreting the factors.

The three processes of factor interpretation.

To assist in the process of interpreting a factors and selecting final factor solution, three fundamental processes are defined.

1. Estimate the factor matric. First, the un rotated factor matrix is computed, containing the factor loadings for each variable on each factor. Factor loadings are the correlation of each variable and the factor.
2. Factor rotation. Factor rotation should simplify the structure. Rotation is applied to achieve simpler and theoretically more meaningful factor solutions. Ambiguities are reduced.
3. Factor interpretation and re specification. As a final process the researcher evaluates the factor loadings for each variable in order to determine that variables role and contribution in determining the factor structure. The need may arise to respecify the factor model owing to the deletion of a variable from the analysis, the desire to employ

a different rotational method, the need to extract a different number of factors or the desire to change from one extraction method to another.

Rotation of factors.

Perhaps the most important tool in interpreting is factor rotation. The reference axes of the factors are turned about the origin until some other position has been reached. The ultimate effect of rotating the factor matrix is to redistribute the variance from earlier factors to later ones to achieve a simple, theoretically more meaningful factor pattern. The simplest case of rotation is an orthogonal factor rotation, in which the axes are maintained at 90 degrees. When a rotation is not orthogonal, it is called oblique factor rotation. Those different types of rotation are shown in figure 3.7 and 3.8. Oblique rotation represents the clustering of variables more accurately and the oblique solution provides information about the extent to which the factors are actually correlated with each other.

- **Orthogonal Rotation Methods**
  By simplifying the rows, we mean bringing as many values as possible as close to 0 as possible. Three major approaches have been identified:
  1. Quartimax rotation is used to simplify the rows of a factor matrix. It focusses on rotating the initial factor so that a variable loads high on one factor and as low as possible on all other factors. It has not proved especially successful in producing simpler structures. It tends to produce a general factor as the first factor, on which most of the variables have high values.
  2. Varimax criterion centers on simplifying the columns of the factor matrix. The maximum simplification is reached when there are only 0s and 1s in a column. This method maximizes the sum of variances of required loadings of the factor matrix. Varimax seems to give a clear separation of the factors.
  3. Equimax approach is a compromise between the Quartimax and Varimax approaches. It tries to accomplish both simplification of rows and simplification of columns. It is, however, used infrequently.
- **Oblique Rotation Methods**
  These are similar to orthogonal methods, except that these allow correlated factors instead of remaining independence. Most statistical programs only offer limited options for oblique methods. (SPSS: OBLIMIN)
  No specific rule has been developed for choosing the right method. In most cases the researcher uses what the computer program offers.

Judging the significance of factor loadings

- **Ensuring practical significance**. A factor loading is the correlation of the variable and the factor, so the squared loading is the amount of the variables total variance accounted for by the factor. The larger the absolute size of the factor loading, the more important the loading in interpreting the factor matrix. We can assess the loadings as follows:
  - +/- .30 to +/- .40 meet the minimal level for interpretation of the structure.
  - +/- .50 or greater are practically significant
  - exceeding .70 are indicative of well-defined structure and are the goal of any factor analysis.
- **Assessing statistical significance**. Research has demonstrated that factor loadings have larger standard errors than typical correlations. The researcher can use the concept of statistical power in assessing significance for different sample sizes.
- **Adjustments based on the number of variables**. A disadvantage of both methods is that

both of the approaches do not consider the number of variables As the researcher moves from the first factor to later factors, the acceptable level for a loading to be judges significant should increase.

Interpreting a factor matrix.

The researcher must sort through all the factor loadings to identify those most indicative of the underlying structure. Interpreting the complex interrelationships represented in a factor matrix represented in a factor matrix requires a combination of applying objective criteria with managerial judgment. The process can be simplified by following a 5-step process:

Step 1: examine the factor matrix of loadings

*T*ypically, the factors are arranged as columns thus, each column of numbers represents the loadings of a single factor. If an oblique method is used, two matrices of factor loadings are provides. The first is the factor pattern matrix, which has loadings that represent the unique contribution of each variable to the factor. The second is the factor structure matrix, which has simple correlations between variables and factors, but these loadings contain both the unique variance between the factors and the correlation among factors.

Step 2: identify the significant loadings for each variable

the interpretation should start with the first variable and move horizontally from left to right, looking for the highest loading for that variable on any factor. When the highest loading is found, it should be underlined if significant . Most factor solutions do not result in a simple structure solution. When a variable has more than one significant loading, it is termed a cross-loading.

Step 3: Assess the communalities of the variables.

Once the significant loadings are identifies, the researcher should look for any variables that are not adequately accounted for by the factor solution the researcher should view to communalities to assess whether the variables meet acceptable levels of explanation.

Step 4: respect the factor model if needed.

The researcher may find one of the following problems: (a) a variable has no significant loadings (b) even with a significant loading, the communality is deemed too low (c) a variable has cross-loading. The researcher can apply the following remedies:

- Ignore those problematic variables and interpret the solution as it is.
- Evaluate each of these variables for possible deletion.
- Employ an alternative rotation method.
- Decrease/ increase the number of factors retained.
- Modify the type of factor model used.

Step 5: Label the factors

The researcher will examine all the significant variables for a particular factor and will attempt to assign a name or label selected to represent a factor that accurately reflects the variables loading on that factor. On each factor, like signs mean the variables are positively correlated and the opposite sign mean they are negatively correlated.

Stage 6- validation of factor analysis

In this stage the degree of generalizability is tested. It is especially relevant for the interdependence methods, because they describe a data structure that should be representative of

the population as well.

- Use of a confirmatory perspective. The most direct method of validating the results is moving to a confirmatory perspective and assess the replicability of the results, either with split sample in the original data set or with a separate sample. The comparison between two or more factor model results has always been problematic, but several methods exist to make a comparison. Confirmatory factor analysis is one option, but several other options have also been proposed, ranging from a simple matching index to programs designed specifically to assess the correspondence between factor matrices.
- Assessing factor structure stability. Another aspect of generalizability is the stability of factor model results, factor stability is dependent on the sample size and the number of cases per variable. If the sample size permits, the researcher may wish to randomly split the sample into two subsets and estimate the factor model for each subset. Comparison of the two resulting factor matrices will provide an assessment of the robustness of the solution across the sample.
- Detecting influential observations. In addition to generalizability, another issue is important to the validation of factor analysis: The detection of influential observations. The researcher is encouraged to estimate the model with and without the observations identified as outliers to assess their impact on the results.

Stage 7- additional uses of factor analysis results

If the objective is to identify appropriate variables for subsequent application on other statistical techniques, some form of data reduction will be employees. The two options include the following.

- Selecting the variable with the highest factor loading as a surrogate representative for a particular factor dimension.
- Replacing the original set of values with an entirely new, smaller set of variables created either from summated scales or factor scores.

Selecting surrogate variables for subsequent analysis

If the objective is to identify appropriate variables for subsequent analysis with statistical techniques, the researcher can chose to examine the factor matrix and select the variable with the highest factor loading on each factor to act as surrogate variable that is representative for that factor. This method has several disadvantages:

- It does not address the issue of measurement error encountered when using single measures.
- It also runs the risk of potentially misleading results by selecting only a single variable to represent a perhaps more complex result.

Creating a summated scale

A summated scale creates two benefits:

1. It provides a means of overcoming to some extent the **measurement error**. It reduces measurement error by using multiple **indicators** to reduce the reliance on a single response.
2. It has the ability to represent the multiple aspects of a concept in a single measure. The summated scale combines the multiple indicators into a single measure representing what is held in common across the set of measures.

Four issues basic to the construction of the summated scale:

- **Conceptual definition**. This specifies the theoretical basis for the summated scale by defining the concept being represented in terms applicable to the research context. **Content validity** is the assessment of the correspondence of the variables to be included in the summated scale and its conceptual definition. This form of validity, also known as face validity, subjectively assesses the correspondence between the individual items and the concept through ratings by expert judges.
- **Dimensionality**. The items should be unidimensional, meaning that they are strongly associated with each other and represent a single concept. The researcher can assess unidimensionality with either exploratory factor analysis, or confirmatory factor analysis.
- **Reliability**. Reliability is an assessment of the degree of consistency between multiple measurements of a variable. One form of reliability is test- retest, by which consistency is measured between the responses for an individual at two points in time. A second measure is internal consistency, which applies to the consistency among variables in a summated scale. Because no measure is perfect, we must rely on series of diagnostic measures to assess internal consistency.
  - The first measures relate to each seperable item
  - The second type of diagnostic measure is the reliability coefficient, **Cronbach's alpha**.
  - Also available are reliability measures derived from confirmatory factor analysis.
- **Validity**. The extent to which the scale or set of measures accurately represents the subject of interest. The three most widely used forms of validity are:
  - Convergent validity. It assesses the the degree to which two measures of the same concept are correlated.
  - Discriminant validity is the degree to which two conceptually similar concepts are distinct.
  - Nomological validity refers to the degree that the summated scale makes accurate predictions of the other concepts in a theoretical based model.

Calculating summated scales.

The most common approach is to take the average of the items in the scale. Reverse scoring is the process whereby the data values of a variable are reversed so that its correlations are now positive within the factor.

Computing factor scores

The third option for creating a smaller set of variables to replace the original set of variables is the computation of factor scores. Factor scores are also composite measures of each factor computed for each subject. The factor score is computed based on the factor loadings of all variables on the factor, whereas the summated scale is calculated by combining only selected variables. The only disadvantage is that they are not easily replicated across studies because they are based on the factor matrix.

Selecting among the three methods of data reduction:

- If data are used in the original sample or orthogonality must be maintained, factor scores are suitable
- If generalizability is desired, summated scale or surrogate variables are more appropriate.
- If a summated scale is untested and exploratory, surrogate variables should be considered.

For an example of a factor analysis, refer to page 127-145