

PROC FACTOR: How to Interpret the Output of a Real-World Example

Rachel J. Goldberg, Chamblee, GA

ABSTRACT

This paper summarizes a real-world example of a factor analysis with a VARIMAX rotation utilizing the SAS® System's PROC FACTOR procedure. Each step that one must undergo to perform a factor analysis is described – from the initial programming code to the interpretation of the PROC FACTOR output. The paper begins by highlighting the major issues that you must consider when performing a factor analysis using the SAS System's PROC FACTOR. This is followed by an explanation of sample PROC FACTOR program code, and then a detailed discussion of how to interpret the PROC FACTOR output. The main focus of the paper is to help SAS software beginning and average skill level users learn how to interpret programming code and output from PROC FACTOR. Some knowledge of Statistics and/or Mathematics would be helpful in order to understand parts of the paper. All of the results discussed utilize Base SAS and SAS/STAT® software.

INTRODUCTION

Factor analysis can be performed for various reasons, such as:

- ◆ Exploratory data analysis prior to further analysis
- ◆ Broad explanation of the data
- ◆ Testing hypothesized explanations of the data
- ◆ Data reduction

Regardless of what the particular motive is for conducting a factor analysis, the results of a factor analysis identify the factors, or dimensions, that are responsible for the covariation in the data. The SAS System's PROC FACTOR provides an efficient manner in which to perform a factor analysis, no matter what the specific interests are of the user.

To glean meaningful results from a factor analysis several issues need to be addressed before running PROC FACTOR, correct SAS software code for running PROC FACTOR has to be written, and proper interpretation of the output from PROC FACTOR must take place. This paper first will introduce the preliminary analysis phase, which includes a discussion of the major issues that need to be considered *before* running PROC FACTOR. Following the preliminary analysis discussion, the paper will explain an example of a PROC FACTOR program code. Then, the paper details how to interpret the results of the example PROC FACTOR output. Finally, the paper concludes with suggestions for validating the PROC FACTOR results.

PRELIMINARY ANALYSIS

It should be noted that it is beyond the scope of this paper to discuss the detailed statistical and mathematical explanations for the different types of factor analysis. Thus, this paper provides only basic explanations and does not attempt to give the reader all of the information needed to make an educated decision about what type of factor analysis to perform. For more theoretical information, see the references noted at the end of this paper.

THE STEPS

As with any data analysis and interpretation, there are certain data cleaning procedures that should be performed prior to beginning an in-depth look at the data. The steps to consider prior to running PROC FACTOR include (but are not limited to) the following:

- ◆ Obtaining a large enough sample size (generally, the number of observations should be greater than or equal to five times the number of variables),
- ◆ Verifying that all data values are valid,
- ◆ Removing of any data outliers,
- ◆ Reducing data to the relevant analysis variables,
- ◆ Deciding what type of factor analysis to perform, and
- ◆ Ensuring that the proper assumptions for the statistical procedure have been met.

THE METHOD

The SAS System provides several options for the PROC FACTOR initial factoring method that is used for analysis, such as Alpha, Maximum-Likelihood, Principal Components Analysis, and Principal Factor Analysis. Each of these initial factoring methods generates uncorrelated factors. Arguments exist supporting all of the different initial factoring methods available. For the example discussed in this paper, it was assumed that the main objective during the initial factor analysis was to determine the minimum number of factors that would adequately account for the covariation among the larger number of analysis variables. This objective can be achieved by using any of the initial factoring methods. Therefore, the simplest method was used, Principal Factor Analysis, which is further discussed in the EXAMPLE PROC FACTOR PROGRAM section below.

THE ASSUMPTIONS

As with every statistical procedure, certain assumptions must be satisfied in order for a factor analysis to be considered statistically sound. The mathematical computations for a factor analysis are performed on a matrix of correlations. Therefore, the same assumptions that must be met when analyzing correlation statistics, such as the Pearson Correlation Coefficient, must be met when a factor analysis is performed. As long as the sample size is greater than 25, there are four major assumptions which should be met for a factor analysis:

- 1) the data sample should be **RANDOMLY SELECTED** from the population,
- 2) the variables should be **NUMERICAL**, either from the interval scale and/or the ratio scale of measurement,
- 3) the variables should each have an approximately **NORMAL DISTRIBUTION**, and
- 4) the variables should have a **LINEAR RELATIONSHIP**.

If the sample size is less than 25, there is one additional assumption: the pairs of variables should have a **BIVARIATE NORMAL DISTRIBUTION**.

TO ROTATE, OR NOT TO ROTATE?

It is generally considered that using a rotation in factor analysis will produce more interpretable results. If the factor analysis is being performed specifically to gain an explanation of what factors or groups exist in the data or to confirm hypothesized assumptions about the data, rotation can be especially helpful. "Rotating" the factor pattern simply means performing a linear transformation on the factor solution. Factor patterns can be rotated through two different ways:

- ◆ Orthogonal rotations which retain uncorrelated factors
- ◆ Oblique rotations which create correlated factors

Arguments exist supporting both types of rotation methods. Factor analysis which uses an orthogonal rotation often creates a solution that is simpler to interpret, and, thus, one that is easier to grasp than a solution obtained from an oblique rotation.

The factor analysis example discussed in this paper was performed for exploratory data analysis purposes to discover simplified factor or dimension descriptions that exist in the data. For such exploratory data analysis, any rotation method can be used. One of the orthogonal methods, VARIMAX, was used in this example, because it was easily interpretable. More details about the VARIMAX rotation method are discussed in the EXAMPLE PROC FACTOR PROGRAM and OUTPUT sections below.

GOALS FOR RUNNING PROC FACTOR

For the example included in this paper, the overall goals for running PROC FACTOR are the following:

- ◆ Determine the minimum number of factors that can adequately account for the variance in the data,
- ◆ Find simpler, easier to interpret factors through rotating the factors, and
- ◆ Determine a meaningful interpretation of the factors that provides insight into the data for potential use with further analysis.

BACKGROUND

The real-world example that is discussed is based on a factor analysis performed on data from a very large energy services company. The data that is analyzed is the customer and energy usage information that was available for a population that is serviced by the energy services company.

EXAMPLE PROC FACTOR PROGRAM

This section describes each of the options invoked in the following PROC FACTOR program.

```
PROC FACTOR DATA = SAVE.EXAMP
  METHOD = PRINCIPAL PRIORS = SMC SCREE
  NFACTOR = 4 PREPLOT
  ROTATE = VARIMAX REORDER PLOT
  OUT = SAVE.EXAMPFAC;
  VAR X43 - X78;
  RUN;
```

DATA = data set

This names the data set that contains the observations to be factored. Many special (TYPE = x) SAS System created data sets can be used for input into the PROC FACTOR program. For this example, a simple RAW data set was used.

METHOD = method

This specifies what type of method is to be used to extract the initial factors. As discussed earlier, the Principal Factor Analysis method (PRINCIPAL) was chosen for the initial factor extraction method. Without the PRIORS = SMC option also specified (the PRIORS option is discussed in the subsection below), the METHOD = PRINCIPAL would output a Principal Components Analysis (PCA). While there are many similarities between a Principal Factor Analysis and a Principal Components Analysis, one major difference between them is that a Principal Components Analysis outputs components that are linear combinations of the observed variables. However, with the Principal Factor Analysis method, the data is summarized by means of a linear combination of the underlying factors or dimensions.

The Principal Factor Analysis method meets the objective mentioned earlier in this paper: for the factors to account for as much variation as possible in the data. With Principal Factor Analysis, the first factor is defined in such a way that the largest amount of variance in the data is explained by the first factor. The second factor that is identified explains the second most about the variance in the data that has not been explained already AND is perpendicular (thus, uncorrelated) to the first factor. The remaining factors are found in a similar manner.

PRIORS = prior communality estimates

Variance found in the data set can be broken into three parts. First, the common variance is the variance that all of the variables share with each other. Secondly, is the unique variance. The unique variance is what each variable contributes on its own and does not share with the other variables. Finally, the total variance is simply the sum of the common variance plus the unique variance.

In Principal Factor Analysis, the factors account for the common variance of the variables, and they do not account for the portion of the variance introduced by each individual variable. The portion of the common variance that is accounted for by the factors is called each variable's communality. This is another difference between Principal Factor Analysis and Principal Components Analysis. While Principal Factor Analysis accounts just for the *common* variance of the variables, Principal Components Analysis accounts for the *total* variance of the variables.

The extraction of factors through the Principal Factor Analysis method that only accounts for the *common* variance of the variables is produced by the use of an *adjusted* correlation matrix. The adjusted correlation matrix places the communality estimates on the central diagonal of the matrix. However, the prior communality estimates are unknown. Thus, a method must be selected to estimate them. One common method used for these estimates is the squared multiple correlation between each variable and all other variables, referred to as SMC.

The default option for the prior communality estimate in the SAS System's PROC FACTOR procedure is ONE. By using values of one (1.0) for the prior communality estimates, which is setting the diagonal values on the correlation matrix to one (1.0), an *unadjusted* correlation matrix is used. This would be appropriate for conducting a Principal Components Analysis since all of the variance would be

accounted for, not just the variance that the variables have in common. While there are reasons to conduct such an analysis, those reasons do not include one of the underlying purposes for conducting exploratory factor analysis: to account for the factor structure underlying the variables. This factor structure can be determined through a Principal Factor Analysis. Thus an *adjusted* correlation matrix is used to extract the factors, and the communalities must be estimated. The common SMC method is used in this example PROC FACTOR program.

SCREE

This option requests that a scree plot of the eigenvalues be printed, which will be discussed in the EXAMPLE PROC FACTOR OUTPUT section.

NFACTOR = n

This option specifies the number of factors to be extracted from the data. The SAS System's PROC FACTOR default value, if this option is not specified, is the number of factors which contribute to explaining 100% of the variance. In the example PROC FACTOR program discussed in this paper, the default for the NFACTOR option initially was used. However, after further analysis, just four factors were extracted from the data by specifying NFACTOR = 4. This number was selected by first running PROC FACTOR without the NFACTOR option and analyzing the eigenvalues and scree plot. Based on this eigenvalue and scree plot analysis, which will be discussed in more detail in the EXAMPLE PROC FACTOR OUTPUT section, four factors were kept in the final solution.

PREPLOT

This option specifies that a plot be produced of the factor pattern prior to any rotation. Each factor will be plotted individually with every other factor. Therefore, it is best to include this option after determining a reduced number of factors to retain for further iterations. In this example program, the PREPLOT option was included after determining that four factors would be kept (using NFACTOR=4) in the final factor analysis solution. Plotting the factors prior to rotation demonstrates why it is best to apply a rotation method to the factor analysis. This will be explained more thoroughly in the EXAMPLE PROC FACTOR OUTPUT section.

ROTATE = method

This option invokes one of the rotation methods available. If no method is specified, the SAS default is to use no rotation. Therefore, if no rotation method is selected, the initial method that is selected will be the only method used during the factoring procedure. As discussed earlier, this paper's example PROC FACTOR program specifies the VARIMAX orthogonal rotation method. VARIMAX rotation is a transformation that simplifies the interpretation of the factors by maximizing the variances of the squared loadings for each factor. That is, it maximizes the variance of the columns of the factor pattern. The reason that the VARIMAX rotation method was chosen is further discussed in the EXAMPLE PROC FACTOR OUTPUT section.

REORDER

This option reorders the output from the factor procedure, so the variables that explain the largest amount of the variance for factor one are printed in descending order down to those that explain the smallest amount of the variance for factor one. The remaining factors are printed in a similar manner.

PLOT

This option specifies that a plot be produced of the factor pattern after the rotation method has been applied. In this example, the factor pattern that is plotted will reflect the VARIMAX rotation method. Each factor will be plotted individually with every other factor. Therefore, as with the PREPLOT option, it is best to include PLOT option after determining a reduced number of factors to retain for further iterations. In this example program, the PLOT option was included after determining that four factors would be kept (using NFACTOR=4) in the final factor analysis solution. Plotting the factors after rotation and comparing them to the plots produced prior to rotation demonstrates why it is best to apply a rotation method to the factor analysis. This will be explained more thoroughly in the EXAMPLE PROC FACTOR OUTPUT section.

OUT = data set

This specifies that an output data set be created. This data set will contain all of the data from the original input data set and the new factor variables, which are called FACTOR1, FACTOR2, etc. These new variables contain the estimated factor scores, which can be used in further statistical analysis.

EXAMPLE PROC FACTOR OUTPUT

PRIOR COMMUNALITY ESTIMATES

This is a listing of the prior communality estimates. These were estimated due to the specification PRIOR = SMC (the squared multiple correlations). The estimates were derived by applying a regression technique that regresses a variable's squared multiple correlation (SMC) on the other variables in the data.

EIGENVALUES OF THE CORRELATION MATRIX

What exactly are eigenvalues? They are values that consolidate the data variance in a matrix (the eigenvalues) while providing the linear combination of variables (the eigenvector) to do it. PROC FACTOR was initially run by allowing all of the variables entered into the procedure to be possible factors. For exploratory data analysis purposes, this is a preferred way to run the initial factor analysis, so that all of the factors can be analyzed for potential inclusion in the final iterated solution. Determining how many factors will be included in the final factor program partially is based on the analysis of the eigenvalues. Thus, in this paper's example PROC FACTOR program, NFACTOR = 4 was specified after analyzing, among other criteria, the eigenvalues. The eigenvalues for the four factors retained for the final solution are shown below.

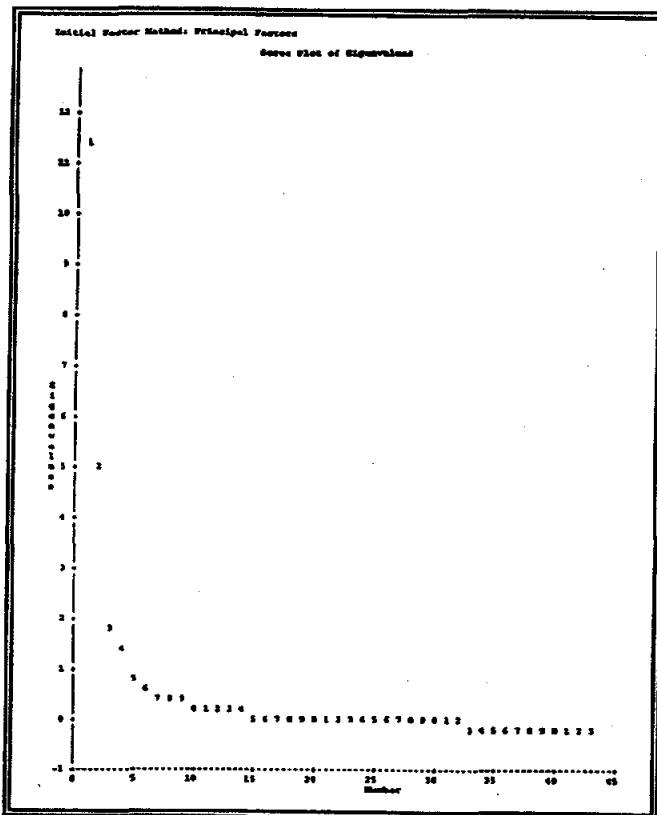
EIGENVALUES OF THE CORRELATION MATRIX					
	1	2	3	4	5
Eigenvalue	11.3937	4.9335	1.7807	1.3238	0.7417
Difference	6.4602	3.1529	0.4568	0.5822	0.0955
Proportion	0.5482	0.2374	0.0857	0.0637	0.0357
Cumulative	0.5482	0.7856	0.8713	0.9350	0.9707

The above eigenvalues are a small portion of the chart of the eigenvalues from the full PROC FACTOR output. In the full eigenvalue chart, the sum of the eigenvalues is displayed, which is the sum of the estimated communalities.

For the eigenvalues displayed in the previous chart, the DIFFERENCE between successive values, the PROPORTION of the variation represented, and the CUMULATIVE proportion of the variation represented is shown. Judgmental criteria usually are set for determining how many factors to retain in the final solution. For example, how much variance a particular eigenvalue must show proportionally (that is, how much variance the factor represented by the eigenvalue must account for), in order to be kept for further analysis can be decided before running the initial factor program. Another such criteria is to set a minimum amount of the common variance that must be accounted for overall, which can be determined by reviewing the eigenvalue's CUMULATIVE proportion of variance numbers. While such criteria are very subjective methods for determining how many factors should be kept, they can be used in conjunction with an analysis of the scree plot to make a more educated decision (the SCREE PLOT is discussed in the next subsection).

In this example, when analyzing the eigenvalue chart alone to determine how many factors to keep in the solution, the CUMULATIVE proportion of explained variance was reviewed. Factors were kept that contributed to a total summed CUMULATIVE proportion of explained variance of at least 90%, and that demonstrated the factor explained at least a five percent (5%) PROPORTION of the variance. As the previous sample eigenvalue output shows, the last factor that had an eigenvalue PROPORTION greater than or equal to five percent (5%) and that had contributed to the 90% total summed CUMULATIVE proportion of explained variance was Factor 4. Hence, this is one argument for keeping the first four factors for the remainder of the factor analysis.

SCREE PLOT OF EIGENVALUES



Another contributing criteria for determining how many factors should be kept in the remainder of the analysis is an examination of the SCREE PLOT. What is a SCREE PLOT? The SCREE PLOT simply plots the eigenvalues for each of the factors, from the first eigenvalue (the one that explains the most variance) to the last eigenvalue.

On the sample SCREE PLOT produced by PROC FACTOR, a line can be drawn that runs between each of the eigenvalues. Such a line would show where the slope levels off as the amount of variance that is explained by each eigenvalue becomes minimal. Thus, towards the bottom of the slope, the eigenvalues are not that much different from each other. Where the line levels off is where the "scree" really shows in the data: that is, it is where the debris, or unnecessary information, collects. Hence, where the line levels off indicates the point of diminishing returns. It is this general area that should be analyzed to determine which eigenvalues are providing enough information to warrant inclusion in further analysis. It is worth noting that it is considered "OK" to have too many as opposed to too few factors included in exploratory factor analysis.

For the example discussed here, by analyzing the SCREE PLOT alone, the slope starts to level off after the fourth eigenvalue. It really levels off after the sixth eigenvalue. Therefore, Factor 4, 5, and 6 were analyzed to determine which one should be the last one retained. After this determination based on just the SCREE PLOT, the actual eigenvalues were reviewed again. As stated previously, criteria had been set prior to running the PROC FACTOR program. The criteria were that a 90% CUMULATIVE proportion of the variance should be accounted for and that each factor should have a PROPORTIONAL contribution of at least five percent (5%) of the variance explained. Factor 4 was the last factor to meet these criteria, and Factor 4 was one of the factors that the SCREE PLOT analysis showed to be close to the point of diminishing returns. Hence, only four factors were retained for the final solution.

FACTOR PATTERN

The elements of the Factor Pattern reflect the variance each factor contributes to the variance that an observed variable has in common with the other observed variables. The actual "loadings", or numbers listed for each factor, are the bivariate correlations between each variable and the factor in which the loading is listed. The loading can therefore range from a negative one (-1) to a positive one (+1), depending on whether the relationship is negative or positive. Variables that have at least a moderately high "loading" on a factor are said to help explain the meaning of that factor. This makes sense since a moderately high "loading" is equivalent to a moderately high correlation between a variable and a factor.

In an ideal solution, the variables should "load" moderately to highly (have a value that approaches a positive or negative one) on just one factor each. In this paper's example, only factor loadings of a positive or negative point five (+/- 0.5) were analyzed. For other theories on appropriate factor loadings, see the references listed at the end of this paper.

The reason factor analysis often is not stopped after this initial factoring stage, without rotating the factors, is that the factors as they currently exist are not easily interpretable. That is, prior to rotation, variables often load moderately high on more than one factor. This would indicate that the variable has a relationship with more than one factor, making it difficult to determine unique factor descriptions.

INITIAL FACTOR METHOD: PRINCIPAL FACTORS

FACTOR PATTERN

	FACTOR1	FACTOR2	FACTOR3	FACTOR4
X35	0.64412	0.48963	-0.10616	-0.03723
X30	0.59214	0.58449	-0.03765	0.04250

In the previous sample of the PROC FACTOR output, the initial factor pattern is displayed. By analyzing variable X35, it is apparent why this factor pattern must be rotated. Variable X35 has a loading of 0.64412 (a moderately high loading) in Factor 1 and a loading of 0.48963 (a moderate loading) in Factor 2. Ideally, this variable would have a moderate to high loading in only one factor, which would make it easily interpretable. Because many variables loaded in a similar manner as X35, the factor pattern was rotated, using the VARIMAX rotation method.

VARIANCE EXPLAINED BY EACH FACTOR

The output below simply summarizes the amount of variance in the data that is explained by each factor. The numbers listed here, prior to rotating the factor pattern, are equal to the corresponding eigenvalues.

INITIAL FACTOR METHOD: PRINCIPAL FACTORS

VARIANCE EXPLAINED BY EACH FACTOR

FACTOR1	FACTOR2	FACTOR3	FACTOR4
11.393712	4.933538	1.780656	1.323822

As you can see by looking at the above sample output, FACTOR1 starts with a variance value of over 11, and Factor 4 has a value of just over one (1). When this example program was first run with no restriction on the number of factors, the factors past Factor 4 had values less than one (1). During further iterations of the program, the factors that were not explaining very much of the data variance were eliminated.

FINAL COMMUNALITY ESTIMATES

The table of the final communality estimates simply shows the squared multiple correlations (SMC's) for predicting the variables from the estimated factors. The communality for a variable can be derived by taking the sum of squares of a variable's loading on each factor and summing these squares. The communality is the variance of the observed variables that is accounted for by each factor.

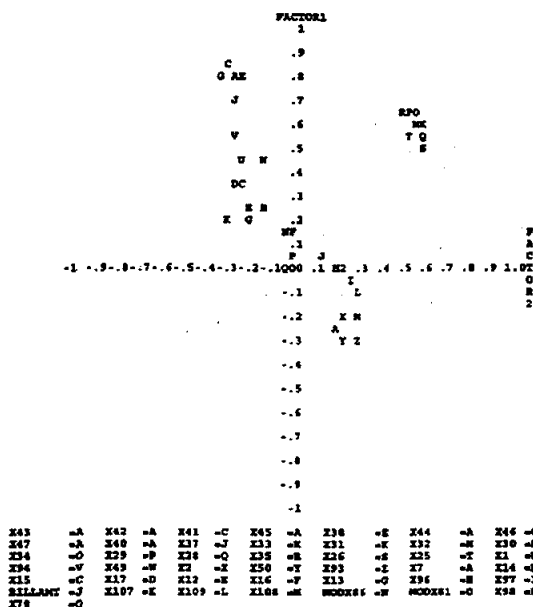
PREPLOT

By invoking the PREPLOT option, individual plots were produced of each factor versus each other factor prior to rotation. For this example PROC FACTOR program, the PREPLOT option was not invoked until it had been determined that just four factors would be retained for the final solution. The plots produced in this example showed the four factors' loadings prior to rotation.

The example output that follows shows the plot of Factor 1 versus Factor 2 prior to rotation. It is obvious that the factors are not quite explaining the variance in the data: the points do not fall very close

to the factors. If the variables loaded highly on just one factor each, the points would fall closer to the factors, making the solution easier to interpret. Producing the plots is a great way to visually understand why it is difficult to interpret factors prior to rotation.

Initial Factor Method: Principal Factors
Plot of Factor Pattern for FACTOR1 and FACTOR2



ORTHOGONAL TRANSFORMATION MATRIX

This output is simply the orthogonal transformation matrix that was used during the factor rotation. Because an orthogonal rotation was specified, VARIMAX in this example, the rotated factors will remain uncorrelated.

ROTATED FACTOR PATTERN

Because a rotated factor analysis was specified, another factor pattern was output. As stated earlier, the orthogonal rotation methods are considered to produce the most easily interpreted results. Therefore, the PARSIMAX, EQUAMAX, and VARIMAX orthogonal rotation methods were tested for this example factor analysis.

While the VARIMAX rotation method was finally chosen, it should be noted that the sample output that follows was produced after running several iterations of the PROC FACTOR program. After just one run, some variables had a moderate to high loading on more than one factor. These variables were deleted from the procedure, and then PROC FACTOR again was run. This process was repeated until all of the final variables loaded moderately to highly on just one factor each. To ensure that the rotation method was not the cause of some variables loading moderately to highly on more than one factor, the factor loadings were analyzed from all three orthogonal rotation methods that initially were tested. All of the rotation methods required several iterations with variables deleted in each step of the process. Based on the final interpretation of the factors, the size of the loadings, and the number of variables that loaded moderately to highly on one factor each, it was decided that the VARIMAX rotation method was the "best" choice for this example program.

ROTATION METHOD: VARIMAX				
ROTATED FACTOR PATTERN				
	FACTOR1	FACTOR2	FACTOR3	FACTOR4
X35	0.26063	0.77412	0.00329	-0.00953
X30	0.16019	0.81256	0.03465	0.09147

The output above is a sample of this paper's example rotated factor pattern. To determine whether or not the rotation made interpretation easier, look again at variable X35. As the sample PROC FACTOR output above shows, variable X35 now loads on FACTOR1 as 0.26063 and loads on FACTOR2 as 0.77412. This is much closer to the ideal solution in which each variable loads high (approaching a value of 1) on only one factor.

VARIANCE EXPLAINED BY EACH FACTOR

ROTATION METHOD: VARIMAX				
VARIANCE EXPLAINED BY EACH FACTOR				
	FACTOR1	FACTOR2	FACTOR3	FACTOR4
	8.087913	6.897154	2.419619	2.027042

The above sample PROC FACTOR output shows the variance explained by each factor after rotation. While Factor 1 is explaining less of the variance than it did prior to rotation, Factors 2, 3 and 4 explain more of the variance than prior to rotating the factors. The overall amount of the variance being explained has remained the same; it has simply been redistributed.

FINAL COMMUNALITY ESTIMATES

The final communality estimates have not changed since the factors were rotated. This is because rotating the factors does not affect how much covariance is explained; it simply produces more interpretable results.

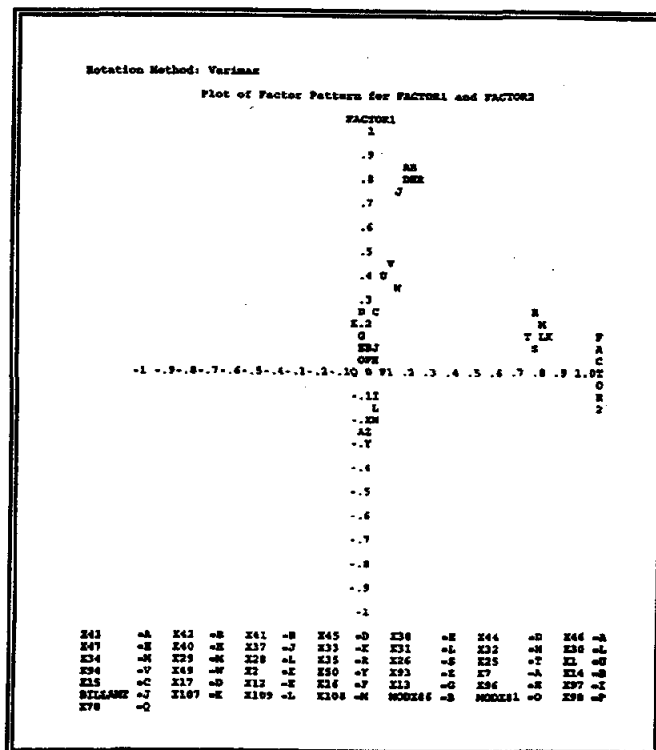
STANDARDIZED SCORING COEFFICIENTS

If one of the goals in performing factor analysis was to find factors to use in subsequent analyses, the estimated values of the factors, which are the factor scores, would need to be produced. In this paper's energy services example, since an output data set was requested, the factors have been output into the data set along with the original raw data. Additionally, the standardized scoring coefficients of each variable for all four factors are automatically printed. Generally, the individual variables' factor scores are not used, the scores are used as a whole to represent the factor and are called factor scores.

PLOT

By invoking the PLOT option, plots were produced for each rotated factor by each other rotated factor. As with the PREPLOT option, the PLOT option was not specified until the decision had been made regarding the final number of factors to retain. The plots that are produced show the factor loadings after rotation.

The example output that follows shows the plot of the rotated Factor 1 versus the rotated Factor 2. If this plot is compared to the PREPLOT example output of the same two factors prior to rotation, it is obvious that the factors are much more closely aligned with the factor loadings after rotation has taken place. Hence, because after rotation the variables are more likely to load highly on just one factor each, the rotated factor solution should be easier to interpret. Printing the plots of the final rotated factors allows for a visual analysis of the results. This can often contribute to a certain level of comfort in the final rotated factor analysis solution.



INTERPRETATION OF PROC FACTOR OUTPUT

NEXT STEPS

From the PROC FACTOR output given directly from the SAS System, several things can be done, such as:

- ◆ Using the factor scores in subsequent data analyses
- ◆ Analyzing the factors to determine if each factor describes a particular dimension in the data
- ◆ Determining if a preliminary hypothesis about the data is correct
- ◆ Deleting variables that do not load highly in any factor, which eliminates unnecessary information from the data

In this paper's energy services example, as is often the case when factor analysis is applied in a real situation (as opposed to in the classroom), the next steps involved more than one of the above items. First, the factors were analyzed to determine what unique dimensions existed in the data. Second, the factor scores were used in subsequent analyses. The remainder of this paper will focus on the first use of this example PROC FACTOR output, creating a description of the dimensions in the energy services data.

DESCRIBING THE FACTORS

It is necessary to describe the four factors to ensure that meaningful insight has been gained into the dimensions in the energy services data. To create a description of each factor, the summary table shown below was created to make the interpretation of the factors easier.

As the Factor Summary Table shows, each of the four factors have multiple variables loading highly on just that one factor, with minimal loadings on the other factors. Often, when many factors are retained, only one variable will have a reasonably high loading. For exploratory factor analysis, when one is just trying to gain an understanding of the data, one variable can adequately explain a factor. In such cases, the one variable should have an exceptionally high factor loading (± 0.7 or greater) on the factor it is explaining and have exceptionally low factor loadings (± 0.2 or less) on the other factors.

However, when factor analysis is performed as a prelude to further analyses, at least two variables, and preferably a minimum of four variables, should load moderately high (± 0.5 or greater) on the factor. Because this example output was used in further analyses, each factor retained had more than one variable that loaded highly. Also of note is that all four factors have variable loadings of at least a positive or negative point five (± 0.5).

FACTOR SUMMARY TABLE					
VARS	FACTORS				DEFINITIONS
	1	2	3	4	
X43	.9	.2	.2	0	Performance of personnel's expertise
X46	.9	.2	.2	-.1	Performance of customer caretaking
X42	.8	.2	.2	0	Performance of repairing on first visit
X41	.8	.2	.2	-.1	Perform. of customer needs assessment
X45	.8	.2	.1	0	Performance of consideration
X47	.8	.2	.2	0	Performance of company management
X44	.8	.2	.1	0	Performance of polite/courteous staff
X38	.8	.2	.2	0	Perform. of servicing at scheduled time
X40	.8	.2	.2	0	Perform. of reliable/consistent service
X37	.7	.2	.2	0	Performance of prompt responses
X94	.5	.1	.4	-.1	Satisfaction with communication
X33	.2	.8	0	0	Importance of consideration
X31	.2	.8	.1	0	Importance of personnel's expertise
X34	.2	.8	0	0	Importance of customer caretaking
X30	.2	.8	0	.1	Importance of repairing on first visit
X29	.2	.8	0	0	Import. of customer needs assessment
X32	.2	.8	0	0	Importance of polite/courteous staff
X28	.1	.8	0	.1	Import. of reliable/consistent service
X26	.1	.8	.1	.1	Import. of servicing at scheduled time
X35	.3	.8	0	0	Importance of company management
X25	.1	.7	.1	0	Importance of prompt responses
X15	.2	0	.6	-.1	Utility service vs. telephone company
X14	.1	0	.6	.1	Utility service vs. water company
X17	.2	0	.5	0	Utility service vs. previous electric co.
X12	.1	0	.5	-.1	Utility service vs. previous gas company
X97	-.1	.1	0	.6	Number of full-time employed adults
X109	-.1	.1	0	.6	Income of head-of-household
X107	.2	0	.1	.6	Age of head-of-household
X108	-.2	0	-.1	.5	Head-of-household years of education
X96	0	0	0	.5	Number of people in household

The next step is to review the meaning of each variable that loaded highly to moderately highly in the factors, as shown in the Factor Summary Table. Based on the variables' definitions, descriptions of

the four factors are surmised. These descriptions are shown in the Factor Explanation Table that follows.

FACTOR EXPLANATION TABLE	
FACTORS	EXPLANATION
1	Performance in Customer Service Areas
2	Importance of Customer Service Areas
3	Utility Company's Performance Compared to Other Service Providers
4	Household Affluence Level

The factor explanations provide a general understanding of the unique dimensions that exist in the energy services data. These explanations can be used to:

- ◆ Confirm/reject preliminary hypotheses about the data
- ◆ Provide a first assessment of the dimensions/key issues in the data to be the focus of further analyses

Since the above Factor Explanation Tables was produced after this factor analysis actually was refined and final output was generated from this energy services data, the assumption can be made that these factors are describing the majority of the variance in the data.

VALIDATION OF FINDINGS

Now that a factor analysis solution has been found, how can one be certain that it is the "right" final result? No matter what methods are used, conducting a factor analysis can be paralleled with producing a work of art. There are general guidelines to follow and statistical assumptions that must be met, but the ultimate factor analysis result will vary from artist to artist. As long as the statistical assumptions are met, *there is no right or wrong factor analysis solution*, just different viewpoints of the same set of data. Hence, there is no proven and universally accepted test that determines if the produced factoring solution is final or valid. However, in the attempt to validate the findings of a factor analysis to gain a degree of comfort, the results obtained from a factor analysis can be subjected to a series of informal tests like the ones in the following list:

- ◆ Perform another factor analysis using a different initial factor extraction method (in this case, use something other than Principal Factor Analysis) and then use the same rotation method (in this case, a VARIMAX rotation)
- ◆ Perform another factor analysis using the same initial factor extraction method (in this example, Principal Factor Analysis) and then use a different type of rotation method (in this example, use something other than the VARIMAX rotation method)
- ◆ Compare the results from the above two steps to the findings achieved through the original PROC FACTOR analysis with the original initial factoring method and the original rotation method (in this example, the initial factoring method was Principal Factor Analysis and the rotation method was VARIMAX)
- ◆ Repeat the factor analyses performed in the above three steps using a different number of final factors
- ◆ For data sets that are large enough, randomly split the data in half and perform factor analysis on both sets of data, using the same initial extraction methods and rotation methods on the two halves, and compare the two solutions

CONCLUSIONS

The example discussed in this paper provides the energy services company with initial dimensions in its data to further explore through more statistical analyses and by researching and confirming the findings. While factor analysis remains an imprecise science, the procedures discussed in this paper serve as a basic approach to discovering the unique factors that may exist in a set of data. To read more about the statistical and mathematical theories behind factor analysis and the other types of factor analysis, review the books listed in the references section of this paper.

Good Luck and Happy Factoring!

REFERENCES

Hatcher, Larry. 1994. *A Step-by-Step Approach to Using the SAS® System for Factor Analysis and Structural Equation Modeling*. Cary, NC: SAS Institute Inc.

Johnson, Richard A. and Wichern, Dean W. 1988. *Applied Multivariate Statistical Analysis, Second Edition*. Englewood Cliffs, NJ: Prentice Hall.

Kim, Jae-On and Mueller, Charles W. 1978. *Introduction to Factor Analysis: What Is It and How to Do It*. Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-013. Newbury Park, London, and New Delhi: Sage Publications.

Kim, Jae-On and Mueller, Charles W. 1978. *Factor Analysis: Statistical Methods and Practical Issues*. Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-014. Newbury Park, London, and New Delhi: Sage Publications.

SAS Institute Inc. 1990. *SAS/STAT® User's Guide, Volume 1, ACECLUS-FREQ, Version 6, Fourth Edition*. Cary, NC: SAS Institute Inc.

ACKNOWLEDGMENTS

SAS and SAS/STAT are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

AUTHOR'S ADDRESS

The author may be contacted at:

Rachel J. Goldberg
3814 Captain Dr.
Chamblee, GA 30341
Phone: (770) 457-3243
Email: TjOski@aol.com