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

Rachel J. Goldberg, Guideline Research/Atlanta, Inc., Duluth, 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 you 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

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 broadly discuss the issues that need to be considered before running PROC FACTOR. 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. Following the preliminary analysis discussion, the paper will explain an example of a PROC FACTOR program code. Finally, the paper concludes with how to interpret the results of example PROC FACTOR output.

PRELIMINARY ANALYSIS

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:

- Verification that all data values are valid,
- Removal of any data outliers,
- Reduction of data to the relevant analysis variables, and
- Decision about what type of factor analysis to perform.

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, and Principal Components Analysis. Each of these initial factoring methods generates uncorrelated factors. Arguments exist supporting all of the different initial factoring methods available. In this paper, it is assumed that the objective during the initial factor analysis is to determine the minimum number of factors that will 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 PROC FACTOR default was used, Principal Components Analysis, which is further discussed in the EXAMPLE PROC FACTOR PROGRAM section below.

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. Factor patterns can be rotated through two different ways:

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

While arguments exist supporting both types of rotation methods, factor analysis which uses an orthogonal rotation often creates a solution that is easier to grasp and interpret than a solution obtained from an oblique rotation.

The factor analysis example discussed in this paper is performed for exploratory data analysis purposes and to discover simplified factor or dimension descriptions that exist in the data. Therefore, one of the common orthogonal rotation methods, VARIMAX, is discussed in the EXAMPLE PROC FACTOR PROGRAM section below.

PROC FACTOR EXAMPLE

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 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 sub-population that is serviced by the energy services company.

EXAMPLE PROC FACTOR PROGRAM

```
PROC FACTOR DATA = SAVE.EXAMP
METHOD = PRINCIPAL SCREE MINEIGEN = 0 NFACTOR = 16
ROTATE = VARIMAX REORDER OUT = SAVE.EXAMPFAC;
VAR X2 -- GAPLL;
RUN;
```

The options invoked in the above PROC FACTOR program are described below:

DATA = data set

This names the data set that contains the observations to be factored.

METHOD = method

This specifies what type of method is to be used to extract the initial factors. As discussed earlier, the Principal Components Analysis (PCA) method (PRINCIPAL) was chosen for the initial factor extraction. The PCA method simply transforms the set of variables into another set of variables; that is, the data is summarized by means of a linear combination of the observed data. This transformation is performed because of the objective mentioned earlier: to account for as much variation as possible in the data. With PCA, the first "component" or factor is defined in such a way that the largest amount of variance in the data is explained by the first component. The second "component" or factor that is identified explains the second most about the variance in the data AND is perpendicular (thus, uncorrelated) to the first component. The remaining components, or factors, are found in a similar manner.

SCREE

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

MINEIGEN = n

This option specifies the smallest eigenvalue for which a factor is retained. In this example, "0" was selected, which retains factors with any eigenvalue.

NFACTOR = n

This option specifies the number of factors to be extracted from the data. The default value, if this option is not specified, is the number of variables in the data. In this example, 16 factors were specified to be extracted from the data. 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, 16 initial factors were kept.

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, the above 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, which are the columns of the factor pattern.

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.

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

EIGENVALUES OF THE CORRELATION MATRIX

What exactly are eigenvalues? They are values that consolidate the data variance in a matrix (the eigenvalue) 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 factor analysis, so that all of the eigenvalues can be analyzed. Based on this analysis, it is determined how many factors will be included in the final factoring program. Thus, in the example PROC FACTOR program shown earlier, NFACTOR = 16 was specified.

EIGENVALUES OF THE CORRELATION MATRIX

	1	2	3. . .	16
Eigenvalue	19.1691	7.1064	3.7441. . .	1.0139
Difference	12.0627	3.3623	0.3293. . .	0.0426
Proportion	0.2590	0.0960	0.0506. . .	0.0137
Cumulative	0.2590	0.3551	0.4057. . .	0.6959

The above eigenvalue chart is a small portion of the chart of the eigenvalues from the PROC FACTOR output. In the full eigenvalue chart in the PROC FACTOR OUTPUT, the sum of the eigenvalues is displayed, which equals the number of variables. As previously explained, for the example PROC FACTOR program in which NFACTOR = 16 was specified, 16 eigenvalues were output into the eigenvalue chart.

For the eigenvalues displayed in the chart above, the DIFFERENCE between successive values, the PROPORTION of the variation represented, and the CUMULATIVE proportion of the variation represented is shown. Generally, eigenvalues of 1 or greater are accepted as explaining an adequate amount of the data variance. Thus, when analyzing the eigenvalue chart to determine how many factors to keep in the solution, eigenvalues that were greater than 1 were kept. As the sample output above shows, the last factor that had an eigenvalue greater than 1 is Factor 16. Hence, one reason to keep the first 16 factors for the remainder of the factor analysis.

SCREE PLOT OF EIGENVALUES

Now, another way to determine how many factors should be kept in the remainder of the analysis is to analyze the SCREE PLOT. What is a SCREE PLOT? The SCREE PLOT simply displays the eigenvalues for each of the factors in a plot, from the first eigenvalue (the one that explains the most variance) to the last eigenvalue.

On a scree plot produced by PROC FACTOR, if you draw a line that runs between each of the eigenvalues, you will see the slope level off as the amount of variance that is explained by each eigenvalue is less and less. Thus, the further down the slope you go, you will see that the eigenvalues are not that much different from each other. Where the line levels off is where the "scree" really shows in your data: that is, it is where the debris, or unnecessary information, collects. Also, where the lines 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 and which ones should be eliminated. It is worth noting that it is considered "OK" to have too many as opposed to too few factors included in exploratory factor analysis.

FACTOR PATTERN

The elements of the Factor Pattern reflect the unique variance each factor contributes to the variance of an observed variable. The reason factor analysis is not stopped after this initial factoring stage, without rotating the factors, is that the factors as they currently exist are not easily interpretable. In an ideal solution, the variables should "load" highly (have a high value that approaches 1) on just one factor each.

INITIAL FACTOR PATTERN: PRINCIPAL COMPONENTS

FACTOR PATTERN

	FACTOR1	FACTOR2	FACTOR3	FACTOR4
X33	0.42684	0.73425	-0.12051	-0.00024
X27	0.31823	0.66500	-0.08290	0.03436

In the above example, a sample of the PROC FACTOR output, look at variable X33. This variable has a loading of 0.42684 (a moderate loading) in Factor 1 and a loading of 0.73425 (a moderately to high loading) in Factor 2. Ideally, this variable would have a moderately high to high loading in only one factor, which would make it easily interpretable. Because many variables loaded in a similar manner as X33, the factor pattern was rotated, using the VARIMAX rotation method.

VARIANCE EXPLAINED BY EACH FACTOR

INITIAL FACTOR METHOD: PRINCIPAL COMPONENTS

VARIANCE EXPLAINED BY EACH FACTOR

FACTOR1...FACTOR13	FACTOR14	FACTOR15	FACTOR16
19.169108... 1.114298	1.110884	1.040456	1.013916

This output simply summarizes the amount of variance in the data that is explained by each factor. As you can see by looking at the above sample PROC FACTOR output, each of FACTOR13 through FACTOR16, only explains 1.1 or less of the variance, compared to FACTOR1 which explains 19.2. If this data had been factored in another iteration, it may have been useful to eliminate the factors that are not explaining very much of the data variance.

FINAL COMMUNALITY ESTIMATES

The table of the final communality estimates simply shows the squared multiple correlations for predicting the variables from the estimated factors. It can be derived by taking the sum of squares of each row of the factor pattern, or a weighted sum of squares if variable weights have been used. This is the variance of the observed variables that is accounted for by each factor.

ROTATED FACTOR PATTERN

Because a rotated factor analysis was specified, another factor pattern was output. As stated earlier, the VARIMAX rotation method was selected since it is considered to produce the most easily interpreted results. Thus, by rotating the factor pattern, a better explanation of the data should be gained.

ROTATION METHOD: VARIMAX

ROTATED FACTOR PATTERN

	FACTOR1	FACTOR2	FACTOR3	FACTOR4
X33	-0.12981	0.84878	-0.00189	0.01275
X27	-0.06003	0.74221	-0.00771	0.05439

To determine whether or not the rotation made interpretation easier, look again at variable X33. As the sample PROC FACTOR output above shows, variable X33 now loads on FACTOR1 as -0.1 and loads on FACTOR2 as 0.8. 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...FACTOR13	FACTOR14	FACTOR15	FACTOR16
16.555348... 1.518652	1.324254	1.212009	1.188257

As the above sample PROC FACTOR output shows, while the lower factors (15 & 16) explain more of the variance than prior to rotating the factors, they still do not provide much of an explanation. However, FACTOR13 now explains 1.5 of the variance compared to just 1.1 prior to factor rotation. As previously stated, if this data had been factored in another iteration, it may have been useful to eliminate the factors that are not explaining very much of the data variance.

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 standardized scoring coefficients are automatically printed. Generally, the individual variables' factor scores are not used, the scores are used as a whole to represent the factor.

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
- Eliminating variables that do not load highly in any factor, which reduces redundancy in the data and unnecessary information

In this paper's energy services example, as is often the case in the real-world, 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, such as with the SAS System's PROC FASTCLUS and PROC DISCRIM. 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 16 factors to ensure that insight has been gained into the dimensions in the energy services data. To create a description of each factor, the highest loading variables on each of the rotated factors are summarized, as shown in the Factor Summary Table that follows in the next column.

As the Factor Summary Table shows, in some of the factors, only one variable had a reasonably high loading. Many researchers believe that at least two variables are necessary to validly explain a factor (load highly). For exploratory factor analysis, just trying to gain an understanding of the data, it has been argued that one variable can adequately explain a factor. Also, some of the factors in the above example only have a moderate to moderately high loading, such as FACTOR15's loading of 0.5. Because of some weaker factor loadings and the minimal amount of variance explained by some of the factors, as previously discussed, the current factors would need to be further analyzed and refined before taking the factor scores that were generated from this run of the PROC FACTOR program into subsequent SAS software procedures.

FACTOR SUMMARY TABLE

	FACTORS															
VARS	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
gapgg	.9															
gapkk	.9															
gapjj	.9															
gapff	.9															
gapee	.9															
gaphh	.8															
gapii	.8															
gapdd	.8															
X33		.9														
X31		.8														
X34		.8														
X30		.8														
X29		.8														
X28		.8														
X32		.8														
X35		.8														
modX57		1														
modX73		.9														
modX55		.9														
X54		-1														
modX86			.8													
modX85			.8													
Q29-8				-.8												
Q29-2					.9											
X21						.9										
X24						-.9										
X49							.6									
X811-2								.8								
X96									.7							
X811-3										.8						
X3-6-3										.8						
X3-6-6											.8					
X811-1												.8				
X3-6-2												.7				
X13													.7			
X37														.5		
X95															.8	

However, for simple descriptive purposes, the current PROC FACTOR output is adequate. Hence, the meaning of each variable that loaded highly to moderately highly in the factors, as shown in the Factor Summary Table, is reviewed. Based on the variables' definitions, descriptions of the 16 factors are surmised, as shown in the Factor Explanation Table on the next page.

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

FACTOR EXPLANATION TABLE

FACTORS	EXPLANATION
1	Expectations Gap
2	Importance of Service Issues
3	Customer Contact
4	Bill Services
5	No Modern Technology at Home
6	Presence of Modern Technology at Home
7	Customer Service
8	Community Services
9	Cost/Rates Issues
10	Number of People per Household
11	Power Quality Issues
12	Positive Service Perception
13	Negative Service Perception
14	Power Company Vs. Other Utilities
15	Prompt Service Call Response
16	Type of House

Given that a refined and final output was generated from this energy services data, the assumption can be made that these factors are describing the variance in the data. Thus, to learn more about why future changes in the data may occur, the issues that are described in the Factor Explanation Table can be scrutinized as the possible instigators of the change or variance.

VALIDATION OF FINDINGS

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 analyses, the results obtained from a factor analysis can be subjected to a series of informal tests like the ones listed below:

- Perform another factor analysis using a different initial factor extraction method, instead of PCA, and then use a VARIMAX rotation
- Perform another factor analysis using PCA for the initial factor extraction method, but then use a different type of rotation
- Compare the results from the above two steps to the findings achieved through running PROC FACTOR with an initial factoring method of PCA with a VARIMAX rotation
- Repeat the factor analyses performed in the above three steps using a different number of final factors
- For data sets that are large enough, split the data in half and perform factor analysis on both sets of data, using the same initial extraction methods and rotation methods, 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

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
Guideline Research/Atlanta, Inc.
3675 Crestwood Parkway, N.W., Suite 520
Duluth, GA 30136
Phone: (770) 717-7844
Fax: (770) 717-7876
Email: GRCRJG@aol.com