A Comparison of Distribution Free and Non-distribution Free Factor Analysis Methods

Nicola L. Ritter

Texas A&M University

Paper presented at the annual meeting of the Southwest Educational Research Association, New Orleans, February 2, 2012.

Abstract

Many researchers recognize that factor analysis can be conducted on both correlation matrices and variance-covariance matrices. Although most researchers extract factors from non-distribution free or parametric methods, researchers can also extract factors from distribution free or non-parametric methods. The nature of the data dictates the method selected. The purpose of this paper is to differentiate between the questions asked by Pearson product-moment correlations and Spearman’s rho coefficients. Additionally, the paper compares distribution free and non-distribution free methods for extracting factors using correlational and covariance matrices, and describes the advantages of each method.

A Comparison of Distribution Free and Non-distribution Free Factor Analysis Methods

Factor analytic methods are used for various purposes. Thompson (2004) mentions factor analyses can be used to: 1) evaluate score validity, 2) develop theory on the nature of constructs, and 3) summarize relationships that can be used in postliminary analyses. In factor analysis, factors are extracted from a matrix of associations rather than the raw data set. All steps in a factor analysis, with the exception of calculating factors scores, can be completed given a matrix of associations. Researchers should report matrices of associations so that researchers can replicate and evaluate a study's findings. Scores from measured or observed variables are used to compute a bivariate matrix of associations. There are two types of factor analysis: exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). EFA does not require the researcher to have a priori assumptions about the nature of the construct. Conversely, in CFA the researcher must have some presumptions about the constructs.

Either matrix of association can be used in both cases of factor analysis. Although in EFA, a correlation matrix is most commonly used, while a covariance matrix is most commonly used in CFA. Coincidentally, the default of the statistical packages

used to run an EFA and CFA are also the correlation matrix and the covariance matrix, respectively. One reason for this phenomenon is that researchers tend to use the defaults of statistical software packages, assuming the program knows something about the data the researcher is inputting. Unfortunately, the statistical software packages do not know anything about the data to make decisions within the analysis. Instead, researchers should base analysis methods on the characteristics of the data. Only the researcher will know if the data meets certain analytical assumptions. If the researcher runs an analysis on a data set that does not meet the assumptions of that analysis, then the results can be misleading or completely incorrect. Before conducting a factor analysis, researchers should have some information about the data set. The researcher should consider the scale of data and the type of questions the researcher is asking. The level of scale tells the researcher whether or not the data is distribution-free (non-parametric) or non-distribution free (parametric) in nature. In turn, the scale of the data will establish the type of questions that can be answered.

Instead of relying on imperfect statistical software packages, researchers should consider the characteristics of the data at hand and select the appropriate matrices of associations

accordingly. To select the appropriate matrices of association, researchers must understand that the factors are sensitive to the information available in a given matrix. Correlation matrices are comprised of bivariate statistics. Bivariate statistics describe the relationship between two variables. There are multiple types of bivariate statistics, but the most commonly known correlation statistic is the Pearson r correlation coefficient. However, there are other bivariate statistics that are just as useful such as Spearman's rho, phi, and point-biserial correlation. Each of these bivariate statistics corresponds to a different scale of measurement (Thompson, 2006). There are four scales of measurement: nominal(categorical), order(rank), interval(continuous), and ratio scales. Figure 1 provides a practical diagram to demonstrate the relationship between the level of scale and the bivariate statistic.

Figure 1 Relationship Between Level of Scale and the Bivarate Statistic

	nominal	ordinal	interval
continuous	r_{pb}		r
rank		ρ	
categorical	Φ		r_{pb}

Just as levels of scales dictate which descriptive statistics can be computed, scales of measurement also influence

bivariate relationships. In the case of multivariate statistics, researchers must consider the level of scale for two variables rather than only one scale when working with univariate statistics. Because the factors are extracted from a matrix of associations, the factors are sensitive to the information available in the statistic used to measure the bivariate relationship. For example, if a statistic only considers the rank of the measured variables, such as Spearman's rho, then the factors will also be based on rank. Likewise, if a statistic considers both order and distance, such as the Pearson r, then the factors will also be based on order and distance. Therefore, researchers should pay careful attention to the assumptions of bivariate correlation coefficients. The Pearson r coefficient requires that both variables be at least intervally scaled, while the Spearman's rho coefficient assumes that both variables are at least ordinally scaled. The "at least" aspect of these assumptions is an important point to emphasize. If the data are intervally scaled, then either the Pearson r coefficient or the Spearman's rho can be used to build a correlation matrix.

Although either coefficient can be used, researchers should not assume that the correlation coefficients are equal. Consider the following intervally scaled data set for heuristic purposes.

Small numbers are used for easy computation. Suppose we have the following data set.

Table 1
Intervally Scaled Heuristic Data

Participant	X	Y
1	3	3
2	4	4
3	5	92

Table 1 presents three individuals scoring three, four, and five respectively on the independent variable (X), and scoring three, four, and 92, respectively, on the dependent variable (Y). Table 2 presents the descriptive statistics of Table 1. Variable X has a mean of 4.0 with a standard deviation of 1.0, while variable Y has a mean of 33.0 with a standard deviation of 1.0. Variable Y has a much larger mean than variable X due to participant 3’s score on the Y variable relative to the other two participant’s scores. The mean here reflects the distance between scores. The covariance of variable X and Y is 44.5 and the Pearson r is 0.87. The Pearson r coefficient indicates a positive correlation between variable x and variable y.

Table 2
Calculating the Pearson r Correlation Coefficient

Participant	X	$\bar{X}$	x	Y	$\bar{Y}$	y	xy
1	3	4.0	-1.0	3	33.0	-30.0	30.0
2	4	4.0	0.0	4	33.0	-29.0	0.0
3	5	4.0	1.0	92	33.0	59.0	59.0
Sum	12.00			99.00			89.0
Mean	4.00			33.00			
SD	1.00			51.10			
COV	44.50						
r	0.87						

Now consider the same data set that only considers the rank of the participant's scores.

Table 3
Ordinally Scaled Data Heuristic Data

Participant	X	Y
1	1	1
2	2	2
3	3	3

Table 3 presents all three individuals ordered based on their score on both variables. Table 4 presents the descriptive statistics from Table 3. Now both variable X and Y have a mean of 2.0 with a standard deviation of 1.0. Variable Y no longer has a larger mean than variable X because the data only considers the individuals' relative standing to one another. In

addition, the covariance of variable X and Y is 1.0 and the Pearson r has perfect positive correlation.

Table 4
Spearman's rho Correlation Coefficient

Participant	X	$\bar{X}$	x	Y	$\bar{Y}$	y	xy
1	1	2.0	-1.0	1	2.0	-1.0	1.0
2	2	2.0	0.0	2	2.0	0.0	0.0
3	3	2.0	1.0	3	2.0	1.0	1.0
Sum	6.00			6.00			2.0
Mean	2.00			2.00			
SD	1.00			1.00			
COV	1.00						
r	1.00						

As previously demonstrated, although either correlation coefficient can be used, the values of these coefficients are not equal. Thus, selecting the appropriate bivariate statistic for a correlation matrix is an important decision when extracting factors.

Researchers may be unsure of which coefficient to select because either correlation coefficient can be used. Correlational statistics address different research questions. Researchers should select the bivariate statistic that answers the researcher's questions. Spearman's rho addresses one question: "How well do the two variables order the cases in exactly the same (or the opposite) order?" (Thompson, 2004, p.

130). Conversely, Pearson r addresses two questions: 1) “How well do the two variables order the cases in exactly the same (or the opposite) order?” and 2) “To what extent do the two variables have the same shape?” (Thompson, 2004, p. 130). Pearson r extends beyond Spearman’s rho to answer questions about the relationship between each variables’ distribution. Because Pearson r considers assumptions about the variables’ distributions, Pearson r is a parametric statistic.

Because the level of scale directs each bivariate statistic, each bivariate statistic answers different research questions. In factor analysis, researchers use a matrix of associations to extract factors. Different factors may be extracted based on the matrix of associations selected. This section demonstrates how different factors are extracted using two types of correlation matrices, Pearson r and Spearman’s rho, and a variance-covariance matrix. The first six cases and first five independent variables from Thompson (2004) Appendix A are used to create the three matrices under discussion. Principle components extraction method and varimax rotations are used to create each of the matrices. The principle component method is used here because the method “extracts the maximum amount of variance that can be possibly extracted by a given number of factors” (Gorsuch, 1983, p. 95). Principle components extraction

and varimax rotation methods can be used when factors are orthogonal. Researchers desiring to use extraction and rotation methods that do not allow factors to be correlated will see different factors extracted similarly to the heuristic examples that follow. Additionally, the syntax to produce the matrix of associations is shown here because the Spearman's rho matrix cannot be created in statistical packages such as SPSS by pointing and clicking through the menu options.

Table 5
Thompson (2004) Appendix A Data Subset

ID	ROLETYPE	PER1	PER2	PER3	PER4	PER5
1	2	8	7	5	5	3
2	2	5	7	5	5	4
3	2	6	5	5	6	5
4	2	5	5	4	6	4
5	2	5	5	5	5	5
6	2	7	7	7	8	7
7	2	8	8	7	7	6

Pearson product-moment correlation matrix

As mentioned earlier, the Pearson r correlation matrix is the default correlation matrix in many statistical software packages. To run a factor analysis using a Pearson r correlation matrix, researchers can either input the syntax in Figure 1 or can point and click through the available software menus.

Figure 1 SPSS Syntax for Factor Analysis with Pearson r Correlation Matrix

```
FACTOR
  /VARIABLES PER1 PER2 PER3 PER4 PER5
  /MISSING LISTWISE
  /ANALYSIS PER1 PER2 PER3 PER4 PER5
  /PRINT INITIAL EXTRACTION ROTATION
  /CRITERIA MINEIGEN(1) ITERATE(25)
  /EXTRACTION pc
  /CRITERIA ITERATE(25)
  /ROTATION varimax
  /METHOD=CORRELATION.
```

Once the syntax is run, the output will report a large amount of information. For the purpose of this paper, researchers should focus on the number of factors extracted and the communality coefficient, h^2. The communality coefficient provides information about the reliability of each variable loading on a given factor. Table 6 presents a consolidated output for the information that is of interest here.

Table 6
Varimax-Rotated Factor Coefficients From Principal Components Analysis Using Pearson r Correlation Matrix

Variable	Factor 1	Factor 2	h^2
PER1	0.169	0.902	0.841
PER2	0.161	0.92	0.872
PER3	0.749	0.627	0.955
PER4	0.911	0.242	0.888
PER5	0.985	0.064	0.974

There are two factors extracted from the Pearson r correlation matrix. Variables PER3, PER4, and PER5 load on the first factor; while variables PER1 and PER2 load on the second factor. In addition, the communality coefficient indicates a high reliability. The same process must be completed using the Spearman's rho correlation matrix to compare the factors produced by each type of correlation matrix.

Spearman's rho correlation matrix

As mentioned earlier, the Spearman's rho cannot be created in SPSS by pointing and clicking through the software menus. A solution to running a factor analysis using a Spearman's rho correlation matrix is to run the syntax from Figure 2. When the syntax is run, SPSS will create a Spearman's rho matrix in a new window and then run the factor analysis. The NONPAR CORR command instructs SPSS to create the correlation matrix with Spearman's rho, presented in Table 7. The RECODE command instructs SPSS to use the Spearman's rho matrix in place of CORR command default, Pearson r. The IN(cor=*) command instructs SPSS to use as input the Spearman's rho correlation matrix. The remaining syntax is the same as the syntax used to analyze the Pearson product-moment correlation matrix. Once the matrix is analyzed, the output will be organized similarly to the previous output. Table 8 presents the information about factor extraction in discussion

here. There are two factors extracted from the Spearman's rho correlation matrix. Variables PER1, PER2, and PER3 load on the first factor, while variables PER4 and PER5 load on the second factor. In addition, the communality coefficient indicates a high reliability.

Figure 2 SPSS Syntax for Factor Analysis with Spearman's rho Correlation Matrix

```
NONPAR CORR
/VARIABLES=PER1 PER2 PER3 PER4 PER5
/PRINT=SPEARMAN
 /MATRIX=OUT(*)
/MISSING=LISTWISE .
RECODE  rowtype_  ('RHO'='CORR')  .
EXECUTE .
FACTOR
/MATRIX=IN(cor=*)
/ANALYSIS PER1 PER2 PER3 PER4 PER5
/PRINT INITAL EXTRACTION ROTATION
/CRITERIA MINEIGEN(1) ITERATE(25)
/EXTRACTION pc
/CRITERIA ITERATE(25)
/ROTATION varimax
/METHOD=CORRELATION .
```

Table 7
Spearman's rho Correlation Matrix

Variable	PER1	PER2	PER3	PER4	PER5
PER1	1				
PER2	0.687	1			
PER3	0.647	0.732	1		
PER4	0.392	0.283	0.574	1	
PER5	0.21	0.216	0.761	0.781	1

Table 8
Varimax-Rotated Factor Coefficients From Principal Components Analysis Using Spearman's rho Correlation Matrix

Variable	Factor 1	Factor 2	h^2
PER1	0.882	0.159	0.804
PER2	0.924	0.117	0.868
PER3	0.697	0.644	0.9
PER4	0.199	0.881	0.815
PER5	0.107	0.969	0.951

Now that factors have been extracted from the Pearson r and Spearman’s rho correlation matrix, the extracted factors may be compared. Both matrices extracted two factors. However the five variables loaded differently on these two factors. The Pearson r correlation matrix loaded variables PER1 and PER2 on one factor and variables PER3, through PER5 on another factor. Conversely, the Spearman’s rho correlation matrix loaded variables PER1 through PER3 on one factor and variables PER4 and PER5 on another factor. When the Spearman’s rho matrix is used, in general, the values of the variables that do not contribute to a factor tend to attenuate. The difference between the factors extracted is due to the Pearson r matrix accounting for order and distance, and the Spearman’s rho matrix only accounting for order. Most of the discussion to this point emphasizes extracting factors from different types of correlation matrices. In addition, variance-covariance matrices can also be used to extract factors.

Variance-covariance matrix

A correlation matrix provides different information than a covariance matrix. A variance-covariance matrix is most commonly used in confirmatory factor analysis. A covariance statistic is computed using the Pearson r coefficient, but removes the standard deviations of the variables. Equation 1 and Equation 2 shows the Pearson r and covariance statistic arithmetically.

$$r_{XY} = COV_{XY} / (SD_X * SD_Y) \quad (1)$$

$$COV_{XY} = r_{XY} * SD_X * SD_Y \quad (2)$$

Here we see the Pearson r is a function of the covariance divided by the standard deviation of both variables, while the covariance coefficient describes the bivariate relationship as a function of the Pearson r and the standard deviation of each variable. The covariance statistic is jointly influenced by three aspects of the variables: 1) correlation between the two variables, 2) variability of the first variable, and 3) variability of the second variable (Thompson, 2004). Extracted factors from a covariance matrix can be problematic because factors are a function of correlations and standard deviations. The Thompson (2004) data set can be used to create a covariance matrix. Then the factors extracted using the covariance matrix

can be compared with the factors produced from the correlation matrices.

To run a factor analysis using a covariance matrix, researchers can input the syntax in Diagram 4 or point and click through the available software menus.

Figure 3 SPSS Syntax for Factor Analysis with Covariance Matrix

```
FACTOR
/VARIABLES PER1 PER2 PER3 PER4 PER5
/MISSING LISTWISE
/ANALYSIS PER1 PER2 PER3 PER4 PER5
/PRINT INITIAL EXTRACTION ROTATION
/CRITERIA MINEIGEN(1) ITERATE(25)
/EXTRACTION pc
/CRITERIA ITERATE(25)
/ROTATION varimax
/METHOD=cov.
```

Table 9 presents a consolidated output for the information that is of interest here. There are two factors extracted from the variance-covariance matrix. Variables PER1 and PER2 load on the second factor, while variables PER3 through PER4 load on the first factor. Additionally, the communality coefficient indicates a high reliability. As previous demonstrated with the Pearson r and Spearman’s rho correlation matrices, different factors may be extracted based on the matrix of associations selected. The extracted factors from the covariance matrix

differ because the factors are a function of correlations and standard deviations. Just as seen with the correlation matrices, factors are sensitive to the information available in the bivarate statistic.

Table 9
Varimax-Rotated Factor Coefficients From Principal Components Analysis Using Spearman's rho Correlation Matrix

Variable	Factor 1	Factor 2	h^2
PER1	0.163	0.923	0.878
PER2	0.174	0.898	0.836
PER3	0.760	0.615	0.955
PER4	0.900	0.250	0.873
PER5	0.990	0.055	0.982

Researchers should keep in mind that factor analysis uses a matrix of associations to extract factors instead of raw data. Because factor analysis uses a matrix of associations, the factors extracted are sensitive to the information available in the bivariate statistic used in the matrix. Likewise, bivariate statistics represented in matrices address different questions. Matrices should be selected according to the questions asked, not the default settings of statistical software packages. The researcher can select the appropriate matrix based on the questions the researcher is asking. Researchers should be conscientious about which matrix of associations to use because

different factors may be extracted based on the matrix of associations selected.

References

Gorsuch, R.L. (1983). *Factor Analysis* (2nd ed.). Hillsdale, New Jersey: Erlbaum.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, DC: American Psychological Association.

Thompson, B. (2006). *Foundations of behavioral statistics: An insight-based approach*. New York: Guilford.