DOCUMENT RESUME

ED 355 281 TM 019 637

AUTHOR	Mittag, Kathleen Cage
TITLE	Scale-Free Nonparametric Factor Analysis: A User-Friendly Introduction with Concrete Heuristic Examples.
PUB DATE	Jan 93
NOTE	21p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (Austin, TX, January 28-30, 1993).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Correlation; *Factor Analysis; *Heuristics; Mathematical Models; *Matrices; *Nonparametric Statistics; *Scores
IDENTIFIERS	*Pearson Product Moment Correlation; *Spearman Rank Correlation Coefficient

ABSTRACT

Most researchers using factor analysis extract factors from a matrix of Pearson product-moment correlation coefficients. A method is presented for extracting factors in a non-parametric way, by extracting factors from a matrix of Spearman rho (rank correlation) coefficients. It is possible to factor analyze a matrix of association such that neither means, variances, nor monotonicity of distances between ordered pairs of scores will affect the results, by factor analyzing a matrix of Spearman's rho coefficients. Since Spearman's rho is a scale-free non-parametric procedure, the benefits of non-parametric statistics can be applied to this application. An example of six variables and seven cases demonstrates regular factor analysis using Pearson's "r" for the correlation matrix and non-parametric factor analysis using Spearman's rho for the correlation matrix. Advantages of the non-parametric strategy are discussed. Three tables contain analysis data, the correlation matrix, and the varimax rotated factor matrix. (SLD)

Scale-Free Nonparametric Factor Analysis: A User-Friendly Introduction with Concrete Heuristic Examples

Kathleen Cage Mittag
Texas A&M University 77843-4232

Paper presented at the annual meeting of the Southwest Educational Research Association, Austin, TX, January 29, 1993.

ABSTRACT

Most researchers recognize that many matrices of association may be subjected to factor analysis, including the variance-covariance matrix and the correlation matrix. Which data characteristics influence the results depends on what matrix is subjected to analysis. For example, when the variance-covariance matrix is analyzed, both the variables' variances and their relationships influence the factor structure, but the variables' means do not. Most researchers using factor analysis extract factors from a matrix of Pearson product-moment correlation coefficients. This paper presents a method for extracting factors in a nonparametric method, by extracting factors from a matrix of Spearman rho coefficients, and describes the advantages of this strategy.

Both nonparametric statistics and factor analysis are extremely useful statistical methods. This paper will first present a general discussion of the correlation coefficients Pearson's r (Pearson's product-moment correlation coefficient) and Spearman's rho, and then go to a discussion of factor analysis and nonparametric methods. It will end with a discussion concerning the factor analysis of a matrix of Spearman's rho coefficients, as against the product-moment correlations traditionally analyzed by commonly used statistical packages. The benefits of nonparametric factor analysis as against parametric factor analysis will be discussed and illustrated with some actual data.

Correlation Coefficients

The two correlation coefficients to be discussed in this paper are Pearson's r and Spearman's rho. The purposes for calculating correlation coefficients are (Daniel, 1990, p. 357):

1. They measure the strength of the relationship among the sample observations.
2. They provide a point estimate of the measure of the strength of the relationship between the variables in the population.
3. They provide the basis for construction a confidence interval for the measure of the strength of the relationship between the variables in the population.
4. They allow the investigator to reach a conclusion about the presence of a relationship in the population from which the sample was drawn.

Pearson's Product Moment Correlation Coefficient (r)

The usual correlation coefficient used for factor analysis is Pearson's r.

Studies using this statistic need to be designed so that the variables have continuous, approximately normal distributions and the regression for every pair of variables should be rectilinear, so that Pearson's product-moment correlation coefficient is applicable (Comrey, 1973). Nunnally (1967) also discussed assumptions necessary for *r*. These assumptions are:

1. There is a linear relationship between the two variables.
2. The two variables are distributed similarly.
3. The relationship between the two variables must be homoscedastic, which means that the spread about the regression line in a scattergram of the two variables must be about the same at all levels of the variables.

Violations of the assumptions can result in a reduction of the maximum value of the calculated *r* (McCallister, 1991). The *r* coefficient is also attenuated by less than perfect reliability of the data, and cannot exceed the square root of the product of the reliability coefficients of the two sets of scores. Dolenz (1992) discussed attenuation influences on *r*, including: departures from linearity, departures from both variables being similarly distributed, using instruments with lower reliability, and using data in which either variable has a restricted range.

The standard formula for Pearson's *r* is (Comrey, 1973, p. 204):

$$r = [n\Sigma XY - \Sigma X \Sigma Y] \div \{[\sqrt{(n\Sigma X^2 - (\Sigma X^2))}]\,[\sqrt{(n\Sigma Y^2 - (\Sigma Y^2))}]\} \qquad \ldots .1$$

The summations are from 1 to *n*, the number of cases.

Pearson's *r* is the most stable of the correlation coefficients. It is a descriptive measure of strength of linear association between two variables (Busby & Thompson, 1990). Values of *r* always are $-1 \leq r \leq 1$ and if the values are close to 1 or -1, there will be a strong positive or negative linear association respectively. If *r* is near 0, then the points are probably widely scattered about the regression line (Weiss & Hassett, 1991).

Spearman's rho (ρ)

Spearman's rho (Rank Correlation Coefficient) was introduced by Spearman in 1904. The assumptions needed for rho are:

1. The data consist of a random sample of n pairs of numeric or nonnumeric observations.

2. Each pair of observations represents two measurements taken on the same object or individual, called the *unit of association.*

3. There is at least an ordinal scale for the two variables.

4. There is independence between pairs but not within a pair.

The computational formula for rho is:

$$\rho = 1 - \{[6\Sigma(R(X_i) - R(Y_i))^2] \div [n(n^2 - 1)]\} \qquad \ldots.\ 2$$

The summation is from 1 to n.

If there are no ties, Spearman's rho is equivalent to Pearson's r if Pearson's r is computed by replacing the data by their ranks (Conover, 1980). When teaching Spearman's rho, it is easier to rank the X's and rank the Y's so you get the ordered pairs in the form $[R(X_i), R(Y_i)]$. Next, simply use a calculator or computer to calculate r on the ranked scores. The asymptotic relative efficiency (ARE) of Spearman's rho to Pearson's r is $9/\pi^2 \cong 0.921$, when both tests are applied under the same conditions in which the assumptions for the parametric test are met (Daniel,1990). Practically, this ARE would mean that for the tests to have the same power, the sample size for the Spearman's rho test must be about 100 if the sample size for the Pearson's r is about 92. The definition of the ARE is:

$$ARE = \lim_{\underline{n}_1 \to \infty} \underline{n}_2 / \underline{n}_1 \quad \ldots 3$$

where $\underline{n}_1$ = the sample size for test 1 and $\underline{n}_2$ = the sample size for test 2.

Various Bivariate Correlation Coefficients

McCallister (1991) considered five bivariate correlation coefficients which were: phi, rho, product moment, biserial and point-biserial. The coefficients depend on different scales of measurement. These coefficients and scales are:

1. Pearson's $\underline{r}$ -- both variables are either interval or ratio scales.
2. Rho -- both variables are ranked.
3. Phi -- both variables are nominal dichotomies.
4. Point-biserial -- one variable is on an interval or ratio scale and the other is a nominal (dichotomous) variable.
5. Biserial -- one variable is on an interval or ratio scale and the other variable is dichotomous but with underlying continuity.

The various conditions created for use with each correlation coefficient were: perfect correlation, restriction of range, measurement error (one variable), measurement error (both variables), extreme scores (one outlier), extreme scores (two outliers), and heterogeneity of sample distribution (one variable). It was shown that (McCallister, 1991) "the correlation coefficients were reduced in each of these six conditions and that the reductions differed by both condition and coefficient."

Correlation Matrix

The correlation matrix is a square matrix whose elements are correlation coefficients between all possible pairings of variables in a data set. The size of the matrix is $\underline{n}$ by $\underline{n}$, where $\underline{n}$ is the number of variables. The values for the main diagonal elements are always equal to 1, since these elements are correlations of a variable to itself. The matrix is symmetric with respect to the main diagonal, since the correlation coefficient of x_1 to x_2 is the same as the correlation coefficient of x_2 to x_1. There are $[n\ (n - 1)] \div 2$ different possible correlation coefficient values above the main diagonal. For example, if there are 10 variables, then there are 45 (10(10-1)/2) possible correlation coefficient to consider and if there are 20 variables, then there are 190 possible correlation coefficients.

Visual inspection of the correlation matrix can lead to several conclusions. Two variables which are highly correlated to each other can be identified, which variables correlate most highly with each of the individual variables can be identified, and cluster of variables which are highly correlated to each other or those which are relatively independent of one another can also possibly be identified. Visual inspection cannot assess the joint effects of two or more variables on another variable nor can it determine to what extent the correlation between two variables is due to the effects of a third, fourth, etc., confounding variable. We need a method to identify and summarize the many inter-relationships that exist between the individual variables, which is factor analysis (Kachigan, 1986).

Factor Analysis

History and Definition

Factor analysis is a branch of statistical science even though many people consider it a part of psychological theory. The reason for this discrepancy is because it was largely developed in mathematical psychology and is used extensively in psychology (Harman, 1976). It was originally developed as the fundamental means of devising multidimensional scales. Factor analysis can be defined as:

> A method for reformulating a set of natural or observed independent variables into a new set (usually fewer in number, but necessarily not more in number) of independent variables,such that the later set has certain desired properties specified by the analyst. (Stopher & Meyburg, 1979, p. 237)

Factor analysis can identify and summarize inter-relationships that exist between many individual variables. Kachigan (1986, p. 378) wrote that the factor analysis procedure can be thought of "as removing the duplicated information from among a set of variables, or as the grouping of similar variables."

Charles Spearman is considered the "father of factor analysis." He began developing his Two-Factor Theory in 1904 then devoted the remaining 40 years of his life to the development of factor analysis. In the 1930's, it was decided that the Two-Factor Theory was not always adequate, so multiple factor analysis was begun. L. L. Thurstone did much work in multiple factor analysis. Factor analysis' main concern is "the resolution of a set of variables linearly in terms of (usually) a small number of categories or factors" which is accomplished by "the analysis of the correlations among the variables" (Harman, 1976, p. 4). Harman (1976, p. 4) wrote the "chief aim is to attain scientific parsimony or economy of description." Factor analysis is not be be thought of as a modeling procedure but it is used to manipulate data before models are developed (Stopher & Meyburg, 1979).

Applications

The study of factor analysis was not limited to psychology because many disciplines could make use of this technique. Some of the uses are (Kachigan, 1986):

1. Identification of underlying factors. Factor analysis clusters a large number of variables into a smaller number of new variables called "factors." This simplifying data: makes it easy to see insights into the subject matter, makes the problem more manageable, and makes interpretation easier.

2. Screening of variables. Factor analysis can be used the screen variables to be included in further statistical studies such as regression analysis or discriminate analysis. Factor analysis groups several variables under one factor so one of these variables can be used later and avoid collinearity.

3. Summary of data. Using factor analysis, as few or as many factors as warranted can be extracted from a set of variables. In one type of factor analysis, consecutive extracted factors each account for decreasing amounts of variance; therefore, only the first few factors can be used to explain the variance.

4. Sampling of variables. A small group of representative variables, though uncorrelated, can be selected from among a larger set to solve a variety of practical problems, such as advertising and stocks.

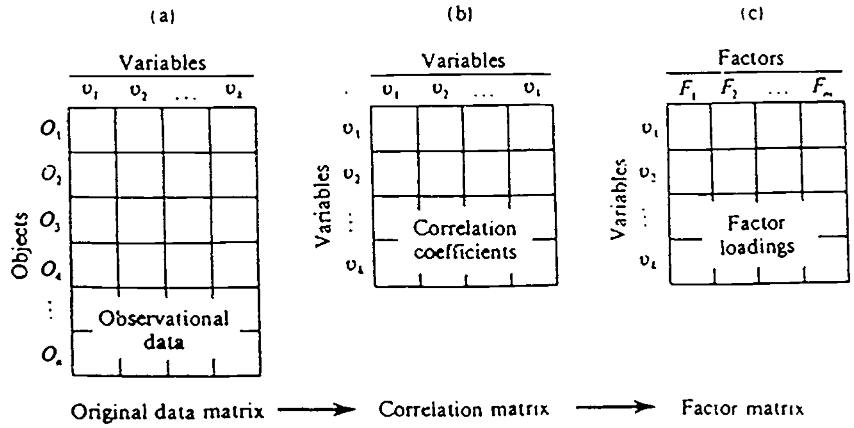
5. Clustering of objects. A cluster of people (or other objects) can be measured on a number of random variables and then clustered into homogeneous groups based on inter-correlations.

Harman (1976) illustrates uses for factor analysis in economics, medicine, physical sciences, political science, sociology, regional science, taxonomic applications, human engineering, and archaeology.

Procedure

Figure 1 shows the key stages in the factor analysis procedure (Kachigan, 1986, p. 383).

Figure 1



The first stage (a) is the original data matrix. There are k different variables and n different objects. Typically, it is better if n> k. Some experts advise that there be ten times as many objects as variables with a minimum of 100 objects. The second stage (b) is the correlation matrix with size k by k which was discussed in the previous section. After the correlation matrix if found, matrix algebra is used to derive the factor matrix (c). In the factor matrix, the columns represent the derived factors and the rows represent the original variables. The factor loadings are the elements of the factor matrix and represent the correlation coefficients between the original variables and the newly derived factors. Values for the factor loadings range from -1 to 1, inclusive. There will be some high loadings and some low loadings in each column. Variables with high loadings on a factor provide meaning and interpretation of the factor and variables with low loadings will not contribute to the meaning that that factor. According to Kachigan (1986, p. 385), "An objects score on a factor represents a weighted combination of its scores on each of the input variables." Highly correlated variables will form a factor and uncorrelated variables will form separate factors.

Principal components analysis is a variation of factor analysis in which the number of possible factors is exactly equal to the number of variables. The first extracted factor accounts for the largest part of the total variance in the data and succeeding factors account for less and less of the total variance. The final factors account for less variance than an individual variable. Principal components analysis is often used as a primary step to determine how many factors to use in factor analysis. According to Kachigan (1086, p. 386), the basic difference between principal components analysis and factor analysis is:

> In principal components analysis, each factor or "component" is viewed as *a weighted combination of the input variables*, with as many components derived as there are variables. In the mainstream factor analysis models, on the other hand, each input variable is viewed as *a weighted combination of factors*, with the number of factors being fewer in number than the original set of input variables.

In principal components analysis, an eigenvalue is calculated for each extracted factor. This eigenvalue can be used to decide how many factors to retain in the analysis. A rule of thumb often used is to retain factors whose eigenvalues are greater than or equal to one. Other methods used to decide which factors are retained are scree curves and variance explained. After deciding which factors to extract, then factor rotation can be used to redefine the factors to make sharper distinctions in the meanings of the factors. Often the high-loading variables of each factor are studied in order to name the factors descriptively. This naming of factors can be very difficult but also very helpful in analysis.

Criticisms of Factor Analysis

There are several criticisms that have been made against factor analysis. One criticism is the danger of labeling the factors wrongly. This can lead to misinterpretation. Another criticism is that sometimes the derived factors are very obvious, thus complicated computer analyses are not always needed. A third criticism is "Garbage In -- Garbage Out." This is not really a defect in the analysis itself, since factor analysis does not create any new variables, it only organizes, summarizes, and quantifies existing information, but the quality of results depend on the quality of the input.

Nonparametric Statistics

Definition

In the 1930's, a different approach to statistics gained momentum and it was called "nonparametric statistics." According to Conover (1980, p. 2), this approach involved "making few, if any, changes in the model, and using simple and unsophisticated methods to find the desired probabilities, or at least a good approximation to those probabilities." The formal definition of nonparametric statistics is:

A statistical method is nonparametric if it satisfies at least one of the following criteria.

1. The method may be used on data with nominal scale of measurement.
2. The method may be used on data with an ordinal scale of measurement.
3. The method may be used on data with an interval or ratio scale of measure, where the distribution function of the random variable producing the data is either unspecified or specified except for an infinite number of unknown parameters (Conover, 1980, p. 92).

Nonparametric methods make no hypothesis about the value of the parameters and are distribution-free, which means there are no assumptions of the precise form of the sample population.

Advantages and Disadvantages

The advantages of nonparametric methods are:

1. Since there are very few assumptions, it is generally not as often misapplied.
2. The calculations are usually quick and easy to do.
3. The basis of the procedure is usually fairly easy to understand.
4. These procedures can be applied when you have weaker measurement scales.
5. The theory behind nonparametric statistics is much easier to understand since it is basically algebra and counting.

The disadvantages of nonparametric methods are:

1. Information may be wasted.
2. If the parametric assumptions are valid, the nonparametric method may not be as powerful as the parametric method used in that situation.

When to Use Nonparametric Procedures

The following are situations where it is appropriate to use nonparametric procedures (Daniel, 1990):

1. There is no population parameter involved in the hypothesis to be tested.
2. The data is measured on a scaler weaker than that required for the associated parametric procedure.
3. The assumptions needed for the parametric procedure are not valid.
4. Results are needed in a hurry and the calculations must be done by hand since a computer is not available.

Procedures

There are many different parametric procedures. Type of data, type of problem, and type of measurement scale all determine which procedure to use.

Some types of data are: one sample, two independent samples, two related samples, three or more independent samples, and three or more related samples. Some types of problems are: location, dispersion, goodness of fit, association, regression, binomial, trends and confidence intervals. Some types of measurement scales are: nominal, ordinal, and interval. The procedures have special names such as: binomial test, chi-square test, McNemar test, Cochran test, quantile test, Cox and Stuart test, Spearman's rho, Kolmogorov test, sign test, Kendall's tau, Friedman test, Mann-Whitney test, Smirnov test, squared-ranks test, median test, Kruskal-Wallis test, Wilcoxon test, and Quade test to name a few.

Nonparametric Factor Analysis

Many matrices of association may be subjected to factor analysis, including the variance-covariance matrix and the correlation matrix. Which data characteristics influence the results largely depend on what matrix is subjected to analysis. For example, when the variance-covariance matrix is analyzed, both the variables' variances and their relationships influence the factor structure, but the variables' means do not.

When the product-moment correlation matrix (Pearson's $\underline{r}$) is analyzed, neither variances nor means affect the identification of factor structure, but both the covariances among the variables and the monotonicity of distances between ordered pairs of scores will affect the results (Dolenz, 1992). For example, the product-moment $\underline{r}$ between 1, 2, and 3, and 1, 3, and 5 is +1.0, but the product-moment $\underline{r}$ between 1, 2, and 3 and 1, 3, and 4 is less than +1.0. However, both sets of data yield Spearman's rho of +1.0.

However, it is possible to factor analyze a different matrix of association, such that neither means, variances, nor monotonicity of distances between ordered pairs of scores will affect the results. This can be accomplished by factor analyzing a matrix of Spearman's rho coefficients, as against product-moment correlations. Since Spearman's rho is a scale-free nonparametric procedure, the benefits of nonparametric statistics can be applied to this application.

An example was used to demonstrate regular factor analysis using Pearson's $\underline{r}$ for the correlation matrix (procedure 1) and nonparametric factor analysis using Spearman's rho for the correlation matrix (procedure 2). The principal components were compared in each of the two procedures. Table 1 contains the data and a correlation matrix which includes the Pearson's $\underline{r}$ correlation coefficients above the main diagonal and the Spearman's rho correlation coefficients below the main diagonal. Table 2 is the varimax-rotated factor matrix for $\underline{r}$ (procedure 1) and Table 3 is the varimax-rotated factor matrix for rho (procedure 2). In both procedures, the eigenvalues indicate that three factors should be extracted, since three of the eigenvalues were greater than or equal to 1. Referring to the varimax-rotated factor matrix in procedure 1, variables 5 and 6 contribute to Factor I, variables 1 and 2 contribute to Factor II, and variables 3 and 4 contribute to Factor III. Referring to the varimax-rotated factor matrix in procedure 2, variables 5 and 6 contribute to Factor I, variables 3 and 4 contribute to Factor II, and variables 1 and 2 contribute to Factor III. The values of the variables which do not contribute to the factors in each column tend to zero-out in procedure 2. This indicates that these differences in the loadings are due to taking the distances of scores from each other into account (procedure 1), or not (procedure 2). Distances are involved in the parametric procedure and not involved in the nonparametric procedure. Researchers need to consider this fact and the fact that parametric assumptions could be violated when performing parametric analyses of their data.

References

Busby, D., & Thompson, B. (1990, January). Factors attenuating Pearson's r: A review of basics and some corrections. Paper presented at the annual meeting of the Southwest Education Research Association, Austin, TX.

Comrey, A. L. (1973). A first course in factor analysis. New York: Academic Press, Inc.

Conover, W. J. (1980). Practical nonparametric statistics (2nd ed.). New York: John Wiley & Sons.

Daniel, W. W. (1990). Applied nonparametric statistics (2nd ed.). Boston: PWS-KENT Publishing Company.

Dolenz, B. (1992, January). Factors that attenuate the correlation coefficient and its analogs. Paper presented at the annual meeting of the Southwest Educational Research Association, Houston, TX. (ERIC Document Reproduction Service No ED 347 173)

Harman, H. H. (1976). Modern Factor Analysis (3rd ed.). Chicago: The University of Chicago Press.

Kachigan, S. K. (1986). Statistical analysis. New York: Radius Press.

McCallister, C. (1991, January). Phi, rho, p.m., biserial, and point-biserial "r": A review of linkages. Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, TX. (ERIC Document Reproduction Service No. ED 336 394)

Nunnally, J. C. (1967). Psychometric theory. New York: McGraw-Hill Book Company.

Stopher, P. R., & Meyburg, A. H. (1979). Survey sampling and multivariate analysis for social scientists and engineers. Lexington, MA: D. C. Heath and Company.

Thorndike, R. M. (1978). Correlational procedures for research. New York: Gardner Press, Inc.

Weiss, N. A., & Hassett, M. J. (1991). Introductory statistics (3rd ed.). Reading, MA: Addison-Wesley Publishing Company.

TABLE 1

DATA: Number of variables = 6, Number of cases = 7

1.	4	5	9	6	-2	-8
2.	-7	-7	1	1	-9	-7
3.	1	1	-2	-3	1	4
4.	10	6	-4	-9	0	1
5.	0	2	-3	-2	1	6
6.	-8	-9	0	0	2	8
7.	3	4	2	9	4	9

CORRELATION MATRIX: Pearson's r above main diagonal (procedure 1) and Spearman's rho below main diagonal (procedure 2).

1.00	0.95	-0.05	-0.22	0.34	-0.04
0.96	1.00	0.08	-0.01	0.37	-0.04
-0.04	-0.07	1.00	0.77	-0.21	-0.53
-0.11	-0.07	0.93	1.00	0.07	-0.04
-0.07	0.00	-0.14	0.11	1.00	0.84
-0.32	-0.29	-0.14	0.11	0.93	1.00

TABLE 2

PROCEDURE 1

Varimax-rotated factor matrix with h2 as last column:

0.05	0.98	-0.12	0.99
0.09	0.98	0.07	0.98
-0.34	0.08	0.91	0.95
0.12	-0.12	0.96	0.95
0.93	0.31	0.05	0.96
0.97	-0.12	-0.2	0.99

TABLE 3

PROCEDURE 2

Varimax-rotated factor matrix with h2 as last column:

0.11	-0.04	0.98	0.98
0.05	-0.04	0.99	0.99
0.14	0.98	-0.03	0.98
-0.12	0.98	-0.04	0.98
-0.99	-0.01	0.05	0.98
-0.96	-0.12	-0.23	0.98