ISSN 2320-6802

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS.....

Multivariate Factor Analysis A Method for Psychological Analysis

Vandana Kaushik[1], Dr.Deepti Gaur[2]

ITM University Sec 23 A,Gurgaon,122001
Haryana ,India
[1]vandana12sep020@itmindia.edu

Associate Professor
ITM University Sec 23 A,Gurgaon,122001
Haryana, India
[2]deeptigaur@itmindia.edu

Abstract: *This review paper discusses about part of multivariate data clustering techniques and their subparts. Factor analysis is a type of multivariate statistical approach commonly used in psychology, education, and more recently in the health-related professions. Factor analysis is an important tool that can be used in the development, refinement, and evaluation of tests, scales. The objective of the paper is to provide an detailed exploratory discussion about Factor analysis.. Factor analysis includes both component analysis and common factor analysis*
.

Keywords: *Factor analysis, covariance relation, conventional fuzzy clustering algorithm, Principal component analysis, Factor rotation, Factor loading, Factor score.*

1. INTRODUCTION

Multivariate analysis is a statistical technique that simultaneously analyzes multiple measurements [1] on individuals or objects under investigation. Need of multivariate analysis arised due to inability of conventional fuzzy clustering algorithms to form fuzzy clusters when only co-occurrence relations among individuals and categories are given and the criterion to obtain clusters is not available [3].Earlier, people were extracting complex patterns from data manually but the ever increasing volume of digital data in modern times has provoked research towards more automated approaches. Data mining tools could be designed to answer business questions [7], which took traditionally much time to resolve. Simultaneous analysis of more than two variables can be considered multivariate analysis [1]. The most important thing is to find the applicable criteria for multivariate analysis. The criterion is as follows.

Many multivariate techniques are extensions of univariate analysis and bivariate analysis [2]. Some multivariate techniques provide a means of performing in a single analysis. Factor analysis includes both component analysis and common factor analysis. [4-5].

2. A CLASSIFIACTION OF MULTIVARIATE TECHNIQUES

The classification of Multivariate technique is based on three judgments [1]

I. The research objective and nature of data.
II. Can the variable be divided into independent and dependent .If they can, how many variables are treated as dependent in a single analysis?
III. How are the variable, both dependent and independent measured?

A dependence technique may be defined as one in which a variable or set of variables is identified as the dependent variable. There are various types of multivariate techniques. Multivariate analysis is an ever-expending set of techniques for data analysis for data analysis that encompasses a wide range of possible research situations as evidenced by the classification scheme. [1]

I. Principal component and common factor analysis [8](Principal Component Analysis (PCA) is a well-known tool often used for the exploratory analysis of a numerical data set)
II. Multiple regression and multiple correlation
III. Multiple discriminated analysis & logistic regression
IV. Canonical correlation analysis
V. Multivariate analysis of variance and covariance
VI. Co-joint analysis
VII. Cluster analysis
VIII. Perceptual mapping
IX. Correspondence analysis
X. Structural equation modeling and confirmatory factor analysis
XI. Factor Analysis.

3. APPLICATION OF MULTIVARIATE CLUSTERING

I. Used as tool for characterization in medical science e.g Characterization of anti-inflammatory compounds[9], Rapid characterization of biomass using near infrared spectroscopy[11]

II. Used as a a novel PAT tool to improve process understanding in fluid bed granulation in Combining microwave resonance technology to multivariate data analysis [10].

III. Used in cell culture bioprocess data- Lactate consumption as process indicator[12]

IV. Application for a meta-bionomics study.

4. FACTOR ANALYSIS

Factor analysis is an interdependence technique, whose primary purpose is to define the underlying structure among the variables in the analysis[4]. Factor analysis is first multivariate technique because it can play unique role in the application of other multivariate techniques. Factor analysis provides the tools for analyzing the structure of the interrelationship among a large number of variables. The set of variable that are highly inter related known as factors. These factors are assumed to represent dimensions within the data sets.

The starting point of factor analysis is a correlation matrix. Which the inter correlations between the variables are presented. The dimensionality of this matrix can be reduced by "looking for variables that correlate highly".

- Psychologist Charles Spearman, who hypothesized that the enormous variety of tests of mental ability--measures of mathematical skill, vocabulary, other verbal skills, artistic skills, logical reasoning ability, One underlying "factor" of general intelligence is that he called *g*. He hypothesized that if *g* could be measured and you could select a subpopulation of people with the same score on *g*, in that subpopulation you would find no correlations among any tests of mental ability. He hypothesized that *g* was the only factor common to all those measures. It was an interesting idea, but it turned out to be wrong.
- Consider various measures of the activity of the autonomic nervous system--heart rate, blood pressure, etc. Psychologists have wanted to know whether, except for random fluctuation, those entire measures move up and down together—the "activation" hypothesis. Or do groups of autonomic measures move up and down together, but separate from other groups? Or are all the measures largely independent? An unpublished analysis of mine found that in one data set, at any rate, the data fitted the activation hypothesis quite well
- Suppose many species of animal (rats, mice, birds, frogs, etc.) are trained that food will appear at a certain spot whenever a noise--any kind of noise--comes from that spot. You could then tell whether they could detect a particular sound by seeing whether they turn in that direction when the sound appears. Then if you studied many sounds and many species, you might want to know on how many different dimensions of hearing acuity the species vary. One hypothesis would be that they vary on just three dimensions--the ability to detect high-frequency sounds, ability to detect low-frequency sounds, and ability to detect intermediate sounds. On the other hand, species might differ in their auditory capabilities on more than just these three dimensions. For instance, some species might be better at detecting sharp click-like sounds while others are better at detecting continuous hiss-like sounds.
- Suppose each of 500 people, who are all familiar with different kinds of automobiles, rates each of 20 automobile models on the question, "How much would you like to own that kind of automobile?" We could usefully ask about the number of dimensions on which the ratings differ. A one-factor theory would posit that people simply give the highest ratings to the most expensive models. A two-factor theory would posit that some people are most attracted to sporty models while others are most attracted to luxurious models. Three-factor and four-factor theories might add safety and reliability. Or instead of automobiles you might choose to study attitudes concerning foods, political policies, political candidates, or many other kinds of objects.
- Rubenstein (1986) studied the nature of curiosity by analyzing the agreements of junior-high-school students with a large battery of statements such as "I like to figure out how machinery works" or "I like to try new kinds of food." A factor analysis identified seven factors: three measuring enjoyment of problem-solving, learning, and reading; three measuring interests in natural sciences, art and music and

new experiences in general; and one indicating a relatively low interest in money.

5. MAKING SENSE OF FACTOR ANALYSIS

1. The Use of Factor Analysis for Instrument Development in Health Care Research presents a straight forward explanation of the complex statically procedures involved in factor analysis.[3]
2. The Use of Factor Analysis for Instrument Development in Health Care Research offers a practical method for developing tests, validating instruments and reporting outcomes through the use of factor analysis.[3]

6. STEPWISE TREATMENT OF FACTOR ANALYSIS

It consists of seven main steps [5]:
Reliable measurements, Correlation matrix, Factor analysis versus principal component analysis, the number of factors to be retained, Factor rotation, and use and interpretation of the results.

6.1 PROCEDURE:

Factor analysis from a correlation matrix of variables that are measured at an interval level. Secondly, the variables should roughly be distributed; this makes it possible to "generalize the results of your analysis beyond the sample collected". Thirdly, the sample size should be taken into consideration, as correlations are not resistant, and can hence seriously influence the reliability of the factor analysis

6.2 CORRELATION MATRIX:

When the data are appropriate, it is possible to create a correlation matrix by calculating the correlations between each pair of variables.

6.3 FACTOR ANALYSIS VERSUS PRINCIPAL COMPONENT ANALYSIS:

After obtaining the correlation matrix [6], it is time to decide which type of analysis to use: factor analysis or principal component analysis. The main difference between these types of analysis lies in the way the communalities are used. In principal component analysis it is assumed that the communalities are initially. In other words, principal component analysis assumes that the total variance of the variables can be accounted for by means of its components or factors, and hence that there is no error variance. On the other hand, factor analysis does assume error variance4. This is reflected in the fact that in factor analysis the communalities have to estimated, which makes factor analysis more complicated than principal component analysis, but also more conservative.

6.4 NUMBER OF FACTORS TO BE RETAINED

The number of factors [6] to be retained is similar to the number of positive eigen values of the correlation matrix. This may, however, not always lead to the right solutions, as it is possible to obtain eigen value that are positive but very close to zero. Therefore, some rules of thumb have been suggested for determining how many factors should be retained
1. Retain only those factors with an eigenvalue larger than 1 (Guttman-Kaiser rule);
2. Keep the factors which, in total, account for about 70-80% of the variance;
3. Make a scree-plot; keep all factors before the breaking point or elbow.

6.5 FACTOR ROTATION:

After factor extraction [6] it might be difficult to interpret and name the factors / components on the basis of their factor loadings. Remember that the criterion of principal component analysis that the first factor accounts for the maximum part of the variance; this will often ensure that "most variables have high loadings on the most important factor, and small loadings on all other factors" . Thus, interpretation of the factors can be very difficult. A solution for this difficulty is factor rotation. Factor rotation alters the pattern of the factor loadings, and hence can improve interpretation. Rotation can best be explained by imagining factors as axes in a graph, on which the original variables load. By rotating these axes, then, it is possible to make clusters of variables load optimally.

6.6 RESULTS: FACTOR LOADING AND FACTOR SCORES

Factor loadings [6] are important for the interpretation of the factors, especially the high ones. One can wonder, however, how high a loading has to be in order to determine the interpretation of the factor in a significant way.
This is dependent of the sample size: the bigger the sample the smaller the loadings can be to be significant. Stevens (1992) made a critical values table to determine this significance states, on the other hand, that "the significance of a loading gives little indication of the substantive importance of a variable to a factor". For this to determine, the loadings have to be squared. Stevens then "recommends interpreting only factor loadings with an absolute value greater than 0.4 (which explain around 16% of variance)". This is only possible in principal component analysis, though. In factor analysis the amount of explained variance is calculated in a different

way. The second result of factor analysis is the factor scores. These factor scores can be useful in several ways. Field and Rietveld & Van Hout name the following:

1. If one wants to find out "whether groups or clusters of subjects can be distinguished that behave similarly in scoring on a test battery [and] the latent, underlying variables are considered to be more fundamental than the original variables, the clustering of factor scores in the factor space can provide useful clues to that end"
2. The factor scores can serve as a solution to multi-co linearity problems in multiple regression. After all, the factor scores are uncorrelated (in the case of orthogonal rotation.
3. The factor scores can also be useful in big experiments, containing several measures using the same subjects. If it is already known in advance that "a number of dependent variables used in the experiment in fact constitute similar measures of the same underlying variable, it may be a good idea to use the scores on the different factors, instead of using the scores on the original variables".

7. EXTRACTION METHODS COMMONLY USED IN FACTOR ANALYSIS:

- Principal components analysis (PCA)
- Principal axis factoring (PAF)
- Maximum likelihood
- Un-weighted least squares
- Generalized least squares Alpha factoring

8. THE GOAL: UNDERSTANDING OF CAUSES

Many statistical methods are used to study the relation between independent and dependent variables.
Factor analysis [4]is different; it is used to study the patterns of relationship among many dependent variables, with the goal of discovering something about the nature of the independent variables that affect them, even though those independent variables were not measured directly. Thus answers obtained by factor analysis are necessarily more hypothetical and tentative than is true when independent variables are observed directly. The inferred independent variables are called *factors*. A typical factor analysis suggests answers to four major questions:

I. How many different factors are needed to explain the pattern of relationships among these variables?
II. . What is the nature of those factors?
III. How well do the hypothesized factors explain the observed data?
IV. How much purely random or unique variance observed variable include?

9. FACTOR ANALYSIS VERSUS CLUSTERING AND MULTIDIMENSIONAL SCALING

Another challenge to factor analysis[4] has come from the use of competing techniques such as cluster analysis and multidimensional scaling. While factor analysis is typically applied to a correlation matrix, those other methods can be applied to any sort of matrix of similarity measures, such as ratings of the similarity of faces. But unlike factor analysis, those methods cannot cope with certain unique properties of correlation matrices, such as reflections of variables. For instance, if you reflect or reverse the scoring direction of a measure of "introversion", so that high scores indicate "extroversion" instead of introversion, then you reverse the signs of all that variable's correlations: -.36 becomes +.36, +.42 becomes -.42, and so on. Such reflections would completely change the output of a cluster analysis or multidimensional scaling, while factor analysis would recognize the reflections for what they are; the reflections would change the signs of the "factor loadings" of any reflected variables, but would not change anything else in the factor analysis output. Another advantage of factor analysis over these other methods is that factor analysis can recognize certain properties of correlations. For instance, if variables A and B each correlate .7 with variable C, and correlate .49 with each other, factor analysis can recognize that A and B correlate zero when C is held constant because .72 = .49. Multidimensional scaling and cluster analysis have no ability to recognize such relationships, since the correlations are treated merely as generic "similarity measures" rather than as correlations. We are not saying these other methods should never be applied to correlation matrices; sometimes they yield insights not available through factor analysis. But they have definitely not made factor analysis obsolete.

10. FACTORS "DIFFERENTIATING" VARIABLES VERSUS FACTORS "UNDERLYING" VARIABLES

When someone says casually that a set of variables seems to reflect "just one factor", there are several things they might mean that have nothing to do with factor analysis. If we word statements more carefully, it turns out that the phrase "just one factor *differentiates* these variables" can mean several different things, none of which corresponds to the factor analytic conclusion that "just one factor *underlies* these variables". One possible

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS…..

meaning of the phrase about "differentiating" is that a set of variables all correlate highly with each other but differ in their means. A rather similar meaning can arise in a different case. Consider several tests A, B, C, D which test the same broadly-conceived mental ability, but which increase in difficulty in the order listed. Then the highest correlations among the tests may be between adjacent items in this list (rAB, rBC and rCD) while the lowest correlation is between items at the opposite ends of the list (rAD). Someone who observed this pattern in the correlations among the items might well say the tests "can be put in a simple order" or "differ in just one factor", but that conclusion has nothing to do with factor analysis. This set of tests would *not* contain just one common factor. A third case of this sort may arise if variable A affects B, which affects C, which affects D, and those are the only effects linking these variables. Once again, the highest correlations would be rAB, rBC and rCD while the lowest correlation would be rAD. Someone might use the same phrases just quoted to describe this pattern of correlations; again it has nothing to do with factor analysis. A fourth case is in a way a special case of all the previous cases: a perfect Guttman scale. A set of dichotomous items fits a Guttman scale if the items can be arranged so that a negative response to any item implies a negative response to all subsequent items while a positive response to any item implies a positive response to all previous items.

For a trivial example consider the items

- Are you above 5 feet 2 inches in height?
- Are you above 5 feet 4 inches in height?
- Are you above 5 feet 6 inches in height?

To be consistent, a person answering negatively to any of these items must answer negatively to all later items, and a positive answer implies that all previous answers must be positive. For a nontrivial example consider the following questionnaire items:

- Should our nation lower tariff barriers with nation B?
- Should our two central banks issue a single currency?
- Should our armies become one?
- Should we fuse with nation B, becoming one nation?

If it turned out that these items formed a perfect Guttman scale, it would be easier to describe peoples' attitudes about "nation B" than if they didn't. When a set of items does form a Guttman scale, interestingly it does not imply that factor analysis would discover a single common factor. A Guttman scale implies that one factor *differentiates* a set of items (e.g, "favorableness toward cooperation with nation B"), not that one factor *underlies* those items. Applying multidimensional scaling to a correlation matrix could discover all these simple patterns of differences among variables. Thus multidimensional scaling seeks factors which *differentiate* variables while factor analysis looks for the factors which *underlie* the variables. Scaling may sometimes find simplicity where factor analysis finds none, and factor analysis may find simplicity where scaling finds none.

11. COMPARISON OF TWO FACTOR ANALYSIS

Since factor loadings [4] are among the most important pieces of output from a factor analysis, it seems natural to ask about the standard error of a factor loading, so that for instance we might test the significance of the difference between the factor loadings in two samples. Unfortunately, no very useful general formula for such a purpose can be derived, because of ambiguities in identifying the factors themselves. To see this, imagine that "math" and "verbal" factors explain roughly equal amounts of variance in a population.

The math and verbal factors might emerge as factors 1 and 2 respectively in one sample, but in the opposite order in a second sample from the same population. Then if we mechanically compared, for instance, the two values of the loading of variable 5 on factor 1, we would actually be comparing variable 5's loading on the math factor to its loading on the verbal factor. More generally, it is never completely meaningful to say that one particular factor in one factor analysis "corresponds" to one factor in another factor analysis. Therefore we need a completely different approach to studying the similarities and differences between two factor analyses. Actually, several different questions might be phrased as questions about the similarity of two factor analyses. First we must distinguish between two different data formats:

1. Same variables, two groups. The same set of measures might be taken on men and women, or on treatment and control groups. The question then arises whether the two factor structures are the same.

2. One group, two conditions or two sets of variables. Two test batteries might be given to a single group of subjects, and questions asked about how the two sets of scores differ. Or the same battery might be given under two different conditions

11.1 Comparing Factor Analyses in Two Groups

In the case of two groups and one set of variables, a question about factor structure is obviously not asking whether the two groups differ in means; that would be a question for MANOVA (multivariate analysis of variance). Unless the two sets of means are equal or have somehow been made equal, the question is also not asking whether a correlation matrix can meaningfully be computed after pooling the two samples, since differences in means would destroy the meaning of such a matrix. The question, "Do these two groups have the same factor structure?" is actually quite different from the question, "Do they have the same factors?" The latter question is closer to the question, "Do we need two different factor analyses for the two groups?" To see the point, imagine a problem with 5 "verbal" tests and 5 "math" tests. For simplicity imagine all correlations between the two sets of tests are exactly zero. Also for simplicity consider a component analysis, though the same point can be made concerning a common factor analysis. Now imagine that the correlations among the 5 verbal tests are all exactly. 4 among women and .8 among men, while the correlations among the 5 math tests are all exactly .8 among women and .4 among men. Factor analyses in the two groups separately would yield different factor structures but identical factors; in each gender the analysis would identify a "verbal" factor which is an equally-weighted average of all verbal items with 0 weights for all math items, and a "math" factor with the opposite pa tern. In this example nothing would be gained from using separate factor analyses for the two genders, even though the two factor structures are quite different.

Another important point about the two-group problem is that an analysis which derives 4 factors for group A and 4 for group B has as many factors total as an analysis which derives 8 in the combined group. Thus the practical question may be not whether analyses deriving *m* factors in each of two groups fit the data better than an analysis deriving *m* factors in the combined group. Rather the two separate analyses should be compared to an analysis deriving 2*m* factors in the combined group. To make this comparison for component analysis, sum the first *m* eigenvalues in each separate group, and compare the mean of those two sums to the sum of the first 2*m* eigenvalues in the combined group. It would be very rare that this analysis suggests that it would be better to do separate factor analyses for the two groups. This same analysis should give at least an approximate answer to the question for common factor analysis as well. Suppose the question really is whether the two factor structures are identical. This question is very similar to the question as to whether the two correlation or covariance matrices are identical--a question which is precisely defined with no reference to factor analysis at all. Tests of these hypotheses are beyond the scope of this work, but a test on the equality of two covariance matrices appears in Morrison (1990) and other works on multivariate analysis.

11.2 COMPARING FACTOR ANALYSYS OF TWO SETS OF VARIABLES IN A SINGLE GROUP

One question people often ask is whether they should analyze variable sets A and B together or separately. The answer is usually "together", unless there is obviously no overlap between the two domains studied.

After all, if the two sets of variables really are unrelated then the factor analysis will tell you so, deriving one set of factors for set A and another for set B. Thus to analyze the two sets separately is to prejudge part of the very question the factor analysis is supposed to answer for you. As in the case of two separate samples of cases, there is a question which often gets phrased in terms of factors but which is better phrased as a question about the equality of two correlation or covariance matrices--a question which can be answered with no reference to factor analysis. In the present instance we have two parallel sets of variables; that is, each variable in set A parallels one in set B. In fact, sets A and B may be the very same measures administered under two different conditions. The question then is whether the two correlation matrices or covariance matrices are identical. This question has nothing to do with factor analysis, but it also has little to do with the question of whether the AB correlations are high. The two correlation or covariance matrices within sets A and B might be equal regardless of whether the AB correlations are high or low. Darlington, Weinberg, and Walberg (1973) described a test of the null hypothesis that the covariance matrices for variable sets A and B are equal when sets A and B are measured in the same sample of cases. It requires the assumption that the AB covariance matrix is symmetric. Thus for instance if sets A and B are the same set of tests administered in years 1 and 2, the assumption requires that the covariance between test X in year 1 and test Y in year 2 equal the covariance between test X in year 2 and test Y in year 1. Given this assumption, you can simply form two sets of scores I'll call A+B and AB, consisting of the sums and differences of parallel variables in the two sets. It then turns out that the original null hypothesis is equivalent to the hypothesis that all the variables in set A+B are uncorrelated with all variables in set A-B. This hypothesis can be tested with MANOVA.

12. LIMITATIONS OF FACTOR ANALYSIS

Three of the most frequent cited limitations are as follow:

1. Because many techniques for performing exploratory factor analyses are available controversy exists over which technique is the best. The subjective aspects of factor analysis (i.e. ,deciding how many factors to extract, which technique should be used to rotate the factor axes, which factor loading are significant)are all subject to many differences in opinion.
2. The problem of reliability is real.

REFERENCES

[1] Hair, Black, "Multivariate data analysis by 7th edition, "Pearson publication.
[2] Chi-Hyon Oh, Katsuhiro Honda and Hidetomo Ichihashi "Fuzzy Clustering for Categorical Multivariate Data" Proceedings of Joint 9th IFSA World Congress and 20th NAFIPS International Conference, 2154-2159 (2001) 2000.
[3] Alanzado, A.C., Miyamoto, S., "Fuzzy *c*-means and a mixture distribution for clustering of noisy data", Proceedings of SCIS & ISIS Conference, October 2002, pp.21–25.
[4] Mr. Brett Williams, Andrys Onsman, Ted Brown, "Exploratory factor analysis: A five-step guide for novices", Australasian Journal of Paramedicine: Vol. 8, 2010.
[5] Tsukuba, Pett MA, Lackey NR, Sullivan JJ. "Making Sense of Factor Analysis: The use of factor analysis for instrument development in health care research", California: Sage Publications Inc; 2003.
[6] Darlington, Richard B., Sharon Weinberg, and Herbert Walberg," Canonical variate analysis and related techniques."Reviews of educational researches, pp 453-454, 2004
[7] N. Sivaram, K. Ramar, "Applicability of Clustering and Classification Algorithms for Recruitment Data Mining", International Journal of Computer Applications, Volume 4 – No.5, July 2010
[8] Paolo Giordani, Henk A.L. Kiers," Principal Component Analysis of symmetric fuzzy data", Computational Statistics & Data Analysis 45 ,pp 519 – 54,2004
[9] Kitty C.M. Verhoeckx, Sabina Bijlsma, Sonja Jespersen,Raymond Ramaker, Elwin R. Verheij, Renger F. Witkamp,Jan van der Greef, Richard J.T. Rodenburg, "Characterization of anti-inflammatory compounds using transcriptomics, proteomics, and metabolomics in combination with multivariate data analysis", International immunopharmacology, www.elsevier.com,2004
[10] Vera Lourenço,Thorsten Herdling , Gabriele Reich,José C. Menezes,Dirk Lochmann," Combining microwave resonance technology to multivariate data analysis as a novel PAT tool to improve process understanding in fluid bed granulation", European Journal of Pharmaceutics and Bio-pharmaceutics , 2011, pp-513-521
[11] Kofi Nkansah,Benjamin Dawson-Andoh , Jeffrey Slahor," Rapid characterization of biomass using near infrared spectroscopy coupled with multivariate data analysis: Part 1yellow-"poplar",Bioresource Technology, 2010,pp:4570-4576 Journal homepage : www.elsevier.com/locate/biortech
[12] Huong Le, Santosh Kabbur, Luciano Pollastrini , Ziran Sun, Keri Mills, Kevin Johnson, George Karypis, Wei-Shou Hu," Multivariate analysis of cell culture bioprocess data—Lactate consumption as process indicator", Journal of Biotechnology, 2010,pp 210-233 Journal of Biotechnology journal homepage: www.elsevier.com/locate/jbiotec