PRINCIPAL COMPONENTS AND THE MAXIMUM LIKELIHOOD METHODS AS TOOLS TO ANALYZE LARGE DATA WITH A PSYCHOLOGICAL TESTING EXAMPLE

Markela Muca
Llukan Puka
Klodiana Bani
Department of Mathematics, Faculty of Natural Sciences,
University of Tirana, Albania
Edmira Shahu
Department of Economy and Agrarian Policy,
Faculty of Economy and Agribusiness, Agricultural University of Tirana.

Abstract

Basing on the study of correlations between large numbers of quantitative variables, the method factor analysis (FA) aims at finding structural anomalies of a communality composed of p-variables and a large number of data (large sample size). It reduces the number of original (observed) variables by calculating a smaller number of new variables, which are called factors (Hair, et al., 2010). This paper overviews the factor analysis and their application. Here, the method of principal components analysis (PCA) to calculate factors with Varimax rotation is applied. The method of maximum likelihood with Quartimax rotation is used for comparison purposes involving the statistic package SPSS. The results clearly report the usefulness of multivariate statistical analysis (factor analysis). The application is done by a set of data from psychological testing (Revelle, 2010).

Keywords: Factorial analysis (FA), Principal components analysis (PCA), Maximum likelihood methods, orthogonal rotation

1. Introduction

Factor analysis is a class of multivariate statistical methods whose primary purpose is data reduction and summarization. It addresses the problem of analyzing the interrelationship among a large number of variables and then explaining these variables in terms of their common, underlying factors (Hair et al., 1979).

This method is a summary to the principal component (PCA) method, since the results and the interpretations of methods are similar, but the mathematical models are different. The FA method relates to the correlations between a large numbers of quantitative variables. It reduces the number of primary variables by calculating a smaller number of new variables, which are called factors. This reduction is achieved by grouping variables into factors by which means each variable within each factor is closely correlated and variables which belong to different factors are less correlated (Hair et al., 2010).

In the area of factors calculation, the principal components analysis (PCA), and the maximum likelihood method (ML), are two of the most applied techniques. If PCA is used, then the minimum average partial method can be used (Velicer, 1976) whereas if ML is used, then fit indices can be used as described by (Fabrigar et al. 1999, Browne & Cudeck 1992) For more comprehensive review of options we can see (Fabrigar et al. 1999 and Zwick and Velicer 1986.) There are also other methods named in general as factor rotation, which make an orthogonal transformation of the loadings matrix, (Hair et al., 2010). Here, the loadings of variables of an extracted factor are maximized and the loadings of variables in the other factors are minimized. In this case, the variables are expected to be independent. Consequently, one of the orthogonal rotations is Varimax, which attempts to maximize the variance of squared loadings on a factor (Kim and Mueller, 1978).

The scores of factors are calculated utilizing the regress technique. Their scores values are saved in new variables in the data file and can be used later for statistical analysis. In this way, each factor now embodies a linear combination of the primary variables. We calculate the numerical characteristics for these variables from which the new variables are standardized with mean 0 and variance 1. Achievable by means of correlation matrix, the new variables must be uncorrelated among them.

The present paper overviews factor analysis and related factors extracting methods. In addition, it reports usefulness of factors analysis and, the new variables (factors) are exemplified.

In this paper, use of PCA when calculating factors with Varimax rotation is reported. The maximum likelihood method with Quartimax rotation is applied for calculation purposes. Correlations matrix is applied and, the factors number is chosen by the eigenvalues which are greater than 1.

Factors rotation and interpretation

In the area of FA implementation, usually, the factors found by methods previously described, cannot be easily interpreted. Determining exactly the variables belonging to factor one, two and so on, up to factor q is difficult, as some variables have the tendency to load on some factors. Consequently, many factors can be interpreted by one variable. On contrary, the present paper aims at finding one factor to interpret one or more variables.

There are several techniques, named together as factor rotation, which make an orthogonal transformation of the matrix L (factor loadings matrix), so that the interpretation of factors can be simplified (Johnson, Wichern, 1982). Reporting of the percentage of variables is explained by each axe (factor), the rotation affects loadings (big loadings become bigger and small loadings become smaller) even in individual values. However, the sum of individual values remains unchanged. Once the factors are extracted, these techniques could be used bringing different results for the same communality of data. Nevertheless, all analysts aim at building simple structures where each variable load only in one factor. Consequently, one variable could be interpreted only by one factor.

Here in this paper, both Varimax and Quartimax rotation are reported.

Choosing the number of factors

Choosing of the number of common factors is very important. We draw a graphic of pairs (j, λ_j), the “scree plot”, and we observe the position in which this graphic begin to become “flat” (Cattell, 1966). Another criterion to address the number of factor problem is the Kaiser criterion (Kaiser 1960). With this approach, a factor j is important when the eigenvalues is $\lambda_j > 1$. If the number of factors found by **Kaiser Test** is equivalent with the number of factors, which have resulted even from the “**scree plot”,** than we can continue with the other procedures, or otherwise we have to choose one of the results already obtained. If one of the obtained results from “scree plot” graphic is chosen, the aforementioned procedures and arrange the best number of factors must be repeated. The results change as the number of factors changes. Available options include Kaiser’s (Kaiser 1956) “eigenvalues greater than one” rule, the scree plot, a priory theory and retaining the number of factors that gives a high proportion of variance accounted for or that gives the most interpretable solution.

The application of these two techniques is demonstrated in the paragraphs below.

Calculation of the Factor Scores

The score for each individual chosen in connections with factors is of great importance for FA methods when applied for the calculation of factor loadings. Consequently, being of great importance, the new variables that correspond to the factors could be added to our selection. SPPS builds a column for each factor extracted and then places the scores of these factors for each subject inside this column. Once placed in the column, these scores can be used for statistical analysis or simply to identify groups of subjects. After the evaluation of dispersion errors, the model $\mathbf{x}-\boldsymbol{\mu}=\boldsymbol{\Gamma}\mathbf{z}+\mathbf{e}$ can calculate scores (the factor values). The Statistical package SPSS contains three different methods of evaluation: Regression, Bartlett (Johnson and Wichern, 1982) and Anderson-Rubin.

Data Application of Factorial Analysis (FA) with PCA method

The set of data psychological testing helps report the usefulness of the aforementioned techniques. The set of data in psychological testing (Revelle, 2010) provides information for n=1000 individuals. It contains three dependent variables: Prelim, GPA and MA, and five predictor variables. The variables of the set are in Table 1 reported. Determining a smaller number of uncorrelated variables to describe the data is of great interest.

Table 1 Description of the Variables in the Data Set.

GREV	**GRE(Graduate Record Examinations) VERBAL**
GREQ	**GRE QUANTITATIVE**
GREA	**GRE ADVANCED**
ACH	**ACHIEVEMENT**
ANX	**ANXIETY**
Prelim	**RATED PERFORMANCE**
GPA	**GRADUATE PERFORMANCE**
MA	**MASTERS PERFORMANCE**

Initially the PCA method was used to calculate factors with Varimax rotation method. Correlation matrix reports that the eigenvalues values are greater than 1, which is a means to address the choice of the number of factors. The Table 2 shows couples of variables (e.g. GREV, GREQ and GREA) which are better correlated with each-other. Regarding the Meyer-Olkin (KMO) statistical value which in this case is 0.657, see Table 3, we see the Kaiser- certain correlation of data (Hair *et al*., 2010) is reported.

Table 2 Correlation Matrix of Variables

Variables		GREV	GREQ	GREA	Ach	Anx
Correlation	GREV	1.000				
	GREQ	**.729**	1.000			
	GREA	**.641**	**.596**	1.000		
	Ach	.006	.007	**.453**	1.000	
	Anx	.010	.005	-.390	**-.556**	1.000
Sig. (1-tailed)	GREV					
	GREQ	.000				
	GREA	.000	.000			
	Ach	.430	.414	.000		
	Anx	.374	.431	.000	.000	

a) Determinant = .104

Table 3 KMO and Bartlett's test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		**.657**
Bartlett's Test of Sphericity	Approx. Chi-Square	2256.716
	Df	10
	Sig.	**.000**

Table 4 shows the proportion of variability which is firstly explained with all factors together and then only with the factors before and after rotation. The result shows the two first common factors, which explain 81% of the total variance, a quite good percentage. After rotation method, this percentage does not change, but it changes the percentage that explains each factor. Specifically, these percentages are transformed in order to reduce the differences between them after rotation.

Table 4 Total Variance Explained by Components before and after Rotation Method

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.453	49.063	49.063	2.453	49.063	49.063	2.265	45.292	45.292
2	1.609	32.180	81.243	1.609	32.180	81.243	1.798	35.951	81.243
3	.447	8.945	90.188						
4	.282	5.638	95.826						
5	.209	4.174	100.000						

Another analysis (scree plot, fig 1) confirms the conclusions: the graphic representation of couples is in the same order as reported in the Table 4. Separation processes of the two eigenvalues greater than 1 in conjunction with those remaining.

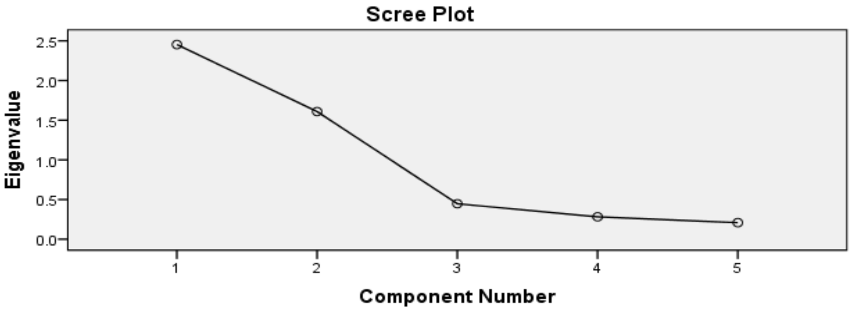

Fig. 1 Scree Plot of the Principal Components Analysis

Table 5 Component Analysis Factor Matrix Before and After Rotation Method

PCA	Components	
	1	2
GREV	.794	**.472**
GREQ	.777	.472
GREA	.920	-.093
Ach	.451	-.757
Anx	-.411	.763

a)

PCA	Components	
(VARIMAX)	1	2
GREV	.923*	**-.040**
GREQ	.907*	-.049
GREA	.767*	.517
Ach	.040	**.880**
Anx	-.001	**-.867**

b)

Reporting correlation with the variables and the role they play when interpreting a variable, the factors loadings before and after rotation is in table 5 shown. Table 5 a) reports the first factor having the highest loadings for variables {GREV, GREQ, and GREA}. Table 5 b) reports the situation after rotation.

The linkage process of variables with the first factor, but with higher loadings is clear. Table 2 (Table of correlation) reports that in general groups of variables found in this way are reasonable as correlations between groups are important.

Table 6 Component Analysis Factor Matrix**Communalistic (PCA)**

Component Matrix PCA

PCA	Initial	Extraction
GREV	1.000	.854*
GREQ	1.000	.825
GREA	1.000	.855
Ach	1.000	.776
Anx	1.000	.751

a)

	Component	
PCA	1	2
GREV	.794*	.472*
GREQ	.777	.472
GREA	.920	-.093
Ach	.451	-.757
Anx	-.411	.763

b)

Confirmed by the results of Table 6 b), (0.854=0.7942+0.4722), Table 6 a), shows how 0.854 (or 85.4%) of variable variance GREV is explained by means of the first and second factors. In general, the communalities show variables, for which the factor analysis is best working or at least, well (Hair et al., 2010).

Implementation of FA with Maximum Likelihood Method (ML)

ML method is here applied to calculate the factor loadings with Quartimax rotation within the same set of data. The number of eigenvalues greater than 1 are a means to address the choosing of the number of factors, while scores are calculated using the regression method. Table 7 reports that they are the two first common factors.

Table 7 Specific variances by factors before and after rotation methods.

Total Variance Explained

Factor	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.453	49.063	49.063	2.202	44.036	44.036	2.017	40.349	40.349
2	1.609	32.180	81.243	1.225	24.499	68.535	1.409	28.186	68.535
3	.447	8.945	90.188						
4	.282	5.638	95.826						
5	.209	4.174	100.000						

The first common factor explains about 69% of total variance, that it is not a good percentage. After rotation this value does not change. It changes only the percentage which explains each factor. The number of factors is the same as the one obtained from the first technique, but the percentage explained by these factors is a bit smaller.

In Table 8 we have the factor loadings before and after rotation. On the left, we can see that variables {GREV, GREQ, GREA} all load highly on factor 1, while on the right, after rotation, we see that the loadings for each variable within each factor has changed.

Table 8 Factor-Loading Matrix Before and After Rotation (ML)

ML	Factors	
	1	2
GREV	.803	**.376**
GREQ	.746	.345
GREA	.895	-.207
Ach	.341	-.716
Anx	-.288	.640

ML	Factors	
QUARTIMAX)	1	2
GREV	.886*	**-.036**
GREQ	.821*	-.030
GREA	.746*	.537
Ach	.038	**.792**
Anx	-.018	**-.701**

Conclusion

Based on the study of correlations between large numbers of quantitative variables, the factor analysis (FA) method aims at finding structural anomalies of a communality composed of p-variables and a large number of data (large sample size). It reduces the number of original (observed) variables by calculating a smaller number of new variables, which are called factors. In PCA the original variables are transformed into the smaller set of linear combination, with all of the variance in the variables being used. In FA (ML), however, factors are estimated using mathematical model, where only the shared variance is analyzed. However, principal components analysis is often preferred as a method for data reduction, while principal factors analysis is often preferred when the goal of the analysis is to detect structure.

Varimax and Quartimax are the two orthogonal approaches, which are used to report on FA method. The most commonly used orthogonal approach is the Varimax method, which aims to minimize the number of variables that have high weights on each factor.

Results are described via a database from a psychological testing. A possible database with information from Albania for similar analysis of practical interest as a future task is being considered.

References:

Browne, M. W. and Cudeck R. 1992. Alternative ways of assessing model fit. Sociological Methods and Research.

Cattell Raymond Bernard, 1966. The scree test for the number of factors. *Multivariate Behavioral Research*.

Fabrigar, L.R... Wegener, D.T. MacCallum, R.C and Strahan, E.J. 1999. Evaluating the use of explanatory factor analysis in psychological Assessment.

Johnson R, Wichern D. 1982. Applied Multivariate Statistical Analysis, Prentice Hall, New Jersey, and 6th ed.

JR Hair JF, Black WC, Babin BJ, Anderson RE. 2010. Multivariate data analysis, 7th ed. Macmillan, New York.

Joseph F. Hair, Jr., Rolph E Anderson, Roland L. Tatham and Bernie J. Grablowsky, 1979. Multivariate Data Analysis, Petroleum Publishing Company, ISBN 0-87814-077-9

Kaiser, H. F., 1956. The varimax methods of factor analysis. Unpublished doctoral dissertation, University of California, Berkeley.

Kaiser, H.F. 1960. The application of electronic computers to factor analysis. Educational and Psychological Measurement.

Kim, J.O. and Mueller, C. W., 1978. Factor analysis: Statistic methods and practical issues. Beverly Hills, CA: Sage.

Velicer, W.F. 1976. Determining the number of components from the matrix of partial correlations. Psychometrical.

William R. 2010. http://personalityproject.org/r/datasets/psychometrics.prob2.txt http://www.ets.org/gre/revised_general/about/content/.

Zwick, W.R. and Velicer W.F. 1986. Comparison if five rules for determining the number of components to retain. Psychological Bulletin.