Journal of Data Science 16(2017), 615-630

Factor Analysis as a tool for Pattern Recognition in biomedical research; a review with application in R software

Dimitris Panaretos[1],George Tzavelas[2], Malvina Vamvakari[3], Demosthenes Panagiotakos[4]

Abstract: Factor Analysis is one of the data mining methods that can be used to analyse, mainly large-scale, multi-variable datasets. The main objective of this method is to derive a set of uncorrelated variables for further analysis when the use of highly inter-correlated variables may give misleading results in regression analysis. In the light of the vast and broad advances that have occurred in factor analysis due largely to the advent of electronic computers, this article attempt to provide researchers with a simplified approach to comprehend how exploratory factors analysis work, and to provide a guide of application using *R*. This multivariate mathematical method is an important tool which very often used in the development and evaluation of tests and measures that can be used in biomedical research. The paper comes to the conclusion that the factor analysis is a proper method used in biomedical research, just because clinical readers can better interpret and evaluate their goal and results.

Keywords: biostatistics; exploratory factor analysis; multivariate analysis; pattern analysis; R programming language

1. INTRODUCTION

In the last years, along with computer development, the need for processing large amount of information, i.e., Big Data concept, arose rapidly in various scientific domains. This created the need to develop, primarily, theoretical methods that could enable researchers to handle these statistical analyses and, more importantly, to apply them in practice through a software program. However, since multivariable analysis is not a solution to all statistical difficulties, the researcher should not only be familiar with all the existing methods in a given period, but also be able to develop the most appropriate way of analyzing the data obtained, by understanding their structure.

The past years, social and biomedical research has incorporated into the analytical methodologies used, the pattern recognition analysis. Pattern analysis is a classical multivariate statistical approach that aims to identify patterns in data in order to show certain attributes. Psychological characteristics, emotions and behaviours, socio-economic dimensions, as well as dietary patterns, are usually extracted through factor analysis (FA), or principal component analysis (PCA) (Panaretos, 2017). PCA was created to be an algebraic method that aims in reducing the dimensionality of multivariate data while preserving as much of the variance of the initial variables as possible (Hotteling, 1933). On the other hand, FA is a statistical method of minimizing the number of variables under investigation while concomitantly maximizing the amount of information in the analysis (Gorsuch, 1983). In recent years FA has been widely employed by biomedical researchers and satisfactory progress and results have been achieved in the field of pattern recognition, from genes to human behaviour.

The purpose of the present review was to briefly discuss exploratory factor analysis as a pattern recognition tool, with an emphasis on biomedical research. Also, using dietary data from the ATTICA epidemiological study, we implemented factor analysis in R open source software - in order to extract dietary patterns and to provide a guide for dietary pattern recognition analysis.

Brief History of Factor Analysis

Factor analysis (FA) has been successfully used in a wide variety of industries and fields, but because of its use was pioneered in the field of psychology the technique itself is often mistakenly considered as psychological theory. The origin of factor analysis is generally ascribed to Charles Spearman (1904), a British psychologist. Spearman and his colleagues (e.g. Burt, 1939, 1941; Garnett, 1919; Ledermann, 1937; 1938; Spearman, 1904, 1922, 1923, 1927, 1928, 1929, 1930; Thomson, 1934, 1936, 1938) pursued the concept of the Two Factor Theory. In particular, Spearman developed a statistical procedure, known as factor analysis, to explain the relationships among various measures of mental ability (memory, physical abilities, and the senses) by means of a single (factor) ability which he called general intelligence, or "g". But it was obvious to Spearman, that g did not account for all the variance. So, in his paper "General Intelligence, Objectively Determined and Measured" was published in the American Journal of Psychology proposed two factors of intelligence: (1) general intelligence (g – factor) that constitute the first and most important aspect of intelligence and (2) specific intelligence (S – factor) that refers to the specific abilities for performing various task. The S – factor varies from

one act to another, while g is available at the same level from all intellectual acts (Spearman, 1927). On the other hand, Luis Thurstone (1938, 1947) an American psychologist, argued that, if the statistical procedure of factor analysis was done in a different way, seven factors would appear. A considerable amount of work on FA followed until today. The principal contributors included Hoteling (1933), Eckart & Young (1936), Holzinger (1937), Thomson (1951), Lawley & Maxwell (1971), Joreskog (1969, 1972) and Velicer (1976). Apparently, factor analysis was used primarily by psychology; however, its use within the health science sector has become much more common during the past few decades (Pett, 2003).

Factor Analysis or Principal Component Analysis

In 1901, Karl Pearson invented a mathematical procedure, which was similar to the method used in principal axes theorem in mechanics and later, Hotelling named (principal components analysis) and independently developed it in the form that has until nowadays (1934). This mean that PCA was created to be a method that aims in reducing the dimensionality of multivariate data while preserving as much of the variance of the initial variables as possible (Hotteling, 1933). That method's main concept is to describe the variance of the total amount of correlated variables

$$X1, X2, \ldots, Xp$$

creating a new total amount of uncorrelated variances

$$Y1, Y2, \ldots, Yp.$$

Each one of the new variables compose linear combination of the initial variables and are created in order to be vertical to each other.

FA has similar goals to PCA, but also many conceptual differences. The basic idea is still that it may be possible to describe a set of n- variables X1, X2, … Xn in terms of a smaller number of variables (factors) and hence elucidate the relationship between these variables. Nevertheless, there are some fundamentals similarities and differences between FA and PCA (Mulaik, 2010; Ogasawara, 2000; Schonemann and Steiger, 1978; Steiger, 1994; Velicer and Jackson, 1990; Widaman, 2007). Both are data reduction techniques and allow us to capture the variance in variables in a smaller set. Also, the steps are the same: extraction, interpretation, rotation, choosing the number of factors or components.

Despite all these similarities, there is a fundamental difference between them. The primary difference between FA and PCA is that the former is based upon a decomposition of the covariance matrix in which the diagonal contains the squared-multiple correlation (or some other initial estimate of explained variance in each observed variable), whereas PCA is a true covariance matrix, with standardized variances on the diagonal. PCA is a linear combination of variables, while FA is a measurement model of a latent variable. In particular, PCA models represent singular value decompositions of random association matrices, whereas a FA incorporates an a priori structure of the error terms. Generally the factor has to be run, if it assumed to test a theoretical model of latent factors causing observed variables, while has to be run PCA if our aim is to simply reduce your correlated observed variables to a smaller set of important independent composite variables (Basilevsky, 2008).

Both methodologies have several applications in research, and particularly in social and biomedical sciences. These multivariate statistical analysis techniques are applied when the interesting is about the inter-relationships of more than one variable at the same time. It is very rare for a researcher to isolate and analyze each variable. The most common procedure that has to be followed is all the variables to be analyzed at the same time in order to reveal the structure of data.

Factor Analysis Model

Factor Analysis is based on the fundamental assumption that some latent variables, which smaller in number than the number of observed variables, are responsible for the co-variation among the observed variables (Kim, 1978).

To illustrate the model, let examine the simplest case where one latent variable, called factor, is responsible for the co-variation between two observed variables. Because F is common to both X1 and X2, it called common factor; likewise, because U1 and U2 are unique to each observed variable, they called unique factors. In algebraic form, the following two equations hold:

$$X1 = \lambda 1F + U1$$

and

$$X2 = \lambda 2F + U2$$

where λi, are called factor loadings, and express the relationship (i.e., in a form of a correlation) of each variable to the common factor F, when the data are standardized.

There are two differences between the common and the unique factors. Firstly, a common factor affects several variables Xi (i = 1, 2, ….n) at the same time – thereby producing one special pattern of relations among the variables – whereas a unique factor affect only one variable at the same time. Secondly, a variable Xi can at the same time be dependent by more than one common factor, but only by one unique factor (Schilderinck, 1977). The factor loadings give an idea about how much the variable has contributed to the factor; the larger the factor loading the more the variable has contributed to that factor (Harman, 1976).

Let's assume a set of observed variables X = [x1, x2, …, xn], supposed to be linked to a smaller number of common factors F = [f1, f2, …, fp], where $p \leq n$ by a regression model of the form (U = [u1, u2, …un] represents the error term):

$$x1 = \lambda 11f1 + \lambda 12f2 + \ldots + \lambda 1kfp + u1$$
$$x2 = \lambda 21f1 + \lambda 22f2 + \ldots + \lambda 2kfp + u2$$
$$\ldots$$
$$xn = \lambda p1f1 + \lambda p2f2 + \ldots + \lambda pkfp + un$$

Also the aforementioned may be written as

$$X = \Lambda * f + U$$

where X, U are column vectors of n components, f is a column vector of p (≤ n) components and Λ is a nXp matrix. Let's assume that the unique factors are uncorrelated with each other and are distributed independently of common factors with zero mean.

The factor analysis model imply that the variance of variable is given by

$$\sigma 2 = Var(xi) = Var(\lambda i1f1 + \lambda i2f2 + \ldots + \lambda ijfk + ui)$$

$$= \quad + \quad + \ldots + \quad + \psi i$$

$$= \quad + \psi i$$

where ψi is the variance of ui.

So the factor analysis' model implies that the variance of each observed variable can be split into two parts. The first given by

$$=$$

called communality of the variable and represent the variance shared with other variables via the common factors. The second part, ψi, called specificity and represent the variance not shared with other variables. Moreover has needed to determine the value of p, the number of factors, and estimate Λ and U.

Rotating the factors

A technique that may help to better retrieve true information from the FA is the rotation of the information axes (factors). It has been suggested that rotation of the axes is required so that the extracting factors can be more interpretable. The rotation maximizes the variance explained of the extracted components and makes the pattern of loadings more well-defined. The largest part of the theory behind the rotation is derived from Thurstone (1947) and Cattell (1978), who advocate that using this theory simplifies FA and makes its interpretation easier and more reliable. The objective of factor rotation is to achieve the most parsimonious and simple structure possible through the manipulation of the factor pattern matrix. Thurstone's guidelines for rotating to simple structure have largely influenced the development of various rotational strategies. The most parsimonious solution, or simple structure, has been explained by Gorsuch (1983) in terms of five principles of factor rotation:

1. Each variable should have at least one zero loading.
2. Each factor should have a set of linearly independent variables whose factor loadings are zero.
3. For every pair of factors, there should be several variables whose loadings are zero for one factor but not the other.
4. For every pair of factors, a large proportion of variables should have zero loadings on both factors whenever more than about four factors are extracted.
5. For every pair of factors, there should only be a small number of variables with nonzero loadings on both.

The rotation can be orthogonal (the factors are uncorrelated) or non-orthogonal (the factors are correlated). Orthogonal rotation shifts the factors in the factor space maintaining 90o angles of the factors to one another to achieve the best simple structure. Orthogonal rotations often do not honour a given researcher's view of reality as the researcher may believe that two or more of the extracted and retained factors are correlated. Secondly, orthogonal rotation of factor solutions may oversimplify the relationships between the variables and the factors and may not always accurately represent these relationships. On the other hand, a non-orthogonal rotation follows the same rotation principles as an orthogonal rotation, but because the factors are not independent, a 900 angle of rotation is not fixed between the axes (Cattel, 1978). The major methods of orthogonal rotation are Varimax, Quartimax and Equimax, while the major methods of non-orthogonal rotation are direct Oblimin and Promax. Nevertheless, there are many additional methods of orthogonal and non-orthogonal rotation like biquartimax, covarimin or biquartimin criterion and oblimax, orthoblique (Harris & Kaiser, 1964) or “Procrustes method”, respectively (in the Greek Mythology, Procrustes was a rogue smith and bandit from Attica region in Ancient Greece, who physically attacked people by stretching them or cutting off their legs, so as to force them to fit the size of an iron bed) (Figure 1).

Varimax, which was proposed by Kaiser (1958), is considered as the most popular rotation method. Each factor has a small number of large loadings and a large number of zero or small loadings. This simplifies the interpretation because after a varimax rotation each original variable tends to be associated with one (or a small number) of the factors, and each factor represents a small number of variables. In essence, higher loadings on a factor are made higher and lower loadings are made lower (Tabachnick & Fidell, 2001). A varimax solution is easily interpreted and provides relatively clear information about which items correlate most strongly with a given factor. A disadvantage of Varimax is that it tends to split up the variances of the major factors among the less important factors,

thus reducing the possibility of identifying an overall general factor. Therefore, if the researcher believes that there will be one general factor that accounts for most of the variance, then Quartimax becomes the logical choice.

Quartimax (Carroll, 1953; Neuhaus & Wrigley, 1954; Saunders, 1960) is an orthogonal rotation which minimizes the number of factors needed to explain each variable. This type of rotation often generates a general factor on which most variables are loaded to a high or medium degree (Hair, 1995). The variables are much easier to interpret in this case, but the factors are more difficult to interpret since all variables are primarily associated with a single factor. Equimax rotation (Saunders, 1962) is a compound between Varimax and Quartimax methods. This method behaves somewhat erratically and should be used only when the number of factors has been clearly identified (Tabachnick & Fidell, 2001).

Direct oblimin (Jennrich & Sampson, 1966) is a non-orthogonal rotation, which results in higher eigenvalues but diminished interpretability of the factors. Promax (Hendrickson & White, 1964) is similar to direct oblimin but it is mostly used for very large datasets. The promax rotation has the advantage of being fast and conceptually simple, by comparison with Oblimin.

Orthogonal or non-orthogonal rotation?

The decision to rotate orthogonally or non-orthogonally is often difficult for researchers and is largely based on the goal of the analysis. If the goal of the analysis is to generate results that best fit the data, then oblique rotation seems to be the logical choice. Conversely, if the reliability of the factor analytic results is the primary focus of the analysis, then an orthogonal rotation might be preferable since results from orthogonal rotation tend to be more parsimonious. As already mentioned, in the orthogonal case the factors are uncorrelated, while in the oblique case the factors are correlated. In this perspective, the decision as to whether one should use an orthogonal or oblique rotation should be based upon the estimated inter-factor correlations from the oblique. If these are all close to 0, then the orthogonal is an appropriate procedure, but if any of them diverge from 0 then an non-orthogonal rotation should be used (Kieffer, Kevin M. 1998).

Exploratory Factors Analysis in Biomedical Research

Exploratory factor analysis (EFA) is the most commonly used method in health care research. EFA is used when the researcher does not know how many factors are necessary to explain the inter-behaviour among a set of medical characteristics. Therefore, the researcher uses the techniques of factor analysis to explore the underlying dimensions of the construct of interest. Moreover, examining single items in biomedical research studies makes the estimation of the effect size measure inaccurate and problematic, mainly due to the multicollinearity effect that was observed because of the high level of the correlations between them; therefore, FA for pattern recognition seems to be the "solution" to this problem. In particular, FA uses in order to examine the relationship between diet and risk of chronic diseases. Therefore, instead of looking at individual nutrients or foods consumed and their relationships with disease outcomes, FA evaluates and examines the effects of overall diet on human health. Moreover, FA is often used to explore the dimensionality of constructs in psychiatry (McKay, 1995, Jacob, 1998) and education (Hoban, 2005).

Application of Factor Analysis using R; using the dietary information from the ATTICA epidemiological study

It would be useful to provide an example of exploratory factor analysis an open-source software system, called R. Briefly, R began as a research project of Ross Ihaka and Robert Gentleman in the 1990s, described in a paper in 1996. It has since expanded into software used to implement and communicate most new statistical techniques. The

software in R implements a version of the S language, which was designed much earlier by a group at Bell Laboratories (John M. Chambers, 1998).

In the following example is used a dataset from the ATTICA Study (Pitsavos, 2003). The "ATTICA" study is a health and nutrition survey, which started collecting data from people in Greece, during 2001-2002, in the greater Athens area. The dietary evaluation was based on a validated semi-quantitative food-frequency questionnaire, which was kindly provided by the Unit of Nutrition of Athens University Medical School (Katsouyanni, 1997). The frequency of consumption was then quantified approximately in terms of the number of times per month a food was consumed. From the entire database 18 foods and food groups of 3042 responses were selected mainly according to their macronutrient composition, i.e., dairy products, fruits, vegetables, legumes, cereals, etc. Factor analysis, using the maximum likelihood method, was applied in order to extract dietary patterns based on foods.

At first, our datasets have been read into R and stored their contents in variables.

```
> my.data <- read.csv("ATTICA_EFA.csv", header=TRUE)
# if data as NAs, it is better to omit them:
# my.data <- na.omit(my.data)
> head(my.data) # to see the first several rows of the data frame and confirm that the data has been stored correctly
```

Next, after installing the psych package by William Revelle (2016), importing that package to the current namespace by calling library() as follows:

```
> install.packages("psych")
> library(psych)
```

Then will be found out the number of factors that selecting for factor analysis. This can be evaluated via methods such as Parallel Analysis and eigenvalue, etc. In our case, a Parallel Analysis function was executed using Psych package's fa.parallel. The following code was run, to find acceptable number of factors and generate the scree plot (Figure 2); the blue line shows eigenvalues of actual data and the two red lines (placed on top of each other) show simulated and resampled data. Simulated data were generated using R software to have characteristics (i.e., mean and variance) like those of the original data, but with no correlations among the observed varioables (i.e., no underlying structure). Resampled data where generated from the original sample using a resampling method. Parallel analysis is a method for determining the number of components or factors to retain from PCA or factor analysis. Parallel analysis is an alternative technique that compares the scree plots of factors of the observed data, with that of a random data matrix of the same size as the original. The correlation matrix is computed from the randomly generated dataset and, then, eigenvalues of the correlation matrix are calculated. If the eigenvalues from the random data are larger than the eigenvalues from the original factor analysis, the extracted factors from the original data are mostly random noise.Also the point of inflection is located – i.e., the point where the gap between simulated data and actual data tends to be the lowest. Looking at the scree plot and parallel analysis, anywhere between 4 to 6 factors should be retained. Thys, the analysis was based on a 5 factors solution. In order to perform factor analysis, psych package's fa() function was used.

```
> model <- fa(foods, nfactors = 5,rotate = "varimax",fm="ml")
```

Then were loadings greater than 0.2 where considered in order to characterize each factor (the threshold used may vary from analysis to analysis, giving to the researcher the

opportunity to better interpret the extracted factors). So let's first establish the cut off to improve visibility in the extracted factors solution:

> print(model,digits=,cut=0.2, sort = TRUE)

Taking into consideration that higher absolute values are indicative of the food's greater contribution to the development of the factor (dietary pattern), the 5 factors (patterns) extracted and heavily loaded with the following food groups or foods (Table 1):

- Factor 1: total meat and red meat
- Factor 2: vegetables, fruits, cereals, legumes, dairy, fish and eggs
- Factor 3: butter, other added fat, alcohol, seed oil and olive Oil
- Factor 4: sweets, soft drinks and potatoes
- Factor 5: poultry

	Factor1	Factor2	Factor3	Factor4	Factor5
SS loadings	1.83	1.75	1.71	1.12	1.06
Proportion Variance	0.10	0.10	0.10	0.06	0.06
Cumulative Variance	0.10	0.20	0.30	0.36	0.42

Factor analysis explained the 42% of the total variation in intake, which is considered a good explanatory ability in nutrition epidemiology. To visualized the factors extracted a special command in R was used, by calling the function fa.diagram(model). (Figure 3) The square boxes are the observed variables, and the ovals are the unobserved factors. The straight arrows are the loadings, the correlation between the factor and the observed variables. The curved arrows are the correlations between the factors. If no curved arrow was present, then the correlation between the factors was not great.

Concluding remarks

Factor analysis is an "old" multivariate statistical procedure for data analysis that has many uses nowadays in various fields of behavioural sciences, like sociology, psychology, molecular biology and genetics, medicine and nutrition, as well as other disciplines, that it is often not possible to measure directly the concepts of primary interest. This method of analysis most generally used to uncover the "hidden" relationships between the assumed latent variables and the initial observed variables. By understanding the concepts and the procedures of factor analysis, readers can better evaluate the reported results.

REFERENCES

[1] Panaretos D., Tzavelas G., Vamvakari M., Panagiotakos D. (2017). Repeatability of dietary patterns extracted through multivariate statistical methods: a literature review in methodological issues. Int J Food Sci Nutr., 68(4), 385-391.

[2] Hotteling H. (1933). Analysis of a complex of statistical variables into principal components. J Educ Psychol. 24, 417–441;498–520.

[3] Gorsuch, Richard L., (1983). Factor Analysis, second edition, Hillsdale: Lawrence Erlbaum Associates

[4] Spearman C. (1904). General intelligence objectively determined and measured. Am J Psychol. 15, 201–293.

[5] Spearman C. (1927). The Abilities of Man: Their Nature and Measurement. New York

[6] Thurstone, L. L. (1938). Primary mental abilities. Chicago: University of Chicago Press

[7] Thurstone, L. L. (1947). Multiple-Factor Analysis. Chicago: University of Chicago Press

[8] Carl Eckart and Gale Young, (1936), The approximation of one matrix by another of lower rank, Psychometrika, 1(3),211-218

[9] Karl J. HolzingerFrances Swineford, 1937, The Bi-factor method, Psychometrika, 2(1),41–54

[10] Thomson, G.H., (1951). The factorial analysis of human ability. University of London Press, London.

[11] Lawley, D.N., & Maxwell, A.E., (1971). Factor Analysis as a statistical method. London: Butterworth.

[12] Joreskog, K.G, (1969). A general approach to confirmatory maximum likelihood factor analysis, Psychometrica, 34, 183-202

[13] Joreskog, K.G. (1972). Factor Analysis by generalized least squares. Psychometrica, 37, 243-250

[14] Wayne F. Velicer. (1976). Determining the number of components from the matrix of partial correlations, Psychometrika, 41(3), 321–327

[15] Pett MA, Lackey NR, Sullivan JJ. (2003). Making Sense of Factor Analysis: The use of factor analysis for instrument development in health care research. California: Sage Publications Inc.

[16] Pearson, Karl (1901). "On Lines and Planes of Closest Fit to Systems of Points in Space," Philosophical Magazine, 6, 559–572

[17] Thurstone, L. L. (1934). The vectors of the mind. Address of the president before the American Psychological Association, Chicago meeting, September, 1933. First published in Psychological Review, 41, 1-32.

[18] Joost C.F. de Winter, Dimitra Dodou. (2015). Common factor analysis versus principal component analysis: a comparison of loadings by means of simulation, Communications in Statistics - Simulation and Computation, 45, 299-321

[19] Alexander Basilevsky. (2008). Statistical Factor Analysis and Related Methods: Theory and Applications, Wiley Series in Probability and Statistics

[20] Jae On Kim, Charles W, Mueller. (1978). Introduction to factor analysis, Sage University Paper

[21] J.H.F. Schilderinck, (1977). Regression and factor analysis applied in econometrics, Springer-Verlag New York Inc

[22] Harry H. Harman. (1976). Modern Factor Analysis, University of Chicago Press.

[23] Thurstone, L.L. (1947) Multiple factor analysis: A development and expansion of vectors of the mind. University of Chicago Press, Chicago, xix 535.

[24] Cattel,R.B. (1978). The scientific use of factor analysis. New York: Plenum.

[25] Kevin M. Kieffer. (1998). Orthogonal Versus Oblique Factor Rotation: A Review of the Literature Regarding the Pros and Cons, Annual Meeting of the Mid-South Educational Research Association, 4-6.

[26] Harris, C. W., & Kaiser, H. F. (1964). Oblique factor analytic solutions by orthogonal transformations, Psychometrika, 29, 347–362.

[27] H.F. Kaiser (1958): "The varimax criterion for analytic rotation in factor analysis." Psychometrika, 23, 187–200.

[28] Tabachnick, B. G., & Fidell, L. S. (2001). Using Multivariate Statistics. Boston: Allyn and Bacon.

[29] Carroll, J. B. (1953). An analytical solution for approximating simple structure in factor analysis, Psychometrika, 18(1), 23-38.

[30] Neuhaus, J. O. & Wrigley, C. (1954). The quartimax method: an analytical approach to orthogonal simple structure. British Journal of Statistical Psychology, 7, 187-191.

[31] Saunders D.R. (1960). A computer program to find the best – fitting orthogonal factors for a given hypothesis, Psychometrica. 25, 199-205

[32] Hair J, Anderson RE, Tatham RL, Black WC. (1995). Multivariate data analysis. 4th ed. New Jersey: Prentice-Hall Inc.

[33] Saunders, D, R. (1962) Trans Varimax: Some properties of the ratiomax and Equimax criteria for blind orthogonal rotation. Paper presented at the meeting of the American Psychological Association, St. Louis.

[34] Jennrich RI, Sampson PF. (1966). Rotation for simple loadings, Psychometrika, 31(3), 313-23.

[35] Hendrickson & White. (1964). Promax: A Quick Method for Rotation to Oblique Simple Structure, British Journal of Mathematical and Statistical Psychology, 17(1), 65-70

[36] McKay D, Danyko S, Neziroglu F, Yaryuratobias JA. (1995). Factor structure of the Yale-Brown obsessive-compulsive scale – A 2-dimensional measure. Behaviour Research and Therapy, 33, 865–69

[37] Jacob KS, Everitt BS, Patel V, Weich S, Araya R, Lewis GH. (1998). The comparison of latent variable models of nonpsychotic psychiatric morbidity in four culturally diverse populations. Psychological Medicine, 28: 145–52.

[38] Hoban JD, Lawson SR, Mazmanian PE, Best AM, Seibel HR. (2005). The Self-Directed Learning Readiness Scale: a factor analysis study. Med Educ. 39(4), 370-9

[39] Ross Ihaka and Robert Gentleman. (1996). R: A language for data analysis and graphics. Journal of Computational and Graphical Statistics, 5, 299–314.

[40] R. A. Becker, J. M. Chambers, and A. R. Wilks. (1988). The New S Language. Chapman & Hall, Boca Raton, FL.

[41] John M. Chambers. (1998). Programming with Data: A Guide to the S Language. Springer, New York

[42] Pitsavos C, Panagiotakos DB, Chrysohoou C, Stefanadis C. (2003). Epidemiology of cardiovascular risk factors in Greece: aims, design and baseline characteristics of the ATTICA study. BMC public health, 3, 32

[43] Katsouyanni K, Rimm EB, Gnardellis C, Trichopoulos D, Polychronopoulos E, Trichopoulou A. (1997). Reproducibility and relative validity of an extensive semi-quantitative food frequency questionnaire using dietary records and biochemical markers among Greek schoolteachers. Int J Epidemiol, 26, 118-127

[44] Revelle, W. (2016). psych: Procedures for Personality and Psychological Research. R package version 1.6.12

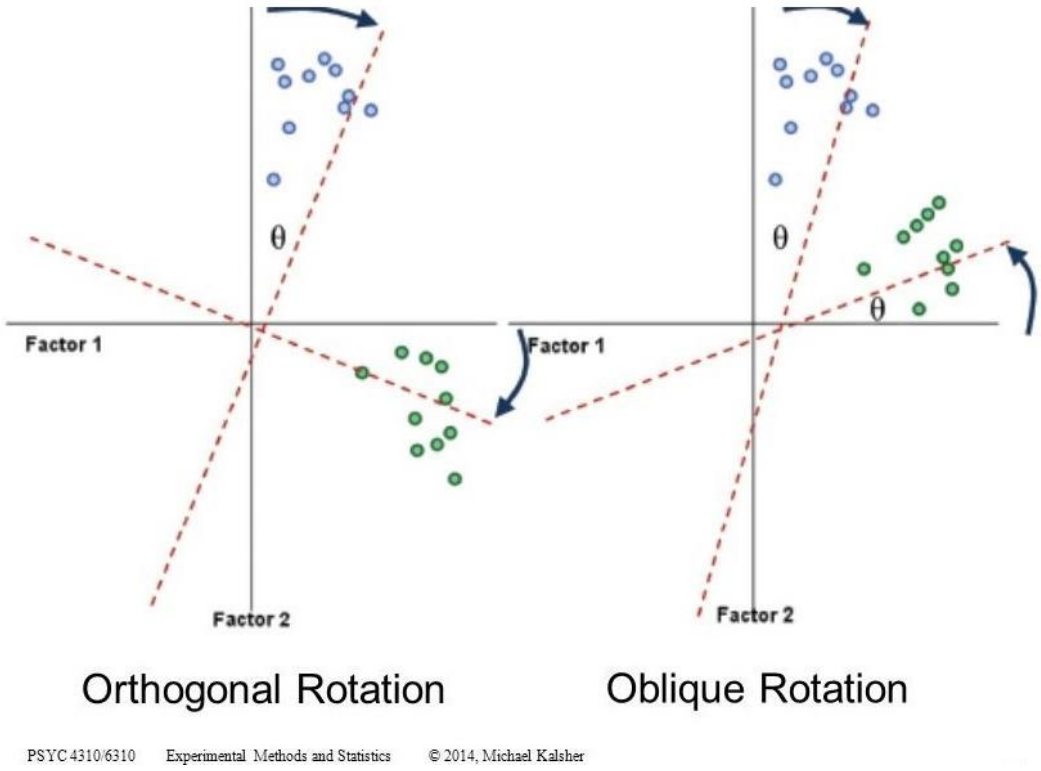

Figure 1: Schematic representations of factor rotation. The left graph displays orthogonal rotation whereas the right graph displays oblique rotation. θ is the angle through which the axes are rotated.

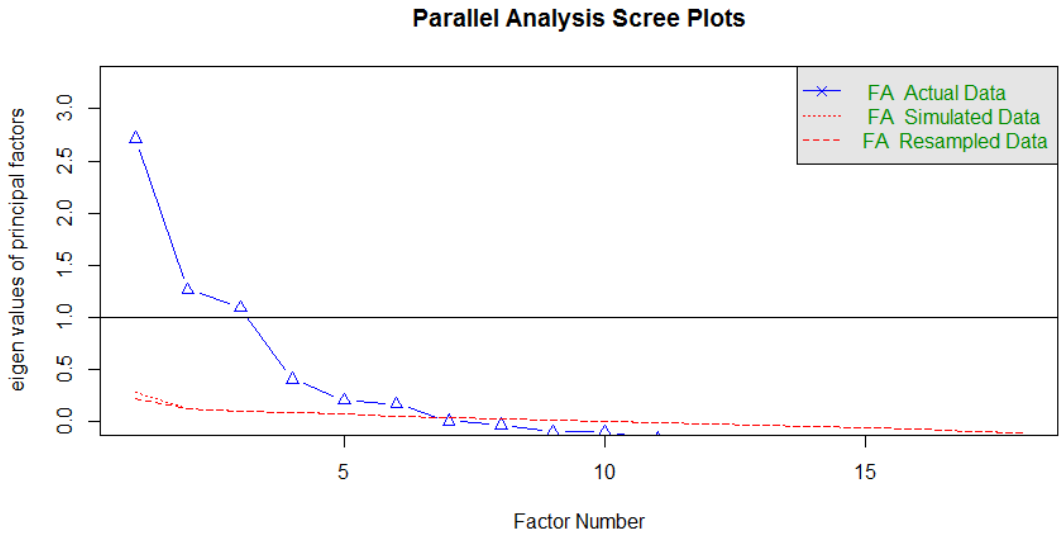

Figure 2. Parallel Analysis Scree Plot for determining the number of factors in Exploratory Factor Analysis, using the ATTICA study dietary database.

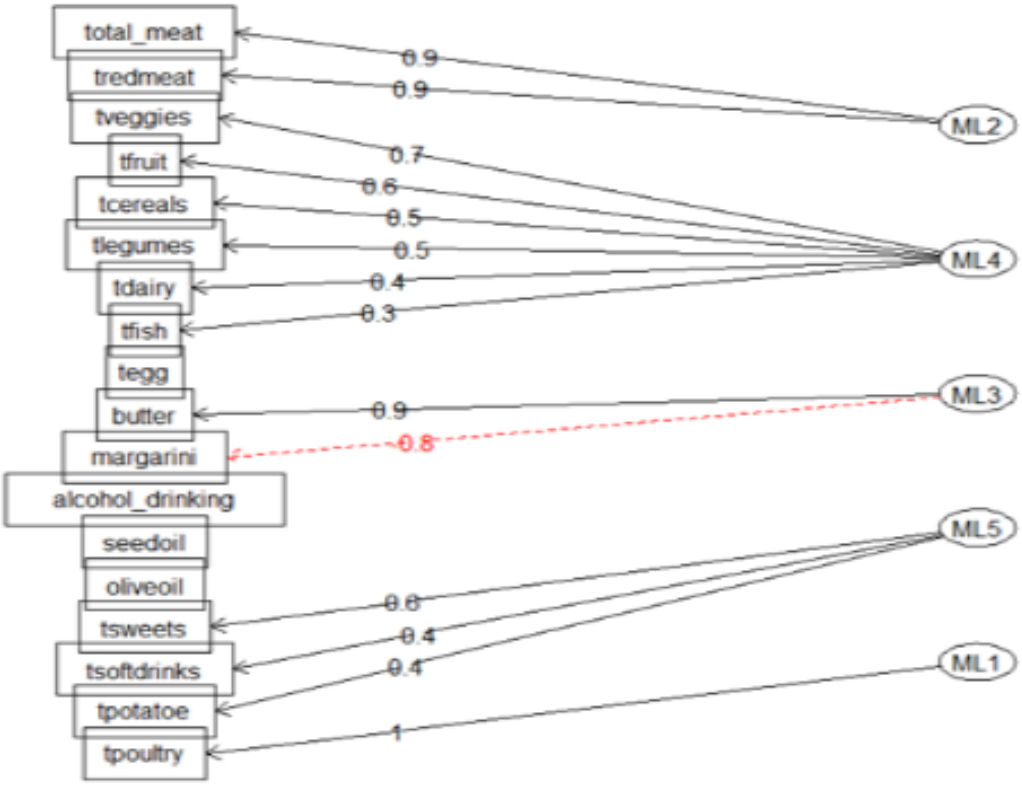

Figure 3. An example of factor analysis using graphic representation of the 5 factors extracted from the ATTICA Study dietary database (using *fa.diagram* command). Factors were transformed to an orthogonal solution using *varimax* rotation.

Table 1. Factor coefficients loadings regarding foods or food groups consumed by Greek ATTICA study participants (n=3042) at baseline

	Loadings				
Food / Food group	*Factor 1*	*Factor 2*	*Factor 3*	*Factor 4*	*Factor 5*
Total meat	0.90				
Red meat	0.89				
Vegetables		0.71			
Fruits		0.61			
Cereals		0.49			
Legumes (lentils, beans, etc)		0.48			
Dairy (milk, yogurt)		0.39			
Fish		0.32			
Eggs		0.23			
Butter			0.93		
Other added fat			-0.81		
Alcohol			0.28		
Seed oil			-0.22		
Olive Oil			0.22		
Sweets				0.61	
Soft drinks				0.44	
Potatoes				0.43	
Poultry					0.97