DOI: 10.1590/1980-549720190041

ORIGINAL ARTICLE / *ARTIGO ORIGINAL*

Principal Component Analysis and Factor Analysis: differences and similarities in Nutritional Epidemiology application

Análise de Componentes Principais e Análise Fatorial: diferenças e similaridades na aplicação em Epidemiologia Nutricional

Roberta de Oliveira Santos[I], Bartira Mendes Gorgulho[I], Michelle Alessandra de Castro[I], Regina Mara Fisberg[I], Dirce Maria Marchioni[I], Valéria Troncoso Baltar[II]

ABSTRACT: ***Introduction:*** Statistical methods such as Principal Component Analysis (PCA) and Factor Analysis (FA) are increasingly popular in Nutritional Epidemiology studies. However, misunderstandings regarding the choice and application of these methods have been observed. ***Objectives:*** This study aims to compare and present the main differences and similarities between FA and PCA, focusing on their applicability to nutritional studies. ***Methods:*** PCA and FA were applied on a matrix of 34 variables expressing the mean food intake of 1,102 individuals from a population-based study. ***Results:*** Two factors were extracted and, together, they explained 57.66% of the common variance of food group variables, while five components were extracted, explaining 26.25% of the total variance of food group variables. Among the main differences of these two methods are: normality assumption, matrices of variance-covariance / correlation and its explained variance, factorial scores, and associated error. The similarities are: both analyses are used for data reduction, the sample size usually needs to be big, correlated data, and they are based on matrices of variance-covariance. ***Conclusion:*** PCA and FA should not be treated as equal statistical methods, given that the theoretical rationale and assumptions for using these methods as well as the interpretation of results are different.

Keywords: Diet. Food. Eating. Nutritional epidemiology.

[I]Department of Nutrition, School of Public Health, Universidade de São Paulo – São Paulo (SP), Brazil.
[II]Department of Epidemiology and Biostatistics, Collective Health Institute, Universidade Federal Fluminense – Rio de Janeiro (RJ), Brazil.
Corresponding author: Roberta de Oliveira Santos. Departamento de Nutrição, Faculdade de Saúde Pública, Universidade de São Paulo. Avenida Dr. Arnaldo, 715, CEP: 01246-904, São Paulo, SP, Brazil. E-mail: oliveira.ros@usp.br
Conflict of interests: nothing to declare – **Financial support:** none.

RESUMO: ***Introdução:*** Métodos estatísticos de análise multivariada, tais como Análise de Componentes Principais e Análise Fatorial, têm sido cada vez mais utilizados nos estudos em Epidemiologia Nutricional, no entanto equívocos quanto à escolha e aplicação dos métodos são observados. ***Objetivos:*** Os objetivos deste estudo são comparar e apresentar as principais diferenças e similaridades conceituais e metodológicas entre Análise de Componentes Principais e Análise Fatorial visando à aplicabilidade nos estudos em nutrição. ***Métodos:*** Análise de Componentes Principais e Análise Fatorial foram aplicadas em uma matriz de 34 grupos de alimentos que expressaram o consumo alimentar médio de 1.102 indivíduos de um estudo populacional. ***Resultados:*** Um total de dois fatores foi extraído e juntos explicaram 57,66% da variância comum entre as variáveis dos grupos alimentares, enquanto um total de cinco componentes foi extraído e juntos explicaram 26,25% da variância total. Entre as principais diferenças envolvendo os dois métodos estão: pressuposto de normalidade; as matrizes de variância-covariância/correlação, com consequente quantidade de variância explicada; a carga fatorial/componente e o erro associado. Entre as similaridades estão: ambas as técnicas são usadas para redução de dados; necessitam de um grande tamanho de amostra; os dados precisam ser correlacionados e são baseadas nas matrizes de variância-covariância/correlação. ***Conclusão:*** Análise de Componentes Principais e Análise Fatorial não devem ser tratadas como métodos estatísticos iguais e intercambiáveis, uma vez que o racional teórico e os pressupostos para o uso dos métodos, assim como a interpretação dos resultados, são diferentes.

Palavras-chave: Dieta. Alimentos. Ingestão de alimentos. Epidemiologia nutricional.

INTRODUCTION

Principal Component Analysis (PCA) and Factor Analysis (FA) are multivariate statistical methods that analyze several variables to reduce a large dimension of data to a relatively smaller number of dimensions, components, or latent factors[1]. These statistical methods are widely applied in nutritional epidemiology to study food combination[2], such as dietary pattern analysis[3]. Despite their widespread utilization, many researchers do not know the assumptions and conceptual differences between PCA and FA, which leads to a misuse of the methods, impairing the interpretation and validity of results.

The selection of PCA or FA should be based on the objective of the research. Both methods are used for data reduction, but PCA aims to describe a large data set in a simpler dimension, preferably a plan. In this case, PCA is used mainly to show graphically the relationships among the variables in some reduced dimension graphs. On the other hand, FA is a statistical model used to build dietary patterns (factors), which are latent variables to predict food choices[4]. PCA is a mathematical procedure that enables the researcher to reduce the number of correlated variables into a smaller number of components (linear combination of such variables), linearly independent of each other, which represents a percentage of the total covariance[1,5]. There is no assumption of normality at this stage. In contrast, FA aims at modeling each original variable through latent factors and random errors, in a way that reduces the number of factors, and, depending

on the extraction method, the assumption of normality becomes necessary[1]. One of the possible estimation methods used in FA is the principal components, hence the confusion between these methods[1].

One of the main differences between PCA and FA in mathematical terms is the values found in the diagonal of the correlation matrix[1,5-7], the basis of both methods. The total variance of each variable is a result of the sum of the shared variance with another variable, the common variance (communality), and the unique variance inherent to each variable (specific variance)[8]. In PCA, all variance is taken into account in the calculations. Consequently, the diagonal of the correlation matrix is 1.00 (sum of the unique variance of each variable, common variance among variables, and error variance) and includes all variance of the variables[1,5,9]. In turn, FA uses only common variance[8]; therefore, the diagonal of the correlation matrix includes only communalities, that is, only the variance shared with other variables will be considered (excluding the unique variance of each variable and error variance)[1,5,9].

PCA is conceptually simpler than FA since it summarizes or aggregates sets of correlated variables and, in that sense, is relatively empirical, being a method of exploratory descriptive analysis[1,6,10]. On the other hand, FA is a more complex method in the sense that factors reflect the causes of observed variables, thereby this analysis assumes a characteristic of the multivariate model by calculating factor loadings and errors assigned to each factor[6,10].

In this regard, the objective of this article was to compare and show the differences and similarities between PCA and FA, presenting an example based on actual data.

METHODS

STUDY POPULATION AND DATA MANAGEMENT

We illustrated the application of PCA and FA in the nutrition field by using both multivariate methods on a matrix of 34 variables expressing the mean food intake (in grams/day) of 1,102 individuals (aged 20 years and older) who responded to two non-consecutive 24-hour dietary recalls (24HDR) in a population-based study[11]. The study had two different objectives: to describe only the multidimensional data in PCA and the derivation of dietary patterns in FA. Castro et al.[11] present a detailed description of the 34 food groups and their composition.

The procedures to group the foods were the same applied by Castro et al.[11]. In brief, a total of 948 different foods consumed on dietary assessment days dropped to 38 food groups, following the criteria:

- similarity in nutrient profile, that is, combining variations of the same food with similar nutrient profile in the same group (*e.g.*, different types of coffee);

- regional dietary habits and culinary usage of foods by the Southeastern Brazilian population.

Next, we analyzed a correlation matrix of the variables to investigate how food groups correlate to each other. Since four food groups did not correlate significantly ($p > 0.05$) with any other food group, they were excluded from analysis, resulting in 34 food groups for FA and PCA.

In food groups with zero augmented distribution, it would be better to treat the data before starting data reduction. Statistical methods to estimate usual intake can be applied to deal with intra-individual variation and zero augmented distribution[12,13]. Another option is the direct analysis of the correlation matrix, using alternative correlation instead of the usual Pearson correlation. After the analysis, the researcher can compare its results to those from the usual analysis and verify if there were relevant differences.

STATISTICAL ANALYSIS

Before using any statistical method, as a first step, the researcher must have a very clear objective. After deciding between possible statistical methods, it is important to verify its assumptions; with FA and PCA, it is not different. First, the sample size needs to be big enough regarding the number of variables that will be analyzed. There is no sample size calculation, and this number is arbitrary, but generally, at least 50 individuals are recommended. Also, the sample size should be at least five times greater than the number of variables, with an ideal proportion of 10 or more individuals for each analyzed variable[5]. In this study, the proportion of individuals to variables considered in the illustrative example was approximately 32:1.

Second, both analyses are based on the covariance / correlation matrix, so assessing sample adequacy according to the multiple correlations of the variables is recommended. It is noteworthy that variables included in both analyses need to be correlated, and if these correlations are low, it is better to have a bigger sample size. Significant correlations of the set of variables indicate sample adequacy for FA or PCA, but looking at correlation magnitudes is always advisable. In FA, sample adequacy should be assessed, and two tests can be applied: the Kaiser-Meyer-Olkin (KMO) test and Bartlett's sphericity test. KMO statistic is a proportion of variance among variables that might be common variance: varies from zero to one, in which zero is inadequate, while close to one is adequate[14]. Bartlett's test compares the observed correlation matrix to the identity matrix (off-diagonal is zero). If they are similar, it will be necessary as many factors as variables, and the analysis is useless[4]. Overall, KMO values above 0.50 and $p < 0.05$ for Bartlett's sphericity test are considered acceptable[5]. Also, FA requires an extra assumption: input variables do not need to present multivariate normal distribution, but normality is assumed for unique factors (regression errors). There is no statistical test to check it properly, but it is recommended to plot histograms or Q-Q plots

of all variables to confirm if they are close to normally distributed and to verify the presence of outliers[15]. Once assumptions were reached, FA and PCA can be applied following the steps in Figure 1.

In the second step of FA, it is necessary to choose one of the several extraction methods available. Principal components, principal factor, and maximum likelihood factor are among the most popular in nutritional epidemiology[1]. The decision about which method to use should combine the objectives of FA with the knowledge about some basic characteristics of the relations between variables[2].

The extraction method of FA used in this study was the principal factor (PF), a default method for some statistical software, such as Stata®, commonly used in health sciences.

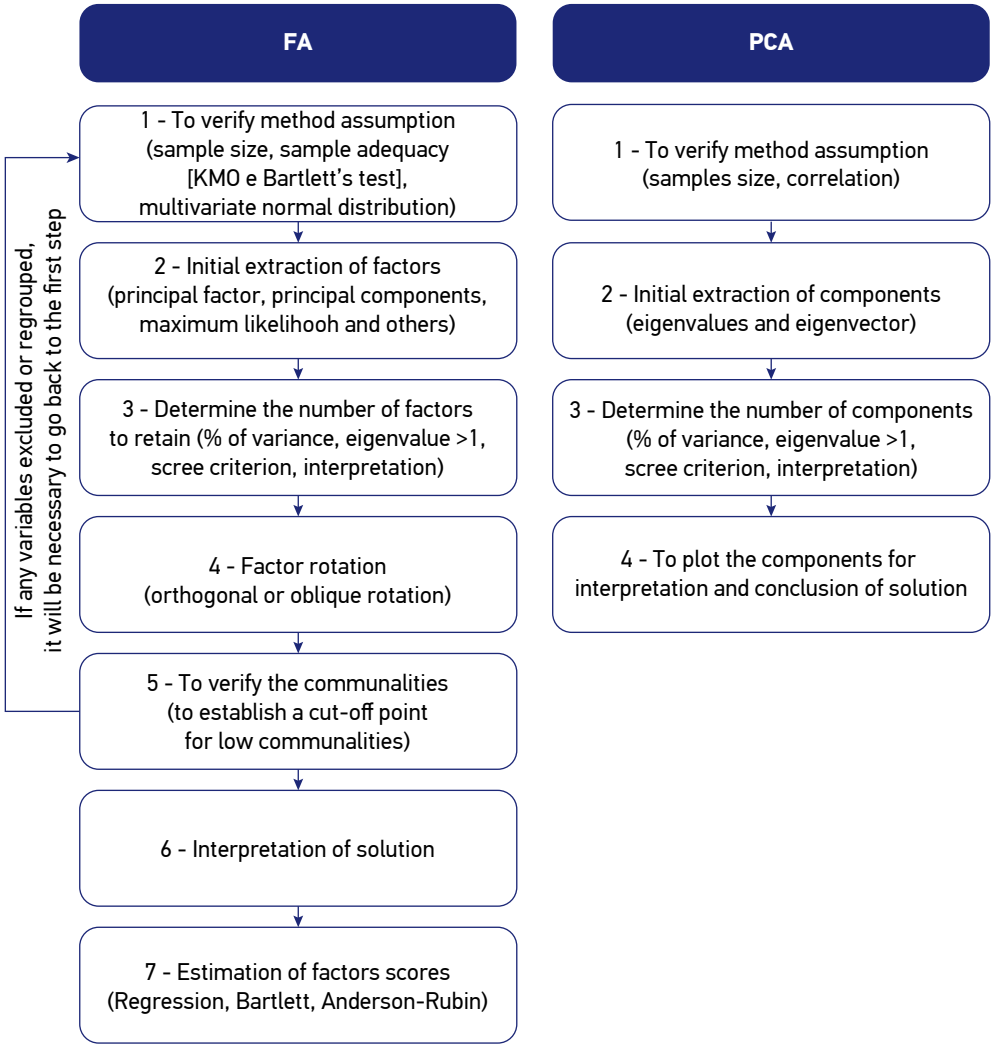

Figure 1. Step-by-step of Factor Analysis and Principal Component Analysis.

This method considers the variance of each observed variable explained by the factor (*i.e.*, communality) to compute factor loadings[1,16]. On the other hand, in the second step of PCA, matrix decomposition is automatic in an exploratory way[5], so there is no need to choose an extraction method.

The third step of PCA and FA settles the number of factors to extract; firstly, the Kaiser criterion was applied[14]. This criterion is based on the rationale that the minimum variance explained by the factor should be equal to or greater than the variance of one single observed variable[17]. Cattell's scree test[5], *i.e.*, a plot of the proportion of variance explained by each component/factor (eigenvalues), was visually inspected to identify breakpoints in the curve trajectory (inflection point) and check the distance between points. The greater the distance between points, the larger the increase in variance explained with the inclusion of the component/factor. Cattell's scree test is useful when deciding on the number of components/factors to extract if a large number of components/factors shows eigenvalues greater than 1.0. The same steps mentioned above were applied to determine the number of components and factors and allow for comparisons. Figure 2 presents the Cattell's scree test for FA and PCA.

The fourth step in PCA is plotting the components for interpretation and concluding the solution. At this point in PCA, it is possible to interpret components or the correlation between components and variables (easily calculated by multiplying component values by the square root of the eigenvalues). Some statistical software plots the graphs with correlations for interpretation. These graphs have two dimensions/plans for interpretation, with vectors corresponding to each food item, and its size shows how well represented they are in such plan. Also, the angle between vectors indicates how correlated these food groups are. If the angle between two food items is small, they have a high positive correlation, if close to 90º, they are not correlated, and if between 90º and 180º, they are negatively correlated. For simplicity, this article will present only the first plan (components 1 and 2), but in conventional analysis, all combinations of the selected components should be plotted.

The fourth step of FA is factor rotation. The orthogonal Varimax rotation was applied to the subset of factors extracted, aiming to estimate uncorrelated factors with a simpler loading matrix, which was considered easier to interpret[14,18]. A simple loading matrix is estimated when the variable loads highly on as few factors as possible, and loadings of the variables across the factors (cross-loadings) are approximately zero[19,20]. The idea of factor rotation is based on the objective of the analysis used to build factors, latent variables representing patterns that predict the intake of food groups. In that sense, the PCA rotation is not appropriate because it is not part of its objective. Factor rotation should be done only to estimate factor when the assumptions for inference were verified.

After identifying factor loadings, as a fifth step, the researcher should look for variables not adequately explained by the factors[5]. Thus, the interpretation of FA must also consider communalities, as estimated communalities represent how much a variable has in common with the remaining variables in the analysis[1,5,21]. If a variable has a high correlation with one

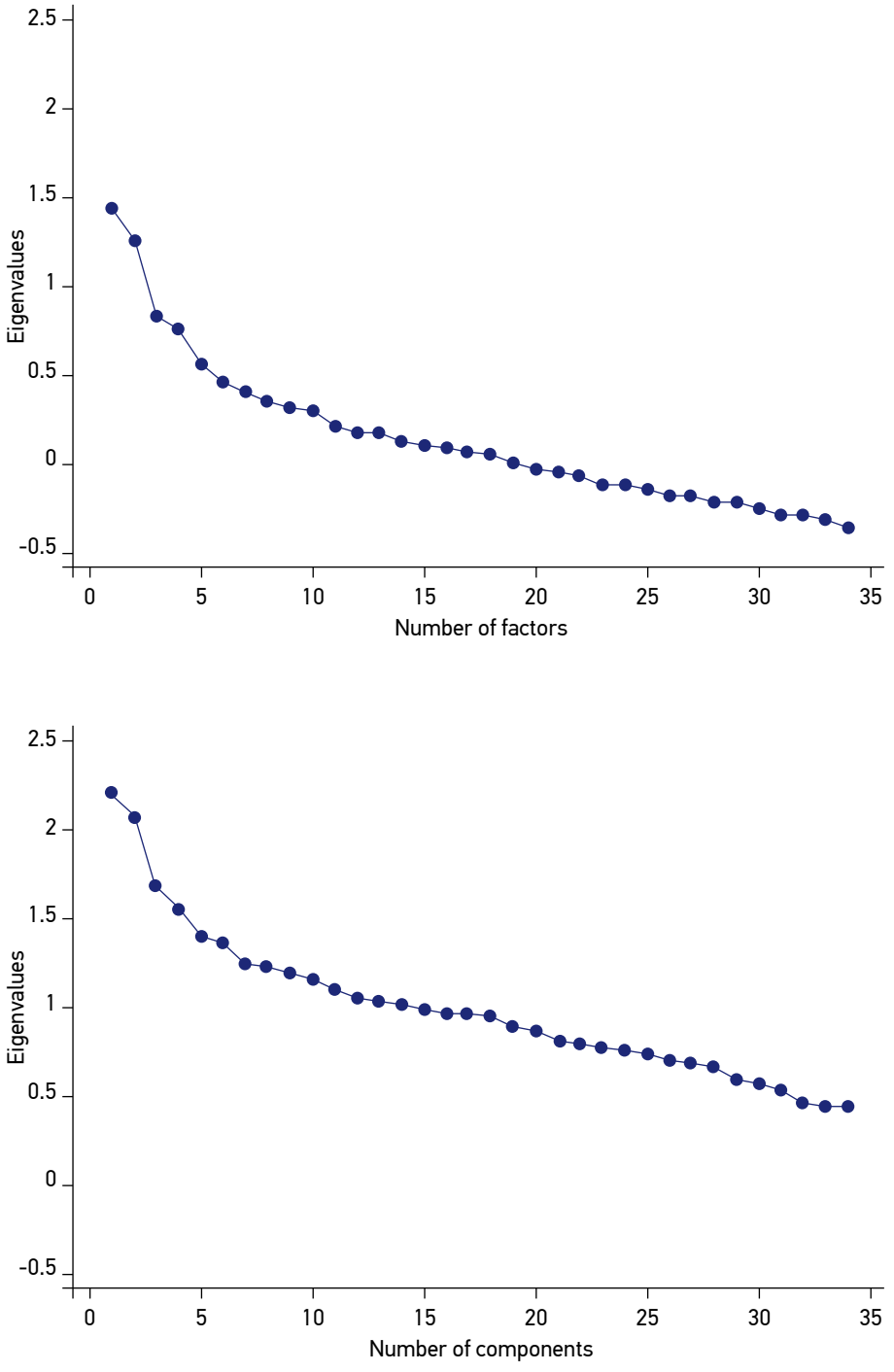

Figure 2. Scree plot of the eigenvalues of factors and components.

or more variables, the communality increases[5], and the set of factors will explain much of the variable variance[22]. Considering that FA seeks to explain variance through common factors, authors usually exclude variables with low communalities and go back to the first step[5,21]. The cut-off point for communality is arbitrary, and each author makes his or her own decision based on the desired explanation level. In the nutrition field, some authors used cut-off values equal to or greater than 0.10[23] and 0.25[24], that is, they considered acceptable variables that explained at least 10 and 25% of variance; however, most articles do not mention it. In this study, we decided to present all communalities.

The sixth step of FA is the interpretability of factors, investigated considering that food groups with positive loadings can be interpreted as being directly correlated to the factor, while food groups with negative loadings can be interpreted as being inversely correlated to the factor.

As a way to facilitate interpretation, authors usually use cut-off points in rotated factor loadings to find factor names[5]. For instance, nutritional epidemiology commonly adopts the cut-off of $|0.30|$, *i.e.*, variables with loadings lower than this cut-off are not considered when creating the name of the factor. In this application, we used a cut-off of $|0.30|$. Nonetheless, we emphasize that all variables/items were included for score calculation, as a way to help to provide some interpretation.

The seventh and last step of FA is the estimation of factor scores. This step is non-compulsory, but it can be useful for the subsequent analysis, given that researchers intend to identify an individual's placement or ranking on the factor; in the nutrition field, the factor could be translated into intake patterns[25].

We performed all analyses using the Stata® software, version 12, and SAS software, version 9.3. The Research Ethics Committee of the School of Public Health at Universidade de São Paulo and the Municipal Secretariat of Health approved the main study.

RESULTS

Table 1 shows the illustrative example of the application of both techniques to the same dietary data. Comparing results from both methods, the number of factors extracted (FA) was, as expected, lower than the number of extracted components (PCA). Two factors were extracted and, together, they explained 57.7% of the *common variance* of food group variables, while five components were extracted, explaining 26.3% of the *total variance* of food group variables. Figure 2 demonstrates that only two factors met the Kaiser criterion (eigenvalues > 1.0). In contrast, fourteen components satisfied the same criterion. However, while performing the visual inspection of the plot, a breakpoint in the curve trajectory of the fifth component was suggested to meet the Kaiser criterion.

Another difference between FA and PCA lies in the food group loadings. Most food groups showed larger loadings, in module, in FA than in PCA. Comparing the two factors with the first two components extracted, the highest loading in FA was 0.55 for

Table 1. Results from principal component analysis (PCA) and factor analysis (FA) before and after Varimax rotation, based on a 2-day mean food group intake, Health Survey of São Paulo, Brazil.

	Principal components (PC)					Correlation between PC and food group intakes					h^2 for 2 PCs	h^2 for 5 PCs	FA (without rotation)		FA (after Varimax rotation)		h^2 for 2 factors
	PC1	PC2	PC3	PC4	PC5	Dim1	Dim2	Dim3	Dim4	Dim5			F1	F2	F1	F2	
Rice	.41	-.14	-.03	-.32	-.04	.61	-.20	-.04	-.40	-.05	.42	.58	.55	-.10	.54	.13	.31
Pasta	.00	.04	.25	.24	-.04	.00	.05	.33	.30	-.05	.00	.20	-.01	.04	-.03	.03	.00
Bread/Toasts/Crackers	.29	-.24	-.06	.29	.18	.43	-.35	-.07	.37	.21	.30	.49	.38	-.25	.45	-.07	.21
Whole bread	-.03	.33	-.09	.07	.08	-.05	.47	-.11	.08	.09	.22	.25	-.07	.35	-.21	.29	.13
Fruits	-.04	.27	-.15	-.01	.02	-.06	.39	-.19	-.01	.02	.16	.19	-.07	.27	-.18	.22	.08
Canned vegetables	.22	.37	-.03	-.08	.04	.32	.53	-.04	-.10	.04	.39	.40	.21	.46	.01	.50	.25
Leafy vegetables	.09	.11	.21	.09	-.09	.13	.16	.28	.11	-.10	.04	.14	.08	.13	.02	.15	.02
Non-leafy vegetables	.19	.37	-.07	.00	.01	.27	.53	-.09	.00	.02	.36	.37	.17	.44	-.03	.47	.22
Beef	.17	.04	.08	-.18	.01	.25	.06	.11	-.22	.01	.06	.12	.18	.07	.14	.14	.04
Pork	.06	-.02	.17	-.06	-.11	.09	-.03	.22	-.07	-.13	.01	.08	.06	-.01	.06	.02	.00
Processed meat	.21	.00	.19	-.06	-.18	.31	.00	.24	-.08	-.21	.09	.20	.22	.03	.19	.12	.05
Poultry	.11	.10	-.06	-.09	.08	.16	.14	-.08	-.11	.10	.05	.08	.11	.12	.05	.16	.03
Chocolate powder	.06	-.02	.06	.04	.58	.09	-.03	.08	.05	.68	.01	.48	.06	-.01	.06	.01	.00
Yellow cheese	.11	.03	.14	.25	-.05	.17	.04	.19	.31	-.06	.03	.17	.11	.04	.08	.08	.01
White cheese	-.03	.23	-.08	.11	.10	-.04	.33	-.10	.14	.12	.11	.15	-.06	.22	-.14	.18	.05
Whole milk	.13	-.12	-.15	.04	.48	.19	-.17	-.20	.04	.57	.07	.43	.16	-.11	.19	-.04	.04
Low-fat and skim milk	-.11	.23	-.06	.12	-.08	-.16	.33	-.08	.15	-.09	.13	.17	-.14	.22	-.22	.14	.07
Other dairy products	.11	.13	-.08	.09	.02	.16	.18	-.11	.11	.02	.06	.08	.10	.14	.03	.17	.03

Continue...

Table 1. Continuation.

	Principal components (PC)					Correlation between PC and food group intakes					h^2 for 2 PCs	h^2 for 5 PCs	FA (without rotation)		FA (after Varimax rotation)		h^2 for 2 factors
	PC1	PC2	PC3	PC4	PC5	Dim1	Dim2	Dim3	Dim4	Dim5			F1	F2	F1	F2	
Eggs	.20	-.02	-.01	-.02	-.02	.30	-.03	-.01	-.03	-.03	.09	.09	.22	.01	.19	.10	.05
Pulses	.00	.05	-.11	.17	-.04	-.01	.08	-.14	.21	-.05	.01	.07	-.01	.05	-.03	.04	.00
Beans	.34	-.18	-.04	-.36	-.09	.51	-.25	-.05	-.45	-.10	.32	.54	.45	-.15	.48	.05	.23
Butter/Margarine	.27	-.15	-.13	.24	.12	.40	-.21	-.17	.30	.14	.21	.35	.34	-.14	.37	.01	.14
Cakes/Confectionery products	.06	.04	.12	.23	-.04	.09	.06	.16	.29	-.05	.01	.12	.05	.05	.03	.06	.00
Salty snacks	.06	.03	.04	.08	.17	.09	.05	.05	.10	.20	.01	.06	.05	.04	.03	.06	.00
Sandwiches	.00	-.05	.38	.11	-.07	.01	-.07	.50	.14	-.08	.00	.28	.00	-.05	.02	-.04	.00
Coffee/Tea	.16	-.12	-.28	.32	-.34	.23	-.17	-.36	.41	-.41	.08	.54	.21	-.13	.24	-.03	.06
Soft drinks	.14	-.08	.44	-.01	.16	.21	-.12	.58	-.01	.19	.06	.43	.16	-.07	.18	.01	.03
Fruit juices	.14	.20	.05	.16	.03	.21	.29	.06	.20	.04	.13	.17	.13	.23	.02	.26	.07
Alcoholic beverages	.09	.04	.30	.00	-.18	.14	.06	.39	.00	-.21	.02	.22	.08	.06	.05	.09	.01
Cold cuts	.07	.04	.11	.12	.16	.11	.05	.15	.15	.19	.01	.09	.07	.04	.04	.07	.01
Salad dressing	.32	.34	-.03	-.06	.00	.48	.50	-.04	-.08	.00	.47	.48	.35	.46	.13	.57	.34
Sugar	.27	-.15	-.22	.32	-.21	.40	-.22	-.28	.40	-.25	.21	.51	.34	-.15	.38	.01	.14
Fatty Sauces/Creams/Mayo	.02	.08	.30	.18	.03	.02	.11	.39	.23	.04	.01	.22	.00	.08	-.03	.08	.01
Spices	.15	.19	.00	-.11	-.09	.23	.27	.00	-.14	-.11	.13	.16	.14	.22	.04	.26	.07
Eigenvalues	2.2	2.1	1.7	1.6	1.4	2.21	2.9	1.7	1.6	1.4	-	-	1.4	1.3	1.4	1.3	-
Explained variance (%)	6.5	6.1	5.0	4.6	4.1	6.5	6.1	5.0	4.6	4.1	-	-	30.8	26.9	30.8	26.9	-
Accumulated variance (%)	6.5	12.6	17.6	22.2	26.3	6.5	12.6	17.6	22.2	26.3	-	-	30.8	57.7	30.8	57.7	-

In bold: loading ≥ |0,30|; KMO = 0,59; Bartlett's sphericity test ($p < 0,001$) / h^2 = communalities.

the rice group (Factor 1), while the highest loading in PCA was 0.41 for the same food group (Component 1).

In FA, the communalities of the variables ranged from 0.00 to 0.34, with seventeen variables explaining less than 5% of the common variance, while in PCA, the communalities of the variables ranged from 0.02 to 0.47 for two components and 0.06 to 0.58 for five components, showing that by extracting a greater number of components, the amount of common variance increases.

Applying a loading cut-off of $|0.30|$ to simplify the interpretation of the factor structure, we can observe two factors: factor one (30.8% of explained variance) showed positive loadings to rice, bread/toasts/crackers, beans, butter/margarine, and sugar; and factor 2 (26.9% of variance) was characterized by canned vegetables, non-leafy vegetables, and salad dressing.

Figure 3 presents the graphic representation of the correlations between the first two components and food group intakes in PCA. This is the first plan to analyze and represents

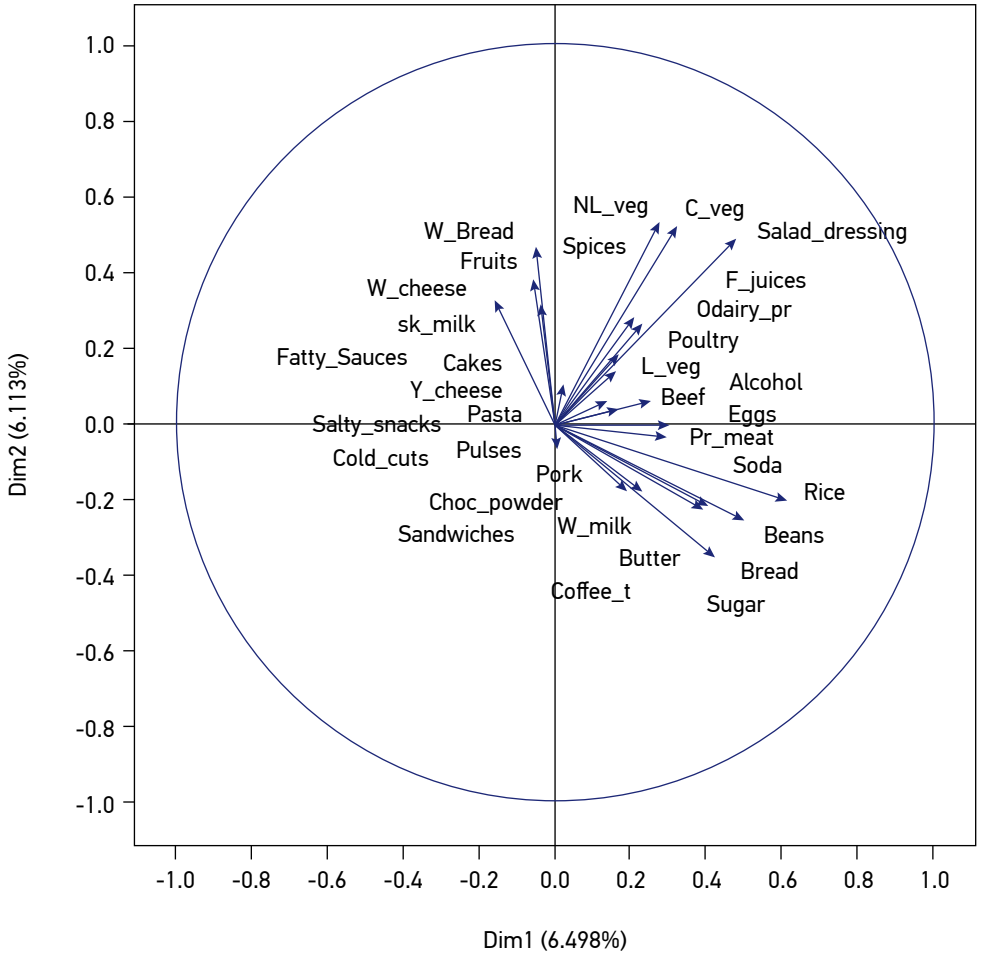

Figure 3. Graph of the correlations between the first two components and food item intakes.

the most important part of the variance. This graph reveals that some food items are well represented in the first plan, such as canned vegetables, salad dressing, and rice, whose vectors are closer to the ray size 1 (maximum correlation). We notice that canned vegetables, salad dressing, and non-leafy vegetables are consumed in association (similar to the results for factor 2). White cheese, whole bread, fruits, and low-fat and skim milk are also consumed in association. Moreover, bread/toasts/crackers, butter/margarine, rice, beans, sugar, and coffee/tea (similar to factor 1) are consumed in association and inversely associated with the intake of white cheese, whole bread, fruits, and low-fat and skim milk (a factor not identified in FA).

DISCUSSION

This work aimed to compare and present the differences and similarities between FA and PCA, highlighting that the choice of method will depend on the study objective: PCA only describes a large data set in a simpler dimension, while FA is a statistical model used to build dietary patterns. Also, our results showed that FA and PCA might lead to different estimates, especially when the common variances of the variables are low. The difference in variable factor loadings between FA and PCA, as observed in this study, might be explained by the low communalities of the variables. In this regard, some authors have suggested that when the number of variables is above 30, common variances exceed 0.60 for most variables[5], and error (unique/specific variance) is close to zero[5,26], FA and PCA can produce similar results. However, even if the final solution (factors and components) in most studies is often similar between the two methods, the interpretation of the findings and data modeling should not be made in the same way.

FA can be applied to studies that aim to analyze the dietary pattern of a certain population since it generates factors that represent a latent variable, which will explain the consumption of food items or food groups. Each food item/group is estimated (with random error) by a linear combination of non-observed variables, the factors (latent variables). The factor scores calculated in FA represent the "pattern" of the individual and not a "real" observation[27].

PCA should be used when the researcher intends to reduce the original data into a smaller set of components for interpretation to reproduce part of the variability in fewer linear combinations of the original variables. The interpretation of the final solution can be made graphically, as shown in this study. Thus, the objective, in this case, is to identify linear combinations of food items or food groups responsible for the larger dietary variability of those individuals and to select food items to elaborate a food frequency questionnaire (FFQ)[6]. Qin et al.[28] used PCA to determine the sensory attributes of apple cider samples based on bi-plot and found that floral and fruity odors were highly correlated to sweet taste and opposed to more complex aroma attributes.

The factors obtained in an FA are latent variables, *i.e.*, random variables whose occurrence is hidden. In other words, the latent variable represents the true measure of the

variables, taking into account the error associated with the measure of the variables originally observed, as the latent variable assumes that each of its items has an associated measurement error and considers this information in its estimation. Castro et al.[29] evaluated the association between dietary patterns and metabolic cardiovascular risk factors in Brazilian adults and, to build the dietary patterns, the authors considered that each food group had measurement errors that could be predicted by dietary patterns.

The latent variable — factor — may represent hypothetical constructs, which contemplate an epistemological aspect, an unobserved concept, such as the characterization of the eating habits of a given population, be it Western, Traditional, Prudent, or Mediterranean. Therefore, FA provides an estimate of the relationship between food and food groups consumed by different individuals (regardless of random error), allowing the identification of food group combinations, or food patterns, that represent the eating habits of the population studied[30].

Although both analyses require attention regarding the sample size, number of variables observed, pattern of covariation/correlation between variables, and number of components/factors that will be formed, the choice of the best method to use will depend on the objective of each study.

CONCLUSION

Researchers need to be aware of the different characteristics of PCA and FA to decide on the most appropriate method to achieve the objectives of their research. Even though in some situations both methods could provide similar results, they are conceptually different, leading to a diverse interpretation of results.

REFERENCES

1. Meyers LS, Gamst G, Guarino AJ. Applied multivariate research: design and interpretation. California: Sage; 2006.
2. Ocké MC. Evaluation of methodologies for assessing the overall diet: dietary quality scores and dietary pattern analysis. Proc Nutr Soc 2013; 72(2): 191-9. http://doi.org/10.1017/S0029665113000013
3. Hu FB. Dietary pattern analysis: a new direction in nutritional epidemiology. Curr Opin Lipidol 2002; 13(1): 3-9.
4. Johnson RA, Wichern DW. Applied Multivariate Statistical Analysis. 6ª ed. Upper Saddle River: Pearson Prentice Hall; 1998.
5. Hair Jr. JF, Black WC, Babin BJ, Anderson RE, Tatham RL. Multivariate Data Analysis. 6ª ed. Upper Saddle River: Pearson Prentice Hall; 2006.
6. Schneeweiss H, Mathes H. Factor analysis and principal components. J Multivar Anal 1995; 55(1): 105-24. http://doi.org/10.1006/jmva.1995.1069
7. Suhr D. Principal component analysis vs. exploratory factor analysis. In: SUGI 30 Proceedings [Internet]. 2005 [accessed on May 18, 2017]. Available from: http://www2.sas.com/proceedings/sugi30/Leadrs30.pdf
8. Park HS, Dailey R, Lemus D. The use of exploratory factor analysis and principal components analysis in communication research. Hum Commun Res 2002; 28(4): 562-77. http://doi.org/10.1111/j.1468-2958.2002.tb00824.x
9. Brown JD. Principal components analysis and exploratory factor analysis – definitions, differences, and choices. Shiken: JALT Testing & Evaluation Sig Newsletter [Internet] 2009 [accessed on Mar. 27, 2017]; 13(1): 26-30. Available from: https://jalt.org/test/PDF/Brown29.pdf

10. Tabachnick BG, Fidell LS. Using multivariate statistics. 5ª ed. Upper Saddle River: Pearson Allyn & Bacon; 2007.
11. Castro MA, Baltar VT, Selem SSC, Marchioni DML, Fisberg RM. Empirically derived dietary patterns: interpretability and construct validity according to different factor rotation methods. Cad Saúde Pública 2015; 31(2): 298-310. http://dx.doi.org/10.1590/0102-311X00070814
12. Rodrigues-Motta M, Galvis Soto DM, Lachos VH, Vilca F, Baltar VT, Verly Junior E, et al. A mixed-effect model for positive responses augmented by zeros. Stat Med. 2015; 34(10): 1761-78. http://dx.doi.org/10.1002/sim.6450
13. Tooze JA, Kipnis V, Buckman DW, Carroll RJ, Freedman LS, Guenther PM, et al. A mixed-effects model approach for estimating the distribution of usual intake of nutrients: the NCI method. Stat Med 2010; 29(27): 2857-68. https://doi.org/10.1002/sim.4063
14. Kaiser HF. An index of factorial simplicity. Psychometrika 1974; 39(1): 31-6. https://doi.org/10.1007/BF02291575
15. Zygmont C, Smith MR. Robust factor analysis in the presence of normality violations, missing data, and outliers: Empirical questions and possible solutions. Quantitative Method for Psychology. 2014; 10(1): 40-55. https://doi.org/10.20982/tqmp.10.1.p040
16. Rencher AC. Methods of multivariate analysis. 2ª ed. New York: John Wiley & Sons; 2002. v. 492.
17. Hayton JC, Allen DG, Scarpello V. Factor retention decisions in exploratory factor analysis: a tutorial on parallel analysis. Organ Res Methods 2004; 7(2): 191-205. https://doi.org/10.1177%2F1094428104263675
18. Kaiser HF. The varimax criterion for analytic rotation in factor analysis. Psychometrika 1958; 23(3): 187-200. https://doi.org/10.1007/BF02289233
19. Floyd FJ, Widaman KF. Factor analysis in the development and refinement of clinical assessment instruments. Psychol Assess 1995; 7(3): 286-99. https://psycnet.apa.org/doi/10.1037/1040-3590.7.3.286
20. Sass DA. Factor loading estimation error and stability using exploratory factor analysis. Education Psychology Measurement 2010; 70(4): 557-77. https://doi.org/10.1177%2F0013164409355695
21. Yong AG, Pearce S. A beginner's guide to factor analysis: focusing on exploratory factor analysis. Tutor Quant Methods Psychol 2013; 9(2): 79-94. http://dx.doi.org/10.20982/tqmp.09.2.p079
22. Kline P. An easy guide to factor analysis. New York: Routledge; 1994.
23. De Oliveira Santos R, Fisberg RM, Marchioni DM, Baltar VT. Dietary patterns for meals of Brazilian adults. Br J Nutr 2015; 114(5): 822-8. https://doi.org/10.1017/S0007114515002445
24. Cunha DB, Almeida RMVR, Pereira RA. A comparison of three statistical methods applied in the identification of eating patterns. Cad Saúde Pública 2010; 26(11): 2138-48. http://dx.doi.org/10.1590/S0102-311X2010001100015
25. DiStefano C, Zhu M, Mîndrilă D. Understanding and Using Factor Scores: Considerations for the Applied Researcher. Pract Assess Res Eval 2009; 14(20). Available from: http://pareonline.net/getvn.asp?v=14&n=20
26. Velicer WF, Peacock AC, Jackson DN. A comparison of component and factor patterns: A Monte Carlo approach. Multivariate Behav Res 1982; 17(3): 371-88. http://dx.doi.org/10.1207/s15327906mbr1703_5
27. Shulze MB, Hoffmann K. Methodological approaches to study dietary patterns in relation to risk of coronary heart disease and stroke. Br J Nut 2006; 95(5): 860-9. http://dx.doi.org/1079/BJN20061731
28. Qin Z, Petersen MA, Bredie WLP. Flavor profiling of apple ciders from the UK and Scandinavian region. Food Res Int 2018; 105: 713-23. https://doi.org/10.1016/j.foodres.2017.12.003
29. Castro MA, Baltar VT, Marchioni DM, Fisberg RM. Examining associations between dietary patterns and metabolic CVD risk factors: a novel use of structural equation modelling. Br J Nutr 2016; 115(Suppl. 9): 1586-97. https://doi.org/10.1017/S0007114516000556
30. Skrondal A, Rabe-Hesketh S. Generalized latent variable modeling: multilevel, longitudinal and structural equation models. London: Chapman & Hall; 2004.

Received on: 03/26/2018
Accepted on: 05/15/2018

Authors' Contributions: Study concept and design: Roberta de Oliveira Santos, Bartira Mendes Gorgulho, Dirce Maria Marchioni and Valéria Troncoso Baltar. Data analysis and interpretation: Roberta de Oliveira Santos, Bartira Mendes Gorgulho, Michelle Alessandra de Castro and Valéria Troncoso Baltar. Writing and critic discussion: Roberta de Oliveira Santos, Bartira Mendes Gorgulho, Michelle Alessandra de Castro, Regina Mara Fisberg, Dirce Maria Marchioni and Valéria Troncoso Baltar.