Focused Principal Component Analysis: a graphical method for exploring dietary patterns

Análise de Componente Principal Focada: um método gráfico para explorar padrões alimentares

Raquel Canuto [1]
Suzi Camey [2]
Denise P. Gigante [3]
Ana M. B. Menezes [3]
Maria Teresa Anselmo Olinto [1]

[1] Programa de Pós-graduação em Saúde Coletiva, Universidade do Vale do Rio dos Sinos, São Leopoldo, Brasil.
[2] Instituto de Matemática, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil.
[3] Programa de Pós-graduação em Edidemiologia, Universidade Federal de Pelotas, Pelotas, Brasil.

Correspondence
M. T. A. Olinto
Programa de Pós-graduação em Saúde Coletiva, Universidade do Vale do Rio dos Sinos.
Av. Unisinos 950, C. P. 275, São Leopoldo, RS 93022-000, Brasil.
mtolinto@unisinos.br

Abstract

The aim of the present study was to introduce Focused Principal Component Analysis (FPCA) as a novel exploratory method for providing insight into dietary patterns that emerge based on a given characteristic of the sample. To demonstrate the use of FPCA, we used a database of 1,968 adults. Food intake was obtained using a food frequency questionnaire covering 26 food items. The focus variables used for analysis were age, income, and schooling. All analyses were carried out using R software. The graphs generated show evidence of socioeconomic inequities in dietary patterns. Intake of whole-wheat foods, fruit, and vegetables was positively correlated with income and schooling, whereas for refined cereals, animal fats (lard), and white bread this correlation was negative. Age was inversely associated with intake of fast-food and processed foods and directly associated with a pattern that included fruit, green salads, and other vegetables. In an easy and direct fashion, FPCA allowed us to visualize dietary patterns based on a given focus variable.

Food Consumption; Principal Component Analysis; Nutritional Epidemiology

Introduction

While dietary patterns reflect an individual's food preferences, they are also influenced by other characteristics, such as economic history, income, schooling and demographic characteristics (sex and age).

The statistical methods most commonly used for identifying dietary patterns among populations, or among specific population groups, include data reductions based on *a posteriori* models, such as cluster analysis and principal component analysis (PCA) [1]. Both clustering and PCA are able to identify underlying structures among different food items, i.e. patterns of reduction and clustering of the dataset. However, investigating the relationship between the dietary patterns identified by these methods and population characteristics requires subsequent dependence analysis, which entails resorting to multivariate regression models that include both the dietary pattern and other characteristics of the sample.

In Brazil, a limited number of studies have attempted to identify dietary patterns and their association with population characteristics. The findings from these studies are consistent in that they indicate the existence of socioeconomic inequities in the dietary patterns of the population. Lenz et al. [2] identified five dietary patterns by means of PCA. Of these, three displayed the characteristics of a healthy diet, such as the pres-

ence of fruits, vegetables, whole-wheat/wholemeal bread, brown rice and nuts, and two were composed of unhealthy forms of carbohydrates and fats. Subsequent multivariate analysis using Poisson regression showed that healthy patterns were more likely followed by women with higher income and schooling, and by older women [2]. Olinto et al. [3], also using PCA followed by Poisson regression, identified dietary patterns among young adults and found that healthy patterns were associated with female sex and higher socioeconomic status.

Focused Principal Component Analysis (FPCA) has emerged as a novel method for *a posteriori* exploratory analysis that is appropriate for scenarios in which explanations are sought for the relationships among a group of variables based on a given characteristic of the sample. Applying FPCA to food data makes it possible to view the correlation between diet and a given variable of interest, at the same time as enabling detection of correlations between the different food items themselves. In FPCA, unlike in PCA, dietary patterns focusing on a particular variable of interest are formed, and are presented exclusively in graphical format [4].

With this in mind, the aim of the present study was to present FPCA analysis by applying this methodology to one food intake dataset. As an example, we will use a demographic variable (age) and two socioeconomic variables (income and schooling) as the focus variables.

Methods

FPCA

FPCA allows one to visualize, simultaneously, both correlations between food items and a particular variable of interest and correlations among food items themselves. For example, suppose that *n* subjects have reported intake of two food items (milk and bread), and we are interested in exploring the correlation between the consumption of both these items, but at the same time focusing on the subject's income. In other words, the major issue is to represent the relationship between income and food consumption without losing the relationship that different food items have with each other. In this case, the FPCA graph faithfully represents, at the same time: (1) the correlation between bread intake and income and (2) the correlation between milk intake and income, in addition to projecting, in a two-dimensional plane, the correlation matrix between income, bread intake, and milk intake (the latter a three-dimensional correlation).

This concept can be extended to the intake of *p* different food items and a variable of interest. In this case, the FPCA graph would faithfully represent the correlation between each food and the variable of interest, presenting a projection of the correlation matrix between all variables [a matrix of (*p+1*) x (*p+1*) dimensions] on a two-dimensional plane.

FPCA presents correlations in graphic format as concentric circles, those of smaller radius representing stronger correlations. The center of these circles (target) contains the variable of interest, which directs the analysis. Negative and positive correlations are differentiated in the graph by use of different colors or patterns. The interpretation of points in an FPCA graph is as follows:

- The closer the point is to the center, the closer to 1 (or -1) is the correlation between intake of the food and the variable of interest;
- Two points close to one another indicate a strong positive correlation between the intake of the foods which they represent;
- Two diametrically opposed points indicate a strong negative correlation between the intake of the foods which they represent;
- Two points placed at a similar distance from the origin, parallel to one of the axes, indicate absence of correlation between the intake of the foods which they represent;
- The dashed circle delimits statistical significance at the 5% level.

Mathematical details of the construction of these models can be found in the work of Falissard [5].

Currently, this type of analysis can be performed using R software (R Development Core Team, Vienna, Austria). R is an open-code free-access software that works by means of libraries. FPCA is included in the psy library (Falissard B. Various procedures used in psychometry. http://cranr.c3sl.ufpr.br, accessed on 12/Jul/2009). Both program and library can be obtained on-line at http://www.r-project.org/.

Information on the installation of R and psy is available at http://carn.r-projetc.org/doc/manuals/R-admin.html; information on importing databases is available at http://cran.r-project.org/doc/manuals/R-data.html; information on database manipulation is available at http://cran.r-project.org/doc/manuals/R-intro.html; and a Portuguese-language tutorial is available at http://leg.ufpr.br/~paulojus/.

To execute FPCA, the following command must be entered in full, in a single line:
fpca(datafile, y, x,cx=0.75,namesvar= attributes (datafile)$names, pvalues="No", partial="Yes", input="data", contraction="No", sample.size=1).

Where; datafile: name of the database; y: column number of the variable of interest; x: vector with the column numbers of food intake variables; cx: font size (0.75 is default; 1 for larger, 0.5 for smaller); namesvar: variable names (the default is to use the name of the variable column); pvalues: vector of pre-specified values (default: pvalues="No") (see manual); partial: default: partial="Yes", corresponds to the original method (see manual); input: "Cor" to input a correlation matrix instead of the data (default: input="data"); contraction: changes the aspect of the graph, contraction="Yes" is convenient for many variables (default: contraction="No"); sample.size: size of the sample. Must be specified if input="Cor".

We will now present an example in which we applied FPCA to investigate dietary patterns, focusing on socioeconomic characteristics and age among the urban population of the city of Pelotas, Rio Grande do Sul State, Southern Brazil.

- **Example**

We used data from a cross-sectional, population-based survey on a representative sample of 1,968 men and women aged 20-70 years who were living in the urban area of the city. Details of the method used in this study have been published elsewhere [6].

Briefly, food intake was collected by means of an interviewer-administered food frequency questionnaire (FFQ). The FFQ assessed the usual intake of a list of 26 food items over the past year. Two scales were used for collecting the intake frequency: one for low and one for high-frequency foods. The response choices for low-frequency foods were: never, once a month, 2-3 times per month, 1-2 times per week, 3-4 times per week, or 5 or more times per week. The response choices for high-frequency foods were: < once a week, once a week, 2-3 times per week, 4-6 times per week, or every day. For analysis purposes, all intake frequency data were adjusted to monthly intake, as follows: in categories that measured weekly intake, we calculated the mean of the lowest and highest weekly intake frequency and multiplied the result by 4. For example, 3-4 times per week is equivalent to 14 times per month. We thus constructed a single numerical ordinal scale for food intake with a range from 0 to 9 (from never to every day).

We defined age, schooling and income as focus variables, or variables that explained dietary patterns. Subject age was collected as full years of life; schooling was collected as full years of study; and income was obtained as total family income, from which per capita income was derived.

Every graph generated by FPCA presents a dashed circle delimiting the statistical significance of the correlation at the 5% level. Light gray dots inside the dashed circle show direct significant associations with the focus variable, and dark gray dots show inverse associations.

Results

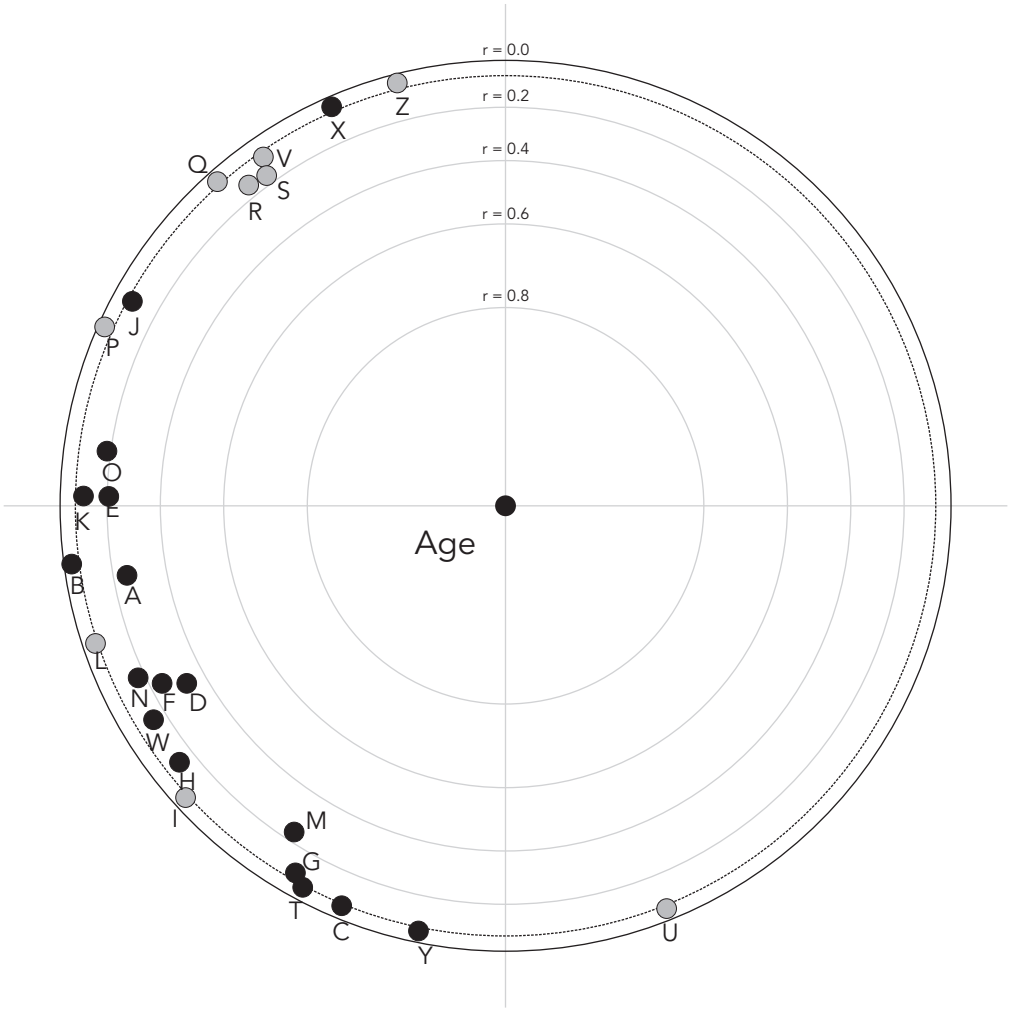
Figure 1 shows the relationship between food items and age. The associations with age were predominantly inverse (dark gray dots) rather than direct (light gray dots). The graph shows that two dietary patterns were formed: one Western-type pattern and another pattern of "prudent" or healthy type. However, the majority of the correlations were weak; i.e. the correlation dots were plotted in concentric circles representing $r \leq 0.2$ (correlation coefficient), except for hot dogs, which presented correlation coefficient values between 0.2 and 0.4. The Western type of diet, with an inverse relationship with age, can be seen at two points in the graph: in the lower left quadrant [hot dogs (D), mayonnaise (F), soda/soft drinks (W), French fries (M), eggs (H) and chips/crisps (N)]; and at the lower limit of the upper left quadrant [sweets/desserts (K), hamburgers (A), ham (E) and ice cream (O)]. Although butter item (G) is close to those foods, it is located on the dashed circle, i.e. it has a borderline correlation. The other type (prudent or healthy) can be seen in the upper left corner of the graph, and this included fruit (R), green salads (S) and other vegetables (V). This pattern was associated positively with age.

A prudent pattern also emerged when the analysis focused on income (Figure 2). This pattern can be seen in the upper left quadrant of the graph. It includes cereals (X), vegetables (V), green salads (S) and fruit (R), and was positively associated with income, i.e. reported intake of these items increased together with income. In the same quadrant, we were also able to identify a second pattern associated with income. Although the food items forming this pattern are spread out, it is possible to discern characteristics of a Western diet in it. In this pattern, income was associated directly with intake of cheese (J), sweets/desserts (K), steak (B) and ham (E). Figure 2 also shows that the intake of both beans (U) and white bread (Y) was inversely associated with income, i.e. the higher the income level was, the less frequent the intake of these foods was among adults.

Figure 3 shows the results from analysis using schooling as the focus variable. The associations with schooling were predominantly direct

Figure 1

Focused Principal Component Analysis (FPCA) on age and intake of different foods among adults in Pelotas, Rio Grande do Sul State, Brazil, 2000.

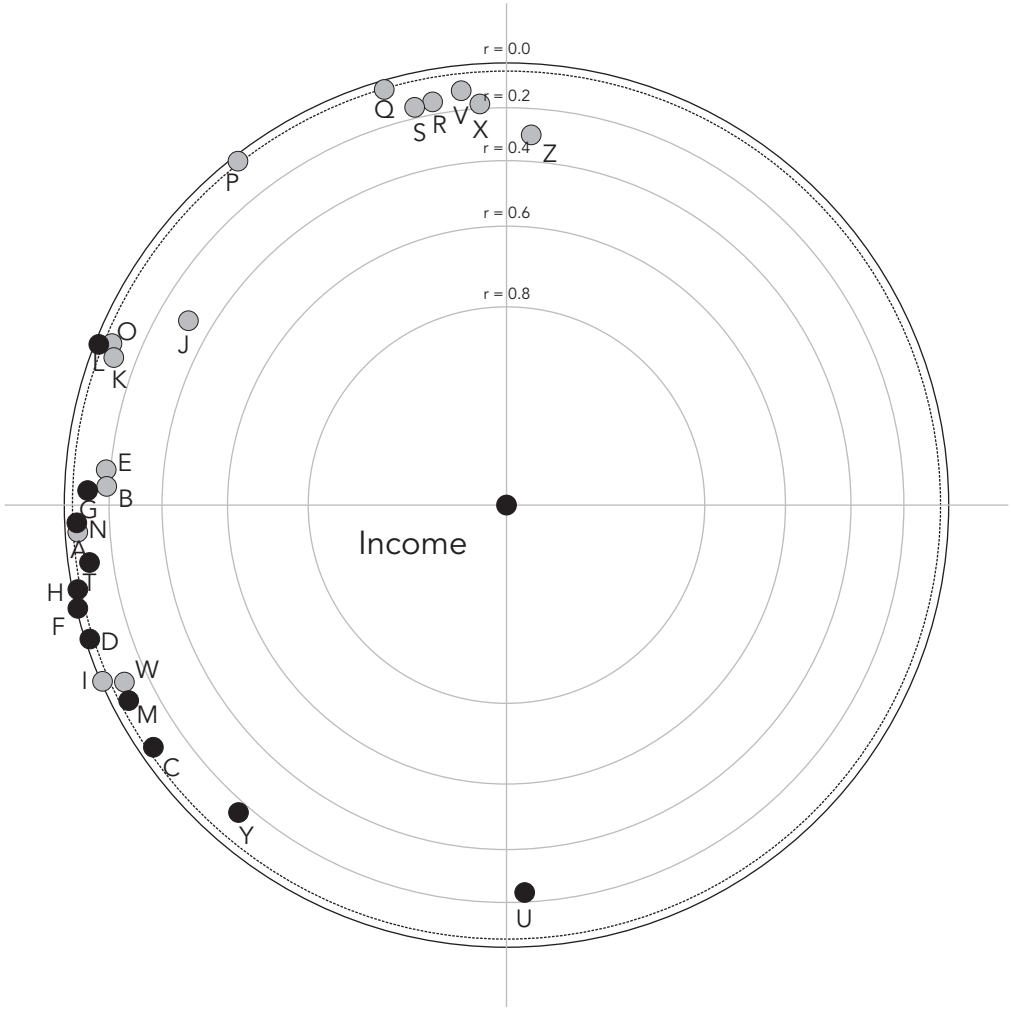


A: hamburger; B: steak; C: chicken; D: hot dog; E: ham; F: mayonnaise; G: butter; H: eggs; I: bacon; J: cheese; K: sweets/deserts; L: milk; M: French fries; N: chips/crisps; O: ice cream; P: cake; Q: orange juice; R: fruits; S: green salad; T: potato; U: black beans; V: vegetables; X: cereals; Z: whole-wheat/wholemeal bread; Y: white bread; W: soda/soft drinks.

(light gray dots) rather than inverse (dark gray dots). Although in a more dispersed manner, we identified two dietary patterns similar to those reported for income, which we were able to name prudent and Western. The prudent or healthy pattern was composed of vegetables (V), green salad (S), fruit (R), cereals (X) and whole-wheat/wholemeal bread (Z). The Western type of diet was composed of cheese (J), sweets/desserts (K), ice cream (O), hamburger (A), steak (B) and ham (E). Yet another Western pattern with weak correlations emerged immediately below this, in the lower left quadrant. This pattern was composed of fast-food items: hot dogs (D), mayonnaise (F) and soda/soft drinks (W), which were inversely associated with income (Figure 3). Although hamburgers (A) were placed closer to the Western characteristic pattern, they were also close to, and in the same quadrant as, the fast-food pattern. Beans (U) were located close to white bread

Figure 2

Focused Principal Component Analysis (FPCA) on income and intake of different foods among adults in Pelotas, Rio Grande do Sul State, Brazil, 2000.



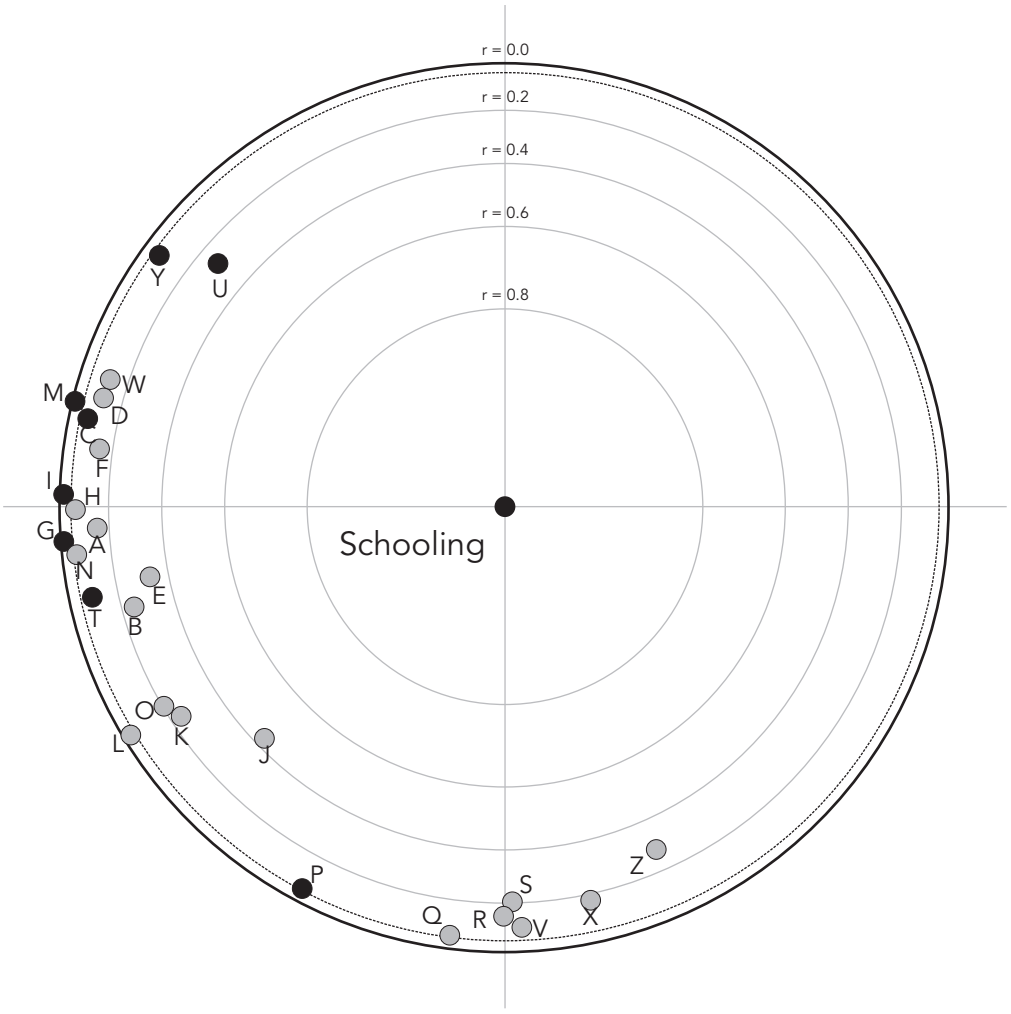
A: hamburger; B: steak; C: chicken; D: hot dog; E: ham; F: mayonnaise; G: butter; H: eggs; I: bacon; J: cheese; K: sweets/deserts; L: milk; M: French fries; N: chips/crisps; O: ice cream; P: cake; Q: orange juice; R: fruits; S: green salad; T: potato; U: black beans; V: vegetables; X: cereals; Z: whole-wheat/wholemeal bread; Y: white bread; W: soda/soft drinks.

(Y), with borderline significance: both of these are part of a common Brazilian dietary pattern that is inversely associated with schooling.

Furthermore, in Figure 3, two diametrically opposed points indicating a negative correlation between the intakes of the foods that they represent can be seen; i.e. the intake of whole-wheat/wholemeal bread (Z) and black beans (U). There was a negative correlation between these when education was the focus variable. In addition, it can be seen in Figure 3 that cheese (J) and beans (U) presented similar distances from the point of origin of the graph and were parallel to its vertical axis. Thus, there was an inverse relationship between the intakes of beans and cheese.

Figure 3

Focused Principal Component Analysis (FPCA) on schooling and intake of different foods among adults in Pelotas, Rio Grande do Sul State, Brazil, 2000.



A: hamburger; B: steak; C: chicken; D: hot dog; E: ham; F: mayonnaise; G: butter; H: eggs; I: bacon; J: cheese; K: sweets/deserts; L: milk; M: French fries; N: chips/crisps; O: ice cream; P: cake; Q: orange juice; R: fruits; S: green salad; T: potato; U: black beans; V: vegetables; X: cereals; Z: whole-wheat/wholemeal bread; Y: white bread; W: soda/soft drinks.

Discussion

FPCA provided us with an easy and direct way to view correlations between different food items and a given focus variable, at the same time as it allowed us to observe correlations that existed between the food items themselves. In the example presented, dietary patterns emerged based on socioeconomic variables (income and schooling). The graphs indicated the presence of socioeconomic inequities relating to food intake patterns in this dataset. Although the primary goal of the present article was to introduce FPCA to the field of nutritional epidemiology, the patterns we detected are also worth commenting on.

Briefly, FPCA was used in conjunction with three focus variables: age, schooling and income. Age was inversely associated with fast-food and processed items (hamburgers, hot dogs, French fries, etc.), and directly associated with a healthy

dietary pattern that included fruit, green salads and other vegetables. Therefore, increasing age seems to lead to a reduction in the prevalence of unhealthy dietary patterns. Other studies have shown similar direct relationships with age, i.e. older individuals of both sexes tend to adopt more healthy diets [7,8,9,10,11]. We draw attention to the fact that we only investigated correlations between the variables, and therefore no causal relationships can be inferred.

Socioeconomic inequities could be detected using both income and schooling as focus variables. Patterns comprising whole foods as well as fruit and vegetable intake were directly associated with income and schooling. Foods of lower cost, such as white bread and black beans, were inversely associated with these socioeconomic variables. Similar findings have been reported in recent studies carried out in southern Brazil that used principal component analysis [2,3,12]. Income and schooling have been shown to be limiting factors for a healthy diet, as well as for other lifestyles [13].

It is noteworthy that although each graph made it possible to form dietary patterns focusing on one specific variable at a time, these dietary patterns were complementary, i.e. their findings were not mutually exclusive. Intuitively, it is implausible that one variable alone would influence the formation of dietary patterns, and therefore the graphs need to be interpreted in a complementary manner. For example, Figure 1 shows that age presented an inverse relationship with fast foods (hamburgers, soda/soft drinks, French fries, etc.), but on the other hand, the same graph indicates the existence of a prudent pattern that is directly related to increasing age. Likewise, Figures 2 and 3 show that a similar prudent pattern was set up by using income and schooling as the focus variables. Therefore, the interpretation that can be made from this is that with increasing age, healthier dietary habits emerge, but that these habits are influenced by income and schooling, such that the higher the income and schooling levels were, the greater the adherence to a prudent dietary pattern.

Focused principal component analysis was first proposed by Falissard et al. [4], working in the field of psychiatry. This analysis was intended to evaluate the relationship between a series of explanatory variables and a variable of interest. This is not a form of predictive analysis; rather, it is an instrument for exploratory analysis of data. To date, we are unaware of any studies using this methodological approach for identifying dietary patterns, or even for describing the intake of different foods according to the characteristics of individuals or social groups. Unlike the other two methods for exploring dietary patterns, the present method generates results in exclusively graphical form (diagrams).

Certain limitations of our dataset and methodology should be borne in mind. Regarding the dataset, our major difficulty was the manner in which the intake frequency was obtained, i.e. in categories that differentiated between high and low-frequency intake. This made it necessary to create equivalence of intake frequency in order to transform food item consumption data into a single continuous scale. Regarding the analysis method (FPCA) three limitations are noteworthy. First, the number of food items can influence the patterns formed. When an item is removed or included in the analysis, the correlations change such that new relationships emerge and the direction of existing ones change. Moreover, including a large number of food items can make graphical visualization more difficult. In cases of extensive lists of food items, we recommended that they should be aggregated before analysis. Another limitation is that, unlike PCA, FPCA does not generate factorial scores, and, unlike cluster analysis, it does not classify subjects into groups. Furthermore, the focus variable has to be continuous. Finally, this analysis can currently be performed only on R; however, the 17th version of SPSS (SPSS Inc., Chicago, USA) will allow R routines to be used in the SPSS platform.

In conclusion, in the present study, FPCA was used to demonstrate the influence of socioeconomic characteristics such as income, schooling and age on the formation of dietary patterns. This statistical method proved to be a promising analytical instrument for exploring dietary patterns from the standpoint of different variables of interest in the population, and at the same time, it may contribute towards generating hypotheses in surveys on diet and health-related outcomes.

Resumo

O presente estudo teve objetivo de apresentar a Análise de Componentes Principais Focada (ACPF) como um método exploratório para investigar padrões alimentares a partir de características da amostra. Para exemplificar utilizou-se as variáveis idade, renda e escolaridade de um banco de dados de 1.968 adultos. O consumo alimentar foi obtido através questionário de frequência alimentar (QFA) com 26 itens alimentares. As análises foram realizadas no programa R. Os gráficos gerados evidenciaram iniquidades socioeconômicas na conformação dos padrões alimentares. Alimentos integrais, frutas e verduras foram diretamente correlacionados com renda e escolaridade, e cereais refinados, gordura animal e pão branco tiveram associação inversa. A idade mostrou-se como associada inversamente a alimentos fast-food *e industrializados e, diretamente, a um padrão "saudável" que inclui frutas, salada verde e outros vegetais. De maneira fácil e direta, a ACPF permitiu a visualização de correlações entre alimentos a partir de variáveis escolhidas como foco.*

Consumo de Alimentos; Análise de Componente Principal; Epidemiologia Nutricional

Contributors

M. T. A. Olinto conceived and wrote the article, and was responsible for data analysis. R. Canuto and S. Camey participated in the data analysis and write up. D. P. Gigante and A. M. B. Menezes reviewed the article.

Acknowledgments

To the National Research Council (CNPq-PQ nº. 308833/2006-6).

References

1. Tucker N. Empirically derived eating patterns using factor or cluster analysis: a review. Nutr Rev 2004; 5:177-203.
2. Lenz A, Olinto MTA, Dias-da-Costa JS, Alves AL, Balbinotti M, Pattussi MP, et al. Socioeconomic, demographic and lifestyle factors associated with dietary patterns of women living in Southern Brazil. Cad Saúde Pública 2009; 25:1297-306.
3. Olinto MTA, Willet W, Gigante D, Victora C. Sociodemographic and lifestyle characteristics in relation to dietary patterns among young Brazilian adult. Public Health Nutr; in press.
4. Falissard B, Corruble E, Mallet L, Hardy P. Focused Principal Component Analysis: a promising approach for confirming findings of exploratory analysis? Int J Methods Psychiatr Res 2006; 10:191-5.
5. Falissard B. Focused Principal Components Analysis: looking at a correlation matrix with a particular interest in a given variable. J Comput Graph Stat 1999; 8:906-12.
6. Dias-da-Costa JS, Barcellos FC, Sclowitz ML, Sclowitz I, Castanheira M, Olinto MTA, et al. Prevalência de hipertensão arterial em adultos e fatores associados: um estudo de base populacional urbana em Pelotas, Rio Grande do Sul, Brasil. Arq Bras Cardiol 2007; 88:59-65.
7. Sadakane A, Tsutsumi A, Gotoh T, Ishikawa S, Ojima T, Kario K, et al. Dietary patterns and levels of blood pressure and serum lipids in a Japanese population. J Epidemiol 2008; 18:58-67.
8. Nanri A, Yoshida D, Yamaji T, Mizoue T, Takayana GIR, Suminori K. Dietary patterns and C-reactive protein in Japanese men and women. Am J Clin Nutr 2008; 87:488-96.
9. Sánchez-Villegas A, Delgado-Rodríguez M, Martínez-González MA, De Irala-Estévez J; Seguimiento Universidad de Navarra Group. Gender, age, socio-demographic and lifestyle factors associated with major dietary patterns in the Spanish Project SUN (Seguimiento Universidad de Navarra). Eur J Clin Nutr 2003; 57:285-92.
10. Cunha DB. Padrões de consumo alimentar e excesso de peso em adultos de Duque de Caxias [Masters Thesis]. Rio de Janeiro: Universidade Federal do Rio de Janeiro; 2008.
11. Paradis AM, Pérusse L, Vohl M. Dietary patterns and associated lifestyles in individuals whit and without familial history of obesity: a cross-sectional study. Int J Behav Nutr Phys Act 2006; 3:38.
12. Henn RL. Padrão alimentar e excesso de peso em uma população adulta da cidade de Porto Alegre, RS, 2005 [Doctoral Dissertation]. Porto Alegre: Universidade Federal do Rio Grande do Sul; 2006.
13. Whitehead M. The concepts and principles of equity and health. Copenhagen: World Health Organization; 1991.

Submitted on 28/Aug/2009
Final version resubmitted on 11/May/2010
Approved on 17/May/2010