The application of exploratory factor analysis in the identification of dietary patterns: some notes from an analysis of gastric cancer

Paola Bertuccio[1,2], Francesca Bravi[1,2], Valeria Edefonti[2]

[1] Istituto di Ricerche Farmacologiche "Mario Negri", Milan, Italy
[2] Istituto di Statistica Medica e Biometria "Giulio A. Maccacaro", Università degli Studi di Milano, Milan, Italy

Corresponding Author:
Paola Bertuccio, ScD
Istituto di Ricerche Farmacologiche "Mario Negri"
Via Giuseppe La Masa 19, 20156 Milan, Italy
Tel.: +390239014667 - Fax: +390233200231
E-mail: paola.bertuccio@marionegri.it

Summary

Objectives. Our aims are to describe the steps through which we identified nutrient-based dietary patterns using factor analysis, and to compare results from three different factor analysis solutions.

Methods. We derived data from an Italian case-control study of gastric cancer, including 230 cancer cases and 547 frequency matched controls. We applied exploratory principal component factor analyses on a selected set of 28 micro- and macronutrients. We estimated odds ratios and corresponding 95% confidence intervals using conditional multiple logistic regression models.

Results. The cumulative explained variances were 75%, 80%, 84% for the four-, five-, and six-factor solutions, respectively. The patterns shared across the three solutions were named: *Animal products*, *Vitamins and fiber*, *Starch-rich*, and *Vegetable unsaturated fatty acids*. Consistent associations emerged between the identified patterns and gastric cancer, but their significance varied across the three solutions.

Conclusions. All of the solutions were characterized by a fair proportion of total variance explained and had an appealing interpretation. The four-factor solution led to a higher number of retained factors significantly related to gastric cancer, whereas the five-factor solution identified an extra pattern with a consistent interpretation. The six-factor solution may not be adopted, since it showed a pattern based only on a single nutrient.

KEY WORDS: *exploratory factor analysis; dietary patterns; gastric cancer; internal consistency; measures of sampling adequacy; nutritional epidemiology.*

Introduction

Factor analysis, originated in psychometrics, is commonly used in socio-economic studies. The scope of this multivariate statistical method is to describe the variance-covariance structure among a potentially large set of variables in terms of a few underlying, unobservable and randomly varying factors (1).

Single nutrients and foods are consumed in combination and their joint effects may be better investigated by considering the key aspects of the whole eating profile. The application of factor analysis in nutritional epidemiology may allow to identify those dietary patterns that overall represent the whole eating profile of a given population. A set of different nutrients, foods or food groups represents the variables, different subjects are the observations, and the so-called dietary patterns are the factors summarizing dietary information. This is of interest particularly when many dietary components are relevant for a disease.

Nevertheless, factor analysis combines both methodological and subjective aspects, so the investigators have to make many arbitrary decisions.
In this paper, we describe the steps through which we identified nutrient-based dietary patterns in the context of a case-control study of gastric cancer conducted in northern Italy. We also evaluated a potential association between the identified dietary patterns and gastric cancer. We compared results derived from three factor analysis solutions, characterized by a different number of retained factors.

Methods

Design and participants

We derived data from a case-control study of gastric cancer conducted between 1997 and 2007 in the Greater Milan area, Italy (2). Briefly, cases were 230 patients (143 men and 87 women; median age 63 years, range 22-80 years), admitted to major teaching and general hospitals in the study area with incident, histologically confirmed gastric cancer (ICD IX, 151.0-151.9), diagnosed no longer than 1 year before the interview, and with no previous diagnosis of cancer. The control group included 547 patients (286 men and 261 women; median age 63 years, range 22-80 years) frequency matched to cases by age and sex (with a ratio of 2:1 for men, and 3:1 for women), admitted to the same hospitals as cases for a wide spectrum of acute, non neoplastic conditions, unrelated to known or potential risk factors for gastric cancer and long term diet modification.
For both cases and controls, data were collected during their hospital stay by centrally trained interviewers. The questionnaire included information on socio-demographic characteristics, anthropometric measures, selected lifestyle habits, such as tobacco smoking and alcohol consumption, a personal medical history and a family history of cancer. A satisfactorily reproducible (3) and valid (4) food frequency questionnaire (FFQ) was used to assess the patients' usual diet in the two years preceding diagnosis (for cases) or hospital admission (for controls). The FFQ included questions on 78 foods and beverages, including a range of the most common recipes in Italian diet. Subjects were asked to indicate the average weekly raw frequency and corresponding portion size (small, medium, large) of consumption for each dietary variable. To estimate micro- and macro-nutrients, an Italian food composition database was used, integrated with other sources, when needed (5, 6).

Factorability of the original matrix

We identified a preliminary set of 28 variables, selected among micro- and macro-nutrients for potential application of factor analysis. Throughout the literature a strong correlation among nutrients is widely documented (7).
We calculated the correlation matrix R of the original data to assess its factorability. We checked variables that were:

1) too highly correlated ($r \geq 0.80$); this reflects problems of multicollinearity, so that one or more of these variables would be dropped from the analysis;
2) not sufficiently correlated ($r < 0.30$) with one another; this means these variables will not share much of the common variance, thus potentially leading to solutions with as many factors as variables.

Then, we evaluated matrix factorability through statistical procedures. Measures of sampling adequacy that compare the simple and partial correlation coefficients may be defined either overall or for single variables. The overall measure, called Kaiser-Meyer-Olkin statistic (*KMO*), is defined as follows (8):

$$KMO = \frac{\sum\sum_{i \neq j} r_{ij}^2}{\sum\sum_{i \neq j} r_{ij}^2 + \sum\sum_{i \neq j} a_{ij}^2}$$

where $\sum\sum$ are the sum over all variables in the matrix when variable $i \neq$ variable j, r_{ij} is the Pearson correlation coefficient between i and j, and a_{ij} the partial correlation coefficient between i and j. Individual measures of sampling adequacy are computed using only the simple and partial correlation coefficients involving the specific variable under consideration. The overall and individual measures range between 0 and 1. Smaller values indicate that the squared pearson correlation coefficient is small relative to the squared partial correlation coefficient and therefore a factor analysis may be imprudent. If the sum of the squared partial correlation coefficients is small compared with the sum of the squared correlation coefficients, the measures approach 1. Bartlett's test of sphericity tests the null hypothesis that the correlation matrix is an identity matrix. It is a chi-square test (8), whose statistic is defined as follows:

$$\chi^2 = -\left[(N-1)-\left(\frac{2k+5}{6}\right)\right]\log|R|$$

where χ^2 is the calculated chi-square value for Bartlett's test, N is the sample size, k is the number of variables in the matrix and $|\boldsymbol{R}|$ the determinant of the correlation matrix. The degrees of freedom for this chi-square statistic are $k(k-1)/2$. Larger values of the test suggest that the null hypothesis should be rejected.

Since Bartlett's test statistic depends explicity on the sample size, N, for larger samples this test tends to indicate that the correlation matrix is not an identity matrix. For this reason, it should be used only as a minimum standard for assessing the quality of the correlation matrix. We performed this test using the statistical software R (9).

Identification of factors through factor analysis

We applied exploratory principal component factor analyses (PCFA) on the overall original dataset. The analyses were conducted using the PROC FACTOR procedure, provided by SAS software, version 9.1 (SAS Institute, Inc., Cary, NC). This approach assumes that the variables included in the analysis can be calculated by the extracted components or factors. Because each standardized variable has a mean of 0 and variance of 1, the initial estimate of communality for each variable is 1. This is what will be placed initially on the diagonal of the correlation matrix. The first principal component is a linear combination of the original variables, such that it explains the maximum amount of the variance among the variables. After the first extraction, a residual correlation matrix is created. This matrix contains the variances not explained by the first factor on the diagonal and the partial correlations of the variables with each other after extracting the first factor on the off-diagonal. The second one is extracted from this residual matrix, so it will be uncorrelated to the first one. This process of extracting principal components is repeated on subsequent residual matrices, until the elements in the residual variance-covariance matrix are reduced to random error.

Choosing the number of factors to retain

A crucial aspect of factor analysis is the choice of the number of factors to retain. The choice was based on three main criteria. The first one is to retain those factors with eigenvalues greater than 1. In SAS, we choose the option MINEIGEN=1, so that the unity represents the smallest eigenvalue for which a factor is retained. The second criterion is to add successive factors until the cumulative percentage of variance explained by the retained factors is satisfactory. To terminate the factor extraction process, we considered 75-80% to be a valid threshold for the cumulative variance extracted. The third one, suggested by Cattell (10), is to plot, by the option SCREE in SAS, the extracted factors against their eigenvalues in descending order of magnitude to identify distinct breaks in the slope of the plot, called "scree plot". To determine where the break occurs, a straight line should be drawn with a ruler through the lower values of the plotted eigenvalues. That point where the factors curve above the straight line drawn through the smaller eigenvalues identifies the optimal number of factors to retain.

Finally, to determine the number of factors to retain, a researcher should also consider factor interpretability. In nutritional epidemiology, the identified factors represent potentially uncorrelated dietary habits that, considered altogether, summarize the overall dietary profile of a given population.

In the following we will present results coming from three different solutions: four-, five- and six-factor. We will compare them in terms of eigenvalues, explained variances, scree plot, and factor interpretation.

Estimating factor scores

Factor scores were estimated for each subject and factor. They indicate the degree to which each subject's diet conforms to one of the identified factors (11). In the main analysis we calculated them using the weighted least square method, where variables that have lower loadings on the factor are given less weight than those with higher loadings, in the calculation of factor scores.

Rotating the identified factors

To improve the interpretation of the generated factors, suggestions have been made to rotate them. If a rotation is not performed, the first unrotated factor is most often a general factor on which most variables load highly in absolute value. The rotation consists in turning the reference axes of the factors about their origin to achieve a simple structure where variables should load highly (in absolute value) on one factor only, and each factor should have high absolute loadings only on some of the variables.

There are two types of rotation: *orthogonal* and *oblique*. In the first one, pairs of axes are kept at right angles (90°) to one another during rotation, so that they

are still uncorrelated after rotation. In the second one, each axis may be rotated independently, so that they are not necessarily perpendicular after rotation.
We preferred forms of *orthogonal* rotation. This is a crucial aspect in nutritional epidemiology, where one may deal with severe multicollinearity problems. Another property of *orthogonal* rotation is that the amount of total variance accounted for by the factors under consideration is unaffected by the rotation itself (12).
In detail, we performed a varimax rotation that consists in rotating the axes to orientations that maximize variances of the loadings within the factors, while maximizing differences between the high and low loadings on a particular factor.

Naming the identified factors

To name the identified factors, we considered only those ones having factor loadings greater or equal to |0.63| on a given factor. The contribution that a factor gives to a nutrient's sample variance is equal to the square of its loading on that factor, so if we choose a |0.63| cut-off, we expect a minimum contribution of the factor on the nutrient's variance of approximately 0.40 (11).

Evaluating the identified solution

To determine the internal consistency of the identified factors we considered those variables having rotated factor loading greater or equal to |0.40| on any factor, and we examined Cronbach's coefficient alphas (α) (8, 13). We also calculated standardized *Cronbach's coefficient alpha when variable deleted* for each factor and for each variable. This measure of reliability represents the proportion of total variance in a given scale that can be attributed to a common source.
The general formula for α is given as follows:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum \sigma_i^2}{\sum \sigma_i^2 + 2\sum \sigma_{ij}}\right)$$

where k is number of variables, $\sum\sigma_i^2$ is the sum of the variances of the variables and $\sum\sigma_{ij}$ is the sum of the covariances of all possible pairs of variables.
When the variances of the items are all equal, the formula for standardized coefficient is given as follows:

$$\alpha = \frac{k\bar{r}_{ij}}{1 + (k-1)\bar{r}_{ij}}$$

where $\bar{r}_{ij}$ is the average correlation among the k variables. Values for α should range between 0 and 1. If there is little correlation among the variables, α will equa l to 0. The higher the correlation among the variables, the higher will be the value of α. Its value is influenced not only by the size of the correlation among the variables but also by the number of variables in the set. Indeed, increasing the number of variables will increase the size of α, even when the correlations among the variables are small.
To evaluate the robustness of the factors identified with PCFA, we carried out a series of checks on the identified solution. First, we applied another solution method for factor analysis, specifically principal axis factoring (PAF). Briefly, it consists in adopting the squared multiple correlation coefficients of each variable with all the other ones as an estimate of the initial communality. Then, the analysis is undertaken in the same way as that outlined for PCFA. This approach gave essentially the same results as PCFA. Second, we performed subgroup analysis by sex and quinquennia of period of interview. Third, we calculated factor scores applying the multiple regression method, as follows:

$$\hat{F}_{ij} = \sum_{k=1}^{p} W_{jk} z_{jk}$$

$\hat{F}_{ij}$ = estimated standardized score for respondent i on factor j
W_{jk} = factor score coefficient for variable k on factor j and standardizing the results (11).
The correlations between scores referring to the same factor calculated with different methods were 0.99 for all the comparisons. Fourth, to confirm internal reproducibility of the identified factors, individuals were randomly placed into one of two equally sized groups, or split-samples, and factor analysis was performed separately in both split-samples using the same approach of the main analysis. Each split sample contained cases selected by chance together with the corresponding matched controls. We also performed a stability analysis, where split samples were created separately within quinquennia of period of interview.

Risk estimate

For each solution (four-, five- and six-factor), participants were grouped into 4 categories according to quartiles of factor scores calculated among the control population. We estimated the odds ratios (OR) and the corresponding 95% confidence intervals (CI) for each quartile, using conditional multiple regression models

(14) conditioned on age and sex, and adjusted for quinquennia of period of interview, education, body mass index, tobacco smoking and family history of stomach cancer. We fitted a composite model allowing for all the identified factors simultaneously. We also fitted separate models for each factor and obtained comparable results. Tests for linear trend were also calculated assigning to each subject the median value of each factor within the quartile class.

Results

Identification of dietary patterns

The correlation matrix of the selected nutrients resulted to be amenable to factor analysis (15). From visual inspection of the matrix, all the nutrients showed at least ten correlation coefficients above 0.30 in absolute value, with retinol and vitamin D having a more limited number of correlations above 0.30 in absolute value. The KMO statistic was equal to 0.82, suggesting that we have a good sample size relative to the number of nutrients. The individual measures of sampling adequacy were generally very high, suggesting that overall the correlations among the individual nutrients were strong enough to proceed with a factor analysis. Accordingly, Bartlett's test was highly significant (p-value<0.0001). We then decided to carry out the analyses on the entire set of originally selected nutrients. Given the satisfactory results obtained by the complementary checks, we carried out a factor analysis with principal component method on the overall sample and

Table 1. Factor loading matrix[1], communalities and explained variances from principal component factor analysis: four-factor solution.

Nutrient	Factor 1	Factor 2	Factor 3	Factor 4	Communality
Animal protein	**0.80**	0.10	0.41	0.23	0.873
Vegetable protein	0.15	0.39	0.29	**0.80**	0.902
Cholesterol	**0.72**	-	0.41	0.30	0.780
Saturated fatty acids	0.56	0.15	0.50	0.41	0.758
Monounsaturated fatty acids	0.20	0.29	**0.72**	0.28	0.730
Linoleic acid	0.19	0.16	**0.71**	0.33	0.677
Linolenic acid	0.33	0.27	**0.68**	0.34	0.763
Other polyunsaturated fatty acids	0.48	-	**0.75**	-	0.791
Soluble carbohydrates	0.40	**0.66**	-	0.17	0.625
Starch	0.18	0.11	0.26	**0.88**	0.892
Sodium	0.41	-	0.16	**0.80**	0.846
Calcium	**0.65**	0.34	-	0.28	0.622
Potassium	0.42	**0.76**	0.29	0.28	0.908
Phosphorus	**0.70**	0.37	0.31	0.45	0.922
Iron	0.42	0.48	0.39	0.37	0.705
Zinc	**0.63**	0.29	0.45	0.47	0.913
Thiamin	0.53	0.51	0.30	0.45	0.827
Riboflavin	**0.76**	0.47	0.10	0.26	0.877
Vitamin B6	0.53	0.58	0.41	0.29	0.871
Total folate	0.40	**0.71**	0.22	0.28	0.796
Niacin	0.54	0.37	0.47	0.21	0.693
Vitamin C	0.12	**0.85**	0.13	-0.11	0.763
Retinol	0.47	-	-	-	0.223
Beta-carotene equivalents	-	**0.67**	0.20	-	0.494
Lycopene	-	0.26	0.49	0.32	0.417
Vitamin D	0.54	-	0.54	-0.23	0.635
Vitamin E	-	0.53	**0.74**	0.22	0.877
Total fiber	-	**0.85**	0.15	0.31	0.845
Proportion of variance explained (%)	21.67	20.30	18.02	15.10	
Cumulative variance explained (%)	21.67	41.97	59.99	75.09	

[1] *Loadings greater or equal to 0.63 (in absolute value) were shown in bold typeface; loadings smaller than 0.10 (in absolute value) were not shown.*

Table 2. Factor loading matrix[1], communalities and explained variances from principal component factor analysis: five-factor solution.

Nutrient	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Communality
Animal protein	0.11	**0.63**	0.21	0.32	0.57	0.878
Vegetable protein	0.36	0.15	**0.85**	0.24	0.11	0.950
Cholesterol	-	0.55	0.30	0.32	0.53	0.780
Saturated fatty acids	0.13	0.62	0.28	0.59	0.24	0.880
Monounsaturated fatty acids	0.27	0.19	0.24	**0.75**	0.22	0.777
Linoleic acid	0.14	0.23	0.26	**0.77**	0.16	0.757
Linolenic acid	0.25	0.37	0.25	**0.75**	0.20	0.871
Other polyunsaturated fatty acids	-	-	0.11	0.47	**0.78**	0.850
Soluble carbohydrates	**0.65**	0.42	0.13	-	0.11	0.634
Starch	-	0.21	**0.90**	0.22	-	0.926
Sodium	-	0.49	**0.75**	0.19	0.10	0.846
Calcium	0.31	**0.83**	-	0.21	-	0.837
Potassium	**0.76**	0.31	0.32	0.22	0.30	0.913
Phosphorus	0.36	**0.63**	0.41	0.29	0.38	0.928
Iron	0.49	0.22	0.48	0.23	0.45	0.767
Zinc	0.29	0.47	0.51	0.34	0.50	0.922
Thiamin	0.50	0.47	0.44	0.26	0.31	0.827
Riboflavin	0.46	**0.71**	0.21	-	0.35	0.889
Vitamin B6	0.59	0.33	0.37	0.27	0.48	0.894
Total folate	**0.71**	0.31	0.32	0.15	0.27	0.804
Niacin	0.39	0.18	0.38	0.20	**0.68**	0.834
Vitamin C	**0.85**	-	-	0.12	-	0.763
Retinol	0.10	0.20	0.12	-0.17	0.47	0.310
Beta-carotene equivalents	**0.67**	-	-	0.23	-	0.503
Lycopene	0.27	-0.23	0.46	0.33	0.27	0.517
Vitamin D	-	0.10	-	0.25	**0.79**	0.713
Vitamin E	0.51	-	0.18	**0.78**	0.14	0.929
Total fiber	**0.84**	-	0.34	0.15	-	0.850
Proportion of variance explained (%)	19.93	15.50	15.32	14.61	14.46	
Cumulative variance explained (%)	19.93	35.43	50.75	65.36	79.82	

[1] Loadings greater or equal to 0.63 (in absolute value) were shown in bold typeface; loadings smaller than 0.10 (in absolute value) were not shown.

calculated factor scores applying the weighted least squares method.

Tables 1, 2, 3 present the factor loading matrix for the four-, five-, and six-factor solutions, respectively. The cumulative percentages of variance explained by these factor solutions were approximately equal to 75%, 80% and 84%, respectively, with single factors accounting for a minimum of about 6% to a maximum of about 22% of the total variance. The communalities generally indicate that the retained factors account for a large percentage of the sample variance of each variable. Retinol, beta-carotene equivalents and lycopene nutrients presented the lowest values of communalities across the three factor solutions. Retinol reached a satisfactory value in the six-factor solution, in accordance with the presence of a sixth factor where it loaded highly.

The larger the loading of a given nutrient to the factor, the greater the contribution of that nutrient was on that factor. In each table, all the examined nutrients showed at least one loading greater than 0.30 on any dietary pattern, thus confirming a role for each nutrient in the original set.

Table 4 compared the three selected solutions, providing names for the identified factors in terms of those nutrients having factor loadings greater or equal to |0.63|. The *Animal products* pattern was characterized by a consistent core of selected nutrients given by calcium, riboflavin, animal protein and phosphorus (five-factor solution), integrated by cholesterol and zinc, in the four-factor solution, and by saturated fatty acids, in the six-factor solution. The *Vitamins and fiber* pattern was consistent across the three solutions, with the greatest loadings on

Table 3. Factor loading matrix[1], communalities and explained variances from principal component factor analysis: six factor solution.

Nutrient	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Communality
Animal protein	0.11	**0.69**	0.21	0.25	0.55	0.10	0.903
Vegetable protein	0.35	0.17	**0.85**	0.24	-	-	0.950
Cholesterol	-	0.58	0.29	0.31	0.45	0.25	0.782
Saturated fatty acids	0.11	**0.64**	0.28	0.57	0.23	-	0.880
Monounsaturated fatty acids	0.26	0.19	0.23	**0.76**	0.25	-	0.796
Linoleic acid	0.12	0.22	0.25	**0.79**	0.19	-	0.785
Linolenic acid	0.23	0.37	0.24	**0.77**	0.21	-	0.892
Other polyunsaturated fatty acids	-	0.14	0.13	0.37	**0.83**	-	0.863
Soluble carbohydrates	**0.66**	0.45	0.14	-	0.10	-	0.669
Starch	-	0.23	**0.90**	0.23	-	-	0.927
Sodium	-	0.50	**0.74**	0.19	-	-	0.848
Calcium	0.31	**0.85**	-	0.19	-	-	0.862
Potassium	**0.76**	0.36	0.33	0.19	0.30	-	0.931
Phosphorus	0.35	**0.67**	0.41	0.26	0.34	0.11	0.939
Iron	0.48	0.23	0.47	0.25	0.36	0.29	0.779
Zinc	0.28	0.50	0.51	0.31	0.45	0.18	0.924
Thiamin	0.49	0.51	0.45	0.24	0.27	-	0.836
Riboflavin	0.44	**0.69**	0.18	0.17	0.18	0.40	0.923
Vitamin B6	0.58	0.37	0.37	0.24	0.45	0.15	0.899
Total folate	**0.69**	0.28	0.30	0.24	0.13	0.39	0.869
Niacin	0.39	0.23	0.39	0.16	**0.63**	0.24	0.836
Vitamin C	**0.85**	-	-	0.11	0.10	-	0.767
Retinol	-	-	-	-	0.11	**0.95**	0.930
Beta-carotene equivalents	**0.66**	-	-	0.27	-	-	0.515
Lycopene	0.27	-0.18	0.48	0.25	0.36	-0.11	0.543
Vitamin D	0.10	0.19	-	0.10	**0.85**	-	0.789
Vitamin E	0.50	-	0.18	**0.79**	0.19	-	0.948
Total fiber	**0.84**	0.10	0.35	0.15	-	-	0.856
Proportion of variance explained (%)	19.36	16.77	15.27	13.93	12.80	5.60	
Cumulative variance explained (%)	19.36	36.13	51.40	65.33	78.13	83.73	

[1] *Loadings greater or equal to 0.63 (in absolute value) were shown in bold typeface; loadings smaller than 0.10 (in absolute value) were not shown.*

vitamin C, total fiber, potassium, total folate, beta-carotene equivalents and soluble carbohydrates. The *Starch-rich* pattern was also consistent across the three solutions, with the greatest loadings on starch, vegetable protein and sodium. The *Vegetable unsaturated fatty acids (VUFA)* pattern was characterized by a consistent core of selected nutrients given by vitamin E, linoleic acid, linolenic acid and monounsaturated fatty acids, with other polyunsaturated fatty acids included in the four-factor solution only, and in the successive solutions included in the *Animal unsaturated fatty acids* (*AUFA) pattern.* The *AUFA pattern* was consistent across the two solutions where it was present, with the greatest loadings on vitamin D, other polyunsaturated fatty acids and niacin. Finally, the *Retinol* pattern, selected in the six-factor solution only, had only one nutrient with a loading greater or equal to 0.63, that is retinol. Table 5 gives the values of the standardized Cronbach's coefficient alpha for each factor. They were calculated considering only those nutrients that loaded greater or equal to |0.40| on any factor. Standardized Cronbach's coefficient alphas were generally very high, indicating that more than 90% of the variance of the total scores on these subscales for each factor can be attributed to reliable, systematic variance. Only the *Retinol* pattern in the six-factor solution showed a modest coefficient alpha of about 0.67. Moreover, most standardized *coefficient alpha when item deleted*, were lower than the corresponding standardized coefficient alpha for the factor, although the differences were generally limited (data not shown). These findings indicate that almost all of the nutrients were contributing

Table 4. Description of the identified dietary patterns through nutrients with absolute loadings greater or equal to 0.63: four-, five- and six-factor solutions.

	Four-factor solution		*Five-factor solution*		*Six-factor solution*	
	Dietary pattern Proportion of variance explained (%)	**Nutrients**	**Dietary pattern** Proportion of variance explained (%)	**Nutrients**	**Dietary pattern** Proportion of variance explained (%)	**Nutrients**
Pattern 1	Animal products (21.67)	animal protein riboflavin cholesterol phosphorus calcium zinc	Vitamins and fiber (19.93)	vitamin C total fiber potassium total folate beta-carotene equivalents soluble carbohydrates	Vitamins and fiber (19.36)	vitamin C total fiber potassium total folate beta-carotene equivalents soluble carbohydrates
Pattern 2	Vitamins and fiber (20.30)	vitamin C total fiber potassium total folate beta-carotene equivalents soluble carbohydrates	Animal products (15.50)	calcium riboflavin animal protein phosphorus	Animal products (16.77)	calcium animal protein riboflavin phosphorus saturated fatty acids
Pattern 3	VUFA[1] (18.02)	other polyunsaturated fatty acids vitamin E monounsaturated fatty acids linoleic acid linolenic acid	Starch-rich (15.32)	starch vegetable protein sodium	Starch-rich (15.27)	starch vegetable protein sodium
Pattern 4	Starch-rich (15.10)	starch vegetable protein sodium	VUFA[1] (14.61)	vitamin E linoleic acid linolenic acid monounsaturated fatty acids	VUFA[1] (13.93)	vitamin E linoleic acid linolenic acid monounsaturated fatty acids
Pattern 5			AUFA[1] (14.46)	vitamin D other polyunsaturated fatty acids niacin	AUFA[1] (12.80)	vitamin D other polyunsaturated fatty acids niacin
Pattern 6					Retinol (5.60)	retinol
Cumulative variance explained (%)	75.09		79.82		83.73	

[1] *VUFA: Vegetable unsaturated fatty acids; AUFA: Animal unsaturated fatty acids.*

Table 5. Standardized Cronbach's coefficient alpha for each factor: four-, five-, and six-factor solutions.

	Four-factor solution		*Five-factor solution*		*Six-factor solution*	
	Dietary pattern	**Standardized alpha**	**Dietary pattern**	**Standardized alpha**	**Dietary pattern**	**Standardized alpha**
Pattern 1	Animal products	0.961	Vitamins and fiber	0.948	Vitamins and fiber	0.948
Pattern 2	Vitamins and fiber	0.948	Animal products	0.960	Animal products	0.954
Pattern 3	VUFA[1]	0.947	Starch-rich	0.936	Starch-rich	0.936
Pattern 4	Starch-rich	0.948	VUFA[1]	0.926	VUFA[1]	0.938
Pattern 5			AUFA[1]	0.925	AUFA[1]	0.935
Pattern 6					Retinol	0.665

[1] *VUFA: Vegetable unsaturated fatty acids; AUFA: Animal unsaturated fatty acids.*

Table 6. Odds ratios (OR)[1] of gastric cancer and corresponding 95% confidence intervals (CI) on quartiles of factor scores from principal component factor analysis: four-, five-, six-factor solutions.

	Quartile category, OR (95% CI)				
Dietary pattern	**I[2]**	**II**	**III**	**IV**	**p_{trend}[3]**
Four-factor solution					
Animal products	1	1.08 (0.64-1.80)	1.47 (0.90-2.40)	2.13 (1.34-3.40)	0.0003
Vitamins and fiber	1	0.84 (0.53-1.32)	1.00 (0.64-1.56)	0.60 (0.37-0.99)	0.0861
VUFA[4]	1	0.84 (0.53-1.34)	0.89 (0.56-1.42)	0.89 (0.56-1.42)	0.7325
Starch-rich	1	1.37 (0.83-2.25)	1.37 (0.82-2.28)	1.67 (1.01-2.77)	0.0463
Five-factor solution					
Vitamins and fiber	1	0.99 (0.63-1.55)	0.99 (0.63-1.56)	0.69 (0.42-1.14)	0.1625
Animal products	1	1.30 (0.80-2.11)	1.01 (0.62-1.65)	1.50 (0.93-2.40)	0.1246
Starch-rich	1	1.09 (0.66-1.82)	1.55 (0.96-2.52)	1.47 (0.89-2.42)	0.0557
VUFA[4]	1	0.92 (0.58-1.45)	0.76 (0.48-1.22)	0.71 (0.44-1.14)	0.1396
AUFA[4]	1	1.43 (0.88-2.31)	1.03 (0.62-1.71)	1.81 (1.13-2.90)	0.0302
Six- factor solution					
Vitamins and fiber	1	1.10 (0.70-1.72)	0.90 (0.56-1.44)	0.76 (0.45-1.27)	0.2312
Animal products	1	1.19 (0.73-1.94)	1.25 (0.77-2.04)	1.61 (1.01-2.57)	0.0448
Starch-rich	1	1.11 (0.66-1.86)	1.62 (0.99-2.64)	1.61 (0.98-2.65)	0.0673
VUFA[4]	1	1.22 (0.77-1.93)	0.91 (0.57-1.46)	0.77 (0.47-1.26)	0.2028
AUFA[4]	1	1.38 (0.84-2.26)	1.33 (0.82-2.18)	1.69 (1.04-2.76)	0.1401
Retinol	1	1.23 (0.75-2.02)	1.30 (0.76-2.20)	1.47 (0.92-2.36)	0.0300

[1] Estimates from conditional multiple logistic regression models conditioned on age and sex and adjusted for quinquennia of interview, education, body mass index, tobacco smoking and family history of gastric cancer. Results refer to the composite model including all the four factors simultaneously.
[2] Reference category.
[3] p-value for linear trend.
[4] VUFA: Vegetable unsaturated fatty acids; AUFA: Animal unsaturated fatty acids.

to a high reliability, and none of the nutrients appreciably reduced the value of coefficient alpha when removed from the factor.

Gastric cancer risk estimate

A relevant use of the dietary patterns obtained by the three different factor solutions is in the estimation of a potential association between them and gastric cancer risk.

Table 6 gives the OR and the corresponding CI for gastric cancer according to quartiles of factor scores for the four- five- and six-factor solutions, respectively. Results refer to the composite model including all the four patterns simultaneously, and the same confounding variables as well. In the four-factor solution, we observed an increased risk of gastric cancer with the *Animal products* pattern (OR=2.13, 95% CI: 1.34, 3.40, for the highest *vs* the lowest quartile of factor score; *p*-value for trend: 0.0003), and the *Starch-rich* one (OR=1.67, 95% CI: 1.01, 2.77; *p*-value for trend: 0.046). In contrast, the *Vitamins and fiber* pattern was inversely associated with gastric cancer, with an OR of 0.60 (95% CI: 0.37, 0.99; *p*-value for trend: >0.05), and the *VUFA* pattern was not significantly inversely related to it. Considering the five-factor solution, there was a significant positive association between the *AUFA* pattern and gastric cancer risk (OR=1.81, 95% CI: 1.13, 2.90; *p*-value for trend: 0.03). Also the *Animal products* (OR=1.50, 95% CI: 0.93, 2.40; *p*-value for trend: >0.05) and the *Starch-rich* patterns (OR=1.47, 95% CI: 0.89, 2.42; *p*-value for trend: >0.05) were positively related to gastric cancer risk, but these associations were not significant. For the *Vitamins and fiber* and *VUFA* patterns, no significant inverse association emerged. The six-factor solution showed that the *Animal products* and the *AUFA* patterns were significantly positively related to gastric cancer, with OR=1.61, 95% CI: 1.01, 2.57 and OR=1.69, 95% CI: 1.04, 2.76, respectively. No significant association emerged for the *Vitamins and fiber*, *Starch-rich*, *VUFA* and *Retinol* patterns. Test for trend was significant for the *Animal products* pattern (*p*-value=0.04)

and for the *Retinol* pattern (p-value=0.03). Consistent results were observed for the single models including one dietary pattern at a time.

Conclusions

The application of factor analysis in nutritional epidemiology has become increasingly popular in the last fifteen years, as a way to overcome both conceptual and methodological problems inherent to the definition of diet as an exposure measure. In the current paper, we described the steps through which we applied this method to a set of nutrients derived from an Italian case-control study of gastric cancer. There are many decisions that must be taken in any factor analysis. Probably the most important one is the choice of the number of factors to retain. Most often, this final choice is based on some combination of factor eigenvalue greater or equal to 1, proportion of sample variance, and factor interpretability. The simultaneous use of more than one criterion may avoid specifying a higher number of factors to retain.

In the current paper, we presented three alternative solutions, with four- five- and six-factors. All of them explained a fair proportion of the total variance of the original nutrients and had an appealing interpretation. In the six-factor solution, we retained a pattern based only on a single nutrient. This could be a reason not to choose this solution.

We also estimated the association between dietary patterns and gastric cancer risk. The *Animal products* pattern was positively related to gastric cancer in all the three solutions, although it was not statistically significant in the five-factor one. The *Vitamins and fiber* and *Starch-rich* patterns were inversely associated with gastric cancer, but the ORs were significant only in the four-factor solution. The VUFA pattern was not significantly associated to gastric cancer in any factor solution. There was a significant positive association with the *AUFA* pattern in the five- and six-factor solutions. The four-factor solution led to a higher number of retained factors significantly related to gastric cancer, whereas the five-factor solution identified an extra pattern, the AUFA one, with consistent interpretation.

Reproducibility and validity of an identified factor solution need to be assessed, though strongly influenced by unsolved methodological issues. Confirmatory factor analysis is used when the researcher has some knowledge about the underlying structure of the construct under investigation, when he wants to compare factor structures across studies, and to test specific theories or hypotheses concerning the linear structural relationships among a set of factors (8,16). It may be used in combination with exploratory factor analysis (EFA) to test the utility of the underlying dimensions of a construct identified through EFA. This is the purpose of future analyses.

Acknowledgements

This work was conducted with contribution from the Italian Association for Cancer Research (AIRC), the Italian League Against Cancer and the Italian Ministry of Education (PRIN 2007). Paola Bertuccio was supported by a fellowship from the Italian Foundation for Cancer Research (FIRC). The authors thank Ms. I. Garimoldi for editorial assistance.

References

1. Johnson R, and Wichern D. Applied multivariate statistical analysis. Upper Saddler River, NJ: Prentice Hall, 2002.
2. Lucenteforte E, Scita V, Bosetti C, Bertuccio P, Negri E, La Vecchia C. Food groups and alcoholic beverages and the risk of stomach cancer: a case-control study in Italy. Nutr Cancer 2008;60:577-84.
3. Franceschi S, Barbone F, Negri E, et al. Reproducibility of an Italian food frequency questionnaire for cancer studies. Results for specific nutrients. Ann Epidemiol 1995;5:69-75.
4. Decarli A, Franceschi S, Ferraroni M, et al. Validation of a food-frequency questionnaire to assess dietary intakes in cancer studies in Italy. Results for specific nutrients. Ann Epidemiol 1996;6:110-8.
5. Gnagnarella P, Parpinel M, Salvini S, Franceschi S, Palli D, Boyle P. The update of the Italian food composition database. J Food Comp Analysis 2004;17:509-522
6. Salvini S, Parpinel M, Gnagnarella P, Maisonneuve P, Turrini A. Banca dati di composizione degli alimenti per studi epidemiologici in Italia. Milano, Italia: Istituto Europeo di Oncologia; 1998
7. Willett W. Nutritional Epidemiology. Oxford University Press, 1998, 2nd edition.
8. Pett MA, Lackey NR, and Sullivan JJ. Making sense of factor analysis: the use of factor analysis for instrument development in health care research. CA: Sage, 2003.
9. R Development Core Team: a language and environment for statistical computing. R Fundation for Statistical Computing, Vienna, Austria, 2009. Available at URL: http://www.R-project.org. 2009.

10. Cattell RB. The scree test for the number of factors. Multivariate Behavioral Research. 1966;1:245-76.
11. Comrey AL, Lee HB. A first course in factor analysis. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers, 1992, 2nd edition.
12. Kleinbaum DG, Kupper LL, Muller KE, and Nizam A. Applied Regression Analysis and Other Multivariable Methods. Duxbury Press, 1998.
13. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951;16:297-334.
14. Breslow NE, and Day NE. Statistical methods in cancer research. Vol. I. The analysis of case-control studies. IARC Sci Publ No. 32. Lyon, France: IARC, 1980.
15. Bertuccio P, Edefonti V, Bravi F, Ferraroni M, Pelucchi C, Negri E, Decarli A, La Vecchia C. Nutrient dietary patterns and gastric cancer risk in Italy. Cancer Epidemiol Biomarkers Prev 2009;18:2882-6.
16. Tatsuoka MM. Multivariate Analysis. New York: John Wiley & Sons, Inc. 1971.