Assessment of dietary patterns in nutritional epidemiology: principal component analysis compared with confirmatory factor analysis

Raphaëlle Varraso, Judith Garcia-Aymerich, Florent Monier, Nicole Le Moual, Jordi De Batlle, Gemma Miranda, Christophe Pison, Isabelle Romieu, Francine Kauffmann, and Jean Maccario

ABSTRACT

Background: In the field of nutritional epidemiology, principal component analysis (PCA) has been used to derive patterns, but the robustness of interpretation might be an issue when the sample size is small. The authors proposed the alternative use of confirmatory factor analysis (CFA) to define such patterns.

Objective: The aim was to compare dietary patterns derived through PCA and CFA used as equivalent approaches in terms of stability and relevance.

Design: PCA and CFA were performed in 2 different studies: the Epidemiological Study on the Genetics and Environment of Asthma 2–France (EGEA2-France; n = 1236) and the Phenotype and Course of Chronic Obstructive Pulmonary Disease study–Spain (n = 274). To check for stability, PCA and CFA were also performed in 2 subsamples from the EGEA2 study (n = 618 and 309). Statistical properties were evaluated by 1000 bootstrapped random sets of observations for each of the 4 subsamples. For each random set of observations, the distribution of the factor loading for each pattern was obtained and represented by using box-plots. To check for relevance, partial correlations between different nutrients and the different patterns derived by either PCA or CFA were calculated.

Results: With the use of CFA, 2 consistent dietary patterns were derived in each subsample (the Prudent and the Western patterns), whereas dietary factors were less interpretable with the use of PCA (smaller median of factor loadings and higher dispersion), especially for the smallest subsample. Higher correlations were reported among total fiber, vitamins, minerals, and total lipids with patterns derived by using CFA than with patterns derived by using PCA.

Conclusion: The current study shows that CFA may be a useful alternative to PCA in epidemiologic studies, especially when the sample size is small. Am J Clin Nutr 2012;96:1079–92.

INTRODUCTION

The assessment of diet in nutritional epidemiology is a complex issue (1). Because we eat meals instead of isolated foods or nutrients, several authors have proposed assessing dietary patterns to have a broader picture of diet. Exploratory approaches based on statistical dimension-reduction methods have been widely used to derive dietary patterns (2). The most popular dimension-reduction method was to derive dietary patterns from principal component analysis (PCA)4, which groups correlated food groups into uncorrelated factors, ie, dietary patterns (3–7). Another method is cluster analysis, which clusters either individuals (8) or food groups (9, 10). All of these exploratory factor analyses do not require a hypothesis and involve several arbitrary decisions, including the construction of the food groups, the number of factors or clusters to retain, the method of computation, and even the labeling of the components (2–7). More recently, it has been proposed to derive dietary patterns by using other exploratory methods, such as reduced rank regression (11, 12) or partial least-squares regression (12) and also to combine the strength of PCA and cluster analysis by using the treelet transform dimension-reduction method (13). Despite the differences in the goal of these methods, the methods are similar regarding their mathematical foundation.

Structural equation modeling has been widely used in social and psychiatric areas to specify latent variables associated with psychological measurements (observed variables) (14–16). According to the type of observed and latent variables (qualitative or quantitative), different statistical methods have been proposed...
[latent classes, latent profiles, latent traits, and confirmatory factor analysis (CFA)]. Even if structural equation modeling is intuitively appealing, because it is based on theory and also reduces some of the subjectivity involved in the exploratory procedures, very few studies have used CFA in nutritional epidemiology, and up to now, it has been used only to examine the robustness and goodness-of-fit of the factor structures derived from the PCA (17–21). None of these studies have considered deriving patterns by using CFA in which all foods load on all factors (such as in PCA) and comparing patterns obtained by each method.

The aim of this article was to improve our knowledge regarding the use of CFA in nutritional epidemiology to derive dietary patterns. Our purpose was to use CFA in the same context as PCA (ie, to define a model comparable with the PCA by using all of the food groups available, by specifying that latent variables might depend on all of the food groups) and to compare PCA and CFA used as equivalent approaches. Because most of the published studies looking at dietary patterns were large surveys, including several thousands of participants, we also wondered about the relevance of PCA in small studies (22). Therefore, dietary patterns were derived by using both PCA and CFA in small epidemiologic surveys and by using all of the food groups available. Then, dietary patterns obtained by CFA were compared with those obtained by PCA, in terms of stability (assessed by bootstrap) and relevance.

SUBJECTS AND METHODS

Studies

The analyses were performed by using 2 different studies from different countries and with different food habits: the French EGEA2 (Epidemiologic Study on the Genetics and Environment of Asthma 2) and the Spanish PAC-COPD study (Phenotype and Course of Chronic Obstructive Pulmonary Disease). Briefly, EGEA2, a cross-sectional study including 1601 men and women, is the 12-y follow-up of the EGEA study—a case-control and family asthma study (23–25). None of these studies have considered deriving patterns by using CFA in which all foods load on all factors (such as in PCA) and comparing patterns obtained by each method.

The FFQ was administered by 2 trained interviewers over the telephone; therefore, all the items from the FFQ were filled in for all the participants (no missing data). A Spanish food-composition table of the Centre for Superior Studies in Nutrition and Dietetics was used to estimate nutrient intake from each food in the FFQ. Nutrients were derived by using the same method used in EGEA2.

To prepare for PCA and CFA, and because both methods can break down in the presence of outliers or missing data, participants in EGEA2 who left >4 items blank on the diet questionnaire were excluded (n = 60; 4%). For participants with 1–3 blank items [198 participants (14%) with 1 blank item, 94 (7%) with 2 blank items, and 34 (2%) with 3 blank items], each missing value was replaced by the median of the food item in the remaining population. To prepare for PCA and CFA, the theory of both methods assume multivariate normality, foods from the FFQ were classified into a priori food groups based on nutrient profiles or culinary use (Table 1). Foods were grouped into 46 food groups for EGEA2 and into 43 food groups for PAC-COPD (sandwiches, sorbet, and water were not available in PAC-COPD). This classification follows that of the Etude épidémiologique auprès des femmes de la Mutuelle Générale de l’Education Nationale (E3N) study—the French part of the European Prospective Investigation into Cancer and Nutrition (EPIC) study, with slight modifications (30). Foods that did not fit into any of the groups or that may have represented distinctive dietary behaviors were left as individual categories (eg, French fries, tea, and wine). Each type of alcoholic beverage (beer, cider, red wine, white wine, alcohol) was analyzed separately as a food group in factor analysis.

Dietary assessment

In EGEA2, dietary intake information was collected by using a 118-item food-frequency questionnaire (FFQ) designed to assess average food intake during the previous 12 mo. It was set up based on a French validated dietary questionnaire (28). The participants indicated their average frequency of consumption for the 118-item FFQ over the past year in terms of the specified serving size by checking 1 of 8 frequency categories ranging from “never or <1 time/month” to “≥4 times/day.” Standard portion sizes were listed with each food. The selected frequency category for each food item was converted to a daily intake. For example, a response of “1 serving/week” was converted to 0.14 servings/d. Through use of the French food-composition data from the Vitamin and Antioxidant Element Supplementation Study (SU.VI.MAX) survey (29), the average daily intake of nutrients was calculated by multiplying the frequency of consumption of each item by its nutrient content per serving and totaling the nutrient intake for all food items. In PAC-COPD, dietary intake information was collected by using a 122-item FFQ designed to assess average food intake during the previous 2 y (27). The FFQ was administered by 2 trained interviewers over the telephone; therefore, all the items from the FFQ were filled in for all the participants (no missing data). A Spanish food-composition table of the Centre for Superior Studies in Nutrition and Dietetics was used to estimate nutrient intake from each food in the FFQ. Nutrients were derived by using the same method used in EGEA2.

Statistical analysis

For exploratory analysis, PCA was conducted by using the procedure Factor in SAS (version 9; SAS software). We used all the food groups in our models (ie, 46 for EGEA2 and 43 for PAC-COPD). In nutritional epidemiology, the most used method to derive dietary pattern is PCA with varimax rotation; therefore, the factors were rotated by an orthogonal transformation (varimax rotation function in SAS) to enhance the difference between loadings, which allowed easier interpretability. The number of factors to retain was determined by using the diagram of eigenvalues (the Scree plot) and the interpretability of the factors. The Scree plots are presented in the online supplement (see supplemental Figure S1 under “Supplemental data” in the online issue). The factor score for each pattern was constructed by summing up observed intakes of the component food items weighted by the factor loading.
<table>
<thead>
<tr>
<th>Food groups (abbreviation)</th>
<th>Items in EGEA2 (France)</th>
<th>Items in PAC-COPD (Spain)</th>
</tr>
</thead>
<tbody>
<tr>
<td>High-fat dairy products (HFat Dai)</td>
<td>Whole milk, high-fat cottage cheese, yogurt, milky cream dessert, fromage frais, goat cheese, hard cheese, creamy cheese, blue cheese, Roquefort, sour cream, chocolate milk, ice cream</td>
<td>Milkshakes, whole milk, condensed milk, whole yogurt, creamy cheese, hard cheese, blue cheese, milky cream dessert, sour cream, ice cream, chocolate milk</td>
</tr>
<tr>
<td>Low-fat dairy products (LFatDai)</td>
<td>Low-fat milk, skim milk, fat-free cottage cheese, organic yogurt, probiotic yogurt</td>
<td>Low-fat milk, skim milk, fat-free yogurt, probiotic yogurt</td>
</tr>
<tr>
<td>Refined-grain cereals (RefiCer)</td>
<td>White bread, cereal, corn flakes, pasta, white rice, semolina, wheat, other cereals</td>
<td>White bread, dry bread, white rice, white pasta</td>
</tr>
<tr>
<td>Whole-grain cereals (WholCer)</td>
<td>Whole bread, whole rice</td>
<td>Whole bread</td>
</tr>
<tr>
<td>Cakes (Cakes)</td>
<td>Pastry, cake, fruit pie</td>
<td>Pastry, doughnuts, butter cakes (mantecado)</td>
</tr>
<tr>
<td>Sandwiches (Sandwich)</td>
<td>Cheese sandwich, processed meat sandwich, tuna sandwich</td>
<td>—</td>
</tr>
<tr>
<td>Citrus fruit (CitrFrt)</td>
<td>Oranges, grapefruit, mandarin, kiwi</td>
<td>Oranges, grapefruit, mandarin, kiwi</td>
</tr>
<tr>
<td>Fruit with flavonoids (FlavFrt)</td>
<td>Apples, pears, plums, cherries, strawberries, raspberries, grapes</td>
<td>Apples, pears, cherries, strawberries, grapes</td>
</tr>
<tr>
<td>Fruit with β-carotene (BetaFrt)</td>
<td>Melons, peaches, apricots, nectarines</td>
<td>Bananas, baked fruit, figs</td>
</tr>
<tr>
<td>Other fruit (OthrFrt)</td>
<td>Bananas, baked fruit, pineapple</td>
<td>Cooked spinach, Swiss chard, watercress, lettuce, raw endive</td>
</tr>
<tr>
<td>Leafy vegetables (LeafyVg)</td>
<td>Salad, raw endive, watercress, raw spinach, cooked endive, cooked spinach</td>
<td>Onion, squash, carrots, beetroot</td>
</tr>
<tr>
<td>Dark-yellow vegetables (DrYeVeg)</td>
<td>Raw carrots, cooked carrots, raw/cooked onion</td>
<td>White fish, fried fish</td>
</tr>
<tr>
<td>Cruciferous vegetables (CruciVg)</td>
<td>Cabbage salad, cabbage</td>
<td>White fish, fried fish</td>
</tr>
<tr>
<td>Other vegetables (OtherVg)</td>
<td>Other raw cudités, vegetable soup, other vegetables, leek</td>
<td>White fish, fried fish</td>
</tr>
<tr>
<td>Fruity vegetables (FruityVg)</td>
<td>Tomato salad, avocado (one-half), zucchini, eggplant, cooked tomatoes, tomato sauce, ratatouille, green beans</td>
<td>Zucchini, cucumber, eggplant, green beans, tomatoes, gazpacho, tomato sauce</td>
</tr>
<tr>
<td>Pods and peas (Pods)</td>
<td>Green peas, corn</td>
<td>Green peas</td>
</tr>
<tr>
<td>Legumes (Legumes)</td>
<td>Soy, lentils, chickpeas</td>
<td>Lentils, red beans, white beans, chickpeas, broad beans</td>
</tr>
<tr>
<td>Potatoes (Potatoes)</td>
<td>Potatoes</td>
<td>Mashed potatoes, potatoes</td>
</tr>
<tr>
<td>French fries (FrFries)</td>
<td>French fries</td>
<td>Mashed potatoes, potatoes</td>
</tr>
<tr>
<td>Eggs (egg)</td>
<td>Eggs</td>
<td>Mashed potatoes, potatoes</td>
</tr>
<tr>
<td>Red meat (RedMeat)</td>
<td>Beef, lamb, veal, pork</td>
<td>Beef, veal, pork, lamb, goat</td>
</tr>
<tr>
<td>Poultry (Poultry)</td>
<td>Poultry</td>
<td>Poultry, rabbit</td>
</tr>
<tr>
<td>Offal (Offal)</td>
<td>Liver, offal</td>
<td>Liver, offal</td>
</tr>
<tr>
<td>Processed meat (CuredMe)</td>
<td>Paté, raw or cooked ham, sausage, dried and cured sausage (salami)</td>
<td>Bacon, raw or cooked ham, serrano ham, sausage, Frankfurt sausage, dried and cured sausage (chorizo, salami), paté/fouge gris</td>
</tr>
<tr>
<td>White fish (WhiFish)</td>
<td>White fish, fried fish</td>
<td>White fish, fried fish</td>
</tr>
<tr>
<td>Blue fish (BluFish)</td>
<td>Blue fish, smoked fish, canned tuna/sardine</td>
<td>Blue fish, surimi seafood, smoked fish, canned fish</td>
</tr>
<tr>
<td>Shellfish (ShlFish)</td>
<td>Seashell, shellfish</td>
<td>Seashell, shellfish</td>
</tr>
<tr>
<td>Prepared meal (Ready)</td>
<td>Fish soup, hamburger, ravioli, lasagna, pizza, salty pie, paella, Cassoulet (meat and bean stew), couscous, sauerkraut, pot-au-feu</td>
<td>Ravioli, lasagna, ready-to-eat soup, pizza</td>
</tr>
<tr>
<td>Butter (Butter)</td>
<td>Butter, diet butter, margarine, diet margarine</td>
<td>Butter, margarine, lard</td>
</tr>
<tr>
<td>Sorbet (Sorbet)</td>
<td>Sorbet</td>
<td>—</td>
</tr>
<tr>
<td>Condiments (Condimt)</td>
<td>Mayonnaise, ketchup, prepared salad dressing, raw/cooked garlic</td>
<td>Mayonnaise, salt, garlic, spicy condiments</td>
</tr>
<tr>
<td>Olive oil (OlivOil)</td>
<td>Olive oil</td>
<td>Olive oil</td>
</tr>
<tr>
<td>Other oil (OtherOil)</td>
<td>Other oils</td>
<td>Other oils (corn, soya, sunflower)</td>
</tr>
<tr>
<td>Fruit juice (FrtJuic)</td>
<td>Orange or grapefruit juice, grape juice, apple juice</td>
<td>Orange juice (fresh or canned), tomato juice, horchata</td>
</tr>
<tr>
<td>Soda (Soda)</td>
<td>Soda</td>
<td>Soda</td>
</tr>
<tr>
<td>Water (Water)</td>
<td>Water</td>
<td>—</td>
</tr>
<tr>
<td>Beer/cider (Beer)</td>
<td>Beer, cider</td>
<td>Beer, cider</td>
</tr>
<tr>
<td>Red wine (RedWine)</td>
<td>Red wine</td>
<td>Red wine</td>
</tr>
<tr>
<td>White wine (WhiWine)</td>
<td>White wine</td>
<td>White wine</td>
</tr>
<tr>
<td>Alcohol (Alcohol)</td>
<td>Long drink (whiskey, gin, vodka) liquor</td>
<td>Champagne, long drink (whiskey, gin, vodka) liquor</td>
</tr>
<tr>
<td>Coffee (Coffee)</td>
<td>Coffee</td>
<td>Coffee</td>
</tr>
<tr>
<td>Tea (Tea)</td>
<td>Tea</td>
<td>Tea, herbal tea</td>
</tr>
<tr>
<td>Snack (Snack)</td>
<td>Cookies and scones, chips</td>
<td>Cookies and scones (with or without chocolate)</td>
</tr>
<tr>
<td>Chocolate (Chocolt)</td>
<td>Chocolate bars</td>
<td>Chocolate, nougat (turron)</td>
</tr>
<tr>
<td>Honey/jam (Honey)</td>
<td>Honey, jam</td>
<td>Sugar, jam</td>
</tr>
<tr>
<td>Nuts and seeds (Nuts)</td>
<td>Peanuts, walnuts, almonds, dry fruit</td>
<td>Peanuts, walnuts, almonds, dry fruit, olives</td>
</tr>
</tbody>
</table>

1 EGEA, Epidemiological Study on the Genetics and Environment of Asthma; PAC-COPD, Phenotype and Course of Chronic Obstructive Pulmonary Disease.
CFA was used for its ability to specify latent variable models (15) by using the procedure Calis in SAS (version 9). In our cases, latent variables (constructs) were dietary patterns and measures were food group intakes assessed via their variance-covariance. Because the objective was to set a minimal model, we tested only 2 models: a 3-factor model (ie, with 3 latent variables) and a 2-factor model (ie, with 2 latent variables). CFA also offers the possibility to define a correlation between latent variables. It is a nice modeling approach for dietary patterns that are not necessarily independent between each other. Therefore, 4 models were defined: a 3-factor model with latent variables correlated, a 3-factor model with independent latent variables, a 2-factor model with latent variables correlated, and a 2-factor model with independent latent variables. On the basis of both measures of global fit (chi-square, goodness-of-fit index, and root mean square error of approximation) and on the relevance and interpretability of dietary patterns compared with the literature, we retained a 2-factor model without correlation between latent variables (see supplemental Table S1 under “Supplemental data” in the online issue). As for PCA, we used all of the food groups.

To assess the stability of these approaches (PCA and CFA), we used 4 different subsamples: 1) with the initial size of the EGEA2 survey (n = 1236), 2) with half the size of the EGEA2 survey (n = 618), 3) with one-quarter of the size of the EGEA2 survey (n = 309), and 4) with the initial size of the PAC-COPD survey (n = 274). Statistical proprieties (min, quartile 1, median, quartile 3, max) were evaluated by 1000 bootstrapped random sets of observations for each of the 4 subsamples. For each random set of observations, the distribution of the factor loading for each pattern was obtained and represented by using box-plots. This distribution was represented by using box-plots, and foods that loaded with a median of 0.74. Therefore, the first pattern was loaded by a high intake of vegetables, fruit, oil, legumes, and fish and was labeled the “Prudent” pattern; the second factor was loaded by a high consumption of prepared meals, French fries, processed meats, condiments, alcohol, beer/cider, sandwiches, potatoes, pods and peas, cake, condiments, high-fat dairy products, and potatoes and was labeled the “Western” pattern; and the third factor was loaded by a high intake of alcoholic beverages and a low intake of low-fat dairy products and was labeled the “Alcohol and Wine” pattern.

Stability of the approaches
Among the first sample: EGEA2 survey

Among the 1000 randomly selected samples from EGEA2 (n = 1236), 3 distinct major dietary patterns were identified by using PCA (Figure 1A). The distribution of the 3 factor scores (loadings of each food groups) is presented in Figure 1A, where each box-plot represented the distribution of the loading of each food group to each factor, for each of the 1000 bootstraps. For example, the correlation for fruity vegetables to the first factor (first box-plot of Figure 1A) ranged from 0.32 to 0.97 with a median of 0.74. Therefore, the first pattern was loaded by a high intake of vegetables, fruit, oil, legumes, and fish and was labeled the “Prudent” pattern; the second factor was loaded by a high consumption of prepared meals, French fries, processed meats, sandwiches, snack, soda, pods and peas, cakes, condiments, high-fat dairy products, and potatoes and was labeled the “Western” pattern; and the third factor was loaded by a high intake of alcoholic beverages and a low intake of low-fat dairy products and was labeled the “Alcohol and Wine” pattern.

For CFA, the first factor was loaded by a high intake of vegetables, fruit, oils, whole-grain cereals, and fish and was labeled the “Prudent” pattern (Figure 1B), and the second factor was loaded by a high consumption of prepared meals, French fries, processed meats, condiments, alcohol, beer/cider, sandwiches, potatoes, pods and peas, snack, soda, cakes, red meats, high-fat dairy products, nuts and seeds, offal, shellfish, sorbet, high-fat dairy products, coffee, fruit juice, refined cereals, butter, chocolate, and red wine and was labeled the “Western” pattern.

### RESULTS

#### Characteristics of participants

Participants from EGEA2 were 51% men and 49% women and were middle-aged; ~25% of the participants were ex-smokers, ~50% had a university-level education, and 39% were overweight.

#### Table 2

<table>
<thead>
<tr>
<th></th>
<th>EGEA2 (n = 1236)</th>
<th>PAC-COPD (n = 274)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex, male (%)</td>
<td>51</td>
<td>93</td>
</tr>
<tr>
<td>Age (y)</td>
<td>43 ± 16</td>
<td>68 ± 8</td>
</tr>
<tr>
<td>Smoking habits (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Never smokers</td>
<td>49</td>
<td>1</td>
</tr>
<tr>
<td>Ex-smokers</td>
<td>28</td>
<td>66</td>
</tr>
<tr>
<td>Current smokers</td>
<td>23</td>
<td>33</td>
</tr>
<tr>
<td>Educational level (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Primary</td>
<td>24</td>
<td>87</td>
</tr>
<tr>
<td>Secondary</td>
<td>27</td>
<td>9</td>
</tr>
<tr>
<td>University</td>
<td>49</td>
<td>4</td>
</tr>
<tr>
<td>BMI (kg/m²)</td>
<td>24.5 ± 4.3</td>
<td>28.3 ± 4.6</td>
</tr>
<tr>
<td>Energy intake (kcal/d)</td>
<td>2416 ± 653</td>
<td>2026 ± 610</td>
</tr>
</tbody>
</table>

1. EGEA, Epidemiological Study on the Genetics and Environment of Asthma; PAC-COPD, Phenotype and Course of Chronic Obstructive Pulmonary Disease.

2. Mean ± SD (all such values).
Among the second sample: EGEA2 survey

Among the 1000 randomly selected samples of half of the population \((n = 618)\), 4 major dietary patterns were identified by using PCA (Figure 2A). The first factor was loaded by a high intake of vegetables, oil, and fish and was labeled the “Vegetables, Oil, and Fish” pattern; the second factor was loaded by a high consumption of prepared meals, French fries, processed meats, sandwiches, snack, soda, cakes, pods and peas, beer, condiments, high-fat dairy products, and fruit juice and was labeled the “Western” pattern; the third factor was loaded by a high intake of alcoholic beverages, shellfish, and coffee and was labeled the “Alcohol” pattern; and the fourth factor was loaded by a high intake of fruit and therefore was labeled the “Fruit” pattern.

For CFA, the first factor was loaded by a high intake of vegetables, fruit, oils, whole-grain cereals, and fish and was labeled the “Prudent” pattern (Figure 2B), and the second factor was loaded by a high consumption of prepared meals, French fries, processed meats, condiments, alcohol, sandwiches, potatoes, pods and peas, snack, soda, cakes, beer/cider, high-fat dairy products, red meats, sorbet, nuts and seeds, coffee, fruit juice, refined cereals, butter, chocolate, and red wine and was labeled the “Western” pattern.

Among the third sample: EGEA2 survey

Among the 1000 randomly selected samples of 25% of the population \((n = 309)\), 3 major dietary patterns were identified by using PCA (Figure 3A). The first factor was loaded by a high intake of vegetables, oil, and fruit and was labeled the “Vegetables, Oil, and Fruit” pattern; the second factor was loaded by a high consumption of prepared meals, French fries, processed meats, sandwiches, soda, snack, cakes, beer/cider, pods and peas, and condiments and was labeled the “Western” pattern; and the third factor was loaded by a high intake of alcoholic beverages and was labeled the “Alcohol” pattern.

For CFA, the first factor was loaded by a high intake of vegetables, fruit, oils, whole-grain cereals, and fish and was labeled the “Prudent” pattern (Figure 3B), and the second factor was loaded by a high consumption of prepared meals, French fries, processed meats, condiments, alcohol, sandwiches, potatoes, legumes, poultry, pods and peas, snack, soda, cakes, beer/cider, high-fat dairy products, red meats, sorbet, nuts and seeds,
offal, shellfish, coffee, fruit juice, egg, refined cereals, butter, chocolate, and red wine and was labeled the “Western” pattern.

Among the fourth sample: PAC-COPD survey

Among the 1000 randomly selected samples of the population \((n = 274)\), 2 major dietary patterns were identified by using PCA (Figure 4A). The first factor was loaded by a high intake of other oils, fruity vegetables, red meats, offal, cured meats, and potatoes and was labeled the “Vegetables and Meats” pattern, and the second factor was mainly loaded by a high intake of leafy vegetables and low-fat dairy products and was labeled the “Leafy Vegetables and Low-Fat Dairy” pattern.

For CFA, the first factor was loaded by a high intake of fruity vegetables, other vegetables, blue fish, leafy vegetables, white fish, other oil, red meats, pods and peas, and dark-yellow vegetables and was labeled the “Prudent” pattern (Figure 4B); the second factor was loaded by a high intake of high-fat dairy products, chocolate, potatoes, soda, snack, nuts and seeds, butter, and refined cereal and a low intake of low-fat dairy products and citrus fruit and was labeled the “Western” pattern.

Consistency of the approaches

Most of the selected nutrients were more correlated with patterns derived by using CFA (F1 \(_{\text{CFA}}\) and F2 \(_{\text{CFA}}\)) than with the patterns derived by using PCA (F1 \(_{\text{PCA}}\) and F2 \(_{\text{PCA}}\)). The size of the box-plots was smaller for factors derived by using CFA than for the ones derived by using PCA, and correlations for CFA were more stable across the 4 subsamples than were correlations for PCA.

For each of the 4 subsamples, positive correlations were found between F1 \(_{\text{CFA}}\) and F1 \(_{\text{PCA}}\) with total fiber, water (Figure 5A), PUFAs, linoleic acid, EPA (Figure 5B), \(\beta\)-carotene, vitamin B-6, folic acid, vitamin C, vitamin E (Figure 5C), iron, potassium, magnesium, and phosphorus (Figure 5D). The highest correlations were reported for total fiber, vitamins (C, B-6, E, and folic acid), \(\beta\)-carotene, and minerals (potassium, magnesium, and iron). The median correlation between total fiber and F1 \(_{\text{CFA}}\) was 0.83 in the first subsample \((n = 1236), 0.82\) in the second subsample \((n = 618), 0.82\) in the third subsample \((n = 309), and 0.64 in the fourth subsample \((n = 274)\); corresponding values for F1 \(_{\text{PCA}}\) were 0.75, 0.48, 0.53, and 0.24, respectively (Figure 5A).

For folic acid, the median correlation with F1 \(_{\text{CFA}}\) was 0.84 in the first subsample \((n = 1236), 0.83\) in the second subsample \((n = 618), 0.82\) in the third subsample \((n = 309), and 0.69 in the fourth subsample \((n = 274)\); corresponding values for F1 \(_{\text{PCA}}\) were 0.74, 0.50, 0.54, and 0.25, respectively (Figure 5C).

For the 3 subsamples from EGEA2, negative correlations were found between F2 \(_{\text{CFA}}\) and F2 \(_{\text{PCA}}\) with total fiber, water (Figure 5D).
vitamin B-6, folic acid, vitamin C, vitamin E (Figure 6C), calcium, iron, potassium, magnesium, and phosphorus (Figure 6D) and positive correlations with total lipids, SFAs, cholesterol (Figure 6B), and sodium (Figure 6A). For the fourth subsample ($n = 274$), the correlations between total fiber, water, vitamin B-6, folic acid, vitamin C, vitamin E, calcium, iron, potassium, magnesium, and phosphorus with F2CFA were negative, whereas they were positive with F2PCA. The same inversion was reported for total lipids, SFAs, and cholesterol.

**DISCUSSION**

We reported in both French and Spanish populations that dietary patterns derived by using CFA were more meaningful in terms of stability and relevance than were dietary patterns derived by using PCA. With CFA, we characterized 2 dietary patterns—Prudent and Western—consistent with patterns already reported in the literature, which correlated more with nutrient intakes than patterns derived with PCA. Our hypothesis was that CFA, used in the same context as PCA by using all of the food groups available and by specifying that latent variables might depend on all of the food groups, would be more appropriate for deriving dietary patterns than would PCA, especially when the sample size is small. Although it was not the classical use of CFA, we thought it to be an original and interesting approach, allowing comparisons between PCA and CFA. The terminology of CFA (ie, confirmatory) is also misleading because it is usually used as a second step to confirm dietary patterns derived by using PCA. In our case, we propose to use CFA as a first step to define dietary patterns.

The 2 methods (PCA and CFA) differ from the statistical point of view regarding their objective, method, and estimation. PCA is a descriptive method for data reduction, whereas CFA is a statistical model that can be tested. PCA can be applied when one knows little or nothing about the data, whereas CFA requires some previous knowledge, either theoretically or empirically derived. With PCA, factors are always uncorrelated, whereas with CFA, factors may be correlated or not. With PCA, components are linear combinations of observed variables, whereas with CFA, factors are linear combinations of unobservable variables. Regarding the method of estimation, singular value decomposition is used for PCA, whereas CFA is based on the likelihood ratio test. To our knowledge, there is no statistical theory perspective for why the CFA results in much smaller CIs by using bootstrapping than does PCA. Sample sizes in factor analyses are a contentious area. Indeed, the use of clustering methods has become widespread in the context of high-dimensional
low–sample size data, such as in microarray data analysis, where the ratio of the number of subjects to the number of variables is often as low as 0.01 (31). It has been suggested that the minimum sample size should be dependent on other aspects of design, such as the communality of the variables (32). With the lack of general rules, we decided to perform a quasiempirical analysis.

Some studies have compared solutions from exploratory analyses with those from CFA, but CFA was used only to examine the robustness and goodness of fit of the factor structures derived from the PCA (17–21). Among 3296 participants from the Inter99 study (Danish population), Lau et al (21) derived 2 dietary patterns with the use of PCA: a “Traditional” and a “Modern” pattern, consistent with the so-called Western and Prudent patterns (3, 5, 33). Then, they used CFA to derive patterns but only by using the food groups with loadings \( \geq 0.40 \) from the PCA (ie, only 18 food groups). They reported that loadings in CFA were very similar to those in PCA. Maskarinec et al (19) used the same approach to validate the factor model obtained from PCA with CFA. Among 514 women from Hawaii, they derived 4 dietary patterns with the use of PCA on 39 food groups and then used CFA with the 16 food groups with loadings \( \geq 0.60 \) from the PCA. Togo et al (18), of roughly 2000 participants to the Danish MONICA study, derived 3 dietary patterns in men and 2 in women with the use of PCA; then, confirmatory factor models were tested based on the foods loading \( \geq 0.30 \) in PCA. Newby et al (20) also used this approach, first with PCA they derived 6 dietary patterns among 33,840 Swedish women. They then confirmed patterns that had \( \geq 4 \) items with factor loadings of at least the absolute value of 0.20; thus, 4 of 6 patterns were confirmed. Using this approach, they concluded that the benefit from CFA over exploratory factor analysis in diet studies was not yet clear. Our approach allows a comparison of PCA with CFA, not only from the point of view of repeatability or stability, but also from the point of view of distributional properties of estimates, crudely summarized by box-plots.

### FIGURE 4

Factor-loading matrix for the major factors (dietary patterns), 1000 bootstraps of sample (\( n = 274 \), PAC-COPD Spain, 2004–2007). Each box-plot represents the distribution of the loading of each food group to each factor for each of the 1000 bootstraps by using either PCA (A) or CFA (B) to derive dietary patterns. CFA, confirmatory factor analysis; F1, factor 1; F2, factor 2; PAC-COPD, Phenotype and Course of Chronic Obstructive Pulmonary Disease study; PCA, principal component analysis.
FIGURE 5. Partial Pearson correlations between the first factor of dietary pattern derived by using both PCA and CFA, with selected nutrients, in each of the 4 subsamples (n = 1236, 618, 309, and 274). Each box-plot represented the distribution of partial Pearson correlations between the first factor of dietary pattern derived using PCA (white boxes) and CFA (gray boxes) with total proteins, total lipids, total carbohydrates, total fiber, alcohol, and water (A); SFAs, MUFAs, PUFAs, linoleic fatty acid, EPA, and cholesterol (B); retinol, β-carotene, vitamin B-6, folic acid, vitamin C, and vitamin E (C); and calcium, iron, potassium, magnesium, sodium, and phosphorus (D). CFA, confirmatory factor analysis; FA, fatty acid; F1, factor 1; Monounsatur., monounsaturated; PCA, principal component analysis; Polyunsat., polyunsaturated; Satur., saturated.
To account for country-specific foods and habits, we tested our hypothesis in 2 different surveys, one in France and one in Spain. Most studies that have used FFQs have not described how non-responses to items have been handled, but to assume no consumption (null value imputation) or to impute the median value, or a combination of the 2, appears to be the most common practice after the initial exclusion of subjects (34). In the French survey, because of the small number of missing values to impute,
FIGURE 6. Partial Pearson correlations between the second factor of dietary pattern derived by using both PCA and CFA with selected nutrients, in each of the 4 subsamples ($n = 1236, 618, 309$, and $274$). Each box-plot represented the distribution of partial Pearson correlations between the second factor of dietary pattern derived by using PCA (white boxes) and CFA (gray boxes) with total proteins, total lipids, total carbohydrates, total fiber, alcohol, and water (A); SFAs, MUFAs, PUFAs, linoleic fatty acid, EPA, and cholesterol (B); retinol, β-carotene, vitamin B-6, folic acid, vitamin C, and vitamin E (C); and calcium, iron, potassium, magnesium, sodium, and phosphorus (D). CFA, confirmatory factor analysis; FA, fatty acid; F2, factor 2; Monounsatur., monounsaturated; PCA, principal component analysis; Polyunsat., polyunsaturated; Satur., saturated.
we a priori chose the traditional median value imputation, knowing that it may introduce a bias toward weaker factors. The Spanish study presented the advantage of a small sample size. To reduce subjectivity due to the a priori food grouping, similar food groups were used in both surveys (excepted for 3 food groups not available in the Spanish data). We used all of the food groups (46 in EGEA2 and 43 in PAC-COPD) to derive dietary patterns with PCA and CFA. For CFA, we examined whether or not existing...
data were consistent with a priori hypotheses regarding the number of
dietary patterns (latent variables). Our goal was to set up a
minimal model relevant regarding interpretability and previous
studies. Dietary patterns varied across countries because of dif-
ferent regional and cultural food habits; however, there is con-
 sistency across studies to identify a Prudent pattern and a Western
pattern. To test for stability, the analyses with PCA and CFA were
repeated 1000 times by using bootstrapped sets in 4 subsamples.

PCA raised several methodologic issues and arbitrary de-
 cisions, including the number of factors to extract. The method
most used to retain the numbers of factors in PCA is to use
eigenvalues > 1 (4). As did Togo et al (18), we opted for the
diagram of eigenvalues, ie, the Scree plot, because if we had
opted for eigenvalues > 1, a high number of factors not easily
interpretable would have been retained. In the PAC-COPD
study, in which the sample size was small (n = 274), 2 dietary
patterns were derived by using PCA, and the labeling of patterns
was not really easy because of the small number of food groups
included in each pattern. Previous studies have faced the same
issue (35). By using data from the Baltimore Longitudinal Study
of Aging, Newby et al (35) derived 6 dietary patterns among 459
participants. Their “Eggs, Bread and Soup” pattern was loaded
by a high intake of eggs, white bread, and refined grains; mis-
cellular fats and soups; and a low intake of potatoes. This
pattern included various foods with different potential health
benefits (36–40). In this particular case, when the sample size
was small and when the number of factors was high and not
easily labeled, CFA offered a nice alternative for deriving di-
etary patterns.

To test for the relevance of patterns, we assessed the corre-
lation between dietary patterns and some nutrients. This method
has often been used to test the validity and reproducibility of
patterns (21, 35, 41, 42). The Prudent and the Western pattern
derived from CFA correlated more with nutrients (higher cor-
relation, smaller variability, and higher stability across sub-
samples) than did the Prudent or the Western pattern derived from
PCA. Each of the derived patterns was related differently with
nutrients intake, which suggested that our patterns factor solution
yielded different meaningful patterns. The highest correlation
between our Prudent pattern derived by using CFA with nutrients
was for fiber intake; this correlation was stronger than the one
reported in previous studies between their so-called Prudent
pattern and fiber (17, 35, 41, 42). We also reported a strong
correlation between our Prudent pattern from CFA with vitamins
(folic acid, C, B-6, and E), β-carotene, and minerals (potassium,
magnesium, and iron). These correlations were similar to those
reported in the Swedish Mammography Cohort between the
“Healthy” pattern and folate, vitamin D, and β-carotene (21),
but were higher than those reported in other studies (35, 41, 42).
Regarding our Western pattern derived with CFA, the highest
positive correlation was with total lipids. This correlation was of
a magnitude similar to that reported in the Health Professionals
Follow-Up Study in the United States (33) for their “Western”
pattern and to those reported in the Baltimore Longitudinal
Study of Aging for their “Fatty Meats” pattern (36), but higher
than those reported in other studies (21, 41, 42).

In conclusion, CFA brought more stable factors, with less
variability than those derived with PCA. Our data support the use
of CFA to derive patterns by using all food groups available,
especially when the sample size is small.

We thank Serge Hercberg, who provided us the nutrients database from
the SU.VI.MAX survey; Katelle Joly, Anne Forhan, and Peggy Drouillet for com-
putation of the nutrients in the EGEA2 survey; and Jeremie Botton for his help
with statistical issues. We also thank all those who participated in the setting
of the 2 studies (EGEA and PAC-COPD), the EGEA Cooperative Group, and
the individuals who participated. EGEA Cooperative Group: coordination
[F Kauffmann, V Siouxs (epidemiology), F Demenais (genetics), I Pin (clin-
ical aspects), and R Nadif (biology). Respiratory epidemiology: M Korobaffen
(EGEA1) and F Neukirch (EGEA1) (Inserm U 700, Paris); I Annesi-Maesano
(Inserm U 707, Paris); F Kauffmann, N Le Moual, R Nadif, and MP Oryszczyn
(Inserm CESPU 1018, Villejuif); and V Siouxs (Inserm U 823, Grenoble).
Genetics: J Feingold (Inserm U 393, Paris); E Bouzigon, F Demenais, and
MH Dzierz (Inserm U 946, Paris); I Gut (CNG, Evry; now CNAG); and M
Lathrop (CNG, Evry; now CEPH/McGill). Clinical centers: I Pin and C Pison
(Grenoble); D Ecocard (EGEA1), F Gormand, and Y Pacheco (Lyon); D
Charpin (EGEA1) and D Vervloet (Marseille); J Bousquet (Montpellier);
A Lockhart (EGEA1) and R Matran (now in Lille) (Paris Cochon); E Pavy
and P Scheimann (Paris Necker); and A Grimfeld and J Just (Paris-Trousseau).

Data and quality management: J Hochez (EGEA1) (Inserm ex-U155), N Le Moual
(Inserm CESPU U 1018, Villejuif), C Ravault (Inserm ex-U780), N Chatteigner
(Inserm ex-U794), and J Ferran (Grenoble).

The authors’ responsibilities were as follows—RV: data collection, study
conception and planning, statistical programming, data analysis, data inter-
pretation, primary manuscript preparation, and funding; JG-A and NLM:
data collection, statistical expertise, and data interpretation; FM and GM:
refinement of dietary pattern exposures and data interpretation; JDB: data
collection, refinement of dietary pattern exposures, and data interpretation;
CP and IR: data collection and interpretation; FK: data collection, data in-
terpretation, and funding; and JM: study conception and planning, statistical
programming, data analysis, statistical expertise, data interpretation, and
funding. All authors contributed to the drafting of the report and approved
the final version. RV received a grant from the French Society of Nutrition
for the present work. CP received a grant, gave expert testimony, and
received payment for lectures from Nutricia Europe. None of the other authors
reported a conflict of interest.

REFERENCES
2. Imamura F, Jacques PF. Invited commentary: dietary pattern analysis.
3. Newby PK, Tucker KL. Empirically derived eating patterns using factor
4. Hu FB. Dietary pattern analysis: a new direction in nutritional epide-
5. Kant AK. Dietary patterns and health outcomes. J Am Diet Assoc
2004;104:615–35.
7. Slattery ML. Analysis of dietary patterns in epidemiological research.
statistical approaches to diminish subjectivity of cluster analysis to
derive dietary patterns: the tomorrow project. Am J Epidemiol 2011;
Dietary patterns in middle-aged Irish men and women defined by
10. Smith AD, Emmett PM, Newby PK, Northstone K. A comparison of
3 methods for identifying dietary patterns associated