METHODOLOGICAL PAPER

Food or nutrient pattern assessment using the principal component analysis applied to food questionnaires. Pitfalls, tips and tricks

Cristian Ricci, Jeannine Baumgartner, Edelweiss Wentzel-Viljoen and Cornelius M. Smuts
Centre of Excellence for Nutrition (CEN), North-West University, Potchefstroom, South Africa

ABSTRACT
We considered the Blom’s transformation, a statistical method aimed to normalise and standardise food intakes before principal component analysis. A simulation study was performed to evaluate the eigenvalue distribution of a correlation matrix under common conditions in food questionnaire analysis. The scree plot visual inspection and the Guttman–Kaiser (GK) criterion were compared to Horn’s parallel analysis to evaluate their efficacy in food pattern identification. The scree plot results as a monotone continuous series when no food patterns are present. In this situation, about 50% of the eigenvalues assume a value higher than one, showing a first fallacy of the GK. When three food patterns are simulated a clear discontinuity appears after the third eigenvalue, showing that the scree-plot visual inspection is a suitable method to identify food patterns. Finally, according to the present work it appears that the GK generates a number of false-positive food patterns.

ARTICLE HISTORY
Received 19 June 2018
Revised 12 December 2018
Accepted 14 December 2018

KEYWORDS
Principal component analysis; foods/nutrients pattern recognition; food intake questionnaires; scree plot visual inspection; Guttman–Kaiser criterion

Introduction
Principal component analysis (PCA) is a multivariate technique aimed to synthesise the information carried by a multivariate set of correlated variables to a lower subspace of independent, or orthogonal, variables commonly named as principal component or latent factors (Hair et al. 1998; Johnson and Wichern 2014). The PCA is performed by the factorisation of the covariance or the correlation matrix by spectral decomposition, so that the resulting principal components are the eigenvectors of the correlation or covariance matrix of the original set of data. When the PCA is applied to data from food intake questionnaires, for example a quantified food frequency questionnaire, the relation between food or nutrient intakes defines patterns that represent foods or nutrients that are consumed in combination (Hu 2002; Jacobs and Steffen 2003). In nutritional epidemiology, food or nutrient patterns can be used to define clusters of subjects according to a certain nutritional behaviour. Another use of the PCA in nutritional epidemiology is to evaluate the association between a given food or nutrient pattern and the risk of a health outcome (Trichopoulos and Lagiou 2001; Hu 2002).

PCA is nowadays becoming very popular in nutritional research and statistical packages made its use simple. Nevertheless, the PCA is not free from pitfalls (Henson and Roberts 2006).

First, defining the adequate sample size is the first challenge that the researcher meet when applying PCA. Unfortunately, PCA applicability is not merely a matter of statistical power and more factors should be taken into account along with the sample size. According to first pioneering and theoretical papers on PCA, the number of patterns, the number of variables associated to a given pattern, the correlation between variables defining a pattern and variables not associated to any pattern (white noise) play a major role in determining the PCA applicability (Dziuban and Shirkey 1974). Second, the correct application of the PCA is based on the assumption of a multivariate normal distribution of the data (Rummel 1988; Williams et al. 2010). This condition is almost never observed in data deriving from food intake questionnaires, so the data normalisation before the PCA is a critical step. Third, the PCA can be applied to the
correlation or the covariance matrix as well (Hair et al. 1998). Notably, these two approaches have some remarkable differences and it is neither clear nor widely accepted, which method should be preferred (Hair et al. 1998; Henson and Roberts 2006). Finally, once the analysis is performed, the approach for the recognition of the patterns is still a matter of discussion since the commonly used techniques have some remarkable limitations. The visual inspection of the scree plot, a plot of the eigenvalue versus its rank, may carry a certain degree of subjectivity because it is based on the visual recognition of the discontinuity in the trend of the eigenvalues.

On the other side, when the PCA is applied to the correlation matrix the eigenvalues can be retained when assuming a value higher than 1, a commonly accepted rule named the Guttman–Kaiser (GK) criterion. The GK criterion is widely used and even recommended in nutritional research (Moeller et al. 2007). By the way, it was reported that this simple rule may result in an overestimation of the number of patterns, especially in the condition of limiting sample size and weak correlations among the variables determining a pattern (Kaiser 1974; Cerny and Kaiser 1977; Yeomans and Golder 1982). Nevertheless, the GK criterion remains widely used in nutritional research.

The present work aims to highlight the pitfalls of the PCA use in nutrition research and to provide potential solutions on how to overcome these problems. To this aim, we reported about some numerical simulations aimed to disentangle the complex relations between the elements determining the sample size adequacy for PCA. According to these simulations, we proposed a recursive algorithm to determine data adequacy to PCA analysis. Afterwards, the Blom’s transformation (Blom 1958), a statistical method aimed to normalise and standardise food or nutrient intakes before PCA was described along with the advantages of performing PCA on normalised and standardised data. A second simulation study was performed to evaluate the eigenvalue distribution of a correlation matrix, or a covariance matrix of standardised variables, in the presence of simulated patterns under certain conditions that are commonly found in food intake questionnaire analysis. The number of correctly identified patterns according to the GK criterion and the scree plot visual inspection were compared to the empirical threshold defined by the Horn’s parallel analysis. Finally, we evaluated pattern recognition also in conditions of partial to full factoriability of the covariance/correlation matrix according to the Kaiser–Meyer–Olkin (KMO) statistic (Kaiser 1974).

**Methods**

**Data simulation**

The data simulation programme follows three steps. First, the programme performs the starting parameters, a vector of foods or nutrients means and the correlation matrix between foods or nutrients. Then, a set of normally distributed random variables, representing foods or nutrients with the above correlation and means, are simulated using random number generator functions. Finally, the set of foods or nutrients resulting from the previous steps are merged to create a dataset that is analysed to perform the PCA and to compute statistics such as the KMO or eigenvalues. The variables were simulated having a specific correlation structure representing given patterns. First, the variables related to a pattern were simulated as uncorrelated to variables related to any other pattern so the patterns were defined as uncorrelated or orthogonal. Afterwards, residual correlation between variables related to other patterns was introduced to define patterns that are partially correlated. All results were obtained by Monte Carlo studies with 5000 replicates (Mooney 1997). Data simulation was based on the IML procedure of the SAS software vers. 9.4 (Wicklin 2013). The programme to perform simulation was reported on Supplemental File S1.

**Sample size adequacy**

The sample size adequacy for PCA was evaluated using the KMO statistic (Kaiser 1970; Dziuban and Shirkey 1974). Amongst others, the KMO was chosen according to its computational simplicity and because it is popular being performed by all statistical packages of common use. Moreover, the KMO is widely accepted to assess the post-hoc sample adequacy in PCA. Finally, the KMO is preferable to any sphericity test because of its ability to rank sample adequacy gradually, on a numerical scale, instead of on a threshold as for the type-I error given by a sphericity tests. The KMO represents the proportion of variance that is caused by latent factors and is expressed by the ratio between the squared correlations between a set of variables and the sum of those squared correlations with their anti-image squared correlations (Kaiser 1974). The KMO statistics assumes values $\leq 0.5$ (asymptotically 0.5) when the correlation matrix of the variables is null being approximable to an identity matrix.
matrix. This represents the situation where no underly-lying latent factors are present and all variables are uncorrelated (Shirkey and Dziuban 1976). The value of 1 is instead reached when the sum of the anti-image squared correlations are null so that latent factors are perfectly identified. Intermediate values are commonly acknowledged as 0.5–0.59 miserable, 0.6–0.69 mediocre, 0.7–0.79 middling and > 0.8 meritorious (Dziuban and Shirkey 1974; Shirkey and Dziuban 1976).

**Data normalisation before the PCA**

Data deriving from food intake questionnaires are generally skewed and normalisation is a necessary step because the PCA applied to skewed data will result in skewed principal component scores (Rummel 1988). Numerous approaches are commonly used to normalise data. Among these, the logarithmic transformation is the most popular by far. The logarithmic transformation has some remarkable advantages, and it is easy to compute but it is not appropriate in many situations, like, for example, for negatively skewed data. On the other side, the so called inverse rank transformations are a less common class of transformations that efficiently normalise and standardise the data, despite the shape of the distribution of the target variable (Blom 1958). There are at least three common inverse rank transformations, and among others, the Blom’s transformation was here discussed because it is the most common, even if results from other similar transformation are fully overlapping (Conover and Iman 1976). Figure 1 reports about SAS, STATA, R and SPSS codes to perform the Blom’s transformation and its efficacy in normalising skewed variable in comparison to the logarithm. Principal component scores from skewed and Blom’s transformed data was reported in Figure 1.

**Empirical eigenvalues thresholds**

Empirical eigenvalues thresholds were defined according to the 95th centile of the first eigenvalue of a set of uncorrelated normally distributed data obtained by the previously described Monte Carlo simulation. This method is also known as Horn’s parallel analysis and is widely acknowledged among the most reliable and

---

**Figure 1.** Blom’s transformation efficacy in comparison with logarithmic transformation applied to variables and principal component score distribution.
methodological correct (Dinno 2009). A programme to perform Horn’s parallel analysis was provided on Supplemental File S2.

**Food pattern recognition**

The applicability of the PCA and its ability to identify patterns of associations among variables depends on the number of latent factors, the correlation between the variables associated to a latent factor and the sample size (Kaiser 1974; Shirkey and Dziuban 1976; Cerny and Kaiser 1977; Guadagnoli and Velicer 1988; Hogarty et al. 2005). All these determinants of the PCA applicability were set to mimic certain characteristics common in food intake questionnaires and nutritional research. Firstly, the sample size ranged between 150 and 1000 observations in a setting with 75 variables related to three patterns so that intakes of 25 foods or nutrients were simulated to define a pattern. The number of simulated patterns and number of variables related to each pattern were defined according to food intake questionnaires commonly used in nutritional research where, generally, 50–100 foods or nutrients are investigated to define 2–3 patterns. Correlations between variables ranged between 0.2 and 0.6 according to weak to strong correlation between foods or nutrients. The condition of no correlation between foods or nutrients, hence defining the absence of patterns, was also investigated because it portrays a theoretical situation of particular interest (Table 1).

The GK criterion was performed for all simulations and the number of patterns to be retained was reported as the mean of retained patterns among the 5000 simulation runs.

To define an automatic procedure aimed to play as the scree plot visual inspection, a piecewise-like regression analysis was used. Briefly, a regression having the eigenvalue as response and its rank as predictor was performed including a second predictor indicating the position of the discontinuity point in the scree plot. This procedure was conducted for all of the possible positions of the discontinuity point. Finally, the discontinuity point was identified according to the model with the highest R-square. The plots of eigenvalues versus its rank (scree plot) were reported in terms of the sample means calculated over the 5000 simulations (Figure 2). The data adequacy to PCA was reported using the KMO statistic where commonly accepted thresholds were used to define sample adequacy (0.5–0.59 poor, 0.6–0.69 mediocre, 0.7–0.79 middling and > 0.8 meritorious) (Dziuban and Shirkey 1974).

**Supplementary analyses**

Supplementary analyses were conducted to evaluate most likely and more extreme scenarios. First, we evaluated the food pattern recognition in a situation with unequal number of variables associated to a pattern. To this aim, the number of variables associated to the three simulated patterns was set to 40, 25 and 10, respectively, with the correlation between the variables set to 0.2. Afterwards, we evaluated the effect given by the mixed pattern of correlations among variables associated to a pattern. Here, the 25 variables associated to patterns had correlations of 0.4, 0.2 and 0.1, respectively. Afterwards, the situation of spurious correlation between variables associated to different factors was investigated considering the above reported conditions of number of patterns, variables

| Sample size | Pattern correlation | Patterns simulated | Patterns by GK criterion | Patterns by Scree plot* | Patterns by Horn’s analysis |
|-------------|---------------------|--------------------|--------------------------|------------------------|----------------------------|
| 150         | 0                   | 0                  | 27.6                     | 0 (NA)                 | 3 (3)                      |
| 250         | 0                   | 0                  | 26.6                     | 0 (NA)                 | 3 (3)                      |
| 500         | 0                   | 0                  | 25.6                     | 0 (NA)                 | 3 (3)                      |
| 1000        | 0                   | 0                  | 14.3                     | 0 (NA)                 | 3 (3)                      |
| 150         | 0.2                 | 3                  | 20.8                     | 3 (3)                  | 3 (3)                      |
| 250         | 0.2                 | 3                  | 20.0                     | 3 (3)                  | 3 (3)                      |
| 500         | 0.2                 | 3                  | 17.9                     | 3 (3)                  | 3 (3)                      |
| 1000        | 0.2                 | 3                  | 14.3                     | 3 (3)                  | 3 (3)                      |
| 150         | 0.4                 | 3                  | 13.4                     | 3 (3)                  | 3 (3)                      |
| 250         | 0.4                 | 3                  | 10.6                     | 3 (3)                  | 3 (3)                      |
| 500         | 0.4                 | 3                  | 5.6                      | 3 (3)                  | 3 (3)                      |
| 1000        | 0.4                 | 3                  | 3.2                      | 3 (3)                  | 3 (3)                      |
| 150         | 0.6                 | 3                  | 4.2                      | 3 (3)                  | 3 (3)                      |
| 250         | 0.6                 | 3                  | 2.8                      | 3 (3)                  | 3 (3)                      |
| 500         | 0.6                 | 3                  | 2.8                      | 3 (3)                  | 3 (3)                      |
| 1000        | 0.6                 | 3                  | 2.8                      | 3 (3)                  | 3 (3)                      |

*Note: Outcomes assessed by 5000 simulated samples. Patterns were defined by visual inspection of the eigenvalues obtained as mean of the 5000 simulations (Figure 3).

*Patterns defined according to the piecewise regression algorithm proposed.
associated to a pattern and correlation plus a residual correlation of 0.05 between variables associated to different patterns. Finally, the situation of small samples resulting in poor to null factoriability of the data matrix was investigated in the above settings setting where the sample size ranged between 50 and 125 subjects.

**Results**

**Sample size and other factors determining PCA applicability**

A plateau curve linking the sample size to the KMO was observed in all of the simulation settings showing that increasing the sample size is effective in improving PCA applicability. When looking at single determinants of PCA sample adequacy, we observed that reduced number of patterns, increased number of foods or nutrients and increased correlations between foods or nutrients determining a given pattern resulted in higher KMO values.

We observed that in a setting given by three patterns, 75 foods or nutrients (25 foods or nutrients for each pattern) and a correlation of 0.2 between foods or nutrients, a minimum sample size of about 350 individuals is necessary to have a KMO higher than 0.8. On the other hand, when the correlation between foods or nutrients associated to a given pattern is higher, a lower sample is sufficient to achieve an adequate factoriability.

The number of foods or nutrients associated to the pattern should be considered when determining the sample size. According to our simulations, higher numbers of foods or nutrients associated to a pattern improve data factoriability. On the other side, at lower sample sizes, an increase of number of foods or nutrients is counterproductive resulting in a reduction of the KMO due to the approaching to a condition of rank deficiency. Finally, in the case of strong correlations between foods or nutrients ($r \geq 0.4$), smaller sample sizes can be considered.

**Operative recursive procedure to determine the sample size for PCA**

A possible operative procedure to obtain a dataset suitable to be analysed by PCA was reported in Figure S1. The first step is the definition of the starting sample size to perform a preliminary evaluation aimed to determine the correlation between the foods or nutrients associated to a given pattern and, in case of missing a priori knowledge about the questionnaire
structure, the number of patterns, the number of foods or nutrients associated to a pattern, and eventually the number of foods or nutrients determining the white noise. This starting sample size could be quantified as 
\[ N = \frac{1.5}{C_0} \left( \frac{2}{C^2} k \right) \]
where 
\[ k \]
is the number of foods or nutrients obtained by the dietary assessment tool. This sample size is sufficient to avoid rank deficiency and to permit a preliminary PCA as for the evaluation of the correlations among foods or nutrients concurring in determining a pattern. Once determined, the number of patterns, the number of foods or nutrients associated to the pattern and the number of foods or nutrients determining the white noise the procedure to define the number of subjects to be considered can start. If the number of foods or nutrient patterns is less than 3–4 and the correlation between the foods or nutrients that determine the patterns is \( \geq 0.6 \), then the sample size of \( 1.5 - 2 \times k \) could be sufficient, the KMO can be performed to confirm the sample adequacy. Otherwise, if the KMO < 0.8 the sample can be augmented by steps of \( 0.5 - 1 \times k \). In the most likely condition for which the correlation between foods or nutrients determining a factor is less than 0.6 then the sample can be augmented by steps of \( 1 - 1.5 \times k \) individuals and the KMO statistics can be used to confirm sample size adequacy.

Applicability of the Blom’s transformations

The Blom’s transformation appeared to be a suitable approach to normalise and standardise food or nutrients intakes deriving from food questionnaires, and performed much better than the logarithmic transformation also in case of positive skewness (Figure 1). It should be noticed that this transformation holds a second advantage; resulting in a series of standardised transformed variables, it allows the PCA to be applied uniquely avoiding discussions about the use of the correlation or the covariance matrix. Finally, the use of data normalisation resulted in a better shape of principal component score, while PCA applied to skewed variables results in skewed principal component scores (Figure 1).

Eigenvalues distribution and comparison between the GK criterion, the scree plot visual inspection and results from Horn’s parallel analysis

The series of the eigenvalues plotted versus its rank appeared as a monotone continuous series without any discontinuity when no patterns are present (Figure 3). Notably, in this situation about 50% of the eigenvalues assumed a value higher than one, showing already a first fallacy of the GK criterion. When three

![Figure 3](image-url)
food patterns were simulated, a clear discontinuity appeared after the third eigenvalue showing that the scree-plot visual inspection is a suitable method to identify food patterns. This discontinuity was well recognised by the proposed algorithm as it is easily recognisable by visual inspection of the plot. Notably the scree plot discontinuity was clearly present also in the simulations with the smallest sample sizes and the weakest correlations between the foods associated to a pattern. The scree plot discontinuity became even more accentuated when the number of subjects and the correlation between the foods increases and runs parallel to the KMO statistic. The Horn’s parallel analysis confirms the results from the scree-plot visual inspection. On the other side, in the presence of three simulated food patterns, and despite other conditions, it appeared that the GK criterion generated a number of false-positive food patterns. For low to medium sample sizes and irrespectively of the correlation between the foods, the GK criterion generated multiple spurious patterns. On the contrary, the reliability of the GK criterion appeared to be satisfactory in the conditions of medium to high sample sizes, especially when the variables are strongly correlated. Notably, results from the scree-plot visual inspection and Horn’s parallel analysis are satisfactory also in conditions of poor to null factoriability of the data matrix.

**Food pattern recognition and the KMO statistic**

The KMO statistic tends to 0.5 when sample size increases in a situation where no food patterns were present according to theory (Kaiser 1974). On the other side, when three patterns were present the KMO statistic ranged between 0.59 and 0.99 representing a situation of poor to excellent factoriability of the correlation/covariance matrix. Notably, even in situations of poor to mediocre values of the KMO statistic the scree plot visual inspection resulted in a reliable outcome without any identification of spurious food patterns. For example, we reported that in condition of poor correlation between variables \( r = 0.2 \) and limiting sample size \( n = 150 \) defining a poor KMO statistic \( \text{KMO} = 0.59 \) the performance of the scree plot visual inspection is excellent. On the contrary, the GK criterion applied in the same condition resulted in a mean of 20.8 patterns over the 5000 simulations (of which only 3 are not spurious patterns). Again, when the sample size and the correlation between the simulated foods associated to a pattern increased the number of spurious false-positive patterns decreased in parallel. Finally, in conditions of excellent KMO statistic \( \text{KMO} \geq 0.9 \) the GK criterion did not results in the recognition of false-positive food patterns.

**Supplementary analyses**

Reducing number of variables or the correlations associated to a factor results in a lower eigenvalue and in a parallel reduced ability of pattern recognition. When 40, 25 and 10 variables are associated to three patterns with a correlation of 0.2, the eigenvalue of the third pattern associated to 10 variables is dramatically reduced compared to the first two. A similar effect is observed also for the third pattern when 25 variables are related to the three patterns with a correlation of 0.4, 0.2 and 0.1 (Figures 4 and 5, panels A and B). In both this conditions, the scree plot visual inspection correctly identifies the three patterns only when the sample size is higher than 150 (Figure 4, panels A and B), the performances of the GK criterion are disappointing resulting in numerous false-positive patterns. Notably, for smaller sample sizes, the scree-plot visual inspection and Horn’s parallel analysis resulted in poor ability to correctly identify the third pattern, especially when the number of observations is below the number of variables. Introducing nuisance in the correlation structure of the variables in the form of white noise and spurious correlations among variables associated to different factors reduces the performances of both the scree-plot visual inspection and Horn’s parallel analysis (Figures 4 and 5, panels C and D). The ability to identify the first two patterns is maintained by the scree-plot visual inspection and Horn’s parallel analysis for sample size higher than 150 in the presence of spurious correlations and white noise (Figure 4, panels C and D). On the contrary, pattern identification resulted limited to the first pattern when spurious correlations and white noise where present with lower sample sizes (Figure 5, panels C and D). Horn’s parallel analysis seems to perform better in these conditions.

**Discussion**

We here reported about certain pitfalls of the PCA applied to food intake questionnaires like those commonly used in current nutritional research. Some solutions to these fallacies were also provided. First, the present work reinforced some of the knowledge about sample size adequacy to PCA (Guadagnoli and Velicer 1988; Hogarty et al. 2005; Williams et al. 2010). According to our simulations, a minimum sample size
of about 450–600 individuals should be considered in common conditions where the correlation between foods or nutrients is moderate to weak and number of foods or nutrients patterns is up to four. Higher correlations between foods or nutrients can permit a reduction of the minimal sample size required by PCA, with a natural lower limit to sample size given by a number of observations exceeding at least 1.5–2 times the number of variables. According to previous studies, we confirmed that PCA requires an adequate sample size that, in common operative conditions based on the use of questionnaires, should never be less than 300 individuals at least, the so-called Tabachnick’s rule of thumb (Tabachnick and Fidell 2007). Afterwards, it was reported that the problem of variable distribution can be addressed using the Blom’s or other inverse rank transformations. Inverse rank transformation has some remarkable advantages and reduces skewness standardising target variables uniquely, despite the distribution of the target variable. This class of transformations are not as popular and easy to be computed as some more common transformations, like the logarithm. Nevertheless, inverse rank transformations perform far better than logarithm when it comes to reducing skewness and standardising the variable, thereby avoiding the problem of having to decide on whether to apply the PCA to the correlation or the covariance matrix. Moreover, we showed how Blom’s transformation reduces the skewness of principal component score with clear advantages when using principal component scores to define clusters among subjects as for related techniques like reduced rank regression. Notably, other inverse rank transformations could be evaluated. Nevertheless, we do not expect the results would change significantly. It was reported that inverse rank transformations do not bring any advantage when used as outcomes in univariate tests because type I and II errors are comparable to those in non-parametric tests (Beasley et al. 2009). This does not affect their applicability before the PCA and their use should not be discouraged.

It is well known that the GK criterion may result in an overestimation of the number of latent factors (Yeomans and Golder 1982). Here, we reported that this threat is remarkable also in those operative conditions that are commonly found when the PCA is applied to data from food intake questionnaires. On
the other side, the scree plot visual inspection is a more reliable method to individuate food patterns even if it is based on a subjective evaluation by the researcher. Furthermore, we showed for the first time that the scree plot visual inspection is reliable also in conditions of limited sample size and correlation structure as commonly found in nutritional research. The present results are in strong agreement with theoretic and computational works in the field of applied statistics and psychometry (Yeomans and Golder 1982; Zwick and Velicer 1986). In fact, we reported that the GK criterion is a suitable way to determine number of food patterns under certain specific circumstances. When the sample size increases and when the correlation between variables determining a factor is moderate to strong the GK criterion correctly determines the underlying factors. This evidence is reported in the early works around this topic (Kaiser 1970, 1974). It should be highlighted that the reliability of the GK criterion runs parallel with the value of the KMO statistic showing that a strong correlation structure and sample size is recommended for its use (Kaiser 1974). In the present study, we showed that the GK criterion is reliable when the KMO statistics have a value of 0.9 or higher, but unfortunately this condition is rarely found in nutritional research. Finally, we showed that the scree-plot visual inspection is robust also in the presence of white noise and spurious correlations among variables associated to different factors. This represents a further demonstration of PCA power to detect patterns of association among variables. Nevertheless, the ability to identify the components of the scree-plot visual inspection is lost when white noise and spurious correlation among variables are associated with small sample sizes and poor KMO. In this condition, the Horn’s parallel analysis seems to perform better. Nevertheless, PCA should be applied only when the KMO statistics have values higher than 0.8, in conditions of partial to null data factoriability PCA should then be avoided.

The present work has some remarkable strengths. First, this is the first time that simulation studies about sample size adequacy and pattern recognition are conducted using settings similar to those found in nutritional research. Second, the present work is based on a robust and up to date methodology using a much more modern simulation approach and using much more powerful computers than those in use.
during the 1980s, when the literature around the topic developed. Third, the present work provides some useful practical suggestions about how to manipulate variables before PCA and perform the pattern recognition.

Furthermore, it should be highlighted that the results from the present work are applicable to all of those PCA-related techniques, like exploratory factor analysis, cluster analysis, and confirmatory factor analysis where latent factors are defined by means of eigenvalue decomposition of the covariance/correlation matrix.

The present study also has some weaknesses. Mainly, it should be noticed that in common operative conditions the patterns are not so-well delineated as the ones from data simulation. This happens because all foods are consumed in association and it is common that foods or nutrients may belong to more than one single pattern.

There is little information around this topic and simulation approaches should be used to evaluate the ability of the PCA to disentangle some much more complex situation where a food is correlated to more than one pattern at the time. For now, it is clear that the GK criterion results in an overestimation of the food patterns when food patterns are independent. In the condition of partially correlated patterns the GK criterion is not expected to improve its performances, rather it is expected to make it worse.

The fallacies of the GK criterion should be well known to all researchers in any field and it is not clear why this method is still in use, especially because the limitations of the scree plot visual inspection appear as negligible (Streiner 1998).

**Conclusion**

The sample size alone does not guarantee the applicability of PCA and other factors such as number of patterns, number of variables associated to a pattern and the strength of the association between the variables associated to a pattern should be taken into account before conducting the PCA. The Blom’s transformation appears to be a suitable method to normalise and standardise data before the PCA and results in an appreciable improvement of the distribution of principal component scores. The GK criterion applied to PCA performed on data from food intake questionnaires results in false-positive patterns; this is more evident as the KMO statistic decreases under the value of 0.8. The scree plot visual inspection, on the contrary, results in a better performance also in conditions of limited sample size and correlation structure providing results comparable to those from Horn’s parallel analysis.

**Acknowledgements**

The present work was inspired by the wisdom and great charisma of professor Silvano Milani, recently deceased.

**Disclosure statement**

The authors have no conflict of interest to declare. The present work has to be considered as spontaneous research and no one of the authors received any grant.

**References**

Beasley TM, Erickson S, Allison DB. 2009. Rank-based inverse normal transformations are increasingly used, but are they merited? Behav Genet. 39:580.

Blom G. 1958. Statistical estimates and transformed beta-variables. Uppsala: Almqvist & Wiksell.

Cerny BA, Kaiser HF. 1977. A study of a measure of sampling adequacy for factor-analytic correlation matrices. Multivariate Behav Res. 12:43–47.

Conover W, Iman RL. 1976. On some alternative procedures using ranks for the analysis of experimental designs. Commun Stat Theory Methods. 5:1349–1368.

Dinno A. 2009. Exploring the sensitivity of Horn’s parallel analysis to the distributional form of random data. Multivariate Behav Res. 44:362–388.

Dziuban CD, Shirley EC. 1974. When is a correlation matrix appropriate for factor analysis? Some decision rules. Psychol Bull. 81:358.

Guadagnoli E, Velicer WF. 1988. Relation of sample size to the stability of component patterns. Psychol Bull. 103:265.

Hair JF, Black WC, Babin BJ, Anderson RE, Tatham RL. 1998. Multivariate data analysis. Upper Saddle River, NJ: Prentice-Hall.

Henson RK, Roberts JK. 2006. Use of exploratory factor analysis in published research: common errors and some comment on improved practice. Educ Psychol Meas. 66:393–416.

Hogarty KY, Hines CV, Kromrey JD, Ferron JM, Mumford KR. 2005. The quality of factor solutions in exploratory factor analysis: the influence of sample size, communality, and overdetermination. Educ Psychol Meas. 65:202–226.

Hu FB. 2002. Dietary pattern analysis: a new direction in nutritional epidemiology. Curr Opin Lipidol. 13:3–9.

Jacobs DR, Steffen LM. 2003. Nutrients, foods, and dietary patterns as exposures in research: a framework for food synergy. Am J Clin Nutr. 78:508S–513S.

Johnson RA, Wichern DW. 2014. Applied multivariate statistical analysis. Upper Saddle River, NJ: Prentice-Hall.

Kaiser HF. 1970. A second generation little jiffy. Psychometrika. 35:401–415.

Kaiser HF. 1974. An index of factorial simplicity. Psychometrika. 39:31–36.
Moeller SM, Reedy J, Millen AE, Dixon LB, Newby P, Tucker KL, Krebs-Smith SM, Guenther PM. 2007. Dietary patterns: challenges and opportunities in dietary patterns research an Experimental Biology workshop, April 1, 2006. J Am Diet Assoc. 107:1233–1239.
Mooney CZ. 1997. Monte Carlo simulation. London: Sage Publications.
Rummel RJ. 1988. Applied factor analysis. Evanston: Northwestern University Press.
Shirkey EC, Dziuban CD. 1976. A note on some sampling characteristics of the measure of sampling adequacy (MSA). Multivariate Behav Res. 11:125–128.
Streiner DL. 1998. Factors affecting reliability of interpretations of scree plots. Psychol Rep. 83:687–694.
Tabachnick BG, Fidell LS. 2007. Using multivariate statistics. London: Allyn & Bacon/Pearson Education.
Trichopoulos D, Lagiou P. 2001. Dietary patterns and mortality. Br J Nutr. 85:133–134.
Wicklin R. 2013. Simulating data with SAS. Cary, NA: SAS Institute.
Williams B, Onsman A, Brown T. 2010. Exploratory factor analysis: a five-step guide for novices. Au J Paramed. 8.
Yeomans KA, Golder PA. 1982. The Guttman-Kaiser criterion as a predictor of the number of common factors. J Royal Stat Soc. 31:221–229.
Zwick WR, Velicer WF. 1986. Comparison of five rules for determining the number of components to retain. Psychol Bull. 99:432.