Use of principal component analysis (PCA) and hierarchical cluster analysis (HCA) for multivariate association between bioactive compounds and functional properties in foods: A critical perspective

Daniel Granato, Jânio S. Santos, Graziela B. Escher, Bruno L. Ferreira, Rubén M. Maggio

PII: S0924-2244(17)30636-2

DOI: 10.1016/j.tifs.2017.12.006

Reference: TIFS 2133

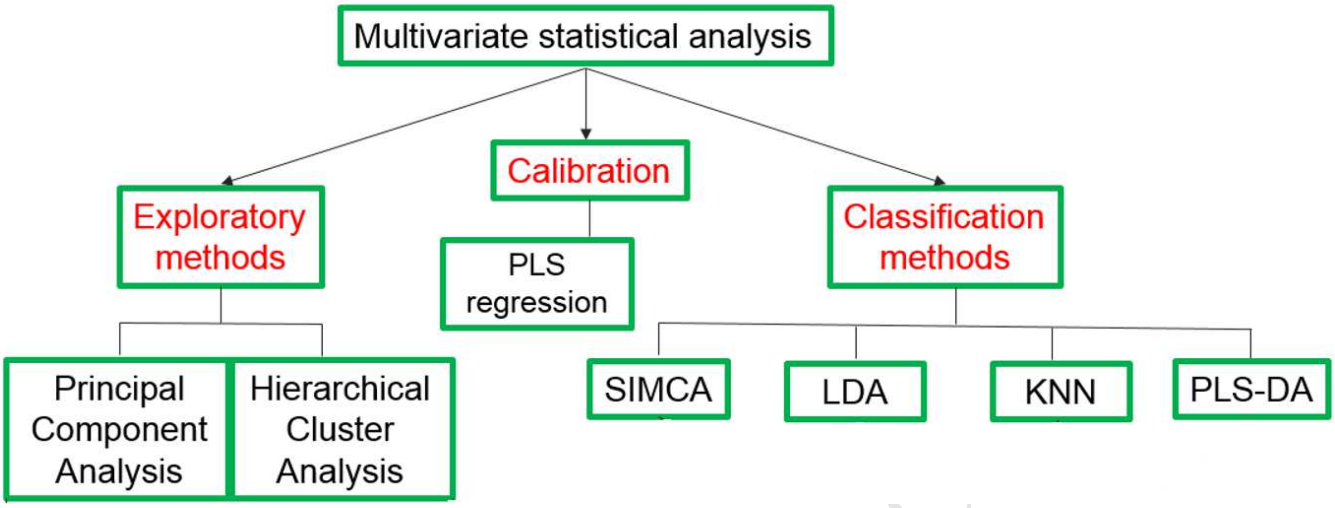
To appear in: *Trends in Food Science & Technology*

Received Date: 19 October 2017

Revised Date: 10 December 2017

Accepted Date: 13 December 2017

GRAPHICAL ABSTRACT


Multivariate statistical analysis
Exploratory methods
Calibration
Classification methods
PLS regression
Principal Component Analysis
Hierarchical Cluster Analysis
SIMCA
LDA
KNN
PLS-DA

Use of principal component analysis (PCA) and hierarchical cluster analysis (HCA) for multivariate association between bioactive compounds and functional properties in foods: a critical perspective

Daniel Granato[1]*, Jânio S. Santos[1], Graziela B. Escher[1], Bruno L. Ferreira[2], Rubén M. Maggio[3*]

[1]Graduation Program in Food Science and Technology, State University of Ponta Grossa. Av. Carlos Cavalcanti, 4748, 84030-900, Ponta Grossa, Brazil. E-mail: dgranato@uepg.br

[2]Graduation Program in Food Science, Federal University of Santa Catarina. Rodovia Admar Gonzaga, 1346, Itacorubi, Florianópolis, Brazil.

[3]Área Análisis de Medicamentos, Facultad de Ciencias Bioquímicas y Farmacéuticas, Universidad Nacional de Rosario, Instituto de Química de Rosario (IQUIR-CONICET), Suipacha 531, S2002LRK Rosario, Argentina. E-mail: maggio@iquir-conicet.gov.ar

Abstract

Background: The development of statistical software has enabled food scientists to perform a wide variety of mathematical/statistical analyses and solve problems. Therefore, not only sophisticated analytical methods but also the application of multivariate statistical methods have increased considerably. Herein, principal component analysis (PCA) and hierarchical cluster analysis (HCA) are the most widely used tools to explore similarities and hidden patterns among samples where relationship on data and grouping are until unclear. Usually, larger chemical data sets, bioactive compounds and functional properties are the target of these methodologies. *Scope and approach:* In this article, we criticize these methods when correlation analysis should be performed and results analyzed. *Key findings and conclusions:* The use of PCA and HCA in food chemistry studies has increased because the results are easy to interpret and discuss. However, their indiscriminate use to assess the association between bioactive compounds and *in vitro* functional properties is criticized as they provide a qualitative view of the data. When appropriate, one should bear in mind that the correlation between the content of chemical compounds and bioactivity could be duly discussed using correlation coefficients.

Keywords: chemometrics; principal component analysis; cluster analysis; correlation analysis; bioactive compounds; functional properties.

Abbreviations

ABTS - 2,2'-azino-bis(3-ethylbenzothiazoline-6-sulphonic acid)

ANN - artificial neural networks

CAIMAN - classification and influence matrix analysis

DD-SIMCA - data-driven soft independent modeling of class analogy

DPPH - 2,2-diphenyl-1-picrylhydrazyl

FRAP - ferric reducing antioxidant power

FuRES - fuzzy rule-building expert system

HCA – hierarchical cluster analysis

HPLC – high performance liquid chromatography

IMS - ion mobility spectrometry

k-NN - *k*-nearest neighbors

LDA – linear discriminant analysis

NMR – nuclear magnetic resonance

OPLS-DA - orthogonal partial least squared discriminant analysis

ORAC – oxygen radical absorbance capacity

PCA – principal component analysis

PLS-DA - partial least squared discriminant analysis

PRIMA - pattern recognition by independent multi-category analysis

QDA - quadratic discriminant analysis

RF - random forests

SIMCA - soft independent modeling of class analogy

sPLS-DA - super partial least squared discriminant analysis

SVM - support vector machine

UHPLC-MS – ultra-high performance liquid chromatography – mass spectrometry

Introduction

As well stressed by Ropodi, Panagou, and Nychas (2016), in the 21st century, governmental, industrial, and academic problems need to be addressed by using sophisticated analytical tools with proper data collection, analysis and interpretation. In this sense, data mining and data analysis are two interrelated approaches developed rapidly to address problems related to engineering and technology, as well as medicine, economics, biology, and food science (Brown, 2017).

Chemometrics is an interfacial discipline that extracts useful information from large chemical and biochemical data sets using different mathematical and statistical methods (Nunes et al., 2015, Brown, 2017). In applied chemistry, the use of chemometrics has been spread and well recognized since 1960 (Brereton, 2014), but in food sciences and technology the applications of chemometrics and sensometrics (multivariate methods applied to sensory data and studies consumers) are somewhat new (Munck, Nørgaard, Engelsen, Bro, & Andersson, 1998; Aquino et al., 2014; Qannari, 2017). Conversely, the application of chemometrics for assessing the adulteration and geographical origin of foods based on chemical markers is well established in food science (Granato, Koot, Schnitzler, & van Ruth, 2015; Granato, Margraf, Brotzakis, Capuano, & van Ruth, 2015; Paneque, Morales, Burgos, Ponce, & Callejón, 2017; Giannetti, Mariani, Mannino, & Marini, 2017; Opatić et al., 2018). For example, Garrido-Delgado, Muñoz-Pérez, and Arce (2018) used ion mobility spectrometry (IMS) to determine the origin of the olive oil, quality and adulteration with low-cost vegetable oils. Using different statistical tools, authors were able to predict the level of contaminating oil in olive oil. Therefore, there is no doubt that chemometric tools is of fundamental importance to solve real life problems.

Granato, Nunes, and Barba (2017) stated that the use of design of experiments together with appropriate statistical data analysis is of pivotal importance to assess the association between nutrition, biology, pharmacology, functional properties and the chemical components of foods and their extracts. In this sense, chemometric tools and other statistical methodologies may be of interest when different food extracts and bioactivities need to be evaluated (Granato, de Araújo Calado, & Jarvis, 2014).

In real life applications, chemometrics may be employed in food science and technology studies either to assess similarities/differences between multiple objects (samples) or to project the objects in a two/three-dimensional factor-plane based on various characteristics. Therefore, clusterings can be observed and the reasons for the grouping can be pinpointed (Jandrić, & Cannavan, 2017; Lund, Brown, & Shipley, 2017; Erasmus, Muller, Butler, & Hoffman, 2018). Additionally, multivariate techniques have been widely used to authenticate/trace the geographical origin of foods, to verify the farming system employed by a company and check whether it complies to the information declared on the label, and to check for adulterations (intentional or not) of foods and raw materials (Granato, Koot, & van Ruth, 2015; Chiesa et al., 2016; Müller-Maatsch, Schweiggert, & Carle, 2016; Tavares et al., 2016; Zhu, Wang, & Chen, 2017; Karabagias et al., 2017; Chung et al., 2017; Giannetti, Mariani, Mannino, & Marini, 2017; Acierno et al., 2018).

For example, Luo, Shi, and Feng (2017) aimed to characterize the metabolites of Zhi-Zi-Hou-Po decoction, a traditional Chinese medicine, in rat bile, urine and feces after oral administration, using untargeted liquid chromatography time of flight mass spectrometry combined with orthogonal partial least squared discriminant (OPLS-DA). After analyzing the experimental data, authors were able to identify 83 compounds, in which 39 were metabolites, in the biological samples. In addition, the metabolic pathway (glucoronidation) by which these metabolites formed after oral administration of the decoction was identified by using OPLS-DA. This research is an example on how chemometric tools are important aids in not only in the food chemistry field but also in the experimental nutrition studies.

According to Brereton (2015), chemometrics users tend to 'follow the crowd' and use indiscriminately the available software without knowing the principles and fundamentals of each method applied in their research data analysis. In food chemistry studies, Principal Components Analysis (PCA) and Hierarchical Cluster Analysis (HCA) are widely (and, sometimes, improperly) applied as "*unsupervised classification"* methods to assess the association between bioactive compounds and *in vitro* functional properties (*i.e*., antioxidant and inhibition of enzymes). Herein, a critical perspective on these display techniques (PCA and HCA) is made together with some comments on their use in the field of bioactive compounds.

Study of bioactive compounds and *in vitro* potential functional properties with the use of chemometrics

Chemometrics may be used for both *qualitative* and *quantitative* analysis of experimental data (Szymanska et al., 2015; Martínez et al., 2017). Determining whether a rice sample comes from European countries or elsewhere based on the NMR spectra or the presence or absence of a chemical compound in a HPLC chromatogram are two typical examples of *qualitative* data. On the other hand, assessing the correlation between the content of chlorogenic acid derivatives and antioxidant activity of coffee brews represents a *quantitative* approach. A summary of selected multivariate statistical methods is shown in Figure 1.

Overall, chemometrics may be divided into *calibration*, *classification* and *exploratory* methods. According to Oliveri and Simonetti (2016), chemometrics may be divided into *supervised* and *unsupervised* methods. The first class encompasses a varied number of methods/algorithms, including both qualitative and quantitative approaches. Among qualitative methods, *k*-nearest neighbors (*k*-NN), partial least squares discriminant analysis (PLS-DA), super PLS-DA (sPLS-DA), fuzzy rule-building expert system (FuRES), soft independent modeling of class analogy (SIMCA) and linear or quadratic discriminant analysis (LDA or QDA) are the most used techniques. However, some methods, such as classification and influence matrix analysis (CAIMAN), pattern recognition by independent multi-category analysis (PRIMA), support vector machine (SVM), random forests (RF), and artificial neural networks (ANN), show several applications in food science and technology, especially in the classification and authentication problems (Tian et al., 2017; Torkashvand, Ahmadi, & Nikravesh, 2017; Aloglu et al., 2017; Mehretie, Al Riza, Yoshito, & Kondo, 2018).

Unsupervised methods, also named *clustering* or *displays methods*, are used to study the data structure, look for similarities between multiple objects, and check for outliers in the data set (Liu, Koot, Hettinga, de Jong, & van Ruth, 2018). Mixture models, self-organizing maps, *k*-means, HCA and PCA are representatives of unsupervised methods. However, PCA and HCA are the most used in food and chemistry field, representing both sub-classes visualization and agglomerative algorithms, respectively (Wang, Zeng, Contreras, & Wang, 2017).

The goal of multivariate unsupervised methods is to evaluate whether clustering exists in a dataset without using class membership information in the calculations (Beebe, Pell, & Seasholtz, 1998). Natural clustering of samples/objects is the result of understanding the measurement system used to characterize the samples and this union between statistical analysis and analytical methods aids in elucidating the physical reasons for the presence/absence of clustering in the data. For further information on these methods, the reader is referred to existing literature (Oliveri & Downey, 2012; de Oliveira et al., 2015).

Here we show some recent applications of unsupervised multivariate techniques in the field of bioactivity of food components. When it comes to studies relating bioactive compounds, almost all reports aim to associate the level of certain chemical compounds, *i.e.*, phenolic compounds and carotenoids, with antioxidant activity and other functionalities. Additionally, a critical perspective on the use of display techniques (PCA and HCA) is made together with some comments on their use in the field of bioactive compounds.

Principal component analysis

The term PCA is statistical test that belongs to a group of factor analysis. PCA is a mathematical tool that aims to represent the variation present in the dataset (*i.e.*, responses used to characterize the samples) using a small number of factors. For visual analysis, usually two-dimensional or three-dimensional projection of samples is constructed having the axes (principal components, PC) as the factors. Each PC is a linear combination of the original responses (that retain some correlation among) and PCs are orthogonal to each other. PCs iteratively calculated hold as much variation from original data set as possible, in a way that PC1 explains more the data variation than PC2, and PC2 explain more data variation than PC3 and so on. That is why a few PCs explain the variation of a large number of original responses. One possible way to determine the number of PC is based on the Kaiser criterion (Kaiser, 1960): eigenvalues higher than 1 are considered as "significant" in the PCA analysis. In addition, the use of Bartlett's test of sphericity is of interest to check for correlation between responses. This test indicates that the responses are (un)related and therefore (un)suitable for structure detection.

Figure 2 contains an example of PCA of fruit juices (*i.e.,* orange, lemon and grape) based on chemical composition and antioxidant activity: the responses used to generate the 2D-scatter plot are based on correlation analysis of each response with the first three PCs. As first step an exploration of cumulative variance explained should be carried out and the Kaiser criterion (eigenvalues higher than 1) may be used to define the number of significant PC. Usually this decision is taken according to pre-established level of variance (90, 95, 99, or 99.9%) or based on experimental error.

Using a factor loadings analysis (Table 1), PC1 retained about 50% of data variation and differentiate the juice samples according to the contents of caffeic acid, (-)-epicatechin, (+)-catechin, quercetin, luteolin and antioxidant activity (2,2-diphenyl-1-picrylhydrazyl – DPPH, 2,2'-azino-bis (3-

ethylbenzothiazoline-6-sulfonic acid – ABTS, and ferric reducing antioxidant power - FRAP). Similarly, PC2 explained another 30% of variability in the original responses and separates the juices based on FRAP, gallic acid, and 5-*O*-caffeoylquinic acid. PC3 and PC4 explain only 11% of data variance and barely does not differentiate the juice samples. The factor loadings from PC3 and PC4 were very low (except for quercetrin/luteolin and ellagic acid, respectively). Factor loadings lower than 0.60 indicate that those variables that do not fit well with the factor solution should possibly be dropped from the analysis, especially if the projection of samples on a factor-plane is based on a 2-dimensional graph. As a final comment, the first two PCs explain about 81% of data variance but there remains room for about 19% unexplained variation.

Once the representative PCs were found, on the basis of samples differentiation/grouping and variance explained, loading analysis is started in order to find the underlying relationships in the original data structure. In this step loading could be visualized as a regression vector (a vector of correlation coefficients between the original variables with each PC-score). The positive factor loadings indicate that the factor will be higher in the positive axis of that PC. For example, for DPPH, a factor loading of 0.69 was obtained with PC1, which means that the samples located in the right-hand side (*i.e*., violet stars) of the graph have higher mean DPPH values than the samples located in the left-hand side (*i.e*., red stars). Similarly, the negative factor loadings indicate that the factor will be higher in the positive axis of that PC. For example, for (-)-epicatechin a factor loading of -0.75 was obtained for PC1, meaning that the samples located in the right-hand side (*i.e*., violet stars) of the graph have lower mean concentrations than the samples located in the left-hand side (*i.e*., red stars).

As a complementary analysis, as an illustrative example, PCA data may be compared to correlation coefficients (Table 2). As shown, the antioxidant activity measured by three different assays (*i.e*., ABTS, FRAP, and DPPH) is mainly correlated ($p < 0.05$) to caffeic acid, (-)-epicatechin, (+)-catechin, quercetin, and luteolin. FRAP also correlated significantly with gallic acid and 5-O-caffeoylquinic acid. However, if the main objective is to check for association between bioactive compounds and functional properties, correlation analysis should be carried out.

For instance, Pearson's correlation coefficients or Spearman's rank correlation coefficients are the choices for normally distributed data and for data do not conform to the normal distribution, respectively (de Oliveira et al., 2015).

As a final comment on this topic, there is no scientific need to perform PCA or HCA for data sets that have a similar conclusion as the one shown in the above-mentioned example. However, if the number of responses and samples is quite large and data are quite complex (*i.e.,* NMR spectra), PCA is highly indicated.

Dos Santos et al. (2017) quantified 13 phenolic compounds in 96 guava fruit pulps (*Psidium guajava* L.) by HPLC, including (+)-catechin, gallic, ferulic, *trans*-cinnamic, chlorogenic, caffeic, *p*-coumaric, syringic, vanillic, and ellagic acid, rutin, quercetin, and kaempferol. The extraction procedure was optimized using different concentrations of ethyl alcohol and methyl alcohol for 15 to 90 min using a sample to solvent ratio between 1:30 and 1:100 w/v. The extracts were also analyzed for total phenolic content, ascorbic acid, and flavonoids, together with the antioxidant activity toward DPPH and ABTS radicals. PCA was able to explain only 60% of data variability with 2 PC, but a clear separation between ripe and green guava fruits was observed from the scatter plot. The main responses that separated the groups were syringic acid, (+)-catechin, *p*-coumaric acid, caffeic acid, ellagic acid, *trans*-cinnamic acid and rutin for the green guava, while for ripe and white guava, the better markers were gallic acid and chlorogenic acid. As rational subsequent step, authors applied ANN (a supervised algorithm) on same data set to obtain a reliable methodology to classify their samples. ANN showed a suitable separation between not only green and white variety but also ripe and unripe guava fruits. It should be stressed that as data were successfully analyzed by PCA, a linear algorithm, LDA or PLS-DA was the logical way to try.

However, in some cases, the differentiation between classes is not so clear (Figure 3A) and outliers (one or more observation point(s) that is/are unusually distant from the other observations) can be detected in the dataset. In this case, the researcher cannot expect a straightforward separation between classes. Almost perfect segregation was obtained when all samples are analyzed after outliers removal (in synthetic data) using only two principal components (PCs), as shown in Figure 3B.

Fidelis et al. (2017) evaluated multiple juices from different botanical origins (fruits and other vegetables) in relation to some classes of phenolics/bioactive compounds (tannins, total phenols, flavonoids, *ortho*-diphenols, flavonols, total anthocyanins, and betalains), physicochemical properties (pH, soluble solids, and acidity), and antioxidant effects (Fe^{2+} chelating properties, antiradical effect (DPPH, ABTS, and FRAP), Folin-

Ciocalteu's reducing capacity, and total reducing capacity. A total of 570 data points (38 juices and 15 responses) were analyzed for patterns using PCA, which explained 72% of data variability with 2 PC and it was possible to pinpoint the juices with higher bioactive compounds and antioxidant activity. PLS-DA was used to discriminate juice groups and authors were able to separate *Citrus* juices from *Super juices* (made with berries) with correct classification rates above 73%, while data-driven SIMCA, which is a one-class classification method, was able to discriminate the juices samples with accuracy higher than 86%. In this research, authors concluded that the use of DD-SIMCA may be of interest when the authentication of juices based on phenolic compounds and antioxidant activity need to be performed, especially in quality control programs in the juice industry.

Kalaycıoğlu, Kaygusuz, Döker, Kolaylı, & Erim (2017) used PCA to explore only n = 10 Turkish honeybee pollens from distinct origins based on organic acids, carbohydrates, 14 minerals, total phenolic content, and antioxidant activity measured by the DPPH assay. Not surprisingly, the first three principal components explained 71% of data variability and authors claim they "classified" the pollen samples according to the geographical origin of the samples (less than five samples per class, in which, n = 2 chestnuts, n = 1 oak, n = 1 Abana, n = 1 Bayburt, n = 1 Balikesir, n = 1 buckwheat, and n = 3 Anzer). However, results are untrustworthy when such low number of samples are available, so conclusions based on the PC plots should be pondered. According to de Oliveira et al. (2015), for PCA, at least five responses and five objects (samples) need to be part of the dataset.

Santos et al. (2016) used PCA to reveal the effects of time (5 – 10 min) and extraction temperature (65 – 85 ^{o}C) on phenolic composition and functional properties of aqueous extracts of fermented rooibos (*Aspalathus linearis*). For this purpose, a 2^2 factorial design with three central points was used to manufacture beverages in which some phenolic acids and flavonoids were quantified using LC-MS/MS, the antioxidant activity (ABTS, FRAP, and total reducing capacity), and the inhibition of α-amylase and α-glucosidase were determined. As a large amount of data were generated (210 data points), authors performed a PCA to reduce dimensionality of the data. Authors verified that rooibos extracted at 85 ^{o}C, regardless of the extraction time, presented the highest levels of phenolic compounds, *in vitro* antioxidant activity, and highest inhibition of the digestive enzymes. Although correlation coefficients were calculated to know which compounds exerted the *in vitro* antioxidant effect, PCA was effective in showing the best technological conditions to produce the infusions with higher bioactive compounds.

Farag, Ezzat, Salama, and Tadros (2016) studied the anti-acetylcholinesterase activity and bioactive compounds of four sweet basil species (*Ocimum basilicum, Ocimum africanum, Ocimum americanum* and *Ocimum minimum*) by ultra-performance liquid chromatography quadrupole time of flight mass spectrometry (UPLC/qTOF/MS), PCA was used as exploratory tool and OPLS-DA was used for its further analysis. Twenty one hydroxycinnamic acids, 4 benzoic acid conjugates, 14 flavonoid conjugates, 2 alcohols, 5 acyl sugars, 4 triterpenes and 12 fatty acids were identified in the extracts. Using these responses, authors applied PCA and HCA to pinpoint the sweet basis species with higher anti-acetylcholinesterase activity: *O.*

americanum, O. africanum, and *O. basilicum.* Additionally, OPLS-DA was used to distinguish between *O. basilicum* (official drug) from *O. americanum*, with more than 96% of data variability explained by the classification model.

Hierarchical cluster analysis

HCA is a clustering method which explore the organization of samples in groups and among groups depicting a hierarchy (Lee & Yang, 2009). The result of HCA is usually presented in a dendrogram, a plot which shows the organization of samples and its relationships in tree form. There are two main approaches to resolve the grouping problem in HCA, agglomerative or divisive (Figure 4).

In the first one, each sample is initially considered a cluster, and subsequently pairs of clusters are merged. In divisive approach algorithm start with one cluster including al samples, recursive splits are performed. Clustering is achieved by use of an appropriate metric of samples distance (usually, Euclidean, Mahalanobis or Manhattan distance) and linkage criterion among groups. Complete, single and average and Ward’s linkage are the more common variants of linkage criterions. Ward’s method, based in optimal value of a target function, is a possible choice (Granato, Karnopp, & van Ruth, 2015).

HCA has also been extensively used to evaluate the multivariate association between bioactive compounds and bioactivity of foods, beverages and their extracts. For instance, Viapiana et al. (2016) used HCA aiming to associate the relationship between phenolic composition measured by HPLC with the *in vitro* antioxidant activity (FRAP and DPPH assays) of 19 chamomile commercial samples (*Matricaria chamomilla* L.). Overall, caffeic, ferulic, and syringic acids were the most effective phenolics in exerting antioxidant activity in the herbal extracts (MeOH:H_2O, 80:20 v/v). Linear correlation coefficients were also calculated to display a mathematical proof of such findings ($r>0.70$, $p<0.05$). HCA and PCA were used with the aim to tentatively “classify” the commercial samples based on the HPLC fingerprint but no differentiation between samples was achieved. This study shows that PCA/HCA methods not always provide sufficient means to group samples according to the concentrations of bioactive compounds and antioxidant activity indices.

Sánchez-Salcedo et al. (2016) used HCA as a tool to propose a polyphenolic fingerprint of white (*Morus alba* L., n=4) and black (*Morus nigra* L., n=4) mulberry leaves clones. UHPLC-MS identified 31 phenolic compounds, mostly important 20 flavonoids, in more than 120 spectrums analyzed, a very high number of variables for such low number of samples. Ward’s method based on Euclidean distance generated three major groups, first characterized by 4 clones of both species presenting high amount of caffeic acid-hexoside, caffeoylquinic acid and kaempferol-malonyl-rutinoside and low content of O-hexoside flavonols. Only one clone of *Morus nigra* formed the second group, representing caffeic acid and cryptochlorogenic acid as characteristics. The last group was formed of 3 clones of two mulberry species, presenting high flavonols containing O-hexoside and a low content of caffeic acid.

To study the geographical influence on phenolic content and antioxidant activity in Napirira bean (*Phaseolus vulgaris* L.), Fan and Beta (2017) applied an unsupervised pattern recognition method (HCA) based on the Euclidean distance and Ward’s method. Total phenolic content, antioxidant activity, and phenolic compounds (protocatechuic acid, *p*-hydroxybenzoic acid, catechin, *p*-

coumaric acid, ferulic acid and sinapic acid) were analyzed in eighteen bean samples from four locations in Malawi. HCA was able to differentiate 3 major groups: group 1 clustered samples with high contents of phenolic compounds and antioxidant activity which were from the high-altitude region; group 2 clustered samples that presented low contents of phenolic compounds and antioxidant activity which were from a lower-altitude region; and group 3 contained samples with intermediate values of phenolic compounds and antioxidant activity and included samples from both intermediate regions of Malawi. As a conclusion, HCA was a useful tool to associate the phenolic compounds/antioxidant activity with the cultivation region.

A good example where algorithm configuration could be decisive to obtain a valid conclusion is illustrated by Kaškonienė et al. (2015). Authors analyzed the total phenolic and flavonoids contents, antioxidant activity and individual phenolic compounds (gallic acid, caffeic acid, ferulic acid, 2-hydroxycinnamic acid, rutin naringenin and quercetin) in 14 pollen samples collected in the Baltic region (Latvia and Lithuania) and two others from Spain and China. Data were treated by HCA using both Spearman's distance and Euclidean distance. Samples were clustered in two groups according to the antioxidant activity. Similarly, Euclidean distance clustered the samples into three major groups according to the geographical regions with clear differences in the phenolic composition. As a conclusion the choice of distance function is not a trivial matter and should be tested when HCA is applied. The use of the only one clustering technique (*i.e.,* k-means or tree-clustering), amalgamation rule (*i.e.,* single linkage, complete linkage, or Ward's method), and distance measure (*i.e*, Euclidean, Manhattan, 1- Pearson r) is not recommended.

Nayik and Nanda (2016) analyzed the minerals, phenolic composition and antioxidant activity of n = 37 unifloral honeys from Kashmir, India. PCA and HCA were used to assess the effects of the botanical origins of those samples based on the quality parameters and verified that PCA was able to group the samples according to the origin (apple, cherry, saffron and wild bush). The authors claimed that “minerals presented the highest discriminating power” in PCA while samples were “classified” using HCA. The terms “discriminating power” and “classification” are related to supervised chemometric tools, such as LDA/QDA or SIMCA, among other techniques (Popek, Halagarda, & Kursa, 2017; Mapelli-Brahm, Hernanz-Vila, Stinco, Heredia, & Meléndez-Martínez, 2018; Kasprzyk, Depciuch, Grabek-Lejko, & Parlinska-Wojtan, 2018). Therefore, such terms should be avoided when PCA or HCA are employed.

Overall comments on PCA and HCA

Both PCA and HCA are usually used concomitantly in studies covering bioactive compounds and functional properties. To illustrate what is widely seen in published articles, consider the following: n = 20 samples coming from two fruits (A and B) are analyzed for the concentrations of total phenolics, carotenoids, antioxidant activity measured by the oxygen radical absorbance capacity (ORAC) assay, and inhibition of amylase and lipase. Results were analyzed using PCA and the 2D projection is given in Figure 5A: it is possible to see a defined cluster containing fruit “B” and another group containing most “A” fruits. However, there are n = 3 “A” samples that are far from the main “A” group. One could say they are outliers simply by looking at the projection, but this cannot be done as PCA does not “classify” objects. In Figure 5B, HCA was applied using the Ward's method as the amalgamation rule and Euclidean distances were calculated between fruits. Using a linkage distance of 15, only two groups are formed, one containing the "A" fruits and the other containing all "B" fruits. Similarly, if a distance of 5 is considered, there are 1 group containing the "B" fruits and two other groups containing the "A" fruits, which is similar to the PCA results. However, if a linkage distance of 1.5 is considered, a total of 6 small groups can be visualized. Using this simple example, it is possible to conclude that HCA is an arbitrary method and should be used for exploratory purposes only. Additionally, neither PCA nor HCA creates a "mathematical model" for classification and authentication purposes. Rather, they only project or display the objects under investigation based on selected responses and grouping of samples may be identified by the user. Moreover, neither PCA nor HCA provides a statistical significance of such similarities (Andrić, Bajusz, Rácz, Šegan, & Héberger, 2016).

If the aim is to find an association between bioactive compounds and functional properties using HCA, the method may be applied (Figure 5C). It is an easy and straightforward result: total phenolics and carotenoids are associated with ORAC and inhibition of α-amylase. Conversely, the inhibition of lipase does not seem to be associated with any of the responses. Although HCA shows the existence of association between responses, however it does not provide a measure of the association (qualitative approach). One alternative to overcome this limitation is to calculate the correlation coefficient and provide a quantitative measure of the correlation between responses. As a matter of fact, the inhibition of lipase is not correlated to the concentrations of total phenolics ($r = -0.022$, $p = 0.927$), carotenoids ($r = 0.213$, $p = 0.367$), and ORAC (r = 0.304, p = 0.193). In this case, the use of HCA is near meaningless as correlation coefficients are robust enough to draw the association between the chemical composition and the functional properties of the fruits.

Although PCA and HCA are very useful to study the data structure and find similarities among samples, in most cases, linear correlation coefficients would render very similar interpretations of the results. Indeed, it is widely known and recognized that higher levels of phenolic compounds will render a higher antioxidant activity measured by chemical reactions *in vitro* (Guo, Sun, Yu, & Qi, 2017; Lv, Zhang, Shi, & Lin, 2017). Another main disadvantage of using PCA/HCA in those studies is the real applicability of the observations: it seems that most researchers only use PCA and HCA to increment their data analysis rather than to explain the mechanisms of action and have a strong and in-depth discussion based on a solid hypothesis. In fact, in the field of bioactive compounds, when *in vitro* assays are used, it is somewhat obvious that almost all carotenoids and phenolic compounds will exert antioxidant activity. In this case, correlation coefficients should be calculated and results analyzed.

Final comments and recommendations

The use of PCA and HCA in food chemistry studies has increased in the past years because the results are easy to interpret and discuss, especially of a large data set is analyzed. However, the indiscriminate use of multivariate exploratory statistical techniques (PCA and HCA) to assess the association between bioactive compounds and *in vitro* functional properties is criticized as the results will be, in most cases, a *sine qua non* observation. When appropriate, the researcher should bear in mind that the correlation between the content of chemical compounds and bioactivity could be duly discussed using simple linear correlation coefficients.

Acknowledgements

Daniel Granato acknowledges CNPq for a productivity grant (process 303188/2016-2). J. S. Santos and G. B. Escher thank CAPES/Fundação Araucária for their Ph.D scholarships.

References

Acierno, V., Alewijn, M., Zomer, P., & van Ruth, S. M. (2018). Making cocoa origin traceable: fingerprints of chocolates using Flow Infusion - Electro Spray Ionization - Mass Spectrometry. *Food Control*, 85, 245-252.

Aloglu, A. K., Harrington, P. B., Sahin, S., Demir, C., & Gunes, M. E. (2017). Chemical profiling of floral and chestnut *honey* using high-performance liquid chromatography-ultraviolet detection. *Journal of Food Composition and Analysis*, 62, 205-210.

Andrić, F., Bajusz, D., Rácz, A., Šegan, S., & Héberger, K. (2016). Multivariate assessment of lipophilicity scales—computational and reversed phase thin-layer chromatographic indices. *Journal of Pharmaceutical and Biomedical Analysis*, 127, 81-93.

Aquino, L. F. M. C., Silva, A. C. O., Freitas, M. Q., Felicio, T. L., Cruz, A. G., & Conte-Junior, C. A. (2014). Identifying cheese whey an adulterant in milk: Limited contribution of a sensometric approach. *Food Research International*, 62, 233-237.

Beebe, K. R., Pell, R. J., & Seasholtz, M. B. Chemometrics: a practical guide. 1^{st} ed. New York: Wiley & Sons, 1998, 348 p.

Brereton, R. G. (2014). A short history of chemometrics: a personal view. *Journal of Chemometrics*, 28(10), 749-760.

Brereton, R. G. (2015). Pattern recognition in chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 149, 90–96

Brescia, M. A., Alviti, G., Liuzzi, V., & Sacco, A. (2003). Chemometric classification of olive cultivars based on compositional data of oils. *Journal of the American Oil Chemists' Society*, 80(10), 945-950.

Brown, S. D. (2017). The chemometrics revolution re-examined. *Journal of Chemometrics*, 31(1), e2856.

Chiesa, L., Panseri, S., Bonacci, S., Procopio, A., Zecconi, A., Arioli, F., Cuevas, F. J., & Moreno-Rojas, J. M. (2016). Authentication of Italian PDO lard using NIR spectroscopy, volatile profile and fatty acid composition combined with chemometrics. *Food Chemistry*, 212, 296-304.

Chung, I. M., Kim, J. K., Yang, J. H., Lee, J. H., Park, S. K., Son, N. Y., & Kim, S. H. (2017). Effects of soil type and organic fertilizers on fatty acids and vitamin E in Korean ginseng (*Panax ginseng* Meyer). *Food Research International*, 102, 265-273.

de Oliveira, C. C., de Araújo Calado, V. M., Ares, G. & Granato, D. (2015). Statistical approaches to assess the association between phenolic compounds and the *in vitro* antioxidant activity of *Camellia sinensis* and *Ilex paraguariensis* teas. *Critical Reviews in Food Science and Nutrition*, 55, 1456-1473.

dos Santos, W. N. L., Sauthier, M. C. S., dos Santos, A. M. P., Santana, D. A., Azevedo, R. S. A., & Caldas, J. C. (2017). Simultaneous determination of 13 phenolic bioactive compounds in guava (*Psidium guajava* L.) by HPLC-PAD with evaluation using PCA and Neural Network Analysis (NNA). *Microchemical Journal*, 133, 583-592.

Erasmus, S. W., Muller, M., Butler, M., & Hoffman, L. C., (2018). The truth is in the isotopes: Authenticating regionally unique South African lamb. *Food Chemistry*, 239, 926-934.

Farag, M. A., Ezzat, S. M., Salama, M. M., & Tadros, M. G. (2016). Anti-acetylcholinesterase potential and metabolome classification of 4 *Ocimum* species as determined via UPLC/qTOF/MS and chemometric tools. *Journal of Pharmaceutical and Biomedical Analysis*, 125, 292-302.

Fidelis, M., Santos, J. S., Coelho, A. L. K., Rodionova, O. Y., Pomerantsev, A., & Granato, D. (2017). Authentication of juices from antioxidant and chemical perspectives: A feasibility quality control study using chemometrics. *Food Control*, 73, 796-805.

Garrido-Delgado, R., Muñoz-Pérez, M. A., & Arce, L. (2018). Detection of adulteration in extra virgin olive oils by using UV-IMS and chemometric analysis. *Food Control*, 85, 292-299.

Giannetti, V., Mariani, M. B., Mannino, P., & Marini, F. (2017). Volatile fraction analysis by HS-SPME/GC-MS and chemometric modeling for traceability of apples cultivated in the Northeast Italy. *Food Control*, 78, 215-221.

Granato, D., de Araújo Calado, V. M., & Jarvis, B. (2014). Observations on the use of statistical methods in food science and technology. *Food Research International*, *55*, 137-149.

Granato, D., Karnopp, A. R., & van Ruth, S. M. (2015). Characterization and comparison of phenolic composition, antioxidant capacity and instrumental taste profile of juices from different botanical origins. *Journal of the Science of Food and Agriculture*, 95 (10), 1997-2006.

Granato, D., Koot, A., Schnitzler, E., & van Ruth, S. M. (2015). Authentication of geographical origin and crop system of grape juices by phenolic compounds and antioxidant activity using chemometrics. *Journal of Food Science*, 80(3), C584-C593.

Granato, D., Margraf, T., Brotzakis, I., Capuano, E., & Ruth, S. M. (2015). Characterization of conventional, biodynamic, and organic purple grape juices by chemical markers, antioxidant capacity, and instrumental taste profile. *Journal of Food Science*, 80(1), C55-C65.

Granato, D., Koot, A., & van Ruth, S. M. (2015). Geographical provenancing of purple grape juices from different farming systems by proton transfer reaction mass spectrometry using supervised statistical techniques. *Journal of the Science of Food and Agriculture*, 95(13), 2668-2677.

Granato, D., Nunes, D. S., & Barba, F. J. (2017). An integrated strategy between food chemistry, biology, nutrition, pharmacology, and statistics in the development of functional foods: A proposal. *Trends in Food Science and Technology*, *62*, 13-22.

Guo, Y., Sun, L., Yu, B., & Qi, J. (2017). An integrated antioxidant activity fingerprint for commercial teas based on their capacities to scavenge reactive oxygen species. *Food Chemistry*, 237, 645-653.

Jandrić, Z., & Cannavan, A. (2017). An investigative study on differentiation of citrus fruit/fruit juices by UPLC-QToF MS and chemometrics. *Food Control*, 72, 173-180.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20,141–151.

Kalaycıoğlu, Z., Kaygusuz, H., Döker, S., Kolaylı, S., & Erim, F. B. (2017). Characterization of Turkish honeybee pollens by principal component analysis based on their individual organic acids, sugars, minerals, and antioxidant activities. *LWT – Food Science and Technology*, 84, 402-408.

Karabagias, I. K., Louppis, A. P., Karabournioti, S., Kontakos, S., Papastephanou, C., & Kontominas, M. G. (2017). Characterization and geographical discrimination of commercial *Citrus* spp. honeys produced in different Mediterranean countries based on minerals, volatile compounds and physicochemical parameters, using chemometrics. *Food Chemistry*, 217, 445-455.

Kaškonienė, V., Ruočkuvienė, G., Kaškonas, P., Akuneca, I., & Maruška, A. (2014). Chemometric analysis of bee pollen based on volatile and phenolic compound compositions and antioxidant properties. *Food Analytical Methods*, 8(5), 1150-1163.

Kasprzyk, I., Depciuch, J., Grabek-Lejko, D., & Parlinska-Wojtan, M. (2018). FTIR-ATR spectroscopy of pollen and honey as a tool for unifloral honey authentication. The case study of rape honey. *Food Control*, 84, 33-40.

Lee, I., Yang, J. (2009). Common clustering algorithms, in: S.D. Brown, R. Tauler, B. Walczak (Eds.), Comprehensive Chemometrics*,* Elsevier, Oxford, England, 2009, pp. 577-618.

Liu, N., Koot, A., Hettinga, K., de Jong, J., & van Ruth, S. M. (2018). Portraying and tracing the impact of different production systems on the volatile organic compound composition of milk by PTR-(Quad)MS and PTR-(ToF)MS. *Food Chemistry*, 239, 201-207.

Lv, H., Zhang, Y., Shi, J., & Lin, Z. (2017). Phytochemical profiles and antioxidant activities of Chinese dark teas obtained by different processing technologies. *Food Research International*, 100, 486-493.

Lund, J. A., Brown, P. N., Shipley, P. R. (2017). Differentiation of *Crataegus* spp. guided by nuclear magnetic resonance spectrometry with chemometric analyses. *Phytochemistry*, 141, 11-19.

Luo, K., Shi, Q., & Feng, F. (2017). Characterization of global metabolic profile of Zhi-Zi-Hou-Po decoction in rat bile, urine and feces after oral administration based on a strategy combining LC–MS and chemometrics. *Journal of Chromatography B*, 1040, 260-272.

Mapelli-Brahm, P., Hernanz-Vila, D., Stinco, C. M., Heredia, F. J., & Meléndez-Martínez, A. J. (2018). Isoprenoids composition and colour to differentiate virgin olive oils from a specific mill. *LWT - Food Science and Technology*, 89, 18-23.

Martínez, E. B., Ramos, E. F., Hernández, N. P., Vallejo, L. G. Z., Ruano, N. V., Ponce, M. V., Mendoza, F. G., & Hernández, A. E. B. (2017). ^{1}H NMR-based metabolomic fingerprinting to determine metabolite levels in serrano peppers (*Capsicum annum* L.) grown in two different regions. *Food Research International*, In Press.

Mehretie, S., Al Riza, D. F., Yoshito, S., & Kondo, N. (2018). Classification of raw Ethiopian honeys using front face fluorescence spectra with multivariate analysis. *Food Control*, 84, 83-88.

Müller-Maatsch, J., Schweiggert, R. M., & Carle, R. (2016). Adulteration of anthocyanin- and betalain-based coloring foodstuffs with the textile dye

‘Reactive Red 195’ and its detection by spectrophotometric, chromatic and HPLC-PDA-MS/MS analyses. *Food Control*, 70, 333-338.

Munck, L., Nørgaard, L., Engelsen, S. B., Bro, R., & Andersson, C. A. (1998). Chemometrics in food science – a demonstration of the feasibility of a highly exploratory, inductive evaluation strategy of fundamental scientific significance. *Chemometrics and Intelligent Laboratory Systems*, 44, 31-60.

Nayik, G. A., & Nanda, V. (2016). A chemometric approach to evaluate the phenolic compounds, antioxidant activity and mineral content of different unifloral honey types from Kashmir, India. *LWT - Food Science and Technology*, 74, 504-513.

Nunes, C. A., Alvarenga, V. O., Sant'Ana, A. S., Santos, J. S., & Granato, D. (2015). The use of statistical software in food science and technology: Advantages, limitations and misuses. *Food Research International,* 75, 270–280.

Oliveri, P., & Downey, G. (2012). Multivariate class modeling for the verification of food authenticity claims. *Trends in Analytical Chemistry*, 35, 74-86.

Oliveri, P., & Simonetti, R. (2016). Chemometrics for Food Authenticity Applications. In: Downey, G. Advances in Food Authenticity Testing. Amsterdam: Elsevier. 1st ed., p. 701-728.

Opatić, A. M., Nečemer, M., Lojen, S., Masten, J., Zlatić, E., Šircelj, H., Stopar, D., Vidrih, R. (2018). Determination of geographical origin of commercial tomato through analysis of stable isotopes, elemental composition and chemical markers. *Food Control*, doi.org/10.1016/j.foodcont.2017.11.013

Paneque, P., Morales, M. L., Burgos, P., Ponce, L., & Callejón, R. M. (2017). Elemental characterisation of Andalusian wine vinegars with protected designation of origin by ICP-OES and chemometric approach. *Food Control*, 75, 203-210.

Popek, S., Halagarda, M., & Kursa, K. (2017). A new model to identify botanical origin of Polish honeys based on the physicochemical parameters and chemometric analysis. *LWT - Food Science and Technology*, 77, 482-487.

Qannari, E. M. (2017). Sensometrics approaches in sensory and consumer research. *Current Opinion in Food Science*, 15, 8-13.

Ropodi, A. I., Panagou, E. Z., & Nychas, G. J. E. (2016). Data mining derived from food analyses using non-invasive/nondestructive analytical techniques; determination of food authenticity, quality & safety in tandem with computer science disciplines. *Trends in Food Science and Technology*, 50, 11-25.

Santos, J. S., Deolindo, C. T. P., Fujita, A., Genovese, M. I., Daguer, H., Valese, A., Marques, M. B., Rosso, N. D., & Granato, D. (2016). Effects of time and extraction temperature on phenolic composition and functional properties of red rooibos (*Aspalathus linearis*). *Food Research International*, 89, 476-487.

Szymanska, E., Gerretzen, J., Engel, J., Geurts, B., Blanchet, L., & Buydens, L. M. C. (2015). Chemometrics and qualitative analysis have a vibrant relationship. *Trends in Analytical Chemistry*, 69, 34–51.

Tavares, K. M., Lima, A. R., Nunes, C. A., Silva, V. A., Mendes, E., Casal, S., & Pereira, R. G. F. A. (2016). Free tocopherols as chemical markers for Arabica coffee adulteration with maize and coffee by-products. *Food Control*, 70, 318-324.

Tian, Y., Yan, C., Zhang, T., Tang, H., Li, H., Yu, J., Bernard, J., Chen, L., Martin, S., Delepine-Gilon, N., Bocková, J., Veis, P., Chen, Y., & Yu, J. (2017). Classification of wines according to their production regions with the contained trace elements using laser-induced breakdown spectroscopy. *Spectrochimica Acta Part B: Atomic Spectroscopy*, 135, 91-101.

Torkashvand, A. M., Ahmadi, A., & Nikravesh, N. L. (2017). Prediction of kiwifruit firmness using fruit mineral nutrient concentration by artificial neural network (ANN) and multiple linear regressions (MLR). *Journal of Integrative Agriculture*, 16(7), 1634-1644.

Viapiana, A., Struck-Lewicka, W., Konieczynski, P., Wesolowski, M., & Kaliszan, R. (2016). An approach based on HPLC-fingerprint and chemometrics to quality consistency evaluation of *Matricaria chamomilla* L. commercial samples*. Frontiers in Plant Science*, 7, 1-11.

Wang, X., Zeng, Q., Contreras, M. M., & Wang, L. (2017). Profiling and quantification of phenolic compounds in *Camellia* seed oils: Natural tea polyphenols in vegetable oil. *Food Research International*, 102, 184-194.

Zhu, W., Wang, X., & Chen, L. (2017). Rapid detection of peanut oil adulteration using low-field nuclear magnetic resonance and chemometrics. *Food Chemistry*, 216, 268-274.

FIGURE HEADINGS

Figure 1: Summary of selected multivariate statistical methods applied in food research.

Figure 2: PCA of juice samples based on chemical composition and antioxidant activity: A – represents the number of PCs e the explained variance. B- represents the projection of samples on the factor-plane. For illustration

purposes, red starts represent orange juice, green stars represent lemon juices, and violet stars represent grape juices.

Figure 3: Principal components analysis, PCA, to project different samples (i.e., fruits from different varieties) based on some selected responses: outlier detection with no separation between varieties (A), no outliers with a clear separation between fruit varieties (B).

Figure 4: HCA dendrogram for agglomerative algorithm (A) and divisive algorithm grouping flow (B).

Figure 5: Example of PCA (A) and HCA (B, C) applied to a data set composed of n = 20 fruit samples (A and B) according to the concentrations of total phenolics, carotenoids, antioxidant activity measured by the ORAC assay, and inhibition of lipase and α-amylase.

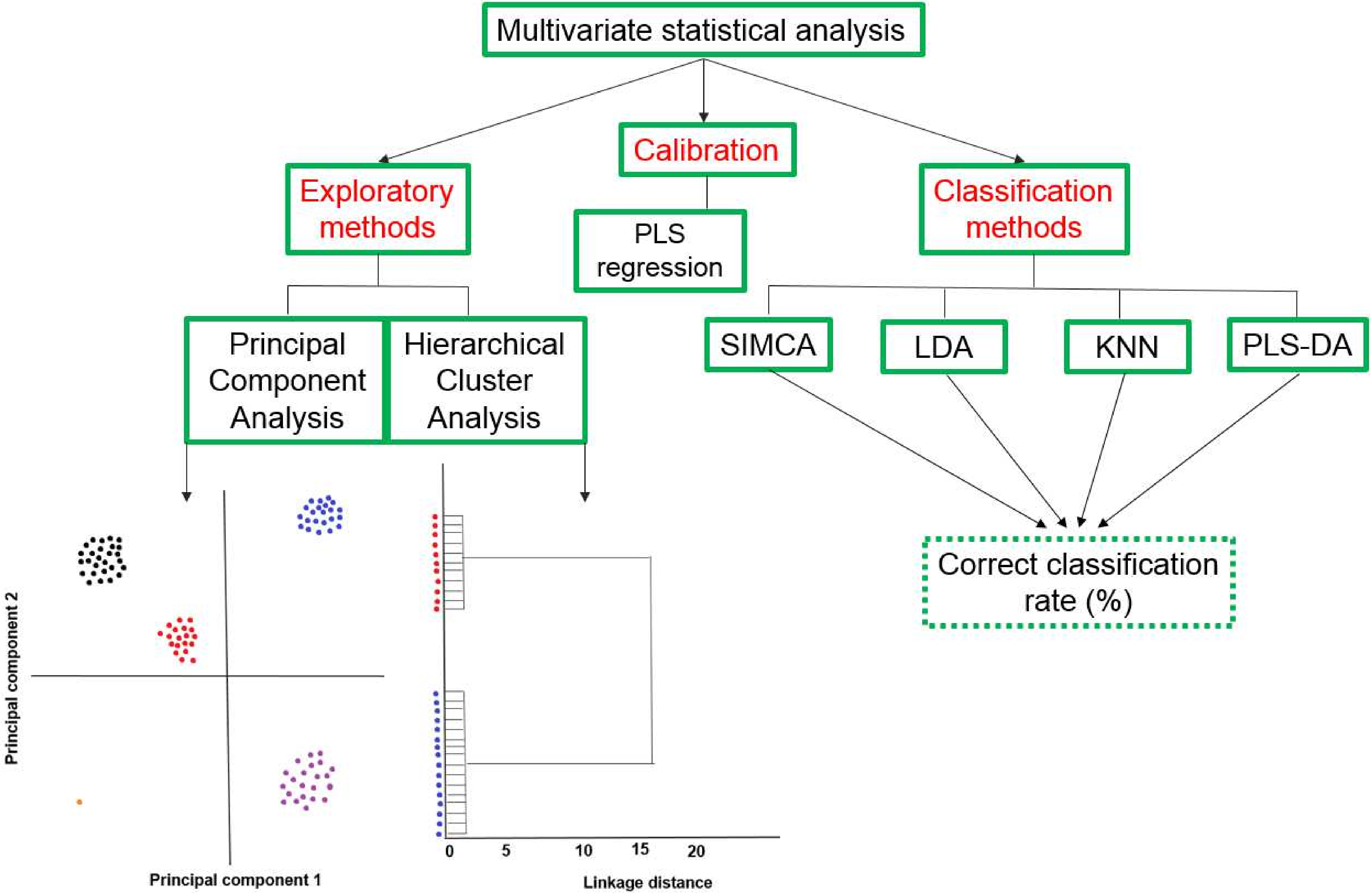
Table 1: Factor loadings for illustrating the interpretation of Figure 2.

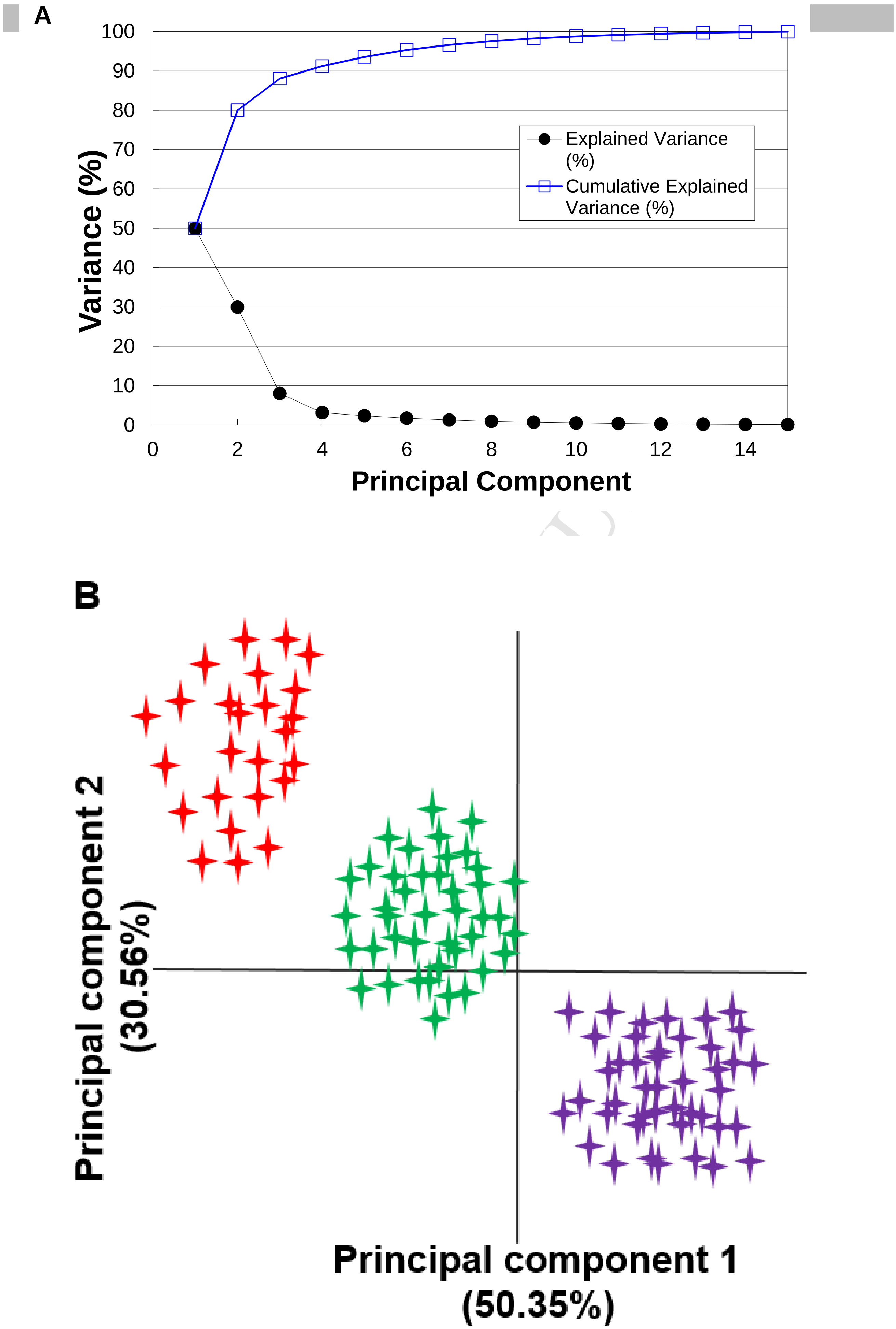
Factor	PC1	PC2	PC3	PC4
DPPH	**0.69**	-0.47	0.16	-0.42
ABTS	**0.68**	0.06	-0.40	0.44
FRAP	**0.63**	**-0.65**	-0.12	-0.02
Gallic acid	0.50	**-0.66**	0.09	-0.30
Caffeic acid	**0.81**	-0.23	0.22	0.15
5-*O*-caffeoylquinic acid	0.04	**-0.70**	-0.50	0.27
(+)-Epicatechin	**-0.75**	-0.54	-0.30	-0.09
(+)-Catechin	**-0.90**	-0.07	0.08	-0.17
Quercetin	**-0.90**	-0.19	0.03	-0.08
Quercetrin	-0.52	-0.26	**0.70**	-0.36
Luteolin	-0.78	-0.05	**-0.65**	0.09
Ellagic acid	0.13	0.48	-0.46	**-0.73**
Eigenvalue	5.39	3.78	0.56	0.23
Explained variance (%)	50.35	30.56	8.05	3.18

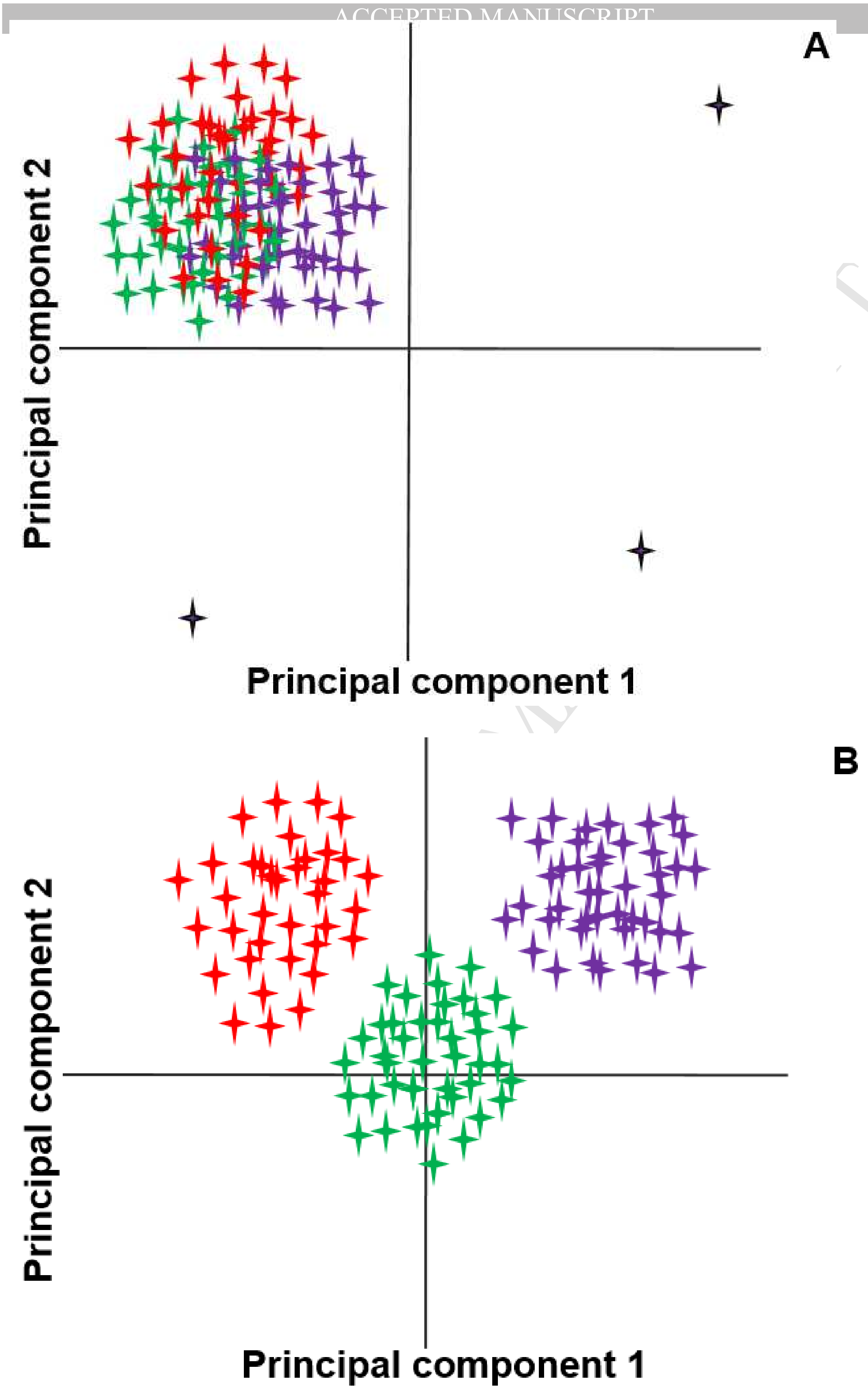
Table 2: Illustrative correlation coefficients to help in the interpretation of the example shown in Figure 2.

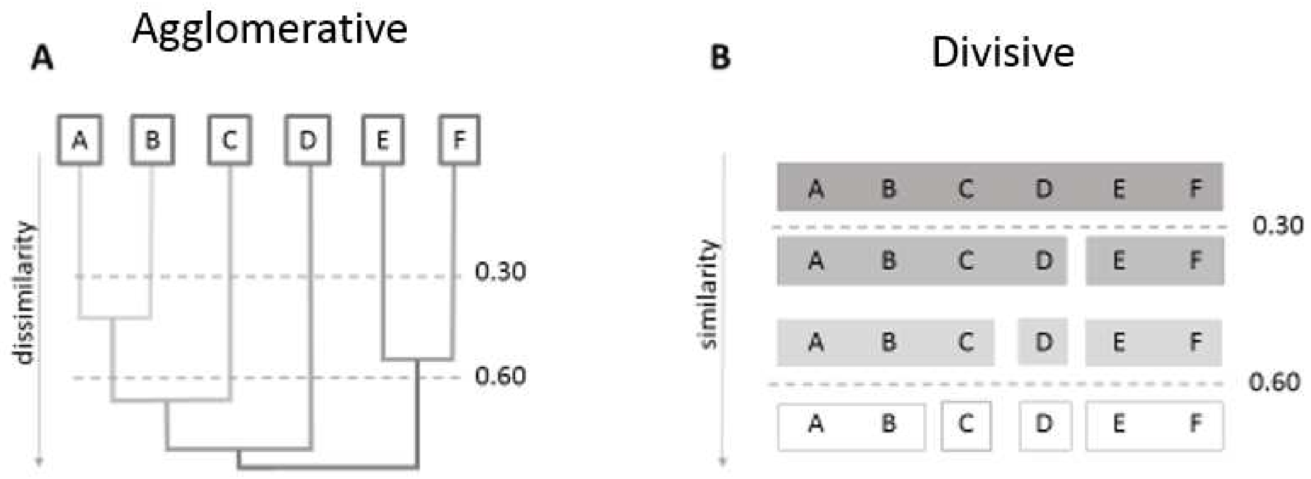
Responses	DPPH	ABTS	FRAP
DPPH	1		
ABTS	0.899	1	
FRAP	0.946	0.947	1
Gallic acid	0.564*	0.529*	0.608
Caffeic acid	0.895	0.911	0.935
5-O-caffeoylquinic acid	0.523*	0.518*	0.622
(+)-Epicatechin	0.875	0.812	0.804
(+)-Catechin	0.926	0.874	0.935
Quercetin	0.873	0.924	0.901
Quercetrin	0.425*	0.378*	0.333*
Luteolin	0.788	0.829	0.845
Ellagic acid	0.238*	0.356*	0.458*

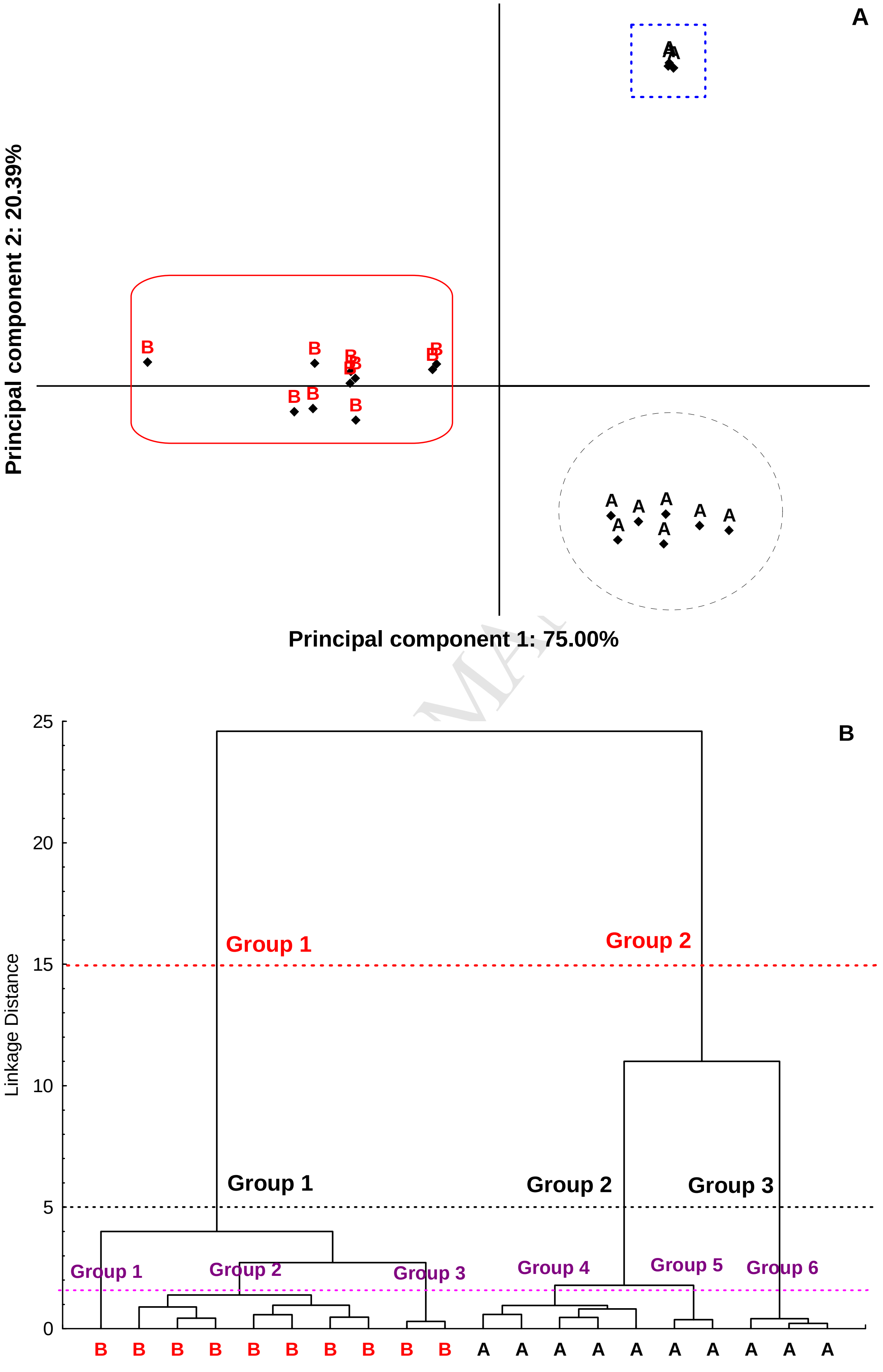
Note: * denotes $p > 0.05$ while the other correlation coefficients present $p < 0.05$.

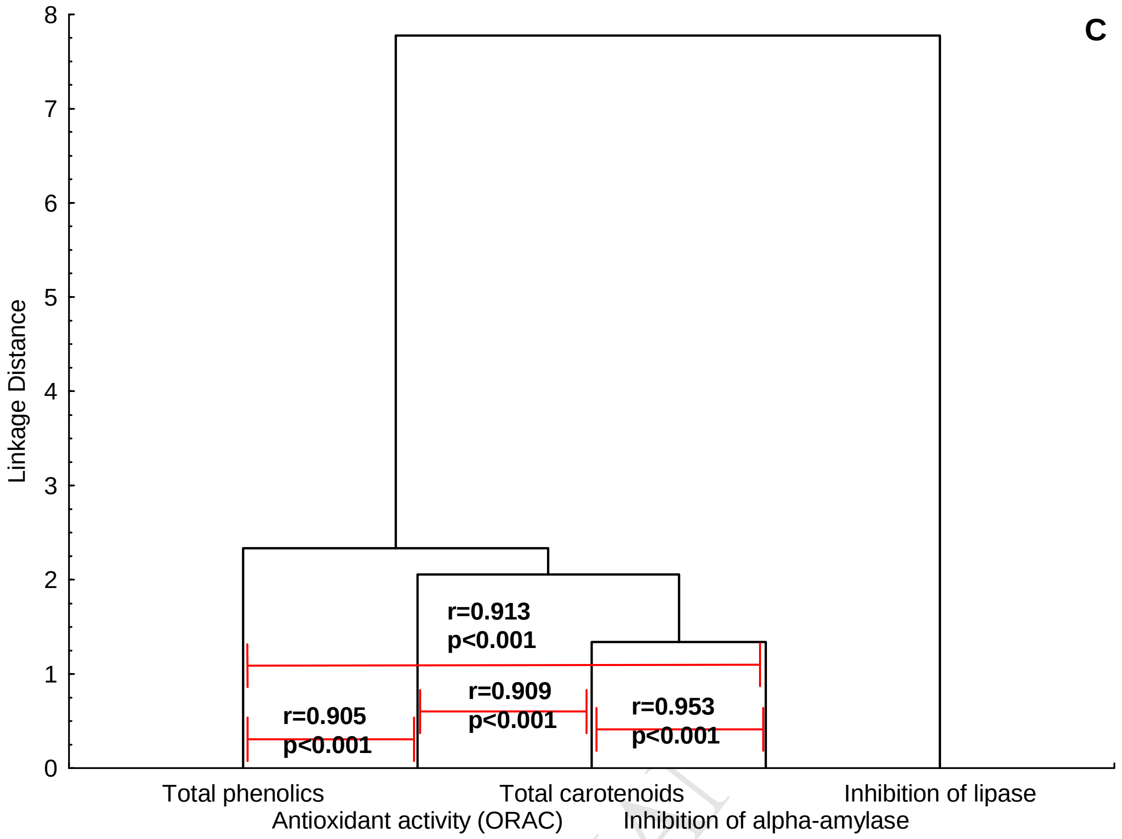

Multivariate statistical analysis
Exploratory methods
Calibration
Classification methods
PLS regression
Principal Component Analysis
Hierarchical Cluster Analysis
SIMCA
LDA
KNN
PLS-DA
Correct classification rate (%)
Principal component 2
Principal component 1
0
5
10
15
20
Linkage distance


A
Variance (%)
100
90
80
70
60
50
40
30
20
10
0
0
2
4
6
8
10
12
14
Principal Component
● Explained Variance (%)
□ Cumulative Explained Variance (%)
B
Principal component 2 (30.56%)
Principal component 1 (50.35%)


A
Principal component 2
Principal component 1
B
Principal component 2
Principal component 1


A
Agglomerative
A
B
C
D
E
F
dissimilarity
0.30
0.60
B
Divisive
similarity
A B C D E F
0.30
A B C D
E F
A B C
D
E F
0.60
A B
C
D
E F


A
Principal component 2: 20.39%
Principal component 1: 75.00%
B
B
B
B
B
B
B
B
B
A
A
A
A
A
A
A
A
A
B
25
20
15
10
5
0
Linkage Distance
Group 1
Group 2
Group 1
Group 2
Group 3
Group 1
Group 2
Group 3
Group 4
Group 5
Group 6
B B B B B B B B B B A A A A A A A A A A


C
Linkage Distance
8
7
6
5
4
3
2
1
0
r=0.913
p<0.001
r=0.909
p<0.001
r=0.953
p<0.001
r=0.905
p<0.001
Total phenolics
Total carotenoids
Inhibition of lipase
Antioxidant activity (ORAC)
Inhibition of alpha-amylase

HIGHLIGHTS

- Chemometric tools are widely used for classification, calibration and exploratory issues
- Unsupervised statistical methods are used to study data structure and look for clusters of samples
- PCA and CA are the most widely used methods
- PCA and CA can be useful in studies regarding bioactive compounds in foods
- We criticize the indiscriminate use of PCA and CA