JOURNAL OF CHEMICAL EDUCATION

pubs.acs.org/jchemeduc

Laboratory Experiment

Teaching Principal Component Analysis Using a Free and Open Source Software Program and Exercises Applying PCA to Real-World Examples

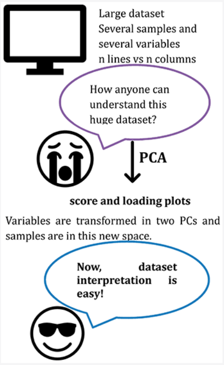
Laís Feltrin Sidou and Endler Marcel Borges*

Cite This: *J. Chem. Educ.* 2020, 97, 1666−1676

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Principal component analysis (PCA) is one of the most important and powerful methods in chemometrics as well as in a wealth of other areas. Running a PCA results in two main elements, the score plot and the loading plot; the score plot provides the location of the samples, and the loading plot indicates correlations among variables, the trends in the grouping of samples, and the most important variables. In the past 10 years teaching chemometrics, we have struggled with not having free software with an easy to use graphical user interface for data handling and calculations. In this paper, we provide a series of examples that students used to carry PCA in R-Project, a free and open source software program. In the first example, students used PCA to find correlations among chemical properties of chemical elements and relate these properties with the periodic distribution of the elements. In the second example, meat samples were grouped using 14 variables, and students could observe how outlier samples might influence the PCA model; in this case, they were also taught how to use the *t* test to choose the variables that were significant to the PCA model. In the third example, healthy patients were differentiated from diabetic patients using 163 lipid concentrations. In the fourth example, Atlantic salmon samples were differentiated from catfish samples. In the fifth and sixth examples, students were able to observe how data treatment affects the classification of natural products and edible oils, respectively.


Large dataset
Several samples and several variables
n lines vs n columns
How anyone can understand this huge dataset?
PCA
score and loading plots
Variables are transformed in two PCs and samples are in this new space.
Now, dataset interpretation is easy!

KEYWORDS: *Upper-Division Undergraduate, Analytical Chemistry, Chemoinformatics, Interdisciplinary/Multidisciplinary, Calculator-Based Learning, Chemometrics, Computational Chemistry*

■ INTRODUCTION

Principal Component Analysis (PCA) is a versatile tool for the interpretation of large data sets,[1,2] in which a data set is transformed into a set of orthogonal variables (components) that account for the greatest degree of variability in the data.[3] PCA is also an important and powerful method in chemometrics as well as in many other areas.[2,4,5]

PCA can be done using a variety of commercially available software, such as Minitab, SAS, Unscrambler, IBM SPSS, Pirouette, Statistica, and Matlab.[6,7] However, in the past 10 years teaching chemometrics, we have struggled with not having free and easy to use software for data handling and calculations.

In this paper, we present a method for doing the PCA analysis using R, a light-weight, free, and open source cross-platform piece of software, and we offer an addition to already existing laboratory experiments dealing with PCA analysis.[4,8−19]

Here, we teach and present PCA using several examples, where these examples may be carried out by students using R and data sets provided in the Supporting Information.

■ STUDENTS' LEARNING GOALS

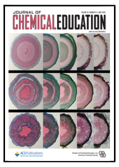
The mathematical background of PCA and its applications in many fields have been previously reviewed by several authors.[20−25] Here, our objective is to run PCA in several data sets using the R project. As a result of performing PCA in those data sets, it is expected that students

- Understand the basic principles of PCA
- Identify the similarities and patterns in the score plots
- Identify correlations and divergences among variables in the loading plots
- Understand the influence of each variable in the position of objects in the score plot.
- Understand and use the PCA

■ MATERIALS AND METHODS

In Table S1, 5 atomic properties [atomic mass (u), electronegativity (Pauling scale), atomic radius (van der Waals; ppm), ionization energy (eV), electron affinity (eV)] of 35 elements (representative elements excluding noble gases) were taken from PubMed's periodic table.[26] In Table S2 (Supporting Information), there are 8 atomic properties [atomic mass (u),

Received: October 1, 2019
Revised: April 25, 2020
Published: May 22, 2020



https://dx.doi.org/10.1021/acs.jchemed.9b00924

electronegativity (Pauling scale), atomic radius (van der Waals; ppm), ionization energy (eV), electron affinity (eV), melting point (K), boiling point (K), and density (g/mL)] for 63 elements, which are transition elements and representative elements, excluding noble gases, lanthanides, and actinides.[26]

In Table S3, data was taken from the Brazilian Table of Food Composition (*Tabela Brasileira de Composição de Alimentos*, TACO).[27,28] In Table S3, there are 15 variables and 71 samples, further divided into bovine (48 samples) and chicken (23 samples). Variables were water and ash percentage, protein, lipids, cholesterol, chemical elements (Ca, Mg, Mn, P, Fe, Na, K, Cu, and Zn), and niacin concentrations. Table S4a contains the same data as Table S3 excluding samples 355, 356, 384, 385, 400, 415, 416, and 417. Table S4b contains the same data as Table S4a excluding 8 out of 17 variables

In Table S5, data was taken from the Supporting Information of Xuan et al.'s paper.[29] It contains 163 lipid concentrations for 30 healthy patients (control) and 30 diabetes patients.[29] Table S5 contains the same data as Table S6, excluding variables with low importance for the PCA model.

In Table S7, there are 73 samples and 68 variables. Samples are of catfish (32 samples) and Atlantic salmon (41). Abbreviations and units of all variables are shown in Table S8.[30]

Tables S9 and S10 were taken from Appendix 2 and Appendix 1, respectively, of Chen et al.'s paper.[31]

Table S11 was taken from Yeh's work.[32] Table S12 contains the same data as Table S11 excluding canola samples.

PCA was carried out using Rcommander, which is a free and open source plugin that provides new functionalities to R. Several questions and exercises were provided in the test questions in Supporting Information.

Reference 33 shows how data can be loaded into R and how to run PCA on it. R can be obtained from CRAN.[34] Reference 35 shows how to obtain, install, and use R. R can be used in Windows, Linux, and MacOS.

■ EXPERIMENTAL OVERVIEW

Using R's basic command line interface might be challenging for students with no prior experience.[32] However, RCommander provides a graphical interface that made students comfortable when using the software, and they claimed that the plugin provided an easy and friendly interface. This study was incorporated into the Advanced Analytical Chemistry course in the Chemistry graduate program in our university.

This experiment was carried out with graduate students in the second semesters of the years 2015−2018; students worked alone using their own laptops. The examples provided in this paper were carried out in a 200 min class. After the class, students worked on the test questions (Supporting Information) as homework, and responses provided by students were worked on in another 200 min class. We evaluated students' learning by their performance in the test questions.

The study began working PCA with the properties of chemical elements, where students could observe the grouping of chemical elements (objects) in the score plot according to the periodic table's classification. They could also correlate the position of chemical elements in the score plot to chemical properties by looking at the loading plot.

In the following examples, students were able to observe again the correlation between objects' position in the score plot with the variables in the loading plot. The score plot also shows correlations between variables and the importance of each variable in sample grouping and sorting.

During the execution of experiments, students were taught how to recognize outliers and understand their effects on PCA, like how the exclusion of outlier samples changes the importance and correlations among variables in the loading plot.

All the examples were carried out autoscaling the data, but some of them were also carried out without data scaling, where students could observe that variables with larger numerical value dominate the first PC and, normally, more than 90% of the total information is retained by PC 1.

■ HAZARDS

This "dry lab" involves no hazards.

■ RESULTS AND DISCUSSION

Grouping 35 Chemical Elements from Block A Using 5 Chemical Properties

The introduction of general information and chemometric concepts was carried out using Besalú's paper.[10] There are several papers in this *Journal* dealing with the periodic table and grouping elements together by similarities in their chemical properties.[36−40] Using atomic properties to teach PCA has a high pedagogical value,[10] since students are familiar with the periodic table and, with the analysis, are able to observe that chemical elements are grouped together according to similarities in their properties.

In the first place, data had to be autoscaled, since some variables were measured in much larger quantities than others. For example, atomic radius is within the 123−343 range, and atomic mass is within the 1−209 range, whereas electronegativity is within the 0.79−3.98 range and electron affinity is within the 0.27−3.33 range. If these scale differences are not properly handled, PCA will only focus on higher numbers.[2,41,42] Autoscaling the data adjusts all columns to the same "size", giving them an equal opportunity to be modeled. Autoscaling means that, from each variable, the mean value is subtracted, and then the variable is divided by its standard deviation. Therefore, our data was autoscaled before the PCA model was built.[2,41,42]

We began the discussion about PCA applying it to 5 atomic properties of 35 elements (Table S1 in Supporting Information). The chemical elements analyzed were from block A (the representative ones) without noble gases. The outcomes of the analysis in R showed correlations among variables (Table 1), and correlations among ionization energy, atomic radius, electron affinity, and electronegativity could be observed.

Table 1. Correlations among 5 Variables for 35 Elements from Block A[a]

Chemical Properties	Atomic Mass	Atomic Radius	Electron Affinity	Electronegativity
Atomic radius	0.31			
Electron affinity	−0.09	−0.34		
Electronegativity	−0.13	−0.72	0.73	
Ionization energy	−0.33	−0.74	0.71	0.92

[a] See Table S1 in the Supporting Information

Correlations shown in Table 1 were already expected by the students. Atomic radius was negatively correlated with electronegativity and ionization energy, because a higher atomic radius represents lower ionization energy and electronegativity.

Principal component analysis employs a mathematical procedure that transforms a set of possibly correlated response variables into a new set of noncorrelated variables, called principal components, PCs.[43] In Figure 1, five atomic properties (i.e., variables) were reduced by a projection of the chemical elements (i.e., objects) into a smaller number of new variables called principal components (PCs). PCs are oriented so that the first PC describes as much original variation as possible between the objects (elements). The second PC is oriented in an orthogonal manner to the first and describes as much of the remaining variation as possible.

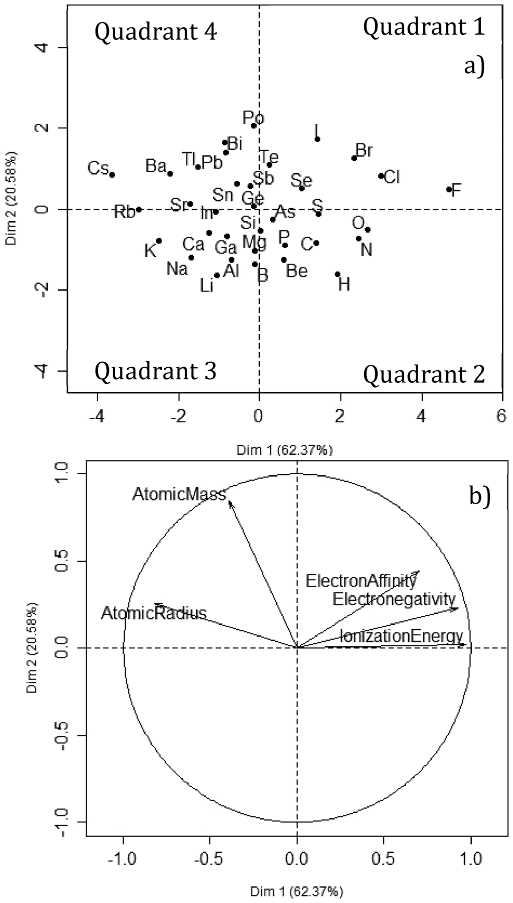


Figure 1. PCA using 5 atomic properties of 35 elements: (a) score plot and (b) loading plot.

PCA is a multivariate method of analysis based on data reduction considering the correlation among the data.[6] Here, the data was well correlated (Table 1); 82.951% of the total information was explained with 2 PCs, and 62.370% of the information was explained by PC1 alone (Table 2).

Table 2. Percentage of Information Retained by Each PC

Principal Components (PCs)	PC1	PC2	PC3	PC4	PC5
% of Var	62.37	20.58	11.90	3.97	1.18
Cumulative % of Var	62.37	82.95	94.85	98.82	100.00

In summary, the application of PCA provides two main elements: the scores and loadings.[43] The loadings are the space constructed with the original variables, showing which variables are important to explain the trends in the grouping of samples.[43]

In Figure 1a, the score plot shows the projection of the data (elements) along with the principal components (PCs).[43] In the score plot (Figure 1a), we are able to see differences and similarities among the objects (i.e., elements): the distance between two chemical elements in the score plot represents the similarity between elements.[1] The amount of each of the original variables included in the PC is described by the loading (Figure 1b). By plotting the loadings for the two PCs, it is possible to assess the relative importance of each of the variables (i.e., atomic properties): the further the variable is from the origin, the more important it is. Correlations among variables are observed in the loading plot (positively correlated variables will be located close together or inversely correlated variables will be at 180° to one another).[1]

PCA must be interpreted using both the score plot (Figure 1a) and the loading plot (Figure 1b). The loading plot shows the importance of original variables in each PC. R also provides information about the importance of each original variable in PCs (Table 3). Table 3 shows that electronegativity, ionization energy, and atomic radius have a higher impact on PC1, while being negligible to PC2. On the other hand, atomic mass was negligible to PC1 and had a higher impact on PC2.

Table 3. Percentages of Each Variable in Each PC

Principal Component (PC)	PC1	PC2	PC3	PC4	PC5
Atomic mass	4.89	69.43	21.67	1.19	2.81
Electronegativity	27.44	5.16	3.43	23.85	40.13
Atomic radius	21.62	6.49	29.41	42.23	0.24
Ionization energy	30.12	0.04	0.17	13.13	56.54
Electron affinity	15.94	18.88	45.32	19.59	0.27

The importance of each original variable in each PC (Table 3) can be visualized in the loading plot (Figure 1b), where we can see that the atomic mass vector has a lower component in PC1 and a higher component in PC2, i.e., a higher impact on PC2 and a lower one on PC1. In the same way, electronegativity, atomic radius, ionization energy, and electron affinity vectors have lower components in PC2 and higher components in PC1, which means a higher impact on PC1 and a lower one on PC2.

The loading plot (Figure 1b) shows correlations among the original variables. In quadrant 1, electronegativity, ionization energy, and electron affinity vectors had closer angles, which means that these variables were correlated. In quadrant 4, atomic radius was inversely correlated to these variables, and they had angles close to 180°. These correlations were also observed in Table 1.

The location of elements in the score plot (Figure 1a) was directly related to the loading plot (Figure 1b). The loading plot shows that elements with higher atomic weight were placed at the top of the score plot, while elements with lower atomic weight were placed at the bottom. Elements with higher atomic radius and lower electronegativity, ionization energy, and electron affinity were placed on the left-hand side, while elements with lower atomic radius and higher electronegativity, ionization energy, and electron affinity were placed on the left-hand side.

In the score plot (Figure 1a), we can see that halogens were placed on the left-hand side and alkaline metals and alkaline earths were placed on the right-hand side. Polonium, bismuth, and lead, which have high atomic weight, were placed at the top of the score plot, while elements of the second period (beryllium, boron, oxygen, nitrogen, and lithium) and hydrogen were placed at the bottom.

Table 4. Correlations among 8 Variables for 64 Elements from Block A and B[a]

Chemical Properties	Atomic Mass	Atomic Radius	Boiling Point	Density	Electron Affinity	Electronegativity	Ionization Energy
Atomic radius	0.29						
Boiling point	0.33	0.10					
Density	0.80	0.05	0.68				
Electron affinity	0.01	−0.31	−0.41	−0.07			
Electronegativity	0.00	−0.61	−0.22	0.06	0.73		
Ionization energy	−0.19	−0.64	−0.34	−0.13	0.70	0.90	
Melting point	0.23	0.03	0.93	0.61	−0.30	−0.08	−0.16

[a] See Table S2 in the Supporting Information.

Metalloids are elements with chemical properties between those of a nonmetal and a metal. In Figure 1a, metalloids were placed between metals and nonmetals in accordance with their periodic properties.

Grouping 63 Chemical Elements from Blocks A and B Using 5 Chemical and 3 Physical Properties

In the next step, PCA was applied to 64 representative elements, excluding noble gases, actinides, and lanthanides, using 8 chemical and physical properties: atomic weight, atomic radius, ionization potential, electron affinity, electronegativity, melting point, boiling point, and density (Table S2 in Supporting Information).

Table 4 shows that, in Table S2, correlations between variables were lower than those in Table S1. Furthermore, the reduction of 8 variables in two PCs results in more information loss than our previous example.

Since this data was less correlated than the previous data, the PCA model (Figure 2) explains only 71.37% of total information using two PCs. At this point, students were tempted to use more than two PCs, but useful information was obtained observing the PC1 vs PC2 plot, and using PC3 did not provide any useful information since it explained only 15.65% of the total variance (Table 5).

In Figure 3, using PC1 vs PC3, only nonmetal and halogens were separated, while transition metals, post-transition metals, alkaline metals, and earth metals were not separated.

In the loading plot (Figure 2b) in quadrant 4, we can observe a high correlation among electron affinity, ionization energy, and electronegativity. In quadrant 2, we can observe that atomic radius is inversely correlated to those chemical properties. Thus, in the score plot (Figure 2a), halogens, oxygen, and nitrogen were placed on the left-hand side, since these elements have a high electron affinity, ionization energy, and electronegativity and they also have a small atomic radius. Alkaline metals and alkaline earths were placed at the bottom of quadrant 2, since these elements have a high atomic radius and low electron affinity, ionization energy, and electronegativity.

In the loading plot (Figure 2b), we can observe a high correlation among atomic mass, melting point, boiling point, and density. Thus, period 6 metals (transition) were placed at the top of quadrant 1 (Figure 2a), rhenium, tantalum, tungsten, osmium, and iridium, which have high density and atomic weight and are also very resistant to corrosion. For example, osmium is the densest of all the elements, and it is twice as dense as lead.[44]

Gold and platinum were placed at the top side of Figure 2a; both elements are very resistant to corrosion and also have high density and atomic weight but present lower melting points than rhenium, tantalum, tungsten, osmium, and iridium.[44] Thus, gold and platinum were placed on the left side of these elements.

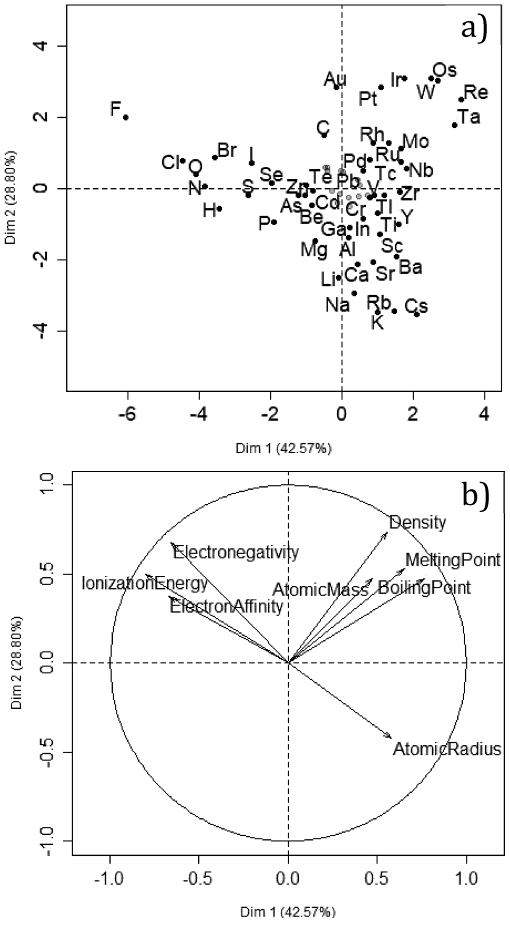


Figure 2. PCA (PC1 vs PC2) using the 8 atomic properties of 64 elements: (a) score plot and (b) loading plot.

Table 5. Percentage of Information Retained by Each PC

Principal Component (PC)	PC1	PC2	PC3	PC4	PC5
% of Var	42.57	28.80	15.64	6.79	3.48
Cumulative % of Var	42.57	71.37	87.01	93.80	97.28

In quadrant 1 of the score plot (Figure 2a), period 5 elements (niobium, molybdenum, technetium, ruthenium, rhodium, palladium) were placed under period 6 elements.

In quadrant 2, aluminum was placed close to gallium, which represents the similarity between these elements. Scandium and yttrium were placed close to aluminum and gallium because aluminum and scandium also have similar properties.

In Figure 1, 5 variables were reduced to 2 PCs, and 83% of the original information was retained, while, in Figure 2, 8 variables were reduced to 2 PCs and only 71.37% of the total information

https://dx.doi.org/10.1021/acs.jchemed.9b00924

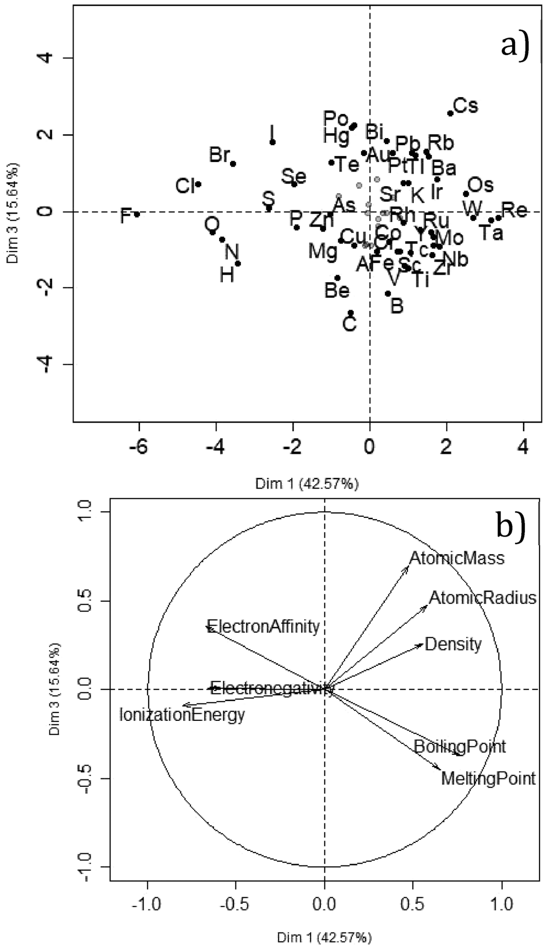


Figure 3. PCA (PC1 vs PC3) using the 8 atomic properties of 64 elements: (a) score plot and (b) loading plot.

was retained. Therefore, variable reduction to PCs represents some information loss, and using PCA with highly uncorrelated variables results in higher information loss.

Meat Differentiation

In this section, Table S3 (Supporting Information) was handed to the students, and they differentiated bovine meat from chicken meat. Table S3 has 71 lines (objects) and 15 columns (variables), and at first glance, it is quite complicated to get an overview of what kind of information was available in the data. Therefore, PCA provided the tool needed to get the new variables which best explain the variation in the whole data set.[41]

When students ran the PCA for this data set (Figure 4), they were able to see that only 43.66% of the total information was retained in PC1 and PC2 due to the uncorrelated data.

In the score plot (Figure 4a), meat samples were not grouped. In quadrant 2, samples 400 (chicken liver) and 355 and 356 (bovine liver) were placed far from other meat samples due to their high Fe, Cu, P, K, cholesterol, niacin, and water levels and low lipid, ash, and Na levels. On the top side, samples 415, 416, and 417 (hamburgers) and 384 and 385 (dried meats) were placed far from the other samples due to their high lipid, Na, and ash levels and low water levels. These samples were very different from the others, and they were called outliers. Outlier samples strongly affect the PCA's results, and their exclusion from the data set can result in a better distribution of samples. After the exclusion of outlier samples from the data set (Table S4a in Supporting Information), it is possible to visualize a new sample distribution in the score plot (Figure 5a). However, meat samples were not grouped. In the loading plot (Figure 5b), the distribution of original variables in PC1 and PC2 dramatically changed after the exclusion of outliers from the data set, which also slightly affected the correlation between variables and total information described by PC1 and PC2 to 48.81%. In the loading plot (Figure 5b), the chemical elements levels (Ca, Zn, Na, Mn, Fe, Cu) were highly correlated, and these variables were inverse correlated with water levels.

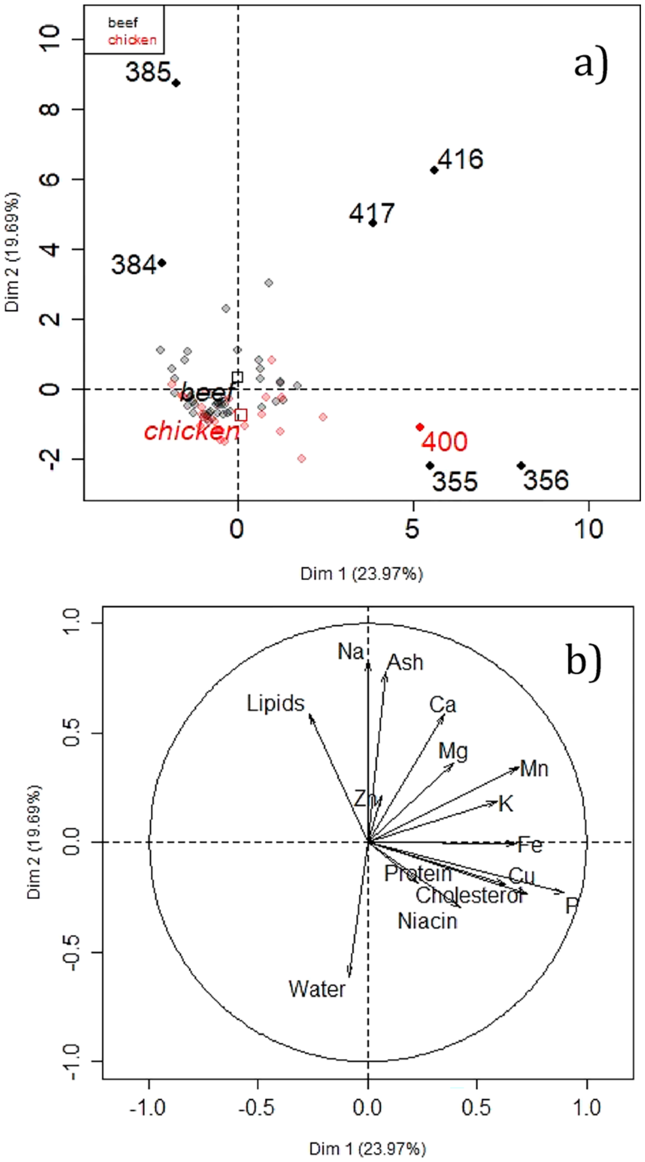


Figure 4. PCA using the 15 chemical properties of bovine (48 samples) and chicken (23 samples) meats: (a) score plot and (b) loading plot.

Here, students could observe that the exclusion of outlier samples strongly affects the results of the analysis and that PCA was not a good choice for highly uncorrelated data. Having that in mind, what could be done?

One of the answers lies in removing variables which do not provide relevant information using hypothesis tests.

The *F* test and *t* test compare the variance and means of concentrations, respectively, obtained with bovine and chicken samples to determine whether there is statistical evidence that the associated means are significantly different ($p < 0.05$).[45−48]

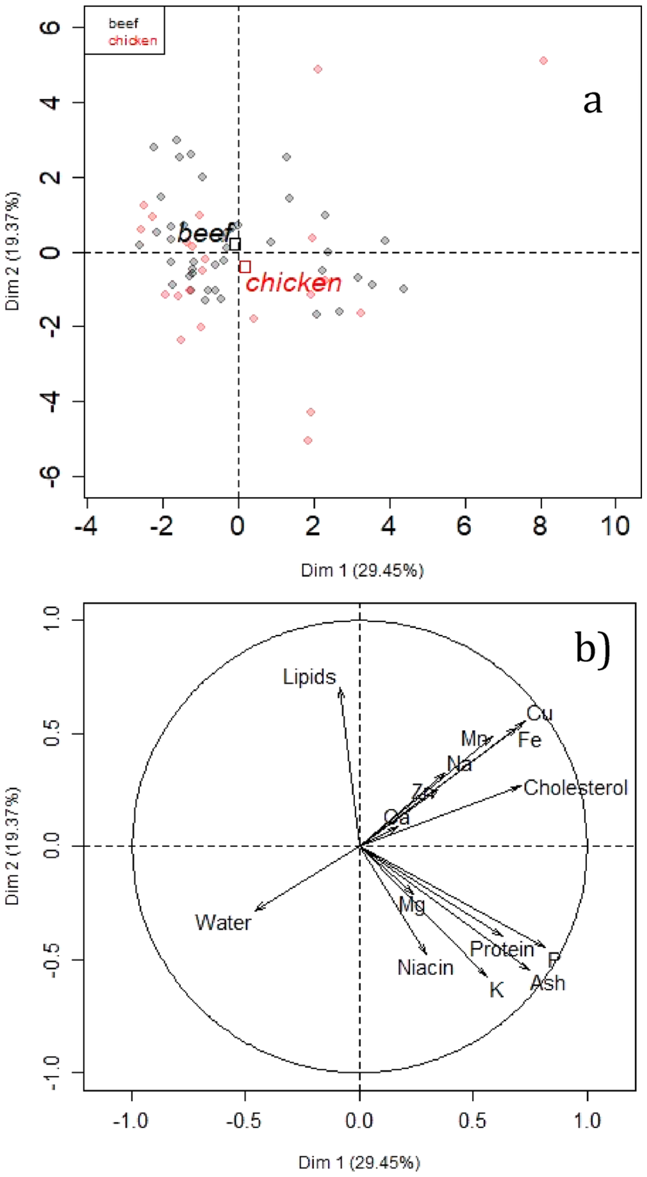


Figure 5. PCA after exclusion of outlier samples: (a) score plot and (b) loading plot.

You can do an *F* test, *t* test, and several other hypothesis tests in R. Nevertheless, we do hypothesis tests using a spreadsheet. You can use functions f.test(MatrixA;MatrixB) and t.test(MatrixA;MatrixB;tails;) in Excel to do hypothesis tests in a simple and fast manner. For example, in Table S4a, you may use an *F* test to see whether two standard deviations are the “same” or whether they are “different”. Entering f.test(C2:C42;C43:C64) in C66, it returns 0.18 which means that there was an 18% probability of observing the water content variance of chicken meat in bovine meat (beef) and vice versa. Thus, assuming a 95% confidence interval ($p = 0.05$), we assume that variances of water percentages in beef and chicken samples are equivalent.

We can also use a *t* test to compare means to decide whether or not they are the “same”.[45] Entering t.test((C2:C42;C43:C64;2;2) in C67, it returns 0.047 which means that at the 95% confidence interval ($p = 0.05$) there is evidence that the water contents in bovine and chicken meat are different.

In the t.test(MatrixA;MatrixB;tails;type) function, in tails, you may use 1 for a one-tailed *t* test and 2 for a two-tailed *t* test. We used the two-tailed *t* test because it concerned testing for a difference between two means in either direction. In type, we used 3 for nonequivalent variances and 2 for equivalent variances.

Appling an *F* test for Ca concentrations, f.test(H2:H42;H43:H64) returns 9.58 × 10^{-10} which means that there is evidence that Ca variances were nonequivalent; then, t.test(C2:C42;C43:C64;2;3) returns 0.024 which means that Ca average concentrations in bovine and chicken samples are significantly different.

The *t* test showed that there was no statistical evidence that 8 out of 15 variables in bovine meat samples were different from chicken meat samples and vice versa. Thus, the PCA model was rebuilt without these variables (Table S4b in Supporting Information), and bovine meat samples were successfully separated from chicken meat samples (Figure 6a).

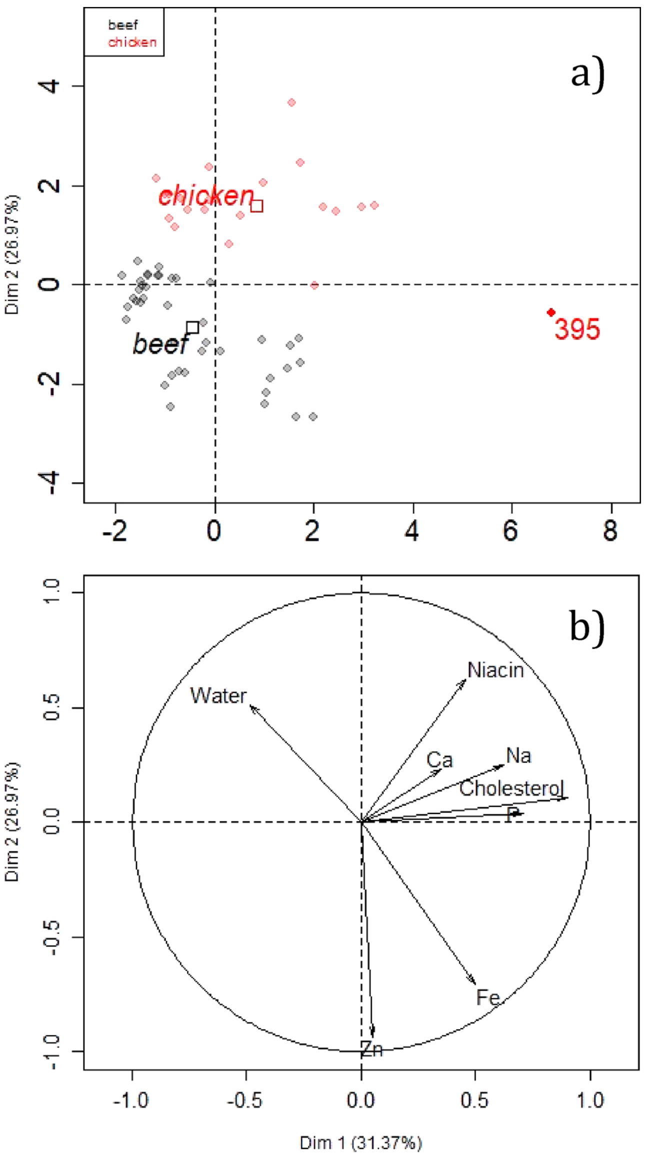


Figure 6. PCA using 8 significant variables: (a) score plot and (b) loading plot.

The loading plot (Figure 6b) shows that meat samples were differentiated by the water, niacin, Zn, and Fe levels. Chicken meat had higher water and niacin levels and lower Zn and Fe levels than bovine meat. Sample 395 (chicken heart) has higher Na, cholesterol, Fe, and P levels and a lower water level than any other sample in the data set.

Differentiation of Healthy Patients from Diabetes Patients

Xuan et al.,[29] found 163 lipids which were representative for differentiating healthy patients (control) from diseased patients (diabetes). Table S5 (Supporting Information) has concentrations of 163 lipids in 30 control and 30 diabetes patients. PCA carried out with this data is shown in Figure 7. In this data set, variables were related. Thus, the reduction of 163 variables (lipids concentrations) to two PCs (PC1 vs PC2) retains 58.26% of the total information (Figure 7). However, there were some false negative and false positive groupings of samples in the score plot (Figure 7).

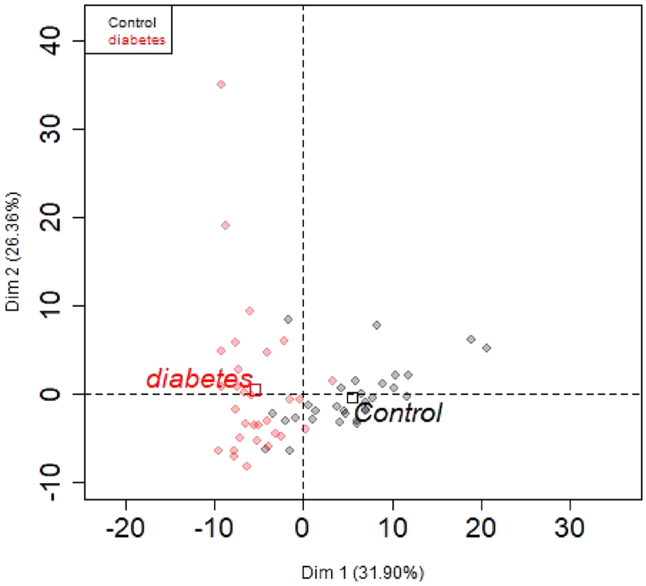


Figure 7. Score plot using concentrations of 163 lipids for 30 healthy patients (control) and 30 diseased patients (diabetes).

The two-tailed t test showed that all variables in Table S5 were not equivalent in both classes for $p < 0.05$. Therefore, we deleted variables with $p < 10^{-5}$ (Table S6 in Supporting Information); we adopted this approach to select only the most discriminating variables. In Table S5a in the Supporting Information, we show how hypothesis tests may be carried out using a spreadsheet. There were just 22 of 123 variables that had $p < 10^{-5}$ in the t test, and the PCA model was rebuilt using just 22 variables (Figure 8).

In Figure 8, patient classes were better separated than in Figure 7. Additionally, 69.68% of the total information was retained with just 2 PCs.

In the PCA carried out using jut 22 variables (Figure 8), we also had false positive, but false negative results were eliminated in comparison with the PCA model carried out using 123 variables (Figure 7).

In this example, students could observe that, in data sets where variables were correlated, most of the information could be explained using two PCs. They could also see that choosing only variables which were extremely different in both classes ($p < 10^{-5}$) increases the differentiation of samples and that hypothesis tests were a good tool to select the most important variables.

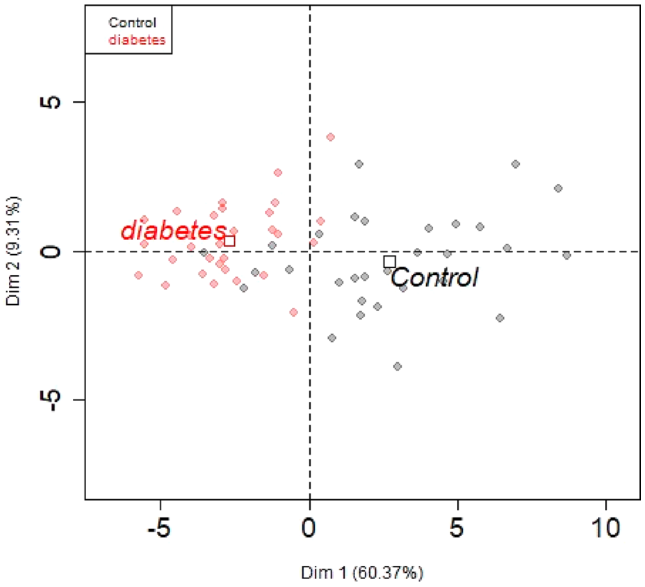


Figure 8. Score plot using concentrations of 22 lipids for 30 healthy patients (control) and 30 diseased patients (diabetes).

Differentiation of Atlantic Salmon and Catfish

In this example, Table S7 was given to students, and they had to differentiate catfish samples from Atlantic salmon samples. This data set has 73 lines (sample) and 68 columns (variables). In this case, variables were correlated, and 65.7% of total information was retained with two PCs (Figure 9).

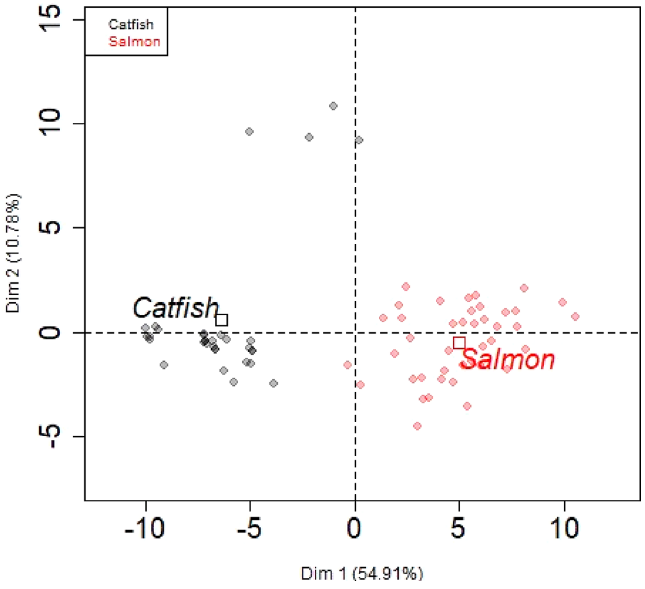


Figure 9. Differentiation of Atlantic salmon from catfish in the score plot.

Atlantic salmon and catfish samples were well-separated in the score plot (Figure 9), where catfish samples were placed on the left-hand side (quadrants 3 and 4) and Atlantic salmon samples were placed on the right-hand side (quadrant 1 and 2).

We did not show the loading plot because 68 variable labels were collapsed in the loading plot, but students observed that

Atlantic salmon samples have higher levels of all variables than catfish samples, except by water content, where catfish samples have a higher water content than Atlantic salmon.

Students observed that variable reduction using PCA is a good choice for correlated data. The analysis also provides differentiation of samples in the score plot (Figure 9), and trends in samples may be found observing original variables in the loading plot.

Chromatographic Fingerprint of *Ginkgo biloba*

In this example, we used the data sets provided by Chen et al.[31] The goal of this section is to show that data treatment can strongly influence PCA.

Ginkgo has been used in traditional Chinese medicine for many centuries to treat ailments.[49] In 2001, the annual worldwide sales of *Ginkgo biloba* were higher than US $21 billion.[50] High performance liquid chromatography, HPLC, is the industry standard for quantification of components in mixtures.[51−55] Chen et al.[31] used HPLC to analyze 14 *Ginkgo biloba* samples (tablets and capsules), an American Herbal Pharmacopoeia (AHP)-verified *G. biloba* leaf sample, and three *G. biloba* standard reference materials (SRM 3246, leaf material; SRM 3247, extracts of *G. biloba* leaves; and SRM 3248, tablet).

Hydrolyzed samples were analyzed by HPLC providing 21 chromatographic peaks. The authors divided all peak areas by rutin's peak area, and then the data was autoscaled. In the PCA score plot (Figure 10a) carried out using this data set (Table S9 in Supporting Information), the samples placed near reference materials (SRM and AHP) were recognized as samples that met the standard requirements.

In Figure 10a, sample 1 is an outlier; samples 5 and 6 were placed close to SRM samples, and sample 7 was placed close to AHP. SRM samples were placed near one another. In this data treatment, rutin was the compound with the highest peak area, and when peak areas were divided by rutin's peak area, the information about it was lost.

When we applied PCA to the original data set (Table S10 in Supporting Information, peak areas were not divided by rutin's peak area), the information about rutin peak area was not lost. Then, we got a score plot (Figure 10b) with a different distribution of objects compared to the score plot shown in Figure 10a.

In Figure 10b (score plot), SRM samples were not as close to each other as in Figure 10a. SHP was not too far from SRM3247 and SRM3248 but was placed in different positions than in Figure 10b. Samples 12, 4, 13, 14, 9, 10, 7, 8, and 2 were closer to SRM 3246 than to SRM 3247. Thus, students could conclude that data treatment strongly influences the distribution of samples in the score plot, and it also influences the interpretation of results.

Identification of Edible Oils by Principal Component Analysis Selecting Featured Peaks of ^{1}H NMR Spectra

Yeh[32] used selected featured peaks of ^{1}H NMR spectra to differentiate edible oils (canola, corn, olive, peanut, sesame seed, and sunflower oil) and sort unknown samples. He did PCA using selected peaks of ^{1}H NMR (Table S11 in the Supporting Information). PCA was carried out using MetaboAnalyst.[56] The data set was treated using Pareto scaling (mean-centered and divided by the square root of the standard deviation of each variable) and normalization (normalization by sum). Then, edible oil samples were well-separated, and unknown samples were placed near corn oil samples.

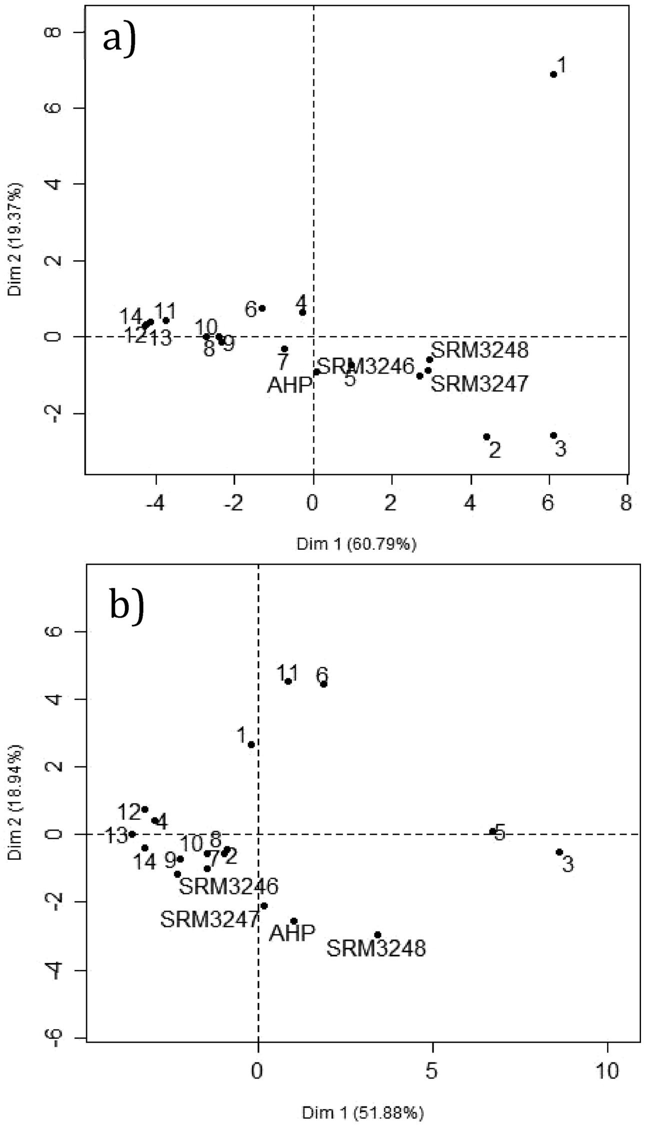


Figure 10. PCA fingerprint of 14 *Ginkgo biloba* samples, AHP, and standard reference materials. (a) Peak areas were divided by rutin's peak area. (b) Original data set.

When the data set was autoscaled only, canola oils presented a very different pattern from other edible oil samples (Figure 11a). After the exclusion of canola oil samples (Table S12 in the Supporting Information), peanut, sunflower, and olive oils were placed near one another and without an appreciable separation (Figure 11b). Unknown samples were placed closer to sesame seed oils than to corn oils (Figure 11b). Therefore, we can conclude that data set treatment is a critical point that may strongly influence objects distribution in the score plot of principal component analysis.

■ CONCLUSION

Students felt comfortable with the graphical interface of RCommander and were able to run PCA in several data sets which were provided in the Supporting Information. They were able to recognize the correlation among variables and their importance to the analysis looking at the loading plots. One of the critical points was associating the importance of the original variables in the score plot looking at the loading plot.

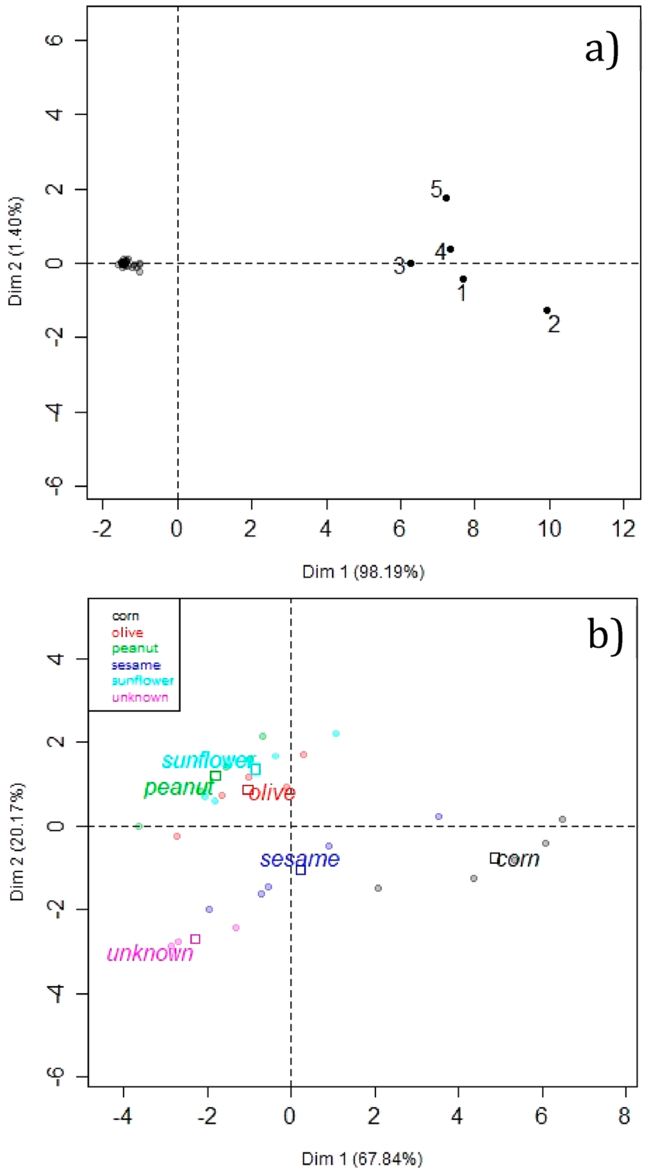


Figure 11. PCA of edible oils (canola, corn, olive, peanut, sesame seed, and sunflower oil): (a) data was autoscaled, canola oil samples 1–5; (b) data was autoscaled, and canola oils samples were excluded.

In the test questions (Supporting Information), there were several examples dealing with classification of chromatographic stationary phases,[57−60] differentiation of organic food from ordinary food,[41,42,48] and evaluation of green tea samples.[61] Students were able to answer most of the questions correctly, and we concluded that the paper was effective for teaching PCA.

Important concepts as recognition of outliers and the effect of outliers in the loading plots and score plots were learned by the students when carrying the analysis. Another critical point was the effect of data treatment in the PCA. Students were able to observe that different treatments of data sets may result in a higher variation in score and loading plots.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available at https://pubs.acs.org/doi/10.1021/acs.jchemed.9b00924.

Tables used in the examples given in the paper and in the test questions (ZIP)

Test questions about the subject (PDF, DOCX)

■ AUTHOR INFORMATION

Corresponding Author

Endler Marcel Borges − *Departamento de Química, Fundação Universidade Regional de Blumenau, FURB, 89012-900 Blumenau, SC, Brazil;* orcid.org/0000-0002-9260-3639; Email: marcelborgesb@gmail.com, embsouza@furb.br

Author

Laís Feltrin Sidou − *Departamento de Química, Fundação Universidade Regional de Blumenau, FURB, 89012-900 Blumenau, SC, Brazil*

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.jchemed.9b00924

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors acknowledge financial support and fellowships from the Brazilian agencies FAPESC (Fundação de Amparo à Pesquisa do Estado de Santa Catarina), CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) Project 399402226/2016-0, and CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior).

■ REFERENCES

(1) Borges, E. M. How to Select Equivalent and Complimentary Reversed Phase Liquid Chromatography Columns from Column Characterization Databases. *Anal. Chim. Acta* **2014**, *807*, 143−152.

(2) Bro, R.; Smilde, A. K. Principal Component Analysis. *Anal. Methods* **2014**, *6* (9), 2812−2831.

(3) Kellogg, J. J.; Paine, M. F.; McCune, J. S.; Oberlies, N. H.; Cech, N. B. Selection and Characterization of Botanical Natural Products for Research Studies: A NaPDI Center Recommended Approach. *Nat. Prod. Rep.* **2019**, *36* (8), 1196−1221.

(4) Anderson, S. L.; Rovnyak, D.; Strein, T. G. Identification of Edible Oils by Principal Component Analysis of 1 H NMR Spectra. *J. Chem. Educ.* **2017**, *94* (9), 1377−1382.

(5) Granato, D.; Santos, J. S.; Escher, G. B.; Ferreira, B. L.; Maggio, R. M. 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. *Trends Food Sci. Technol.* **2018**, *72*, 83−90.

(6) Pérez-Arribas, L. V.; León-González, M. E.; Rosales-Conrado, N. Learning Principal Component Analysis by Using Data from Air Quality Networks. *J. Chem. Educ.* **2017**, *94* (4), 458−464.

(7) Nunes, C. A.; Alvarenga, V. O.; de Souza Sant'Ana, A.; Santos, J. S.; Granato, D. The Use of Statistical Software in Food Science and Technology: Advantages, Limitations and Misuses. *Food Res. Int.* **2015**, *75*, 270−280.

(8) Msimanga, H. Z.; Elkins, P.; Tata, S. K.; Smith, D. R. A Chemometrics Module for an Undergraduate Instrumental Analysis Chemistry Course. *J. Chem. Educ.* **2005**, *82* (3), 415.

(9) Wanke, R.; Stauffer, J. An Advanced Undergraduate Chemistry Laboratory Experiment Exploring NIR Spectroscopy and Chemometrics. *J. Chem. Educ.* **2007**, *84* (7), 1171.

(10) Besalú, E. From Periodic Properties to a Periodic Table Arrangement. *J. Chem. Educ.* **2013**, *90* (8), 1009−1013.

(11) Rusak, D. A.; Brown, L. M.; Martin, S. D. Classification of Vegetable Oils by Principal Component Analysis of FTIR Spectra. *J. Chem. Educ.* **2003**, *80* (5), 541.

(12) Cazar, R. A. An Exercise on Chemometrics for a Quantitative Analysis Course. *J. Chem. Educ.* **2003**, *80* (9), 1026.

(13) Horovitz, O.; Sârbu, C. Characterization and Classification of Lanthanides by Multivariate-Analysis Methods. *J. Chem. Educ.* **2005**, *82* (3), 473.

(14) Ribone, M. É.; Pagani, A. P.; Olivieri, A. C.; Goicoechea, H. C. Determination of the Active Principle in a Syrup by Spectrophotometry and Principal Component Regression Analysis. An Advanced Undergraduate Experiment Involving Chemometrics. *J. Chem. Educ.* **2000**, *77* (10), 1330.

(15) Sandusky, P. O. Introducing Undergraduate Students to Metabolomics Using a NMR-Based Analysis of Coffee Beans. *J. Chem. Educ.* **2017**, *94* (9), 1324−1328.

(16) Stitzel, S. E.; Sours, R. E. High-Performance Liquid Chromatography Analysis of Single-Origin Chocolates for Methylxanthine Composition and Provenance Determination. *J. Chem. Educ.* **2013**, *90* (9), 1227−1230.

(17) De LorenziPezzolo, A. To See the World in a Grain of Sand: Recognizing the Origin of Sand Specimens by Diffuse Reflectance Infrared Fourier Transform Spectroscopy and Multivariate Exploratory Data Analysis. *J. Chem. Educ.* **2011**, *88* (9), 1304−1308.

(18) Gilbert, M. K.; Luttrell, R. D.; Stout, D.; Vogt, F. Introducing Chemometrics to the Analytical Curriculum: Combining Theory and Lab Experience. *J. Chem. Educ.* **2008**, *85* (1), 135.

(19) Skov, T.; Honoré, A. H.; Jensen, H. M.; Næs, T.; Engelsen, S. B. Chemometrics in Foodomics: Handling Data Structures from Multiple Analytical Platforms. *TrAC, Trends Anal. Chem.* **2014**, *60*, 71−79.

(20) Kemsley, E. K.; Defernez, M.; Marini, F. Multivariate Statistics: Considerations and Confidences in Food Authenticity Problems. *Food Control* **2019**, *105*, 102−112.

(21) Esbensen, K. H.; Geladi, P. Principles of Proper Validation: Use and Abuse of Re-Sampling for Validation. *J. Chemom.* **2010**, *24* (3−4), 168−187.

(22) Westad, F.; Marini, F. Validation of Chemometric Models − A Tutorial. *Anal. Chim. Acta* **2015**, *893*, 14−24.

(23) Brereton, R. G.; Jansen, J.; Lopes, J.; Marini, F.; Pomerantsev, A.; Rodionova, O.; Roger, J. M.; Walczak, B.; Tauler, R. Chemometrics in Analytical Chemistry—Part I: History, Experimental Design and Data Analysis Tools. *Anal. Bioanal. Chem.* **2017**, *409* (25), 5891−5899.

(24) Brereton, R. G.; Jansen, J.; Lopes, J.; Marini, F.; Pomerantsev, A.; Rodionova, O.; Roger, J. M.; Walczak, B.; Tauler, R. Chemometrics in Analytical Chemistry—Part II: Modeling, Validation, and Applications. *Anal. Bioanal. Chem.* **2018**, *410* (26), 6691−6704.

(25) Li, C.-Q.; Xiao, N.; Wen, Y.; He, S.-H.; Xu, Y.-D.; Lin, Y.-W.; Li, H.-D.; Xu, Q.-S. Collaboration Patterns and Network in Chemometrics. *Chemom. Intell. Lab. Syst.* **2019**, *191*, 21−29.

(26) PubChem. Periodic Table of Elements. https://pubchem.ncbi.nlm.nih.gov/periodic-table/#view=list (accessed Sep 11, 2019).

(27) TACO. http://www.nepa.unicamp.br/taco/tabela.php?ativo=tabela (accessed Sep 25, 2019).

(28) Coelho, K. S.; BistricheGiuntini, E.; Grande, F.; da Silva Dias, J.; Purgatto, E.; de Melo Franco, B. D. G.; Lajolo, F. M.; de Menezes, E. W. 12th IFDC 2017 Special Issue − Brazilian Food Composition Table (TBCA): Development and Functionalities of the Online Version. *J. Food Compos. Anal.* **2019**, *84*, 103287.

(29) Xuan, Q.; Hu, C.; Yu, D.; Wang, L.; Zhou, Y.; Zhao, X.; Li, Q.; Hou, X.; Xu, G. Development of a High Coverage PseudotargetedLipidomics Method Based on Ultra-High Performance Liquid Chromatography−Mass Spectrometry. *Anal. Chem.* **2018**, *90* (12), 7608−7616.

(30) INFOODS. FAO/INFOODS Databases. http://www.fao.org/infoods/infoods/tables-and-databases/faoinfoods-databases/en (accessed Sep 28, 2019).

(31) Chen, P.; Ozcan, M.; Harnly, J. Chromatographic Fingerprint Analysis for Evaluation of Ginkgo Biloba Products. *Anal. Bioanal. Chem.* **2007**, *389* (1), 251−261.

(32) Yeh, T.-S. Comment on "Identification of Edible Oils by Principal Component Analysis of 1 H NMR Spectra. *J. Chem. Educ.* **2019**, *96* (8), 1790−1792.

(33) Teaching Principal Component Analysis Using a Free and Open Source Software. YouTube. https://youtu.be/zNOZO2Bcsug (accessed Apr 17, 2020).

(34) The Comprehensive R Archive Network (CRAN). https://cran.r-project.org (accessed 2020-03-16).

(35) Sidou, L. F.; Borges, E. M. How to install and use R to perform chemometrics analysis. https://drive.google.com/file/d/1xyLR4K1_8raDKaaiBcIDSRpWk-hvK2be/view (accessed May 2020).

(36) Strong, F. C. The Atomic Form Periodic Table. *J. Chem. Educ.* **1959**, *36* (7), 344.

(37) Bierenstiel, M.; Snow, K. Periodic Universe: A Teaching Model for Understanding the Periodic Table of the Elements. *J. Chem. Educ.* **2019**, *96*, 1367.

(38) Hoffman, A.; Hennessy, M. The People Periodic Table: A Framework for Engaging Introductory Chemistry Students. *J. Chem. Educ.* **2018**, *95* (2), 281−285.

(39) Winter, M. J. Diffusion Cartograms for the Display of Periodic Table Data. *J. Chem. Educ.* **2011**, *88* (11), 1507−1510.

(40) Diener, L. News from Online: The Periodic Table of the Elements. *J. Chem. Educ.* **2009**, *86* (10), 1163.

(41) Borges, E. M.; Gelinski, J. M. L. N.; de Oliveira Souza, V. C.; Barbosa, F.; Batista, B. L. Monitoring the Authenticity of Organic Rice via Chemometric Analysis of Elemental Data. *Food Res. Int.* **2015**, *77*, 299.

(42) Borges, E. M.; Volmer, D. A.; Gallimberti, M.; Ferreira De Souza, D.; Luiz De Souza, E.; Barbosa, F. Evaluation of Macro- and Microelement Levels for Verifying the Authenticity of Organic Eggs by Using Chemometric Techniques. *Anal. Methods* **2015**, *7* (6), 2577.

(43) Cozzolino, D.; Power, A.; Chapman, J. Interpreting and Reporting Principal Component Analysis in Food Science Analysis and Beyond. *Food Anal. Methods* **2019**, *12*, 2469.

(44) Royal Society of Chemistry. Periodic Table. https://www.rsc.org/periodic-table (accessed Apr 17, 2020).

(45) da Silva, R. S.; Borges, E. M. Quantitative Analysis Using a Flatbed Scanner: Aspirin Quantification in Pharmaceutical Tablets. *J. Chem. Educ.* **2019**, *96* (7), 1519−1526.

(46) Volmer, D. A.; Curbani, L.; Parker, T. A.; Garcia, J.; Schultz, L. D.; Borges, E. M. Determination of Titratable Acidity in Wine Using Potentiometric, Conductometric, and Photometric Methods. *J. Chem. Educ.* **2017**, *94* (9), 1296−1302.

(47) Ledesma, C. M.; Krepsky, L. M.; Borges, E. M. Using a Flatbed Scanner and Automated Digital Image Analysis To Determine the Total Phenolic Content in Beer. *J. Chem. Educ.* **2019**, *96* (10), 2315−2321.

(48) Borges, E. M.; Volmer, D. A.; Brandelero, E.; Gelinski, J. M. L. N.; Gallimberti, M.; Barbosa, F. Monitoring the Authenticity of Organic Grape Juice via Chemometric Analysis of Elemental. *Data. Food Anal. Methods* **2016**, *9* (2), 362−369.

(49) Borges, E. M.; Volmer, D. A.; Eberlin, M. N. Comprehensive Analysis of Ginkgo Tablets by Easy Ambient Sonic Spray Ionization Mass Spectrometry. *Can. J. Chem.* **2013**, *91* (8), 671−678.

(50) Rimmer, C. A.; Howerton, S. B.; Sharpless, K. E.; Sander, L. C.; Long, S. E.; Murphy, K. E.; Porter, B. J.; Putzbach, K.; Rearick, M. S.; Wise, S. A.; et al. Characterization of a Suite of Ginkgo-Containing Standard Reference Materials. *Anal. Bioanal. Chem.* **2007**, *389* (1), 179−196.

(51) Stankus, B.; White, R.; Abrams, B. Effective and Inexpensive HPLC Analogue for First-Year Students: Buret Chromatography of Food Dyes in Drinks. *J. Chem. Educ.* **2019**, *96* (4), 739−744.

(52) HPLC. A Practical Technique for Future Chemists. *J. Chem. Educ.* **1992**, *69* (4), 260.

(53) Beussman, D. J.; Walters, J. P. Complete LabVIEW-Controlled HPLC Lab: An Advanced Undergraduate Experience. *J. Chem. Educ.* **2017**, *94* (10), 1527−1532.

(54) Borges, E. M.; Volmer, D. A. Silica, Hybrid Silica, Hydride Silica and Non-Silica Stationary Phases for Liquid Chromatography. Part II: Chemical and Thermal Stability. *J. Chromatogr. Sci.* **2015**, *53* (7), 1107.

(55) Borges, E. M. Silica, Hybrid Silica, Hydride Silica and Non-Silica Stationary Phases for Liquid Chromatography. *J. Chromatogr. Sci.* **2015**, *53* (4), 580.

(56) Chong, J.; Wishart, D. S.; Xia, J. Using MetaboAnalyst 4.0 for Comprehensive and Integrative Metabolomics Data Analysis. *Curr. Protoc. Bioinforma.* **2019**, *68* (1). DOI: 10.1002/cpbi.86

(57) Borges, E. M.; Silva, C. G. A.; Collins, C. H. Chromatographic Evaluation of Some Stationary Phases Based on Poly-(Methyloctylsiloxane) Immobilized onto Silica. *Microchem. J.* **2010**, *96* (1), 120.

(58) Žuvela, P.; Skoczylas, M.; Jay Liu, J.; Baczek, T.; Kaliszan, R.; Wong, M. W.; Buszewski, B. Column Characterization and Selection Systems in Reversed-Phase High-Performance Liquid Chromatography. *Chem. Rev.* **2019**, *119*, 3674−3729.

(59) Euerby, M. R.; Petersson, P. Chromatographic Classification and Comparison of Commercially Available Reversed-Phase Liquid Chromatographic Columns Containing Polar Embedded Groups/Amino Endcappings Using Principal Component Analysis. *In Journal of Chromatography A* **2005**, *1088*, 1−15.

(60) Euerby, M. R.; Petersson, P. Chromatographic Classification and Comparison of Commercially Available Reversed-Phase Liquid Chromatographic Columns Using Principal Component Analysis. *J. Chromatogr. A* **2003**, *994* (1−2), 13−36.

(61) Kellogg, J. J.; Graf, T. N.; Paine, M. F.; McCune, J. S.; Kvalheim, O. M.; Oberlies, N. H.; Cech, N. B. Comparison of Metabolomics Approaches for Evaluating the Variability of Complex Botanical Preparations: Green Tea (Camellia Sinensis) as a Case Study. *J. Nat. Prod.* **2017**, *80* (5), 1457−1466.