Chapter 17

Chemometric Methods in Food Authentication

Riccardo Leardi
Department of Pharmacy, University of Genova, Genoa, Italy

Chapter Outline

1 INTRODUCTION

In this chapter the fundamentals of chemometrics will be presented by means of a quick overview of the most relevant techniques for data display, classification, modeling, multivariate process monitoring, multivariate quality control, and calibration. This chapter is intended to make people aware of the great superiority of multivariate analysis over the commonly used univariate approach. Mathematical and algorithmical details will not be presented, since the chapter is mainly focused on the general problems to which chemometrics can be successfully applied in the field of food authentication.

As a matter of fact, many of the readers of this book may not be familiar with chemometrics, and a significant percentage of them may have never even heard of this "new" science (quite strange that it is still considered a "new" science, when the Chemometrics Society was founded in 1974 and the most basic algorithms date back to the beginning of the 20th century). Furthermore, some of them could be quite put off by anything involving mathematical computations higher than a square root or statistical tests more complex than a *t*-test.

Modern Techniques for Food Authentication. https://doi.org/10.1016/B978-0-12-814264-6.00017-7

Therefore, the goal of this chapter is simply that of being read and understood by the majority of the readers of this book. This goal will be completely achieved if some of them, after having read it, could say: "chemometrics is easy and powerful indeed, and from now on I will always think in a multivariate way."

Of course, to accomplish this goal in the limited space of a chapter the attractive sides of chemometrics must be highlighted. Therefore, the intuitive aspects of each technique will be shown, without giving too much relevance to the algorithms.

First of all, what is chemometrics? According to the definition of the Chemometrics Society, it is "the chemical discipline that uses mathematical and statistical methods to design or select optimal procedures and experiments, and to provide maximum chemical information by analyzing chemical data."

One of the major mistakes that people make about chemometrics is thinking that to use it one has to be a very good mathematician and to know the mathematical details of the algorithms being used. From the definition itself, it is clear instead that a chemometrician is a *chemist* who can *use* mathematical and statistical methods.

If we want to draw a parallel with everyday life, how many of us really know in detail how a TV set, a mobile phone, a car, or a washing machine works? But everybody watches TV programs, makes phone calls, drives a car, and starts a washing machine. Of course, what is important is that people know what each instrument is made for and that nobody tries to drive a TV set, or to speak inside a washing machine or to do the laundry in a car.

Though chemometrics makes available a very wide range of techniques, some of them being very difficult to fully understand and use correctly, the great majority of the real problems can be solved by applying one of the basic techniques, whose understanding, at least from an intuitive point of view, is relatively easy and does not require high-level mathematical skills.

Another big misunderstanding arises when talking about the role of chemometric software. More and more people have access to some (commercial or free) software, and it is widespread opinion that the software is the heart of the chemometric process and that it is enough to be able to run the software to be a chemometrician. The software is just a tool, which should be properly used in order to follow the strategy envisioned by the chemometrician. More and more software "guide" in an automated way (not to say "oblige") the user toward a standard procedure (e.g., automatic outlier removal or automatic detection of the number of significant components). By doing this, the nontrained user ends up doing what the software wants him to do, instead of having the software doing what he wants it to do.

So, people must be well aware that the software is not at all the most important point in a chemometric analysis. Quite provocatively, I like saying that, if properly trained, a chimp would be perfectly able to run a principal component analysis (PCA) (just teach him the correct sequence of buttons to press, and then give him some peanuts as reward when he does it correctly). Instead, a chimp will never be able to interpret the result of a PCA. This step can be properly

performed only by somebody who knows the chemical problem lying behind the data (Brereton, 2013). Note all the elaborations and the plots reported in this chapter (except Figs. 5 and 6) have been performed by an R-based software developed by the Italian Group of Chemometrics, freely downloadable at http://gruppochemiometria.it/index.php/software.

2 DATA COLLECTION

Chemometrics works on data matrices. This means that on each sample a certain number of variables have been measured (in the "chemometrical jargon" we say that each object is described by v variables). Although some techniques can work with a limited number of missing values, a chemometrical data set must be thought of as a spreadsheet in which all the cells are full.

Sometimes, instead, if data are gathered without having any specific project, it happens that the result is a "sparse" matrix containing some blank cells. In such cases, if the percentage of missing data is quite high, the whole data set is not suitable for a multivariate analysis; as a consequence, the variables and/or the objects with the lowest number of data must be removed, and therefore a huge amount of experimental effort can be lost.

It should be obvious that, since "missing value" means that the value has not been measured, the missing values should never be replaced by 0 (0 meaning that the variable has been measured and its value is 0). Instead, the missing value must be coded according to the specific software (e.g., blank cell in Excel, NaN in Matlab, NA in R).

Any chemometrical software allows the import of data from ASCII files or from spreadsheets. It is therefore suggested to organize the data in matrix form from the start, as shown in Table 1, in such a way that the import can be performed in a single step.

If, on the contrary, the data are spread in several files or sheets (e.g., one file for each sample or for each variable), then the import procedure would be much longer and more cumbersome.

Sometimes it also happens that people tend to "overcrowd" the data file with intermediate and therefore "useless" numbers (e.g., in the case of a weighing, the gross weight and the tare). These values, being nonrelevant for the data analysis, should anyway be removed before the chemometrical elaboration. It can be easily understood that a good data collection, resulting in a well-structured data sheet, is the first step toward a successful data analysis.

Together with the numerical variables, that will undergo the statistical analysis, each sample should also have "ancillary variables" or "metavariables," such as date of the analysis, date of production, operator, instrument, geographical origin, etc. Having collected the right metavariables could be the key for a successful interpretation of the results of the elaboration.

It is also very important that the data follow the chronological order of production (or analysis), or that is anyway possible to reconstruct it. In many cases

TABLE 1 The Structure of a Chemometrical Data Set

Sample name	Operator	Date	Origin	var. 1	var. 2	var. 3	var. 4	…	var. *v*
obj1									
obj2									
obj3									
obj4									
obj5									
obj6									
……									
obj*n*									

a chemometrical analysis allows to highlight time trends, which of course can be detected only if the data are stored according to the time sequence.

3 DATA DISPLAY

The human mind can digest much more information when looking at plots rather than numbers. This is easily demonstrated by looking first at the sequence of numbers reported in Table 2, and then at the plots in Fig. 1.

It is very clear that, even in a very simple data set like this one (just 10 samples and only 1 variable) the information obtained by looking at the plot is superior and much more easily available than the information that one can get by analyzing the raw numbers. From the plot, it becomes evident that the samples are clustered into two groups of the same size, the one at higher values being much tighter than the one at low values. Much more time and effort are required when we want to get the same information from the table.

Let us now take into account a more complex data set, that is, the one reported in Table 3, where each object is described by two variables. The same data are plotted in Fig. 2.

In this data set we have 20 samples, supposed to belong to the same population. When looking at univariate plots (Figs. 2A and B) it really seems that the samples constitute one single group. When looking at the bivariate plot shown in Fig. 2C we realize instead that we are in a situation very similar to what we found with the univariate data set. The samples are split into two clusters of the same size, with the objects of the first one more tightly grouped than the objects of the second one. As previously shown, this conclusion cannot be reached when looking at one variable at a time, since neither of the two variables is able to discriminate between the two groups.

This bivariate data set, beyond showing once more that a plot is much more easily handled by the human brain than a data table, demonstrates that when dealing with more than one variable the analysis of just one variable at a time can lead to wrong results.

If we had a data set with three variables it would still be possible to visualize the whole information by a three-dimensional scatter plot, in which the coordinates of each object are the values of the variables. However what should we do if there are more than three variables? What we need therefore is a technique permitting the visualization by simple bi- or tridimensional scatter plots of the majority of the information contained in a highly dimensional data set. This technique is PCA, one of the simplest and most used methods of multivariate analysis. PCA is very important especially in the preliminary steps of an elaboration, when one wants to perform an exploratory analysis in order to have an overview of the data.

It is quite common to have to deal with large data tables with, for instance, a series of samples described by a number (v) of chemicophysical parameters. Examples of such data sets can be samples of olive oils from different origins

TABLE 2 Ten Samples Described by One Variable

Sample	1	2	3	4	5	6	7	8	9	10
Value	25.3	22.1	25.5	25.6	19.4	25.7	20.2	21.3	25.9	21.8

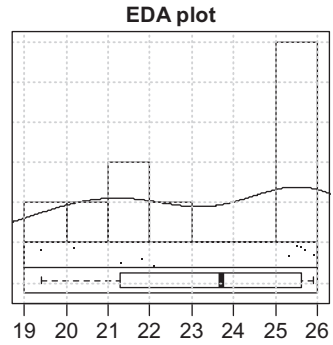

FIG. 1 Univariate plots of the data in Table 2 Exploratory data analysis; from bottom to top: box and whiskers plot, scatter plot, histogram, probability density function.

TABLE 3 Twenty Samples Described by Two Variables

Sample	Variable 1	Variable 2
1	21.2	32.5
2	16.2	21.0
3	13.1	21.7
4	11.6	21.3
5	20.8	29.9
6	10.4	20.6
7	19.5	26.8
8	9.8	25.2
9	15.2	31.2
10	12.0	26.0
11	17.6	28.5
12	24.0	30.0
13	17.8	33.1
14	15.0	24.0
15	11.0	24.2
16	24.8	25.3
17	12.8	23.3
18	26.5	30.6
19	22.9	27.5
20	9.7	22.8

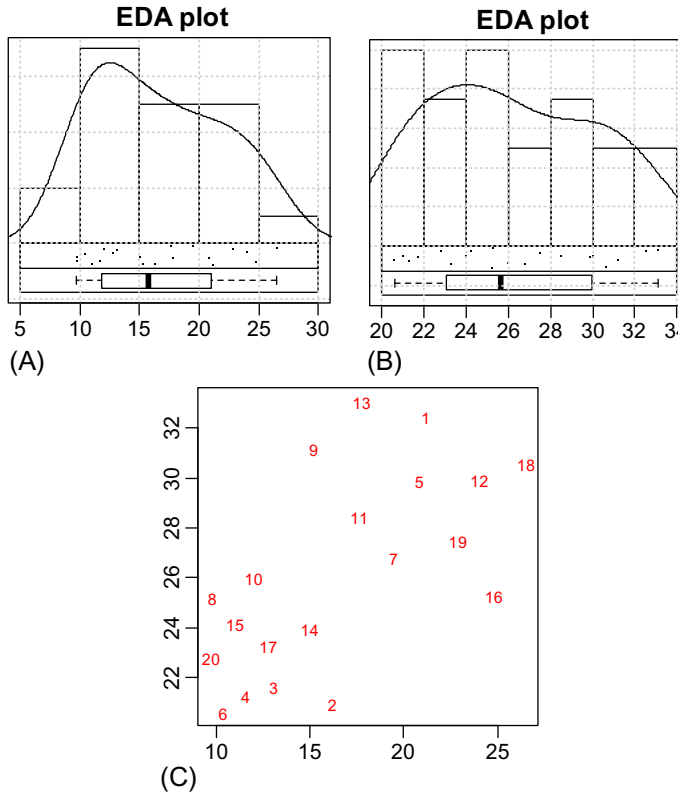

FIG. 2 Plots of the data in Table 3: (A) univariate plots of variable 1; (B) univariate plots of variable 2; and (C) bivariate scatter plot.

described by their content in fatty acids and sterols, or samples of wines described by Fourier-transformed infra-red (FT-IR) spectra. It is easy to realize how, especially in spectral data sets, v can be very high (>1000). In such cases, it would be impossible to obtain valuable information without the help of multivariate techniques.

From a geometrical point of view, we can consider a v-dimensional space, in which each dimension is associated to one of the variables. In this space each sample (object) has coordinates corresponding to the values of the variables describing it.

Since it is impossible to visualize all the information at once, one should be content with the analysis of several bi- or tridimensional plots, each of them showing a different part of the global information.

It is also evident that not all possible combinations of two or three variables will give the same quality of information. For instance, if some variables are very highly correlated, then the information brought by each of them would be almost the same. If two variables are perfectly correlated, then one of them can be discarded, losing no information at all. In this way, the dimensionality of our space

will be reduced from v to $v-1$. If two variables are very highly correlated, then the elimination of one of them would produce only a slight loss of information, while the dimensionality of the space would be reduced to $v-1$. So, one can deduce that the information contained in the "lost" vth dimension was well below the average of the information contained in the other dimensions.

It is quite apparent now that not all the dimensions have the same importance, and that, owing to the correlations among the variables, the "real" dimensionality of our data matrix is somehow lower than v. Therefore, it would be very valuable to have a technique capable of concentrating in a few variables, and therefore in a few dimensions, the bulk of our information. This is exactly what is performed by PCA: it reduces the dimensionality of the data and extracts the most relevant part of the information, placing into the last dimensions the unstructured information, that is, the noise. According to these two characteristics, the information contained in very complex data matrices can be visualized in just one or a few plots.

From the mathematical point of view, the goal of PCA is to obtain, from v variables $(X_1, X_2, \ldots, X_v)$, v linear combinations having two important features: to be uncorrelated and to be ordered according to the explained variance (i.e., to the information they contain). The lack of correlation among the linear combinations is very important, since it means that each of them describes different "aspects" of the original data. As a consequence, the examination of a limited number of linear combinations (generally the first two or three) allows us to obtain a good representation of the studied data set.

From a geometrical point of view, what is performed by PCA corresponds to look for the direction which, in the v-dimensional space of the original variables, brings the greatest possible amount of information (i.e., explains the greatest variance). Once the first direction is identified, the second one is looked for: it will be the direction explaining the greatest part of the residual variance, under the constraint of being orthogonal to the first one. This process goes on until the vth direction has been found.

These new directions can be considered as the axes of a new orthogonal system, obtained after a simple rotation of the original axes. While in the original system each direction (i.e., each variable) brings with it, at least in theory, $1/v$ of total information, in the new system the information is concentrated in the first directions, and decreases progressively so that in the last ones no information can be found except noise.

The global dimensionality of the system is always that of the original data (v), but since the last dimensions explain only a very small part of the information, they can be neglected and one can take into account only the first dimensions (the "significant components"). The projection of the objects in this space of reduced dimensionality retains almost all the information that can now also be analyzed in a visual way, by bi- or tridimensional plots. These new directions, linear combinations of the original ones, are the principal components (PCs) or eigenvectors.

With a mathematical notation, we can write:

$$\mathrm{var}(Z_1) > \mathrm{var}(Z_2) > \ldots > \mathrm{var}(Z_v) \tag{1}$$

where var.(Z_i) is the variance explained by component i. Furthermore, since a simple rotation has been performed, the total variance is the same in the two systems of axes:

$$\Sigma\, \mathrm{var}(X_i) = \Sigma\, \mathrm{var}(Z_i) \tag{2}$$

The first PC is formed by the linear combination

$$Z_1 = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1v}X_v \tag{3}$$

explaining the greatest variance, under the condition

$$\sum a_{1i}^2 = 1 \tag{4}$$

This last condition notwithstanding, the variance of Z_1 could be made greater simply by increasing one of the values of a.

The second PC

$$Z_2 = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2v}X_v \tag{5}$$

is the one having var.(Z_2) as large as possible, under the conditions that

$$\sum a_{2i}^2 = 1 \tag{6}$$

and that

$$\sum a_{1i}a_{2i} = 0 \tag{7}$$

Eq. (7) assures the orthogonality of components 1 and 2.

The lower order components are computed in the same way, always under the two conditions previously reported.

From a mathematical point of view, PCA is solved by finding the eigenvalues of the variance-covariance matrix; they correspond to the variance explained by the corresponding PC. Since the sum of the eigenvalues is equal to the sum of the diagonal elements (trace) of the variance–covariance matrix, and since the trace of the variance-covariance matrix corresponds to the total variance, one has confirmed that the variance explained by the PCs is the same as explained by the original data.

It is now interesting to locate each object in this new reference space. The coordinate on the first PC is computed simply by substituting into Eq. (3), with X_i being the values of the corresponding original variables. The coordinates on the other PCs are then computed in the same way. These coordinates are named scores, while the constants a_{ij} are named loadings.

By taking into account the loadings of the variables on the different PCs, it is very easy to understand the importance of each single variable in constituting each PC. A high absolute value means that the variable under examination plays

an important role for the component, while a low absolute value means that it has a very limited importance.

If a loading has a positive sign, it means that the objects with a high value of the corresponding variable have high positive scores on that component. If the sign is negative, then the objects with high values of that variable will have high negative scores. As already mentioned, after a PCA the information is mainly concentrated on the first components. As a consequence, a plot of the scores of the objects on the first components allows the direct visualization of the global information in a very efficient way. It is now very easy to detect similarity between objects (similar objects have a very similar position in the space) or the presence of outliers (they are very far from all other objects) or the existence of clusters. Taking into account at the same time scores and loadings it is also possible to interpret very easily the differences among objects or groups of objects, since it is immediately understandable which are the variables giving the greatest contribution to the phenomenon under study.

Mathematically speaking, we can say that the original data matrix $\mathbf{X}_{o,v}$ (having as many rows as objects and as many columns as variables) has been decomposed into a matrix of scores $\mathbf{S}_{o,c}$ (having as many rows as objects and as many columns as retained components, with c usually $\ll v$) and a matrix of loadings $\mathbf{L}_{c,v}$ (having as many rows as retained components and as many columns as variables). If, as usual, $c < v$, a matrix of the residuals $\mathbf{E}_{o,v}$, having the same size as the original data set, contains the differences between the original data and the data reconstructed by the PCA model (the smaller the values of this matrix, the higher the variance explained by the model).

We can therefore write the following relationship:

$$\mathbf{X}_{o,v} = \mathbf{S}_{o,c} {}^{*} \mathbf{L}_{c,v} + \mathbf{E}_{o,v}$$

Now, let us see the application of PCA to a real data set (MacNamara, 2005). Twelve variables have been measured by gas chromatography on 43 samples of Irish whiskeys, of two different types. Nineteen samples were from type A, while 24 samples were from type B, with the samples of each type ordered according to the production time. The data are reported in Table 4.

Since a trained assessor can easily differentiate a whiskey of type A from a whiskey of type B, it is interesting to know whether they are different also chemically, just taking into account the variables obtained by a routine analysis. When looking separately at each of the 12 variables, it can be seen that none of them completely separates the two types. Therefore, when thinking on a univariate basis, one could say that it is not possible to state that the two types of whiskey are different. As a consequence, one could look for different (and possible more expensive to determine) variables.

After a PCA (Fig. 3), it is instead evident that the information present in the 12 variables clearly shows that the two whiskeys are different. Once more, it has to be pointed out that taking into account all the variables at the same time gives much more information than just looking at one variable at a time.

TABLE 4 Chemical Composition of 43 Whiskey Samples

Sample	Type	(1) Acetal-dehyde	(2) Ethyl Acetate	(3) Acetal	(4) Propanol	(5) Isobutanol	(6) Isoamyl Acetal	(7) Butanol-1	(8) 2-Me-1-Butanol	(9) 3-Me-1-Butanol	(10) Et Caproate	(11) Ethyl Caprylate	(12) Et Caprate
1	A	80	408	37	583	466	24	15	388	988	3	13	45
2	A	76	327	40	507	483	25	18	396	1033	3	12	46
3	A	79	296	43	467	397	20	17	323	859	4	13	44
4	A	74	415	28	569	407	24	15	352	921	4	13	46
5	A	69	381	29	510	367	21	14	329	870	4	13	46
6	A	66	340	35	428	387	26	13	339	910	4	14	50
7	A	82	373	17	401	337	23	11	297	813	4	13	42
8	A	78	385	34	459	371	19	12	313	843	3	12	41
9	A	67	374	34	458	385	22	12	326	868	3	13	47
10	A	50	331	32	422	345	17	12	307	835	3	12	42
11	A	66	342	30	423	341	17	13	305	846	3	13	43
12	A	54	321	28	408	354	20	13	310	874	4	13	41
13	A	68	344	33	429	333	16	12	300	824	3	11	38
14	A	69	358	37	446	347	17	13	311	855	3	11	37
15	A	78	346	40	411	320	16	12	287	796	3	11	36
16	A	77	387	51	427	345	22	12	290	805	3	10	32
17	A	104	322	72	432	353	18	13	303	823	3	10	35

18	A	84	333	55	421	340	17	13	292	787	3	10	31
19	A	82	382	47	457	328	18	10	278	765	3	10	31
20	B	65	403	18	496	529	19	19	365	1014	3	11	35
21	B	58	352	18	434	457	17	17	312	907	3	8	26
22	B	71	394	25	555	560	18	20	391	1083	3	11	33
23	B	69	369	25	497	500	16	18	349	1005	3	10	29
24	B	83	344	28	489	479	15	17	352	957	3	10	29
25	B	93	344	31	500	481	15	18	352	990	3	10	29
26	B	65	453	18	503	529	21	17	390	1017	3	10	31
27	B	62	405	17	500	488	18	17	357	965	3	9	27
28	B	58	435	16	501	548	21	17	415	1056	3	10	31
29	B	63	459	17	544	575	21	19	426	1100	3	10	28
30	B	99	462	26	490	500	22	16	403	1057	3	10	30
31	B	81	357	21	402	396	16	14	310	814	2	7	17
32	B	80	380	23	497	483	18	17	395	1041	3	10	28
33	B	76	425	22	486	475	22	17	379	1007	4	10	25
34	B	79	446	24	446	418	18	14	319	803	3	9	25
35	B	78	461	24	478	458	19	16	352	908	3	9	23
36	B	108	477	29	493	430	16	14	329	811	3	11	28
37	B	111	481	28	494	429	16	15	330	833	3	9	22
38	B	82	408	22	473	431	18	12	317	774	3	10	27

Continued

TABLE 4 Chemical Composition of 43 Whiskey Samples—cont'd

Sample	Type	(1) Acetal-dehyde	(2) Ethyl Acetate	(3) Acetal	(4) Propanol	(5) Isobutanol	(6) Isoamyl Acetal	(7) Butanol-1	(8) 2-Me-1-Butanol	(9) 3-Me-1-Butanol	(10) Et Caproate	(11) Ethyl Caprylate	(12) Et Caprate
39	B	73	428	20	493	445	18	13	327	804	3	8	20
40	B	102	469	25	490	457	20	11	327	776	3	10	27
41	B	90	463	22	491	452	20	12	324	774	3	9	21
42	B	50	410	14	440	419	19	12	300	704	3	11	28
43	B	61	425	17	445	432	20	12	318	758	3	10	23

Now, let us go one step back and try to understand how this result has been obtained. First, since the variables have different magnitudes and variances, a normalization has to be performed, in such a way that each variable will have the same importance. Autoscaling is the most frequently used normalization, which is done by subtracting from each variable its mean value and then dividing the result by its standard deviation. After that, each normalized variable will have mean = 0 and variance = 1. Table 5 shows the data after autoscaling. The results of the PCA are such that PC1 explains 38.4% of the total variance and PC2 26.4%. This means that the PC1–PC2 plots shown in Fig. 3 explain 64.8% of total variance. Table 6 shows the loadings of the variables on PC1 and PC2. From it, the loading plot in Fig. 3 is obtained.

From the score plot in Fig. 3 it can be seen that the two categories are perfectly separated on the plane PC1-PC2. By looking at the loading plot and at Table 6 it is possible to know which are the variables mainly contributing to each of the PCs. Variables 4, 5, 7, 8, and 9 (propanol, isobutanol, butanol-1, 2-Me-1-butanol, and 3-Me-1-butanol, i.e., the alcohols) have the loadings with the highest absolute value on PC1, all of them being negative. This means that the alcohols are higher in those samples having the highest negative scores on PC1. Variables 6, 10, 11, and 12 (isoamyl acetal, ethyl caproate, ethyl caprylate and ethyl caprate, i.e., the esters) have the loadings with the highest absolute value on PC2, all of them being negative. This means that the esters are higher in those samples having the highest negative scores on PC2. Therefore, it can be said that the esters are the main responsible for the difference between the two types, while the alcohols are the main responsible for the variability inside each type. The fact that all the alcohols have very similar loadings means that they are very much correlated, as is the case for the esters. This is a further demonstration of the superiority of multivariate analysis on univariate analysis. Indeed, it will be possible to adulterate a product in such a way that all the variables, singularly taken, fall inside their individual range of acceptance; much more difficult (not to say impossible) will be to have an adulterated product in which also the correlations among the variables will be preserved. Therefore, adulterated products that will be unnoticed by the "classical" univariate analysis will be easily detected by a multivariate analysis (see Section 7).

Table 7 reports the scores of the objects on PC1 and PC2.

As previously shown, the scores of an object are computed by multiplying the loadings of each variable by the value of the variable. As an example, let us compute the score of sample 1 on PC1 (since the autoscaled data have been used, these are the values that must be taken into account):

$$0.288^{*}0.006 + 0.340^{*}(-0.253) + 0.672^{*}0.261 + 2.507^{*}(-0.363)$$
$$+ 0.553^{*}(-0.452) + 1.743^{*}(-0.067) + 0.187^{*}(-0.385) + 1.348^{*}(-0.429)$$
$$+ 0.936^{*}(-0.378) + (-0.338)^{*}0.071 + 1.441^{*}0.146 + 1.418^{*}0.159 = -1.778$$

TABLE 5 Autoscaled Data

Sample	Type	(1) Acetal-dehyde	(2) Ethyl Acetate	(3) Acetal	(4) Propanol	(5) Isobutanol	(6) Isoamyl Acetal	(7) Butanol-1	(8) 2-Me-1-Butanol	(9) 3-Me-1-Butanol	(10) Et Caproate	(11) Ethyl Caprylate	(12) Et Caprate
1	A	0.288	0.340	0.672	2.507	0.553	1.743	0.187	1.348	0.936	−0.338	1.441	1.418
2	A	0.013	−1.285	0.927	0.791	0.796	2.105	1.335	1.559	1.366	−0.338	0.821	1.536
3	A	0.219	−1.907	1.183	−0.112	−0.435	0.295	0.952	−0.365	−0.297	2.084	1.441	1.301
4	A	−0.125	0.481	−0.095	2.190	−0.291	1.743	0.187	0.399	0.296	2.084	1.441	1.536
5	A	−0.469	−0.202	−0.010	0.858	−0.864	0.657	−0.196	−0.207	−0.192	2.084	1.441	1.536
6	A	−0.675	−1.024	0.501	−0.993	−0.578	2.467	−0.579	0.056	0.190	2.084	2.060	2.005
7	A	0.426	−0.362	−1.032	−1.602	−1.293	1.381	−1.344	−1.051	−0.737	2.084	1.441	1.066
8	A	0.150	−0.121	0.416	−0.293	−0.806	−0.067	−0.961	−0.629	−0.450	−0.338	0.821	0.949
9	A	−0.606	−0.342	0.416	−0.316	−0.606	1.019	−0.961	−0.286	−0.211	−0.338	1.441	1.653
10	A	−1.776	−1.205	0.246	−1.128	−1.178	−0.791	−0.961	−0.787	−0.526	−0.338	0.821	1.066
11	A	−0.675	−0.984	0.075	−1.106	−1.236	−0.791	−0.579	−0.840	−0.421	−0.338	1.441	1.184
12	A	−1.501	−1.406	−0.095	−1.444	−1.050	0.295	−0.579	−0.708	−0.154	2.084	1.441	0.949
13	A	−0.537	−0.944	0.331	−0.970	−1.350	−1.153	−0.961	−0.972	−0.631	−0.338	0.202	0.597
14	A	−0.469	−0.663	0.672	−0.586	−1.150	−0.791	−0.579	−0.682	−0.335	−0.338	0.202	0.480
15	A	0.150	−0.904	0.927	−1.377	−1.536	−1.153	−0.961	−1.314	−0.899	−0.338	0.202	0.363

16	A	0.082	−0.081	1.864	−1.015	−1.178	1.019	−0.961	−1.235	−0.813	−0.338	−0.418	−0.106
17	A	1.939	−1.386	3.654	−0.903	−1.064	−0.429	−0.579	−0.893	−0.641	−0.338	−0.418	0.245
18	A	0.563	−1.165	2.205	−1.151	−1.250	−0.791	−0.579	−1.183	−0.985	−0.338	−0.418	−0.224
19	A	0.426	−0.182	1.524	−0.338	−1.422	−0.429	−1.727	−1.552	−1.195	−0.338	−0.418	−0.224
20	B	−0.744	0.240	−0.947	0.542	1.454	−0.067	1.718	0.742	1.184	−0.338	0.202	0.245
21	B	−1.225	−0.784	−0.947	−0.857	0.424	−0.791	0.952	−0.655	0.162	−0.338	−1.657	−0.810
22	B	−0.331	0.059	−0.351	1.874	1.897	−0.429	2.100	1.427	1.844	−0.338	0.202	0.011
23	B	−0.469	−0.442	−0.351	0.565	1.039	−1.153	1.335	0.320	1.098	−0.338	−0.418	−0.458
24	B	0.494	−0.944	−0.095	0.384	0.739	−1.515	0.952	0.399	0.640	−0.338	−0.418	−0.458
25	B	1.182	−0.944	0.160	0.633	0.767	−1.515	1.335	0.399	0.955	−0.338	−0.418	−0.458
26	B	−0.744	1.243	−0.947	0.700	1.454	0.657	0.952	1.401	1.213	−0.338	−0.418	−0.224
27	B	−0.950	0.280	−1.032	0.633	0.867	−0.429	0.952	0.531	0.716	−0.338	−1.037	−0.693
28	B	−1.225	0.882	−1.117	0.655	1.726	0.657	0.952	2.060	1.586	−0.338	−0.418	−0.224
29	B	−0.881	1.364	−1.032	1.626	2.112	0.657	1.718	2.350	2.006	−0.338	−0.418	−0.575
30	B	1.595	1.424	−0.265	0.407	1.039	1.019	0.570	1.743	1.595	−0.338	−0.418	−0.341
31	B	0.357	−0.683	−0.691	−1.580	−0.449	−1.153	−0.196	−0.708	−0.727	−2.759	−2.276	−1.866
32	B	0.288	−0.222	−0.521	0.565	0.796	−0.429	0.952	1.533	1.442	−0.338	−0.418	−0.575
33	B	0.013	0.681	−0.606	0.317	0.681	1.019	0.952	1.111	1.117	2.084	−0.418	−0.927
34	B	0.219	1.103	−0.436	−0.586	−0.134	−0.429	−0.196	−0.471	−0.832	−0.338	−1.037	−0.927

Continued

TABLE 5 Autoscaled Data—cont'd

Sample	Type	(1) Acetal-dehyde	(2) Ethyl Acetate	(3) Acetal	(4) Propanol	(5) Isobutanol	(6) Isoamyl Acetal	(7) Butanol-1	(8) 2-Me-1-Butanol	(9) 3-Me-1-Butanol	(10) Et Caproate	(11) Ethyl Caprylate	(12) Et Caprate
35	B	0.150	1.404	−0.436	0.136	0.438	−0.067	0.570	0.399	0.171	−0.338	−1.037	−1.162
36	B	2.214	1.725	−0.010	0.475	0.038	−1.153	−0.196	−0.207	−0.756	−0.338	0.202	−0.575
37	B	2.420	1.805	−0.095	0.497	0.023	−1.153	0.187	−0.181	−0.545	−0.338	−1.037	−1.279
38	B	0.426	0.340	−0.606	0.023	0.052	−0.429	−0.961	−0.524	−1.109	−0.338	−0.418	−0.693
39	B	−0.194	0.742	−0.777	0.475	0.252	−0.429	−0.579	−0.260	−0.823	−0.338	−1.657	−1.514
40	B	1.801	1.564	−0.351	0.407	0.424	0.295	−1.344	−0.260	−1.090	−0.338	−0.418	−0.693
41	B	0.976	1.444	−0.606	0.429	0.352	0.295	−0.961	−0.339	−1.109	−0.338	−1.037	−1.396
42	B	−1.776	0.380	−1.288	−0.722	−0.120	−0.067	−0.961	−0.972	−1.778	−0.338	0.202	−0.575
43	B	−1.019	0.681	−1.032	−0.609	0.066	0.295	−0.961	−0.497	−1.262	−0.338	−0.418	−1.162

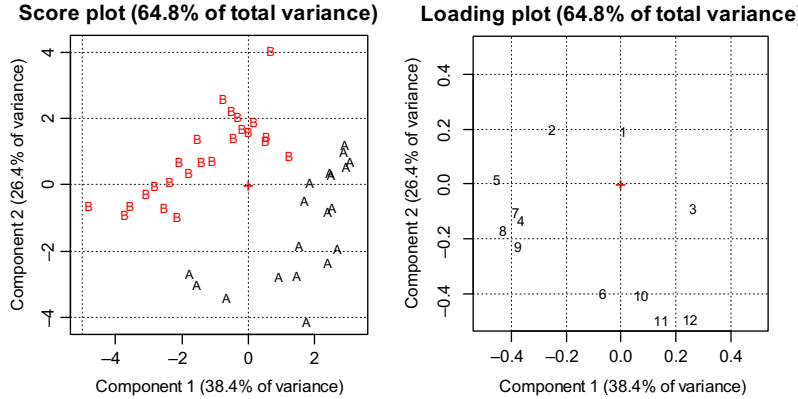

FIG. 3 PCA of the data in Table 4. On the left, the score plot of the objects (coded and colored according to the whiskey type), on the right the loading plot of the variables (coded according to the order in Table 4).

So, we have shown that the two types of whiskeys are really different also from the chemical point of view.

Now, let us look at Fig. 4, the samples are coded according to Table 4, that is, following the production order. It can be seen that for both types there is a trend from the left-hand side of the plot (negative values of PC1) to the right-hand side of the plot (positive values of PC1), with this effect being much clearer for type A. As it has previously been said, PC1 is mainly related to the alcohols. Therefore, it can be concluded that throughout the production period taken into account there has been a progressive decrease of the alcohol content (though it is not possible to say whether this effect is related to a plant issue or to a trend in the analytical system). While the previous finding was the answer to a question that was explicitly formulated by the producer ("are the two types of whiskey different?") this result came out totally unexpected. This shows very well what is mentioned in a paper by Bro et al. (2002): "Usually, data analysis is performed as a confirmatory exercise, where a postulated hypothesis is claimed, data generated accordingly and the data analyzed in order either to verify or reject this hypothesis. No new knowledge is obtained in confirmatory analysis except the possible verification of a prior postulated hypothesis. Using exploratory analysis the data are gathered in order to represent as broadly and as well as possible the problem under investigation. The data are analyzed and through the, often visual, inspection of the results, hypotheses are suggested on the basis of the empirical data. Consequently, exploratory data analysis is an extraordinary tool in displaying thus far unknown information from established and potential monitoring methods."

TABLE 6 Loadings of the Variables on PC1 and PC2

	(1) Acetald-ehyde	(2) Ethyl Acetate	(3) Acetal	(4) Propanol	(5) Isobutanol	(6) Isoamyl Acetal	(7) Butanol-1	(8) 2-Me-1-Butanol	(9) 3-Me-1-Butanol	(10) Et Caproate	(11) Ethyl Caprylate	(12) Et Caprate
PC1	0.006	−0.253	0.261	−0.363	−0.452	−0.067	−0.385	−0.429	−0.378	0.071	0.146	0.159
PC2	0.196	0.206	−0.086	−0.129	0.023	−0.395	−0.096	−0.162	−0.221	−0.404	−0.493	−0.498

TABLE 7 Scores of the Objects on PC1 and PC2

Object	Category	Score on PC1	Score on PC2
1	A	−1.778	−2.654
2	A	−1.581	−2.974
3	A	1.477	−2.730
4	A	−0.679	−3.359
5	A	0.921	−2.744
6	A	1.737	−4.096
7	A	2.675	−1.889
8	A	1.673	−0.432
9	A	1.534	−1.811
10	A	2.526	−0.650
11	A	2.395	−0.793
12	A	2.395	−2.350
13	A	2.488	0.350
14	A	1.850	0.113
15	A	3.080	0.722
16	A	2.446	0.409
17	A	2.955	0.603
18	A	2.891	1.031
19	A	2.903	1.231
20	B	−2.546	−0.658
21	B	−0.426	1.448
22	B	−3.730	−0.862
23	B	−1.805	0.400
24	B	−1.093	0.748
25	B	−1.392	0.723
26	B	−3.069	−0.258
27	B	−2.090	0.724
28	B	−3.556	−0.589
29	B	−4.814	−0.585
30	B	−2.815	0.020

Continued

TABLE 7 Scores of the Objects on PC1 and PC2—cont'd

Object	Category	Score on PC1	Score on PC2
31	B	0.677	4.098
32	B	−2.361	0.140
33	B	−2.148	−0.925
34	B	0.180	1.939
35	B	−1.527	1.444
36	B	−0.173	1.729
37	B	−0.747	2.663
38	B	0.575	1.484
39	B	−0.510	2.283
40	B	0.012	1.648
41	B	−0.314	2.114
42	B	1.251	0.927
43	B	0.514	1.368

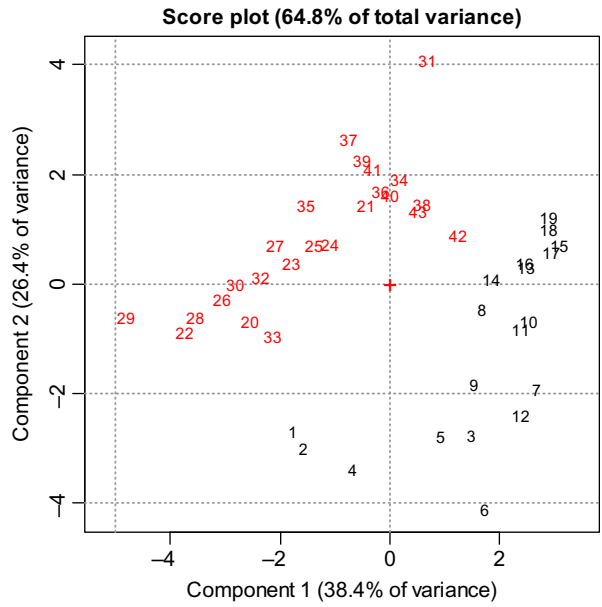

FIG. 4 Score plot of the data in Table 4. The samples are colored according to the whiskey type and coded according to the production order.

4 PROCESS MONITORING AND QUALITY CONTROL

When running a process it is very important to know whether it is under control (i.e., inside its natural variability) or out of control (i.e., in a condition that is not typical and therefore can lead to an accident).

Analogously, when producing a product it is very important to know whether each single piece is inside specifications (i.e., close to the "ideal" product, inside its natural variability) or out of specifications (i.e., significantly different from the "standard" product and therefore in a condition possibly leading to a complaint by the final client).

PCA is the basis for a multivariate process monitoring and a multivariate quality control, much more effective than the usually applied univariate approaches (Kourti and MacGregor, 1995).

After having collected a relevant number of observations describing the "normally operating" process (or the "inside specification" products), encompassing all the sources of normal variability, it will be possible to build a PCA model defining the limits inside which the process (or the product) should stay.

Any new set of measurements (a vector $\mathbf{x}_{1,v}$) describing the process in a given moment (or a new product) will be projected onto the previously defined model using the following equation: $\mathbf{s}_{1,c} = \mathbf{x}_{1,v} * \mathbf{L}_{c,v}$'. From the computed scores, it can be estimated how far from the barycenter of the model, that is, from the "ideal" process (or product) it is.

Its residuals can also be easily computed: $\mathbf{e}_{1,v} = \mathbf{x}_{1,v} - \mathbf{s}_{1,c} * \mathbf{L}_{c,v}$ ($\mathbf{e}_{1,v}$ is the vector of the residuals, and each of its v elements corresponds to the difference between the measured and reconstructed value of each variable). From them, it can be understood how well the sample is reconstructed by the PCA model, that is, how far from the model space (a plane, in case $c = 2$) it lies.

Statistical tests make possible the automatic detection of an outlier in both cases (they are defined as T^2 outliers in the first case and Q outliers in the second case). With these simple tests it will be possible to detect a fault in a process or to reject a bad product by checking just two plots, instead of as many plots as variables, as in the case of the Shewhart charts commonly used when the univariate approach is applied. Finally, the contribution plots will easily outline which variables are responsible for the sample being an outlier.

Fig. 5 shows the output of a process monitoring software (Leardi et al., 2007) when the process (in this case, a continuous two-column whiskey distillation pilot plant, with 26 process variables being monitored) is in an ideal condition.

The actual situation of the plant (star) is well inside both the confidence ellipses of the PCA model (left plot) and the critical limits of the T^2 and Q statistics (right plot). Furthermore, the trajectories in both plots also demonstrate that the variability of the plant in the last hour has been quite small.

From Fig. 6 it is instead very easy to understand that the process is no longer under control. Looking at the PCA plot (left plot), it can be seen that the star is at the border of the external ellipse, corresponding to $p = 0.001$; furthermore, the

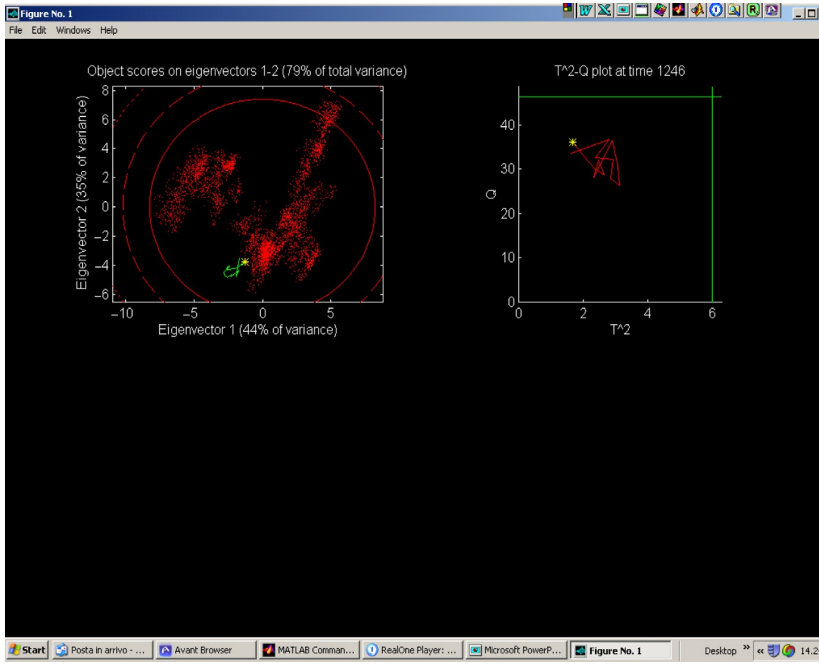

FIG. 5 Output of a process monitoring software when the process is under control.

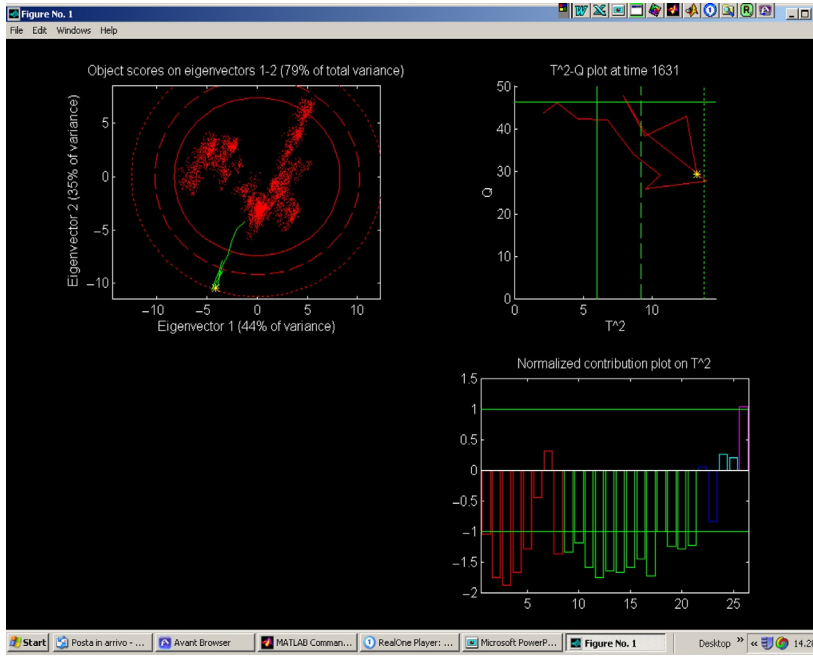

FIG. 6 Output of a process monitoring software when the process is out of control.

trajectory shows the presence of a very clear trend which took place during the last hour. Looking at the T^2-Q plot, one can understand that, while the T^2 value is very close to the $p = 0.001$ critical limit, the Q value is well below any critical limit, this meaning that the correlations inside the plant are still preserved.

The T^2 normalized contribution plot shows that the anomalous situation depends on the fact that the temperatures in the two columns (variables 1–21) are lower than the normal values.

Compared with the standard univariate approach, the multivariate approach is much more robust, since it will lead to a lower number of false negative and false positives, and much more sensitive, since it allows the detection of faults at an earlier stage.

Furthermore, in many cases a process out of control produces a product which is also out of specification. Therefore, the multivariate process monitoring also gives an indirect estimation of the quality of the final product, without having to perform any direct analysis on the product itself, whose results would be known some time later.

Morris and Martin (2003) tried to quantify the advantage brought by multivariate process monitoring and came to the following conclusions: "Better process control not only increases yields and results in more consistent high-quality production, but has contributed to reducing specific energy costs by around 5%–6%, coupled with production increases of up to 10%. It is conjectured that, if 10% of the 100,000 process manufacturing plants in Europe embrace these technologies, a net benefit of the order of €500 million per annum to the European process manufacturing industries could result (EU Project Prognosis)."

5 THREE-WAY PCA

It can happen that the structure of a data set is such that a standard two-way table (objects versus variables) is not enough to describe it. Let us suppose that some food samples have been analyzed by a panel of assessors, each one of them giving scores to different attributes. A third way needs to be added to adequately represent the data set, which can be imagined as a parallelepiped of size $I \times J \times K$, where I is the number of food samples (objects), J is the number of attributes (variables), and K is the number of assessors (conditions) (Geladi, 1989, Smilde, 1992).

To apply standard PCA, these three-way data arrays $\underline{\mathbf{X}}$ have to be somehow transformed to obtain a two-way data table. This can be done in different ways, according to what one is interested in focusing on.

The usual and simplest transformation is to average the scores given by the different assessors. By doing that, a matrix with I rows and J columns is obtained. This is simply done, but the price to be paid is that, since we are now dealing with the scores given by a hypothetical "average" assessor, we have lost every kind of information related to the assessors, such as the

variability with which each attribute is assessed and the systematic effect typical of each assessor.

A different transformation applied to study the food samples consists in matricizing the data array $\underline{\mathbf{X}}$ to $\mathbf{X}'_a$ (I rows, $J \times K$ columns). The interpretability of the score plot is usually very high, but since $J \times K$ is usually a rather large number, the interpretation of the loading plot is very difficult. The same considerations can be made when focusing on the assessors: in this case, $\mathbf{X}'_c$ is obtained (K rows, $I \times J$ columns).

Three-way PCA allows a much easier interpretation of the information contained in the data set, since it directly takes into account its three-way structure. If the Tucker3 model (Tucker, 1966) is applied, the final result is given by three sets of loadings together with a core array describing the relationship among them. If the number of components is the same for each way, the core array is a cube. Each of the three sets of loadings can be displayed and interpreted in the same way as a score plot of standard PCA.

In the case of a cubic core array a series of orthogonal rotations can be performed on the three spaces of the objects, variables, and conditions, looking for the common orientation for which the core array is as much body diagonal as possible. If this condition is sufficiently achieved, then the rotated sets of loadings can also be interpreted jointly by overlapping them.

An example of application of three-way PCA is a data set from the field of sensory evaluation (Cordella et al., 2011). In it, eight types of noodles, each corresponding to a different formulation, were produced in four independent replicates, with each replicate tested by the panel in two independent sessions. Each of the 12 panelists gave a score to eight descriptors [(1) yellow color, (2) translucency, (3) shininess, (4) surface smoothness, (5) firmness, (6) chewiness, (7) surface stickiness, (8) elasticity]. The data set can therefore be seen as a $64 \times 8 \times 12$ data set.

By taking into account the loading plots of the objects (Fig. 7), it can be seen that the regions occupied by the eight samples of each noodle (4 replicates $\times$ 2 sessions) never overlap. This means that the global variability (production + sensory evaluation) of each noodle is always smaller than the differences among the noodles. The fact that the region spanned by each noodle is approximately the same (with the exception of noodle 2) indicates that the global variability can be considered as independent of the type of noodle. It can also be seen that the variability between sessions is smaller than the variability among replicates, this meaning that the "instrumental error" of the judges is smaller than the variability of the production.

On the first axis, noodles 7 and 8 have the lowest loading, followed by noodle 6 and then by the remaining five types, all with very similar loading. This ranking ($7 = 8 > 6$) corresponds to the content of glyceryl monostearate (GMS, 2.8%, 2.8%, and 1.4%, respectively, with the other noodles having no GMS). It can be concluded that the loading of each noodle group on the first axis is directly related to the GMS content.

FIG. 7 Scatter plot of the loadings of the objects. Objects 1–8: noodle 1; objects 9–16: noodle 2; objects 57–64: noodle 8.

On the second axis, the five formulations having no GMS are discriminated, with noodle 3 having the highest loading and noodle 1 having by far the lowest loading. Noodle 3 is made only by durum wheat four (DWF), while noodle 1 is the only one containing wheat starch (WS). On the same axis, noodles 5, 4, and 2 have decreasing loadings, and this corresponds to their amount of wheat gluten (WG, 6%, 3%, and 0%, respectively).

Fig. 8 shows the scatter plot of the loadings of the variables. Variables 5–8 (the texture-related descriptors) have the highest values on the first axis. This means that the first axis is mainly related to the texture of the product. Variables 1 and 4 (color and smooth, both positive attributes) have positive loadings on axis 2, in contrast with variables 2 and 3 (translucent and shiny, both negative attributes). Therefore, the second axis is mainly related to the appearance attributes of the noodles.

It should also be noticed that variables 5–8 (the texture-related descriptors) have very similar loadings on both axes, and therefore are very highly correlated. As a result, it can be concluded that axis 1 is related to the amount of GMS and to the texture of the product; it can be seen that the addition of GMS gives a worse product.

Axis 2 is related to the aspect; it can be seen that noodle 3, obtained with DWF, is the product with the best appearance (the most yellow and the smoothest), while noodle 1, obtained with a large amount of WS, has the worst

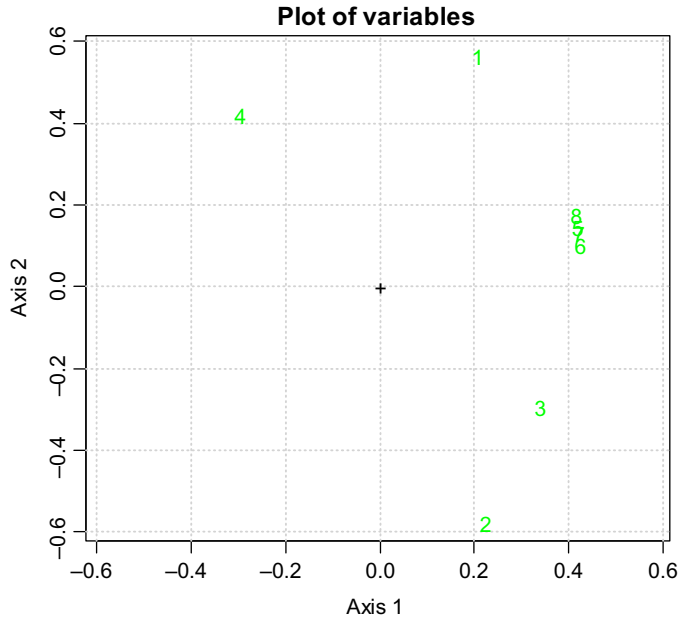

FIG. 8 Scatter plot of the loadings of the variables.

appearance (the most translucent and the most shiny). The addition of WG also improves the appearance, since it results in an increase of the yellow color and of the smoothness. By taking into account both axes, it is easy to detect noodle 3 as the best one. Table 8 shows some types of data sets on which three-way PCA can be successfully applied.

6 DISCRIMINANT CLASSIFICATION

In Section 3 we could verify that the two types of whiskey are indeed well separated in the multivariate space of the variables. Therefore, we can say that we have two really different classes. Let us suppose that we now get some unknown samples and we want to know what their class is. After having performed the chemical analyses, we can add these data to the previous data set, run a PCA and see where the new samples are placed. This will be fine if the new samples fall inside one of the clouds of points corresponding to a category, but what if they fall in a somehow intermediate position? How can we say with "reasonable certainty" that the new samples are from type A or type B? We know that PCA is a very powerful technique for data display, but we realize that we need something different if we want to classify new samples. What we want is a technique producing some "decision rules" discriminating among the possible categories.

TABLE 8 Data Sets on Which Three-Way PCA can be Applied

Field of Application	Objects	Variables	Conditions
Environmental analysis	Air or water samples	Chemicophysical analyses	Time
Environmental analysis	Water samples (different locations)	Chemicophysical analyses	Depth
Panel tests	Food products (oils, wines)	Attributes	Assessors
Food chemistry	Foods (cheeses, spirits, etc.)	Chemical composition	Aging
Food chemistry	Foods (oils, wines, etc.)	Chemical composition	Crops
Sport medicine	Athletes	Blood analyses	Time after effort
Process monitoring	Batches	Chemical analyses	Time

While PCA is an "unsupervised" technique, the discriminant classification methods are "supervised" techniques. In these techniques the category of each of the objects on which the model is built must be specified in advance.

The most commonly used techniques are linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA). They define a set of delimiters (according to the number of categories under study) in such a way that the multivariate space of the objects is divided in as many subspaces as the number of categories, and that each point of the space belongs to one and only one subspace. Rather than describing in detail the algorithms behind these techniques, special attention will be given to the critical points of a classification.

As previously stated, the discriminant classification techniques use objects belonging to the different categories to define boundaries delimiting regions of the space. The final goal is to apply these classification rules to new objects for their classification into one of the existing categories. The performance of the technique can be expressed as classification ability and prediction ability. The difference between "classification" and "prediction", though quite subtle at first glance, is actually very important and its underestimation can lead to very bitter deceptions.

Classification ability is the capability of assigning to the correct category the same objects used to build the classification rules, while prediction ability is the capability of assigning to the correct category objects that have not been used to build the classification rules. Since the final goal is the classification of new

TABLE 9 Example of the Performance of a Classification Technique

Category #	Objects	Correct Class.	% Correct Class.
1	112	105	93.8
2	87	86	98.9
3	21	10	47.6
Total	220	201	91.4 / 80.1

samples, it has to be clear that the predictive ability is by far the most important figure of merit to be looked at.

The results of a discriminant classification method can be expressed in several ways. The most synthetic one is the percentage of correct classifications (or predictions). Note that, in the following, only the term "classification" will be used, but it has to be understood as "classification or prediction". This can be obtained as the number of correct classifications (independently of the category) divided by the total number of objects, or as the average of the performance of the model over all the categories. The two results are very similar when the size of all the categories is very similar, but can be very different if the size is quite different. Let us consider the case shown in Table 9.

The very poor performance of category 3, by far the smallest one, almost does not affect the classification rate computed on the global number of classifications, while it produces a much lower result if the classification rate is computed as the average of the three categories.

A more complete and detailed overview of the performance of the method can be obtained using the confusion matrix, that also allows to know the categories to which the wrongly classified objects are assigned (in many cases the cost of an error can be quite different according to the category the sample is assigned to). In it, each row corresponds to the true category and each column to the category to which the sample has been assigned. Continuing with the previous example, a possible confusion matrix is the one shown in Table 10.

TABLE 10 Example of a Confusion Matrix

Category	1	2	3
1	105	0	7
2	1	86	0
3	11	0	10

From it, it can be seen that the 112 objects of category 1 were classified in the following way: 105 correctly to category 1, none to category 2, and seven to category 3. In the same way, it can be deduced that all the objects of category 3 which were not correctly classified have been assigned to category 1. Therefore, it is easy to conclude that category 2 is well defined and that the classification of its objects gives no problems at all, while categories 1 and 3 are quite overlapping. As a consequence, to have a perfect classification more effort must be put into better separating categories 1 and 3. All this information cannot be obtained from just the percentage of correct classifications.

If overfitting occurs, then the prediction ability will be much worse than the classification ability. To avoid it, it is very important that the sample size is adequate to the problem and to the technique. A general rule is that the number of objects should be more than five times (at least, no less than three times) the number of parameters to be estimated. LDA works on a pooled variance–covariance matrix: this means that the total number of objects should be at least five times the number of variables. QDA computes a variance-covariance matrix for each category, which makes it a more powerful method than LDA, but this also means that each category should have a number of objects at least three times higher than the number of variables. This is a good example of how the more complex, and therefore "better" methods, sometimes cannot be used in a safe way because their requirements do not correspond to the characteristics of the data set.

7 MODELING

In discriminant classification, the space is divided into as many subspaces as categories, and each point belongs to one and only one category. This means that the samples that will be predicted by such methods must belong to one of the categories used to build the models; if not, they will anyway be assigned to one of them. To make this concept clearer, let us suppose that the discriminant classification technique is used to discriminate between water and wine. Of course, this discrimination is very easy. Each sample of water will be correctly assigned to the category "water" and each sample of wine will be correctly assigned to the category "wine." However, what will happen when classifying a sample of orange squash? The sample will be assigned either to the category "water" (if variables such as alcohol are taken into account) or to the category "wine" (if variables such as color are considered). The discriminant classification techniques are therefore not able to define a new sample as being "something different" from all the categories of the training set. This is instead the main feature of the modeling techniques.

Though several techniques are used for modeling purpose, UNEQ (one of the modeling versions of QDA) and SIMCA (soft independent model of class analogy) are the most used. While in classification every point of the space belongs to one and only one category, with these techniques the models (one

for each category) can overlap and leave some regions of the space unassigned. This means that every point of the space can belong to one category (the sample has been recognized as a sample of that class), to more than one category (the sample has such characteristics that it could be a sample of more than one class) or to none of the categories (the sample has been considered as being different from all the classes).

Of course, the "ideal" performance of such a method would be not only to correctly classify all the samples in their category (as in the case of a discriminant classification technique), but also be such that the models of each category could be able to accept all the samples of that category and to reject all the samples of the other categories. The results of a modeling technique are expressed as specificity and sensitivity. For category *c*, its specificity (how much the model rejects the objects of different categories) is the percentage of the objects of categories different from *c* rejected by the model, while its sensitivity (how much the model accepts the objects of the same category) is the percentage of the objects of category *c* accepted by the model.

While the discriminant classification techniques need at least two categories, the modeling techniques can also be applied when only one category is present (in which case also the term "one-class classification" can be used), with the goal of defining if a new sample can be considered as a typical sample of that category or not. This can be very useful in the case of protected denomination of origin products, to verify whether a sample, declared as having been produced in a well-defined region, has indeed the characteristics typical of the samples produced in that region.

The application of a multivariate analysis will greatly reduce the possibility of frauds. While an "expert" can adulterate a product in such a way that all the variables, independently considered, still stay in the accepted range, it is almost impossible to adulterate a product in such a way that its multivariate "pattern" is still accepted by the model of the original product, unless the amount of the adulterant is so small that it becomes unprofitable from the economic point of view.

From what has been previously said, the difference between discriminant classification and modeling should be quite obvious. Unfortunately, many people apply a discriminant classification when instead modeling should be the correct approach. This is particularly true in the case of food authenticity.

Let us suppose that we want to verify the origin of a product labeled as "from region x." Does it mean that if the tree from which the product has been obtained was growing a few meters beyond the political border of that specific region we must reject it as a fraud? Of course we will never do it, also because in many cases the political borders are just lines drawn on a map. What we must be interested to is instead to verify if the characteristics of the product under investigation are "compatible" with those of the products from the region from which it is claimed to come from. Therefore, the correct approach is to build a model with samples whose origin is certified and check if our sample is accepted by it.

Obviously, the most difficult point is the selection of the samples on which the model will be built, since they must constitute a representative sampling of the products produced in that region.

The most common error is to use a discriminant classification approach in which the samples are divided into two categories: samples from the region under investigation and samples from all the other regions. This approach is wrong by definition, since all the samples of the same category must have some characteristic in common (e.g., production area, production technique, age, vintage, etc.). In that case the only thing samples from the second category have in common is a negative attribute (they have not been produced in the region of interest), and therefore they cannot be considered as being members of the same category.

As Oliveri (2017) writes in his tutorial, "Although, in the food analytical field, most of the issues would be properly addressed by class modeling strategies, the use of such techniques is rather limited and, in many cases, discriminant methods are forcedly used for one-class problems, introducing a bias in the outcomes".

8 CALIBRATION

Let us imagine we have a set of wine samples and that on each of them the FT-IR spectrum is measured, together with some variables such as alcohol content, pH, or total acidity. Of course, chemical analyses will require much more time than a simple spectral measurement. It would therefore be very useful to find a relationship between each of the chemical variables and the spectrum. This relationship, after having been established and validated, will be used to predict the content of the chemical variables. It is easy to understand how much time (and money) this will save, since in a few minutes it will be possible to have the same results as previously obtained by a whole set of chemical analyses.

Generally speaking, we can say that multivariate calibration finds relationships between one or more response variables y and a vector of predictor variables $\mathbf{x}$. As the previous example should have shown, the final goal of multivariate calibration is not just to "describe" the relationship between the $\mathbf{x}$ and the y variables in the set of samples on which the relationship has been computed, but to find a real practical application for samples that in a following time will only have the $\mathbf{x}$ variables measured.

The model is a linear polynomial ($y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_Kx_K + f$), where b_0 is an offset, the b_k ($k = 1, \ldots, K$) is regression coefficients and f is a residual. The "traditional" method of calculating $\mathbf{b}$, the vector of regression coefficients, is multiple linear regression (MLR). However, this method has two major limitations that make it inapplicable to many data sets:

- It cannot handle more variables than objects.
- It is sensitive to collinear variables.

It can be easily seen that both these limitations do not allow the application of MLR to spectral data sets, where the samples are described by a very high number of highly collinear variables. If one wants to use MLR to such data anyway, the only way to do it is to reduce the number of variables and their collinearity through a suitable variable selection (see Section 9).

When describing the PCA, it has been noticed that the components are orthogonal (i.e., uncorrelated) and that the dimensionality of the resulting space (i.e., the number of significant components) is much lower than the dimensionality of the original space. Therefore, it can be seen that both the aforementioned limitations have been overcome. As a consequence, it is possible to apply MLR to the scores originated by PCA. This technique is principal component regression (PCR).

It has to be considered that PCs are computed by taking into account only the **x** variables, without considering at all the *y* variable(s), and are ranked according to the explained variance of the "*x* space." This means that it can happen that the first PC has little or no relevance in explaining the response that we are interested in. This can be easily understood by considering that, even when we have several responses, the PCs to which the responses have to be regressed will be the same.

Nowadays, the most favored regression technique is partial least-squares regression (PLS or PLSR). As happens with PCR, PLS is based on components (or "latent variables"). The PLS components are computed by taking into account both the **x** and the **y** variables, and therefore they are slightly rotated versions of the PCs. As a consequence, their ranking order corresponds to the importance in the modeling of the response. A further difference with MLR and PCR is that, while the former must work on each response variable separately, PLS can be applied to multiple responses at the same time.

Because both PCR and PLS are based on latent variables, a very critical point is the number of components to be retained. Though we know that information is "concentrated" in the first components and that the last components explain just noise, it is not always an easy task to detect the correct number of components (i.e., when information finishes and noise begins). Selecting a lower number of components would mean removing some useful information (underfitting), while selecting a higher number of components would mean to incorporate some noise (overfitting).

Before applying the results of a calibration, it is very important to look for the presence of outliers. Three major types of outliers can be detected: outliers in the x-space (samples for which the *x*-variables are very different from that of the rest of the samples; they can be found by looking at a PCA of the *x*-variables), outliers in the *y*-space (samples with the *y*-variable very different from that of the rest of the samples; they can be found by looking at a histogram of the *y*-variable) and samples for which the calibration model is not valid (they can be found by looking at a histogram of the residuals).

The goodness of a calibration can be summarized by two values, the percentage of variance explained by the model and the root-mean-square error in calibration (RMSEC). The former, being a "normalized" value, gives an initial idea about how much of the variance of the data set is "captured" by the model; the latter, being an absolute value to be interpreted in the same way as a standard deviation, gives information about the magnitude of the error.

As already described in the classification section and as pointed out at the beginning of this section, the goal of a calibration is essentially not to describe the relationship between the response and the x-variables of the samples on which the calibration is computed (training, or calibration, set), but to apply it to future samples where only the cheaper x-variables will be measured. In this case too, the model must be validated by using a set of samples different from those used to compute the model (validation, or test, set). The responses of the objects of the test set will be computed by applying the model obtained by the training set and then compared with their "true" response. From these values the percentage of variance explained in prediction and the root-mean-square error in prediction (RMSEP) can be computed. Provided that the objects forming the two sets have been selected flawlessly, these values give the real performance of the model on new samples.

As an example, the results obtained on wheat samples (Kalivas, 1997) are reported. The NIR spectra of 100 samples have been recorded from 1100 to 2500 nm with a step of 2 nm (701 wavelengths), and on the same samples two responses (moisture and protein) have been measured. A total of 75 samples were used as training set, while the remaining 25 samples constituted the test set. The ranges of the two responses were 12.45–17.36 and 7.75–14.28, and the RMSEP obtained by applying PLS to the whole spectrum were 0.28 and 0.43, respectively.

From Fig. 9, showing the predictions on the test set, it can be seen that the accuracy of the estimation is quite good though in the case of protein a systematic bias can be detected.

9 VARIABLE SELECTION

Usually, not all the variables of a data set bring useful and nonredundant information. Therefore, a variable (or feature) selection can be highly beneficial, since from it the following results are obtained:

- removal of noise and improvement of the performance
- reduction of the number of variables to be measured and simplification of the model

The removal of noisy variables should always be looked for. Though some methods can give good results even with a moderate amount of noise disturbing the information, it is clear that their performance will increase when this noise is removed. So, feature selection is now widely applied also for those techniques

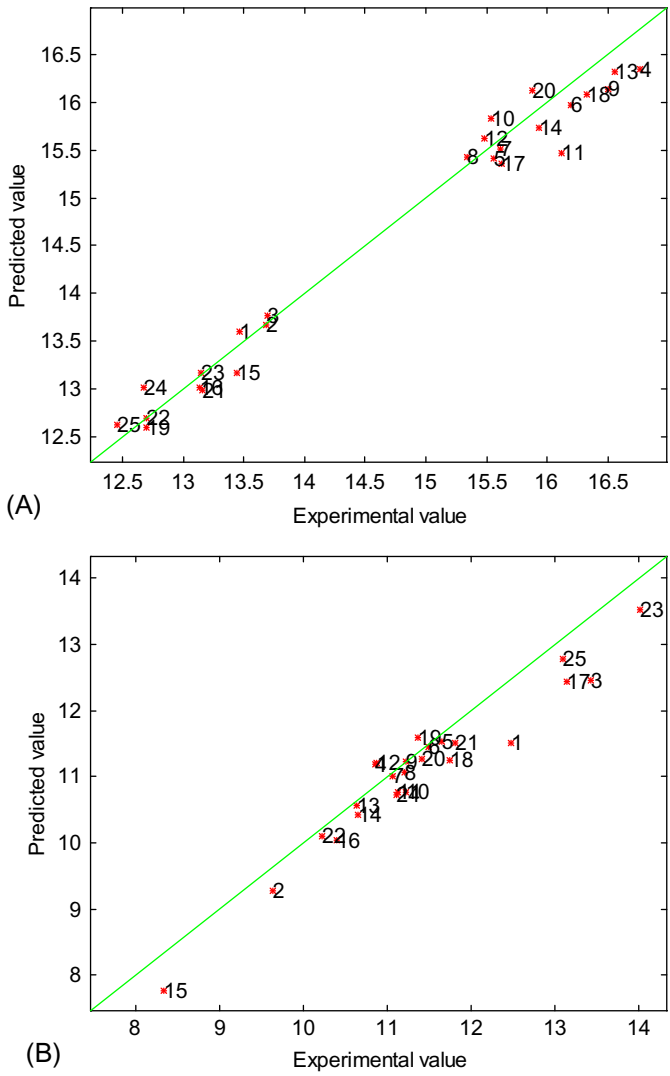

FIG. 9 Experimental vs predicted values of the test set with PLS applied to whole spectrum [(A) moisture and (B) protein].

(PLS and PCR) that in the beginning were considered to be almost insensitive to noise.

While noise reduction is a common goal for any data set, the relevance of the reduction of the number of variables in the final model depends very much on the kind of data constituting the data set, and a very wide range of situations are possible. Let us consider the extreme conditions:

- Each variable requires a separate analysis.
- All the variables are obtained by the same analysis (e.g., chromatographic and spectroscopic data).

In the first case, each variable not selected means a reduction in terms of costs and/or analysis time. The variable selection should therefore always be made on a cost/benefit basis, looking for the subset of variables leading to the best compromise between performance of the model and cost of the analyses. This means that, in the presence of groups of useful but highly correlated (and therefore redundant) variables, only one variable per group should be retained. With such data sets, it is also possible that a subset of variables giving a slightly worse result is preferred, if the reduction in performance is widely compensated by a reduction in costs or time.

In the second case, the number of retained variables has no effect on the analysis cost, while the presence of useful and correlated variables improves the stability of the model. Therefore, the goal of variable selection will improve the predictive ability of the model by removing the variables giving no information, without being worried by the number of retained variables.

Intermediate cases can happen, in which "blocks" of variables are present. As an example, take the case of olive oil samples, on each of which the following analyses have been run: a titration for acidity, the analysis of peroxides, a UV spectroscopy for ΔK, a GC for sterols, and another GC for fatty acids. In such a situation, what counts is not the final number of variables, but the number of analyses one can save.

The only possible way to be sure that "the best" set of variables has been picked up is the "all-models" techniques testing all the possible combinations. Since with k variables the number of possible combinations is $2^k - 1$, it is easy to understand that this approach cannot be used unless the number of variables is really very low (e.g., with 30 variables $>10^9$ combinations should be tested).

The simplest (but least effective) way of performing a feature selection is to operate on a "univariate" basis, by retaining those variables having the greatest discriminating power (in case of a classification) or the greatest correlation with the response (in case of a calibration). By doing that each variable is taken into account by itself without considering how its information "integrates" with the information brought by the other (selected or unselected) variables. As a result, if several highly correlated variables are "good," they are all selected, without taking into account that, owing to their correlation, the information is highly redundant and therefore at least some of them can be removed without any decrease in the performance. On the other hand, those variables that, though not giving by themselves significant information, become very important when their information is integrated with that of other variables, are not taken into account.

An improvement is brought by the "sequential" approaches. They select the best variable first, then the best pair formed by the first and second, and so on in a forward or backward progression. A more sophisticated approach applies a look

back from the progression to reassess previous selections. The problem with these approaches is that only a very small part of the experimental domain is explored and that the number of models to be tested becomes very high in case of highly dimensional data sets, such as spectral data sets. For instance, with 1000 wavelengths, 1000 models are needed for the first cycle (selection or removal of the first variable), 999 for the second cycle, 998 for the third cycle, and so on.

More "multivariate" methods of variable selection, especially suited for PLS applied to spectral data, are available. Among them, we can cite interactive variable selection (Lindgren et al., 1994), uninformative variable elimination (Centner et al., 1996), iterative predictor weighting PLS (Forina et al., 1999), and interval PLS (Nørgaard et al., 2000). The improvements obtained by the application of a variable selection can be checked by looking at Fig. 10.

In Fig. 10, the predictions on the test set of the wheat data after variable selection by genetic algorithms (GA) (see Section 10) are reported (Leardi, 2000). When the variables were reduced from the original 701 to 104 (for moisture) and 64 (for protein), the RMSEP decreased from 0.28 to 0.24 and from 0.43 to 0.30. In the case of protein, it has to be noticed that the bias present when the model was built on the whole spectrum has totally disappeared.

10 FUTURE TRENDS

In future, multivariate analysis should be used more and more in everyday (scientific) life. Until a few decades ago, experimental work resulted in a very limited amount of data, the analysis of which was quite easy and straightforward. Nowadays, it is common to have instrumentation producing an almost continuous flow of data. One example is process monitoring performed by measuring the values of several process variables, at a rate of one measurement every few seconds. Another example is quality control of a final product of a continuous process, on which an FT-IR spectrum is taken every few seconds).

In Section 8 the case of wine FT-IR spectra was cited, from which the most relevant characteristics of the product can be directly predicted. It is therefore clear that the main problem has shifted from obtaining a few data to the treatment of a huge amount of data. It is also clear that standard statistical treatment is not enough to extract the whole information buried in them.

Many instruments have already some chemometric routines built into their software in such a way that their use is totally transparent to the final user (and sometimes the word "chemometrics" is not even mentioned, to avoid possible aversion). Of course, they are "closed" routines, and therefore the user cannot modify them. It is quite obvious that it would be much better if chemometric knowledge were much more widespread, in order that the user could better understand what kind of treatment his data have undergone and eventually modify the routines in order to make them more suitable to his requirements.

As computers become faster and faster, it is nowadays possible to routinely apply some approaches that require very high computing power. Two of them are GAs and artificial neural networks (ANNs).

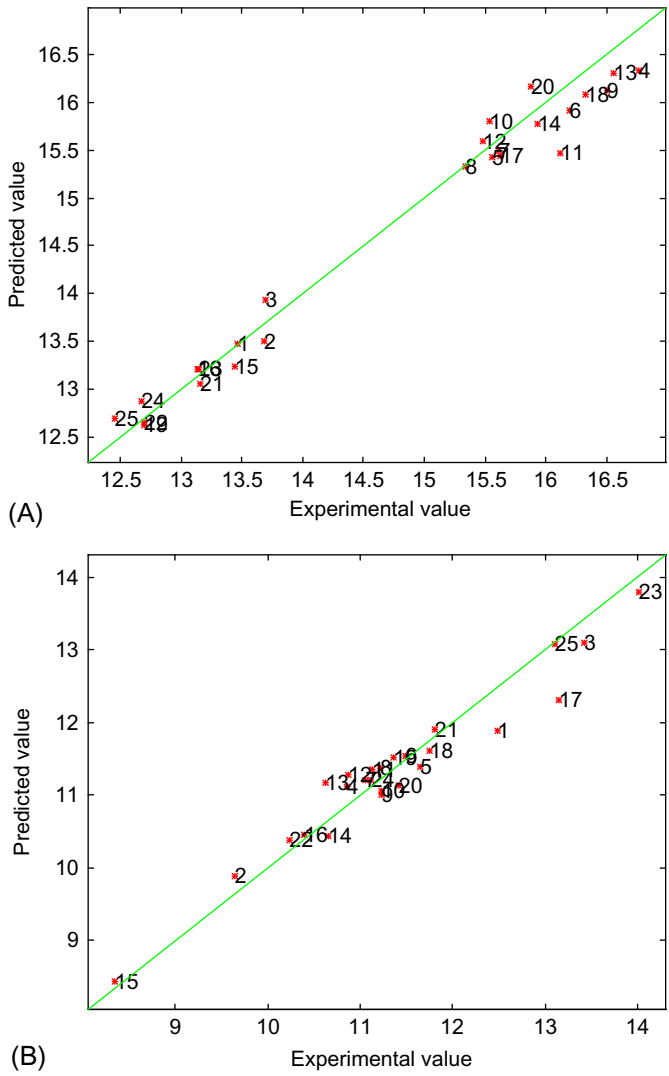

FIG. 10 Experimental vs. predicted values of the test set with PLS applied after variable selection [(A) moisture and (B) protein].

GAs are a general optimization technique with good applicability in many fields, especially when the problem is so complex that it cannot be tackled with "standard" techniques. In chemometrics it has been applied especially in feature selection (Leardi, 2000; Niazi and Leardi, 2012). GAs try to simulate the evolution of a species according to the Darwinian theory. Each experimental condition (in this case, each model) is treated as an individual, whose "performance" (in the case of a feature selection for a calibration problem, it

can be the explained variance) is treated as its "fitness". Through operators simulating the fights among individuals (the best ones have a greatest probability of mating and thus spreading their genome), the mating among individuals (with the consequent "birth" of "offspring" having a genome that is derived by both the parents) and the occurrence of mutations, the GA result in a pattern of search that, by mixing "logical" and "random" features, allows a much more complete search of complex experimental domains.

ANNs try to mimic the behavior of the nervous system to solve practical computational problems. As in life, the structural unit of ANN is the neuron. The input signals are passed to the neuron body, where they are weighted and summed, then they are transformed, by passing through the transfer function into the output of the neuron. The propagation of the signal is determined by the connections between the neurons and by their associated weights. The appropriate setting of the weights is essential for the proper functioning of the network. Finding the proper weight setting is achieved in the training phase. The neurons are usually organized into three different layers: the input layer contains as many neurons as input variables, the hidden layer contains a variable number of neurons, and the output layer contains as many neurons as output variables. All units from one layer are connected to all units of the following layer. The network receives the input signals through the input layer. Information is passed to the hidden layer and finally to the output layer that produces the response.

These techniques are very powerful, but very often they are not applied in a correct way. In such cases, despite a very good performance on the training set (due to overfitting), they will show very poor results when applied to external data sets.

11 THE ADVANTAGES AND DISADVANTAGES OF CHEMOMETRICS

Already 30 years ago, in an issue of the North American International Chemometrics Society Newsletter, Schönkopf (1998) was listing some of the results obtained by applying chemometrics:

- A petroleum producer used chemometrics and could increase productivity with 30% in one oil refinery, earning 14 million USD extra per year.
- A dairy saved $130,000 by not investing in new cooling equipment—a conclusion they draw from smart experiments.
- A petroleum company saves 1 million USD/year by chemometrics-based measurements.
- An agricultural researcher got the same results in 10 min with chemometrics as he got by analyzing his data during 3 months with classical statistics.
- A meat manufacturer saved 150,000 USD/year by reduced waste in a process they optimized.

- A food producer's first chemometric project saved 115,000 USD/year.
- A dispersant developer made 20 experiments and simulated 380 thus reducing their experimental efforts by 95%.
- An oil manufacturer solved a quality problem in 2 weeks with chemometrics after 2 years of failure using traditional methods.

The results are really impressive in terms of money and time (which is the same…) saved. It has also been noticed that four out of the eight examples mentioned are related with the food industry.

In one of his papers Workman Jr. (2002) very efficiently depicts the advantages and disadvantages of multivariate thinking for scientist in industry. From the eight advantages of chemometrics that he clearly outlines, special relevance should be given to the following ones:

1. Chemometrics provides speed in obtaining real-time information from data.
2. It allows high-quality information to be extracted from less resolved data.
3. It promises to improve measurements.
4. It improves knowledge of existing processes.
5. It has very low capital requirements—it is cheap.

The last point especially should convince people to give chemometrics a try. No extra equipment is required: just an ordinary computer and some chemometrical knowledge (or a chemometrical consultancy). It is certain that in the very worst cases the same information as found from a classical analysis will be obtained in a much shorter time and with much more evidence. In the great majority of cases, instead, also a simple PCA can provide much more information than what was previously collected. So, why are people so shy of applying chemometrics? In the same paper previously cited, Workman gives some very common reasons:

1. The perceived disadvantage of chemometrics is that there is widespread ignorance about what it is and what it can realistically accomplish.
2. This science is considered too complex for the average technician and analyst.
3. Chemometrics requires a change in one's approach to problem solving from univariate to multivariate thinking.

So, while chemometrics leads to several real advantages, its "disadvantages" lie only in the general reluctance to use it and accepting the idea that the approach that has been followed over many years can turn out not to be the best one.

12 CONCLUSIONS

This chapter clearly shows that the standard univariate methods are not sufficient to extract the maximum possible information from a data set. To do that, multivariate techniques must be applied. By using them, data display,

classification, modeling, process monitoring, and multivariate calibration can be performed much more efficiently and the results can be interpreted much more easily.

Unfortunately, quite often people devote almost all their efforts to collect the data, while paying almost no attention to the crucial step of transforming them into information.

REFERENCES

Brereton, R., 2013. The evolution of chemometrics. Anal. Methods 5, 3785–3789.

Bro, R., van den Berg, F., Thybo, A., Andersen, C.M., Jørgensen, B.M., Andersen, H., 2002. Multivariate data analysis as a tool in advanced quality monitoring in the food production chain. Trends Food Sci. Technol. 13, 235–244.

Centner, V., Massart, D.L., de Noord, O.E., de Jong, S., Vandeginste, B.M., Sterna, C., 1996. Elimination of uninformative variables for multivariate calibration. Anal. Chem. 68, 3851–3858.

Cordella, C.B.Y., Leardi, R., Rutledge, D., 2011. Three-way principal component analysis applied to noodles sensory data analysis. Chemom. Intell. Lab. Syst. 106, 125–130.

Forina, M., Casolino, C., Pizarro Millán, C., 1999. Iterative predictor weighting (IPW) PLS: a technique for the elimination of useless predictors in regression problems. J. Chemom. 13, 165–184.

Geladi, P., 1989. Analysis of multi-way (multi-mode) data. Chemom. Intell. Lab. Syst. 7, 11–30.

Kalivas, J.H., 1997. Two data sets of near infrared spectra. Chemom. Intell. Lab. Syst. 37, 255–259.

Kourti, T., MacGregor, J.F., 1995. Process analysis, monitoring and diagnosis, using multivariate projection methods. Chemom. Intell. Lab. Syst. 28, 3–21.

Leardi, R., 2000. Application of genetic algorithm-PLS for feature selection in spectral data sets. J. Chemom. 14, 643–655.

Leardi, R., MacNamara, C., MacNamara, K., 2007. Multivariate on-line process monitoring applied to a continuous two-column distillation pilot plant.VI Colloquium Chemiometricum Mediterraneum, Saint-Maximin-la-Sainte-Baume, September 5–7, 2007.

Lindgren, F., Geladi, P., Rännar, S., Wold, S., 1994. Interactive variable selection (IVS) for PLS. 1. Theory and algorithms. J. Chemom. 8, 349–363.

MacNamara, K., 2005. Personal communication.

Morris, J., Martin, E., 2003. Business Briefing. CPI Technology, Pennsylvania.

Niazi, A., Leardi, R., 2012. Genetic algorithms in chemistry. J. Chemom. 26, 345–351.

Nørgaard, L., Saudland, A., Wagner, J., Nielsen, J.P., Munck, L., Engelsen, S.B., 2000. Interval partial least-squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy. Appl. Spectrosc. 54, 413–419.

Oliveri, P., 2017. Class-modelling in food analytical chemistry: Development, sampling, optimisation and validation issues - a tutorial. Anal. Chim. Acta 982, 9–19.

Schönkopf, S., 1998. Chemometrics saves Time & Money. NAmICS Newsletter #17. http://www.namics.nysaes.cornell.edu/news17/money.html.

Smilde, A.K., 1992. Three-way analyses. Problems and prospects. Chemom. Intell. Lab. Syst. 15, 143–157.

Tucker, L.R., 1966. Some mathematical notes on three mode factor analysis. Psychometrika 31, 279–311.

Workman Jr., J., 2002. The state of multivariate thinking for science in industry: 1980–2000. Chemom. Intell. Lab. Syst. 60, 13–23.

FURTHER READING

A. Books

Beebe, K.R., Pell, R.J., Seasholtz, M.B., 1998. Chemometrics: A Practical Guide. Wiley & Sons, New York.

Brereton, R.G., 2003. Chemometrics—Data Analysis for the Laboratory and Chemical Plant. Wiley, Chichester.

Stephen D. Brown, Romà Tauler, and Beata Walczak (editors-in-chief) 2009. Comprehensive Chemometrics—Chemical and Biochemical Data Analysis. Amsterdam: Elsevier.

Leardi, R. (Ed.), 2003. Nature-Inspired Methods in Chemometrics: Genetic Algorithms and Artificial Neural Networks. In: Data Handling in Science and Technology series, vol. 23. Elsevier, Amsterdam.

Manly, B.F.J., 1986. Multivariate Statistical Methods. A Primer. Chapman and Hall, London.

Martens, H., Naes, T., 1991. Multivariate Calibration. Wiley & Sons, New York.

Massart, D.L., Vandeginste, B.G.M., Deming, S.N., Michotte, Y., Kaufman, L., 1990. Chemometrics: A Textbook. Data Handling in Science and Technology series, vol. 2. Elsevier, Amsterdam.

Massart, D.L., Vandeginste, B.G.M., Buydens, L.M.C., de Jong, S., Lewi, P.J., Smeyers-Verbeke, J., 1997. Handbook of Chemometrics and Qualimetrics. Part A. Data Handling in Science and Technology Series, vol. 20A. Elsevier, Amsterdam.

Massart, D.L., Vandeginste, B.G.M., Buydens, L.M.C., de Jong, S., Lewi, P.J., J Smeyers-Verbeke, J., 1998. Handbook of Chemometrics and Qualimetrics. Part B. Data Handling in Science and Technology Series, vol. 20B. Elsevier, Amsterdam.

Meloun, M., Militky, J., Forina, M., 1992. Chemometrics for Analytical Chemistry. PC-aided Statistical Data Analysis, vols. 1. Ellis Horwood, Chichester.

Meloun, M., Militky, J., Forina, M., 1994. Chemometrics for Analytical Chemistry. PC-Aided Regression and Related Methods, vol. 2. Ellis Horwood, Hemel Hempstead.

Pomerantsev, A.L., 2014. Chemometrics in Excel. Wiley & Sons, New York.

Sharaf, M.A., Illman, D.L., Kowalski, B.R., 1986. In: Elving, P.J., Winefordner, J.D. (Eds.), *Chemometrics*, in *Chemical Analysis, a Series of Monographs on Analytical Chemistry and its Applications series*. In: vol. 82. Wiley & Sons, New York.

Varmuza, K., Filzmoser, P., 2008. Introduction to Multivariate Statistical Analysis in Chemometrics. Taylor & Francis, Boca Raton.

Wehrens, R., 2011. Chemometrics with R. Springer-Verlag, Berlin Heidelberg.

B. Web Sites (Tested on December 27, 2017)

https://folk.uio.no/ohammer/past/.

http://gruppochemiometria.it/index.php/software.

http://www.models.life.ku.dk.

http://www.namics.nysaes.cornell.edu.

http://www.statsoft.com/textbook/stathome.html.