

We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,900

Open access books available

186,000

International authors and editors

200M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com



PCA: The Basic Building Block of Chemometrics

Christophe B.Y. Cordella

Additional information is available at the end of the chapter

<http://dx.doi.org/10.5772/51429>

1. Introduction

Modern analytical instruments generate large amounts of data. An infrared (IR) spectrum may include several thousands of data points (wave number). In a GC-MS (Gas Chromatography-Mass spectrometry) analysis, it is common to obtain in a single analysis 600,000 digital values whose size amounts to about 2.5 megabytes or more. There are different methods for dealing with this huge quantity of information. The simplest one is to ignore the bulk of the available data. For example, in the case of a spectroscopic analysis, the spectrum can be reduced to maxima of intensity of some characteristic bands. In the case analysed by GC-MS, the recording is, accordingly, for a special unit of mass and not the full range of units of mass. Until recently, it was indeed impossible to fully explore a large set of data, and many potentially useful pieces of information remained unrevealed. Nowadays, the systematic use of computers makes it possible to completely process huge data collections, with a minimum loss of information. By the intensive use of chemometric tools, it becomes possible to gain a deeper insight and a more complete interpretation of this data. The main objectives of multivariate methods in analytical chemistry include data reduction, grouping and the classification of observations and the modelling of relationships that may exist between variables. The predictive aspect is also an important component of some methods of multivariate analysis. It is actually important to predict whether a new observation belongs to any pre-defined qualitative groups or else to estimate some quantitative feature such as chemical concentration. This chapter presents an essential multivariate method, namely *principal component analysis* (PCA). In order to better understand the fundamentals, we first return to the historical origins of this technique. Then, we will show - via pedagogical examples - the importance of PCA in comparison to traditional univariate data processing methods. With PCA, we move from the one-dimensional vision of a problem to its multidimensional version. Multiway extensions of PCA, PARAFAC and Tucker3 models are exposed in a second part of this chapter with brief historical and bibliographical elements. A PARAFAC example on real data is presented in order to illustrate the interest in this powerful technique for handling high dimensional data.

2. General considerations and historical introduction

The origin of PCA is confounded with that of linear regression. In 1870, Sir Lord Francis Galton worked on the measurement of the physical features of human populations. He assumed that many physical traits are transmitted by heredity. From theoretical assumptions, he supposed that the height of children with exceptionally tall parents will, eventually, tend to have a height close to the mean of the entire population. This assumption greatly disturbed Lord Galton, who interpreted it as a "move to mediocrity," in other words as a kind of *regression* of the human race. This led him in 1889 to formulate his law of *universal regression* which gave birth to the statistical tool of *linear regression*. Nowadays, the word "regression" is still in use in statistical science (but obviously without any connotation of the regression of the human race). Around 30 years later, Karl Pearson,¹ who was one of Galton's disciples, exploited the statistical work of his mentor and built up the mathematical framework of *linear regression*. In doing so, he laid down the basis of the correlation calculation, which plays an important role in PCA. Correlation and linear regression were exploited later on in the fields of psychometrics, biometrics and, much later on, in chemometrics.

Thirty-two years later, Harold Hotelling² made use of correlation in the same spirit as Pearson and imagined a graphical presentation of the results that was easier to interpret than tables of numbers. At this time, Hotelling was concerned with the *economic games* of companies. He worked on the concept of economic competition and introduced a notion of spatial competition in *duopoly*. This refers to an economic situation where many "players" offer similar products or services in possibly overlapping trading areas. Their potential customers are thus in a situation of choice between different available products. This can lead to illegal agreements between companies. Harold Hotelling was already looking at solving this problem. In 1933, he wrote in the *Journal of Educational Psychology* a fundamental article entitled "Analysis of a Complex of Statistical Variables with Principal Components," which finally introduced the use of special variables called *principal components*. These new variables allow for the easier viewing of the intrinsic characteristics of observations.

PCA (also known as Karhunen-Love or Hotelling transform) is a member of those descriptive methods called *multidimensional factorial methods*. It has seen many improvements, including those developed in France by the group headed by Jean-Paul Benzécri³ in the 1960s. This group exploited in particular geometry and graphs. Since PCA is

¹ Famous for his contributions to the foundation of modern statistics (χ^2 test, linear regression, principal components analysis), Karl Pearson (1857-1936), English mathematician, taught mathematics at London University from 1884 to 1933 and was the editor of *The Annals of Eugenics* (1925 to 1936). As a good disciple of Francis Galton (1822-1911), Pearson continues the statistical work of his mentor which leads him to lay the foundation for calculating correlations (on the basis of principal component analysis) and will mark the beginning of a new science of biometrics.

² Harold Hotelling: American economist and statistician (1895-1973). Professor of Economics at Columbia University, he is responsible for significant contributions in statistics in the first half of the 20th century, such as the calculation of variation, the production functions based on profit maximization, the using the t distribution for the validation of assumptions that lead to the calculation of the confidence interval.

³ French statistician. Alumnus of the Ecole Normale Supérieure, professor at the Institute of Statistics, University of Paris. Founder of the French school of data analysis in the years 1960-1990, Jean-Paul Benzécri developed statistical

a descriptive method, it is not based on a probabilistic model of data but simply aims to provide geometric representation.

3. PCA: Description, use and interpretation of the outputs

Among the multivariate analysis techniques, PCA is the most frequently used because it is a starting point in the process of data mining [1, 2]. It aims at minimizing the dimensionality of the data. Indeed, it is common to deal with a lot of data in which a set of n objects is described by a number p of variables. The data is gathered in a matrix X , with n rows and p columns, with an element x_{ij} referring to an element of X at the i^{th} line and the j^{th} column. Usually, a line of X corresponds with an "observation", which can be a set of physicochemical measurements or a spectrum or, more generally, an analytical curve obtained from an analysis of a real sample performed with an instrument producing analytical curves as output data. A column of X is usually called a "variable". With regard to the type of analysis that concerns us, we are typically faced with multidimensional data $n \times p$, where n and p are of the order of several hundreds or even thousands. In such situations, it is difficult to identify in this set any relevant information without the help of a mathematical technique such as PCA. This technique is commonly used in all areas where data analysis is necessary; particularly in the food research laboratories and industries, where it is often used in conjunction with other multivariate techniques such as discriminant analysis (Table 1 indicates a few published works in the area of food from among a huge number of publications involving PCA).

3.1. Some theoretical aspects

The key idea of PCA is to represent the original data matrix X by a product of two matrices (smaller) T and P (respectively the scores matrix and the loadings matrix), such that:

$${}_nX^p = {}_nT^q \cdot {}_p[P^t]^q + {}_nE^p \quad (1)$$

Or the non-matrix version:

$$x_{ij} = \sum_{k=1}^K t_{ik} p_{kj} + e_{ij}$$

With the condition $p_i^t p_j = 0$ and $t_i^t t_j = 0$ for $i \neq j$

This orthogonality is a means to ensure the non-redundancy (at least at a minimum) of information "carried" by each estimated principal component.

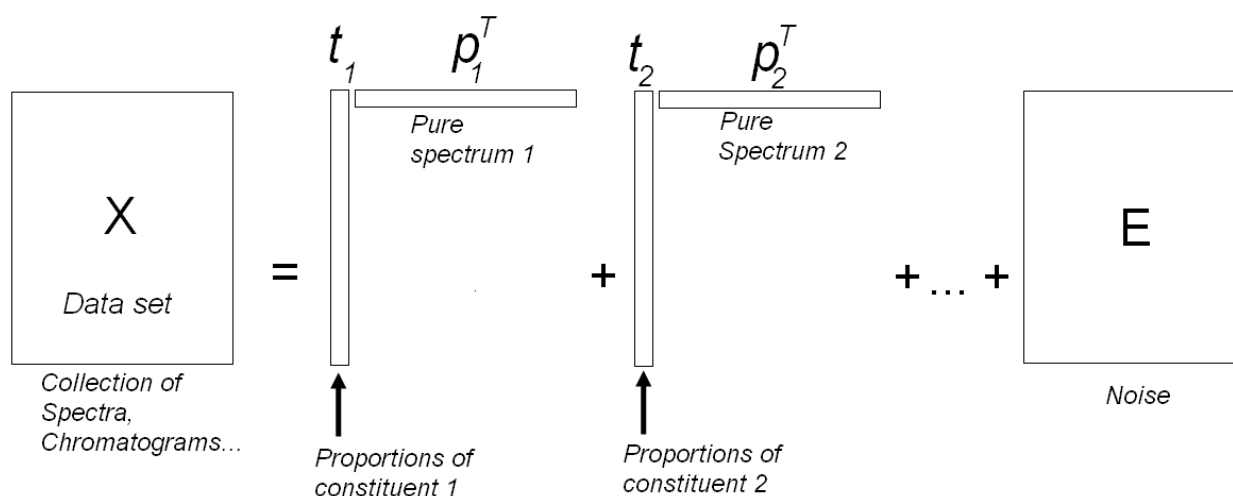
Equation 1 can be expressed in a graphical form, as follows:

tools, such as correspondence analysis that can handle large amounts of data to visualize and prioritize the information.

$$\begin{bmatrix} \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & X & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{bmatrix} = \begin{bmatrix} \cdot & \cdot \\ \cdot & T \\ \cdot & \cdot \\ \cdot & \cdot \end{bmatrix} \begin{bmatrix} \cdot & \cdot & P^t & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{bmatrix} + \begin{bmatrix} \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & E & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{bmatrix}$$

Scheme 1. A matricized representation of the PCA decomposition principle.

Schema 2 translates this representation in a vectorized version which shows how the X matrix is decomposed in a sum of column-vectors (components) and line-vectors (eigenvectors). In a case of spectroscopic data or chromatographic data, these components and eigenvectors take a chemical signification which is, respectively, the proportion of the constituent i for the i^{th} component and the “pure spectrum” or “pure chromatogram” for the i^{th} eigenvector.



Scheme 2. Schematic representation of the PCA decomposition as a sum of "components" and "eigenvectors" with their chemical significance.

The mathematical question behind this re-expression of X is: is there another basis which is a linear combination of the original basis that re-expresses the original data? The term “basis” means here a mathematical basis of unit vectors that support all other vectors of data. Regarding the linearity - which is one of the basic assumptions of the PCA - the general response to this question can be written in the case where X is perfectly re-expressed by the matrix product $T.P^T$, as follows:

$$PX = T \quad (2)$$

Equation 2 represents a *change of basis* and can be interpreted in several ways, such as the transformation of X into T by the application of P or by geometrically saying that P - which is the rotation and a stretch - transforms X into T or the rows of P , where $\{p_1, \dots, p_m\}$ are a set of new basis vectors for expressing the columns of X .

Therefore, the solution offered by PCA consists of finding the matrix P , of which at least three ways are possible:

1. by calculating the eigenvectors of the square, symmetric covariance matrix $X^T X$ (eigenvector analysis implying a diagonalization of the $X^T X$);
2. by calculating the eigenvectors of X by the direct decomposition of X using an iterative procedure (NIPALS);
3. by the singular value decomposition of X - this method is a more general algebraic solution of PCA.

The dual nature of expressions $T = PX$ and $X = TP^T$ leads to a comparable result when PCA is applied on X or on its transposed X^T . The score vectors of one are the eigenvectors of the other. This property is very important and is utilized when we compute the principal components of a matrix using the covariance matrix method (see the pedagogical example below).

3.2. Geometrical point of view

3.2.1. Change of basis vectors and reduction of dimensionality

Consider a p -dimensional space where each dimension is associated with a variable. In this space, each observation is characterized by its coordinates corresponding to the value of variables that describe it. Since the raw data is generally too complex to lead to an interpretable representation in the space of the initial variables, it is necessary to “compress” or “reduce” the p -dimensional space into a space smaller than p , while maintaining the maximum information. The amount of information is statistically represented by the variances. PCA builds new variables by the linear combination of original variables. Geometrically, this change of variables results in a change of axes, called *principal components*, chosen to be orthogonal⁴. Each newly created axis defines a direction that describes a part of the global information.

The first component (i.e. the first axis) is calculated in order to represent the main pieces of information, and then comes the second component which represents a smaller amount of information, and so on. In other words, the p original variables are replaced by a set of new variables, the *components*, which are linear combinations of these original variables. The variances of components are sorted in decreasing order. By construction of PCA, the whole set of components keeps all of the original variance. The dimensions of the space are not then reduced but the change of axis allows a better representation of the data. Moreover, by retaining the q first principal components (with $q < p$), one is assured to retain the maximum of the variance contained in the original data for a q -dimensional space. This reduction from p to q dimensions is the result of the projection of points in a p -dimensional space into a subspace of dimension q . A highlight of the technique is the ability to represent simultaneously or separately the samples and variables in the space of initial components.

⁴ The orthogonality ensures the non-correlation of these axes and, therefore, the information carried by an axis is not even partially redundant with that carried by another.

Food(s)	Analysed compounds	Analytical technique(s)	Chemometrics	Aim of study	Year [Ref.]
Cheeses	Water-soluble compounds	HPLC	PCA, LDA	Classification	1990 [3]
Edible oils	All chemicals between 4800 et 800 cm ⁻¹	FTIR (Mid-IR)	PCA, LDA	Authentication	1994 [4]
Fruit puree	All chemicals between 4000 et 800 cm ⁻¹	FTIR (Mid-IR)	PCA, LDA	Authentication	1995 [5]
Orange juice	All chemicals between 9000 et 4000 cm ⁻¹	FTIR (NIR)	PCA, LDA	Authentication	1995 [6]
Green coffee	All chemicals between 4800 et 350 cm ⁻¹	FTIR (Mid-IR)	PCA, LDA	Origin	1996 [7]
Virgin olive oil	All chemicals between 3250 et 100 cm ⁻¹	FT-Raman (Far- et Mid-Raman)	PCR, LDA, HCA	Authentication	1996 [8]
General ⁵	General	FTIR (Mid-IR)	PCA, LDA + others	Classification & Authentication	1998 [9]
Almonds	Fatty acids	GC	PCA	Origin	1996 [10]
			PCA, LDA		1998 [11]
Garlic products	Volatile sulphur compounds	GC-MS	PCA	Classification	1998 [11]
Cider apples fruits	Physicochemical parameters	Physicochemical techniques + HPLC for sugars	LDA	Classification	1998 [12, 13]
Apple juice	Aromas	Capillary GC-MS + chiral GC-MS	PCA	Authentication	1999 [14]
Meat	All chemicals between 25000 et 4000 cm ⁻¹	IR (NIR+Visible)	PCA + others	Authentication	2000 [15]
Coffees	Physicochemical parameters + chlorogenic acid	Physicochemical techniques + HPLC for chlorogenic acid determination	PCA	Classification according to botanical origin and other criteria	2001 [16]
-	Data from various synthetic substances	MS + IR	PCA, HCA	Comparison of classification methods	2001 [17]
Honeys	Sugars	GC-MS	PCA, LDA	Classification according to floral origin	2001 [18]
Red wine	Phenolic compounds	HPLC	PCA (+PLS)	Relationship between phenolic compounds and antioxidant power	2001 [19]
Honeys	All chemicals between 9000-4000 cm ⁻¹	IR (NIR)	PCA, LDA	Classification	2002 [19]

Table 1. Examples of analytical work on various food products, involving the PCA (and/or LDA) as tools for data processing (from 1990 to 2002). Not exhaustive.

⁵ Comprehensive overview on the use of chemometric tools for processing data derived from infrared spectroscopy applied to food analysis.

3.2.2. Correlation circle for discontinuous variables

In the case of physicochemical or sensorial data, more generally in cases where variables are not continuous like in spectroscopy or chromatography, a powerful tool is useful for interpreting the meaning of the axes: the correlation circle. On this graph, each variable is associated with a point whose coordinate on an axis factor is a measure of the correlation between the variable and the factor. In the space of dimension p , the maximum distance of the variables at the origin is equal to 1. So, by projection on a factorial plan, the variables are part of a circle of radius 1 (the correlation circle) and the closer they are near the edge of the circle, the more they are represented by the plane of factors. Therefore, the variables correlate well with the two factors constituting the plan.

The angle between two variables, measured by its cosine, is equal to the linear correlation coefficient between two variables: $\cos(\text{angle}) = r(V1, V2)$

- if the points are very close (angle close to 0): $\cos(\text{angle}) = r(V1, V2) = 1$ then $V1$ and $V2$ are very highly positively correlated
- if α is equal to 90° , $\cos(\text{angle}) = r(V1, V2) = 0$ then no linear correlation between $X1$ and $X2$
- if the points are opposite, α is 180° , $\cos(\text{angle}) = r(V1, V2) = -1$: $V1$ and $V2$ are very strongly negatively correlated.

An illustration of this is given by figure 1, which presents a correlation circle obtained from a principal component analysis on physicochemical data measured on palm oil samples. In this example, different chemical parameters (such as *Lauric acid* content, concentration of *saponifiable* compounds, iodine index, *oleic acid* content, etc.) have been measured. One can note for example on the PC1 axis, "*iodine index*" and "*Palmitic*" are close together and so have a high correlation; in the same way, "*iodine index*" and "*Oleic*" are positively correlated because they are close together and close to the circle. On the other hand, a similar interpretation can be made with "*Lauric*" and "*Miristic*" variables, indicating a high correlation between these two variables, which are together negatively correlated on PC1. On PC2, "*Capric*" and "*Caprilic*" variables are highly correlated. Obviously, the correlation circle should be interpreted jointly with another graph (named the score-plot) resulting from the calculation of samples coordinates in the new principal components space that we discuss below in this chapter.

3.2.3. Scores and loadings

We speak of **scores** to denote the **coordinates** of the observations on the PC components and the corresponding graphs (objects projected in successive planes defined by two principal components) are called score-plots. **Loadings** denotes the **contributions** of original variables to the various components, and corresponding graphs called **loadings-plot** can be seen as the projection of unit vectors representing the variables in the successive planes of the main components. As scores are a representation of observations in the space formed by the new axes (principal components), symmetrically, loadings represent the variables in the space of principal components.

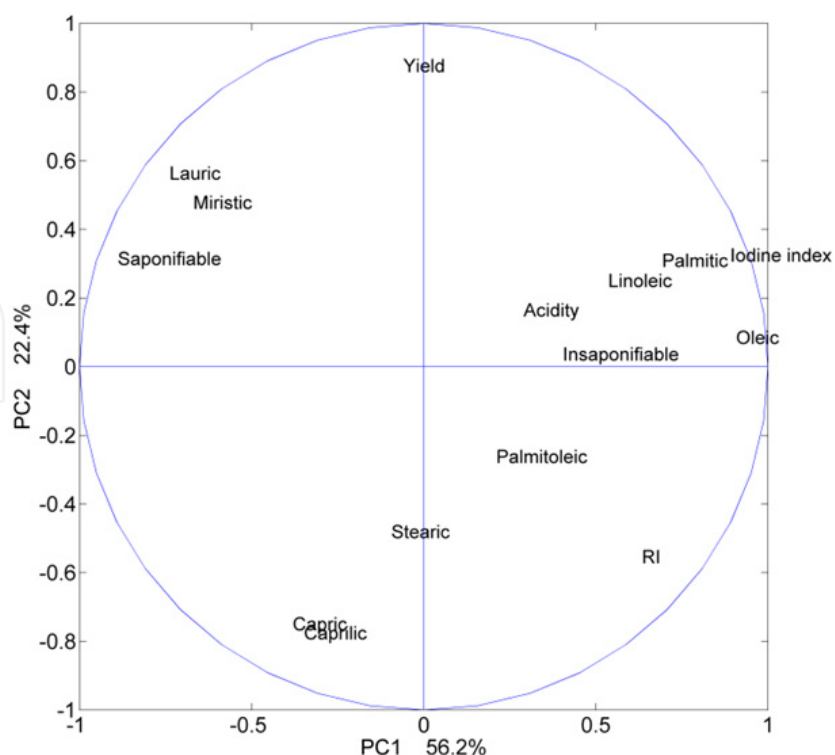


Figure 1. Example of score-plot and correlation circle obtained with PCA.

Observations close to each other in the space of principal components necessarily have similar characteristics. This proximity in the initial space leads to a close neighbouring in the score-plots. Similarly, the variables whose unit vectors are close to each other are said to be positively correlated, meaning that their influence on the positioning of objects is similar (again, these proximities are reflected in the projections of variables on loadings-plot). However, variables far away from each other will be defined as being negatively correlated.

When we speak of *loadings* it is necessary to distinguish two different cases depending on the nature of the data. When the data contains discontinuous variables, as in the case of physicochemical data, the loadings are represented as a factorial plan, i.e. PC1 vs. PC2, showing each variable in the PCs space. However, when the data is continuous (in case of spectroscopic or chromatographic data) loadings are not represented in the same way. In this case, uses usually represent the values of the loadings of each principal component in a graph with the values of the loadings of component PC_i on the Y-axis and the scale corresponding to the experimental unit on the X-axis. Thus, the loadings are like a spectrum or a chromatogram (See § B. Research example: Application of PCA on ¹H-NMR spectra to study the thermal stability of edible oils). Figure 2 and Figure 3 provide an example of scores and loadings plots extracted from a physicochemical and sensorial characterization study of Italian beef [20] through the application of principal component analysis. The goal of this work was to discriminate between the ethnic groups of animals (hypertrophied Piemontese, HP; normal Piemontese, NP; Friesian, F; crossbred hypertrophied PiemontesexFriesian, HPxF; Belgian Blue and White, BBW). These graphs are useful for determining the likely reasons of groups' formation of objects that are visualized, i.e. the

weight (or importance) of certain variables in the positioning of objects on the plane formed by two main components. Indeed, the objects placed, for example, right on the scores' plot will have important values for the variables placed right on the loadings plot, while the variables near the origin of the axes that will make a small contribution to the discrimination of objects. As presented by the authors of this work [20], one can see on the loadings plot that PC1 is characterized mainly by eating quality, one chemical parameter and two physical parameters (ease of sinking, Te; overall acceptability, Oa; initial juiciness, Ji; sustained juiciness, Js; friability, Tf; residue, Tr). These variables are located far from the origin of the first PC, to the right in the loadings plot, and close together, which means, therefore, that they are positively correlated. On the other hand, PC2 is mainly characterized by two chemical (hydroproline, Hy; and ether extract, E) and two physical (hue, H; and lightness, L) parameters. These variables located on the left of the loadings plot are positively correlated. The interpretation of the scores' plot indicates an arrangement of the samples into two groups: the first one includes meats of hypertrophied animals (HP) while the second one includes the meats of the normal Piemontese (NP) and the Friesian (F). Without repeating all of the interpretation reported by the authors, the combined reading of scores and loadings shows, for example, that the meat samples HP and BBW have, in general, a higher protein content as well as good eating qualities and lightness. At the opposite end, the meat samples F and NP are characterized more by their hydroxyproline content, their ether extract or else their Warner-Bratzler shear value. This interpretation may be made with the rest of the parameters studied and contributes to a better understanding of the influence of these parameters on the features of the product studied. This is a qualitative approach to comparing samples on the basis of a set of experimental measurements.

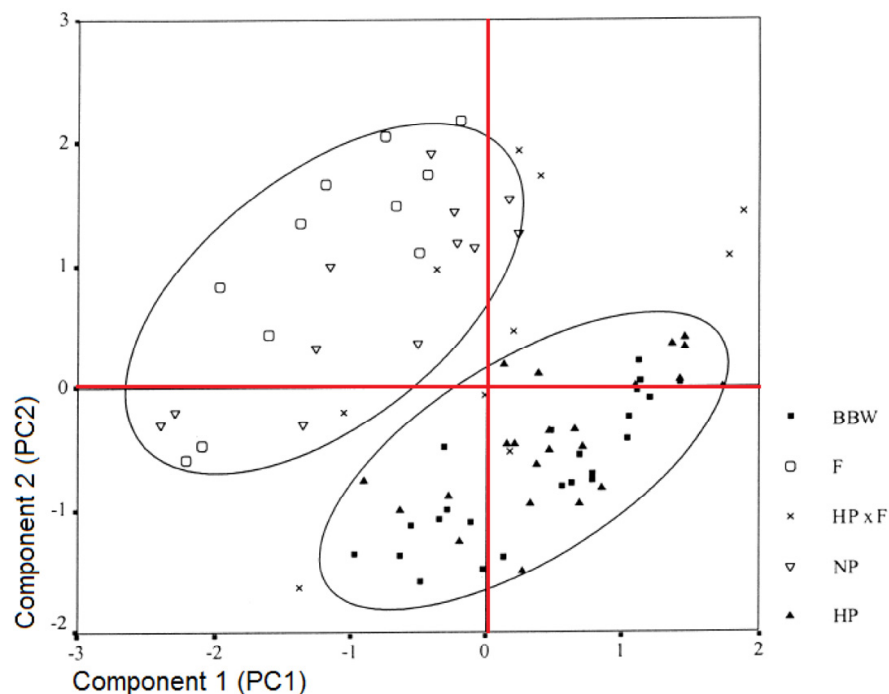
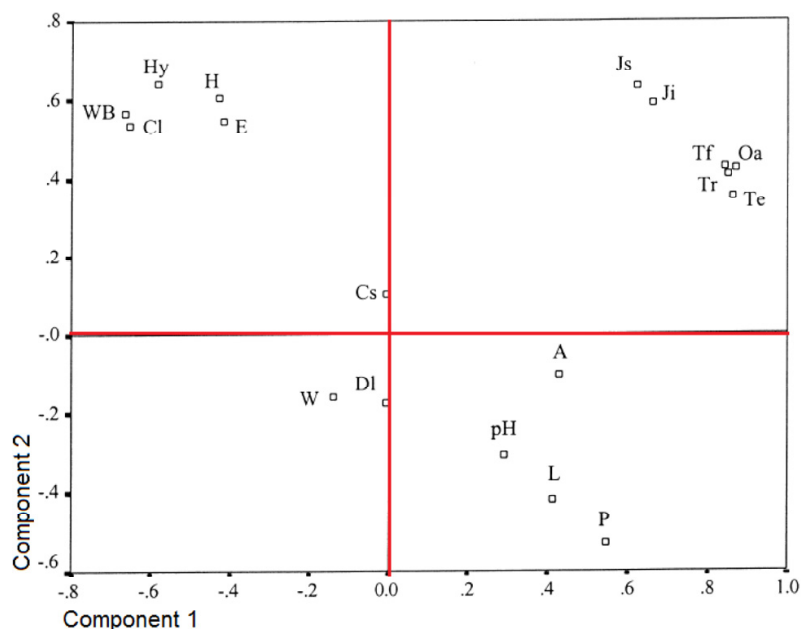


Figure 2. Score-plot obtained by PCA applied on meat samples. Extracted from [20].



Plot of the first two PC loading vectors. Water (W); protein (P); ether extract (E); hydroxyproline (Hy); collagen solubility (Cs); lightness (L); hue (H); drip losses (DI); cooking losses (Cl); Warner±Bratzler shear (WB); appearance (A); ease of sinking (Te); friability (Tf); residue (Tr); initial juiciness (Ji); sustained juiciness (Js); overall acceptability (Oa).

Figure 3. Loadings-plot obtained by PCA applied on meat samples. Extracted from [20].

3.3. Algorithms

There are different ways to achieve PCA, depending on whether one uses an iterative algorithm such as the NIPALS algorithm (Non-linear Iterative Partial Least Squares) or else a matrix factorization algorithm like SVD (Singular Value Decomposition). There are many variants of the SVD algorithm; the most well-known is probably the Golub-Reinsch algorithm [GR-SVD] [21, 22]. Both types of algorithms are well-suited for computing the eigenvalues and eigenvectors of a matrix M and determining, ultimately, the new axes of representation.

3.3.1. NIPALS

Let us start with the description of NIPALS. The most common version is given below.

M	Preferably the mean centred data matrix
$E(0) = M$	E -matrix equals to mean centred M at the beginning
t	Initialization step: this vector is set to a column in M , Scores for PC_i
p	Loadings for PC_i
$threshold = 10^{-4}$	Convergence criterion, a constant in the procedure

Beginning of the main loop (i=1 to Nb of PCs):

1. Project M onto t to calculate the corresponding loading p

$$p = \frac{E_{(i-1)}^T t}{t^T * t}$$

2. Normalise loading vector p to length 1

$$p = p * \sqrt{p^T * p}$$

3. Project M onto p to calculate the corresponding score vector t

$$t = \frac{E_{(i-1)}^T p}{p^T * p}$$

4. Check for convergence.

IF $\tau_{\text{new}} - \tau_{\text{old}} > \text{threshold} * \tau_{\text{new}}$ THEN return to step 1

With $\tau_{\text{new}} = (t^T t)$ and $\tau_{\text{old}} = (t^T t)$ from (i -1)th iteration.

5. Deflating process: remove the estimated PC component from $E_{(i-1)}$:

$$E_{(i)} = E_{(i-1)} - t p^T$$

End of the main loop

3.3.2. SVD

SVD is based on a theorem from linear algebra which says that a rectangular matrix M can be split down into the product of three new matrices:

- An orthogonal matrix U ;
- A diagonal matrix S ;
- The transpose of an orthogonal matrix V .

Usually the theorem is written as follows:

$$M_{mn} = U_{mm} S_{mn} V_{nn}^T$$

where $U^T U = I$; $V^T V = I$; the columns of U are orthonormal eigenvectors of MM^T , the columns of V are orthonormal eigenvectors of $M^T M$, and S is a diagonal matrix containing the square roots of eigenvalues from U or V in decreasing order, called singular values. The singular values are always real numbers. If the matrix M is a real matrix, then U and V are also real. The SVD represents an expression of the original data in a coordinate system where the covariance matrix is diagonal. Calculating the SVD consists of finding the

eigenvalues and eigenvectors of MM^T and M^TM . The eigenvectors of M^TM will produce the columns of V and the eigenvectors of MM^T will give the columns of U .

Presented below is the pseudo-code of the SVD algorithm:

Compute M^T , M^TM (or MM^T)*

6. Compute eigenvalues of M^TM and sort them in descending order along its diagonal by resolving:

$$|M^TM - \lambda I| = 0$$

- Characteristic equation: its resolution gives eigenvalues of M^TM .
1. Square root the eigenvalues of M^TM to obtain the singular values of M ,
 2. Build a diagonal matrix S by placing singular values in descending order along its diagonal and compute S^{-1} ,
 3. Re-use eigenvalues from step 2 in descending order and compute the eigenvectors of M^TM . Place these eigenvectors along the columns of V and compute its transpose V^T .

Compute U as $U = MVS^{-1}$ (**) and compute the true scores T as $T = US$.

3.3.3. NIPALS versus SVD

The relationship between the matrices obtained by NIPALS and SVD is given by the following:

$$TP = USV$$

With the orthonormal US product corresponding to the scores matrix T , and V to the loadings matrix P . Note that with S being a diagonal matrix, its dimensions are the same as those of M .

4. Practical examples

4.1. Pedagogical example: how to make PCA step-by-step

(See Matlab code in appendix)

The data in Table 2 consists of the fluorescence intensities at four different wavelengths for 10 hypothetical samples 1-10.

The data processing presented here was performed with Matlab v2007b. Like all data processing software, Matlab has a number of statistical tools to perform PCA in one mouse click or a one-step command-process, but we choose to give details of the calculation and bring up the steps leading to the representation of the factorial coordinates (scores) and factor contributions (loadings). Then, we interpret the results.

* Demonstrations can be found elsewhere showing that eigenvalues of M^TM and MM^T are the same. The reason is that M^TM and MM^T respond to the same characteristic equation.

** We know that $M = USV^T$. Therefore, to find U knowing S , just post-multiply by S^{-1} to obtain $AVS^{-1} = USS^{-1}$. Hence $U = AVS^{-1}$, because $SS^{-1} = I = 1$.

Sample #	Wavelengths (nm)			
	300	350	400	450
1	15,4	11,4	6,1	3,6
2	16,0	12,1	6,1	3,4
3	14,0	13,1	7,3	4,2
4	16,7	10,4	5,9	3,1
5	17,1	12,1	7,1	2,9
6	81,9	71,3	42,8	32,9
7	94,9	85,0	50,0	39,0
8	69,9	75,7	46,3	50,4
9	70,4	70,5	42,6	42,1
10	61,8	81,9	50,2	66,2
Mean	15.75	61.25	68.92	29.25

Table 2. Fluorescence intensities for four wavelengths measured on 12 samples.

Step 1. Centring the data matrix by the average

This step is to subtract the intensity values of each column, the average of the said column. In other words, for each wavelength is the mean of all samples for this wavelength and subtract this value from the fluorescence intensity for each sample for the same wavelength.

Step 2. Calculation of the variance-covariance matrix

The variance-covariance matrix (Table 3) is calculated according to the $X^T X$ product, namely the Gram matrix. The diagonal of this matrix consists of the variances; the trace of the matrix (sum of diagonal elements) corresponds to the total variance (3294.5) of the original matrix of the data. This matrix shows, for example, that the covariance for the fluorescence intensities at 420 and 520 nm is equal to 665.9.

λ , nm	420	474	520	570
420	1072,2	1092,9	665,9	654,5
474	1092,9	1194,2	731,5	786,6
520	665,9	731,5	448,4	485,5
570	654,5	786,6	485,5	579,8

Table 3. Variance-covariance matrix

As mentioned above, the matrix also contains the variances of the fluorescence intensities at each wavelength on the diagonal. For example, for the fluorescence intensities at 474 nm, the variance is 1194.2.

The technique for calculating the eigenvectors of the variance-covariance matrix, and so the principal components, is called eigenvalue analysis (*eigenanalysis*).

Step 3. Calculate the eigenvalues and eigenvectors

Table 4 shows the eigenvalues and eigenvectors obtained by the diagonalization of the variance-covariance matrix (a process that will not be presented here, but some details on this mathematical process will be found elsewhere [23]).

Note that the sum of the eigenvalues is equal to the sum of the variances in the variance-covariance matrix, which is not surprising since the eigenvalues are calculated from the variance-covariance matrix.

Eigenvalues			
3163,3645	0,0000	0,0000	0,0000
0,0000	130,5223	0,0000	0,0000
0,0000	0,0000	0,5404	0,0000
0,0000	0,0000	0,0000	0,1007
Eigenvectors			
0,5662	0,6661	0,4719	0,1143
0,6141	-0,0795	-0,7072	0,3411
0,3760	-0,0875	-0,1056	-0,9164
0,4011	-0,7365	0,5157	0,1755

Table 4. Eigenvalues calculated using diagonalization processing of the variance-covariance matrix.

The next step of the calculation is to determine what percentage of the variance is explained by each major component. This is done by using the fact that the sum of the eigenvalues corresponding to 100% of the explained variance, as follows:

$$\%Var_{j^{th} PC} = \frac{\lambda_j}{\sum_{i=1}^l \lambda_i} \times 100\%$$

Where λ_j is the j^{th} eigenvalue. We thus obtain for the first component $PC1 = 3163.4 * 1 / 3294.5 * 100 = 96.01\%$. PC2 is associated with 3.96% and so on, as shown in Table 5 below.

PC1	PC2	PC3	PC4
96.02	3.96	0.02	0.00

Table 5. Percentage of the explained variance for each principal component.

Step 4. Calculating factorial coordinates - Scores

The eigenvectors calculated above are the principal components and the values given in Table 6 are the coefficients of each principal component. Thus, component #1 is written: $PC1 = 0.566 * X1 + 0.614 * X2 + 0.376 * X3 + 0.401 * X4$ where $X1$, $X2$, $X3$ and $X4$ are the fluorescence intensities, 420, 474, 520 and 570 nm respectively. The factorial coordinates of each sample in the new space formed by the principal components can now be calculated directly from the equations of the PCs. For example, sample No. 1 has a coordinate on PC1 equal to:

$0.566 \cdot 15.4 + 0.614 \cdot 11.4 + 0.376 \cdot 6.1 + 0.401 \cdot 3.6 = 19.453$. Table 6 presents the factorial coordinates of all the samples of the initial matrix.

Centred data				
	PC1	PC2	PC3	PC4
1	19,455	6,128	0,377	0,678
2	20,121	6,671	0,072	0,983
3	20,420	4,528	-1,256	0,156
4	19,357	7,534	1,494	0,575
5	20,928	7,635	0,259	0,051
6	119,413	20,924	0,667	0,179
7	140,394	23,343	-0,522	0,843
8	123,676	-0,601	0,537	0,228
9	116,052	6,511	0,590	0,397
10	130,754	-18,524	0,056	0,632

Table 6. Factorial coordinates = *scores* of the initial sample matrix in the principal components space.

Therefore, we can now represent, for example, all samples in the space of two first components, as shown in Figure 4 below.

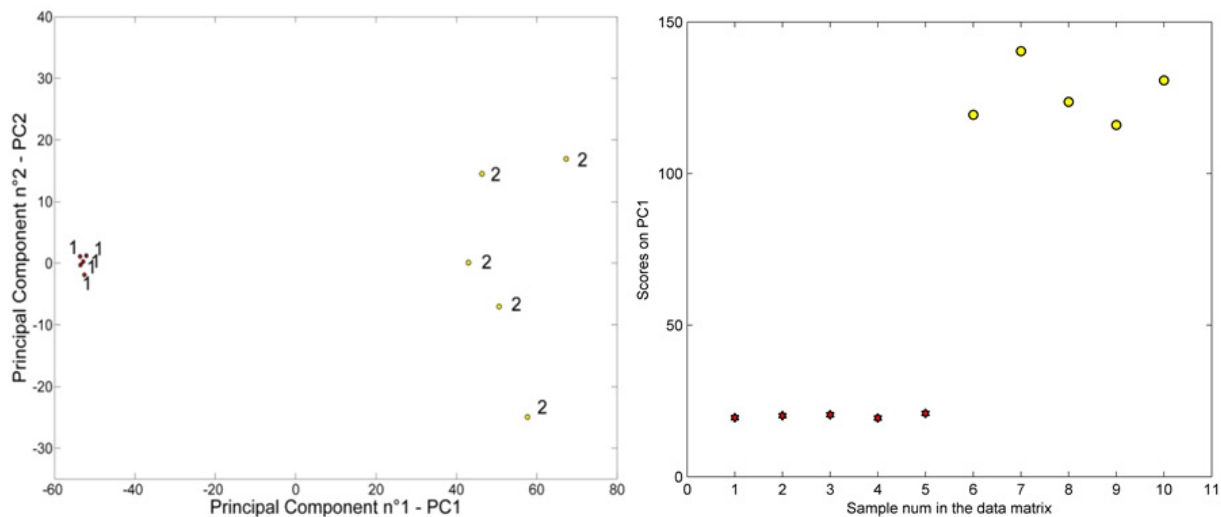


Figure 4. On the left hand side: scores-plot from samples in two first principal components space; on the right hand side: samples-scores on PC1.

It was found that the samples are divided into two distinct groups on the initial data. Another way to visualize this division is to represent only the scores on PC1 (on PC2 or PC_i) for each sample (see Figure 4). In this case, it is also obvious that the samples 1 to 5 have similar values of PC1 and, thus, form a first group (group 1), while samples 6-10 are the second group (group 2).

Step 5. Calculating factorial contributions - Loadings

A complete interpretation of the results of PCA involves the graph of the loadings, i.e. the projection of the variables in the sample space. But how does one get this? Consider what has been calculated so far: the eigenvalues and eigenvectors of the matrix samples and their factorial coordinates in the space of principal components. We need to calculate the factorial coordinates of variables in the sample space. The results are called “loadings” or “factorial contributions”. You just have to transpose the initial data matrix and repeat the entire calculation. Figure 5 shows the results representing the position of the variables of the problem in the plane formed by the first two PCs.

The graph of the loadings allows us to understand what the characteristic variables of each group of samples on the graph of the scores are. We see, in particular, that the samples of Group 1 are distinguished from the others by variables 420 and 474 while the Group 2 samples are distinguished from the others by variables 520 and 570. In other words, for these two sets of variables, the groups of samples have opposite values in quantitative terms: when a group has high values for the two pairs of variables, then the other group of samples has low values for the same pair of variables, and vice versa.

When the measured variables are of structural (mass spectrometry data) or spectral types (infrared bands or chemical shifts in NMR), the joint interpretation of scores and loadings can be extremely interesting because one can be in a position to know what distinguishes the groups of samples from a molecular point of view.

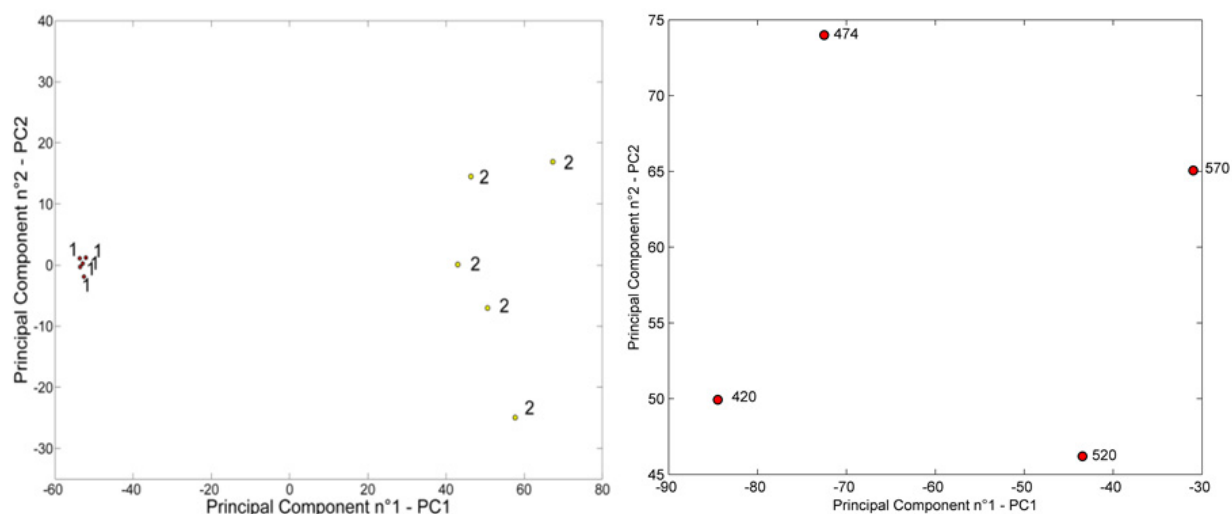


Figure 5. Scores & loadings on the PC1xPC2 plane.

4.2. Research example: Application of PCA on ^1H -NMR spectra for the study of the thermal stability of edible oils

Numerous chemical reactions contribute to the chemical and physical aging of oils. During heating, the oils undergo degradation and their functional and organoleptic features are significantly modified. The heating induces chemical reactions such as oxidation, polymerization, hydrolysis and cis/trans isomerization, which have an impact not only on the

nutritional value of oils but which may also generate toxic compounds injurious to health [24, 25]. The most important cause of the deterioration of oils is oxidation. Among the oxidation products, our interest has focused especially on secondary oxidation products, such as aldehydes, because they are rarely present in natural unheated oil, as described by Choe et al. [26] from the study of secondary oxidation products proposed by Frankel [27]. In this work, an analytical approach was first adopted to calculate a new semi-quantitative criterion of the thermal stability of oils. This new test is based on the assumption that by focusing on a selected portion of the ^1H -NMR spectra and for a relatively short time, we can model the appearance of aldehydes by a kinetic law of order 1, knowing that the mechanisms actually at work are more complex and associated with radical reactions. In the following pages is presented only that part of this work related to the application of PCA to ^1H -NMR data to characterize and follow the effect of temperature and time of heating on the chemical quality of edible oils. For further information about the kinetic study and multiway treatments see [28].

4.2.1. Samples

Three types of edible oils were analysed. Rapeseed, sunflower and virgin olive oils were purchased at a local supermarket and used in a thermal oxidation study. Approximately 12 mL of oil was placed in 10 cm diameter glass dishes and subjected to heating in a laboratory oven with temperature control. Each of the three types of oil was heated at 170 °C, 190 °C and 210 °C, each of which is close to home-cooking temperatures. Three samples of 1 g were collected every 30 min until the end of the heating process, fixed at 180 min, resulting in a total of 189 samples to be analysed. Samples were cooled in an ice-water bath for 4 min, in order to stop thermal-oxidative reactions, and then directly analysed.

4.2.2. ^1H -NMR spectroscopy

Between 0.3 and 0.5 g of oil was introduced into an NMR tube (I.D. 5 mm) with 700 μL of deuterated chloroform for the sample to reach a filling height of approximately 5 cm. The proton NMR spectrum was acquired at 300.13 MHz on a Bruker 300 Advance Ultrashield spectrometer with a 7.05 T magnetic field. A basic spin echo sequence was applied. The acquisition parameters were: spectral width 6172.8 Hz; pulse angle 90–180°; pulse delay 4.4 μs ; relaxation delay 3 s; number of scans 64; plus 2 dummy scans, acquisition time 5.308 s, with a total acquisition time of about 9 min. The experiment was carried out at 25 °C. Spectra were acquired periodically throughout the thermal oxidation process. All plots of ^1H NMR spectra or spectral regions were plotted with a fixed value of absolute intensity for comparison.

4.2.3. Data & data processing

The initial matrix (189 samples \times 1001 variables) contains ^1H -NMR spectra of three oils at three heating temperatures and seven heating times. The computations were performed using the MATLAB environment, version R2007b (Mathworks, Natick, MA, USA).

4.2.4. Results and discussion

The chemical changes taking place in the oils over time and at different temperatures, were monitored by ^1H nuclear magnetic resonance spectroscopy. The spectral region between δ 0 and δ 7.2 was not used in this work because the changes observed were not exclusively specific to heating by-products, as were those in the spectral region between δ 9 and δ 10. Figs. 6.1–6.3 present this spectral region for rapeseed oil for different heating times at 170 °C, 190 °C and 210 °C. The increase in the aldehydic protons peaks over time and as a function of temperature is clearly visible. It is to be noted that the temperature has a similar effect in this spectral region for sunflower oil, but with a different patterns of peaks. This reflects differences in the structure of the aldehydes formed in the two oils. The effect of temperature is much smaller for olive oil due to the composition of the fatty acids and a higher concentration of natural antioxidants such as tocopherols.

PCA score plots show classical trajectories with time and temperature of the three oils, and 66% of the total variance is explained by the first 2 components (see Fig. 7). The score plots show that the thermal degradations are quite different, with that of olive oil being much less extensive, with that of rapeseed being more similar to olive oil than sunflower oil. This observation makes sense because olive oil and rapeseed oil have approximately the same range of oleic acid (respectively 55.0–83.0% and 52.0–67.0% [29]), which is one of the three major unsaturated fatty acids of edible oils, along with linoleic and linolenic acids, while sunflower oil has a lower content (14.0–38.0%). The smaller spread of scores for olive oil at a given temperature indicates that chemical changes caused by heating over time are more pronounced in rapeseed and sunflower oil than in olive oil. Olive oil is also different from the other two oils in that the orientation of its scores' trajectory is different for each temperature. The plot of the PC1 and PC2 loadings suggests that the chemical evolution of the sunflower oil and olive oil is not characterized by the formation of the same aldehyde protons. The scores for PC1 characterize the time factor for sunflower, rapeseed and olive oils at 170 °C, while the loadings for PC1 characterize the progressive appearance of trans-2-alkenals and E,E-2,4-dialkadienals associated with peaks at δ 9.534, 9.507 and 9.554 (see Fig. 7) and the short chain alkanals associated with the peak at δ 9.766.

The scores for PC2 are related more to the time factor for rapeseed oil and olive oil at 190 and 210 °C. The loadings for PC2 are characterized by several peaks: δ 9.497 and 9.506 (possibly related to the increase of trans-2-alkenals), δ 9.763 and δ 9.786 (N-alkanals and short chain alkanals or oxo-alkanals) and δ 9.553 (possibly 4,5-epoxy trans-2-alkenals). According to Guillén et al. [30, 31], these compounds are among the secondary oxidation products. The loadings of each PC give a clear idea of which peaks increase or decrease during the heating and allow a better understanding of what differentiates the oil samples from a chemical point of view. The Euclidean distance in the multivariate space of three first principal components (i.e. those that explain the majority of the variance of the data studied (70%)) allows the proposal of a thermal stability order of studied edible oils. This order is presented in table 7.

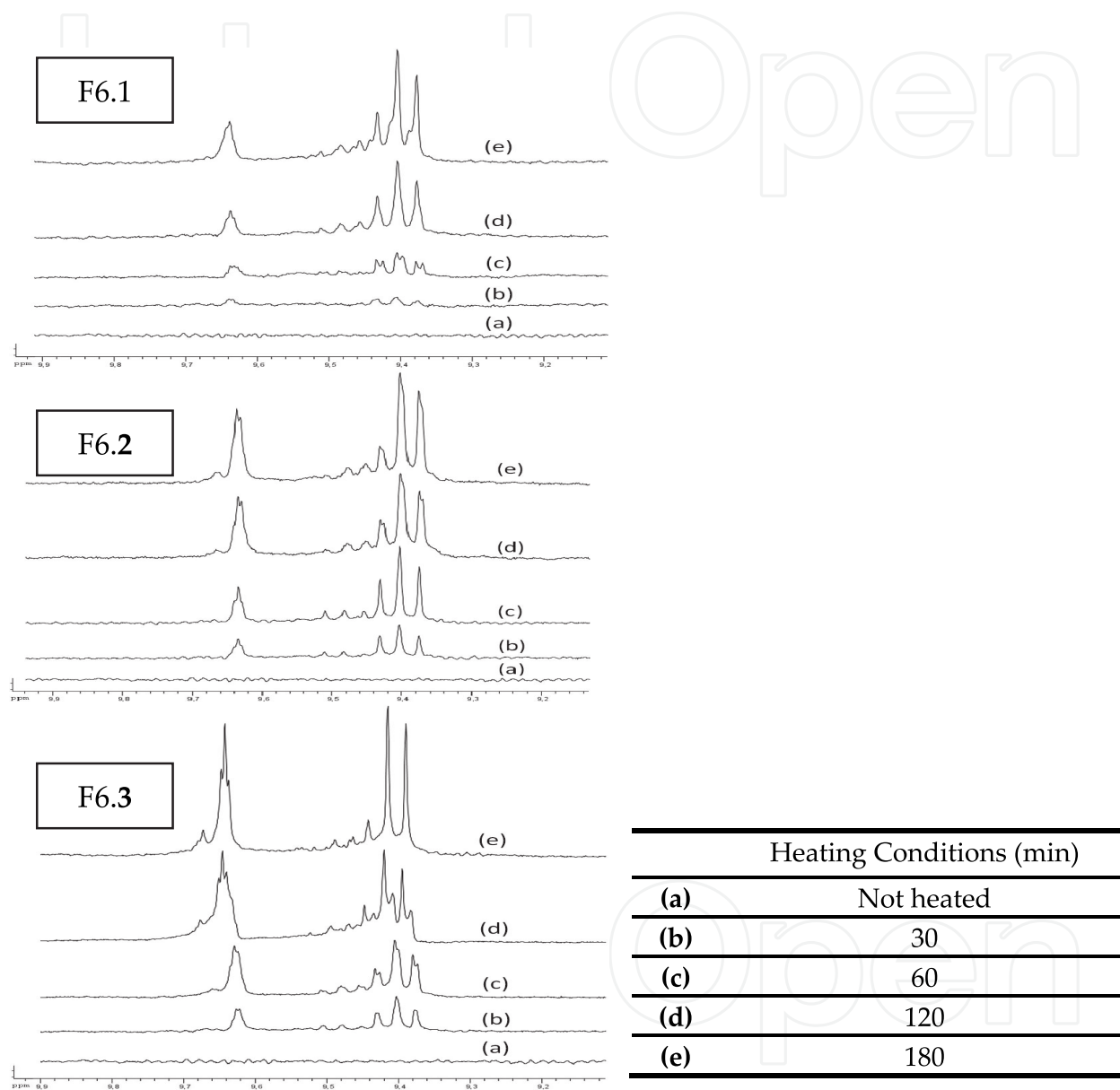


Figure 6. Expanded region between δ 9 and δ 10 of the ^1H NMR spectra of rapeseed oil at different times during heating (F5.1) at 170°C , (F5.2) at 190°C , (F5.3) at 210°C .

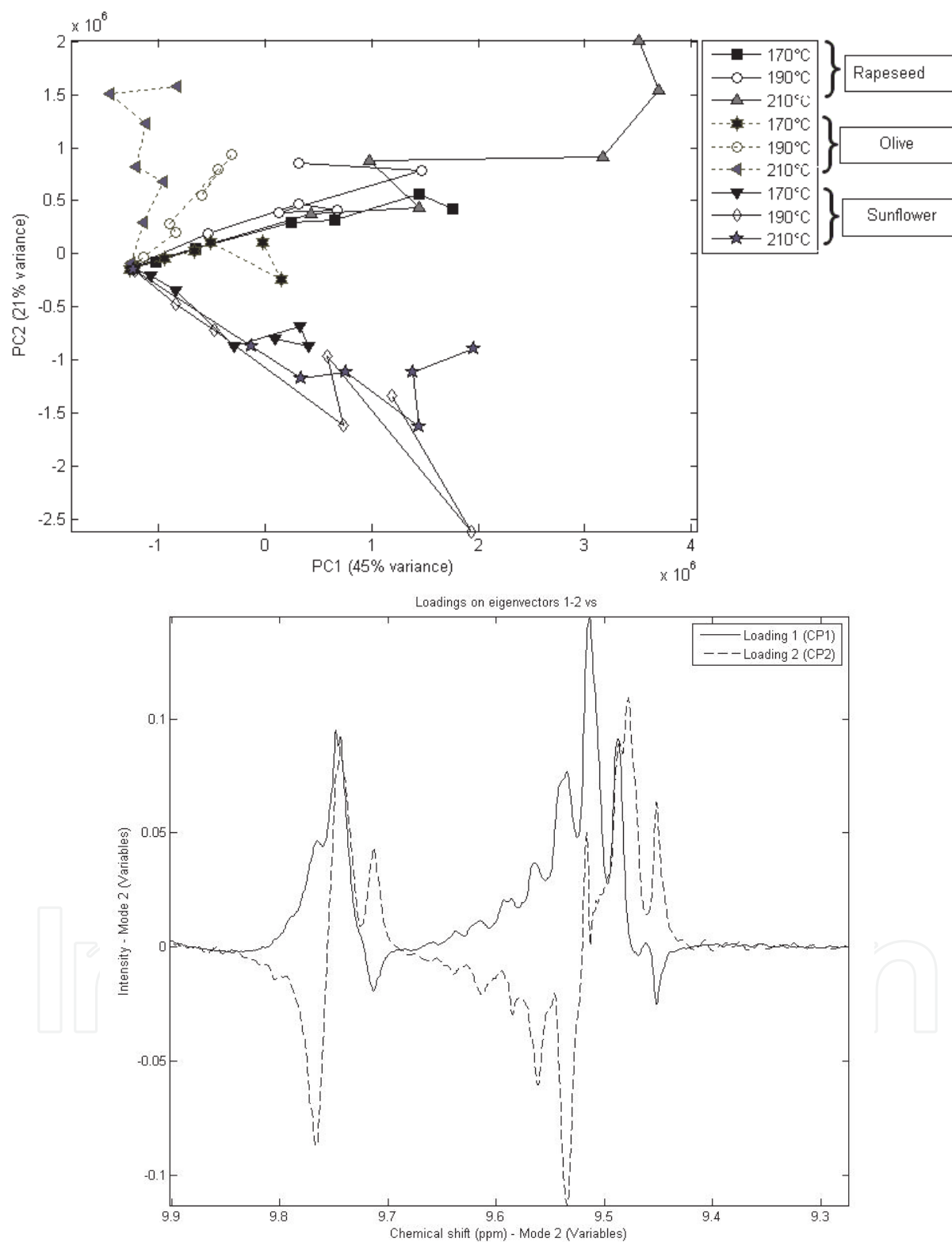


Figure 7. PCA scores (up) and loadings (bottom) of (63 × 1001) data matrix for 3 types of oils and 3 heating temperatures. In both cases, only PC1 and PC2 are represented.

Observed thermal stability order		
Temperature		
190 °C	170 °C	210 °C
OO > SO >= RO	OO > RO >= SO	OO > SO >> RO

Table 7. Thermal stability order of the studied oils given by multivariate distances calculated from coordinates on 3 first PCs. OO: olive oil, RO: rapeseed oil, SO: sunflower oil.

4.2.5. Conclusion

The conclusion of this work indicates that olive oil is more stable than the others (rapeseed and sunflower oil), whatever the heating temperature applied. One reason for this stability is its higher concentration of natural antioxidants, such as tocopherols, and its content of oleic acid with unsaturated bonds, which allow it to participate more easily in the radical oxidation mechanisms which occur during heating and thus protect other fragile compounds in the oil for longer.

This application example of the study of the stability of heated oils shows the power of such tool as PCA in the treatment of multivariate spectroscopic data as used with the characteristic fingerprints of chemical and physical phenomena occurring in the samples over time. The PCA allows the direct use of spectra (or pieces of spectra) or analytical curves instead of the integration values typically used in NMR spectroscopy. The principle presented here is now very common in the field of spectroscopists in general and many monitoring tools or processes for monitoring chemical reactions involve PCA or more robust variants (Robust PCA, RPCA based on robust covariance estimators, RPBA based on projection pursuit, Kernel PCA, etc.).

5. PARAFAC for parallel factor analysis, a generalization of PCA to 3-way data

5.1. Introduction

Usually, physicochemical analysis - or else the more generally parametric monitoring of a number of physicochemical properties of a set of samples - resulted in the construction of a data matrix with samples in rows and physicochemical properties in columns. This data matrix is a mathematical representation of the characteristics of the sample set at time t in preparatory and analytical conditions attached to the instant action. Incidentally, the fact of working on a matrix which is a two-dimensional array allows us to speak of two modes or 2-way data. Now, imagine that you reproduce these measurements on several dates t_1, t_2, \dots, t_n . You no longer have a matrix $X(n, p)$ but the N matrices $X_i(n, p)$ of the same size where i is the number of the matrix corresponding to the time t_i . This is known as 3-mode data or three-way data or a "data cube". Figure 8 below illustrates what has been said.

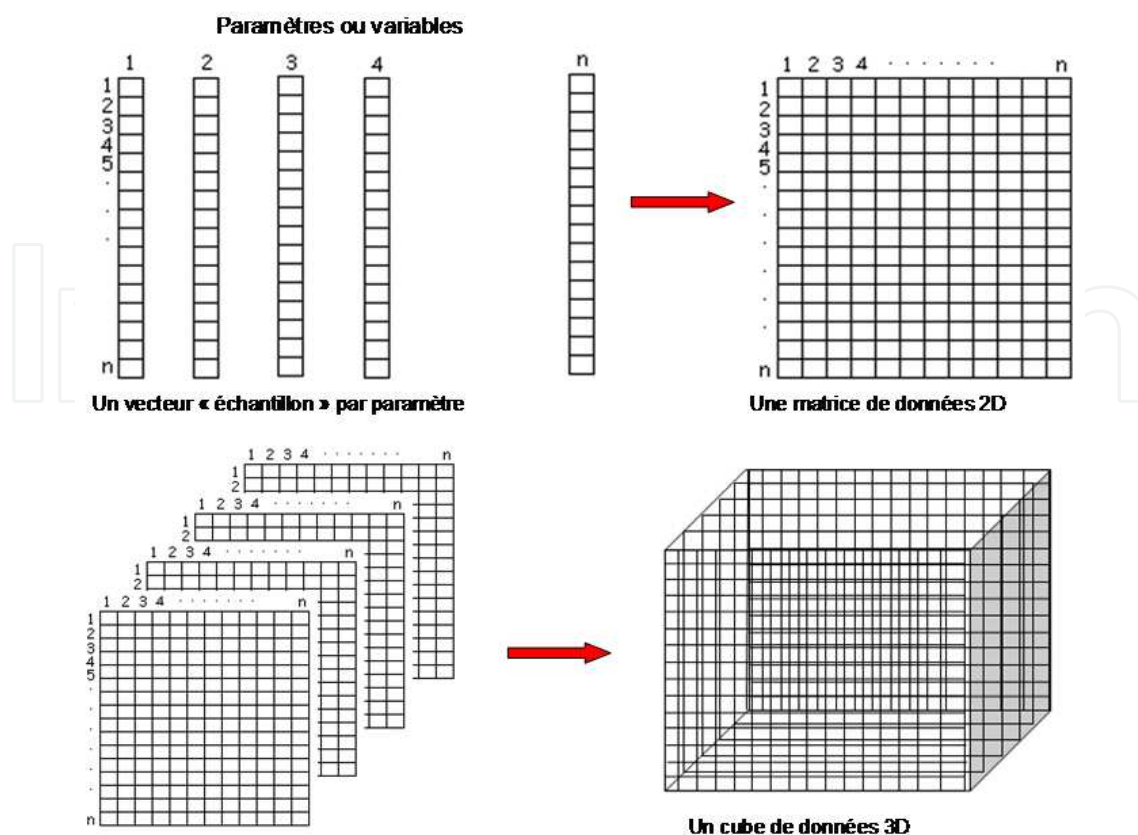


Figure 8. Representation of a matrix and a cube of data from experimental measurements.

At this stage, a key question is raised: how should one analyse all these matrices so as to extract all the relevant information that takes into account the time factor? Several options are available to us. The first would be, for example, an averaging of each parameter per sample measured at the various dates/times. This has the immediate effect of losing the time-based information. By this operation, the effect of time on the measured parameters is no longer observable. We would obtain an average matrix resulting from all of the measured parameters of the study period. The advantage of this in relation to the question is null. Finally, we could consider using the chemometric techniques discussed previously in this book, such as principal component analysis (or others not presented here, such as the hierarchical analysis of each class of matrices stored). But then, it becomes difficult to visualize changes occurring between successive matrices which may correspond to an evolution with time (as in the case of 3D fluorescence spectra which are commonly qualified as second-order data) or to a geographical evolution of the measured parameters (e.g., when studying physicochemical a river port or lagoon in which samples are realized in various places to monitor the level of chemical pollution [32]). The overall interpretation is certainly more difficult.

The solution must be sought to tools capable of taking into account a third or N^{th} dimension(s) in the data while retaining the ability to account for interactions between factors. The following pages describe two of these tools; probably one of the most famous of them is the PARAllel FACtor model (PARAFAC) introduced independently by Harshman

in 1970 [33] and by Carroll & Chang [34] who named the model CANDECOMP (canonical decomposition). The oldest paper we found relating the mathematical idea of PARAFAC was probably published by Hitchcock F.L. in 1927 who presents a method to decompose tensors or polyadics as a sum of products [35]. In this class of tools, the size or dimensions of the dataset are called "modes", hence the name of multi-mode techniques. The most widely used encountered term is "multiway" technique.

5.2. The PARAFAC model

As discussed above, many problems involve chemical 3-ways data tables. Let us come back to the hypothetical example of a liquid chromatography coupled to a fluorescence spectrometer that produces a data set consisting of 3-way "layers" or 2-dimensional matrix sheets coming from the excitation-emission fluorescence spectra collection (Excitation-Emission Matrix: EEM) which vary as a function of elution time.

The fluorescence intensities depend on the excitation and emission wavelengths, and the elution time, these variables will be used to represent the three modes. To exploit these multiway data, the PARAFAC model is an effective possibility to extract useful information. PARAFAC is an important and useful tool for qualitative and quantitative analysis of a set of samples characterized by bilinear data such as EEM in fluorescence spectroscopy. One of the most spectacular examples of the capability of PARAFAC model is the analysis of a mixture of several pure chemical components characterized by bilinear responses. In this case, PARAFAC is able to identify the right pairs of profiles (e.g. emission and excitation profiles) of pure components as well as the right concentration proportions of the mixture. PARAFAC yields unique component solutions. The algorithm is based on a minimizing least squares method. This is to decompose the initial data table in a procedure known as "trilinear decomposition" which gives a unique solution. The trilinear decomposition comes from the model structure and sometimes data itself implies that because of its (their) natural decomposition in 3 modes. The PARAFAC model is a generalization of the PCA itself bilinear, to arrays of higher order (i.e., three or more dimensions). PCA decomposes the data into a two-mode product of a matrix called matrix T scores and a matrix of loadings P describing systematic variations in the data, over a matrix of residues E representing the discrepancies between actual data and the model obtained. Thus, as illustrated by figure 8, the EEM are arranged in a 3-way table (X) of size $I \times K \times J$ where I is the number of samples, J the number of emission wavelengths, and K the number of excitation wavelengths. Similarly, PARAFAC decomposes X into three matrices (see figure 9): A (scores), B and C (loadings) with the elements a_{it} , b_{jt} , and c_{kf} . In other words, N sets of triads are produced and the trilinear model is usually presented as followed in equation 1 [36]:

$$x_{ijk} = \sum_{f=1}^F a_{if} b_{jf} c_{kf} + e_{ijk} \quad (3)$$

($i = 1 \dots I; j = 1 \dots J; k = 1 \dots K$)

where x_{ijk} is the fluorescence intensity for the i^{th} sample at the emission wavelength j and the excitation wavelength k . The number of columns f in the matrices of loadings is the number of PARAFAC factors and e_{ijk} residues, which account for the variability not represented by the model.

Note the similarities between the PARAFAC model and that of the PCA in schema 2, § “II.A Some theoretical aspects”. The PARAFAC model is a specific case of the Tucker3 model introduced by Tucker in 1966. The following paragraph presents the essentials about Tucker3 model and proposes some important papers on the theory and applications of this multiway tool. The subsequent paragraph gives a more complete bibliography related to PARAFAC and Tucker3 models on various application areas.

5.3. Tucker3: a generalization of PCA and PARAFAC to higher order

Conceptually, the Tucker3 model [37] is a generalization of two-way data decomposition methods such as PCA or singular value decomposition (SVD) to higher order arrays or tensors [8] and [9]. In such multiway methods, scores and loadings are not distinguishable and are commonly treated as numerically equivalent. Being a generalization of principal component analysis and PARAFAC to multiway data arrays, the Tucker3 model has for its objective to represent the measured data as a linear combination of a small number of optimal, orthogonal factors. For a 3-way data array, the Tucker3 model takes the following form:

$$x_{ijk} = \sum_{u=1}^r \sum_{v=1}^s \sum_{w=1}^t g_{iu} h_{jv} e_{kw} c_{uvw} + \varepsilon_{ijk}$$

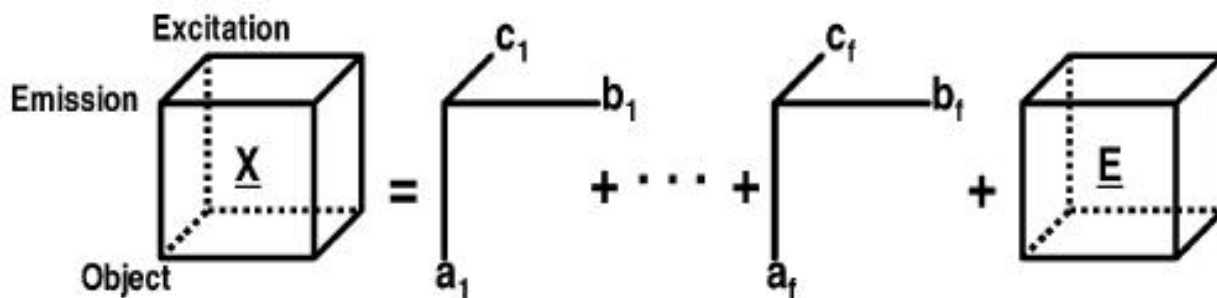


























Figure 9. Principle of the decomposition of a 3-way data cube according to the PARAFAC model.

here, x_{ijk} are the measured data, g_{iu} , h_{jv} and e_{kw} are the elements of the loading matrices for each the three ways (with r , s and t factors, respectively) and c_{uvw} are the elements of the core array (of size $r \times s \times t$), while ε_{ijk} are the elements of the array of the residuals. A tutorial on chemical applications of Tucker3 was proposed by Henrion [33], while Kroonenberg [34] gives a detailed mathematical description of the model and discusses advanced issues such as data preparation/scaling and core rotation. For a complete and very pedagogical comparison of Tucker3 with PARAFAC, another multiway procedure, see Jiang [35]. Nevertheless, some aspects of Tucker3 model which distinguish it from PARAFAC have to be discussed here. The first, the Tucker3 model does not impose the extraction of the same number of factors for each mode. Second, the existence of a core array, C , governing the interactions between factors allows the modelling of two or more factors that might have the same chromatographic profile but different spectral and/or concentration profiles. Third, the presence of the core cube in the Tucker3 model gets it to appear as a non linear model which is not always appropriate for problems having trilinear structure. But this limitation can be overcome by applying constraints to the core cube C . In some cases, with constraints applied to C leadings to have only nonzero elements on the superdiagonal of the cube and a number of factors equal on each mode, then the resulting solution is equivalent to the PARAFAC model [38].

5.4. Applications: Brief review

Although PARAFAC and Tucker3 as factorial decomposition techniques come from the last century, their routine use in analytical chemistry became popular with the Rasmus Bro's thesis in 1998. Of particular note is the remarkable PhD works published and available on the Department of Food Science website of the Faculty of Life Sciences, University of Copenhagen using the PARAFAC model and many other multivariate and multiway methods in many industrial food sectors⁶. Before the Bro's work and more generally of the Danish team, one can list a number of publications on the application of PARAFAC or Tucker3 in various sciences, but the literature does not appear to show significant, focused and systematic production of this type of model in the world of chemistry. Since the 2000s, multiway techniques have become widespread with strength in analytical chemistry in fields as diverse as food science and food safety, environment, sensory analysis and process chemistry. This explosion of applications in the academic and industrial sectors is linked to the popularization of analytical instruments directly producing multiway data and particularly fluorescence spectrometers for the acquisition matrix of fluorescence which are naturally two-modes data, inherently respecting the notion of tri-linearity and therefore suitable for processing by models such as PARAFAC or Tucker3. Applications of three-way techniques are now too numerous to be cited in their entirety. Therefore, some of the more interesting applications are listed in table 8 below and the reader is encouraged to report himself to reviews included in this table.

⁶ Department of Food Science, Faculty of Life Sciences, University of Copenhagen, <http://www.models.kvl.dk/theses>, last visit April, 2012.

Subject	Type	Technique(s)	Year (Ref.)
Foundations of the PARAFAC procedure: Models and conditions		---	1970 [33]
PARAFAC. Tutorial and Applications		General	1997 [39]
Quantification of major metabolites of acetylsalicylic acid		Fluorescence	1998 [40]
Fluorescence of Raw Cane Sugars		Fluorescence	2000 [41]
Determination of chlorophylls and pheopigments		Fluorescence	2001 [42]
Practical aspects of PARAFAC modelling of fluorescence excitation-emission data		Fluorescence	2003 [43]
Evaluation of Two-Dimensional Maps in Proteomics		2D-PAGE imaging	2003 [44]
Evaluation of Processed Cheese During Storage		Front face fluorescence	2003 [45]
Quantification of sulphathiazoles in honeys		Fluorescence	2004 [46]
Evaluation of Light-Induced Oxidation in Cheese		Front face fluorescence	2005 [47]
Olive Oil Characterization		Front face fluorescence	2005 [48]
Detection of Active Photosensitizers in Butter (riboflavine, protoporphyrine, hématoporphyrine, chlorophylle a)		Fluorescence + sensory analysis	2006 [49]
Characterizing the pollution produced by an industrial area in the Lagoon of Venice		Physicochemical analysis	2006 [32]
Calibration of folic acid and methotrexate in human serum samples		Fluorescence	2007 [50]
Quantification of sulphaguanidines in honeys		Fluorescence	2007 [51]
Water distribution in smoked salmon		RMN	2007 [52]
Monitoring of the photodegradation process of polycyclic aromatic hydrocarbons		Fluorescence + HPLC	2007 [53]
Quantification of tetracycline in the presence of quenching matrix effect		Fluorescence	2008 [54]
Fluorescence Spectroscopy: A Rapid Tool for Analyzing Dairy Products		Fluorescence	2008 [55]
Determination of aflatoxin B1 in wheat		Fluorescence	2008 [56]
Study of pesto sauce appearance and of its relation to pigment concentration		Sensory analysis + HPLC-UV	2008 [57]
Determination of vinegar acidity		ATR- IR	2008 [58]
Multi-way models for sensory profiling data		Sensory analysis	2008 [59]
Noodles sensory data analysis		Physicochemical and sensory analysis	2011 [60]
Kinetic study for evaluating the thermal stability of edible oils		¹ H-NMR	2012 [28]




 Books and thesis;
 Scientific papers;
 Reviews

Table 8. Bibliography related to PARAFAC and/or Tucker3 models. Theory and applications in areas such as chemicals and food science, medicine and process chemistry.

5.5. A research example: combined utilization of PCA and PARAFAC on 3D fluorescence spectra to study botanical origin of honey

5.5.1. Application of PCA

The interest for the use of chemometric methods to process chromatograms in order to achieve a better discrimination between authentic and adulterated honeys by linear discriminant analysis was demonstrated by our group previously [61]. An extent of this work was to quantify adulteration levels by partial least squares analysis [62]. This approach was investigated using honey samples adulterated from 10 to 40% with various industrial bee-feeding sugar syrups. Good results were obtained in the characterization of authentic and adulterated samples (96.5% of good classification) using linear discriminant analysis followed by a canonical analysis. This procedure works well but the data acquisition is a bit so long because of chromatographic time scale. A new way for honey analysis was recently investigated with interest: Front-Face Fluorescence Spectroscopy (FFFS). The autofluorescence (intrinsic fluorescence) of the intact biological samples is widely used in biological sciences due to its high sensitivity and specificity. Such an approach increases the speed of analysis considerably and facilitates non-destructive analyses. The non-destructive mode of analysis is of fundamental scientific importance, because it extends the exploratory capabilities to the measurements, allowing for more complex relationships such as the effects of the sample matrix to be assessed or the chemical equilibriums occurring in natural matrices. For a recent and complete review on the use of fluorescence spectroscopy applied on intact food systems see [63, 64]. Concerning honey area, FFFS was directly applied on honey samples for the authentication of 11 unifloral and polyfloral honey types [65] previously classified using traditional methods such as chemical, pollen, and sensory analysis. Although the proposed method requires significant work to confirm the establishment of chemometric model, the conclusions drawn by the authors are positive about the use of FFFS as a means of characterization of botanical origin of honeys samples. At our best knowledge, the previous mentioned paper is the first work having investigated the potential of 2D-front face fluorescence spectroscopy to determine the botanical origins of honey at specific excitation wavelengths. We complete this work by adopting a 3D approach of measurements. We present here below the first characterization of three clear honey varieties (Acacia, Lavender and Chestnut) by 3D-Front Face Spectroscopy.

5.5.1.1. Samples

This work was carried out on 3 monofloral honeys (Acacia: *Robinia pseudo-acacia*, Lavandula: *Lavandula hybrida* and Chestnut: *Castanea sativa*). Honeys were obtained from French beekeepers. The botanical origin of the samples was certified by quantitative pollen analysis according to the procedure of Louveaux et al. [66]. An aliquot part of 10 g of the honey samples was stirred for 10min at low rotation speed (50-80 rpm) after slight warming (40°C, for 1h), allowing the analysis of honeys at room temperature by diminishing potential difficulties due to different crystallization states of samples. Honey samples were pipetted in 3 mL quartz cuvette and spectra were recorded at 20 °C.

5.5.1.2. 3D-Fluorescence spectroscopy

Fluorescence spectra were recorded using a FluoroMax-2 spectrofluorimeter (Spex-Jobin Yvon, Longjumeau, France) mounted with a variable angle front-surface accessory. The incidence angle of the excitation radiation was set at 56° to ensure that reflected light, scattered radiation and depolarisation phenomena were minimised.

The fluorescence excitation spectra were recorded from 280 nm to 550 nm (increment 4 nm; slits: 3 nm, both at excitation and emission), the fluorescence emission spectra were recorded from 280 to 600 nm, respectively. For each sample, three spectra were recorded using different aliquots.

5.5.1.3. Data Processing and Statistical Analysis

The computations were performed using the MATLAB environment, version R2007b [Mathworks, Natick, MA, USA] and with the N-way Toolbox [67]. Each EEM corresponds to a landscape-matrix. For each succession of fluorescence spectra corresponds a collection of matrices which needs to be processed in a specific arrangement. A cube is usually used to organize landscape-matrices as depicted in the figure 10.

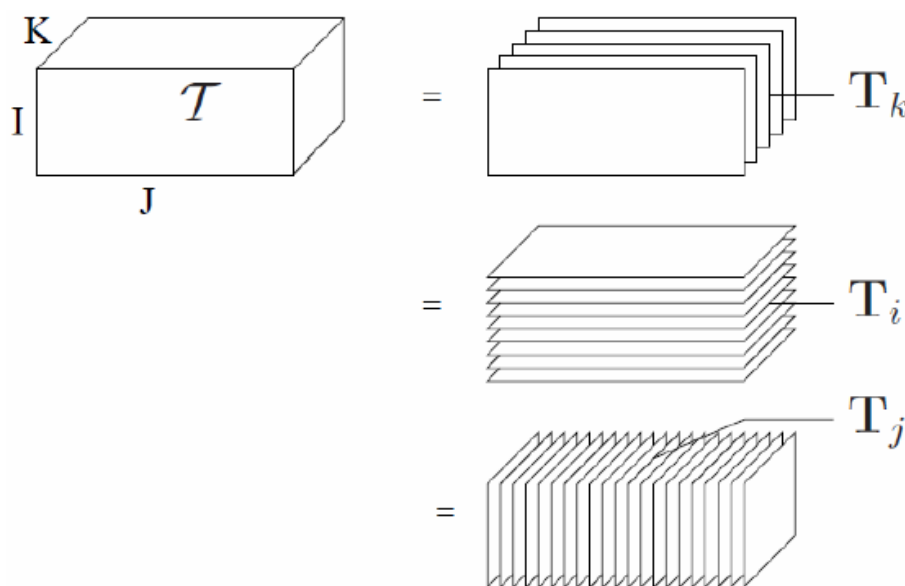


Figure 10. Representation of 3 possible types of arrangements for EEM landscapes processing.

From general point of view, processing this type of data arrangement needs two computing methods: a) 3-way methods such as PARAFAC, Tucker3 or Multiway-PCA and b) 2-way methods such as Principal Component Analysis (PCA) and any other bi-linear technique after data cube unfolding. In the first part of this work, the second approach was used: PCA after the $X^{(1)}$ unfolding of the data cube as illustrated in figure 11. From general point of view, if I is the dimension corresponding to samples, it is interesting to make the unfolding of the cube by keeping this dimension unchanged. One can see that $X^{(1)}$ is the unique unfolding that keep the sample dimension unchanged, therefore we took it thereafter for applying PCA.

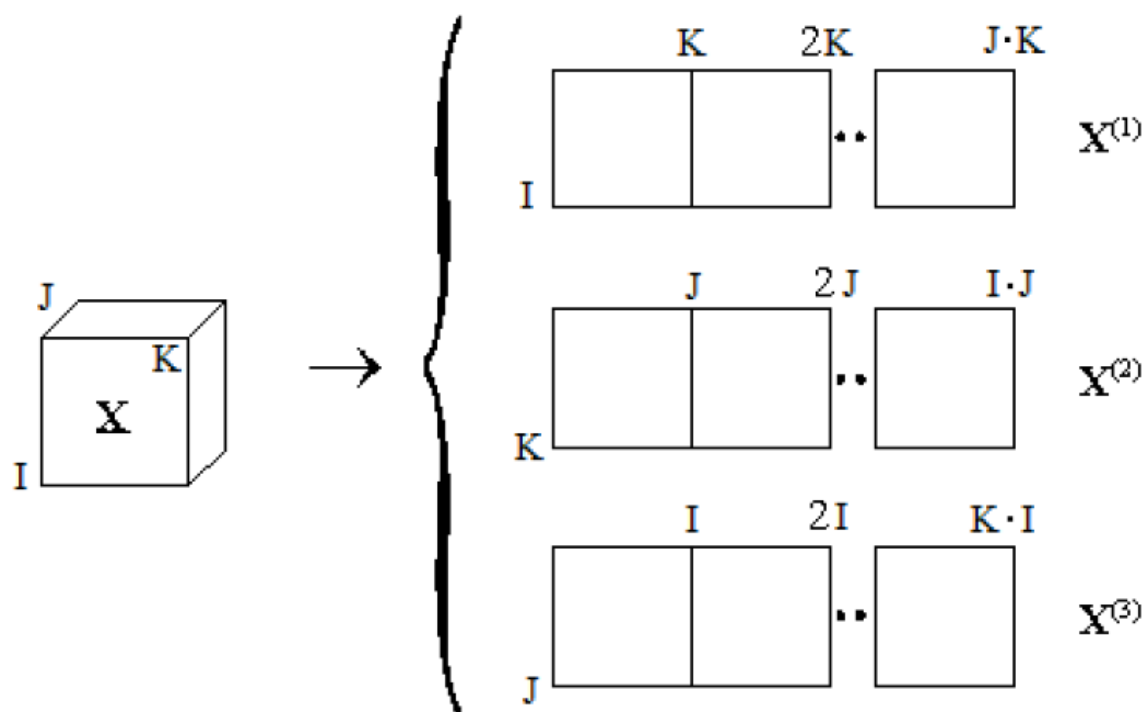


Figure 11. $X^{(1)}$ is obtained by juxtaposition of front slides of X . In the same way, $X^{(2)}$ et $X^{(3)}$ are obtained respectively by juxtaposition of horizontal and lateral slides of X .

5.5.1.4. Results and discussion

Here are below examples of 3D-Front Face Fluorescence spectra of samples of this study. The chemical composition of honey is studied and known to a large extent for nearly 50 years. The presence of compounds beneficial to health for their antioxidant properties or for other reasons is well known, this is the case of polyphenols [68-74] or amino acids [75-83], some of them are good fluorophores like polyphenols. Works evoke the interest of using these fluorophores as tracers of the floral origin of honeys. For example, the ellagic acid was used as a tracer of heather honey *Calluna* and *Erica* species while hesperitin was used to certify citrus honeys [84, 85] and abscisic acid was considered as molecular marker of Australian *Eucalyptus* honeys [86]. Kaempferol was used as marker of rosemary honey as well as quercetin for sunflower honey [87]. As polyphenols, aromatic amino acids were used to characterize the botanical origin or to test the authenticity of honey, this is the case of phenylalanine and tyrosine for honey lavender [88] and the glutamic acid for honeydew honeys [81]. Therefore, recording the overall fluorescence spectrum over a large range of wavelengths allows for taking into account the fluorescence emission of the major chemical components described above. In our case, recorded spectra for Acacia, Lavender and Chestnut honeys are similar for certain spectral regions but differ in others proving the existence of different fluorophores. These chemical characteristics specific of the composition are very useful for the distinction of flower varieties. Recording of 3D fluorescence spectra containing all the emission spectra over a wide range of excitation wavelengths can take into account all the information associated with the fluorophores present in honeys.

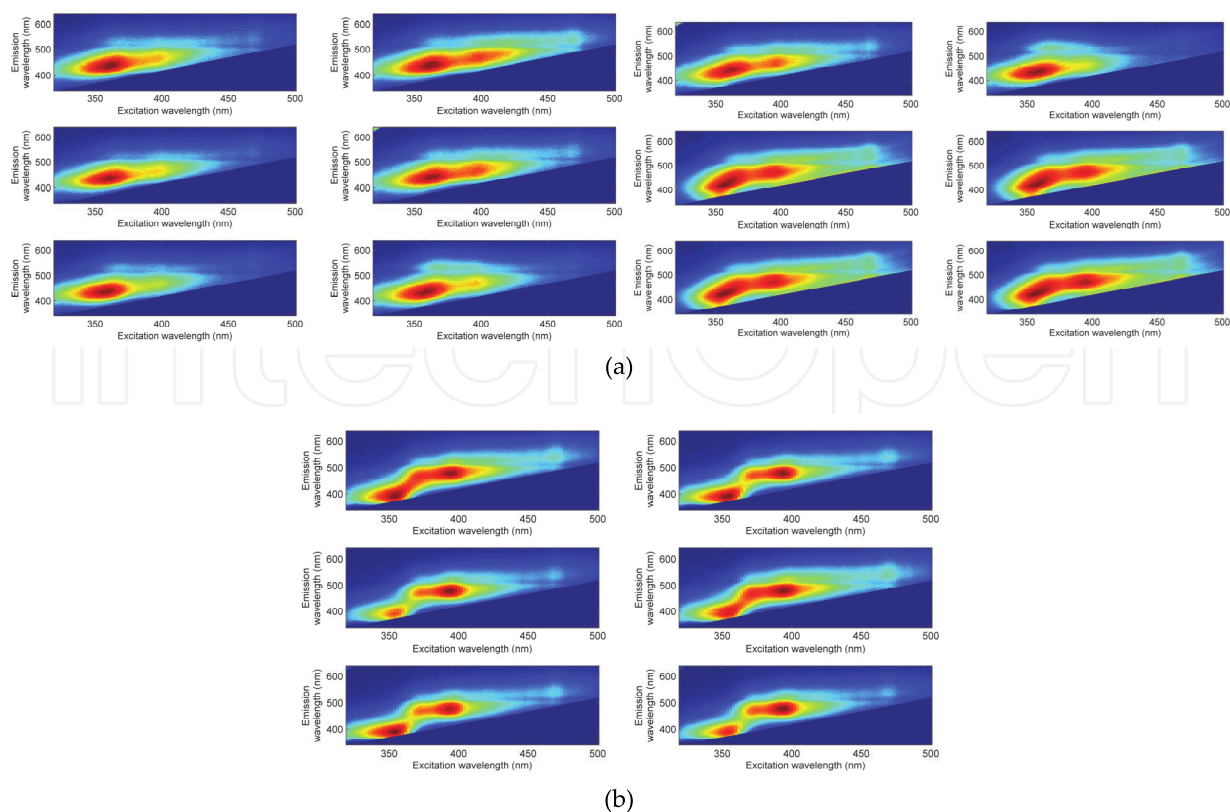
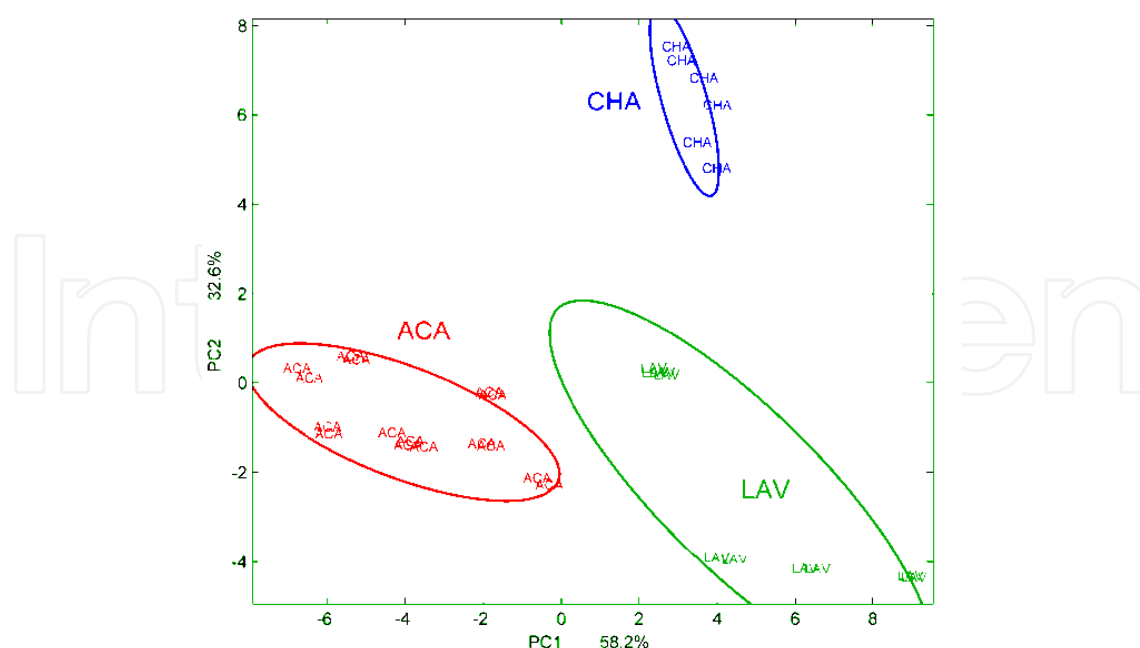


Figure 12. (a): At left, excitation-emission spectra of Acacia samples. At right, excitation-emission spectra of Lavender samples, (b): Excitation-Emission spectra of Chestnut samples

Figure 13 shows the PC1 vs PC2 scores-plot of the honeys dataset. PCA has played his role of data reduction technique by creating new set of axes. One can visualize all samples simultaneously on a simple xy-graph with only the two first principal components without loose of information. The first obvious conclusion is the natural classes of honeys clearly appear on the scores-plot. PC1 accounts for 58.2% of the total variance while PC2 explains 32.6% of the total variance of the initial data set. The Acacia group is separated from Chestnut and Lavender essentially on PC1, while PC2 is more characteristic of the distinction between Acacia + Lavender and Chestnut. The shape and location of fluorescence islets on 3D-spectra are assets in the distinction of samples as they relate to the chemical compounds that distinguish the groups. It is important to note here that other multivariate techniques exist that could be applied to these data successfully. Particular, the application of the technique ICA (Independent Component Analysis) which is a factorial technique similar to PCA, we would associate with each calculated component a signal having a greater chemical significance. ICA is capable from a mixture of signals to extract mutually independent components explaining more particularly the evolution of a "pure" signal [89-92]. In the case presented here, ICA should probably separate more effectively chemical source signals that are causing the observed differences between the honey samples. The chemical interpretation of this result can be assessed using the loadings on PC1 and PC2. As depicted by figure 14, some spectral regions are responsible of these separations observed on the two first axes of PCA. Two excitation/emission maxima (356/376 nm, 468/544 nm) and two excitation/emission minima (350/440 nm, 368/544 nm) are detectable on the PC1 loadings.



Legend: "ACA = Acacia"; "CHA = Chestnut"; LAV = Lavender

Figure 13. Ellipsoids of samples distribution drawn with interval confidence at 95%

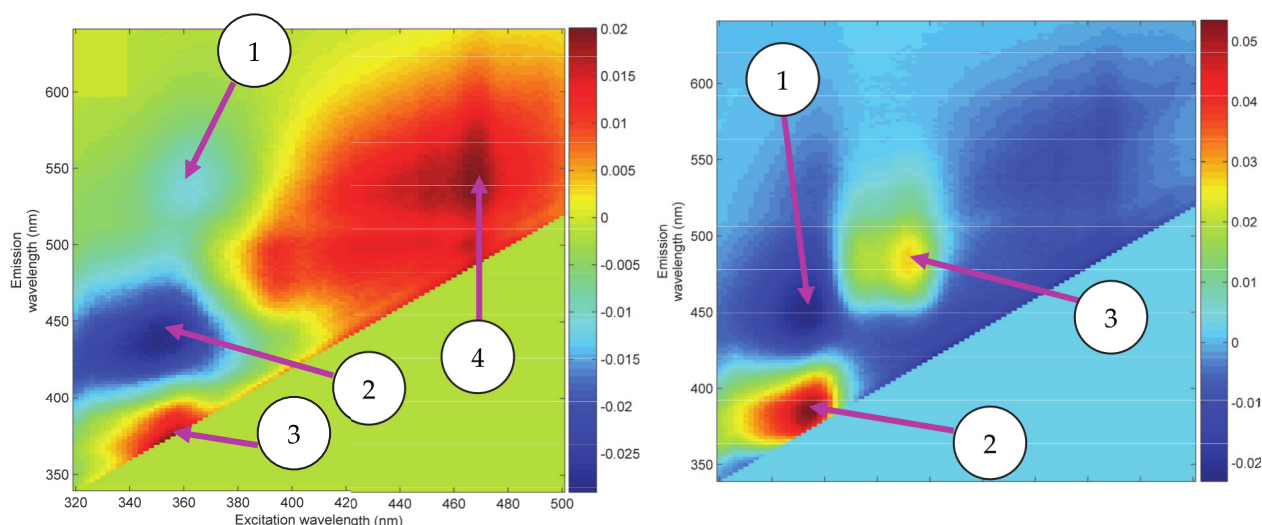


Figure 14. Loadings on PC1 (at left) and PC2 (at right) computed from EEM cube of honey samples after standardization.

In the same manner, the loadings on PC2 show two excitation/emission maxima (356/376 nm, 396/470 nm; respectively zones 2 and 3) and one excitation/emission minima (350/450 nm; zone 1). These maxima and minima are not absolute and cannot be assimilate to true concentrations but represent what spectral zones have largely change in intensity and shape throughout the analysed samples. So, loadings are an overall photography of changes in the samples mainly due to their botanical origins. Figure 15 presents an overview of the main fluorophores potentially present in dairy and food products [63, 93] and some of them are present in honeys too. Here, the chemical interpretation could be made as following. Based

on PC1, Acacia samples are mainly distinguished from Lavender and Chestnut honeys by fluorophores appearing in zone 1 and 2, their content in these fluorophores are greater than other samples. Symmetrically, Lavender and Chestnut honeys are more characterized by fluorophores detected in zone 3 and 4. Loadings on PC2 allow a better understanding of what is more characteristic of Chestnut honeys samples compared with others. Zone 2 and 3 are clearly associated with Chestnut samples because the loading values are positive in these regions as the scores are for these samples on the corresponding graph.

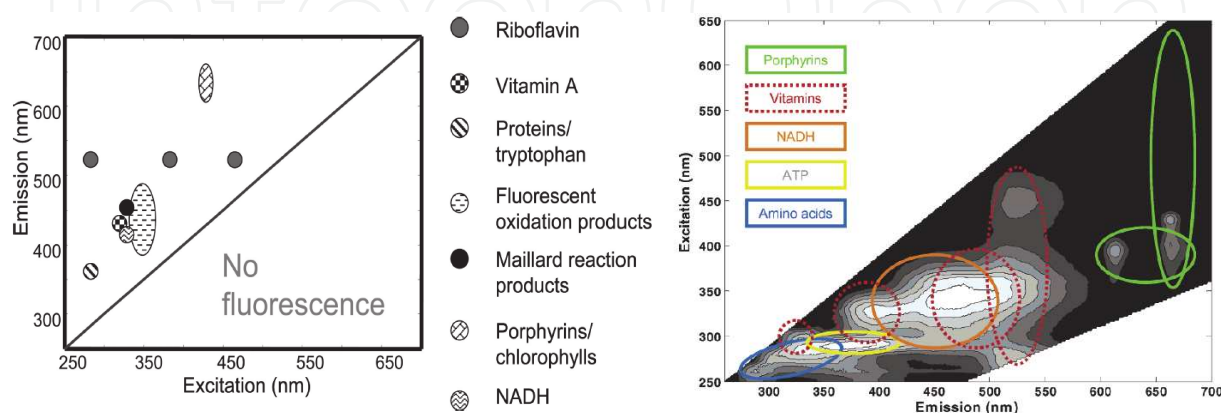


Figure 15. On Left - Excitation and emission maxima of fluorophores present in dairy products (from [93]); On right - Fluorescence landscape map indicating the spectral properties of the selected 11 food-relevant fluorophores (from [63]).

5.5.2. Application of PARAFAC

Let us to consider the previous example on the analysis of three varieties of monofloral honeys. The application of the PARAFAC model (with orthogonality constraint on the 3 modes) to the 3-way EEM fluorescence data can take into account the nature of these trilinear data and get factorial cards similarly to PCA. In the case of PARAFAC, it is usual to identify modes of the PARAFAC model from the dimensions of the original data cube. In our case, we built the cube of fluorescence data as $HoneyCube = [151 \times 91 \times 32]$ which is $[\lambda_{em} \times \lambda_{ex} \times \text{samples}]$.

Therefore, according to the theoretical model presented above we are able to visualize as many factorial cards as components couples in each mode there are. It is particularly interesting in the case of mode 3 which is mode of samples. The user must specify the number of components in each mode that must be calculated in the model. We built a 3-components model by helping us with the *corcondia* criterion associated with the PARAFAC procedure. The *corcondia* criterion was created and used [94] to facilitate the choice of the optimal number of components in the calculated model. It is a number between 0% (worst model fitting) and 100% (best model fitting). The ideal case for choosing the optimal number of components should be to know the exact number of fluorescence sources in the samples, but in our case this number is unknown. We have consider three sources of fluorescence according to our knowledge of the samples (proteins and free amino acids, NADH from cellular materials and oxidation products + Maillard reaction fluorescent products from

heating). The results for mode 1 (emission wavelengths), mode 2 (excitation wavelengths) and mode 3 (samples) are presented below in figures 16 and 17. An interesting parallel between the reconstructed PCA loadings and the PARAFAC loadings in each three modes is possible by forming the good figure arrangement.

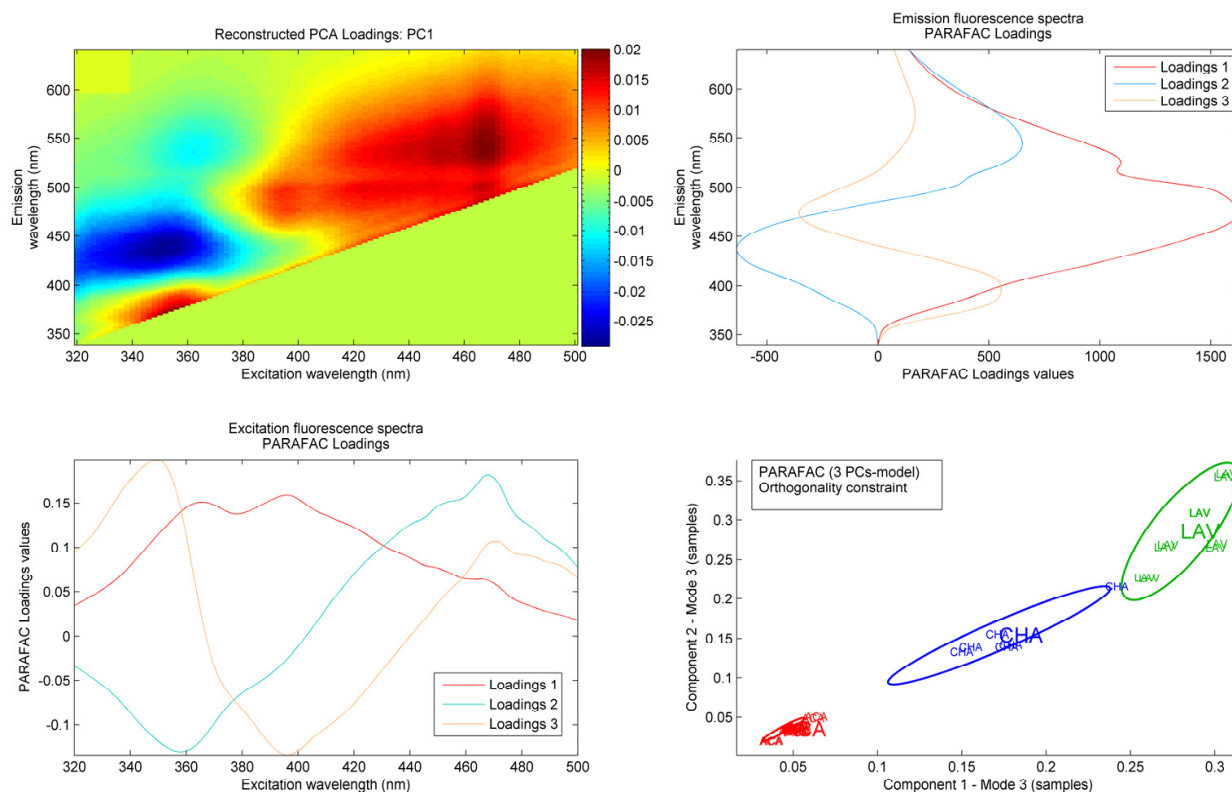


Figure 16. On top left hand side: PCA loadings for component 1; On top right hand side: PARAFAC loadings for Mode 1 (Emission); On bottom left hand side: PARAFAC loadings for Mode 2 (Excitation); On bottom right hand side: PARAFAC loadings for Mode 3 (Samples)

A simultaneous reading of PCA and PARAFAC loadings for honeys EEM matrices gives some elements for understanding what makes distinction between honeys samples on PARAFAC loadings on mode 3. One can see, for example, samples are relatively well separated on the two first PARAFAC loadings in mode 3 (chart on the bottom right). We find good agreement between the PARAFAC profiles and loadings of PCA. Similarly with PCA loadings, those of PARAFAC show the greatest variation of concentration profiles across all samples. And negative areas in the PARAFAC profiles reflect a decrease in fluorescence and therefore a decrease in concentration of the compounds across all the samples. Complementarily, positive PARAFAC profiles indicate an increase in fluorescence and thus the concentration of the corresponding fluorescent compounds in the samples. The first PARAFAC component in mode 3 (chart on the bottom right) allows a good discrimination of the honey samples and the latter is mainly explained by the variation of the fluorescence depicted by the loading 1 (red line) both on the emission (at top and right hand side) and excitation (at bottom and left hand side) graphs. Therefore, Acacia samples have lower content in these compounds than lavender honey samples. Chestnut honeys

have an intermediate level of concentration concerning these compounds. Figure 17 gives some elements to more accurately identify the chemical origin of the observed differences between samples on the 2nd PARAFAC component (in blue) on mode 3 (samples). Particularly, it is possible to associated to the 2nd PARAFAC component of mode 3 to Maillard and oxidation products (excitation wavelength: ≈ 360 nm; emission wavelength: ≈ 445 nm) depicted by the loadings 2 (blue curve) on the excitation and emission loadings graphs. This observation is explained by the method of harvesting, storage and chemical composition of honey. Indeed, lavender honey experienced a hot extraction and uncapping the honey frames also. This is because the lavender honey crystallizes faster than other honeys because of a higher concentration of sucrose while acacia honey is rich in fructose (sugar does not crystallize) and contains no sucrose. The water content and the glucose / fructose ratio of honeys are also important factors in the crystallization process [95]. The harvesting method is to cause a greater concentration of oxidation products and products of the Maillard reaction.

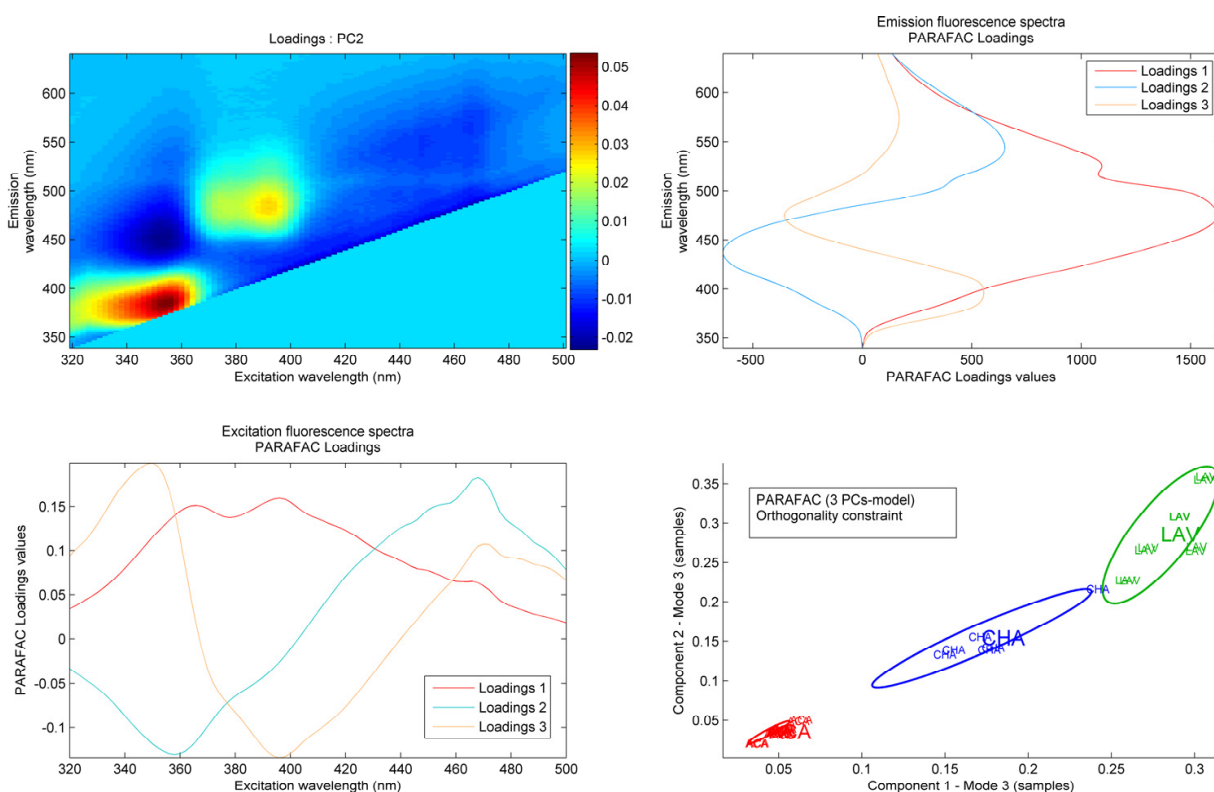


Figure 17. On top left hand side: PCA loadings for component 2; On top right hand side: PARAFAC loadings for Mode 1 (Emission); On bottom left hand side: PARAFAC loadings for Mode 2 (Excitation); On bottom right hand side: PARAFAC loadings for Mode 3 (Samples)

It is clear here that the PARAFAC loadings probably do not reflect pure signals. Indeed, in our case, the appropriate model is probably more complex because we do not know the exact number of fluorophores present in the samples and their contribution is included in each PARAFAC loadings. Thus the chemical interpretation is only partially correct, but nevertheless shows some interesting information on what produces the differences between samples. Another aspect that we have not discussed here is the type of constraint that we have imposed to the model during its construction (orthogonality, non-negativity, unimodality ...). The type of constraints strongly modifies the shape of the PARAFAC loadings obtained and their comparison with PCA is not always possible. We preferred the orthogonality constraint to be in conditions similar to PCA. But this is not necessarily the best choice from spectroscopic point of view, particularly because loadings reflecting concentrations should not be any negative.

5.6. Conclusion

If there were one or two things to note from this application of PARAFAC, it would be, first, the simplicity of the graphical outputs and their complementarities with the results of PCA, and secondly, the possibility offered by the model to process data inherently 3-way or more (without preliminary rearrangement).

In the case of 3D fluorescence data, the loadings of PCA are not easily used directly to interpret and to construct a model of quantitative monitoring, PARAFAC modelling enables this by producing as much as possible loadings representative of changes in concentrations of fluorophores present in the samples, and a good distinction of the studied honey varieties is easily achievable with a simple PARAFAC model.

Appendix

A. Chemometrics books

Here is below a selection of the most interesting books dealing with chemometrics. This list will be convenient as well for the beginner as for the specialist.

- Härdle, W.; Simar, L., *Applied multivariate statistical analysis*. 2nd ed.; Springer: Berlin; New York, 2007; p xii, 458 p.
- Brereton, R. G., *Chemometrics for pattern recognition*. Wiley: Chichester, U.K., 2009; p xvii, 504 p.
- Gemperline, P., *Practical guide to chemometrics*. 2nd ed.; CRC/Taylor & Francis: Boca Raton, 2006; p 541 p.
- Miller, J. N.; Miller, J. C., *Statistics and chemometrics for analytical chemistry*. 5th ed.; Pearson Prentice Hall: Harlow, England; New York, 2005; p xvi, 268 p.
- Beebe, K. R.; Pell, R. J.; Seasholtz, M. B., *Chemometrics : a practical guide*. Wiley: New York, 1998; p xi, 348 p.
- Brown, S. D., *Comprehensive chemometrics*. Elsevier: Boston, MA, 2009

- Jackson, J. E., *A user's guide to principal components*. Wiley-Interscience: Hoboken, N.J., 2003; p xvii, 569 p.
- Martens, H.; Næs, T., *Multivariate calibration*. Wiley: Chichester [England] ; New York, 1989; p xvii, 419 p.
- Massart, D. L., *Handbook of chemometrics and qualimetrics*. Elsevier: Amsterdam ; New York, 1997.
- Sharaf, M. A.; Illman, D. L.; Kowalski, B. R., *Chemometrics*. Wiley: New York, 1986; p xi, 332 p.

B. Matlab codes

B.1 PCA by decomposition of the covariance matrix

```
% Size of the data matrix X
[n,p]=size(X);

% Column centering
meanX = mean(X);
X_Centred = X - ones(n,1) * meanX;

% Compute variance-covariance matrix
% for scores
CovM_scores = cov(X_Centred);

% for variables
CovM_loadings = cov(X_Centred');

% Compute eigenvalue & eigenvectors by keeping the matrix form
[V_scores,D_scores] = eig(CovM_scores);
D_scores = fliplr(D_scores);D_scores = flipud(D_scores);
V_scores = fliplr(V_scores);

[V_loadings,D_loadings] = eig(CovM_loadings);
D_loadings = fliplr(D_loadings);D_loadings = flipud(D_loadings);
V_loadings = fliplr(V_loadings);

% Compute of total variance of the dataset: TotalVP
TotalVP = trace(D_scores);

% Compute the percentage of variance explained by each principal
%component
```

```

Pourcentages = {};
for i=1:size(D_scores,1)
    Perrcentages{i,1} = strcat('PC',num2str(i));
    Perrcentages{i,2} = num2str(D_scores(i,i)/TotalVP*100);
end;
% Compute factorial coordinates (scores) for each sample point on
% components space
Coord = [];
Coord = X_Cent*V_scores;
% Compute factorial contributions (loadings) for each sample point
on
% components space
Contrib = [];
Contrib = X_Cent'*V_loadings;

```

B.2 PCA by SVD

```

% Size of the data matrix X
[n,p]=size(X);
% Column centering
meanX = mean(X);
X_Centred = X - ones(n,1) * meanX;
% Compute scores and loadings simultaneously
[U SingularValues Loadings] = svd(X_Centred);
% Scores T
T = U * SingularValues;

```

C. Chemometrics software list

- CAMO (<http://www.camo.no/>) maker of Unscrambler software, for multivariate modelling, prediction, classification, and experimental design.
- Chemometry Consultancy (<http://www.chemometry.com/>)
- Eigenvector Research (<http://www.eigenvector.com/>) sells a PLS_Toolbox
- Minitab (<http://www.minitab.com/>)
- Statcon (<http://www.statcon.de/>)
- Stat-Ease (<http://www.statease.com/>) maker of Design-Expert DOE software for experimental design.

- Infometrix, Inc., developer of Pirouette, InStep and LineUp packages for multivariate data analysis. We specialize in integrating chemometrics into analytical instruments and process systems. (<http://www.infometrix.com/>)
- Umetrics, maker of MODDE software for design of experiments and SIMCA multivariate data analysis. (<http://www.umetrics.com/>)
- Thermo Electron, offers PLSplus/IQ add-on to GRAMS.
(<http://www.thermo.com/com/cda/landingpage/0,,585,00.html>)
- Grabitech Solutions AB, offering MultiSimplex, experimental design & optimization software. (<http://www.grabitech.se/>)
- PRS Software AS, maker of Sirius software, for data analysis and experimental design. (<http://www.prs.no/>)
- S-Matrix Corp., maker of CARD, design of experiments software. (<http://www.smatrix.com/>)
- Applied Chemometrics, source for chemometrics related software, training, and consulting. (<http://www.chemometrics.com/>)
- DATAN software for multidimensional chemometric analysis. (<http://www.multid.se/>)

D. Chemometrics websites

- Belgian Chemometrics Society (<http://chemometrie.kvcv.be/>)
- Département d'Analyse Pharmaceutique et Biomédicale (FABI) à la Vrije Universiteit Brussel (<http://www.vub.ac.be/fabi>)
- McMaster University Hamilton - Department of Chemical Engineering - Prof. Dora Kourti (<http://www.chemeng.mcmaster.ca/faculty/kourti/default.htm>)
- NAmICS (<http://www.namics.nysaes.cornell.edu/>)
- Research NIR Centre of Excellence (NIRCE) - Prof. Paul Geladi (<http://www.fieldnirce.org/>)
- Université d'Anvers – Groupe de Recherche pour la chimiométrie et l'analyse par rayons x - Prof. Van Espen (<http://www.chemometrix.ua.ac.be/>)
- Université d'Anvers – Département de Mathématique, Statistique et Actuariat - Prof. Peter Goos (<http://www.ua.ac.be/main.aspx?c=peter.goos>)
- Université royale vétérinaire et agronomique de Fredriksberg – Groupe de Recherche en Chimiométrie - Prof. Rasmus Bro (<http://www.models.kvl.dk/users/rasmus/>)
- Université de Louvain - Research Center for Operations Research and Business Statistics - Prof. Martina Vandebroek (<http://www.econ.kuleuven.be/martina.vandebroek>)
- Université de Nimègue - Chemometrics Research Department - Prof. Lutgarde Buydens (<http://www.cac.science.ru.nl/>)
- Université de Silésie – Département de chimiométrie - Prof. Beata Walczak (<http://chemometria.us.edu.pl/researchBeata.html>)
- Johan Trygg (<http://www.chemometrics.se/>)
- Analytical Chemistry Laboratory, AgroParisTech, France (<http://www.chimiometrie.fr>)

Author details

Christophe B.Y. Cordella

UMR1145 INRA/AgroParisTech, Institut National de la Recherche Agronomique,
Laboratoire de Chimie Analytique, Paris, France

Acknowledgments

I thank Dr Dominique Bertrand (INRA Nantes, France) for proofreading and advice about the PCA and for his availability. I also thank my students who have provided much of the data used in this chapter.

6. References

- [1] Leardi, R., Chemometrics: From classical to genetic algorithms. *Grasas Y Aceites* 2002, 53, (1), 115-127.
- [2] Siebert, K. J., Chemometrics in brewing - A review. *Journal of the American Society of Brewing Chemists* 2001, 59, (4), 147-156.
- [3] Molher-Smith, A.; Nakai, S., Classification of Cheese Varieties by Multivariate Analysis of HPLC Profiles. *Can. Inst. Food Sci. Technol. J.* 1990, 23, (1), 53-58.
- [4] Lai, Y. W.; Kemsley, E. K.; Wilson, R. H., Potential of Fourier Transform Infrared Spectroscopy for the Authentication of Vegetable Oils. *J. Agric. Food Chem.* 1994, 42, 1154-1159.
- [5] Defernez, M.; Kemsley, E. K.; Wilson, R. H., Use of Infrared Spectroscopy and Chemometrics for the Authentication of Fruit Purees. *Journal of Agricultural and Food Chemistry* 1995, 43, (1), 109-113.
- [6] Twomey, M.; Downey, G.; McNulty, P. B., The potential of NIR spectroscopy for the detection of the adulteration of orange juice. *Journal of the Science of Food and Agriculture* 1995, 67, (1), 77-84.
- [7] Suchánek, M.; Filipová, H.; Volka, K.; Delgadillo, I.; Davies, A., Qualitative analysis of green coffee by infrared spectrometry. *Fresenius' Journal of Analytical Chemistry* 1996, 354, (3), 327-332.
- [8] Baeten, V.; Meurens, M.; Morales, M. T.; Aparicio, R., Detection of Virgin Olive Oil Adulteration by Fourier Transform Raman Spectroscopy. *Journal of Agricultural and Food Chemistry* 1996, 44, (8), 2225-2230.
- [9] Downey, G., Food and food ingredient authentication by mid-infrared spectroscopy and chemometrics. *TrAC Trends in Analytical Chemistry* 1998, 17, (7), 418-424.
- [10] Garcia-Lopez, C.; Grane-Teruel, N.; Berenguer-Navarro, V.; Garcia-Garcia, J. E.; Martin-Carratala, M. L., Major Fatty Acid Composition of 19 Almond Cultivars of Different Origins. A Chemometric Approach. *Journal of Agricultural and Food Chemistry* 1996, 44, (7), 1751-1755.

- [11] Martín Carratalá, M. L.; García-López, C.; Berenguer-Navarro, V.; Grané-Teruel, N., New Contribution to the Chemometric Characterization of Almond Cultivars on the Basis of Their Fatty Acid Profiles. *Journal of Agricultural and Food Chemistry* 1998, 46, (3), 963-967.
- [12] Mangas, J. J.; Moreno, J.; Picinelli, A.; Blanco, D., Characterization of Cider Apple Fruits According to Their Degree of Ripening. A Chemometric Approach. *Journal of Agricultural and Food Chemistry* 1998, 46, (10), 4174-4178.
- [13] Blanco-Gomis, D.; Mangas Alonso, J. J.; Margolles Cabrales, I.; Arias Abrodo, P., Characterization of Cider Apples on the Basis of Their Fatty Acid Profiles. *Journal of Agricultural and Food Chemistry* 2002, 50, (5), 1097-1100.
- [14] König, T.; Schreier, P., Application of multivariate statistical methods to extend the authenticity control of flavour constituents of apple juice. *Zeitschrift für Lebensmitteluntersuchung und -Forschung A* 1999, 208, (2), 130-133.
- [15] Ding, H. B.; Xu, R. J., Near-Infrared Spectroscopic Technique for Detection of Beef Hamburger Adulteration. *Journal of Agricultural and Food Chemistry* 2000, 48, (6), 2193-2198.
- [16] Maeztu, L.; Andueza, S.; Ibanez, C.; Paz de Pena, M.; Bello, J.; Cid, C., Multivariate Methods for Characterization and Classification of Espresso Coffees from Different Botanical Varieties and Types of Roast by Foam, Taste, and Mouthfeel. *Journal of Agricultural and Food Chemistry* 2001, 49, (10), 4743-4747.
- [17] Schoonjans, V.; Massart, D. L., Combining spectroscopic data (MS, IR): exploratory chemometric analysis for characterising similarity/diversity of chemical structures. *Journal of Pharmaceutical and Biomedical Analysis* 2001, 26, (2), 225-239.
- [18] Terrab, A.; Vega-Pérez, J. M.; Díez, M. J.; Heredia, F. J., Characterisation of northwest Moroccan honeys by gas chromatographic-mass spectrometric analysis of their sugar components. *Journal of the Science of Food and Agriculture* 2002, 82, (2), 179-185.
- [19] Brenna, O. V.; Pagliarini, E., Multivariate Analysis of Antioxidant Power and Polyphenolic Composition in Red Wines. *Journal of Agricultural and Food Chemistry* 2001, 49, (10), 4841-4844.
- [20] Destefanis, G.; Barge, M. T.; Brugiapaglia, A.; Tassone, S., The use of principal component analysis (PCA) to characterize beef. *Meat Sci* 2000, 56, (3), 255-9.
- [21] Chan, T. F., An Improved Algorithm for Computing the Singular Value Decomposition. *ACM Transactions on Mathematical Software* 1982, 8, (1), 72-83.
- [22] Golub, G.; Kahan, W., Calculating the Singular Values and Pseudo-Inverse of a Matrix. *Journal of the Society for Industrial and Applied Mathematics, Series B: Numerical Analysis* 1965, 2, (2), 205-224.
- [23] Mermet, J.-M.; Otto, M.; Widmer, H. M., Multivariate Methods. In *Analytical Chemistry*, Kellner, R., Ed. Wiley-VCH Verlag GmbH: D-69469 Weinheim, Germany, 1998; pp p776-808.
- [24] Choe, E.; Min, D. B., Chemistry and reactions of reactive oxygen species in foods. *Crit Rev Food Sci Nutr* 2006, 46, (1), 1-22.

- [25] Choe, E.; Min, D. B., Chemistry of deep-fat frying oils. *J Food Sci* 2007, 72, (5), R77-86.
- [26] Choe, E.; Min, D. B., Mechanisms and Factors for Edible Oil Oxidation. *Comprehensive Reviews in Food Science and Food Safety* 2006, 5, (4), 169-186.
- [27] Frankel, E. N.; Min, D. B.; Smouse, T. H., Chemistry of autoxidation: products and flavor significance. In *Flavor Chemistry of Fats and Oils, Champaign III*, American Oil Chemist's Society: 1985; pp 1-34.
- [28] Cordella, C. B.; Tekye, T.; Rutledge, D. N.; Leardi, R., A multiway chemometric and kinetic study for evaluating the thermal stability of edible oils by (1)H NMR analysis: comparison of methods. *Talanta* 2012, 88, 358-68.
- [29] Harwood, J. A. R., *Handbook of Olive Oil Analysis and Properties*. Aspen Publishers: Gaithersburg, Maryland, 2000.
- [30] Aguila, M. B.; Pinheiro, A. R.; Aquino, J. C.; Gomes, A. P.; Mandarim-de-Lacerda, C. A., Different edible oil beneficial effects (canola oil, fish oil, palm oil, olive oil, and soybean oil) on spontaneously hypertensive rat glomerular enlargement and glomeruli number. *Prostaglandins Other Lipid Mediat* 2005, 76, (1-4), 74-85.
- [31] Guillén, M. D.; Goicoechea, E., Oxidation of corn oil at room temperature: Primary and secondary oxidation products and determination of their concentration in the oil liquid matrix from 1H nuclear magnetic resonance data. *Food Chemistry* 2009, 116, (1), 183-192.
- [32] Carrer, S.; Leardi, R., Characterizing the pollution produced by an industrial area: chemometric methods applied to the Lagoon of Venice. *Sci Total Environ* 2006, 370, (1), 99-116.
- [33] Harshman, R. A., Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multimodal factor analysis.(University Microfilms, Ann Arbor, Michigan, No. 10,085). *UCLA Working Papers in Phonetics* 1970, 16, 1-84.
- [34] Carroll, J.; Chang, J.-J., Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition. *Psychometrika* 1970, 35, (3), 283-319.
- [35] Hitchcock, F. L., The expression of a tensor or a polyadic as a sum of products. *Journal of Mathematics and Physics* 1927, 6, 164-189.
- [36] Booksh, K. S.; Kowalski, B. R., Calibration method choice by comparison of model basis functions to the theoretical instrumental response function. *Analytica Chimica Acta* 1997, 348, 1-9.
- [37] Tucker, L. R., Some mathematical notes on three-mode factor analysis. *Psychometrika* 1966, 31, (3), 279-311.
- [38] Kiers, H., Hierarchical relations among three-way methods. *Psychometrika* 1991, 56, (3), 449-470.
- [39] Bro, R., PARAFAC. Tutorial and applications. *Chemometrics and Intelligent Laboratory Systems* 1997, 38, (2), 149-171.

- [40] Esteves da Silva, C. G. J.; Novais, A. G. S., Trilinear PARAFAC decomposition of synchronous fluorescence spectra of mixtures of the major metabolites of acetylsalicylic acid. *Analyst* 1998, 123, (10), 2067-2070.
- [41] Baunsgaard, D.; Norgaard, L.; Godshall, M. A., Fluorescence of Raw Cane Sugars Evaluated by Chemometrics. *Journal of Agricultural and Food Chemistry* 2000, 48, (10), 4955-4962.
- [42] Moberg, L.; Robertsson, G.; Karlberg, B., Spectrofluorimetric determination of chlorophylls and pheopigments using parallel factor analysis. *Talanta* 2001, 54, (1), 161-70.
- [43] Andersen, C. M.; Bro, R., Practical aspects of PARAFAC modeling of fluorescence excitation-emission data. *Journal of Chemometrics* 2003, 17, (4), 200-215.
- [44] Marengo, E.; Leardi, R.; Robotti, E.; Righetti, P. G.; Antonucci, F.; Cecconi, D., Application of Three-Way Principal Component Analysis to the Evaluation of Two-Dimensional Maps in Proteomics. *Journal of Proteome Research* 2003, 2, (4), 351-360.
- [45] Christensen, J.; Povlsen, V. T.; Sørensen, J., Application of Fluorescence Spectroscopy and Chemometrics in the Evaluation of Processed Cheese During Storage. *Journal of Dairy Science* 2003, 86, (4), 1101-1107.
- [46] Mahedero, M. C.; Diaz, N. M.; Muñoz de la Peña, A.; Espinosa Mansilla, A.; Gonzalez Gomez, D.; Bohoyo Gil, D., Strategies for solving matrix effects in the analysis of sulfathiazole in honey samples using three-way photochemically induced fluorescence data. *Talanta* 2005, 65, (3), 806-813.
- [47] Andersen, C. M.; Vishart, M.; Holm, V. K., Application of Fluorescence Spectroscopy in the Evaluation of Light-Induced Oxidation in Cheese. *Journal of Agricultural and Food Chemistry* 2005, 53, (26), 9985-9992.
- [48] Guimet, F.; Ferre, J.; Boque, R.; Vidal, M.; Garcia, J., Excitation-Emission Fluorescence Spectroscopy Combined with Three-Way Methods of Analysis as a Complementary Technique for Olive Oil Characterization. *Journal of Agricultural and Food Chemistry* 2005, 53, (24), 9319-9328.
- [49] Wold, J. P.; Bro, R.; Veberg, A.; Lundby, F.; Nilsen, A. N.; Moan, J., Active Photosensitizers in Butter Detected by Fluorescence Spectroscopy and Multivariate Curve Resolution. *Journal of Agricultural and Food Chemistry* 2006, 54, (26), 10197-10204.
- [50] Muñoz de la Peña, A.; Durán Merás, I.; Jiménez Girón, A.; Goicoechea, H. C., Evaluation of unfolded-partial least-squares coupled to residual trilinearization for four-way calibration of folic acid and methotrexate in human serum samples. *Talanta* 2007, 72, (4), 1261-1268.
- [51] Muñoz de la Peña, A.; Mora Diez, N.; Mahedero García, M. C.; Bohoyo Gil, D.; Cañada-Cañada, F., A chemometric sensor for determining sulphaguanidine residues in honey samples. *Talanta* 2007, 73, (2), 304-313.

- [52] Løje, H.; Green-Petersen, D.; Nielsen, J.; Jørgensen, B. M.; Jensen, K. N., Water distribution in smoked salmon. *Journal of the Science of Food and Agriculture* 2007, 87, (2), 212-217.
- [53] Bosco, M. V.; Larrechi, M. S., PARAFAC and MCR-ALS applied to the quantitative monitoring of the photodegradation process of polycyclic aromatic hydrocarbons using three-dimensional excitation emission fluorescent spectra: Comparative results with HPLC. *Talanta* 2007, 71, (4), 1703-1709.
- [54] Rodriguez, N.; Ortiz, M. C.; Sarabia, L. A., Fluorescence quantification of tetracycline in the presence of quenching matrix effect by means of a four-way model. *Talanta* 2009, 77, (3), 1129-1136.
- [55] Andersen, C. M.; Mortensen, G., Fluorescence Spectroscopy: A Rapid Tool for Analyzing Dairy Products. *Journal of Agricultural and Food Chemistry* 2008, 56, (3), 720-729.
- [56] Hashemi, J.; Kram, G. A.; Alizadeh, N., Enhanced spectrofluorimetric determination of aflatoxin B1 in wheat by second-order standard addition method. *Talanta* 2008, 75, (4), 1075-81.
- [57] Masino, F.; Foca, G.; Ulrici, A.; Arru, L.; Antonelli, A., A chemometric study of pesto sauce appearance and of its relation to pigment concentration. *Journal of the Science of Food and Agriculture* 2008, 88, (8), 1335-1343.
- [58] Moros, J.; Iñón, F. A.; Garrigues, S.; de la Guardia, M., Determination of vinegar acidity by attenuated total reflectance infrared measurements through the use of second-order absorbance-pH matrices and parallel factor analysis. *Talanta* 2008, 74, (4), 632-641.
- [59] Bro, R.; Qannari, E. M.; Kiers, H. A. L.; Næs, T.; Frøst, M. B., Multi-way models for sensory profiling data. *Journal of Chemometrics* 2008, 22, (1), 36-45.
- [60] Cordella, C. B. Y.; Leardi, R.; Rutledge, D. N., Three-way principal component analysis applied to noodles sensory data analysis. *Chemometrics and Intelligent Laboratory Systems* 2011, 106, (1), 125-130.
- [61] Cordella, C. B.; Militao, J. S.; Cabrol-Bass, D., A simple method for automated pretreatment of usable chromatographic profiles in pattern-recognition procedures: application to HPAEC-PAD chromatograms of honeys. *Anal Bioanal Chem* 2003, 377, (1), 214-9.
- [62] Cordella, C. B.; Militao, J. S.; Clement, M. C.; Cabrol-Bass, D., Honey characterization and adulteration detection by pattern recognition applied on HPAEC-PAD profiles. 1. Honey floral species characterization. *J Agric Food Chem* 2003, 51, (11), 3234-42.
- [63] Christensen, J.; Norgaard, L.; Bro, R.; Engelsen, S. B., Multivariate autofluorescence of intact food systems. *Chem Rev* 2006, 106, (6), 1979-94.
- [64] Sádecká, J.; Tóthová, J., Fluorescence spectroscopy and chemometrics in the food classification – a review. *Czech J. Food Sci.* 2007, 25, 159-174.
- [65] Ruoff, K.; Luginbuhl, W.; Kunzli, R.; Bogdanov, S.; Bosset, J. O.; von der Ohe, K.; von der Ohe, W.; Amado, R., Authentication of the botanical and geographical origin of

- honey by front-face fluorescence spectroscopy. *J Agric Food Chem* 2006, 54, (18), 6858-66.
- [66] Louveaux J.; Maurizio A.; G., V., Methods of Melissopalynology. *Bee World* 1978, 59, 139-157.
- [67] Andersson, C. A.; Bro, R., The N-way Toolbox for MATLAB. *Chemometrics and Intelligent Laboratory Systems* 2000, 52, (1), 1-4.
- [68] Amiot, M. J.; Aubert, S.; Gonnet, M.; Tacchini, M., Les composés phénoliques des miels: étude préliminaire sur l'identification et la quantification par familles. *Apidologie* 1989, 20, 115-125.
- [69] Andrade, P.; Ferreres, F.; Amaral, M. T., Analysis of Honey Phenolic Acids by HPLC, Its Application to Honey Botanical Characterization. *Journal of Liquid Chromatography & Related Technologies* 1997, 20, (14), 2281-2288.
- [70] Martos, I.; Ferreres, F.; Tomas-Barberan, F. A., Identification of flavonoid markers for the botanical origin of Eucalyptus honey. *J Agric Food Chem* 2000, 48, (5), 1498-502.
- [71] Martos, I.; Ferreres, F.; Yao, L.; D'Arcy, B.; Caffin, N.; Tomas-Barberan, F. A., Flavonoids in monospecific eucalyptus honeys from Australia. *J Agric Food Chem* 2000, 48, (10), 4744-8.
- [72] Yao, L.; Datta, N.; Tomás-Barberán, F. A.; Ferreres, F.; Martos, I.; Singanusong, R., Flavonoids, phenolic acids and abscisic acid in Australian and New Zealand Leptospermum honeys. *Food Chemistry* 2003, 81, (2), 159-168.
- [73] Yao, L.; Jiang, Y.; D'Arcy, B.; Singanusong, R.; Datta, N.; Caffin, N.; Raymont, K., Quantitative high-performance liquid chromatography analyses of flavonoids in Australian Eucalyptus honeys. *J Agric Food Chem* 2004, 52, (2), 210-4.
- [74] Yao, L.; Jiang, Y.; Singanusong, R.; D'Arcy, B.; Datta, N.; Caffin, N.; Raymont, K., Flavonoids in Australian Melaleuca, Guioa, Lophostemon, Banksia and Helianthus honeys and their potential for floral authentication. *Food Research International* 2004, 37, (2), 166-174.
- [75] Wang, J.; Li, Q. X.; Steve, L. T., Chapter 3 - Chemical Composition, Characterization, and Differentiation of Honey Botanical and Geographical Origins. In *Advances in Food and Nutrition Research*, Academic Press: Vol. Volume 62, pp 89-137.
- [76] Senyuva, H. Z.; Gilbert, J.; Silici, S.; Charlton, A.; Dal, C.; Gurel, N.; Cimen, D., Profiling Turkish honeys to determine authenticity using physical and chemical characteristics. *J Agric Food Chem* 2009, 57, (9), 3911-9.
- [77] Rebane, R.; Herodes, K., Evaluation of the botanical origin of Estonian uni- and polyfloral honeys by amino acid content. *J Agric Food Chem* 2008, 56, (22), 10716-20.
- [78] Perez, R. A.; Iglesias, M. T.; Pueyo, E.; Gonzalez, M.; de Lorenzo, C., Amino acid composition and antioxidant capacity of Spanish honeys. *J Agric Food Chem* 2007, 55, (2), 360-5.

- [79] Iglesias, M. T.; Martin-Alvarez, P. J.; Polo, M. C.; de Lorenzo, C.; Gonzalez, M.; Pueyo, E., Changes in the free amino acid contents of honeys during storage at ambient temperature. *J Agric Food Chem* 2006, 54, (24), 9099-104.
- [80] Iglesias, M. T.; Martin-Alvarez, P. J.; Polo, M. C.; de Lorenzo, C.; Pueyo, E., Protein analysis of honeys by fast protein liquid chromatography: application to differentiate floral and honeydew honeys. *J Agric Food Chem* 2006, 54, (21), 8322-7.
- [81] Iglesias, M. T.; De Lorenzo, C.; Del Carmen Polo, M.; Martin-Alvarez, P. J.; Pueyo, E., Usefulness of amino acid composition to discriminate between honeydew and floral honeys. Application to honeys from a small geographic area. *J Agric Food Chem* 2004, 52, (1), 84-9.
- [82] Bergner, K. G.; Hahn, H., [Phenylalanine content of honeys]. *Z Ernahrungswiss* 1972, 11, (1), 47-54.
- [83] Pirini, A.; Conte, L. S.; Francioso, O.; Lercker, G., Capillary gas chromatographic determination of free amino acids in honey as a means of discrimination between different botanical sources. *Journal of High Resolution Chromatography* 1992, 15, (3), 165-170.
- [84] Ferreres, F.; García-Viguera, C.; Tomás-Lorente, F.; Tomás-Barberán, F. A., Hesperetin: A marker of the floral origin of citrus honey. *Journal of the Science of Food and Agriculture* 1993, 61, (1), 121-123.
- [85] Tomás-Barberán, F. A.; Martos, I.; Ferreres, F.; Radovic, B. S.; Anklam, E., HPLC flavonoid profiles as markers for the botanical origin of European unifloral honeys. *Journal of the Science of Food and Agriculture* 2001, 81, (5), 485-496.
- [86] Yao, L.; Jiang, Y.; Singanusong, R.; Datta, N.; Raymont, K., Phenolic acids and abscisic acid in Australian Eucalyptus honeys and their potential for floral authentication. *Food Chemistry* 2004, 86, (2), 169-177.
- [87] Pyrzynska, K.; Biesaga, M., Analysis of phenolic acids and flavonoids in honey. *TrAC Trends in Analytical Chemistry* 2009, 28, (7), 893-902.
- [88] Bouseta, A.; Scheirman, V.; Collin, S., Flavor and Free Amino Acid Composition of Lavender and Eucalyptus Honeys. *Journal of Food Science* 1996, 61, (4), 683-687.
- [89] Comon, P., Independent component analysis, A new concept? *Signal Processing* 1994, 36, (3), 287-314.
- [90] Bouveresse, D. J.; Benabid, H.; Rutledge, D. N., Independent component analysis as a pretreatment method for parallel factor analysis to eliminate artefacts from multiway data. *Anal Chim Acta* 2007, 589, (2), 216-24.
- [91] Mehrabian, H.; Lindvere, L.; Stefanovic, B.; Martel, A. L., A constrained independent component analysis technique for artery-vein separation of two-photon laser scanning microscopy images of the cerebral microvasculature. *Med Image Anal* 2012, 16, (1), 239-51.
- [92] Liu, X.; Liu, F.; Zhang, Y.; Bai, J., Unmixing dynamic fluorescence diffuse optical tomography images with independent component analysis. *IEEE Trans Med Imaging* 2011, 30, (9), 1591-604.

- [93] Andersen, C. M. I.; Mortensen, G., Fluorescence Spectroscopy: A Rapid Tool for Analyzing Dairy Products. *Journal of Agricultural and Food Chemistry* 2008, 56, (3), 720-729.
- [94] Bro, R.; Kiers, H. A. L., A new efficient method for determining the number of components in PARAFAC models. *Journal of Chemometrics* 2003, 17, (5), 274-286.
- [95] White, J. W., *Composition of American honeys*. U. S. Govt. Print. Off.: Washington,, 1962; p iv, 124 p.