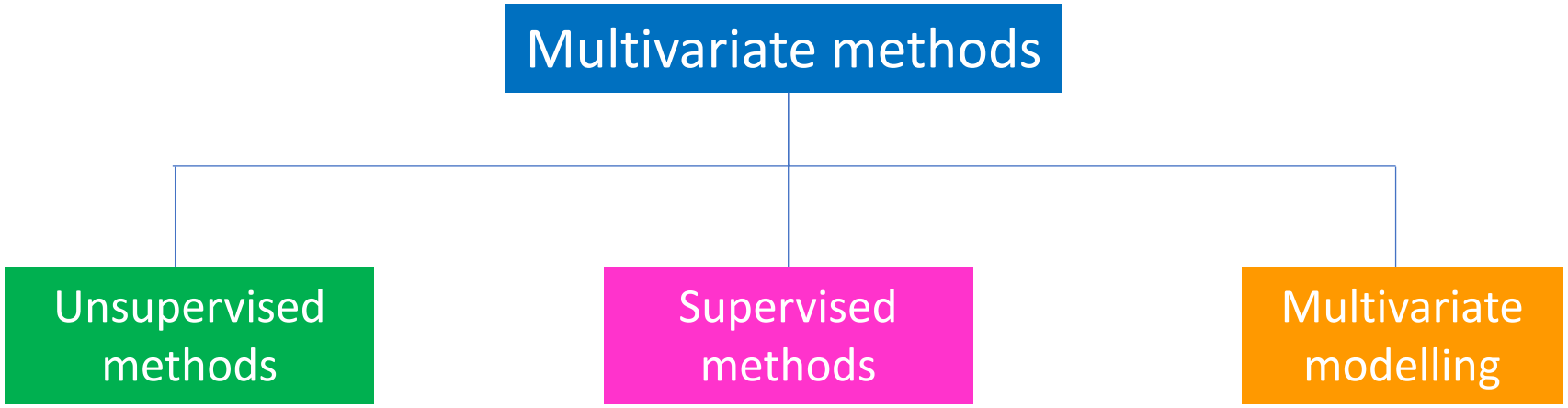
Multivariate methods

Multivariate methods are used to find relationships between sample responses (observations) and variables (properties/features).

A general classification of multivariate methods can be described as follows:

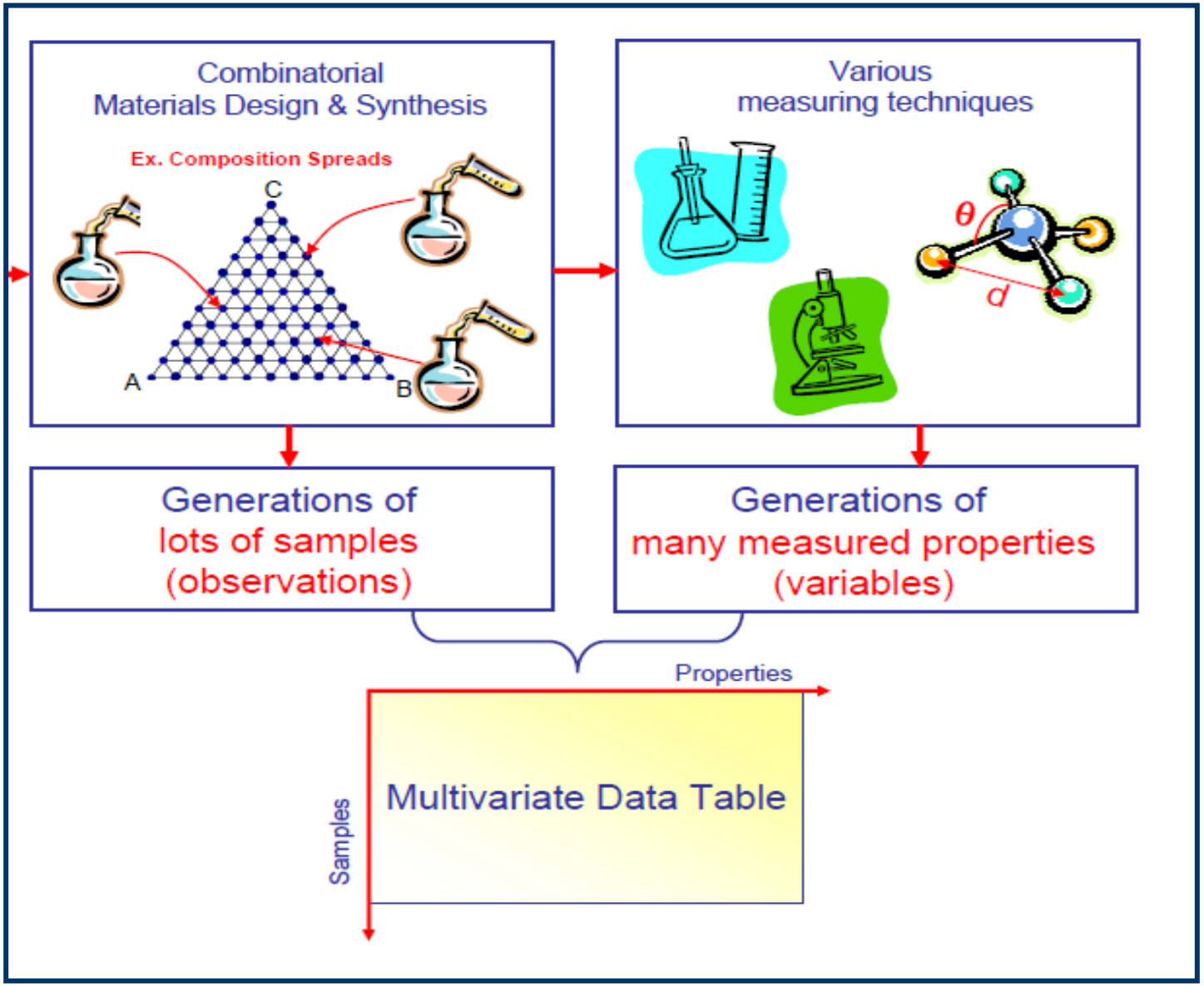


- ✓ Principal Component Analysis (PCA)
- ✓ Cluster Analysis (CA)

- ✓ Linear Discriminant Analysis (LDA)
- ✓ Linear Learning Machine (LLM)
- ✓ Soft Independent Modelling of Class Analogies (SIMCA)

- ✓ Ordinary Least Squares (OLS) Multiple Regression
- ✓ Principal Component Regression (PCR)
- ✓ Partial Least Squares (PLS) Regression

Multivariate data can be obtained in several contexts, including design and synthesis of materials and analysis based on complex methods:



A simple example of a multivariate data set is described in the following table, reporting the concentrations (expressed as ppm) of Cu, Mn, Cl, Br and I in 9 hair samples:

properties / features / variables

samples / objects / observations

Hair No.	Cu	Mn	Cl	Br	I
1	9.2	0.3	1730	12	3.6
2	12.4	0.39	930	50	2.3
3	7.2	0.32	2750	65.3	3.4
4	10.2	0.36	1500	3.4	5.3
5	10.1	0.5	1040	30.2	1.9
6	6.5	0.2	2490	90	4.6
7	5.6	0.29	2940	88	5.6
8	11.8	0.42	867	43.1	1.5
9	8.5	0.25	1620	5.2	6.2

In general terms a multivariate data table can be represented as follows:

Samples (Objects)	Variables (features)					
	1	**2**	**.....**	**j**	**.....**	**p**
1	x_{11}	x_{12}	....	x_{1j}	....	x_{1p}
2	x_{21}	x_{22}	....	x_{2j}	....	x_{2p}
.	.	.	.	.	.	.
.	.	.	.	.	.	.
i	x_{i1}	x_{i2}	....	x_{ij}	....	x_{ip}
.	.	.	.	.	.	.
.	.	.	.	.	.	.
n	x_{n1}	x_{n2}	....	x_{nj}	....	x_{np}

Data matrix **X** will have n rows, corresponding to samples (objects), and p columns, corresponding to variables (features), thus it will be a n × p matrix.

This configuration is called R-mode; in this case covariance (**S**) and correlation (**R**) matrices are both p × p matrices.

An alternative configuration, called Q-mode, has samples in columns and variables in rows, thus a transposed matrix $\mathbf{X}^T$ (p × n) is obtained and **S** and **R** matrices are both n × n matrices.

Matrix **X** can be represented as a set of n points in a p-dimensional space.
The corresponding centroid is represented by the row vector of means:

$$\bar{\mathbf{x}}^T = \left[\bar{x}_1,\ \bar{x}_2\ ,\ldots\ldots,\ \bar{x}_j,\ldots\ldots,\ \bar{x}_p\right] \quad \text{where} \quad \bar{x}_j = n^{-1}\sum_{i=1}^{n} x_{ij}$$

Vector components are thus the means of values included in each column of matrix **X**.

In matricial notation the row vector of means can be expressed as:

$$\underset{(1\times p)}{\bar{\mathbf{x}}^{\mathbf{T}}} = \underset{(1\times n)}{n^{-1}\mathbf{1}^{\mathbf{T}}}\ \underset{(n\times p)}{\mathbf{X}}$$

where vector 1^T is a $(1 \times n)$ row vector whose terms are all equal to 1:

$$\bar{\mathbf{x}}^{\mathbf{T}} = n^{-1}\ [1, 1, \ldots\ldots, 1] \begin{bmatrix} x_{11} & x_{12} & \ldots. & x_{1j} & \ldots. & x_{1p} \\ x_{21} & x_{22} & \ldots. & x_{2j} & \ldots. & x_{2p} \\ . & . & . & . & . & . \\ x_{i1} & x_{i2} & \ldots. & x_{ij} & \ldots. & x_{ip} \\ . & . & . & . & . & . \\ x_{n1} & x_{n2} & \ldots. & x_{nj} & \ldots. & x_{np} \end{bmatrix}$$

Covariance matrix S

Covariance matrix **S** describes the dispersion of data in the p-dimensional space:

$$\boldsymbol{S} = \begin{bmatrix} \sigma^2_1 & \operatorname{cov}(x_1,x_2) & \dots & \dots & \operatorname{cov}(x_1,x_p) \\ \operatorname{cov}(x_2,x_1) & \sigma^2_2 & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ \operatorname{cov}(x_p,x_1) & \dots & \dots & \dots & \sigma^2_p \end{bmatrix}$$

In particular, **S** is a symmetric matrix of rank p (p × p) in which the main diagonal includes variances of variables, whereas other terms correspond to covariances between variables.

Correlation matrix R

Correlation matrix **R** includes linear correlation coefficients between variables:

$$\mathbf{R}=\begin{bmatrix} 1 & r_{12} & \dots & \dots & r_{1p} \\ r_{21} & 1 & . & . & . \\ . & . & 1 & . & . \\ . & . & . & 1 & . \\ r_{p1} & \dots & \dots & \dots & 1 \end{bmatrix}$$

Diagonal values are all equal to 1, since they correspond to correlation coefficients of each variable with itself.

The generic element r_{hk} corresponds to:

$$r_{hk} = \frac{\mathrm{cov}(hk)}{(V_{hh}V_{kk})^{1/2}}$$ with $-1 \leq r_{hk} \leq +1$

Notably, **R** corresponds to the covariance matrix of standardized data: $\frac{x_{ij} - \bar{x}_j}{s_{jj}}$

Principal Component Analysis

Principal Component Analysis (PCA), that is currently one of the most used techniques of multivariate analysis, was proposed by Karl Pearson in 1901 and then developed in its current form by the American statistician Harold Hotelling in 1933.

As a first application, PCA is used to simplify original data, basically by reducing the number of physically measured variables, eventually correlated, into new latent variables, called *principal components*, that are not correlated (i.e., they are orthogonal), can be easily interpreted and are able to synthesize the information embedded in the original data.

Principal components (PC) are linear combinations of the original variables.

Given a column vector $\mathbf{x_i}$, whose terms are values observed for the p variables in the i-th sample (i.e., values reported in the i-th row of the **X** matrix), the value of the k-th principal component for the i-th sample, z_{ik} or PC_{ik}, can be generally expressed as:

$$z_{ik} = \boldsymbol{\alpha}_{\mathbf{k}}^{\mathbf{T}} \mathbf{x_i} = \alpha_{k1} x_{i1} + \alpha_{k2} x_{i2} + \ldots\ldots + \alpha_{kp} x_{ip} = \sum_{j=1}^{p} \alpha_{kj} x_{ij}$$

where:

$$\boldsymbol{\alpha}_{\mathbf{k}}^{\mathbf{T}} = \left[\alpha_{k1}, \alpha_{k2}, \ldots\ldots, \alpha_{kp}\right]$$

The mean of values assumed by the k-th principal component for the n samples can thus be calculated as follows:

$$\bar{z}_k = \frac{1}{n}\sum_{i=1}^{n} z_{ik} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p} \alpha_{kj} x_{ij}$$

Using matricial notation, this mean can be expressed as:

$$\bar{z}_k = n^{-1}\mathbf{1}^{\mathrm{T}}\mathbf{z}_k$$

Where $\mathbf{1}^{\mathrm{T}}$ is a row vector including n times the number 1 and $\mathbf{z_k}$ is a column vector whose components are represented by z_{ik} values, i.e., values assumed by the k-th principal component for each sample. This vector can be thus represented as follows:

$$\mathbf{z_k} = \begin{bmatrix} z_{1k} \\ z_{2k} \\ . \\ z_{ik} \\ . \\ z_{nk} \end{bmatrix}$$

An alternative expression can be used for the average value of the k-th principal component:

$$\bar{z}_k = \frac{1}{n}\sum_{i=1}^{n} z_{ik} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p} \alpha_{kj} x_{ij} = \sum_{j=1}^{p} \alpha_{kj} \sum_{i=1}^{n} \frac{x_{ij}}{n} = \boldsymbol{\alpha}_k^{\mathrm{T}} \bar{\mathbf{X}}$$

where $\bar{\mathbf{X}}$ is a column vector whose components correspond to means of variable values obtained for different samples:

$$\bar{\mathbf{X}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_j \\ \vdots \\ \bar{x}_p \end{bmatrix}$$

The variance of the k-th principal component can thus be calculated as follows:

$$\mathrm{Var}\,(z_k) = \frac{1}{n-1}\sum_{i=1}^{n}\left(z_{ik} - \bar{z}_k\right)^2 =$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\left[\boldsymbol{\alpha}_k\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)^{\mathrm{T}}\boldsymbol{\alpha}_k^{\mathrm{T}}\right] =$$

$$= \boldsymbol{\alpha}_k^{\mathrm{T}}\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)^{\mathrm{T}}\right]\boldsymbol{\alpha}_k = \boldsymbol{\alpha}_k^{\mathrm{T}}\mathbf{S}\boldsymbol{\alpha}_k$$

According to the PCA principle, the first principal component, indicated as z_1, is the one whose variance is maximum:

$$\text{var}\, z_1 = \boldsymbol{\alpha}_1^{\mathrm{T}} \mathbf{S} \boldsymbol{\alpha}_1 \equiv \max$$

under the constrain $\boldsymbol{\alpha}_1^{\mathrm{T}} \boldsymbol{\alpha}_1 = 1$

In other words, the norm of vector α_1 is equal to 1.

It is worth noting that once a vector α_1 able to maximize z_1 variance is found, this variance could be further increased using an alternative vector $c\alpha_1$, with $c > 1$. Infinite possible solutions would thus be found if the constrain described before was not adopted.

The maximum value for var z_i under the described constrain can be found using the method of Lagrange multipliers, i.e., by maximizing the objective function:

$$L = \boldsymbol{\alpha}_i^{\mathrm{T}} \mathbf{S} \boldsymbol{\alpha}_i - \lambda(\boldsymbol{\alpha}_i^{\mathrm{T}} \boldsymbol{\alpha}_i - 1)$$

where λ is the Lagrange multiplier.

The derivative of L with respect to vector $\boldsymbol{\alpha}_i^{\mathrm{T}}$ has to be equalized to 0: $\dfrac{\partial L}{\partial \boldsymbol{\alpha}_i^{\mathrm{T}}} = 0$

According to the rules for vectorial derivation the previous equation can be written as:

$$2\mathbf{S}\boldsymbol{\alpha}_i - 2\lambda\boldsymbol{\alpha}_i = 0 \quad \Rightarrow \quad \mathbf{S}\boldsymbol{\alpha}_i - \lambda\boldsymbol{\alpha}_i = 0$$

Vectors $\boldsymbol{\alpha}_1$ solving the equation are called *eigenvectors* of matrix S, whereas the corresponding λ_i values are called *eingevalues*.

If the first eigenvector and eigenvalue are introduced in the equation, the following equations can be obtained:

$$\mathbf{S}\boldsymbol{\alpha}_1 - \lambda_1\boldsymbol{\alpha}_1 = 0 \quad \Rightarrow \quad \mathbf{S}\boldsymbol{\alpha}_1 = \lambda_1\boldsymbol{\alpha}_1 \quad \Rightarrow \quad \boldsymbol{\alpha}^{\mathrm{T}}{}_1\mathbf{S}\,\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}^{\mathrm{T}}{}_1\,\boldsymbol{\lambda}_1\,\boldsymbol{\alpha}_1 = \lambda_1$$

Since $\boldsymbol{\alpha}^{\mathrm{T}}{}_1\ \mathbf{S}\ \boldsymbol{\alpha}_1$ corresponds to Var(z_1), the first eigenvalue, λ_1, is also the variance of the first principal component.

Once the first principal component is obtained, a second component $\mathbf{z_2 = \alpha_2^T x}$,

not correlated with $\mathbf{z_1}$, is calculated to account for most of the remaining variance of data.

A new eingevector, $\boldsymbol{\alpha_2}$, is thus obtained, respecting the constrains $\boldsymbol{\alpha_2^T \alpha_2} = 1$ and $\boldsymbol{\alpha_2^T \alpha_1} = 0$.

The new objective function now contains two Lagrange multipliers, λ and ϕ:

$$L = \boldsymbol{\alpha_2}^{\mathrm{T}} \mathbf{S} \boldsymbol{\alpha_2} - \lambda(\boldsymbol{\alpha_2}^{\mathrm{T}} \boldsymbol{\alpha_2} - 1) - \phi\,(\boldsymbol{\alpha_2}^{\mathrm{T}} \boldsymbol{\alpha_1} - 0)$$

When the derivative of L with respect to vector $\boldsymbol{\alpha_2^T}$ is calculated and equalized to 0, the equation solution is the eigenvector $\boldsymbol{\alpha_2}$ with the corresponding eigenvalue, λ_2.

The procedure is repeated until the total variance of matrix X is accounted for by the eingevalues:

$$\sum_{K=1}^{P} \lambda_K = \text{ total variance of } \mathbf{X}$$

Principal components resulting from the procedure are usually ordered in terms of decreasing variance and the reduction of complexity occurs by limiting the analysis to the most important components in terms of variance.

Data projection in Principal Component Analysis

Generally speaking, the product between a matrix $\mathbf{X}_{(n \times p)}$ and a vector $\mathbf{v}_{(p \times 1)}$, vector $\mathbf{s} = \mathbf{X}\,\mathbf{v}$, can be interpreted, geometrically, as the projection of a set of n points in a p-dimensional space (points whose co-ordinates are the terms of rows in the **X** matrix) on an axis defined by vector **v**.

In a more general case, the n points defined by matrix **X** can be projected in a k-dimensional space, $\mathbf{T}_{(n,k)}$, defined by k new axes, each represented by a column of the $\mathbf{P}_{(p \times k)}$ matrix, according to the matricial equation:

$$\mathbf{T} = \mathbf{X}\,\mathbf{P}$$

$$(n \times k) = (n \times p)\,(p \times k)$$

T is called score matrix, since points represented by matrix **T** in the new space are called scores.

P is called loadings (or eigenvectors) matrix.

If **X** represents the matrix of input data for PCA, **T** corresponds to a matrix reporting in each row the values of principal components for the n samples and **P** is a matrix reporting in each column the coefficients that need to be multiplied by the p original variables to obtain principal components. Note that k, the number of principal components, was equal to p in calculations shown before, yet PCA can also be performed by choosing $k < p$.

Loading matrix **P** can be obtained from the covariance matrix **S** of original data through an operation called diagonalization:

$$\mathbf{S} = \mathbf{P}\,\Lambda\,\mathbf{P}^T$$

In this equation matrix Λ is a diagonal matrix, i.e., it has values different from 0 only along its main diagonal. Importantly, diagonal terms of matrix Λ correspond to eigenvalues, λ_1, λ_2,…, λ_p, ordered in decreasing order.

As explained before, each column in matrix P correspond to a different principal component and each row to a different original variable:

	PC_1	....	PC_k	.....	PC_p
X_1	α_{11}	....	α_{1k}	....	α_{1p}
	....	....	....	....	....
X_j	α_{j1}	....	α_{jk}	....	α_{jp}
	....	....	....	....	....
X_p	α_{p1}	....	α_{pk}	....	α_{pp}

The generic term α_{jk} of the matrix represent the loading of variable X_j in principal component PC_k, i.e., the coefficient referred to that variable in the calculation of the principal component.

It is worth noting that loadings α_{jk} are standardized linear coefficients, i.e., the sum of their squares is equal to 1:

$$-1 \leq \alpha_{jk} \leq +1 \qquad \sum_j \alpha_{jk}^2 = 1$$

A loading α_{jk} with an absolute value close to 1 indicates that the k-th principal component is represented mainly by the j-th original variable.

On the other hand, a value close to 0 indicates that the variable is almost not represented at all in the principal component.

It is worth noting that eigenvalues assuming very low values are reasonably related to variability due to noise or to non relevant information. In this case the corresponding principal components can be eliminated.

When the number of principal components, k, is equal to that of the original variables, p, the projection discussed so far coincides with a simple rotation, thus **P** is called rotation matrix.

In this case all the information initially contained in matrix **X** is kept.

Geometric interpretation of principal components

As already discussed, principal components individuate a new co-ordinates system, so that the maximum variance, corresponding to PC_1, is located on the first axis, and progressively lower variances are located on other axes.

The new co-ordinates, called scores, are the result of linear combinations in which original variables (usually centered or autoscaled) are combined according to loadings.

As an example, the score of the i-th sample for the k-th principal component is:

$$t_{ik} = \sum_{j=1}^{p} \alpha_{jk} x_{ij}$$

In vectorial terms:

$$t_{ik} = \boldsymbol{\alpha}_{\mathbf{k}}^{\mathbf{T}} \, \mathbf{x}_{\mathbf{i}}$$

where both $\boldsymbol{\alpha}_k$ and $\mathbf{x}_i$ are vectors of length p.

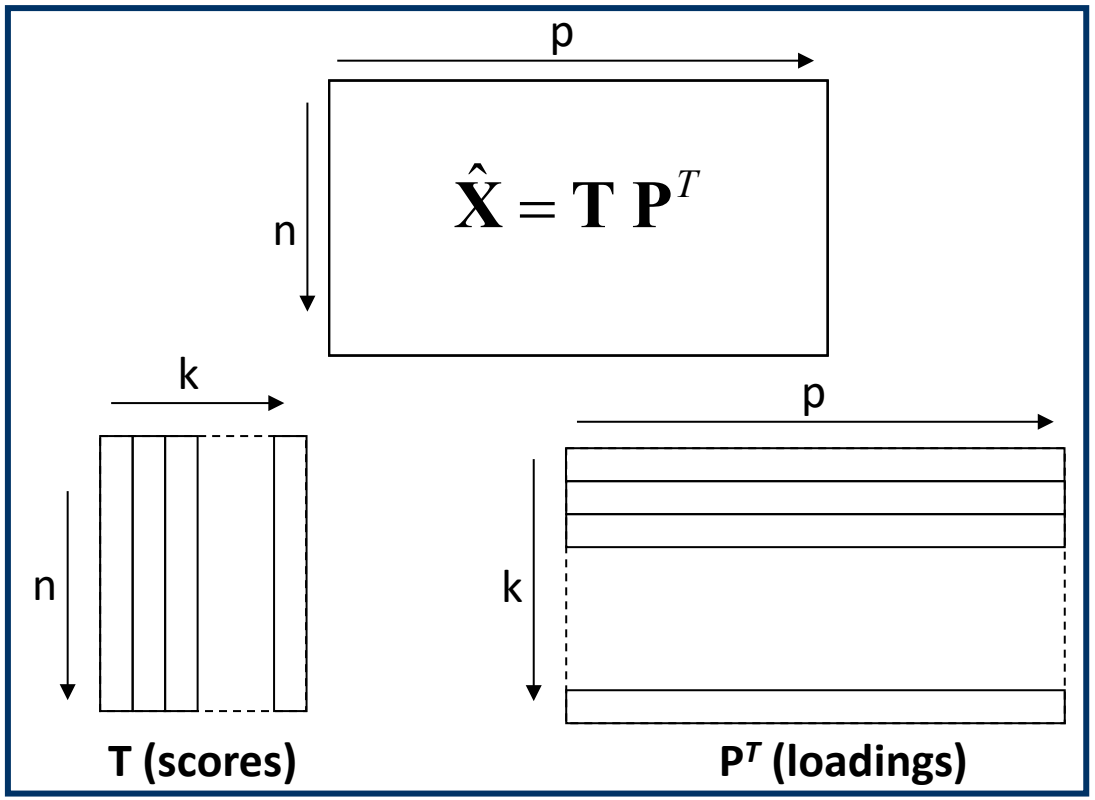
Once principal components have been calculated, the inverse procedure enables the reproduction of the original matrix of data, X, as the product of matrix T and of the transposed version of matrix P:

$$\hat{\mathbf{X}} = \mathbf{T}\,\mathbf{P}^T$$

This equation can be easily demostrated:

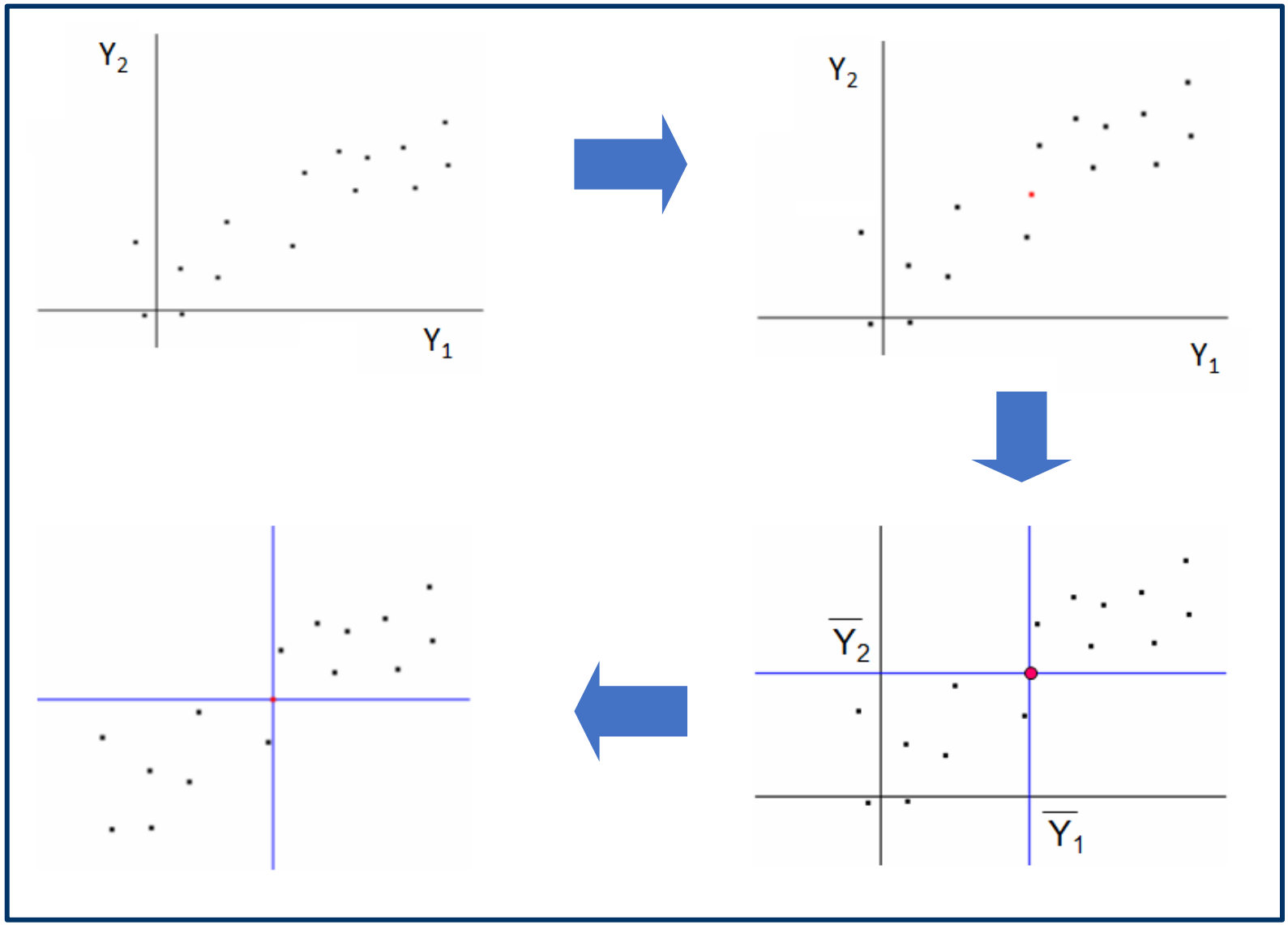
$$\mathbf{T} = \mathbf{X}\,\mathbf{P} \quad \Leftrightarrow \quad \mathbf{T}\,\mathbf{P}^{\mathbf{T}} = \mathbf{X}\,\mathbf{P}\,\mathbf{P}^{\mathbf{T}} \quad \Leftrightarrow \quad \mathbf{T}\,\mathbf{P}^{\mathbf{T}} = \mathbf{X}$$

A graphical representation of the operation is shown in the figure on the right:

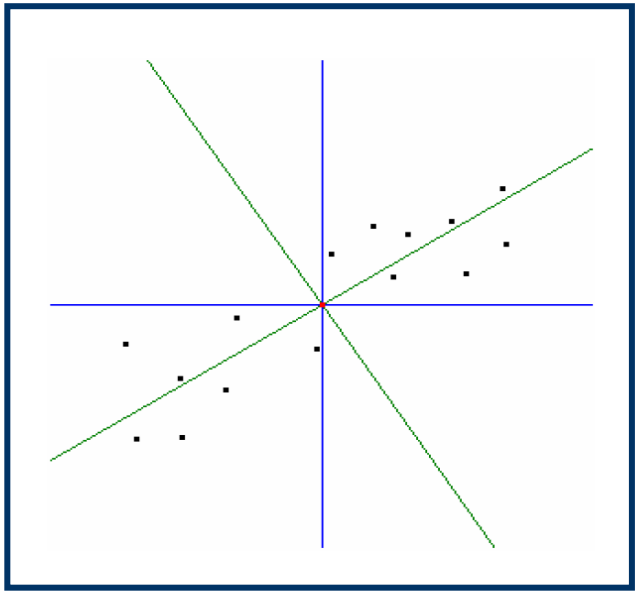


The change in the representation of multivariate data can be easily visualized for bivariate data.

First, the centroid is calculated for original data and is subsequently adopted as the origin of the new reference system (a procedure known as centroidation or mean centering):



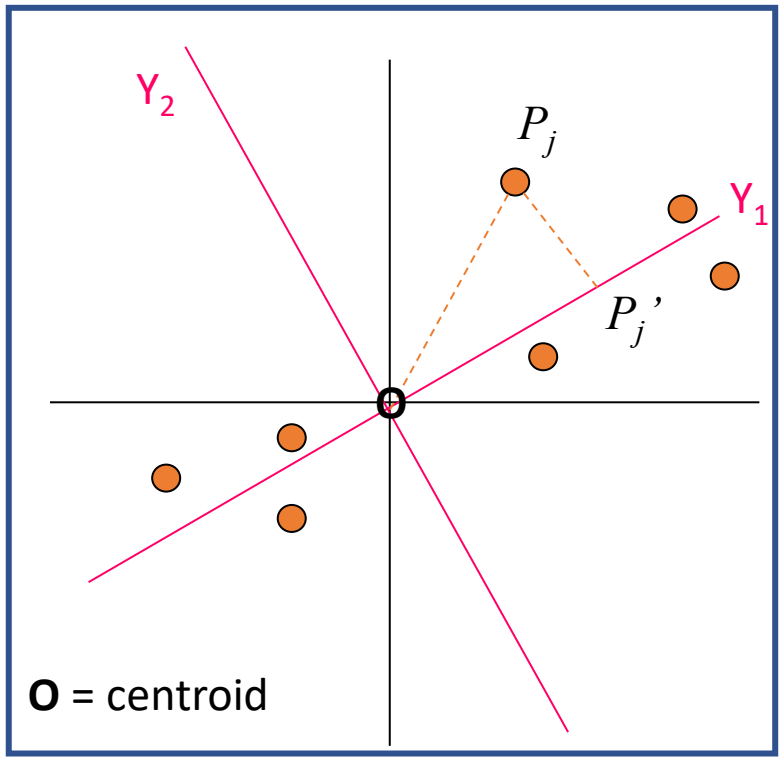
A rotation of axes, aimed at finding the direction characterized by maximum variance, is performed afterwards:



After rotation new co-ordinates are defined for each point with respect to rotated axes.

The optimal rotation is the one able to minimize the sum of squared distances between each point and its projection of the Y_1 axis:

$$\sum_{j=1}^{n} \left(P_j P_j'\right)^2$$

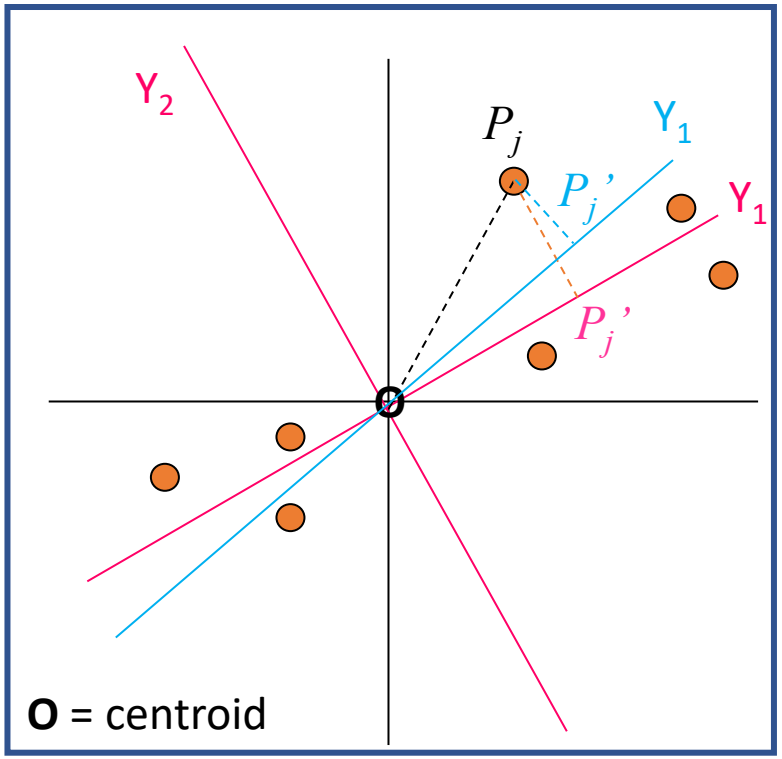


The following equations can be written:

$$(OP_j)^2 = (OP'_j)^2 + (P_jP'_j)^2$$

$$\sum_{j=1}^{n}(OP_j)^2 = \sum_{j=1}^{n}(OP'_j)^2 + \sum_{j=1}^{n}(P_jP'_j)^2$$

$$\frac{1}{n-1}\sum_{j=1}^{n}(OP_j)^2 = \frac{1}{n-1}\sum_{j=1}^{n}(OP'_j)^2 + \frac{1}{n-1}\sum_{j=1}^{n}(P_jP'_j)^2$$



It is worth noting that the choice of axis Y_1 must be able to minimize the second term in the right member of the last equation but, as a consequence, it leads to maximize the first term in the same member.
As a simple graphical demonstration, if the light blue Y_1 axis shown in the figure is considered as an alternative to the purple Y_1 axis, the distance between P_j and the new P_j' projection (the light blue one) is lowered and the distance between the centroid and the new P_j' point is increased.

Notably, the term in the last equation including OP_j' distances corresponds to the variance of projections of points on the Y_1 axis, thus explaining the formulation of the determination of principal components provided by Hotelling.

Main steps of PCA

Based on the considerations made so far, the main steps of Principal Component Analysis are:

1) pretreatment of data matrix $\mathbf{X}_{(n\times p)}$ through centroidation or autoscaling, i.e., centroidation followed by division by standard deviation;

2) calculation of covariance matrix (corresponding to the correlation matrix, if autoscaling of variables is performed preliminarily);

3) calculation of eigenvectors and eigenvalues of the covariance (or correlation) matrix;

4) calculation of the score matrix;

5) graphical representations (scores plot and loadings plot).

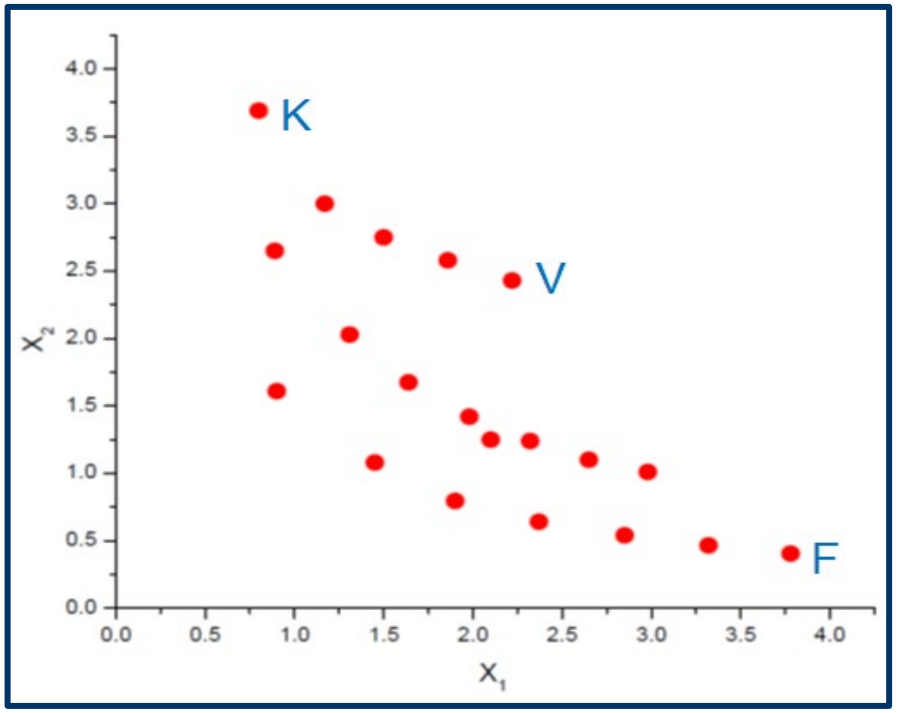
It is worth noting that the use of correlation matrix can be preferred when variables are expressed with different units and/or when they are characterized by quite different variances.

In the latter case principal components based on covariance matrix would emphasize variables having a large variance.

A numerical example of PCA: bivariate data

Let us consider the table on the right, reporting two properties (X_1 and X_2) for 20 chemical elements:

A graphical representation of data can be easily obtained:



Sample NO.	Element	Martynov-Batsanov's Electronegativity (X_1)	Zunger's pseudopotenial core radii sum (X_2)
1	H	2.1	1.25
2	Li	0.9	1.61
3	Be	1.45	1.08
4	B	1.9	0.795
5	C	2.37	0.64
6	N	2.85	0.54
7	O	3.32	0.465
8	F	3.78	0.405
9	Na	0.89	2.65
10	Mg	1.31	2.03
11	Al	1.64	1.675
12	Si	1.98	1.42
13	P	2.32	1.24
14	S	2.65	1.1
15	Cl	2.98	1.01
16	K	0.8	3.69
17	Ca	1.17	3
18	Sc	1.5	2.75
19	Ti	1.86	2.58
20	V	2.22	2.43

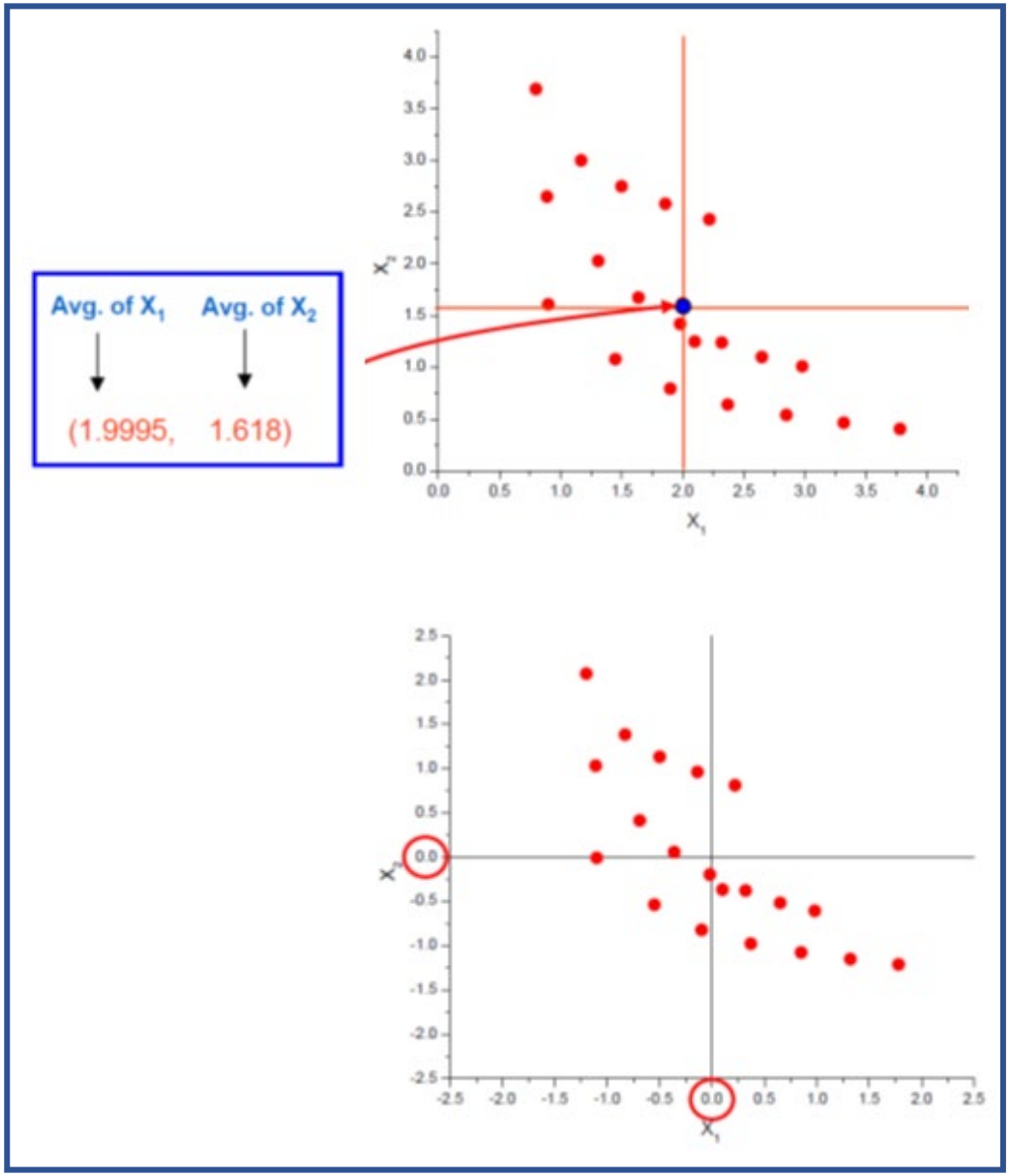
Centroidation of data can be also represented graphically:

$\mathbf{X}(n,p)$

2,1	1,25
0,9	1,61
1,45	1,08
1,9	0,795
2,37	0,64
2,85	0,54
3,32	0,465
3,78	0,405
0,89	2,65
1,31	2,03
1,64	1,675
1,98	1,42
2,32	1,24
2,65	1,1
2,98	1,01
0,8	3,69
1,17	3
1,5	2,75
1,86	2,58
2,22	2,43

$\mathbf{X}_c(n,p)$

0,1005	-0,368
-1,0995	-0,008
-0,5495	-0,538
-0,0995	-0,823
0,3705	-0,978
0,8505	-1,078
1,3205	-1,153
1,7805	-1,213
-1,1095	1,032
-0,6895	0,412
-0,3595	0,057
-0,0195	-0,198
0,3205	-0,378
0,6505	-0,518
0,9805	-0,608
-1,1995	2,072
-0,8295	1,382
-0,4995	1,132
-0,1395	0,962
0,2205	0,812



The covariance matrix **S** is the following: $\begin{bmatrix} 0.6881 & -0.5929 \\ -0.5929 & 0.9026 \end{bmatrix}$

Eingenvalues and eigenvectors can be obtained using the equation shown before:

$$\mathbf{S} = \mathbf{P}\,\Lambda\,\mathbf{P}^{T}$$

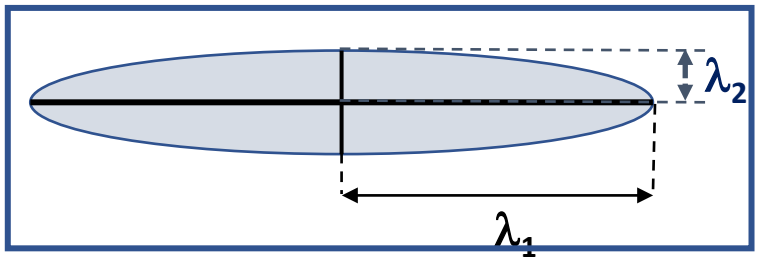
where **P** is an orthonormal matrix and Λ is a diagonal matrix.

Introducing numbers, the equation can be expressed as:

$$\begin{bmatrix} 0.6881 & -0.5929 \\ -0.5929 & 0.9026 \end{bmatrix} = \begin{bmatrix} 0.64108 & -0.76747 \\ -0.76747 & 0.64108 \end{bmatrix} \begin{bmatrix} 1.397805 & 0 \\ 0 & 0.192835 \end{bmatrix} \begin{bmatrix} 0.64108 & -0.76747 \\ -0.76747 & 0.64108 \end{bmatrix}$$

The two eigenvalues are: $\lambda_1 = 1.3978$ and $\lambda_2 = 0.1928$.

From a graphical point of view eigenvectors provide the orientation of the main axes of the covariance ellipse, whereas eigenvalues provide the length of axes:



It is worth noting that the trace of the covariance matrix **S**, i.e., the sum of elements along the main diagonal, correspond to the total variance of the original data, i.e., 1.5907.

$$\mathbf{S} = \begin{bmatrix} 0.6881 & -0.5929 \\ -0.5929 & 0.9026 \end{bmatrix}$$

Variable X_1 contributes to this variance for 0.6881/1.5907 = 43.26%

Variable X_2 contributes to this variance for 0.9026/1.5907 = 56.74%

If the diagonal matrix Λ is considered, its trace corresponds to the total variance explained by principal components and is equal to 1.5906, thus identical (apart from rounding effects) to that of the original data. This is reasonable, since the numbers of original variables and of principal components are the same in this case.

$$\Lambda = \begin{bmatrix} 1.3978 & 0 \\ 0 & 0.1928 \end{bmatrix}$$

Notably, PC_1 and PC_2 contribute for 1.3978/1.5906 = 87.88 % and for 0.1928/1.5906 = 12.12 %, respectively.

It is thus apparent that the transformation of original variables into principal components has led to a different distribution of explained variance, increasing remarkably the one explained by the first component with respect to the one explained by the first original variable.

The matrix of scores, **T**, can be obtained from the $\mathbf{X_c}$ and the **P** ones: $\mathbf{T = X_c\, P}$

PC1	PC2
0,3469	-0,31305
-0,6987	0,84897
0,0606	0,76663
0,5678	0,60397
0,9881	0,34263
1,3726	0,03835
1,7314	-0,2743
2,0724	-0,5889
-1,5033	0,18992
-0,7582	0,26505
-0,2742	0,23936
0,1395	0,1419
0,4956	-0,0036
0,8146	-0,1672
1,0952	-0,3627
-2,3592	-0,4077
-1,5924	-0,2494
-1,189	-0,3424
-0,8277	-0,5097
-0,4818	-0,6898

T

=

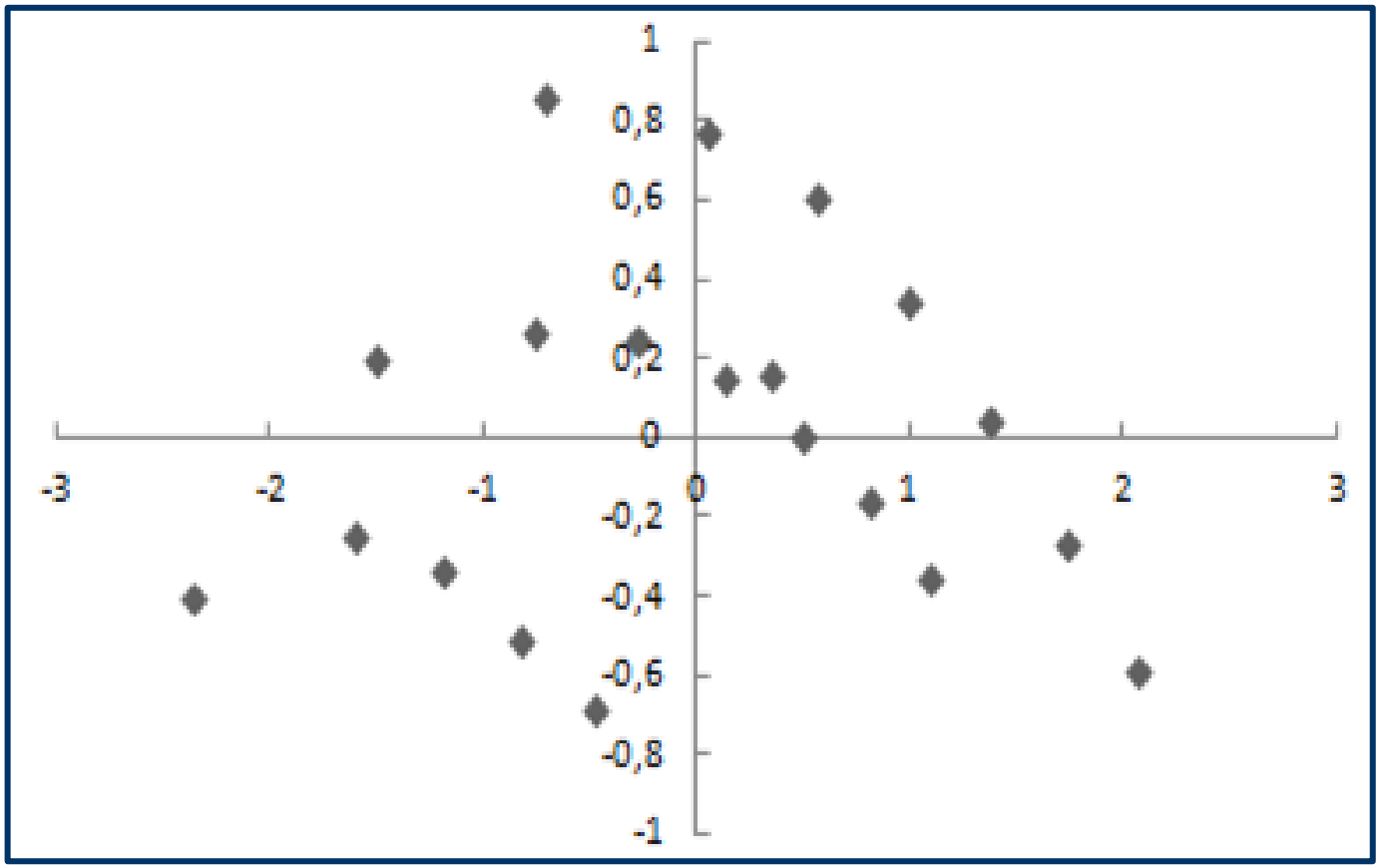
0,1005	-0,368
-1,0995	-0,008
-0,5495	-0,538
-0,0995	-0,823
0,3705	-0,978
0,8505	-1,078
1,3205	-1,153
1,7805	-1,213
-1,1095	1,032
-0,6895	0,412
-0,3595	0,057
-0,0195	-0,198
0,3205	-0,378
0,6505	-0,518
0,9805	-0,608
-1,1995	2,072
-0,8295	1,382
-0,4995	1,132
-0,1395	0,962
0,2205	0,812

$\mathbf{X_c}$

0.64108	-0.76747
-0.76747	0.64108

P

As an example, for the first sample the PC1 score is: 0.1005 * 0.64108 - 0.368*(-0.76747) = 0.3469; the PC2 score is: 0.1005 *(-0.76747) - 0.368 * 0.64108 = -0.31305

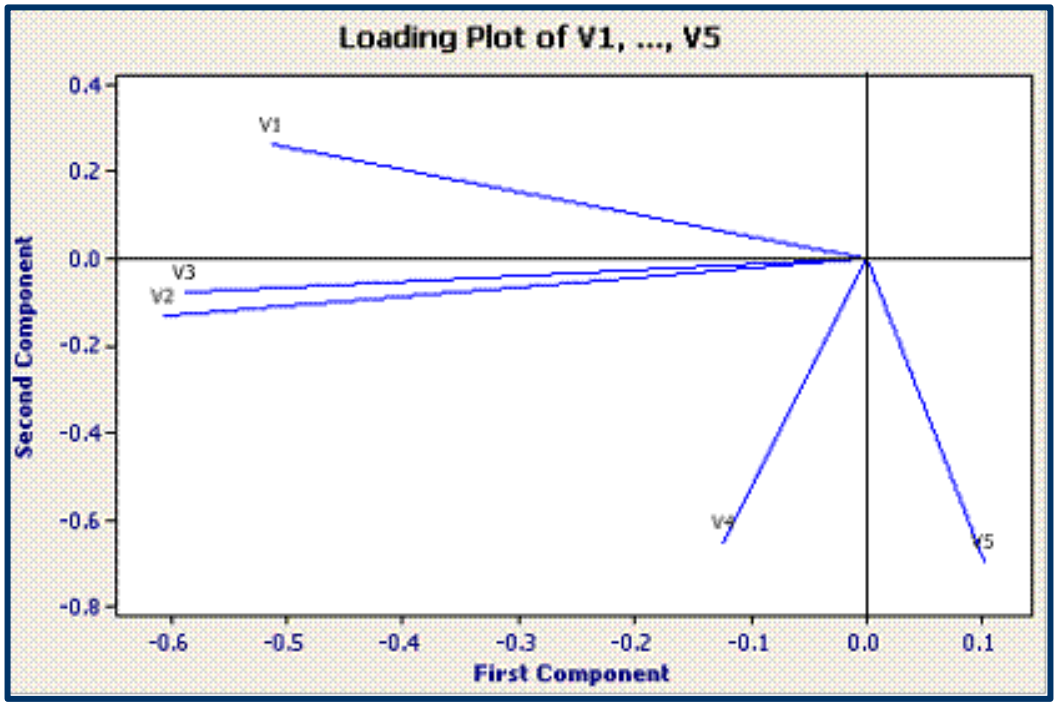


Each row of matrix **T** reports the co-ordinates that each sample will have in the reference system based on the two principal components. Its graphical representation, reported above, is usually called *Score plot*.

It is worth noting that a symmetrical plot with respect to the origin of axes is equivalent to the one shown above. This feature is called mirror effect of PCA.

Loadings plot

As explained before, in PCA loadings represent the contribution of each original variable to a specific principal component. Their graphical representation, usually called loadings plot, can be very informative.

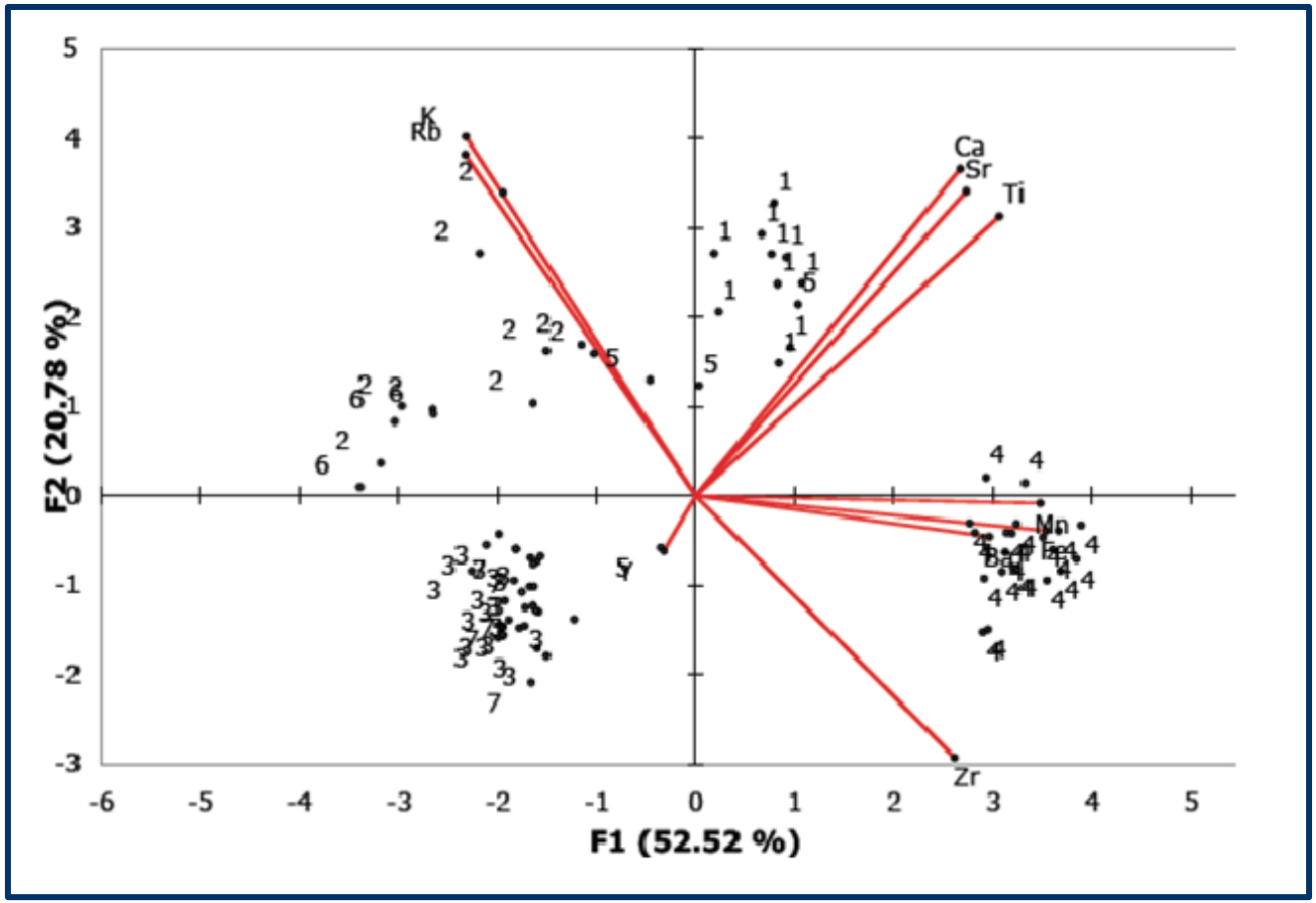


As an example, the plot shown above indicates that variables 1, 2 and 3 have a remarkable (and negative) contribution fo PC_1, whereas variables 4 and 5 contribute to PC_2.

Generally speaking, variables whose points are close to the origin of the loading plot are not relevant for any PC, whereas variables whose points are close to each other are correlated, i.e., they provide a similar information.

Plots including both scores and loadings, called *bi-plots*, can be reproduced after the processing of data based on PCA.

In the following example the scores referred to several samples and the loadings referred to 10 variables, corresponding to the concentrations of different elements, are reported in a bi-plot:



Note that the contribution of each principal component to the total variance is indicated in the axes names.

Choice of the number of principal components

Different criteria can be adopted to select the optimal number of principal components.

1) Predetermined value of the explained variance

According to this approach, the k principal components that explain a cumulative variance of 80-90% can be retained:

$$\frac{\lambda_1 + \lambda_2 + ... + \lambda_k}{\sum_{i=1}^{p} \lambda_i} \approx 80 - 90\%$$

2) Mean eigenvalue

The k principal components whose eigenvalues are greater than the mean eigenvalue are retained:

$$\lambda_1 \geq \lambda_2 \geq ..... \geq \lambda_k \geq \frac{\sum_{i=1}^{p} \lambda_i}{p} = \overline{\lambda}$$

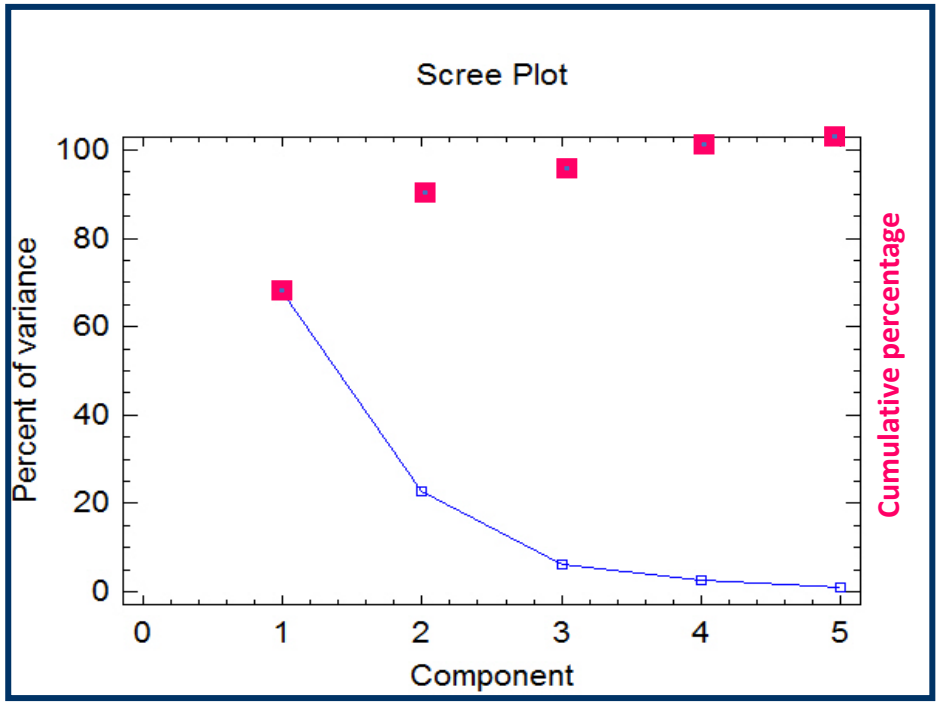
Note that if variables are autoscaled (i.e., the correlation matrix is used for diagonalization) the mean value of eigenvalues is 1. Since the choice of components with eigenvalues greater than 1 is too severe, a threshold of 0.7 is sometimes adopted.

3) Scree plot

The scree plot reports the percentages of variance explained by each principal component.

In the figure on the right cumulative percentages have been also added.

This graphical representation helps the user in choosing which principal component should be considered as the last useful one.



Correlation between principal components and original variables

The degree of correlation existing between the i-th PC, z_i, and the j-th original variable, X_j, is provided by the following correlation coefficient:

$$r(z_i, X_j) = \frac{Cov(z_i, X_j)}{[Var(z_i)Var(X_j)]^{1/2}} = \frac{\lambda_i \alpha_{ij}}{[\lambda_i Var(X_j)]^{1/2}} = \frac{\alpha_{ij}\sqrt{\lambda_i}}{\sqrt{Var(X_j)}}$$

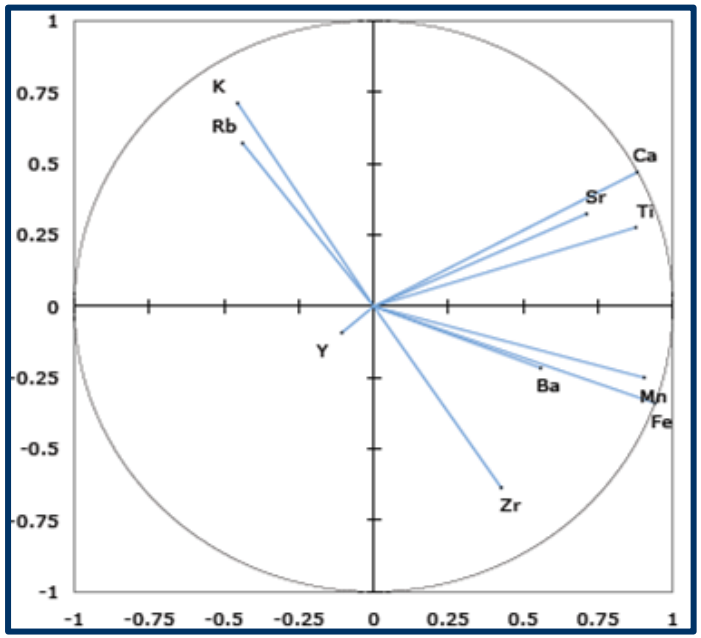
where α_{ij} is the loading of X_j on z_i.

As expected, variables contributing to a PC with a relevant loading (in absolute value) will determine the meaning of that PC.

A graphical representation known as correlation circle enables an easy evaluation of correlations.

Indeed, the circle is drawn with unitary radius in a plot in which each variable corresponds to a point whose co-ordinates are the correlation coefficients between the variable and the principal components on the two axes.

The more a point is close to the circle, the greater is its correlation with the two PCs.



Conceptual considerations on PCA

Due to the approach adopted for the calculation, each principal component represents variations in data due to different intrinsic properties, thus PCs represent macroproperties of the system under study, not directly measurable.

PCA enables the recognition of emerging properties of the system, that can be related to synergic or antagonistic effects of the original variables.

As en example, if some chemical compounds are described contemporarily by a certain number of descriptors (molecular weight, molar volume, etc.) a specific PC in which each of these descriptors is remarkably represented is expected.

This PC becomes a new «macrovariable», whose meaning goes beyond that of the single original variables.

Whether this information is interpreted or not, it represents a new synthetic description of the system under study.

Moreover, not relevant variations or the one caused by experimental noise are not represented in a principal component.

A further numerical example of PCA: data referred to five variables

Let us re-consider the following data matrix, representing the concentrations (ppm) of five elements in nine hair samples:

X: original data matrix

Cu	Mn	Cl	Br	I
9,2	0,3	1730	12	3,6
12,4	0,39	930	50	2,3
7,2	0,32	2750	65,3	3,4
10,2	0,36	1500	3,4	5,3
10,1	0,5	1040	30,2	1,9
6,5	0,2	2490	90	4,6
5,6	0,29	2940	88	5,6
11,8	0,42	867	43,1	1,5
8,5	0,25	1620	5,2	6,2
9,05556	0,33667	1763	43,0222	3,82222
2,32815	0,09152	790,425	33,3582	1,70424

means

standard deviations

A first inspection of variable values and of means and standard deviations indicates that Cl concentrations are quite different in the nine samples. In this case, the autoscaling of data, i.e., the use of correlation matrix, is highly recommended.

1. Centroidation and autoscaling of data

The mean referred to each column is subtracted from all the values of the same column in the original matrix of data, thus obtaining the centered matrix of data:

$\mathbf{X}_c$: centered matrix of data

Cu	Mn	Cl	Br	I
0,14453	-0,03666	-32,9824	-31,0218	-0,22222
3,34454	0,05334	-832,982	6,978208	-1,52222
-1,85547	-0,01666	987,018	22,27821	-0,42222
1,14454	0,02334	-262,982	-39,6218	1,47778
1,04454	0,16334	-722,982	-12,8218	-1,92222
-2,55547	-0,13666	727,018	46,97821	0,77778
-3,45547	-0,04666	1177,02	44,97821	1,77778
2,74454	0,08334	-895,982	0,078208	-2,32222
-0,55547	-0,08666	-142,982	-37,8218	2,37778

Values obtained in each column are divided by the standard deviation referred to that column in the original data, thus obtaining the autoscaled matrix of data:

$\mathbf{X}_a$: autoscaled matrix of data

Cu	Mn	Cl	Br	I
0,06208	-0,400626	-0,04173	-0,93003	-0,13036
1,43656	0,582819	-1,05384	0,209207	-0,89307
-0,79697	-0,182083	1,248718	0,667901	-0,2477
0,49161	0,255004	-0,33271	-1,18786	0,867035
0,44865	1,784807	-0,91468	-0,3844	-1,12775
-1,09764	-1,493343	0,919781	1,408407	0,456345
-1,48421	-0,509898	1,489095	1,348447	1,043045
1,17885	0,910634	-1,13355	0,002345	-1,36243
-0,23859	-0,946984	-0,18089	-1,1339	1,395065

2. Calculation of Correlation Matrix, R, of original data

The second step is the calculation of the covariance matrix for autoscaled data, that is actually equivalent to the correlation matrix of original data:

	Cu	Mn	Cl	Br	I
Cu	1	0,69738	-0,94981	-0,53995	-0,64524
Mn	0,697376	1	-0,692	-0,29135	-0,74884
Cl	-0,94981	-0,692	1	0,61308	0,58073
Br	-0,53995	-0,29135	0,61308	1	-0,04541
I	-0,64524	-0,74884	0,58073	-0,04541	1

3. Calculation of eigenvalues and eigenvectors

Calculation of eigenvalues and eigenvectors is based on the matricial equation:

$$\mathbf{R} = \mathbf{P}\,\Lambda\,\mathbf{P}^T$$

The resulting diagonal matrix Λ, including eigenvalues, is the following:

explained variances

$$\Lambda = \begin{bmatrix} 3.38763 & 0 & 0 & 0 & 0 \\ 0 & 1.1338982 & 0 & 0 & 0 \\ 0 & 0 & 0.301181 & 0 & 0 \\ 0 & 0 & 0 & 0.132766 & 0 \\ 0 & 0 & 0 & 0 & 0.044528 \end{bmatrix}$$

$\Rightarrow$ 3.38763 /5 = 67.75%

$\Rightarrow$ 1.13390 /5 = 22.68%

$\Rightarrow$ 0.30118 /5 = 6.03%

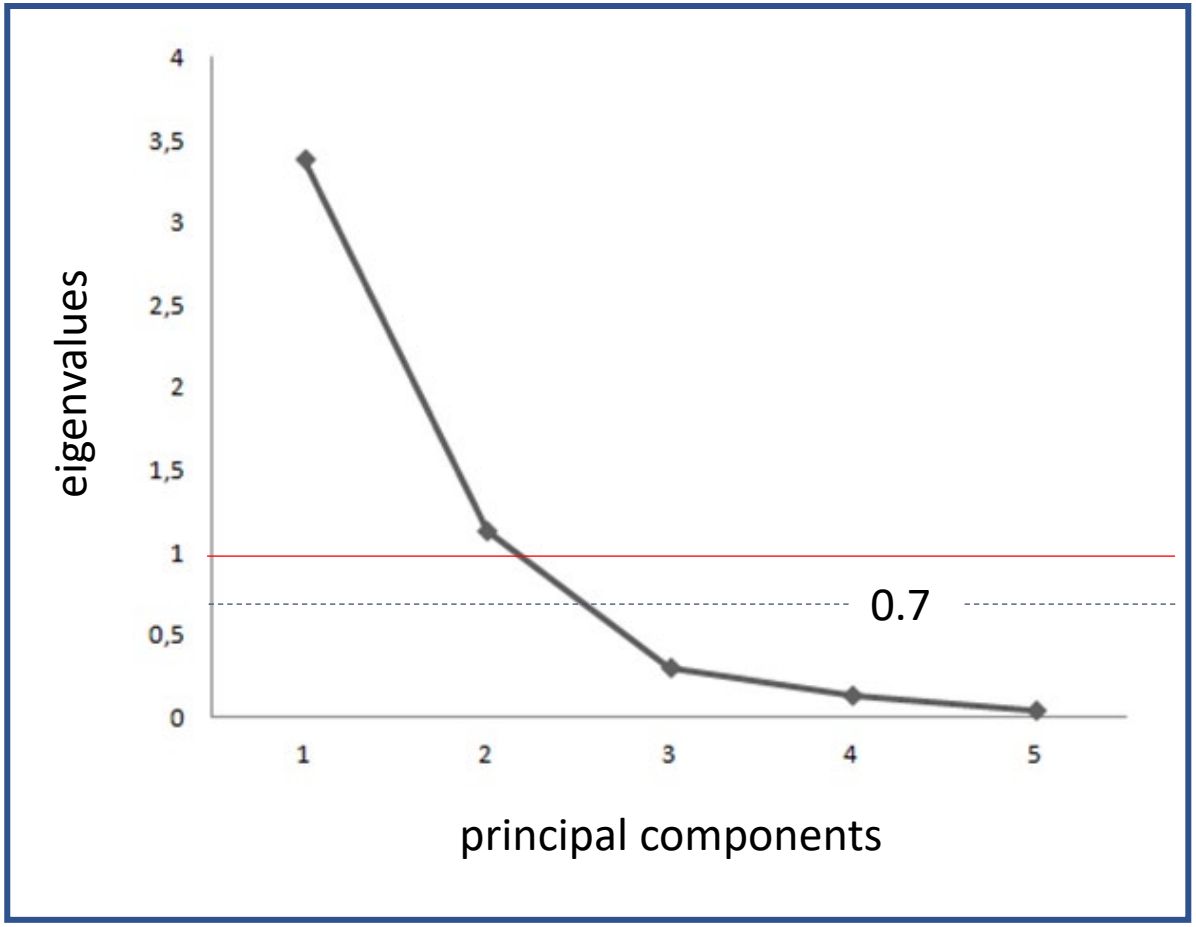
$\Rightarrow$ 0.13277 /5 = 2.66%

$\Rightarrow$ 0.04453 /5 = 0.89%

It is worth noting that when correlation matrix is used the trace of matrix Λ, i.e., the sum of eigenvalues, is equal to the number of variables p:

$$\sum_{i=1}^{p} \lambda_i = tr\,\Lambda = \mathrm{p}$$

The scree plot referred to the five eigenvalues is the following:



Considering both the criteria described before for the selection of principal components based on eigenvalues, i.e., being greater than 1 or 0.7, the number of components to be retained is 2.

The matrix of eigenvectors (loadings) **P** is the following:

	PC_1	PC_2	PC_3	PC_4	PC_5
Cu	-0,517021	0,085173	-0,439943	0,219454	-0,695504
Mn	-0,463559	-0,2639	0,764317	0,357606	-0,058355
Cl	0,514982	-0,17136	0,337149	-0,2954	-0,710282
Br	0,298023	-0,74933	-0,309745	0,502054	0,041036
I	0,404696	0,576401	0,112508	0,696161	-0,08176

Consequently, principal components can be expressed as:

$PC_1 =$ - 0.51702 Cu - 0.46356 Mn + 0.51498 Cl + 0.29802 Br + 0.404696 I

$PC_2 =$ 0.085173 Cu - 0.2639 Mn - 0.1714 Cl - 0.7493 Br + 0.576401 I

$PC_3 =$ - 0.43994 Cu + 0.76432 Mn + 0.33715 Cl - 0.3097 Br + 0.112508 I

$PC_4 =$ 0.219454 Cu + 0.35761 Mn - 0.2954 Cl + 0.50205 Br + 0.696161 I

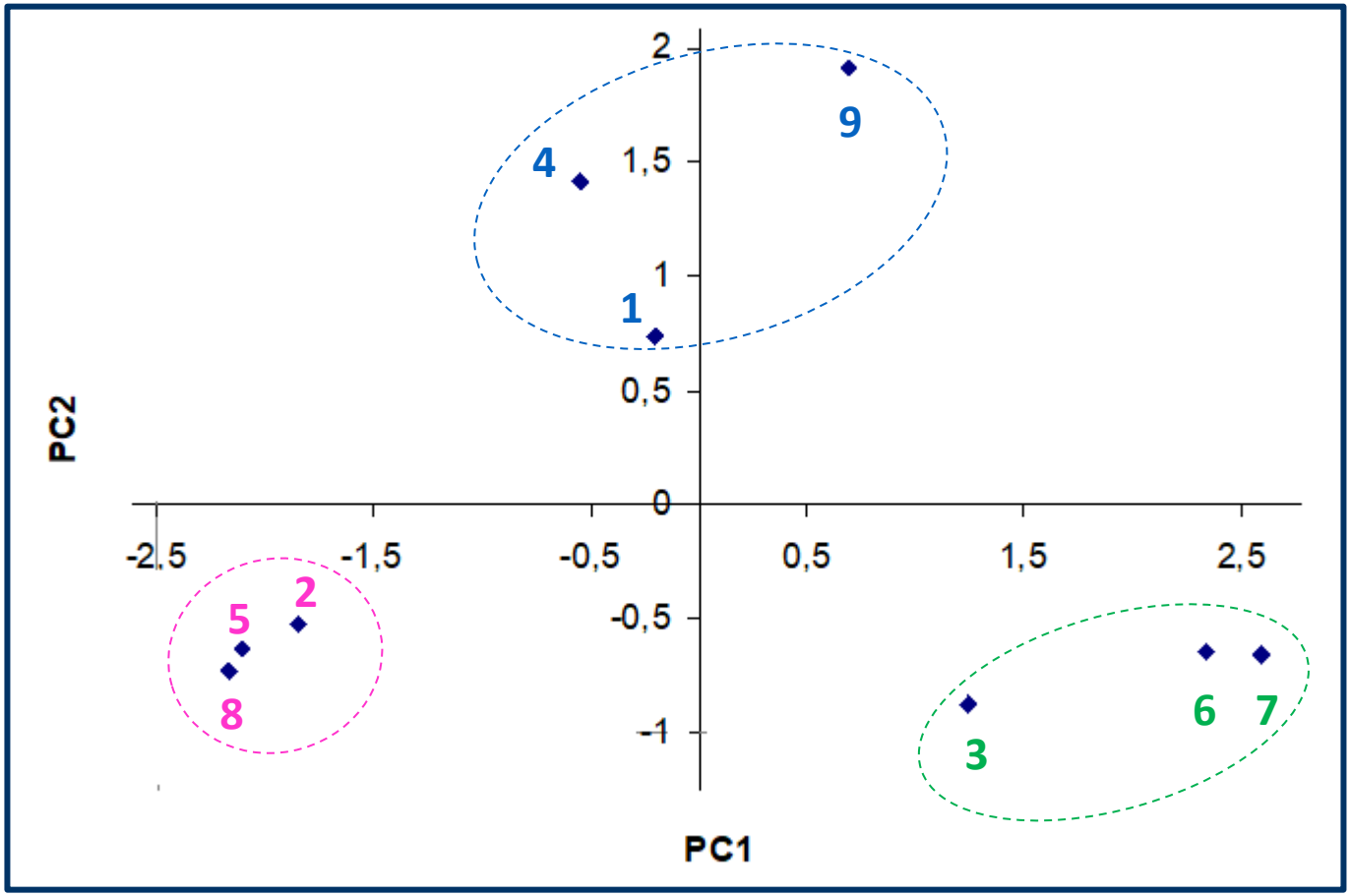
$PC_5 =$ - 0.6955 Cu -0.05835 Mn - 0.7103 Cl + 0.04104 Br - 0.08176 I

4. Calculation of scores matrix

The matrix of scores **T** can be obtained from the equation $\mathbf{T} = \mathbf{X}_a \mathbf{P}$:

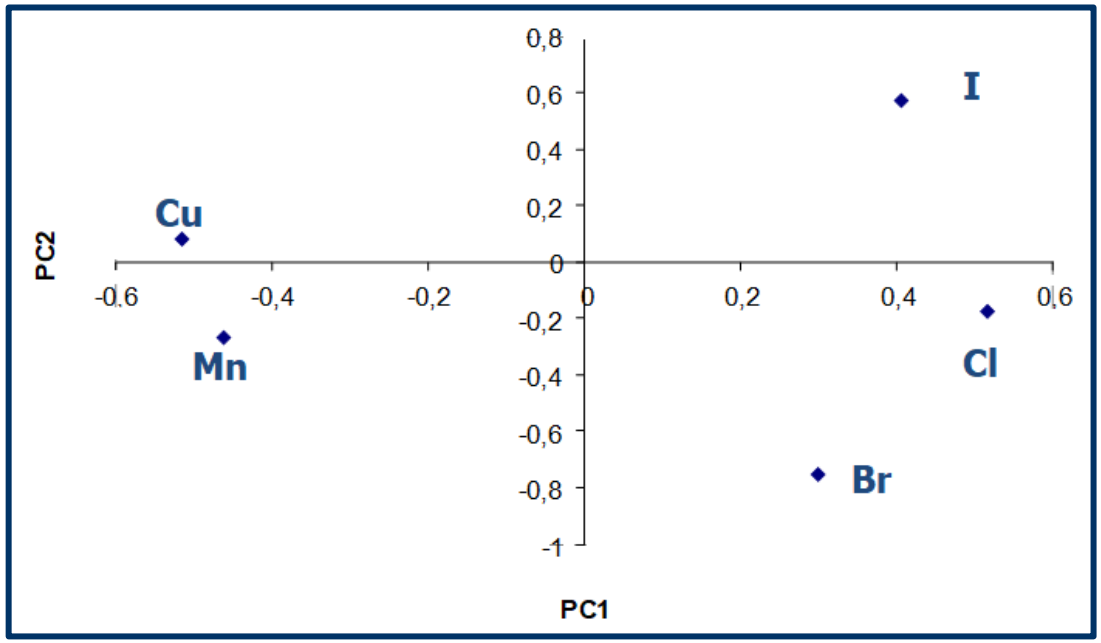
	PC_1	PC_2	PC_3	PC_4	PC_5
1	-0,197798	0,739932	-0,074179	-0,67499	-0,017668
2	-1,854685	-0,52239	-0,707128	0,318296	-0,203019
3	1,238331	-0,87706	0,397711	-0,44599	-0,274362
4	-0,546844	1,42145	0,331934	0,304583	-0,240111
5	-2,101322	-0,63806	0,850578	0,028829	0,309915
6	2,337843	-0,64933	-0,733289	-0,02182	0,217736
7	2,594577	-0,65625	0,46497	0,455188	-0,025593
8	-2,166046	-0,73273	-0,358797	-0,02809	0,043592
9	0,695831	1,914371	-0,171645	0,064341	0,189093

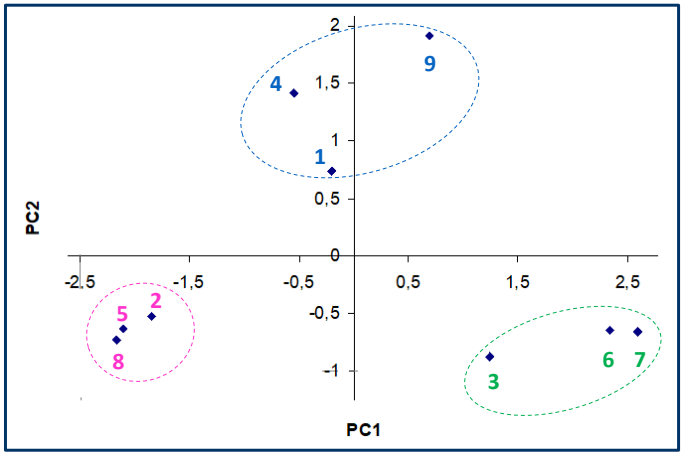
The consequent scores plot for the first two principal components is the following:



Three different clusters of samples can be observed in the score plot.

The loadings plot for the first two principal components is the following:



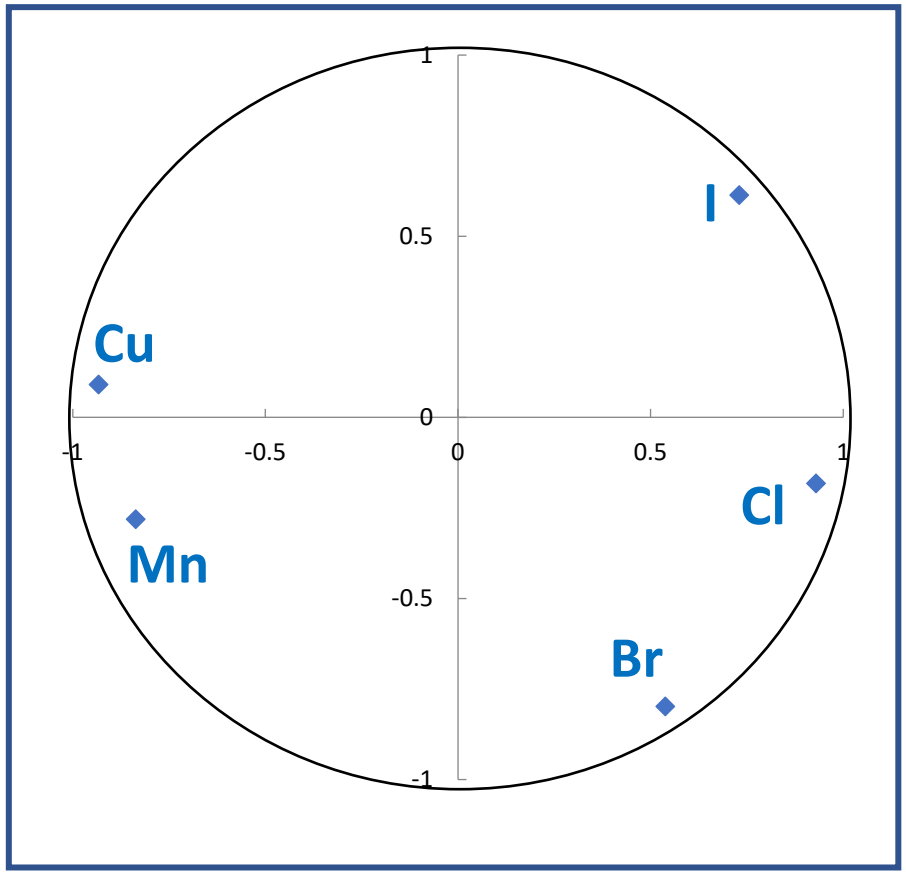


The plot indicates that samples 3, 6 and 7 share relatively high Cl and Br concentrations, whereas samples 2, 5 and 8 are particularly rich in Mn and Cu and samples 4 and 9 are rich in I.

This outcome can be inferred from the original dataset (but only because variables are in a limited number).

	Cu	Mn	Cl	Br	I
1	9,2	0,3	1730	12	3,6
2	12,4	0,39	930	50	2,3
3	7,2	0,32	2750	65,3	3,4
4	10,2	0,36	1500	3,4	5,3
5	10,1	0,5	1040	30,2	1,9
6	6,5	0,2	2490	90	4,6
7	5,6	0,29	2940	88	5,6
8	11,8	0,42	867	43,1	1,5
9	8,5	0,25	1620	5,2	6,2

The correlation circle is obtained using formulas described before:



In this case Cu, Mn and Cl are remarkably correlated with PC1 (the first two elements negatively, the third positively).
Intermediate correlation coefficients towards both principal components are observed for I and Br.

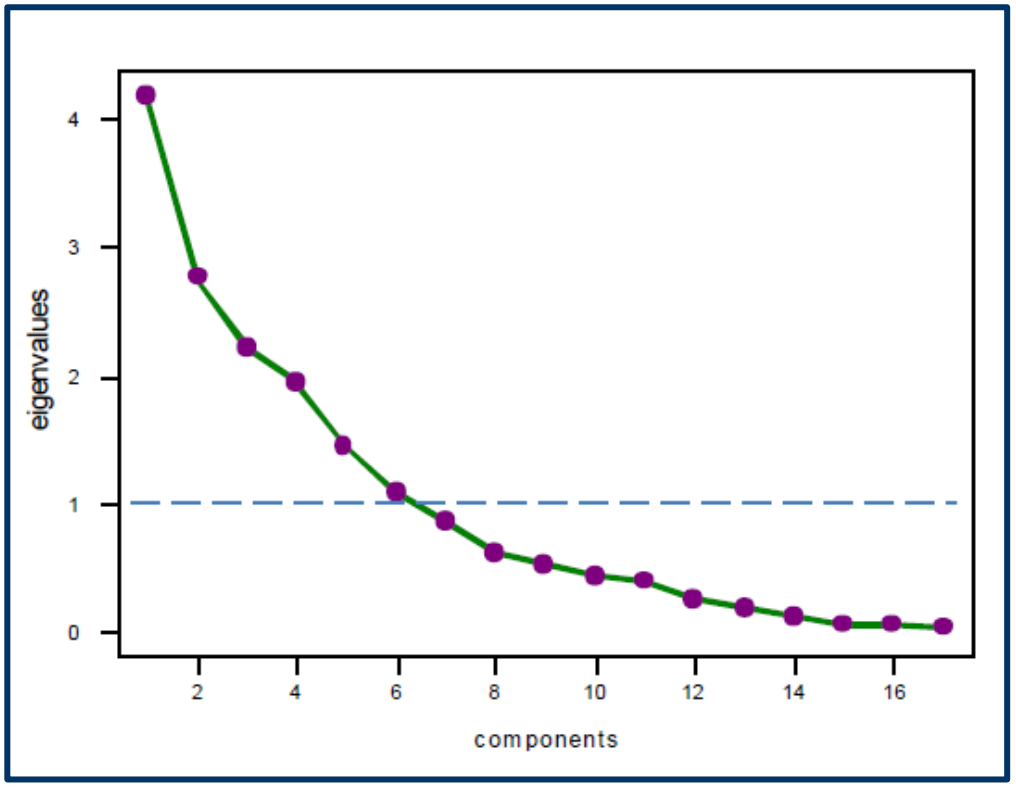
A further numerical example of PCA: data referred to 18 variables

38 wine samples were subjected to the analysis of 17 trace elements and the following data matrix was obtained, including an evaluation of aroma as the 18-th variable:

ID	*Cd*	*Mo*	*Mn*	*Ni*	*Cu*	*Al*	*Ba*	*Cr*	*Sr*	*Pb*	*B*	*Mg*	*Si*	*Na*	*Ca*	*P*	*K*	*Aroma*
1	.005	.044	1.51	.122	.83	.982	.387	.029	1.23	.561	2.63	128	17.3	66.8	80.5	150	1130	3.3
2	.055	.16	1.16	.149	.066	1.02	.312	.038	.975	.697	6.21	193	19.7	53.3	75	118	1010	4.4
3	.056	.146	1.1	.088	.643	1.29	.308	.035	1.14	.73	3.05	127	15.8	35.4	91	161	1160	3.9
4	.063	.191	.959	.38	.133	1.05	.165	.036	.927	.796	2.57	112	13.4	27.5	93.6	120	924	3.9
5	.011	.363	1.38	.16	.051	1.32	.38	.059	1.13	1.73	3.07	138	16.7	76.6	84.6	164	1090	5.6
6	.05	.106	1.25	.114	.055	1.27	.275	.019	1.05	.491	6.56	172	18.7	15.7	112	137	1290	4.6
.....																		
.....																		
.....																		
32	.084	.266	1.28	.087	.071	1.14	.158	.049	.794	1.3	3.77	143	19.7	39.1	128	146	1230	4.2
33	.069	.183	1.94	.07	.095	.465	.225	.037	1.19	.915	2	123	4.57	7.51	69.4	123	943	3.3
34	.087	.208	1.76	.061	.099	.683	.087	.042	.168	1.33	5.04	92.9	6.96	12	56.3	157	949	6.8
35	.074	.142	2.44	.051	.052	.737	.408	.022	1.16	.745	3.94	143	6.75	36.8	67.6	81.9	1170	5
36	.084	.171	1.85	.088	.038	1.21	.263	.072	1.35	.899	2.38	130	6.18	101	64.4	98.6	1070	3.5
37	.106	.307	1.15	.063	.051	.643	.29	.031	.885	1.61	4.4	151	17.4	7.25	103	177	1100	4.3
38	.102	.342	4.08	.065	.077	.752	.366	.048	1.08	1.77	3.37	145	5.33	33.1	58.3	117	1010	5.2

The following results and scree plot were obtained from PCA:

ID	*eigenvalue*	*E.V.%*	*C.E.V.%*
1	4.1785	24.6	24.6
2	2.7468	16.2	40.7
3	2.2098	13.0	53.7
4	1.9349	11.4	65.1
5	1.4355	8.4	73.6
6	1.0813	6.4	79.9
7	0.8527	5.0	84.9
8	0.6082	3.6	88.5
9	0.5129	3.0	91.5
10	0.4287	2.5	94.1
11	0.3711	2.2	96.2
12	0.2542	1.5	97.7
13	0.1682	1.0	98.7
14	0.1151	0.7	99.4
15	0.0495	0.3	99.7
16	0.0333	0.2	99.9
17	0.0193	0.1	100.0



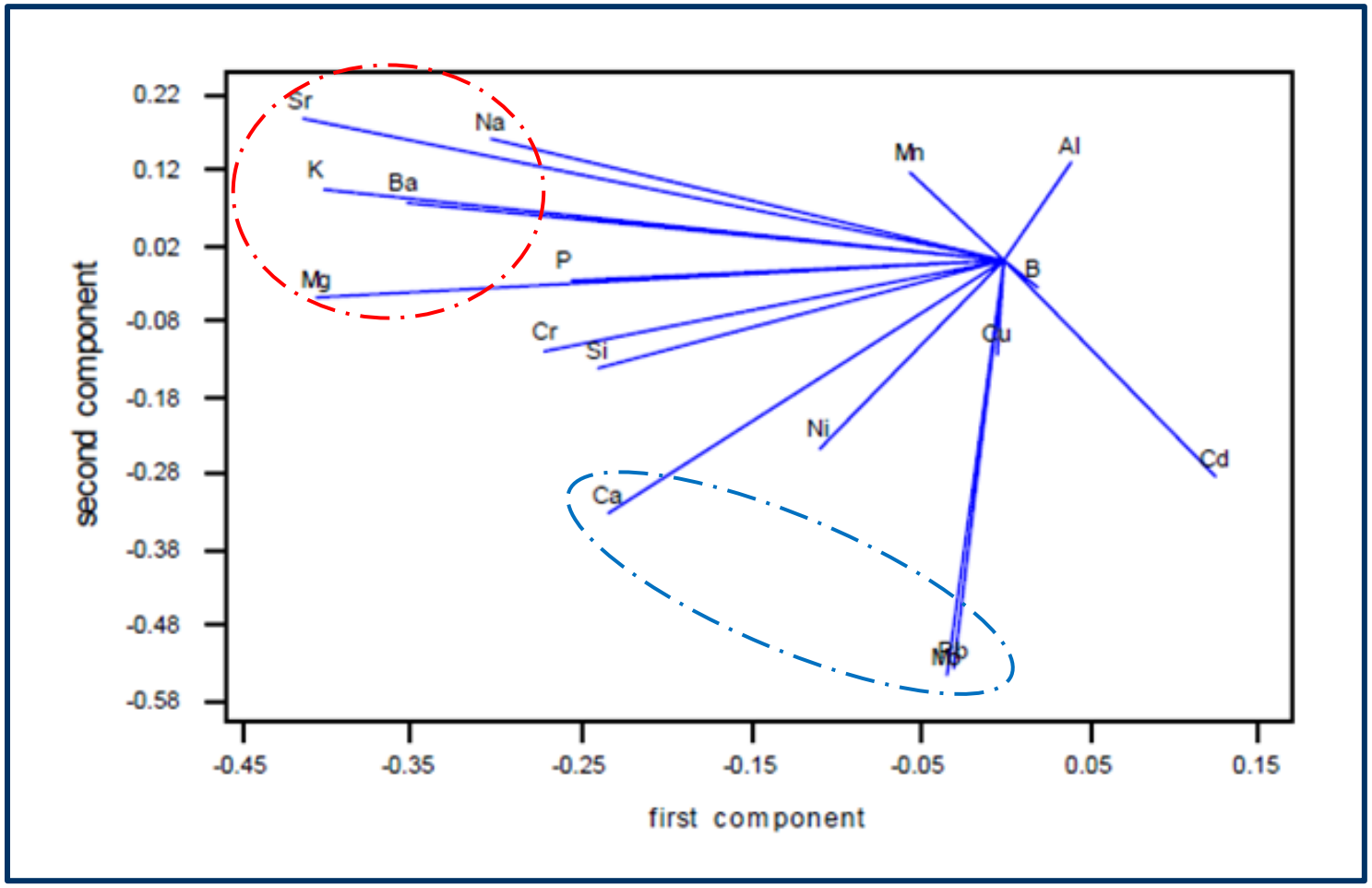
E.V.%: percentage of explained variance
C.E.V.% : percentage of cumulative explained variance

Loadings for the first 6 principal components, those having eigenvalues greater than 1, are reported in the following table:

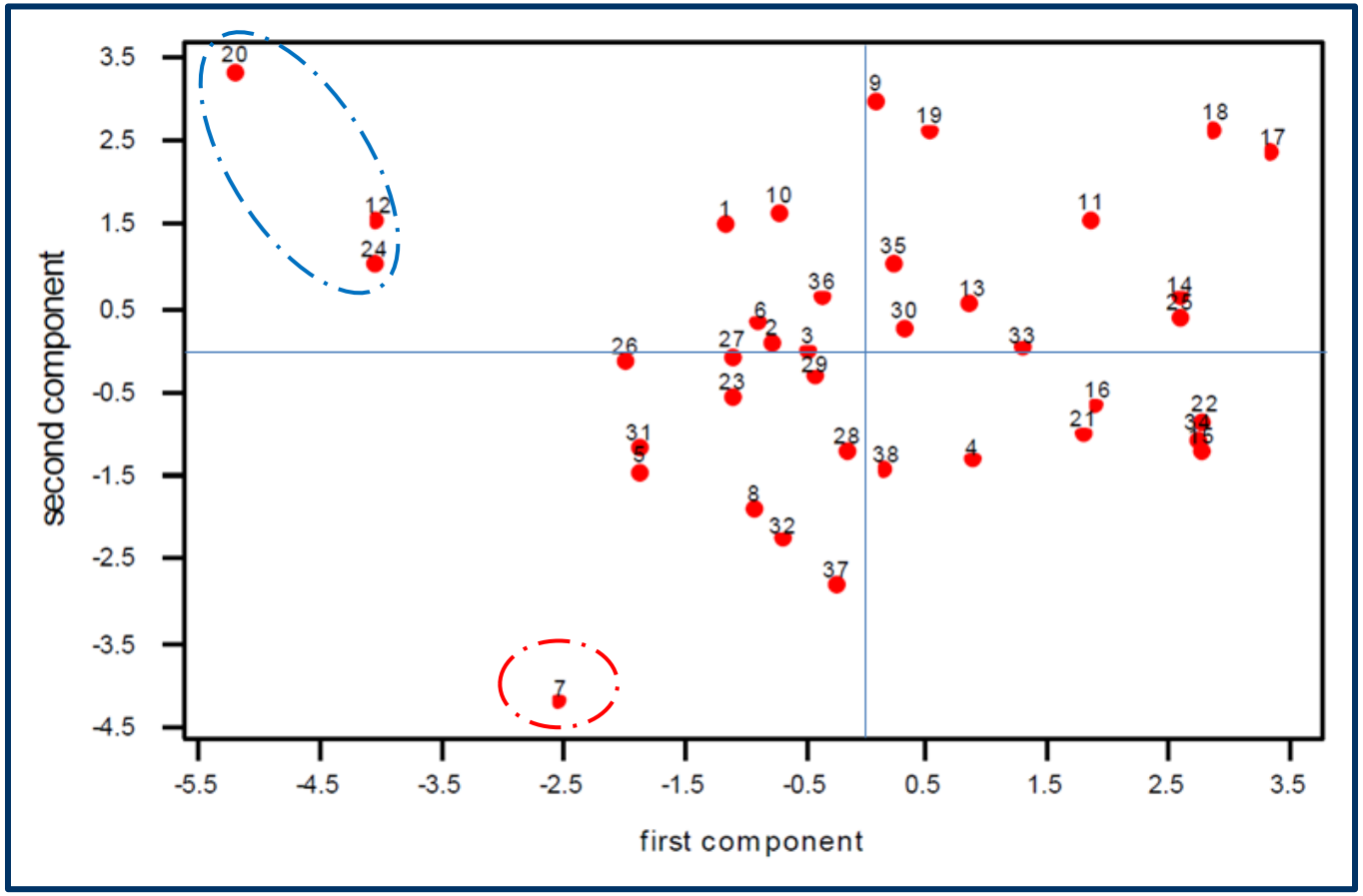
ID	*Var.*	*PC1*	*PC2*	*PC3*	*PC4*	*PC5*	*PC6*
1	Cd	0.125	-0.285	**0.351**	0.055	**-0.369**	-0.233
2	Mo	-0.034	**-0.546**	0.150	-0.125	0.132	-0.096
3	Mn	-0.056	0.118	**0.571**	0.021	0.011	0.072
4	Ni	-0.109	-0.247	-0.268	-0.140	-0.107	**0.552**
5	Cu	-0.004	-0.122	-0.219	-0.065	**0.496**	-0.061
6	Al	0.039	0.130	-0.278	**-0.420**	0.047	**-0.352**
7	Ba	**-0.353**	0.080	0.061	-0.229	**-0.348**	-0.013
8	Cr	-0.271	-0.118	0.266	0.101	**0.394**	-0.087
9	Sr	**-0.415**	0.187	0.134	-0.166	-0.085	0.168
10	Pb	-0.030	**-0.537**	0.168	-0.161	0.064	-0.091
11	B	0.020	-0.034	-0.091	**0.618**	-0.052	-0.224
12	Mg	**-0.405**	-0.048	0.075	-0.084	-0.111	0.115
13	Si	-0.239	-0.142	-0.282	**0.308**	-0.276	-0.123
14	Na	**-0.303**	0.161	-0.019	-0.194	0.228	**-0.438**
15	Ca	-0.233	**-0.333**	**-0.339**	-0.022	-0.140	-0.116
16	P	-0.256	-0.024	-0.015	0.289	**0.368**	**0.342**
17	K	**-0.403**	0.097	-0.011	0.243	-0.029	-0.231

As apparent, elements contributing more to PC1 are all alkaline or alkaline-earth ones, whereas Mo, Pb and Ca are those giving the greatest contribute to PC2.

As apparent from the loadings plot, elements contributing most to PC1 (negatively) are all alkaline or alkaline-earth ones, whereas Mo, Pb and Ca are those having the greatest (negative) contribution to PC2:



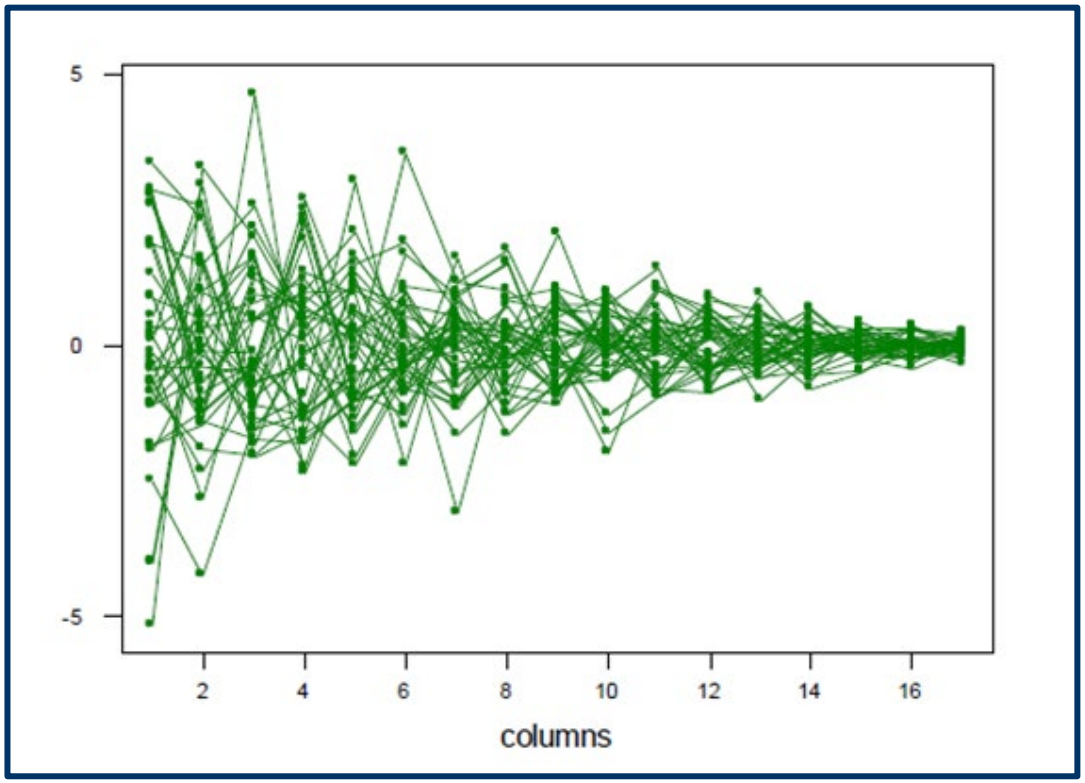
The score plot referred to the first two principal components enables a comparison between samples:



Samples 12, 20 and 24 are isolated on the upper-left zone of the score plot, likely because of the high content in one or more alkaline or alkaline-earth elements.

On the other hand, sample 7 is isolated in the lower zone of the plot, likely due to the presence of Mo and Pb in high concentration.

Another interesting plot that can be generated after PCA calculations is the one reporting scores obtained for all samples in all principal components (corresponding to «columns»):

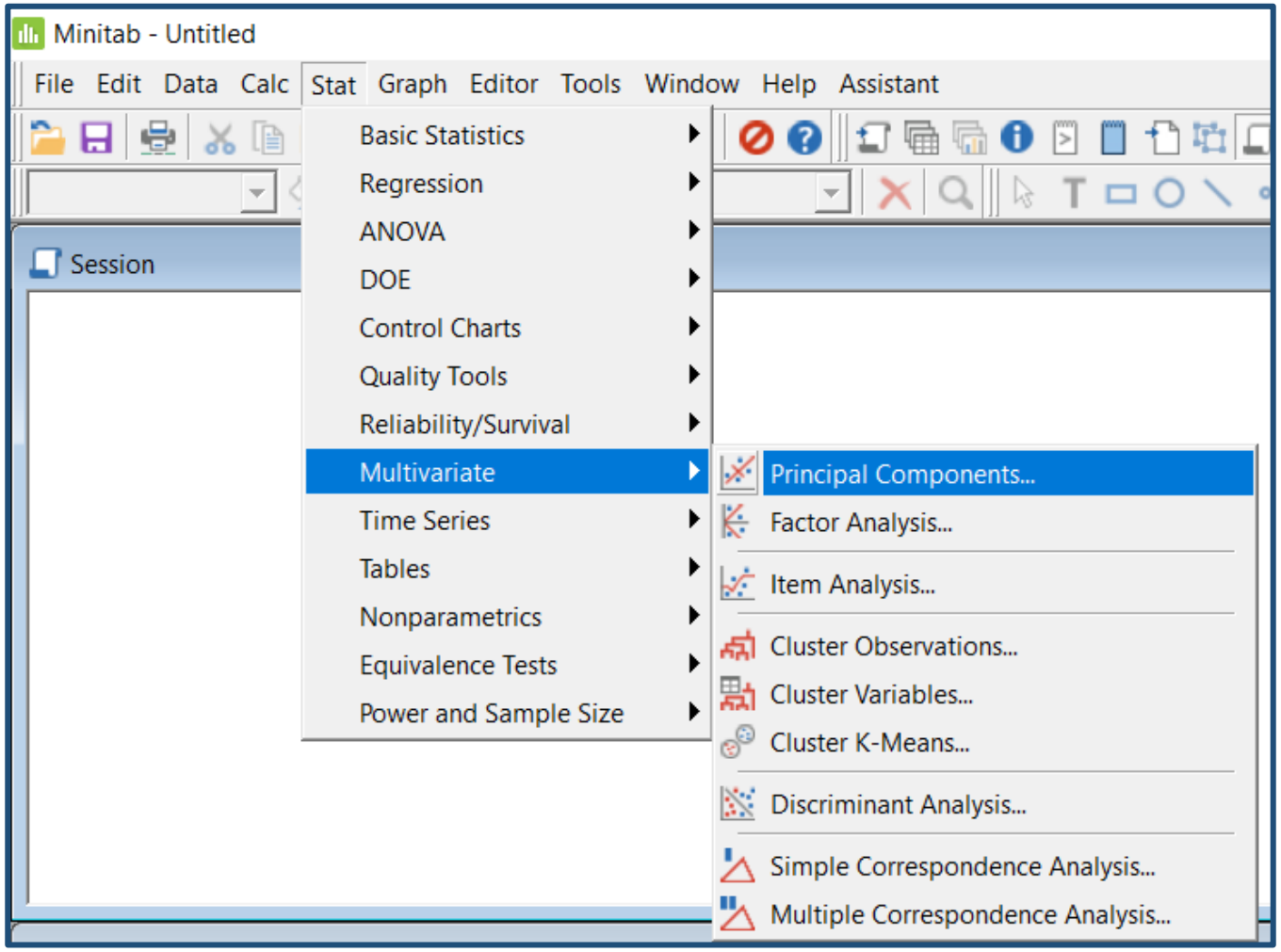


This plot emphasizes the decrease in variance occurring progressively from the first to the last principal component, as expected.

Actually, the variance of some components is still relatively high for the presence of samples whose behavior is anomalous.

Use of Minitab 18 for Principal Components Analysis

PCA can be performed by Minitab 18 by considering the Stat > Multivariate > Principal Components… pathway.



The original dataset is preliminarily transferred into the Minitab worksheet using rows for samples and columns for variables, as shown in this example, where samples are represented by mussel lipid extracts (labelled according to names reported in column C1-T) and variables correspond to free fatty acids detected by LC-MS in those extracts:

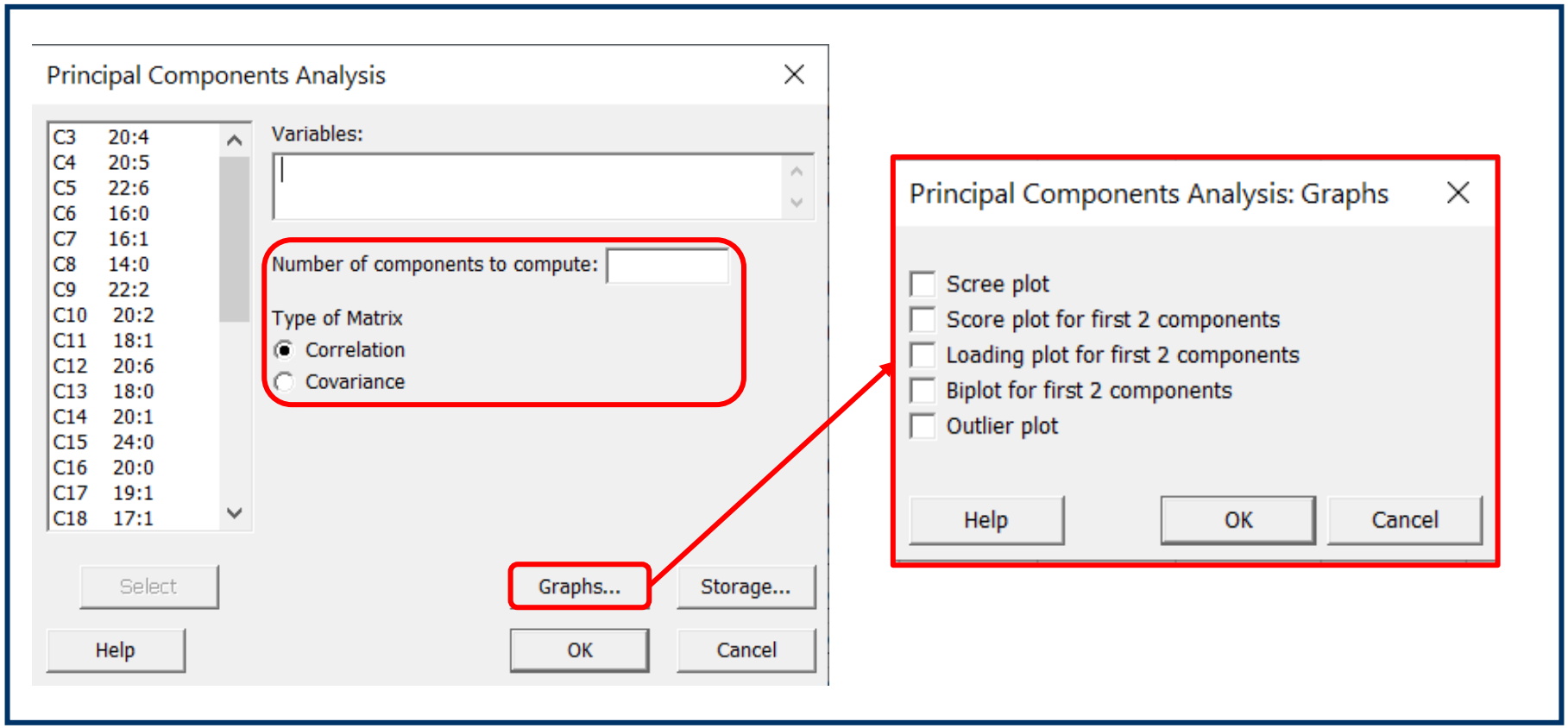
Worksheet 1 ***

↓	C1-T	C2-T	C3	C4	C5	C6	C7	C8	C9	C10
	Sample	Code	20:4	20:5	22:6	16:0	16:1	14:0	22:2	20:2
1	8-7 Fr A	F7A	1.00000	0.78392	0.66834	5.78E-01	0.132161	0.071357	0.110050	0.062814
2	8-7 Fr B	F7B	0.91511	0.91885	0.53308	1.00E+00	0.124844	0.116729	0.068664	0.034332
3	17-2 Fr A	F2A	0.18011	1.00000	0.56452	2.28E-01	0.080645	0.032581	0.031935	0.015591
4	17-2 Fr B	F2B	0.16629	1.00000	0.60571	1.77E-01	0.061714	0.028343	0.026914	0.017943
5	28-4 Fr A	F4A	0.32951	1.00000	0.92350	3.32E-01	0.107104	0.032623	0.073224	0.033989
6	28-4 Fr B	F4B	0.32789	1.00000	0.78421	1.72E-01	0.067895	0.024947	0.051474	0.024000
7	21-4 Fr A	F4A1	0.20792	1.00000	0.61881	1.28E-01	0.045149	0.025050	0.030099	0.013119
8	21-4 Fr B	F4B1	1.00000	0.64255	0.46383	5.79E-02	0.034255	0.009681	0.024043	0.009638
9	10-7 Refr A	R7A	1.00000	0.85890	0.74233	4.86E-01	0.103681	0.071166	0.130675	0.103067
10	10-7 Refr B	R7B	1.00000	0.95041	0.79835	5.85E-01	0.157025	0.090083	0.139669	0.074298
11	14-7 Tamb A	H7A	0.79593	0.80894	0.71870	4.07E-01	0.136585	0.070081	0.144715	0.059837
12	14-7 Tamb B	H7B	1.00000	0.65048	0.36571	4.98E-01	0.091048	0.062571	0.090571	0.062762
13	17-7 Long Refr A	LR7A	0.79399	1.00000	0.84120	4.55E-01	0.190129	0.097425	0.149356	0.078970
14	14-7 Long Refr B	LR7B	0.84444	0.99444	1.00000	6.28E-01	0.164444	0.096667	0.161111	0.077222
15	16-7 sress A	LLR7A	1.00000	0.65969	0.47054	4.23E-01	0.116279	0.052868	0.164341	0.093798
16	16-7 sress B	LLR7B	1.00000	0.92045	0.78409	4.01E-01	0.121591	0.008011	0.111364	0.068750

Note that a further text column (C2-T) can be used to report classification codes that can be useful to recognize samples on the score plot.

Inside the Principal Components Analysis window variables to be considered for calculations are selected, then the number of components to be computed is set (it can be lower than the total number of variables). Last but not least, the choice of data matrix, correlation or covariance, implying the centroidation or autoscaling of data, respectively, is indicated.

Several types of graphical representations can be selected in the Graphs… window. Minitab 18 is able to reproduce only the score and the loading plots for the first two components.



In any case, the storage of scores obtained for further principal components into additional columns of the worksheet can be selected in the Storage… window.

Once the calculation is completed, several types of information are included in the Session window. First, eigenvalues and the fraction of variance explained by each principal component (in the present case 27 principal components were calculated), along with cumulative values, are reported:

Eigenanalysis of the Correlation Matrix

Eigenvalue	9,8295	4,6957	2,4590	2,0375	1,4903	1,3696	1,0905	0,7593	0,6574	0,5156
Proportion	0,364	0,174	0,091	0,075	0,055	0,051	0,040	0,028	0,024	0,019
Cumulative	0,364	0,538	0,629	0,705	0,760	0,810	0,851	0,879	0,903	0,922
Eigenvalue	0,4041	0,3919	0,3499	0,2641	0,1688	0,1604	0,0994	0,0749	0,0496	0,0412
Proportion	0,015	0,015	0,013	0,010	0,006	0,006	0,004	0,003	0,002	0,002
Cumulative	0,937	0,952	0,965	0,975	0,981	0,987	0,990	0,993	0,995	0,997
Eigenvalue	0,0324	0,0232	0,0192	0,0070	0,0039	0,0035	0,0018			
Proportion	0,001	0,001	0,001	0,000	0,000	0,000	0,000			
Cumulative	0,998	0,999	0,999	1,000	1,000	1,000	1,000			

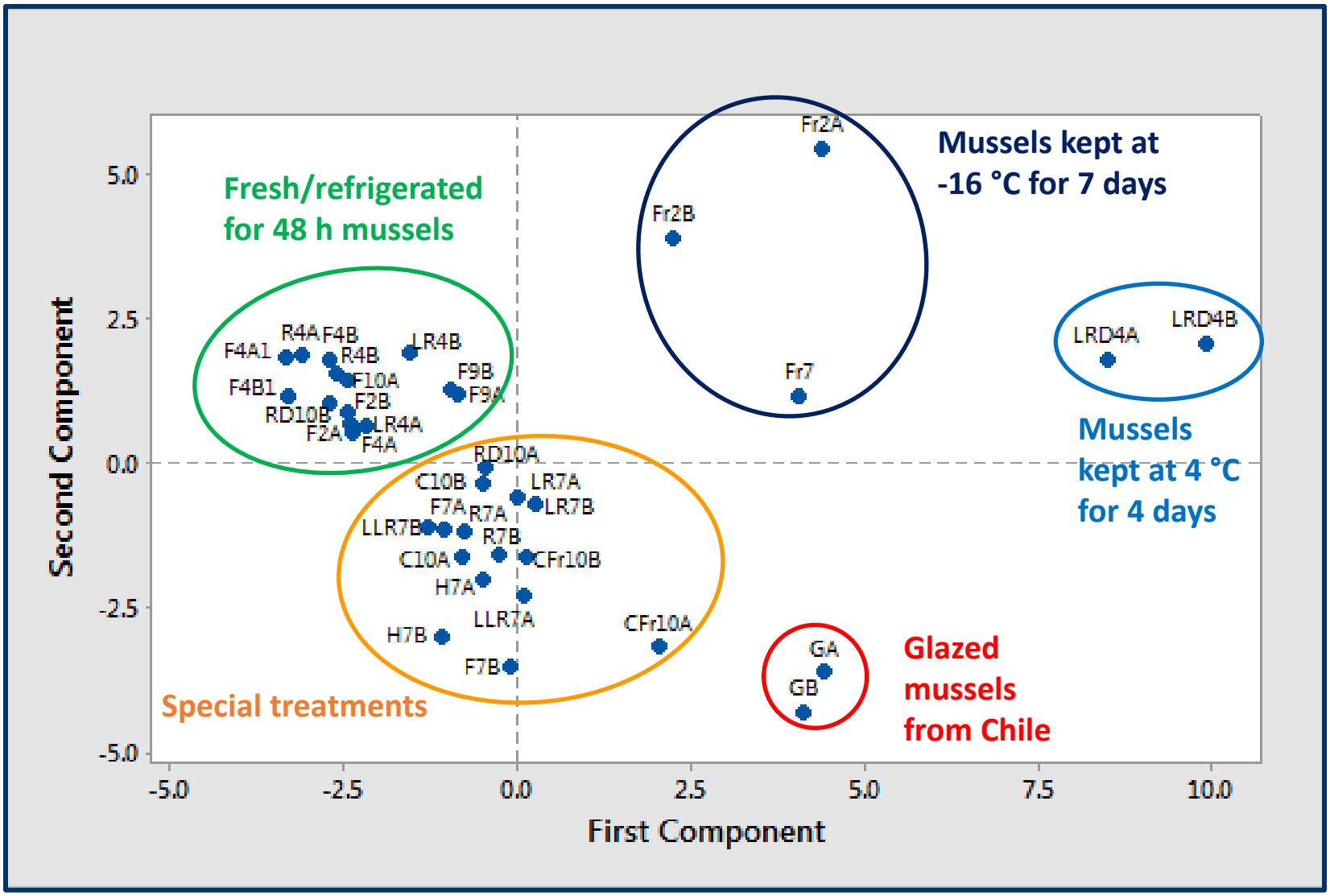
Note that the first six components were able to account for more than the 80% of total variance in this case.

Eigenvectors, i.e., loadings of variables on principal components, are also reported.
In the following figure only those relevant to the first 8 principal components are shown:

Eigenvectors

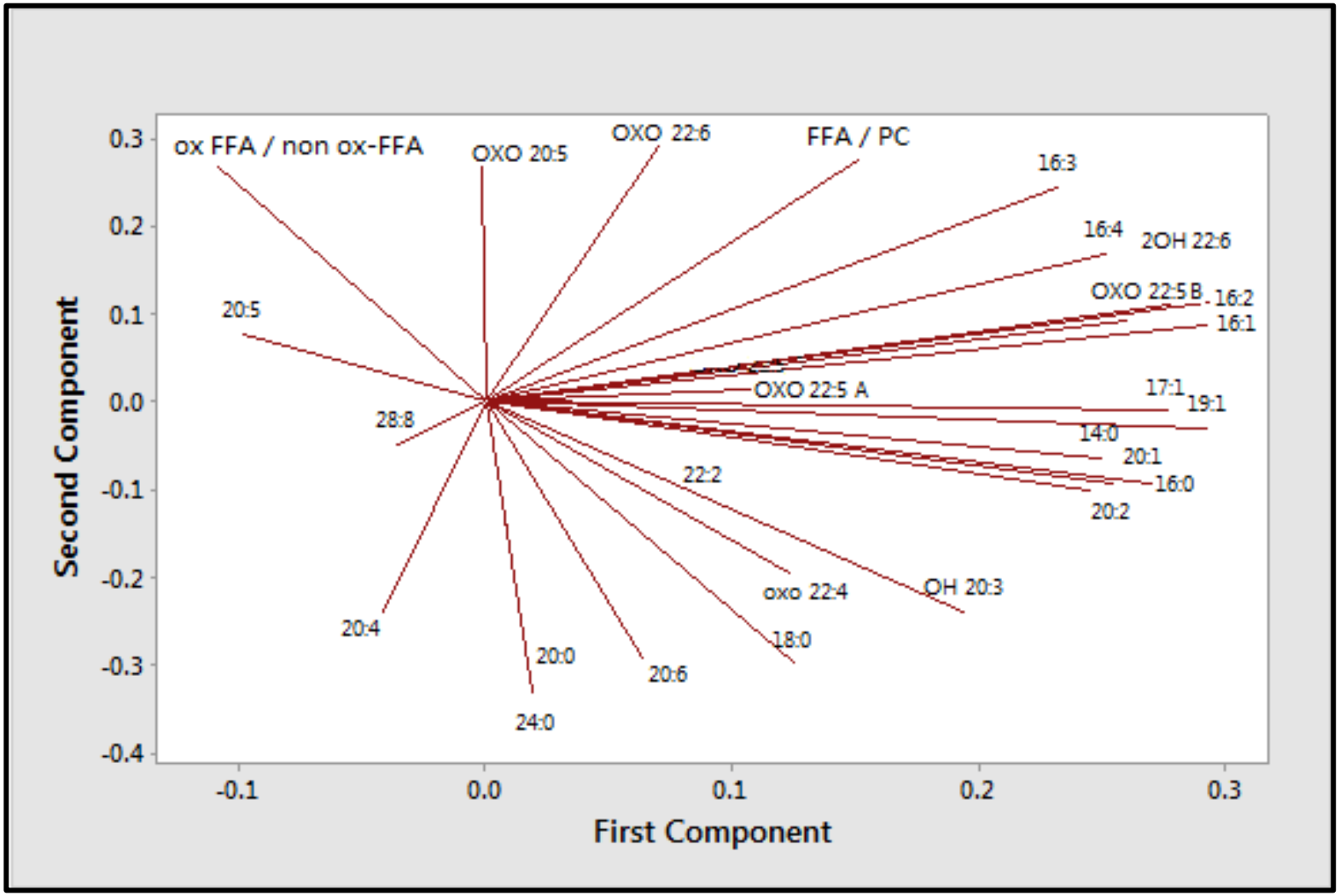
Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
20:4	-0,042	-0,240	-0,263	0,266	-0,032	0,295	-0,128	0,306
20:5	-0,099	0,076	0,123	-0,152	-0,643	-0,144	-0,172	0,192
16:0	0,270	-0,094	0,064	0,193	-0,081	0,174	-0,178	-0,151
16:1	0,293	0,085	0,054	0,041	-0,023	0,081	0,011	0,001
14:0	0,250	-0,066	0,091	0,205	-0,056	0,020	-0,232	-0,362
22:2	0,089	-0,110	-0,448	0,091	-0,142	-0,248	-0,292	-0,313
20:2	0,246	-0,103	-0,243	-0,230	0,003	-0,095	-0,079	0,030
20:6	0,063	-0,293	0,360	0,022	-0,006	-0,001	-0,251	0,102
18:0	0,125	-0,298	0,121	0,109	-0,148	0,331	-0,275	0,093
20:1	0,254	-0,094	-0,158	0,049	-0,128	-0,126	0,161	0,142
24:0	0,019	-0,331	-0,072	0,288	-0,154	-0,134	0,328	0,096
20:0	0,016	-0,281	-0,130	0,299	-0,039	-0,101	0,511	0,073
19:1	0,293	-0,031	-0,012	-0,173	0,034	-0,051	0,063	0,038
17:1	0,277	-0,012	0,080	-0,234	-0,011	0,155	0,038	0,243
16:4	0,252	0,167	0,077	0,075	-0,148	-0,134	0,116	0,221
16:3	0,233	0,242	0,012	0,099	-0,196	-0,013	0,017	0,237
16:2	0,294	0,111	0,081	0,050	-0,099	-0,080	0,063	0,056
OH 20:3	0,194	-0,241	0,319	-0,124	0,131	-0,138	0,080	-0,026
OXO 22:5 A	0,261	0,091	-0,259	-0,069	0,079	0,005	-0,042	0,011
OXO 22:5 B	0,107	0,014	-0,175	-0,313	-0,209	0,325	0,331	-0,387
2OH 22:6	0,278	0,109	-0,050	0,016	0,075	-0,071	-0,070	-0,066
OXO 20:5	-0,001	0,266	0,169	0,311	0,192	-0,098	-0,030	0,145
OXO 22:6	0,071	0,289	-0,007	0,422	0,031	-0,227	-0,060	-0,209
28:8	-0,036	-0,048	-0,251	-0,242	0,121	-0,436	-0,219	0,302
∑aree ox / ∑ aree nativi	-0,108	0,268	0,138	-0,008	-0,427	-0,011	0,129	-0,101
Area ac.grassi / area PC	0,151	0,274	-0,045	-0,009	0,311	0,285	0,096	0,149
OXO 22:4	0,112	-0,200	0,340	-0,147	0,126	-0,342	0,137	-0,224

Score, loadings and scree plots can be easily generated by the program. In the following figure, the score plot, indicating a clear clustering of mussel samples with a different thermal history, is reported:



I. Losito *et al.*, *Food Chemistry*, 255 (2018) 309-322

The loading plot emphazises which variables are responsible for the distinction of sample clusters.



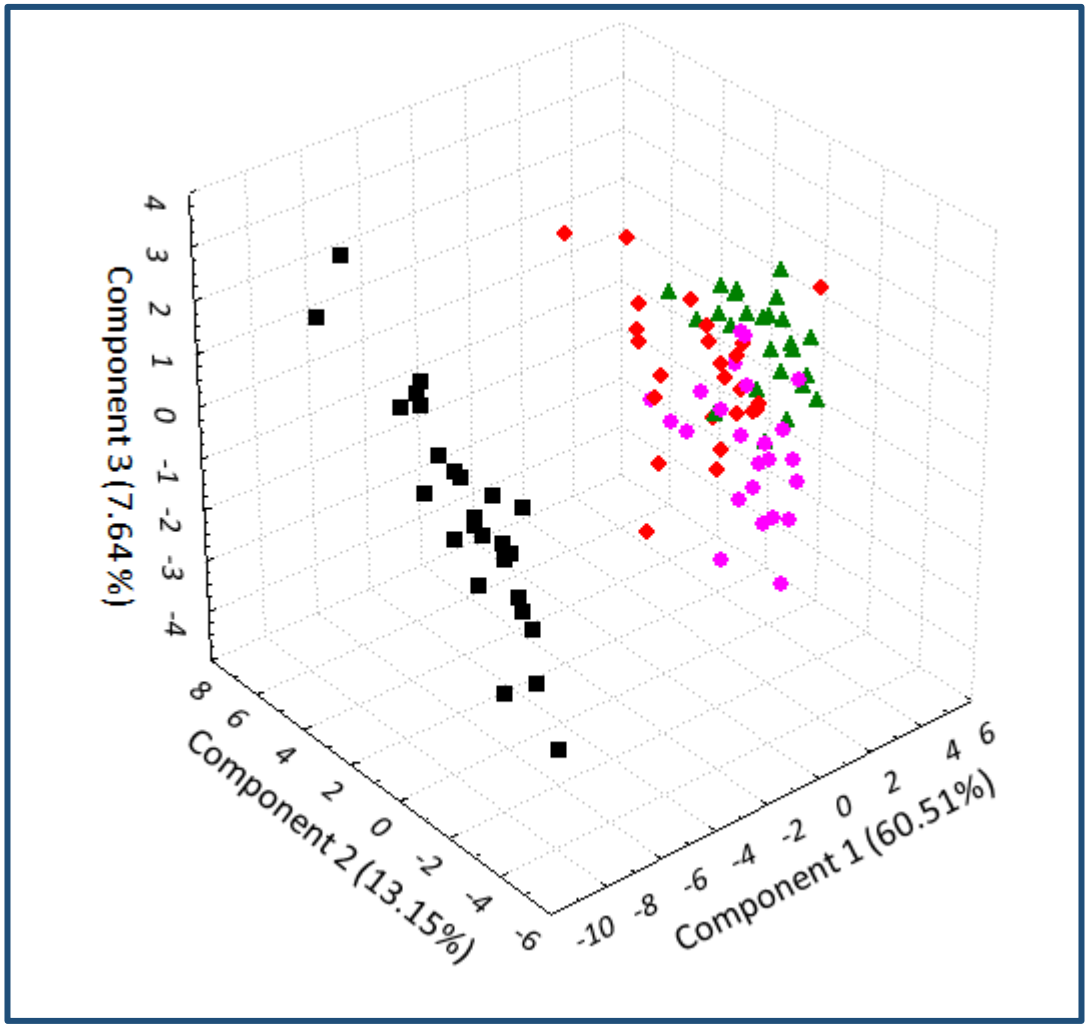
I. Losito *et al.*, *Food Chemistry,* 255 (2018) 309-322

Another example of Principal Components Analysis

Three-dimensional score plot obtained for 100 salmon samples fished in different countries, based on the abundance of 30 different fatty acids in the fish meat (determined by mass spectrometry):

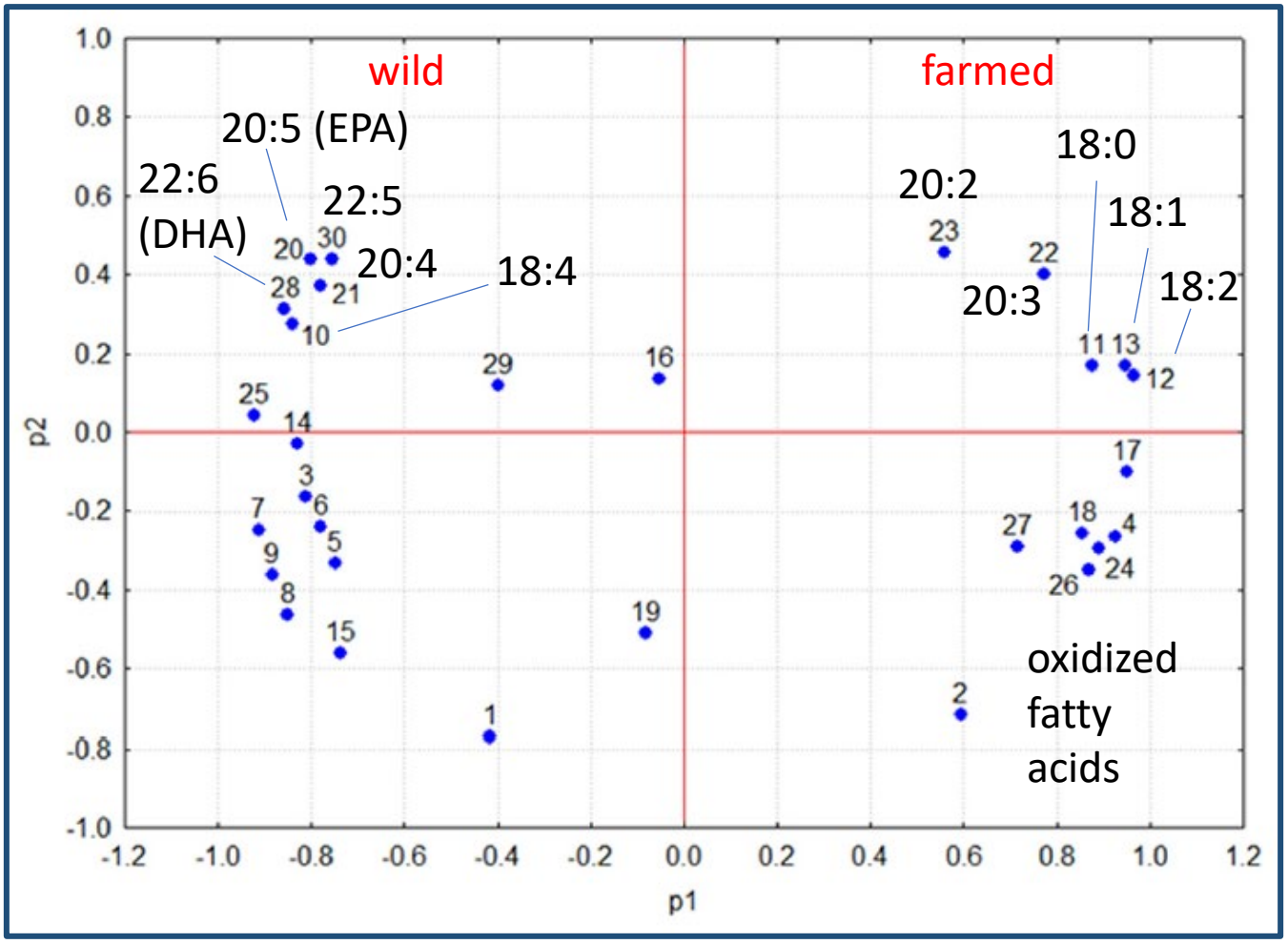
- Farmed – Canada
- Farmed – Norway
- Farmed – Chile
- Wild – Canada

As apparent, wild Canadian salmon samples could be clearly distinguished from all farmed salmon samples, apart from their geographical origin.



G.M. Fiorino, I. Losito *et al.*, *Food Research International, 116* (2019) 1258 - 1265

The loading plot referred to the first two principal components, with variables labelled with the conventional names for fatty acids (C:D, with C = number of carbon atoms, D = number of C=C bonds) provided some explanations for the difference observed between wild and farmed salmons:



Wild salmons contained higher concentrations of long and highly unsaturated fatty acids, including omega-3 ones (like 20:5 and 22:6), whereas fatty acids like 18:1, 18:2 and 18:3, resulting from feeding based on vegetal oils, were more relevant in farmed salmons.