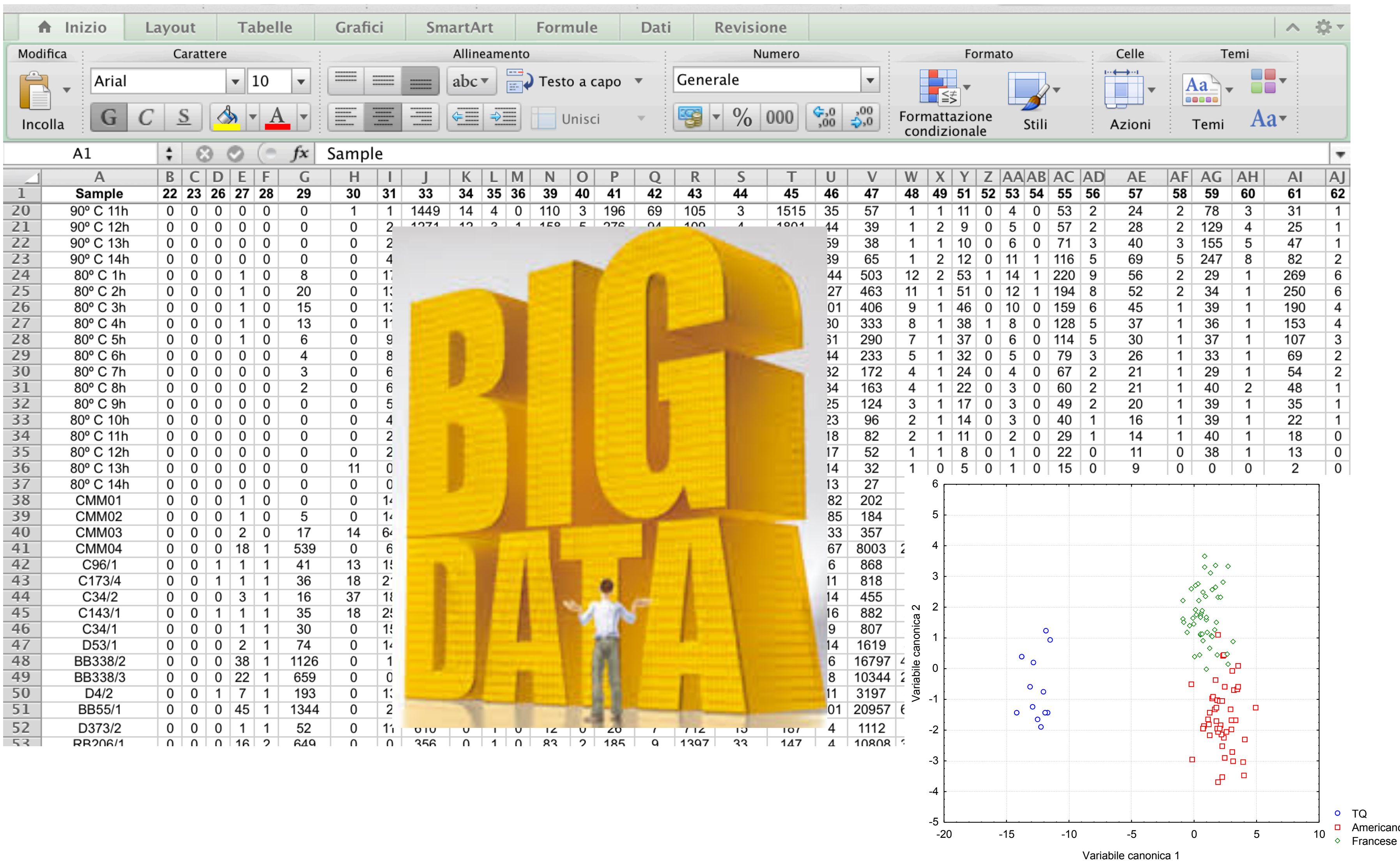
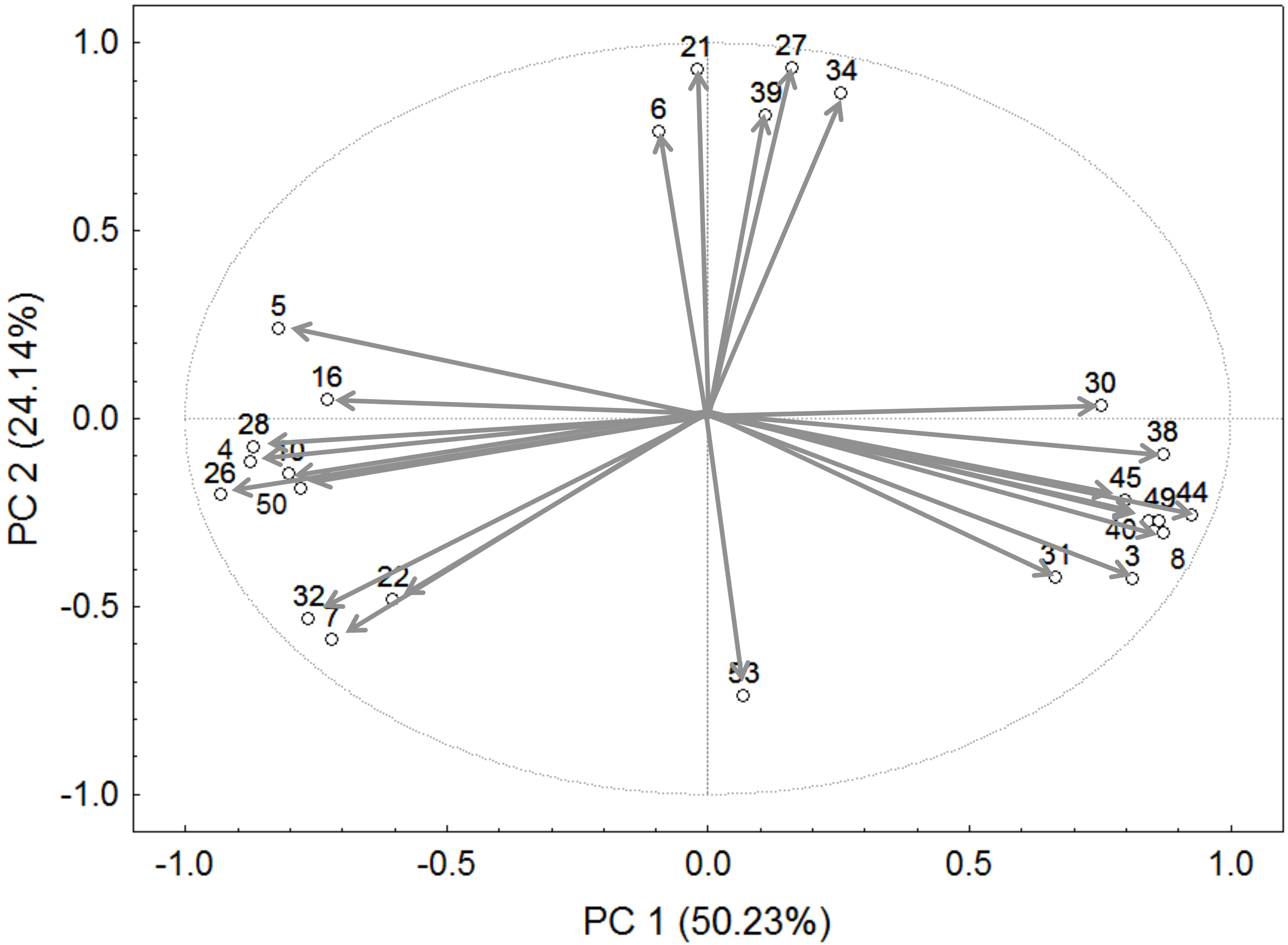
Multivariate data discrimination: LDA



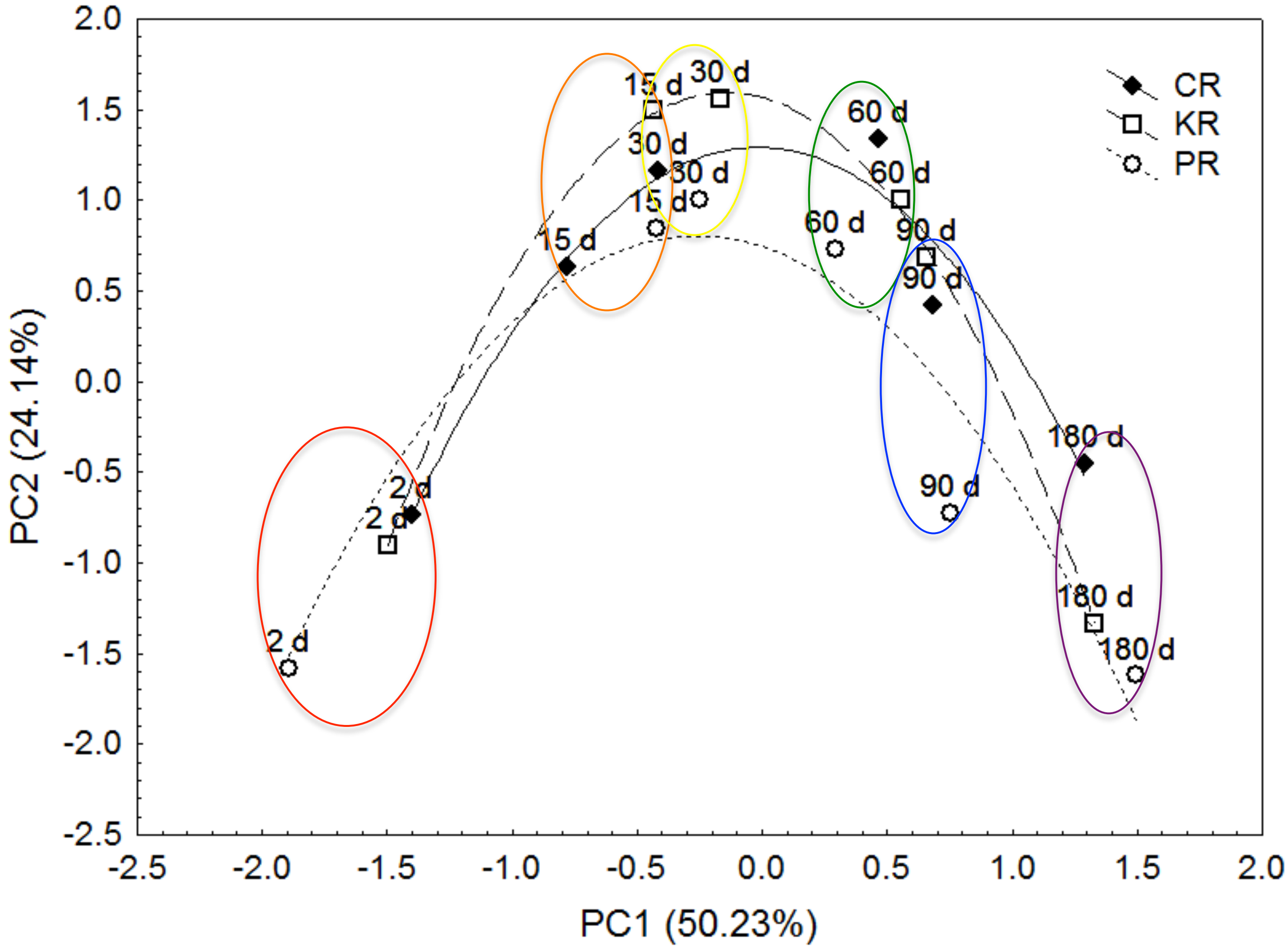
Let’s start from practical PCA results …

- PCA was applied to the data set provided in the EDCF01DEC2016.xlsx file
- 18 samples of cheeses produced with three different rennets (CR, KR and PR) were aged for 180 days and analysed for their volatiles profile.
- A matrix of 18 x 53 was obtained
- PCA was carried out on the data matrix
- The matrix was reduced to a 18 x 25 one by stepwise selection of variables

Results: Loadings plot



Results: Scores plot



Let's analyze PCA results …

- PCA permitted to visualize sample distribution along PC1 on the basis of explained variance maximization.

- Samples appears as if they were ordered along PC1 on the basis of their aging (from 2 to 180 days).

- PCA did not permit to differentiate samples on the basis of the rennet type used for cheesemaking eventhough some variability due to rennet effect could be seen along PC2.

Meaning of results

- Ripening time is the major source of variability (variance) in data structure.
- Rennet type is a secondary source of variability and the type of analysis that was carried out is focused on maximizing explained variance.
- PCA is aimed to highlight data structure as determined by internal data variance.
- PCA is not aimed to discriminate among different group of samples!

Further questions …

- Is it possible to discriminate samples on the basis of rennet type?

- Is it possible to find a latent variable that could serve to this need?

Linear Discriminant Analysis

- LDA is closely related to principal component analysis (PCA) in that they both look for linear combinations of variables (latent variables) which best explain the data.
- LDA explicitly attempts to model the difference and similarities between the classes of data.
- PCA on the other hand does not take into account any difference in class, and builds the feature combinations based on difference (variance) rather than similarities.

Discriminant Analysis

- Discriminant analysis implies a distinction between categorical independent variables (measured variables) and dependent variables (also called criterion variables).
- In the case of our study the criterion variable is a categorical variable consisiting in three categories (CR, KR and PG).

Discriminant Analysis

- Classificatory discriminant analysis is used to test the possibility of attributing a sample to a class (CR, KR or CR) by knowing *a priori* it's classification.
- This attribution/classification is performed starting from the independent variables.
- The success of attribution is measured in terms of probability (%).
- If the 100% of PR sample are attributed to (or classified in) the PR group, the result is optimum.

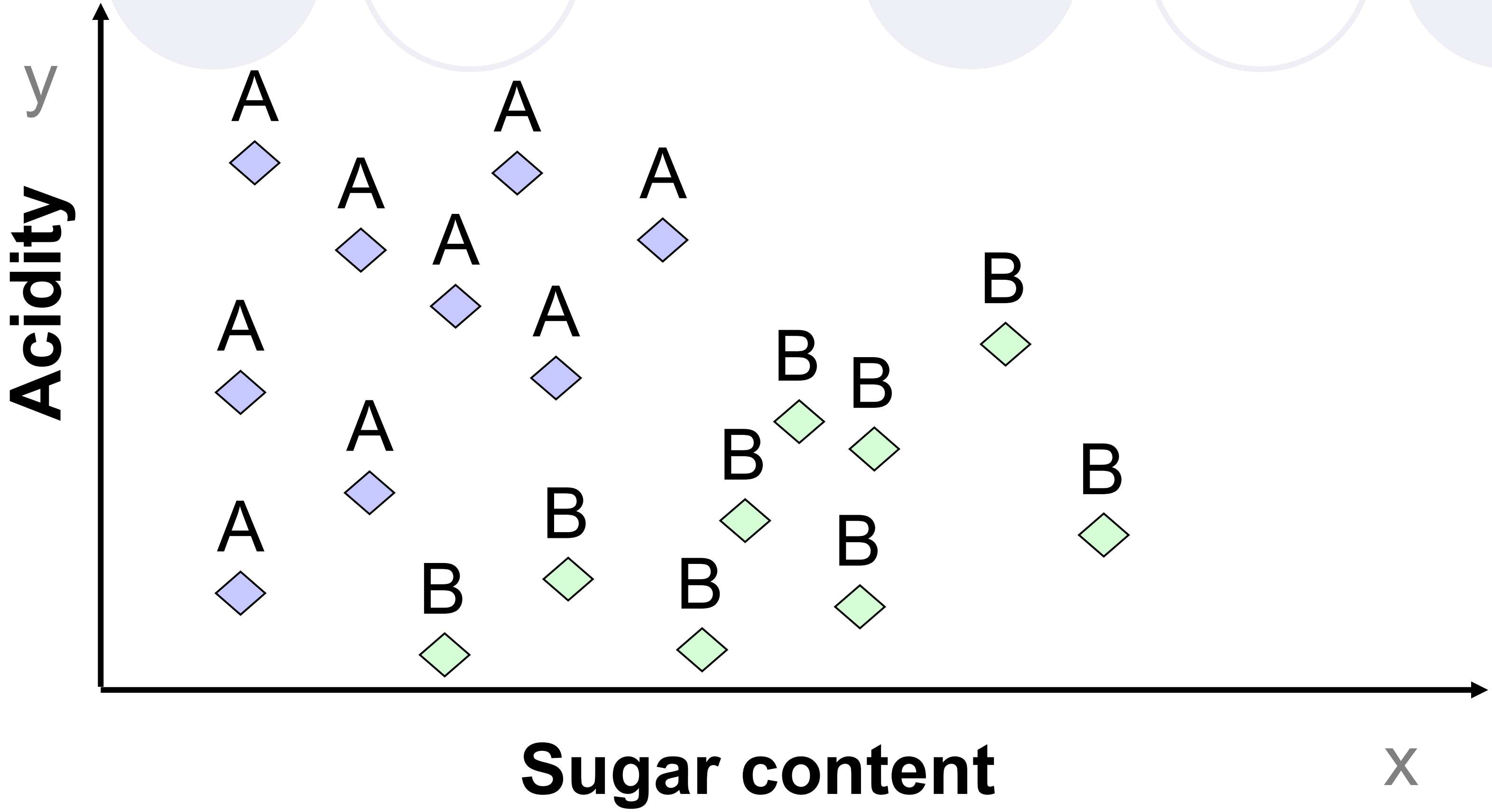
Linear Discriminant Analysis

- Discriminant analysis develop functions (discriminant functions), based on the combination of independent variables, which permits to attribute a sample to a class with the minum possibility of error.

- The discriminant function is a latent structure since it is a combination of original variables. The number of discriminant functions is equal to the number of classes.

- Just in the case that discriminant functions are linear, the analysis tooks the name of Linear Discriminant Analysis (LDA).

Example using two variables and two groups

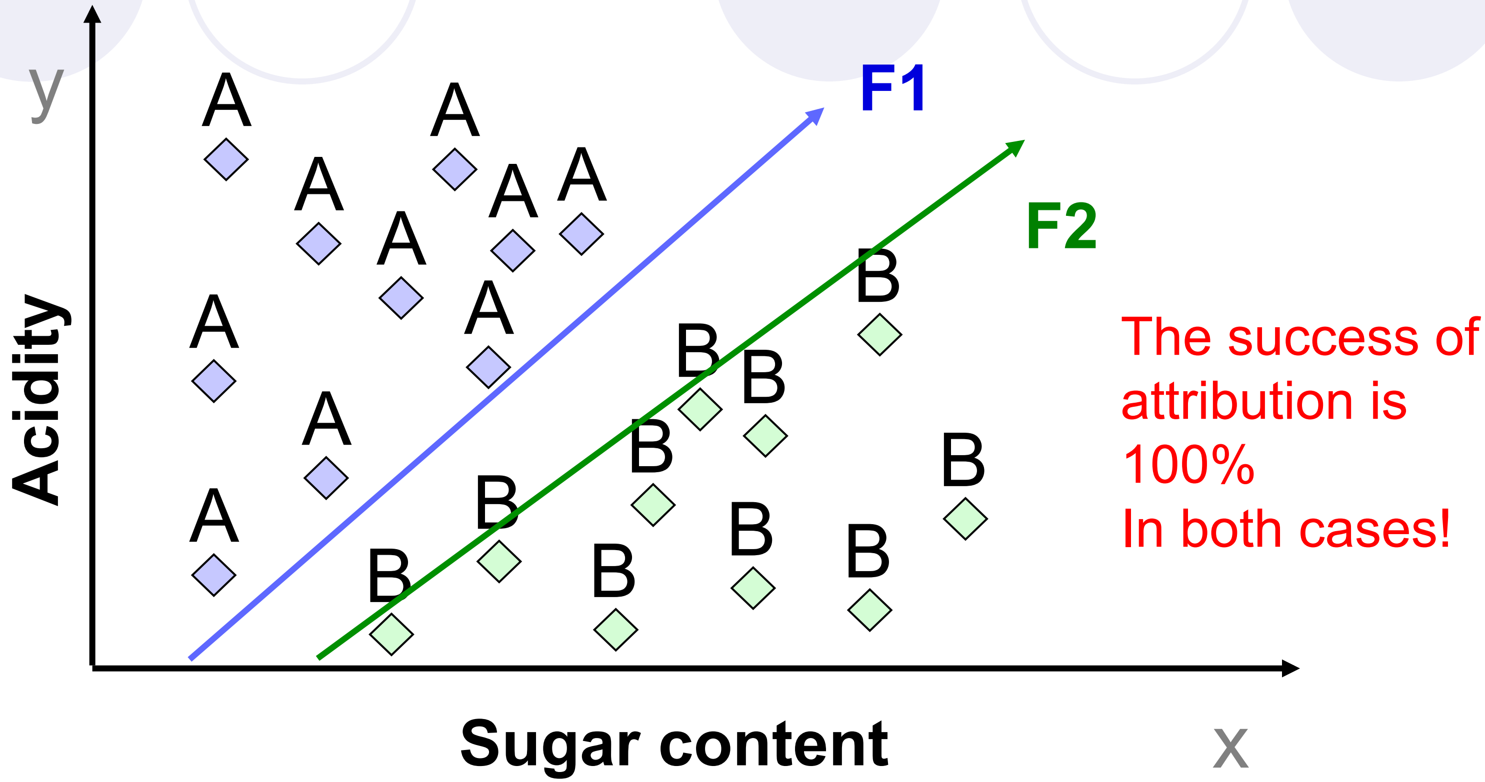


Two groups of samples, A, and B could differ for **two variables**

Two dimensions (x, y) are required for data description;

Each sample is identified by two values (x and y)

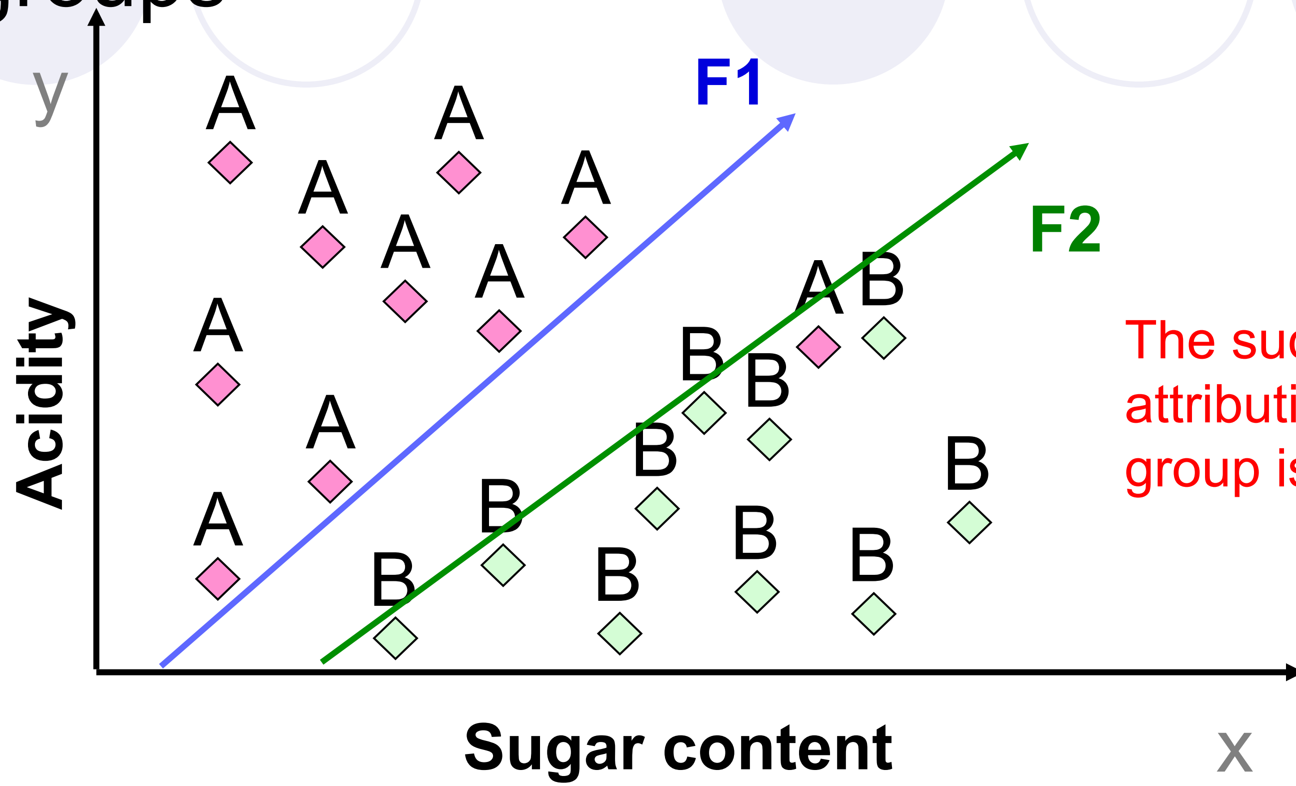
Example using two variables and two groups



LDA permit to calculate two functions, F1 and F2, which are linear combinations of x and y since y = f(x) in both cases.

F1 permits to better discriminate samples of A group from sample of B group, and F2 permits to better discriminate samples of B group from samples of A group.

Another example using two variables and two groups

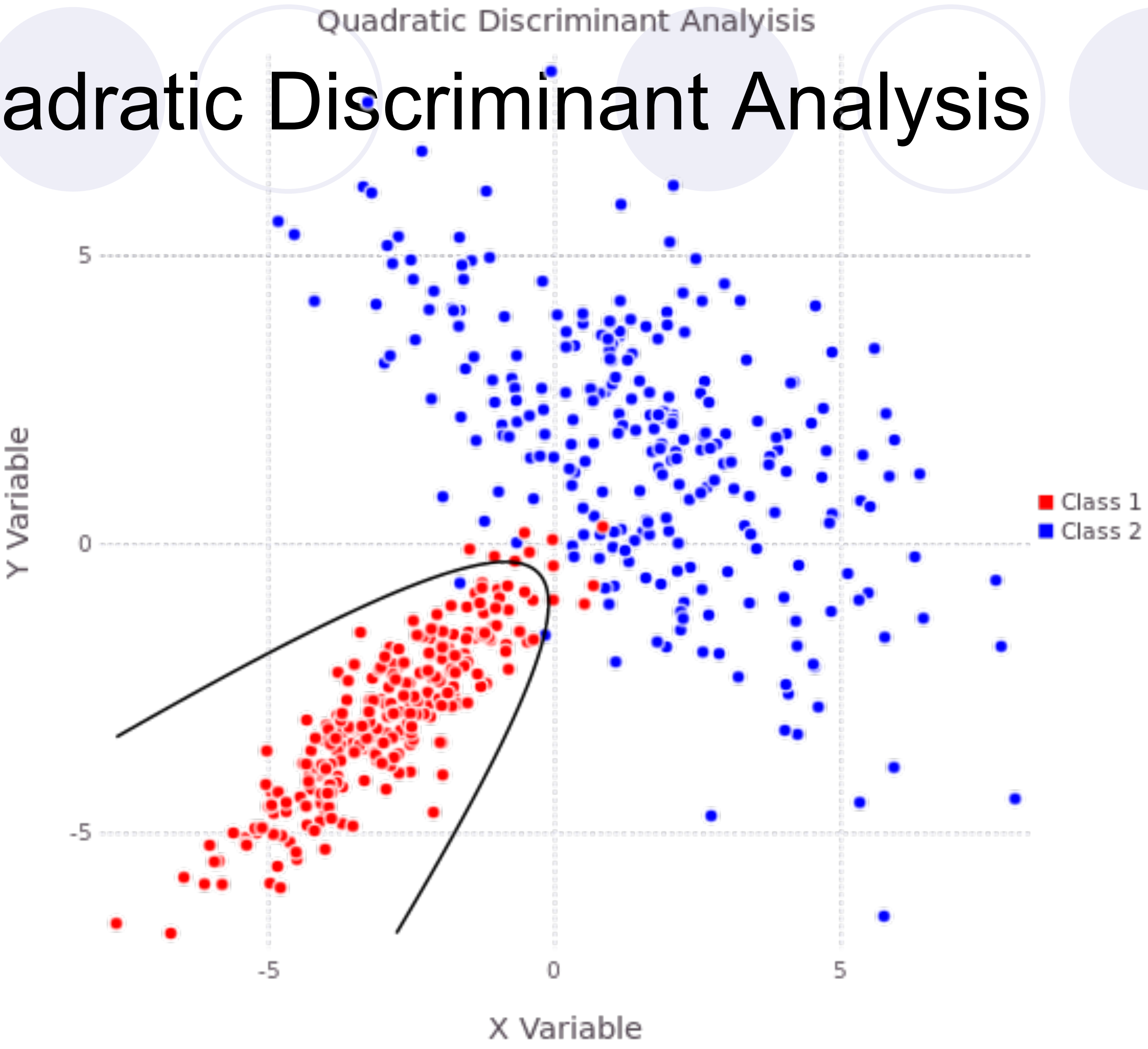


The success of attribution for A group is 90%!

LDA permit to calculate two functions, F1 and F2, which are linear combinations of x and y (y = f(x)).

F1 permits to better discriminate samples of A group from sample of B group, and F2 permits to better discriminate samples of B group from samples of A group.

Quadratic Discriminant Analysis



Canonical discriminant analysis

- This technique permits to extrapolate new variables (Roots or Canonical variables) which synthesise the variability among classes (between-class variation) in much the same way that principal components summarize total variation.

- Root 1 is a linear combination of variables that maximizes the distance among the different classes, Root 2 is the second linear combination that maximizes the distance between different classes and so on, until all the combinations are considered.

Canonical discriminant analysis

- The numbers of roots is equal to that of the original variables similarly to PCA.

- Similarly to PCs, Roots are uncorrelated (orthogonal) among them, but differently from PCs they define a ‘System of Reference’ that maximizes the mean separation among classes and not among single observations, such as PCA.

Canonical discriminant analysis

- Similarly to PCA, Root 1 and 2 could constitute a limited number of variables that could substitute the original variables in order to have a good discrimination among samples.

- The Root 1 versus Root 2 plot is used to adequately visualize the results of CDA by performing a reduction of dimensionality in a way which is similar to that previously studied in PCA.

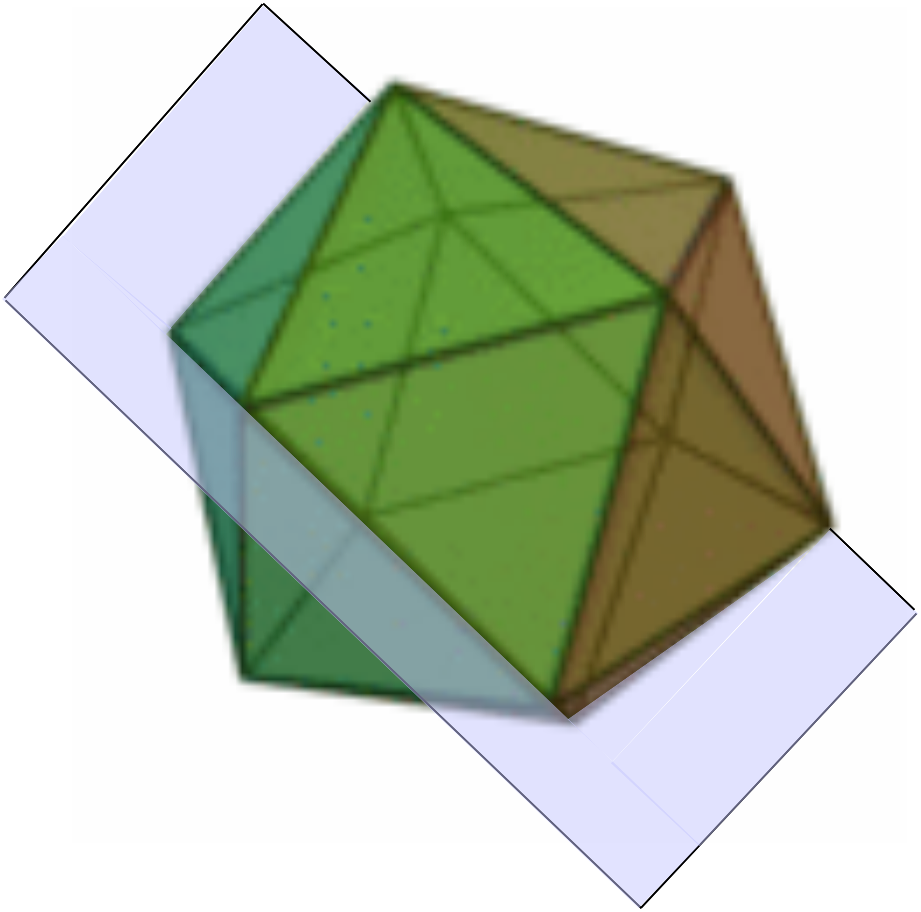
Dimensionality reduction (CDA)

- CDA could be thus used to reduce the dimensionality of a system at *n* (*n* > 3) dimensions by operating the othogonal projection of vectors and scores on a 2D plane or in a 3D space.

- The first 2 or 3 Roots will bring along with them the maximum distance among classes for definition.

- In this case PCA is useful to describe a data set since it ‘summarizes’ the information.

Figurative exemplification



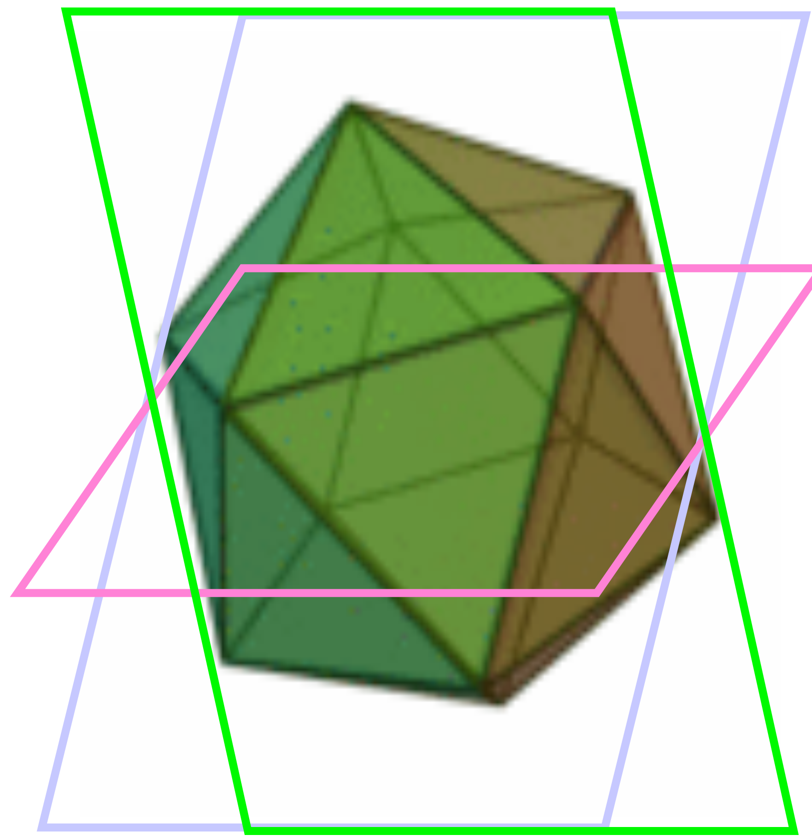
The reduction of dimensionality of a 10D space to a 2D space could be seen as a cut of the solid space with a 2D plane that intersecates the solid by passing trough the origin of axes (geometrical centre of solid).

How CDA operates to reduce dimensionality?

Infinite planes could pass through the central point (geometrical centre of the solid).

How does CDA choose the inclination (slope) of the cutting plane?

CDA choose the **slope** that permits to **maximize distance among classes** in the new bidimensional space.



Maximization distance among classes

- Restart from our data inset provided in the EDCF01DEC2016.xlsx file
- 18 samples of cheeses produced with three different rennets (CR, KR and PR) were aged for 180 days and analysed for their volatiles profile.
- A matrix of 18 x 53 was obtained
- CDA was carried out on the experimental data
- In order to carry out CDA we need two matrices
- The variable matrix (18 x 53)
- The groups matrix

Exercise

- Carry out linear discriminant analysis on the data set
- Individuate the variables that could be used for classes discrimination
- Verify the correct attribution to classes
- Visualize data using CDA

Original table

Group	Age	V1	V2	V3	V4	V5	V6	…	Vn
CR	2	1.05	26.65	3.90	27.19	2.37	1.48	…	1.01
CR	15	0.64	7.23	4.76	35.98	3.01	1.29	…	1.51
CR	30	1.10	6.26	4.90	24.39	4.03	1.53	…	1.34
CR	60	0.89	2.80	1.87	18.21	3.85	6.98	…	1.05
CR	90	0.73	3.99	0.00	20.28	2.07	7.01	…	0.57
CR	180	1.18	1.93	2.03	13.66	4.56	6.55	…	1.15
KR	2	0.18	1.95	3.73	21.31	3.69	4.02	…	0.52
KR	15	0.54	0.04	5.05	16.89	2.29	2.95	…	1.28
KR	30	0.33	2.31	5.39	29.44	3.72	4.23	…	0.81
KR	60	0.43	1.40	9.49	10.38	1.35	2.92	…	0.64
KR	90	0.57	1.18	9.53	9.30	0.84	5.42	…	0.9
KR	180	0.43	1.88	6.65	14.61	2.03	2.75	…	1.03
PR	2	0.35	2.69	11.07	10.21	0.46	2.84	…	0.77
PR	15	0.45	4.65	8.77	11.77	0.33	4.03	…	0.82
PR	30	2.74	0.86	11.35	7.47	0.93	1.27	…	1.18
PR	60	0.87	0.60	16.12	5.22	0.28	3.95	…	1.05
PR	90	0.43	0.68	13.66	6.72	0.35	2.11	…	1.3
PR	180	0.31	1.68	11.77	8.17	0.43	1.47	…	1.76

Data matrix

Group	V1	V2	V3	V4	V5	V6	…	Vn
CR	1.05	26.65	3.90	27.19	2.37	1.48	…	1.01
CR	0.64	7.23	4.76	35.98	3.01	1.29	…	1.51
CR	1.10	6.26	4.90	24.39	4.03	1.53	…	1.34
CR	0.89	2.80	1.87	18.21	3.85	6.98	…	1.05
CR	0.73	3.99	0.00	20.28	2.07	7.01	…	0.57
CR	1.18	1.93	2.03	13.66	4.56	6.55	…	1.15
KR	0.18	1.95	3.73	21.31	3.69	4.02	…	0.52
KR	0.54	0.04	5.05	16.89	2.29	2.95	…	1.28
KR	0.33	2.31	5.39	29.44	3.72	4.23	…	0.81
KR	0.43	1.40	9.49	10.38	1.35	2.92	…	0.64
KR	0.57	1.18	9.53	9.30	0.84	5.42	…	0.9
KR	0.43	1.88	6.65	14.61	2.03	2.75	…	1.03
PR	0.35	2.69	11.07	10.21	0.46	2.84	…	0.77
PR	0.45	4.65	8.77	11.77	0.33	4.03	…	0.82
PR	2.74	0.86	11.35	7.47	0.93	1.27	…	1.18
PR	0.87	0.60	16.12	5.22	0.28	3.95	…	1.05
PR	0.43	0.68	13.66	6.72	0.35	2.11	…	1.3
PR	0.31	1.68	11.77	8.17	0.43	1.47	…	1.76

Results

LDA extraction using all variables (53 volatiles)

The analysis was not feasible because there were a lot of variables and most of them have less than three valid cases (this fact is very important for discriminant analysis).

A stepwise analysis was carried out to select a limited number of variables that permit to carry out the analysis.

Stepwise analysis

- Forward stepwise analysis: variables are inserted one by one in the model starting from the x variable that most contributed to maximize distance among classes, then another variable is inserted. When the first x variable that does not affect distance among classes is identified by the analysis, it is removed from the model. This is removal process was repeated for all the other variables.

- Backward stepwise analysis: all variables are inserted in the model, then the x variables that does not affect distance among classes are removed from the system one by one and the LDA is repeated each time. When the first variable that affects the distance among classes is identified by the analysis, the procedure of removal is stopped.

Results

A forward stepwise analysis was carried out to select a limited number of variables which were used to carry out the analysis.

12 variables were selected

13 / 30 / 34 / 37 / 38 / 42 / 43 / 45 / 47 / 49 / 52 / 53

that permitted a 100% attribution of samples to each class.

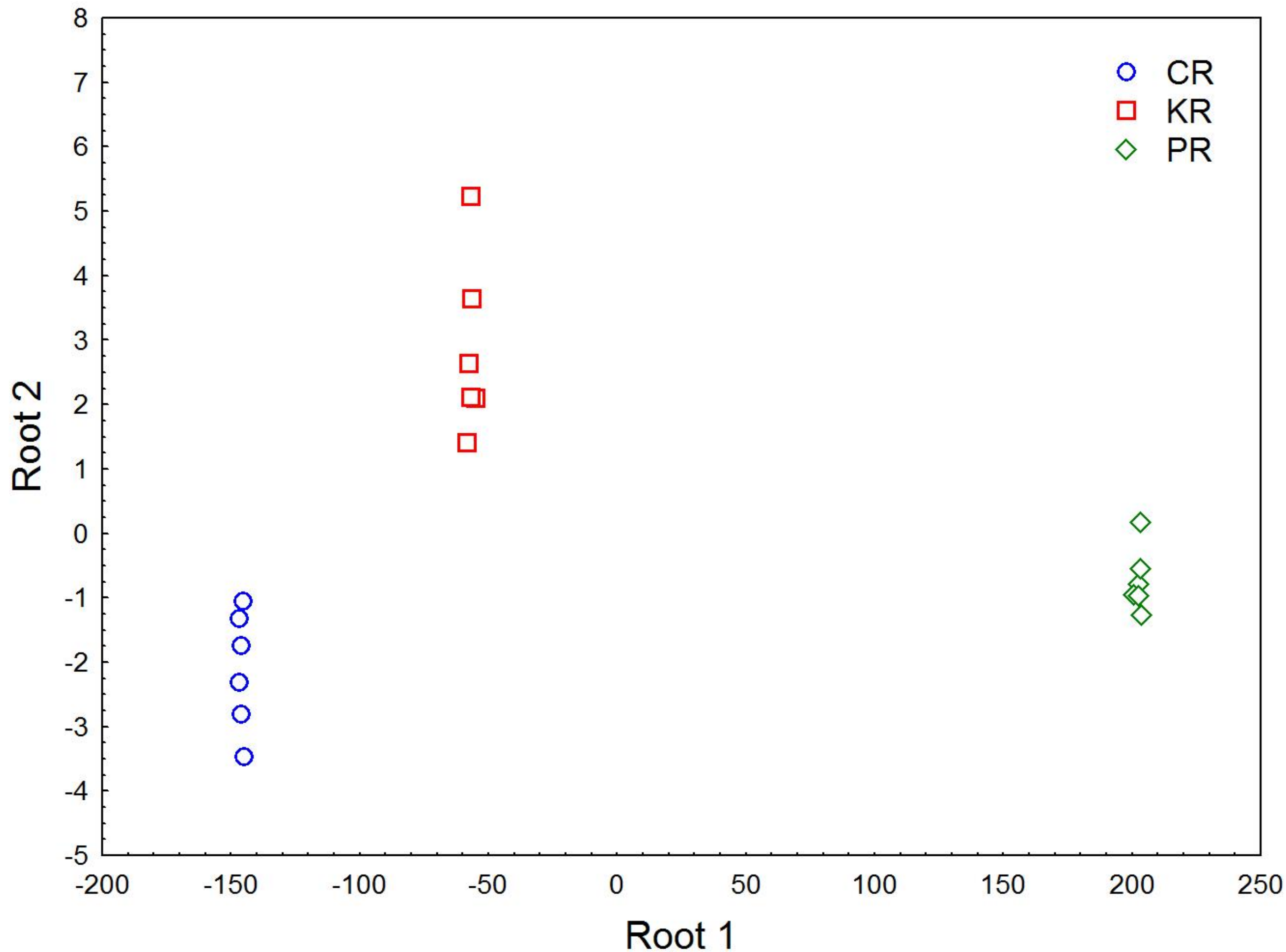
The results were visualized using CDA.

Variables

Variables (volatile compounds) selected by forward stepwise analysis.

Compound	IUPAC name	ID	PC1 loading	PC2 loading
acetone	propan-2-one	1	-	-
ethyl acetate	ethyl acetate	2	-	-
2-butanone	butan-2-one	3	-0.82	0.42
ethyl alcohol	ethanol	4	0.88	0.13
diacetyl	butane-2,3-dione	5	0.83	-0.23
2-pentanone	pentan-2-one	6	0.11	-0.75
1-ethanone	ethan-1-one	7	0.71	0.59
2-butanol	butan-2-ol	8	-0.87	0.30
3-methyl-(2 o 3)-heptanol	3-methylheptan-(2 o 3)-ol	9	-	-
thiophene	thiophene	10	0.80	0.15
1-propyl alcohol	propan-1-ol	11	-	-
ethyl butyrate	ethyl butanoate	12	-	-
methyl butyrate	methyl butanoate	13	-	-
2-hexanone	hexan-2-one	14	-	-
5-methyl-2-hexanone	5-methylhexan-2-one	15	-	-
hexanal	hexanal	16	0.73	-0.05
isobutyl alcohol	2-methylpropan-1-ol	17	-	-
3-methyl-2-butanol	3-methylbutan-2-ol	18	-	-
2-pentanol	pentan-2-ol	19	-	-
butyl alcohol	butan-1-ol	20	-	-
2-heptanone	heptan-2-one	21	0.03	-0.92
heptanal	heptanal	22	0.92	0.28
isoamyl alcohol	3-methylbutan-1-ol	23	-	-
ethyl hexanoate	ethyl hexanoate	24	-	-
2-methyl hexanoate	2-methyl hexanoate	25	-	-
1-pentanol	pentan-1-ol	26	0.93	0.21
2-octanone	octan-2-one	27	-0.16	-0.94
acetoin	3-hydroxybutan-2-one	28	0.87	0.08
octanal	octanal	29	-	-
1-heptanol	heptan-1-ol	30	-0.75	-0.03
isobutyl hexanoate	2-methylpropyl hexanoate	31	-0.70	0.42
hexanol	hexan-1-ol	32	0.76	0.53
2-methyl-3-pentanol	2-methylpentan-3-ol	33	-	-
2-nonanone	nonan-2-one	34	-0.26	-0.88
nonanal	nonanal	35	-	-
ethyl heptanoate	ethyl heptanoate	36	-	-
ethyl octanoate	ethyl octanoate	37	-	-
acetic acid	acetic acid	38	-0.87	0.09
8-nonen-2-one	non-8-en-2-one	39	-0.11	-0.82
propionic acid	propanoic acid	40	-0.84	0.27
2-nonenale	non-2-enal	41	-	-
benzaldehyde	benzaldehyde	42	-	-
2-undecanone	undecan-2-one	43	-	-
butyric acid	butanoic acid	44	-0.92	0.26
isovaleric acid	3-methylbutanoic acid	45	-0.79	0.22
2-thiopheneethanol	2-thiophen-2-yl ethanol	46	-	-
phenylacetaldehyde	2-phenylacetaldehyde	47	-	-
2-thiopheneacetic acid	2-thiophen-2-yl acetic acid	48	-	-
hexanoic acid	hexanoic acid	49	-0.86	0.28
phenethyl alcohol	2-phenylethanol	50	0.78	0.20
octanoic acid	octanoic acid	51	-	-
nonanoic acid	nonanoic acid	52	-	-
decanoic acid	decanoic acid	53	-0.06	0.75

CDA results


CR
KR
PR
Root 2
Root 1
8
7
6
5
4
3
2
1
0
-1
-2
-3
-4
-5
-200
-150
-100
-50
0
50
100
150
200
250

Considerations on Exercise (all variables)

LDA permitted to attribute all samples to their classes in the 100% of cases.

12 variables (volatile compounds) permitted the sample attribution to classes.

The results were well represented on the plane defined by the first two canonical variables (roots).