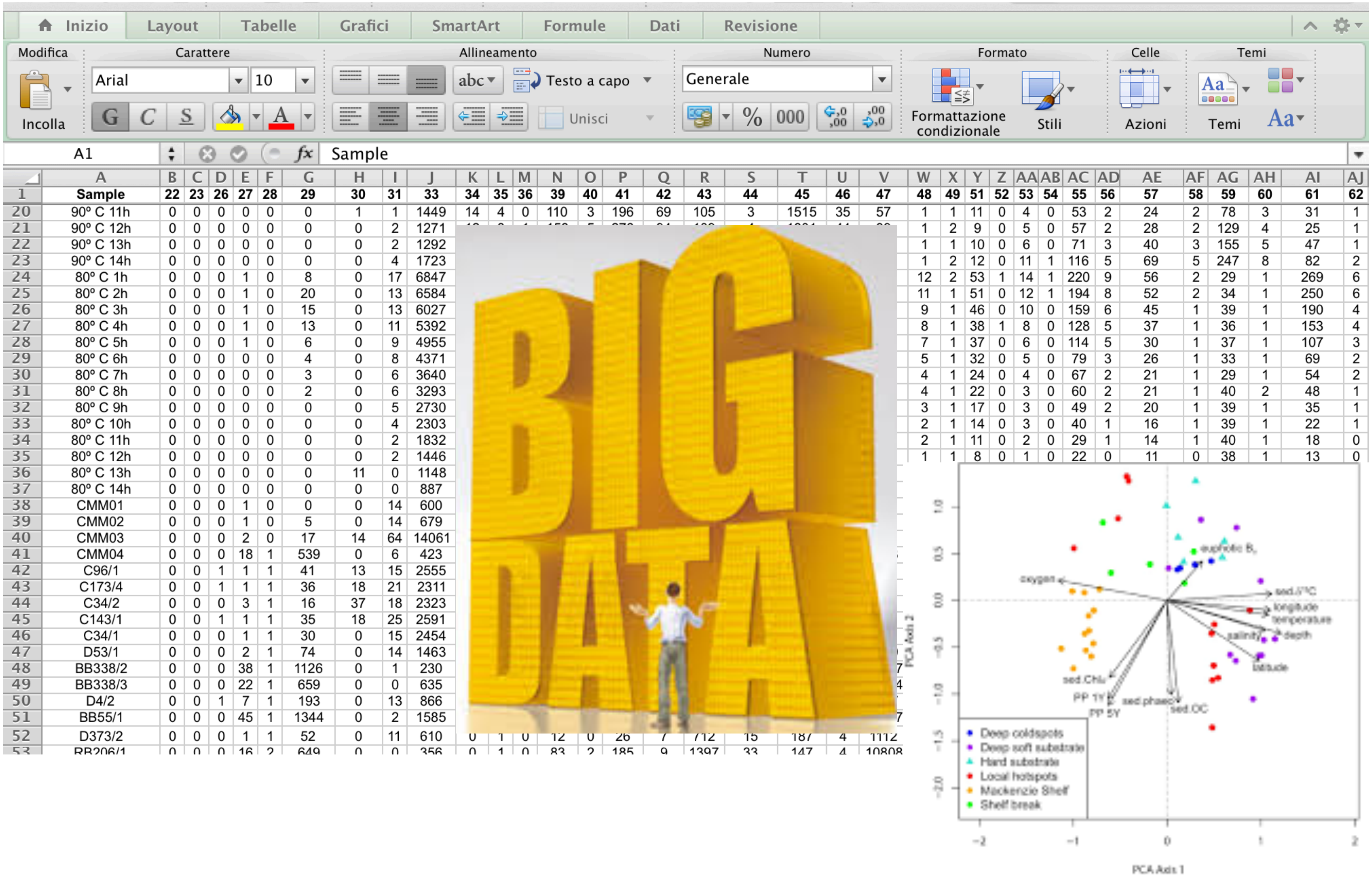
Multivariate data description: PCA

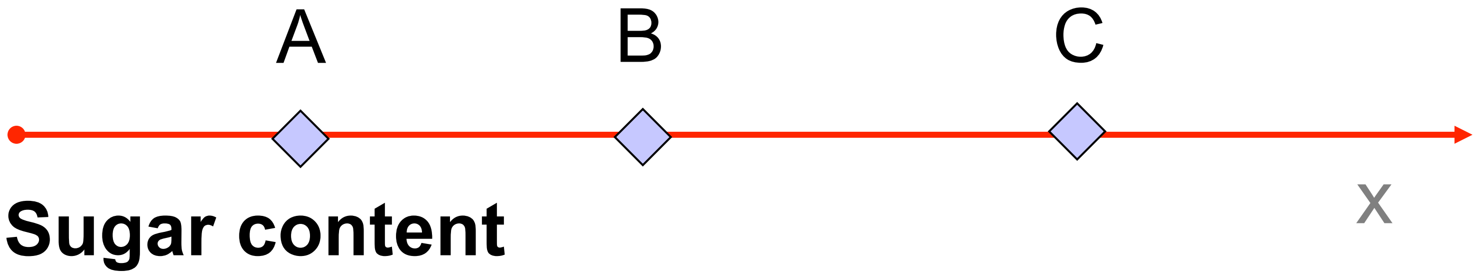


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ
1	**Sample**	**22**	**23**	**26**	**27**	**28**	**29**	**30**	**31**	**33**	**34**	**35**	**36**	**39**	**40**	**41**	**42**	**43**	**44**	**45**	**46**	**47**	**48**	**49**	**51**	**52**	**53**	**54**	**55**	**56**	**57**	**58**	**59**	**60**	**61**	**62**
20	90º C 11h	0	0	0	0	0	0	1	1	1449	14	4	0	110	3	196	69	105	3	1515	35	57	1	1	11	0	4	0	53	2	24	2	78	3	31	1
21	90º C 12h	0	0	0	0	0	0	0	2	1271													1	2	9	0	5	0	57	2	28	2	129	4	25	1
22	90º C 13h	0	0	0	0	0	0	0	2	1292													1	1	10	0	6	0	71	3	40	3	155	5	47	1
23	90º C 14h	0	0	0	0	0	0	0	4	1723													1	2	12	0	11	1	116	5	69	5	247	8	82	2
24	80º C 1h	0	0	0	1	0	8	0	17	6847													12	2	53	1	14	1	220	9	56	2	29	1	269	6
25	80º C 2h	0	0	0	1	0	20	0	13	6584													11	1	51	0	12	1	194	8	52	2	34	1	250	6
26	80º C 3h	0	0	0	1	0	15	0	13	6027													9	1	46	0	10	0	159	6	45	1	39	1	190	4
27	80º C 4h	0	0	0	1	0	13	0	11	5392													8	1	38	1	8	0	128	5	37	1	36	1	153	4
28	80º C 5h	0	0	0	1	0	6	0	9	4955													7	1	37	0	6	0	114	5	30	1	37	1	107	3
29	80º C 6h	0	0	0	0	0	4	0	8	4371													5	1	32	0	5	0	79	3	26	1	33	1	69	2
30	80º C 7h	0	0	0	0	0	3	0	6	3640													4	1	24	0	4	0	67	2	21	1	29	1	54	2
31	80º C 8h	0	0	0	0	0	2	0	6	3293													4	1	22	0	3	0	60	2	21	1	40	2	48	1
32	80º C 9h	0	0	0	0	0	0	0	5	2730													3	1	17	0	3	0	49	2	20	1	39	1	35	1
33	80º C 10h	0	0	0	0	0	0	0	4	2303													2	1	14	0	3	0	40	1	16	1	39	1	22	1
34	80º C 11h	0	0	0	0	0	0	0	2	1832													2	1	11	0	2	0	29	1	14	1	40	1	18	0
35	80º C 12h	0	0	0	0	0	0	0	2	1446													1	1	8	0	1	0	22	0	11	0	38	1	13	0
36	80º C 13h	0	0	0	0	0	0	11	0	1148																										
37	80º C 14h	0	0	0	0	0	0	0	0	887																										
38	CMM01	0	0	0	1	0	0	0	14	600																										
39	CMM02	0	0	0	1	0	5	0	14	679																										
40	CMM03	0	0	0	2	0	17	14	64	14061																										
41	CMM04	0	0	0	18	1	539	0	6	423																										
42	C96/1	0	0	1	1	1	41	13	15	2555																										
43	C173/4	0	0	1	1	1	36	18	21	2311																										
44	C34/2	0	0	0	3	1	16	37	18	2323																										
45	C143/1	0	0	1	1	1	35	18	25	2591																										
46	C34/1	0	0	0	1	1	30	0	15	2454																										
47	D53/1	0	0	0	2	1	74	0	14	1463																										
48	BB338/2	0	0	0	38	1	1126	0	1	230																										
49	BB338/3	0	0	0	22	1	659	0	0	635																										
50	D4/2	0	0	1	7	1	193	0	13	866																										
51	BB55/1	0	0	0	45	1	1344	0	2	1585																										
52	D373/2	0	0	0	1	1	52	0	11	610	0	1	0	12	0	26	7	712	15	187	4	1112														
53	RB206/1	0	0	0	16	2	649	0	0	356	0	1	0	83	2	185	9	1397	33	147	4	10808														

One dimensional data

Three samples, A, B, and C, could differ just for the extent of **one variable** (C > B > A)

Example:



One dimension (x) and three values are sufficient for data description

One dimensional data matrix

A one per three matrix is required to describe the data set:

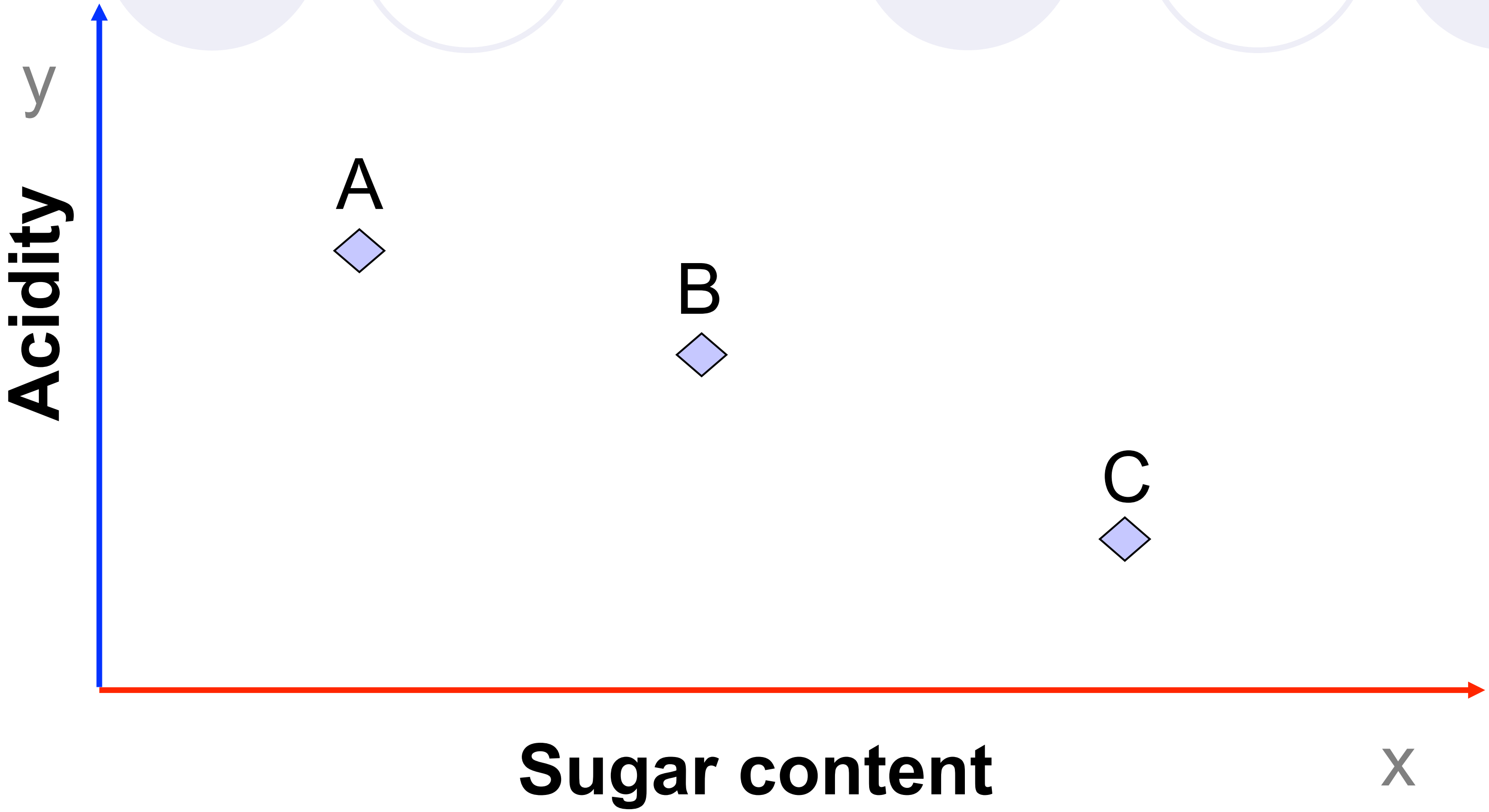
each sample is identified by one values (x) or score

the variable (x) has three cases (A, B, C) corresponding to samples

	Sugar content
A	x_A
B	x_B
C	x_C

Three samples, A, B, and C, could differ for **two variables**

Example:



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

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

Two dimensional data matrix

A two per three matrix is required to describe the data set:

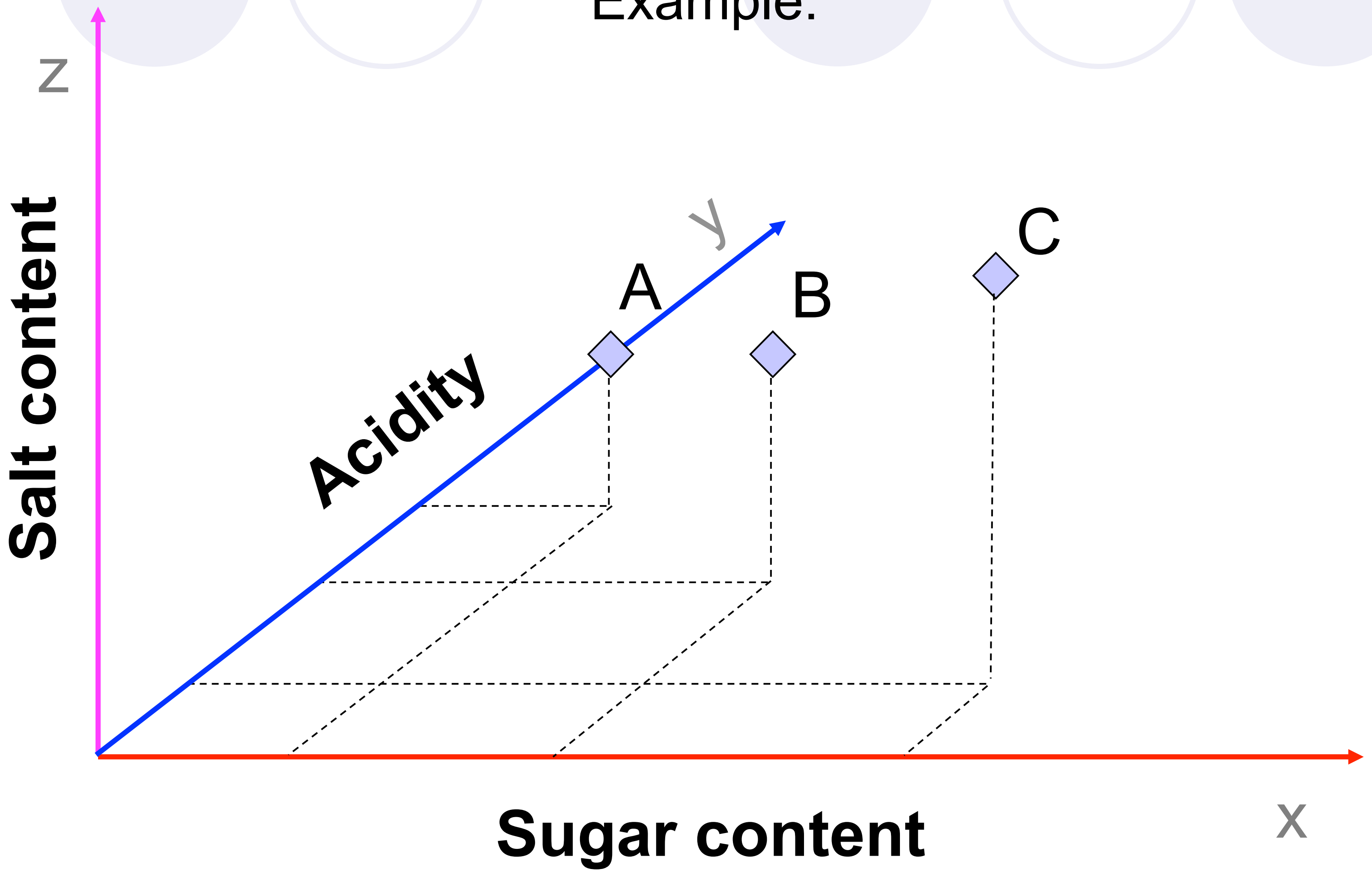
each sample is identified by two values (x, y) or scores

each variable has three cases (A, B, C) corresponding to samples

	Sugar content	Acidity
A	x_A	y_B
B	x_B	y_B
C	x_C	y_C

Three samples, A, B, and C, could differ for **three variables**

Example:



Three dimensions (x, y, z) are required for data description; each sample is identified by three values (x, y, z)

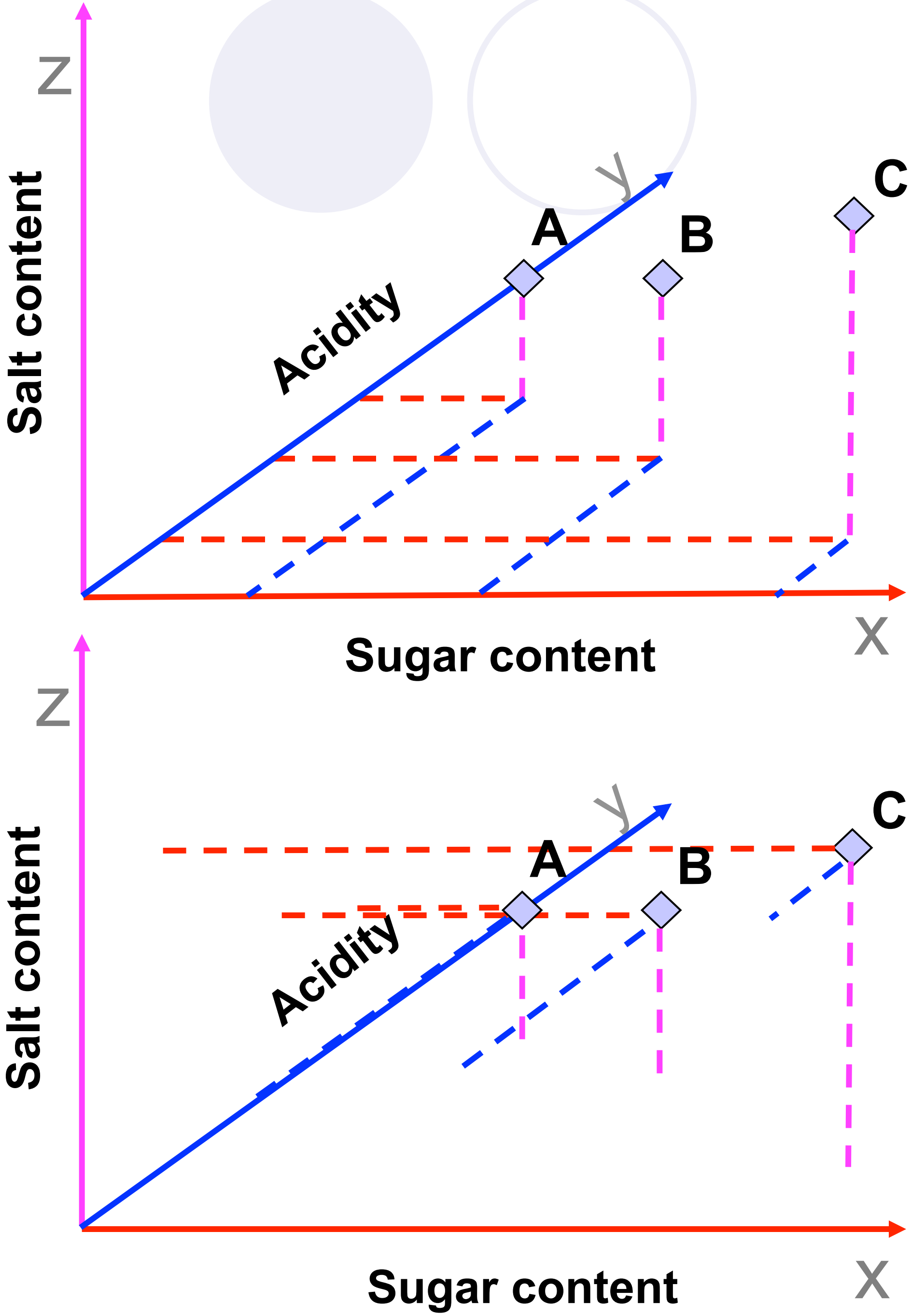
Three dimensional data matrix

A three per three matrix is required to describe the data set:

each sample is identified by three values (x, y, z) or scores

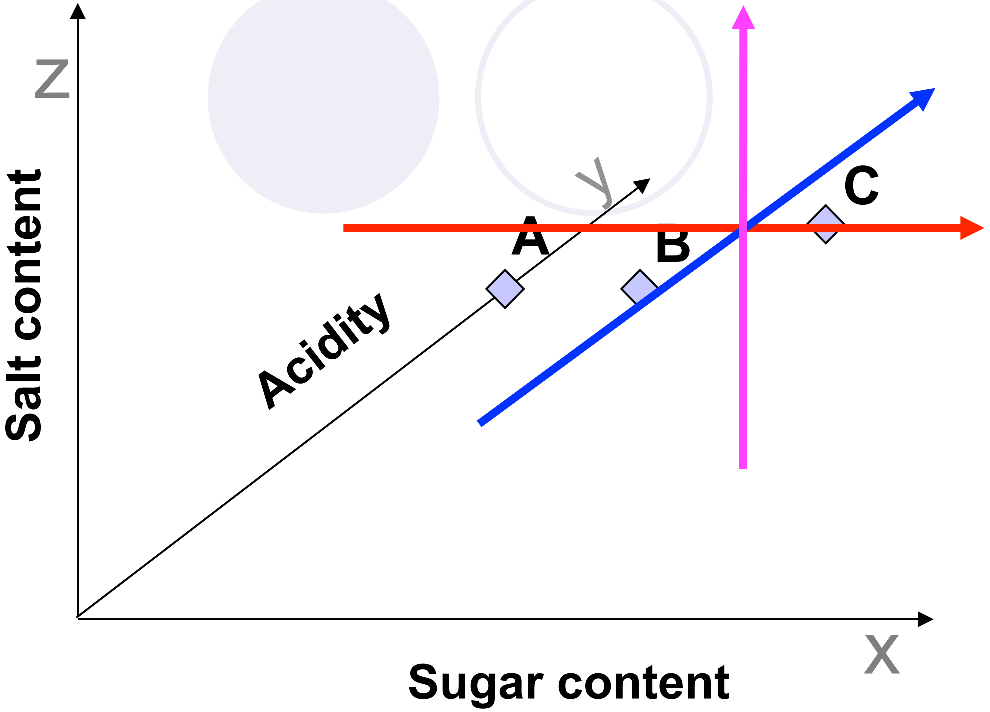
each variable has three cases (A, B, C) corresponding to samples

	Sugar content (x)	Acidity (y)	Salt content (z)
A	x_A	y_A	z_A
B	x_B	y_B	z_B
C	x_C	y_C	z_C



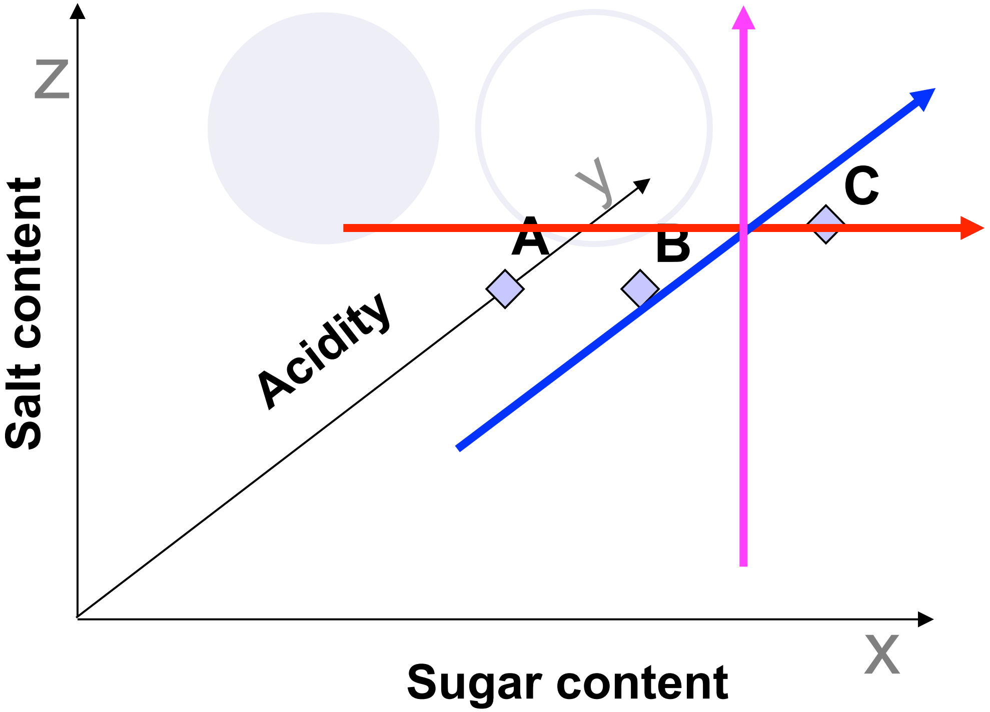
In a tridimensional cartesian space each sample is identified by three points with x, y, z coordinates (scores).

Each coordinate represent the intensity or extent of a variable and is perpendicular to the plane defined by the other two variables

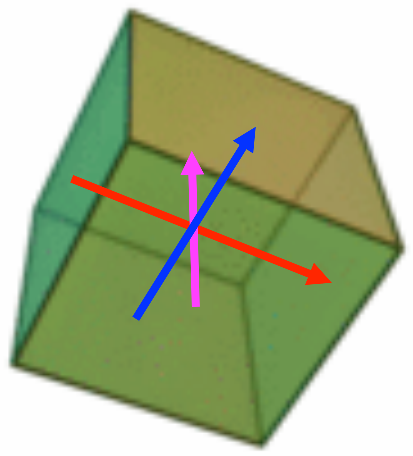


The variables are described by vectors, orthogonals among them, that intersecates in one point defined as 0,0,0.

The vectors correspond to the original variables and are described by a module (vector length) which describes their intensity (extent) and an angle which is of 90°.



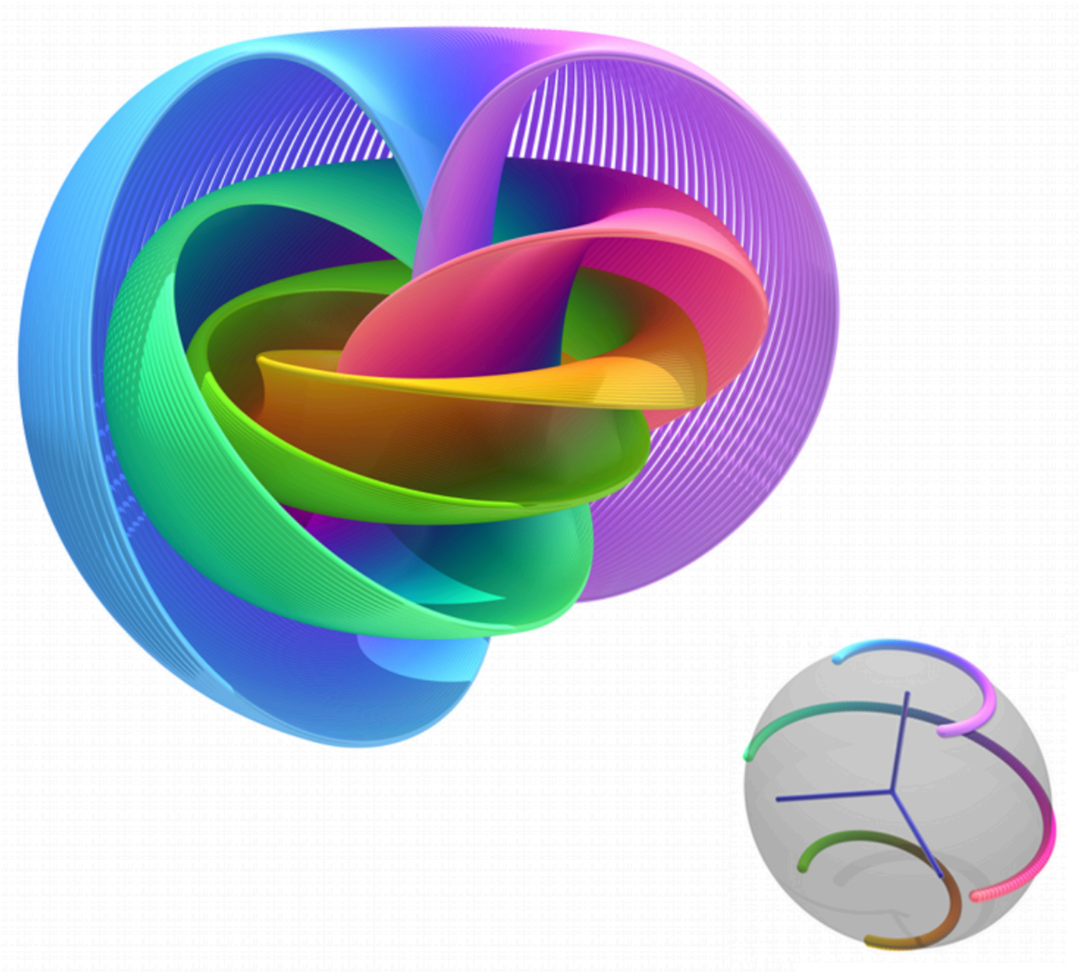
A tridimensional (3D) space is described by a polygonal solid with three opposite 'faces' (six faces in a whole) called *exahedron*. In this solid the opposite faces are orthogonal among them.



Description of multidimensional data (1)

- What if we have a lot of variables that describes our samples?
- In such a case a multidimensional (n dimensional) space occurs to describe data distribution.
- In such a case a n dimensional matrix occurs to describe the data set.

How to describe a n-dimensional space?

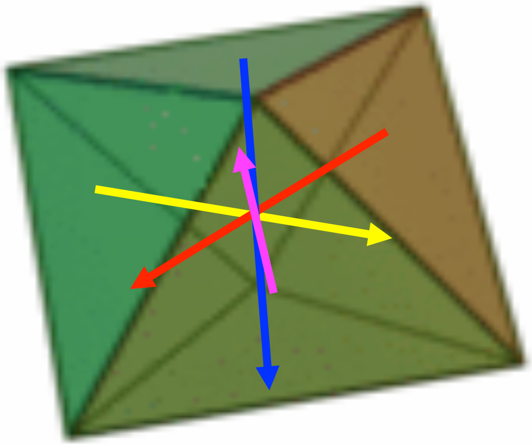


Let’s ask help to geometry ….

If an hexaedron could enable us
to figure out a 3D space,
then another orthogonal solid
could help us anyway?

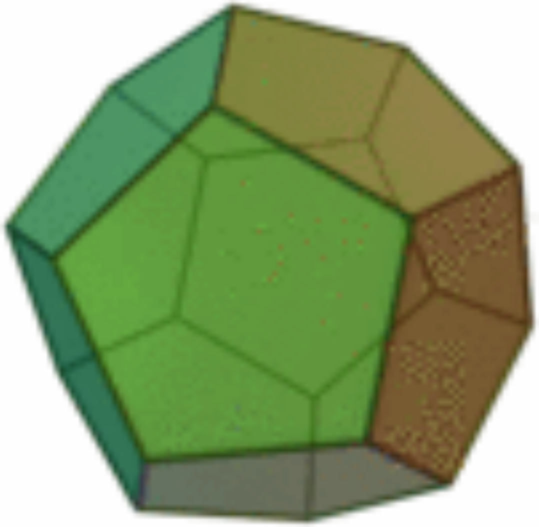
Depiction of vectorial spaces D > 3

octahedron



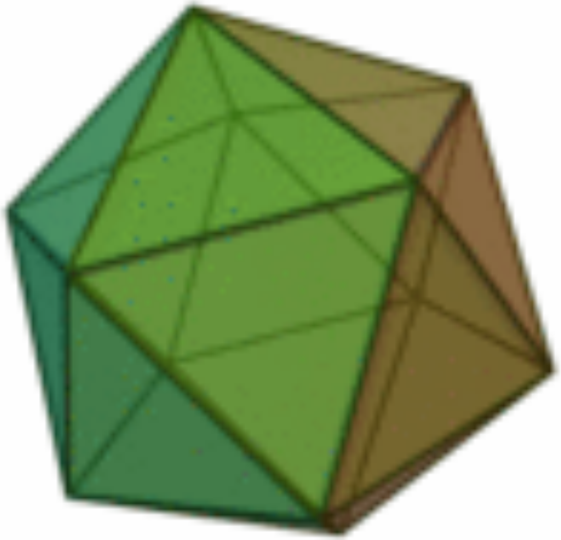
D = 4

dodecahedron



D = 6

icosahedron



D = 10

Description of multidimensional data (2)

- It is not easy to depict a vectorial space with 7, 8, 9 or *n* dimensions (since orthogonal polygonal solids are limited).

- The visual representation of vectors (identifying variables) and coordinates (identifying scores) in a nD space is not easy.

- nD spaces could still be easily mathematically described with data matrices.

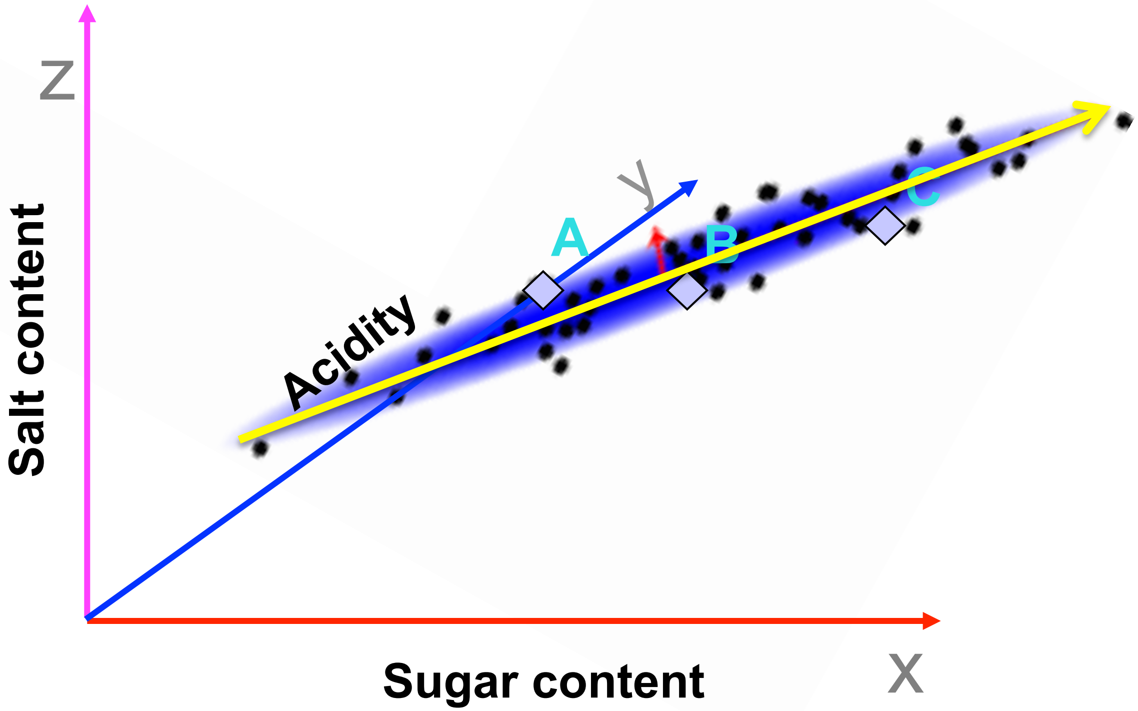
Principal component analysis (PCA)

- PCA is a statistical descriptive analysis that enable the analyst to describe a system by using new variables (latent variables) which are a linear transformation of the original variables and are not correlated among them.

- PCA could be 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.

Latent variables in data structure

Let’s return to our original A, B and C samples and imagine them as a part of a big data set.



The maximum variance of the data set represented in the 3D graph is along the direction indicated by the yellow line.

Explained variance maximization (1)

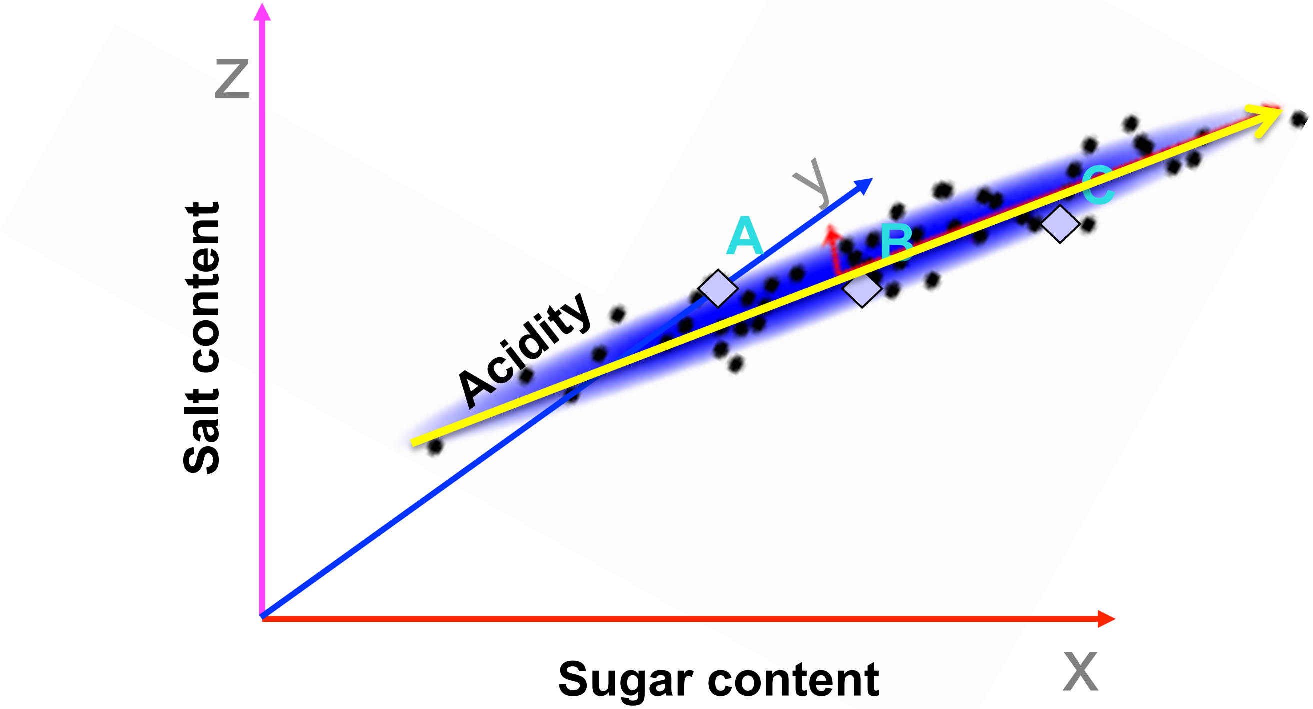
- Each parameter (variable) of a data set could be described by a media and a variance value, which synthesize the information on the distribution of data values.

- Three variables representing three parameters could be described by three media and three variance value, which synthesize the information on the distribution of the data values.

- The maximum variance of multidimensional data could not be along the original variables.

PCA representation

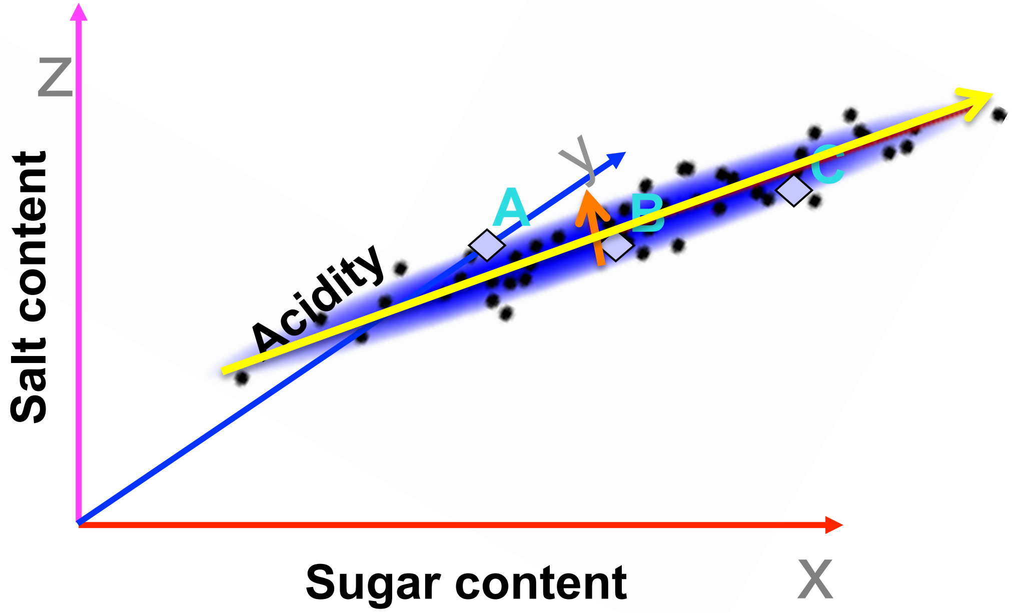
The vector that describes the yellow line is a new variable (factorial variable) that is called PC1 or first principal component.



The PC1 brings always along with it the maximum explained variance.

Explained variance maximization

The vector perpendicular to PC1 is a new variable (factorial variable) that is called PC2 or second principal component.

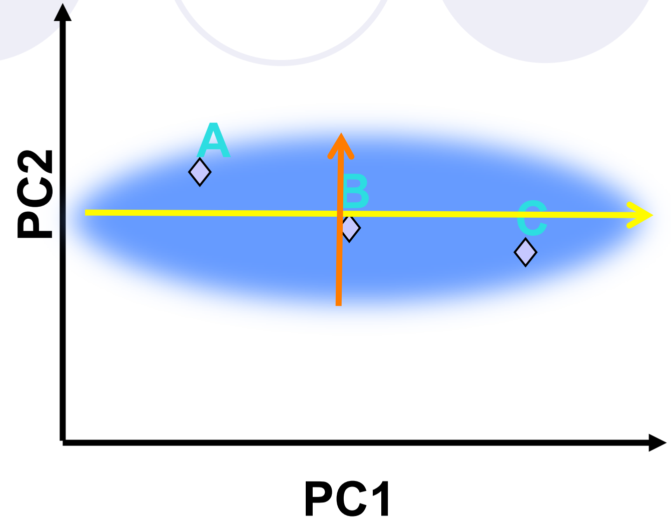


The PC2 (orange line) brings along with it the maximum variance not explained by PC1.

Principal components

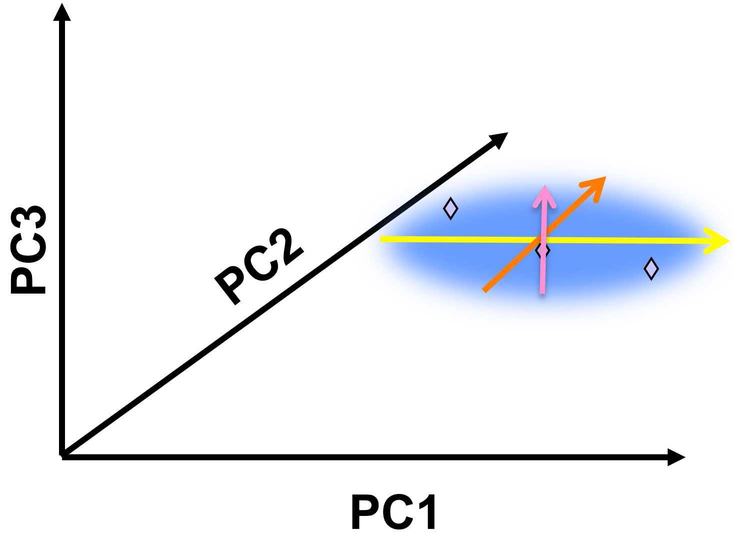
PC1 and PC2 could not explain the same part of variance since they are orthogonal among them by definition, thus they are not correlated among them.

PC1 and PC2 form a 2D plane.



PC1 and PC2 could preceed a PC3 (pink line) orthogonal to both of them.

PC1, PC2 and PC3 form a 3D space.



Big data set

- Starting from 3 variables 3PCs, which describes the 100% of the variance, could be calculated.
- Starting from n variables nPCs, which describes the 100% of the variance, could be calulated.
- In this case, PCA does not help to manage the complexity of the system.

Description of multidimensional data

Multidimensional data could bring along with them a lot of information

Final paradigm:

Too much information = no information

We could not get use of all this information!

Large data set in our mind

- How could we manage large data set in our mind?
- We ‘summarize’ the information
- We keep the most pertinent information

Most pertinent information?

- If there is no variance along a variable, it means that our data could not differ among them for that variable.

- If there is a lot of variance along a variable there is more probability that our data could differ among them for that variable.

- In this case variance could be used as a criterion for ‘pertinence’.

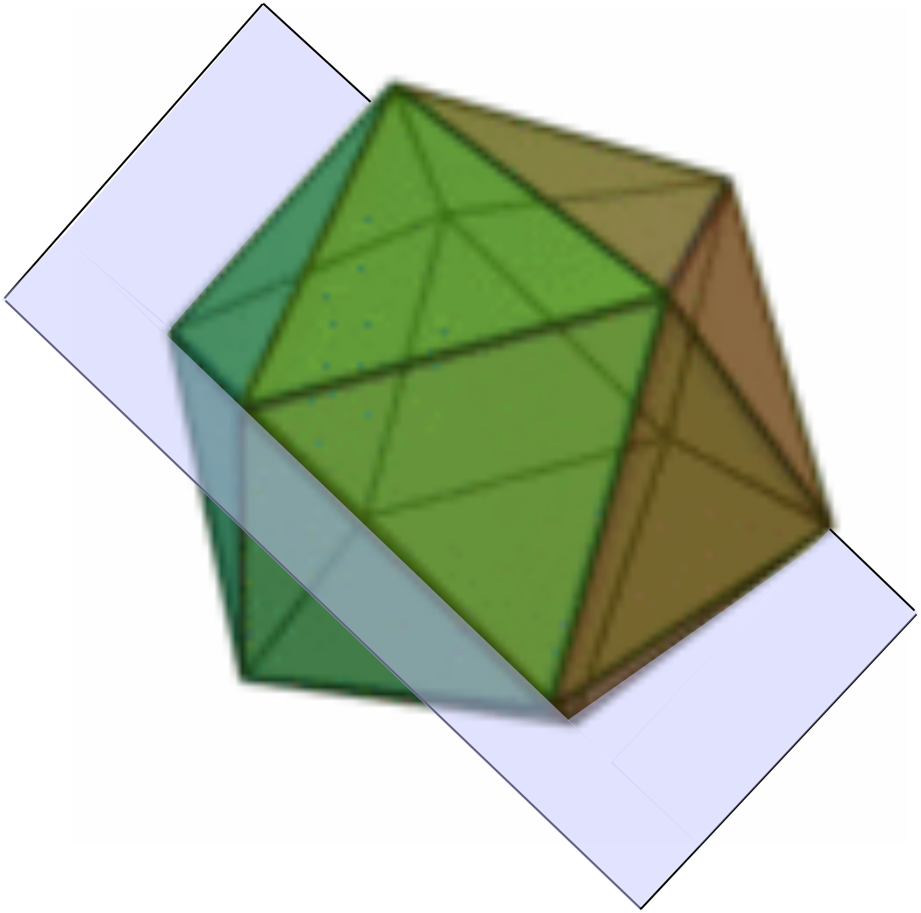
Dimensionality reduction (PCA)

- PCA could be also 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 PCs will bring along with them the maximum explained variance for definition.

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

Figurative exemplification



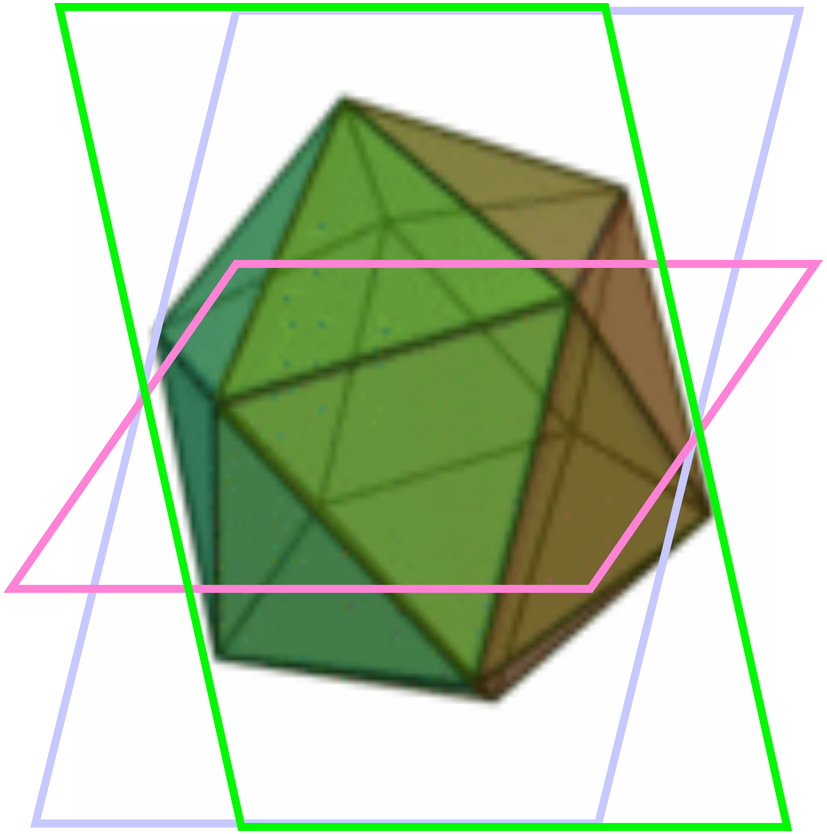
The reduction of dimensionality of a 10D space to a 2D space could be seen as a cut of the solid space with a 2D plane that intersecates the solid by passing trough the origin of axes (geometrical centre of solid).

How PCA operates to reduce dimensionality?

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

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

PCA choose the slope that permits to **maximize the variance** explained from the new bidimensional space.



Maximization of explained variance (2)

- The reduction of dimensionality determines unavoidably a loss of information.

Example: the reduction of a 3D system to a 2D one (with no width) implies a loss of information.

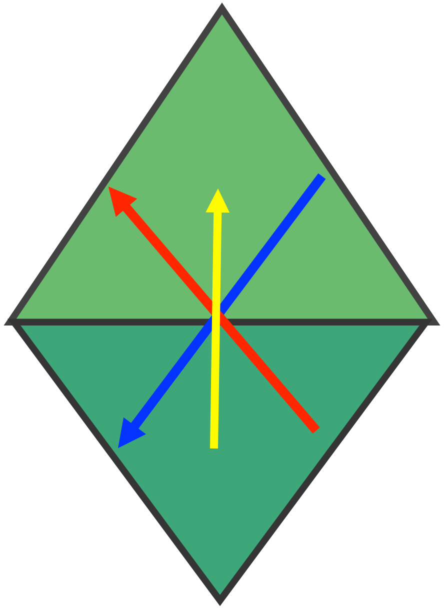
- If a 3D system is reduced to 2D by eliminating one dimension (by observing it orthogonally to width) the third dimension will be completely lost. The information that the third dimension brings along with it will be lost too.

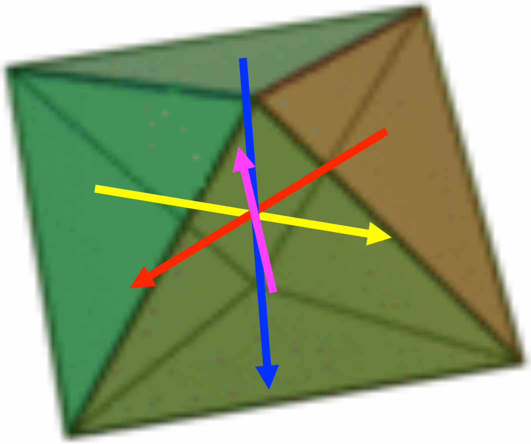
- However, if a 3D system is reduced to 2D by observing it axonometrically, less information is lost since the perception of width will remain in the brain.

Loss of explained variance (4)

low information

high information





orthogonal vision

2 faces and 3 vectors are visible
(yellow vector masks fucsia one)

axonometric vision

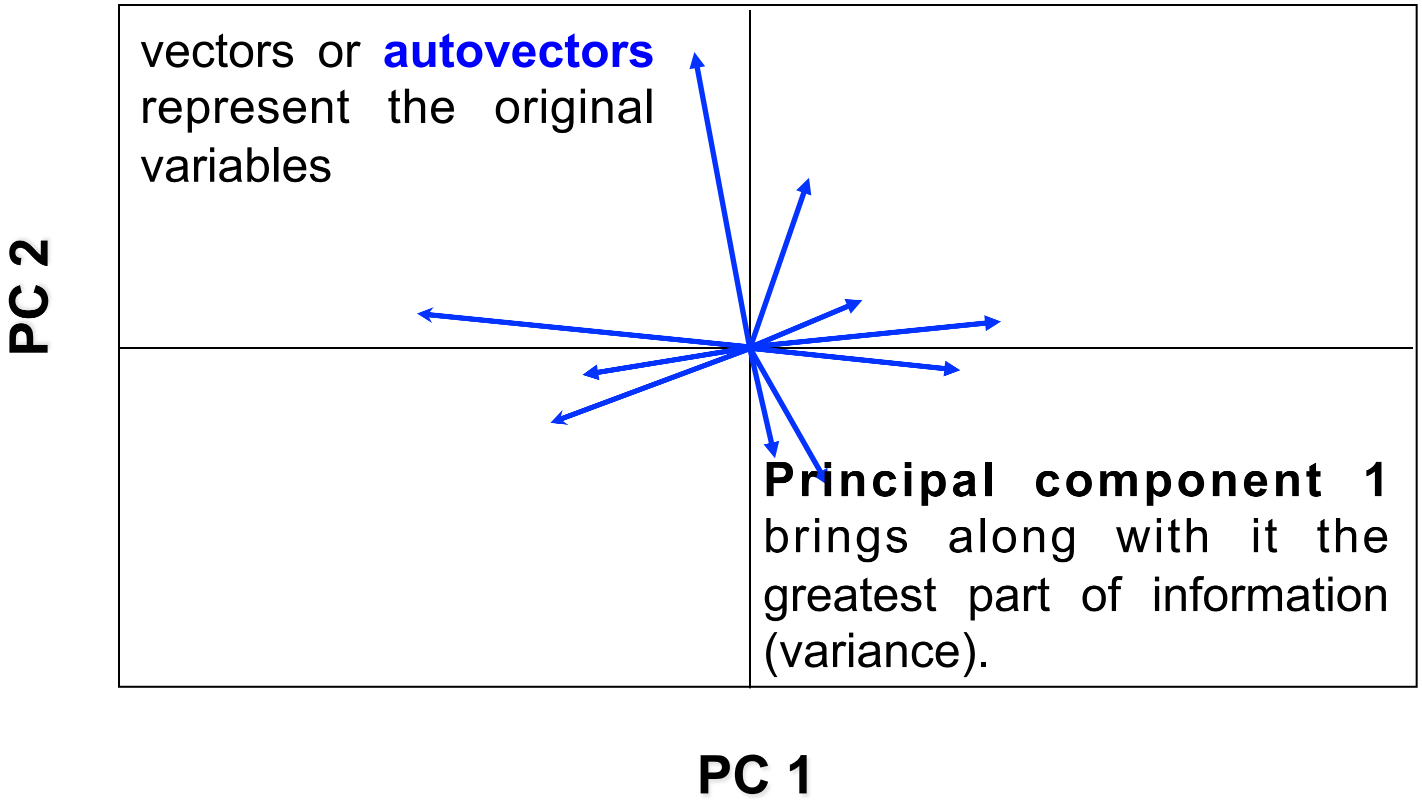
4 faces and 4 vectors are visible

PCA for data description

- Orthogonal projection of original variables and scores on the plane described by the first two PCs (or on the space described by the first three PCs) has been extensively used for big data set description.
- This approach permits to visualize samples scores along the directions depicted by PCs, which explain the maximum variance.

Graphical representation of PCA (1)

- From the orthogonal projection of original vectors on cutting plane, a 2D graph called **loading plot** is obtained
- Each of the two dimensions is called **principal component** or **PC** (x = **1**; y = **2**)

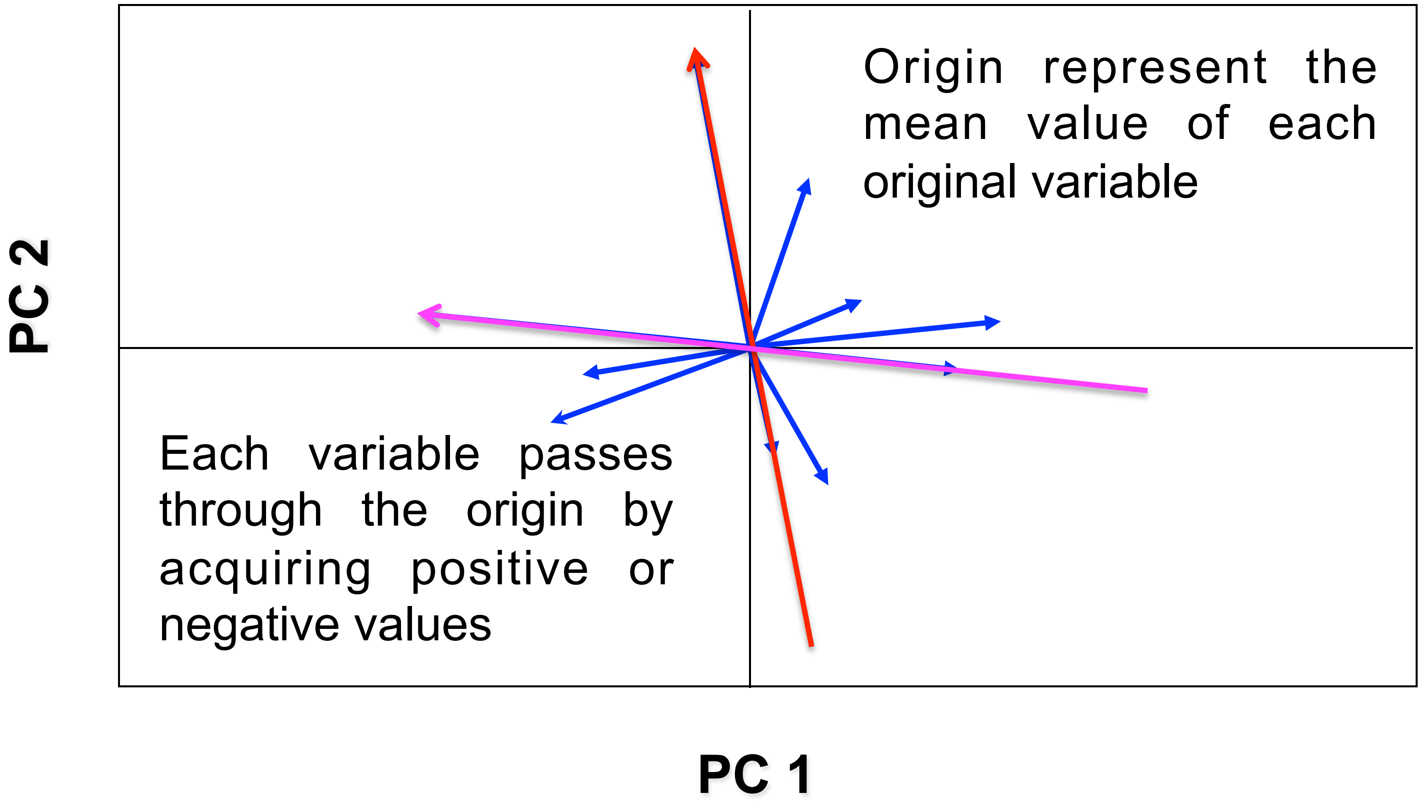


vectors or **autovectors** represent the original variables

Principal component 1 brings along with it the greatest part of information (variance).

Graphical representation of PCA (1)

- In the PCA graph, the origin represent the geometrical centre of the nD space which has been cut by a plane to reduce the dimensionality.

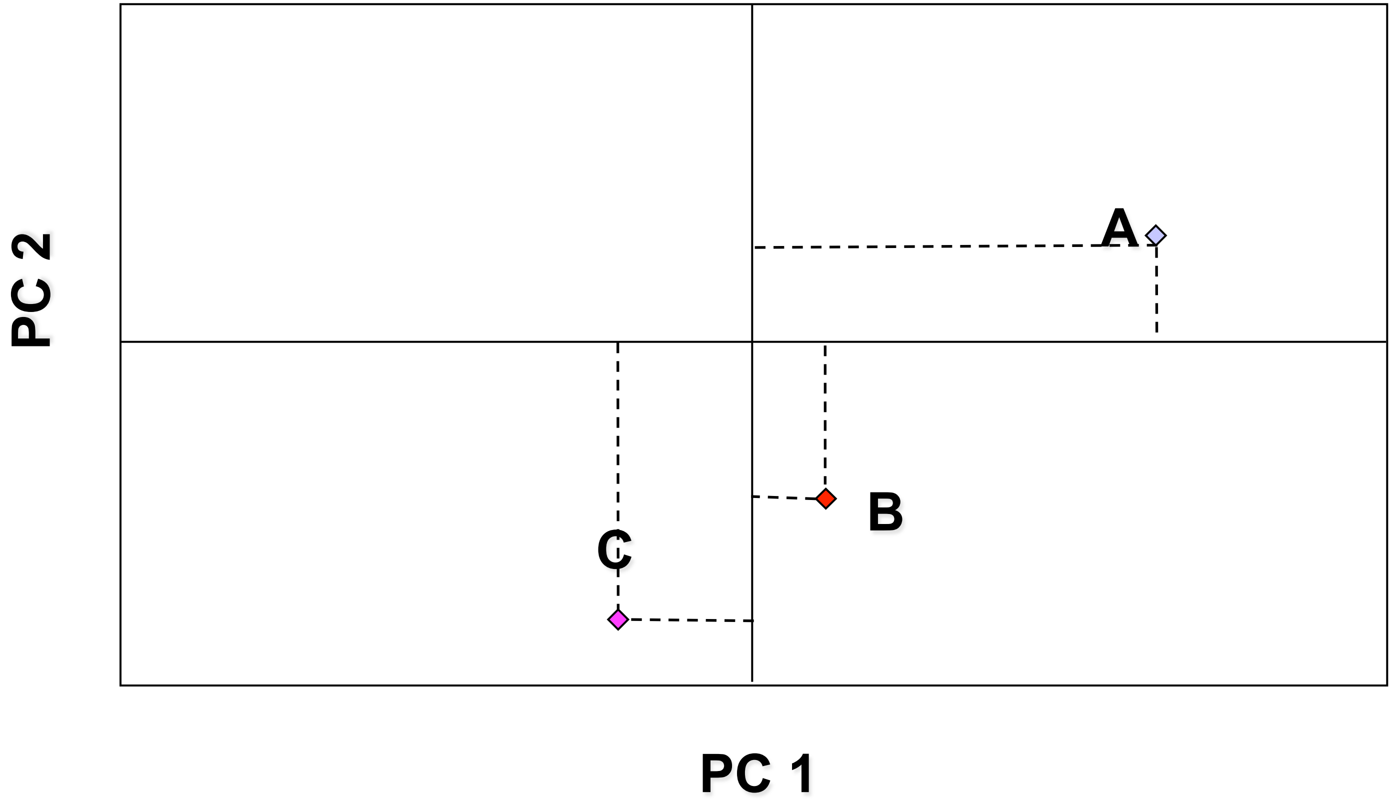


Origin represent the mean value of each original variable

Each variable passes through the origin by acquiring positive or negative values

Graphical representation of PCA (2)

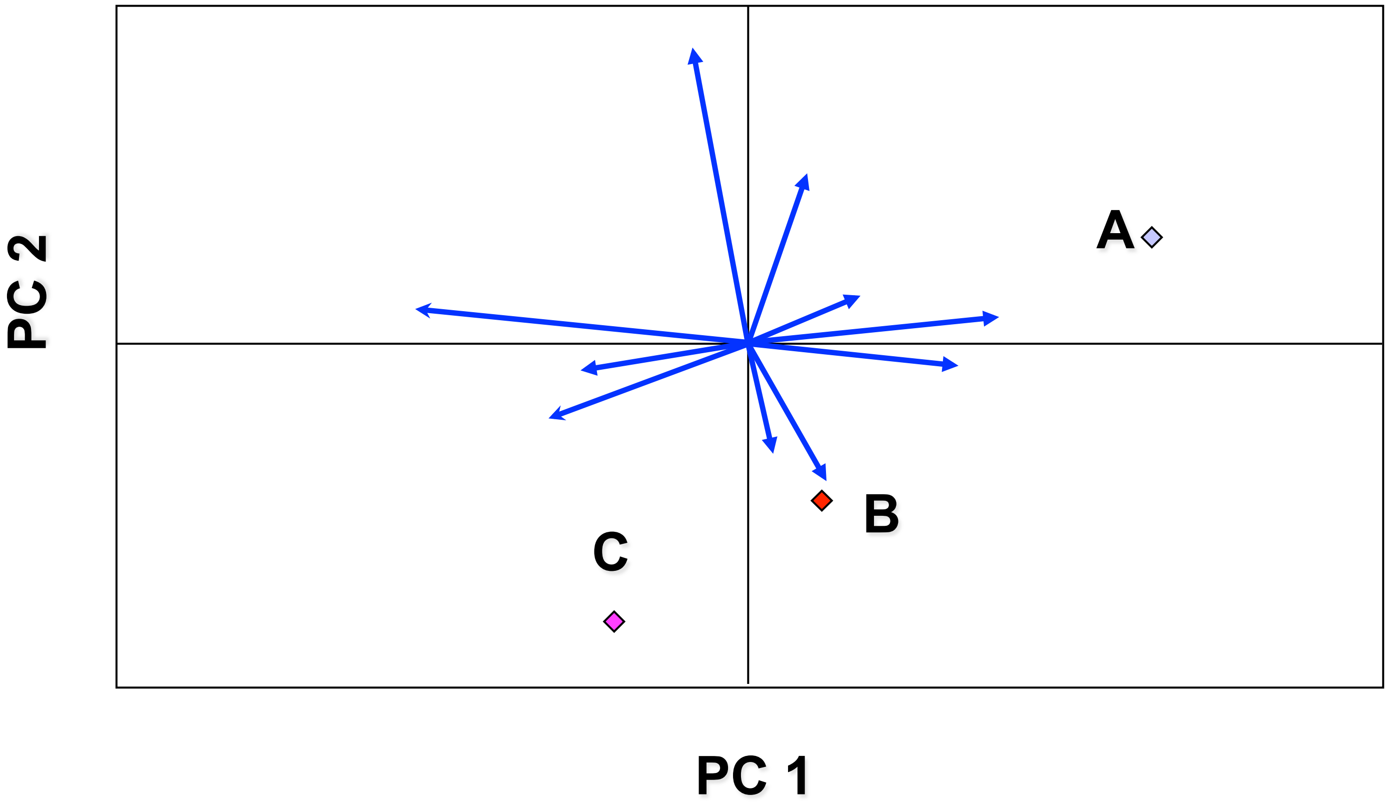
Also the scores of samples could be be identified on the plane of the first two principal components:



This graph is called **scores plot**

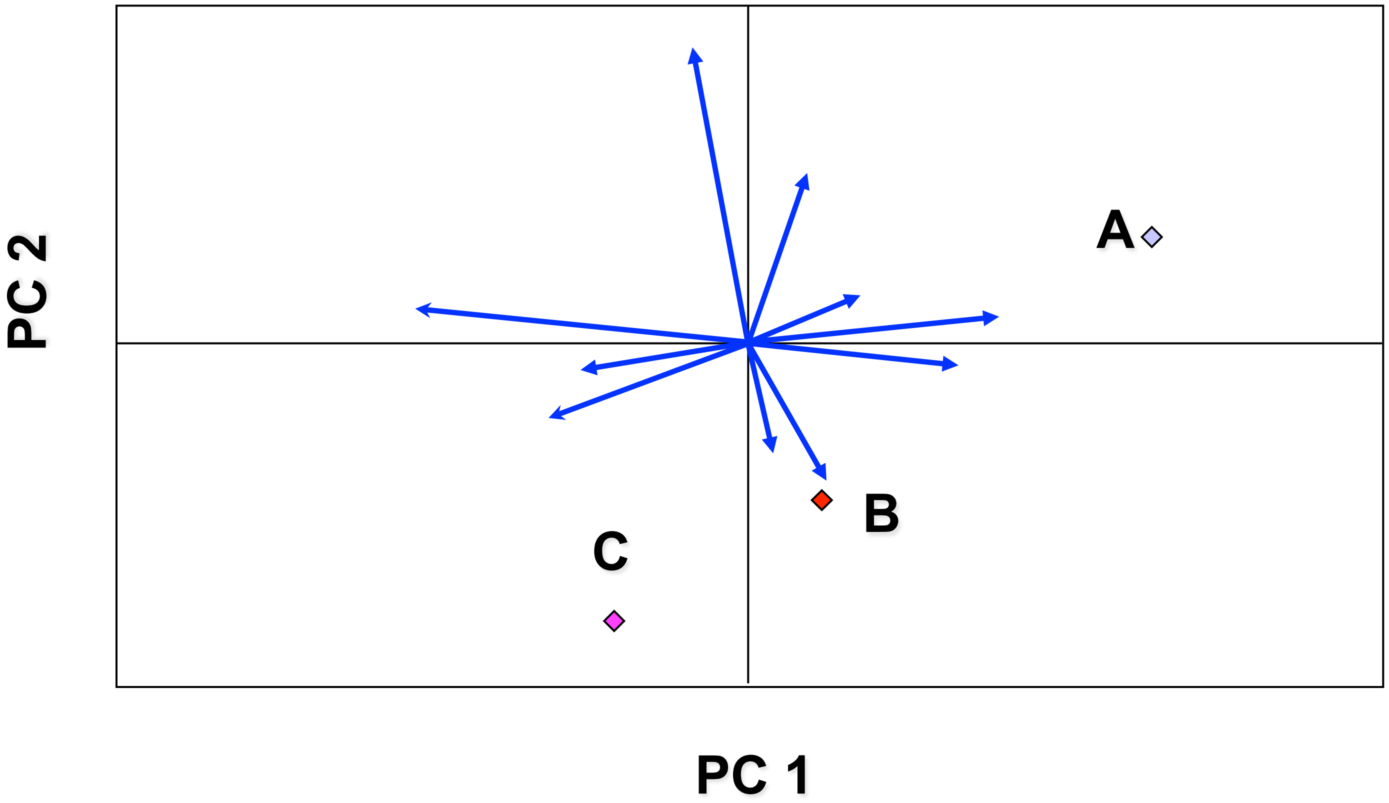
Graphical representation of PCA (3)

Loading and scores biplot



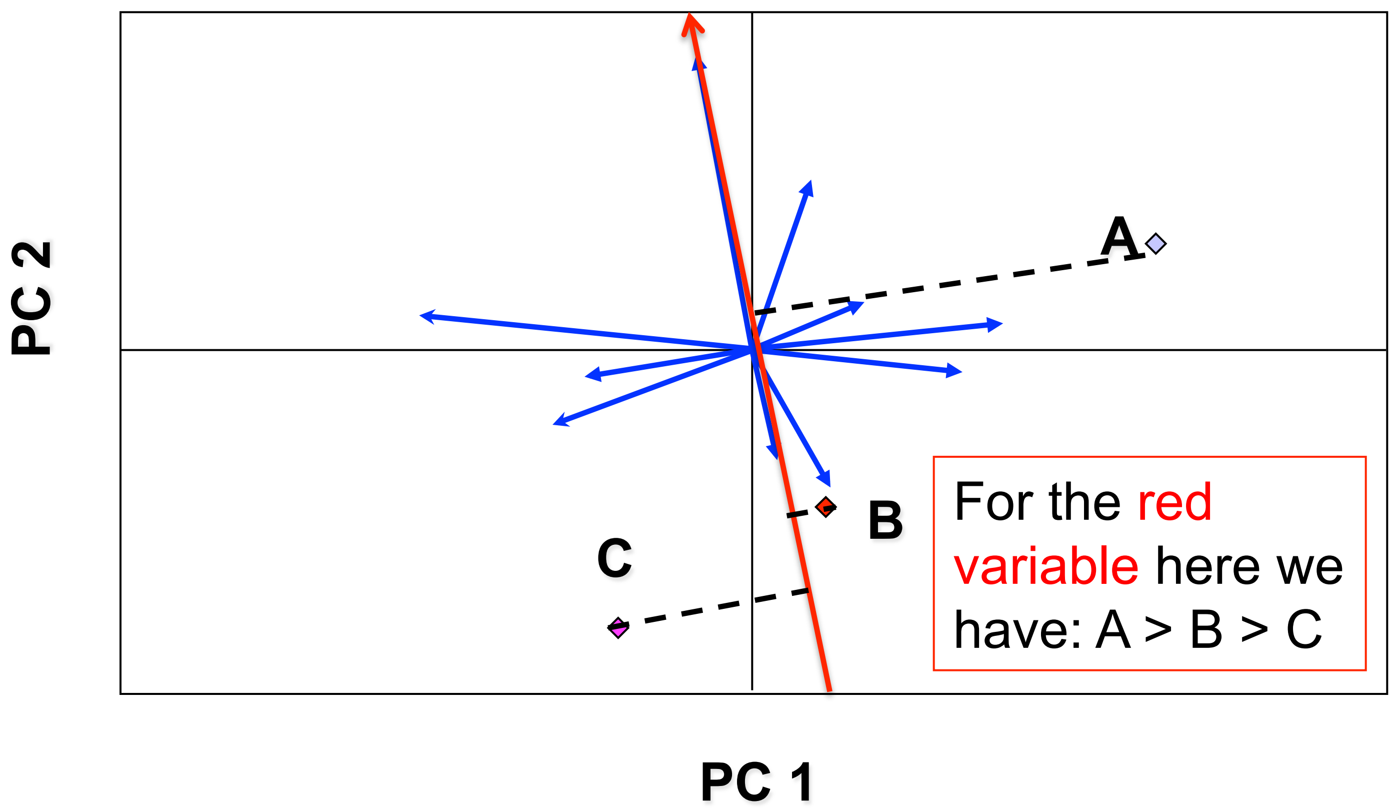
Graphical representation of PCA (3)

How do the scores of the samples on the original variables could be represented on the plane described by PCs?



Graphical representation of PCA (4)

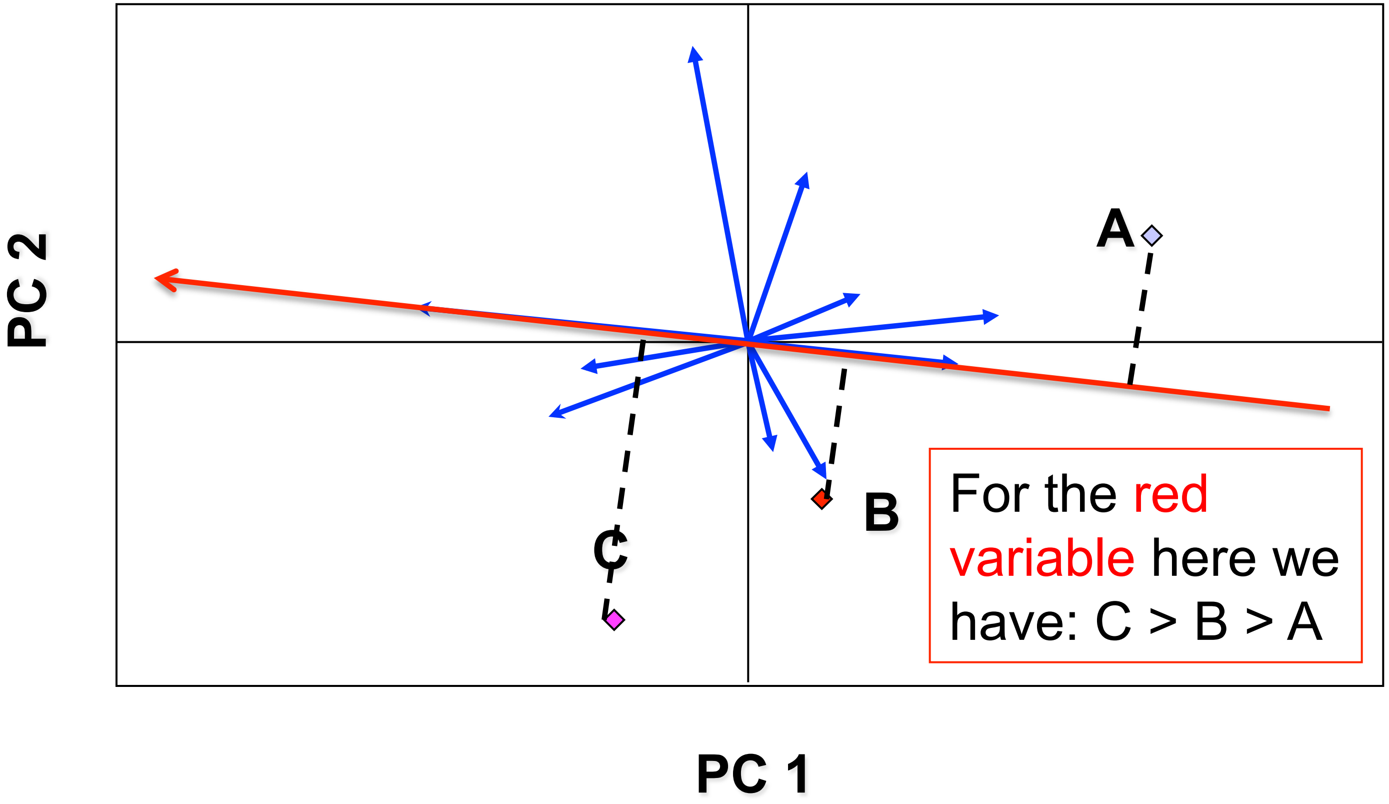
The scores of the samples on each original variables could be visualized by the orthogonal projection of PCs value on the selected variable



For the red variable here we have: A > B > C

Graphical representation of PCA (4)

The scores of the samples on each original variables could be visualized by the orthogonal projection of PCs value on the selected variable



For the red variable here we have: C > B > A

In statistical terms

Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called ‘principal components’.

The number of principal components is less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data set as possible), and each succeeding component, in turn, has the highest possible variance under the constraint that it is orthogonal to the preceding components. The resulting vectors are an uncorrelated orthogonal basis set.

In statistical terms

If a multivariate dataset is visualised as a set of coordinates in a high-dimensional data space (1 axis per variable), it is difficult to interpret.

PCA can supply the user with a lower-dimensional picture, a projection or "shadow" of this object when viewed from its most informative (higher variance) viewpoint.

This is done by using only the first few principal components so that the dimensionality of the transformed data is reduced.

If we consider only two or three principal components to describe a big multivariate data set, we could use PCA as a dimension reducing tool for explorative statistical analysis.

Exercize 1

A data set consisting of a set volatiles compounds from cheese samples samples obtained by using three different rennets (CR, KR, PR) and aged for different times (2 to 180 days) has been provided as EDCF01DEC2016.xlsx file.

Define the variables and the samples.

Carry out PCA analysis;

Calculate the variance explained up to the second PC;

Create the loadings plot and the scores plot using 2 PCs;

Carry out PCA analysis by using only the variables with a loading higher than 0.7 on the first two PCs;

Calculate the variance explained up to the third PC;

Create the loadings plot and the scores plot.

Data set

18 samples

3 rennet type

x

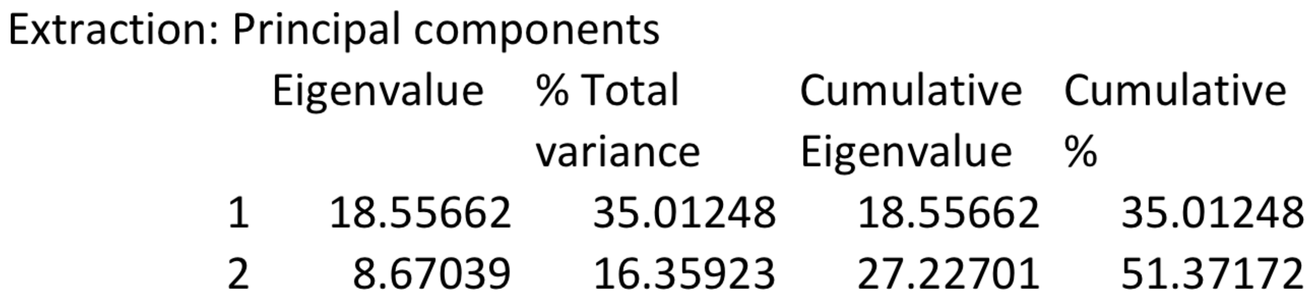
6 aging times

53 variables (volatile compounds)

18 x 53 data matrix

Results

PCA extraction using all variables (53 volatiles)

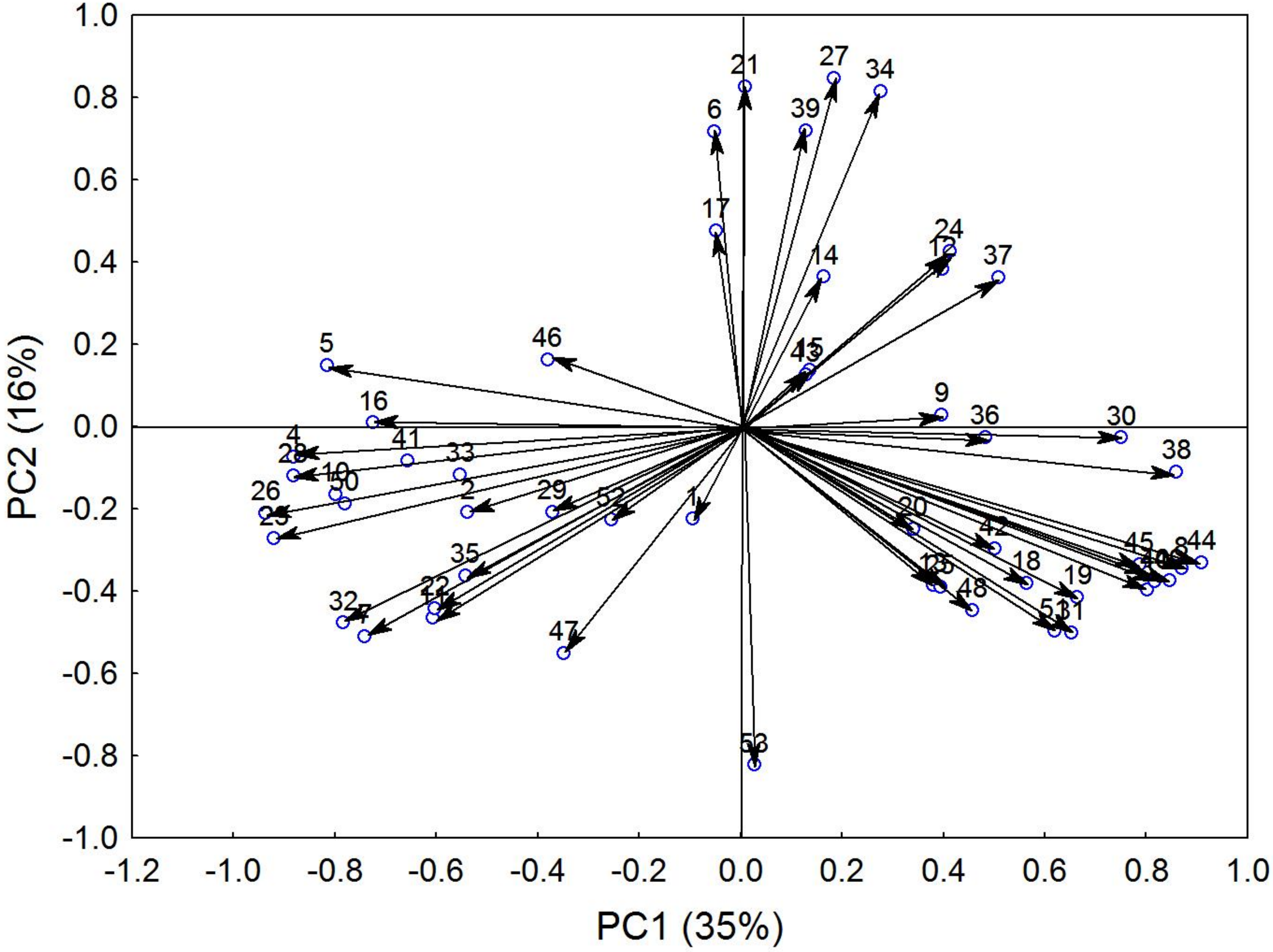


Extraction: Principal components

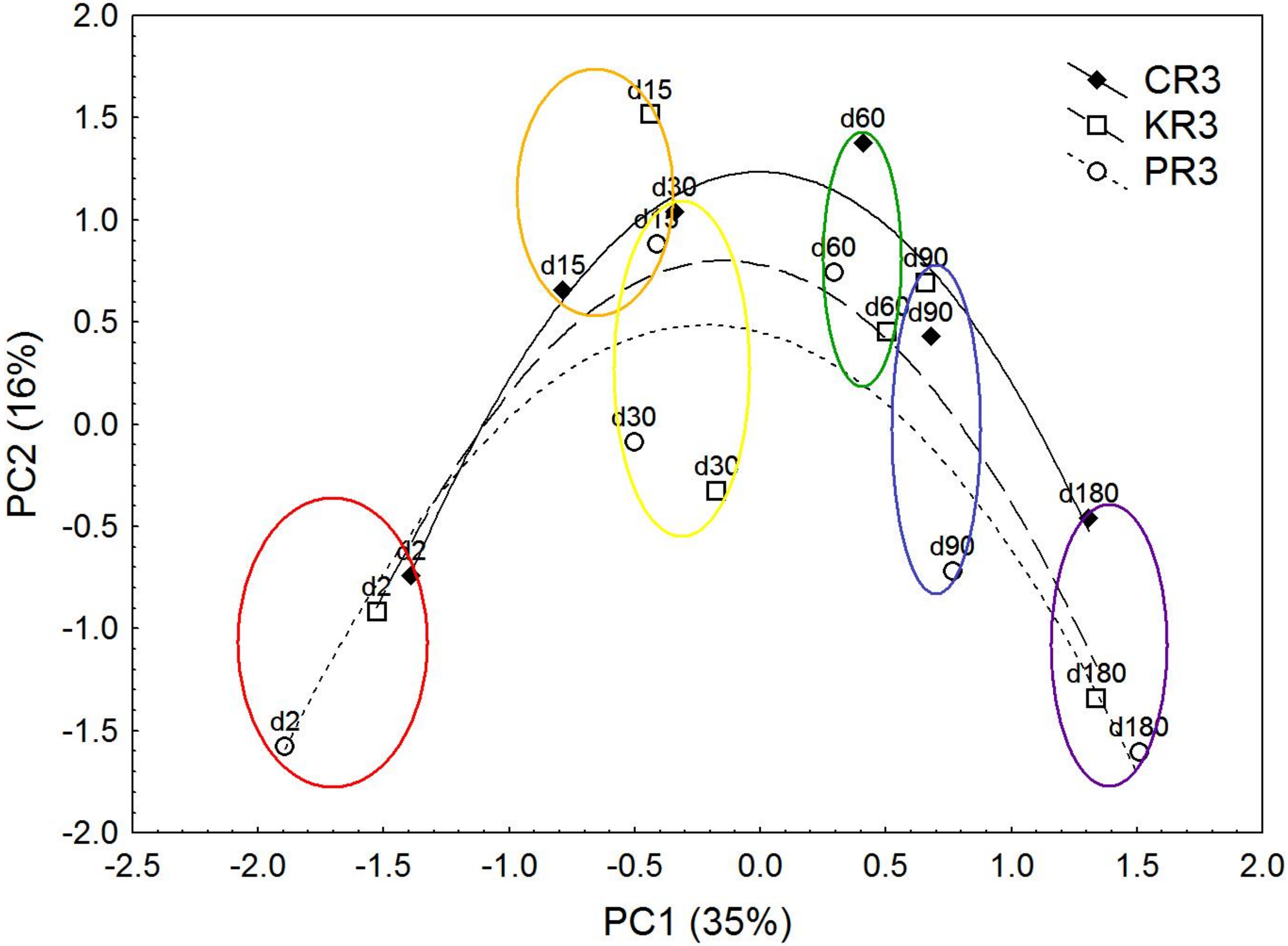
	Eigenvalue	% Total variance	Cumulative Eigenvalue	Cumulative %
1	18.55662	35.01248	18.55662	35.01248
2	8.67039	16.35923	27.22701	51.37172

The percentage of variance explained by the first two principal components is slightly above 50%.

Loadings plot (all variables)



Scores plot (all variables)



Considerations on Exercise 1 (all variables)

PCA permitted to reduce the data dimensionality and to separate samples along PC1 on the basis of ripening time.

PC2 seems to separate the samples on the basis of the rennet used in the cheese-making process.

Ripening time determines the highest variance in data structure and its effect could be observed along PC1.

Ripening time effect overwhelms that of the cheese making technology which could be observed along PC2.

The overall explained variance is low (51%).

How to increase explained variance?

- We could remove variables that gives little contribution to the overall data variance.

- The variance induced by these variables could be intended as a result of intrinsic variability of data.

- This is an assumption and should be taken with care.

Variables selection

- Criterion for selecting variables

- Loading on PCs (generally > 0.70)

- Modality of selection

- Stepwise analysis
  - Forward stepwise (insert the first most important variable and carry out PCA, then insert the second most important and so on, until the first non important appears in the model)
  - Backward stepwise (remove the less important variable and carry out PCA, then remove the second less important and so on, until only important remain in the model)

Data set 2 (variables selection)

18 samples

3 rennet type

x

6 aging times

25 variables (volatile compounds with loading > 0.70 on the first two PCs)

18 x 25 data matrix

Results

PCA extraction using only variables with a loading > 0.7 on the first two PCs (25 volatiles)

Variables selection was peformed by backward stepwise analysis

Extraction: Principal components

	Eigenvalue	% Total variance	Cumulative Eigenvalue	Cumulative %
1	12.55856	50.23426	12.55856	50.23426
2	6.03499	24.13995	18.59355	74.37421

The percentage of variance explained by the first two principal components is above 74%.

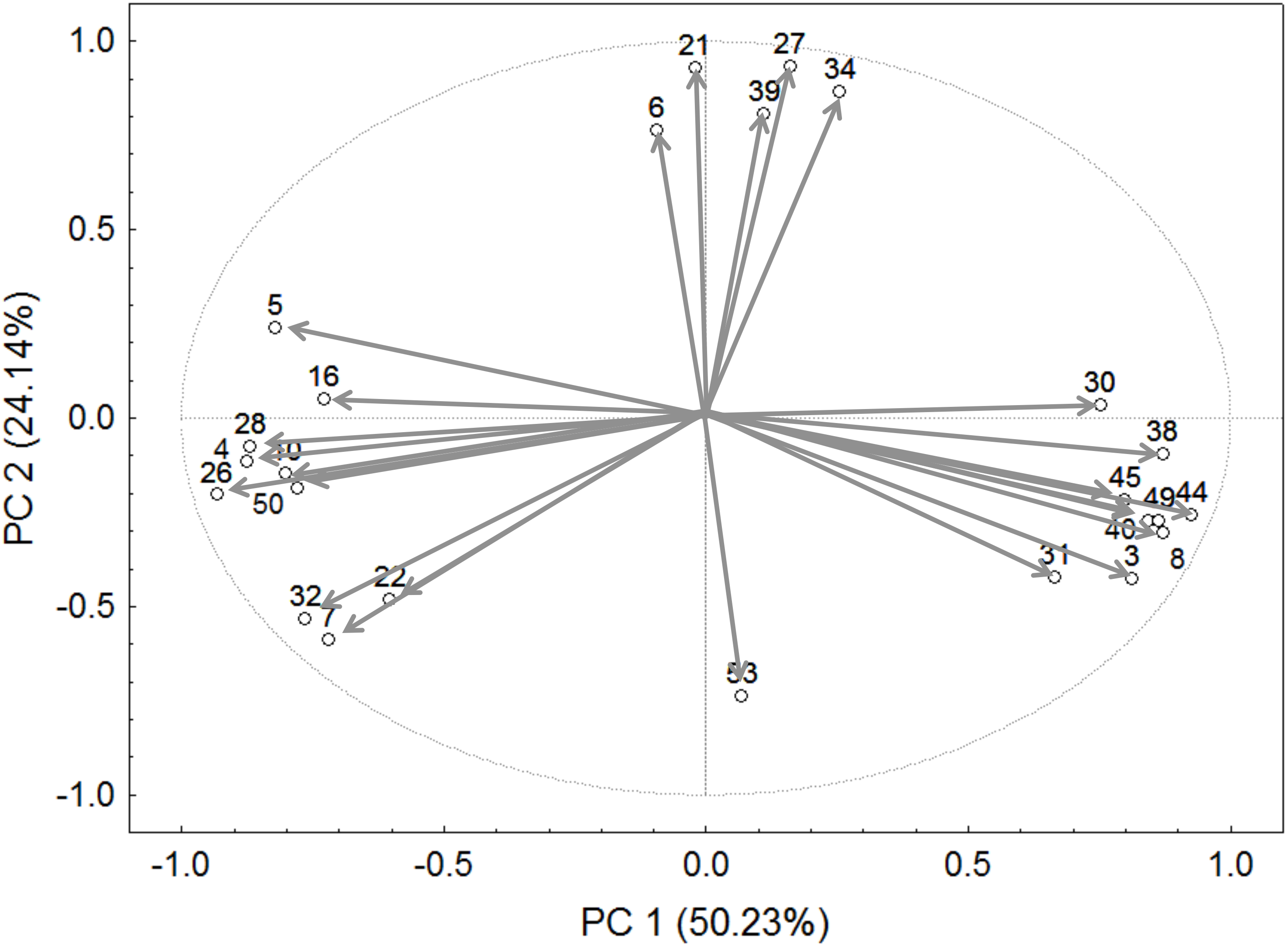
Loadings

Variables (volatile compounds) with a loading higher than 0.7 on the first two PCs.

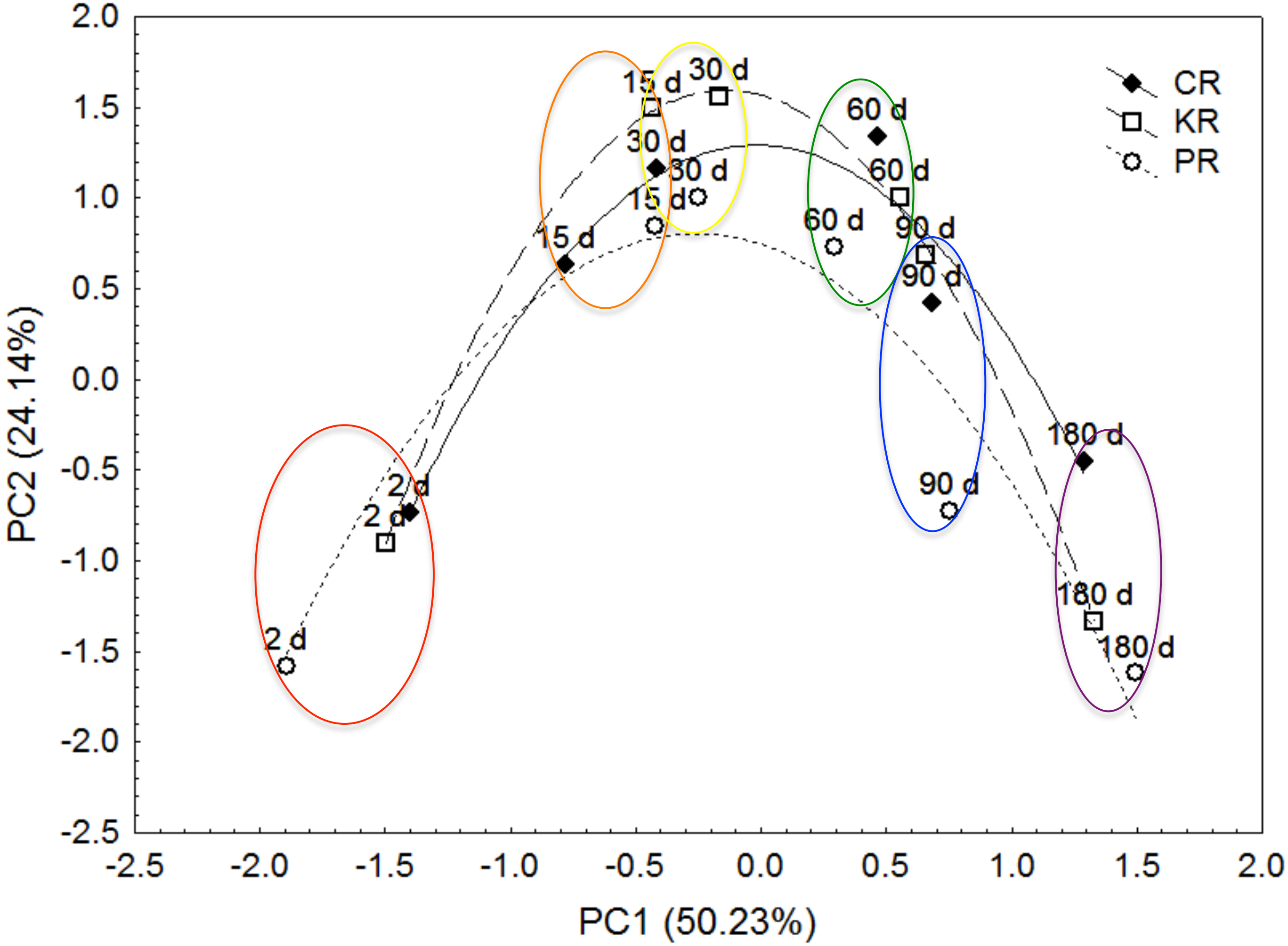
Variables were selected by backward stewise analysis.

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

Loadings plot



Scores plot



2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
PC2 (24.14%)
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
PC1 (50.23%)
◆ CR
□ KR
○ PR
15 d
30 d
30 d
30 d
15 d
15 d
60 d
60 d
60 d
90 d
90 d
90 d
2 d
2 d
2 d
180 d
180 d
180 d

Considerations on Exercise 1 (selected variables)

PCA permitted to reduce the data dimensionality and to separate samples along PC1 on the basis of ripening time.

The overall explained variance is high (74%).

PC2, and PCA in generis, did not pemit to separate the samples on the basis of the rennet used in the cheese-making process.

Ripening time determines the highest variance in data structure and its effect overwhelms that of the cheese making technology.

Final consideration

The selection of variables to contruct the final PCA has its pros and cons.

PCA could be used to describe the data set but could not always permit to discriminate sample among them because it is not a classification technique.

This because samples are distributed in space on the basis of the maximum explained variance criterion.

Homeworks

PCA is retained to be sensitive to the relative scaling of the original variables (this is very important when variables coming from different analysis are considered).

Let's repeat PCA using variables normalized on variance at home and discuss the results in class.