

Prinsipalkomponentanalyse (PCA)

Ingrid Måge
CAMO Software

The Unscrambler®



CAMO

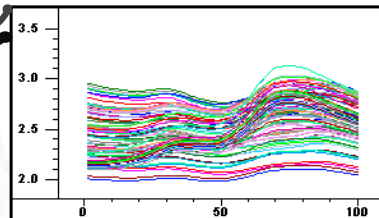
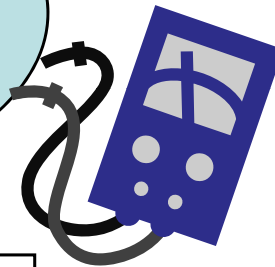
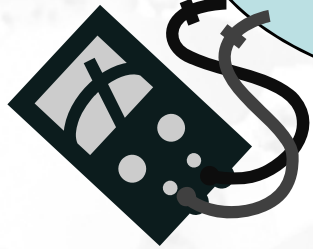
Delivering Tomorrow's Innovations, Today

Oversikt

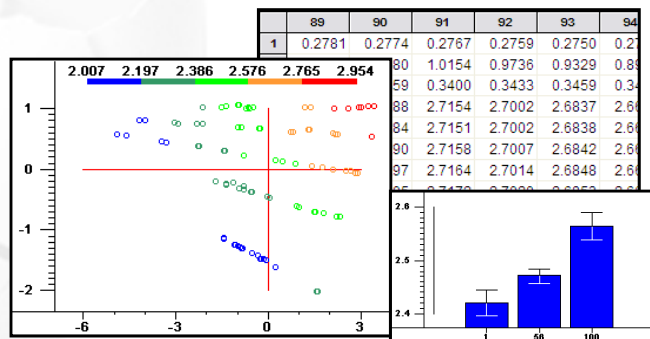
- Hvorfor multivariate metoder?
- Kovarians – hva er det?
- PCA – projeksjonsprinsippet
- Eksempel
 - Bringebærsyltetøy
- Utliggerdeteksjon med PCA
- Eksempel
 - Oljer

Why Multivariate data analysis ?

All real processes are multivariate until otherwise proved



Multivariate processes need multivariate sensors



Multivariate sensors need multivariate mathematics (chemometrics)

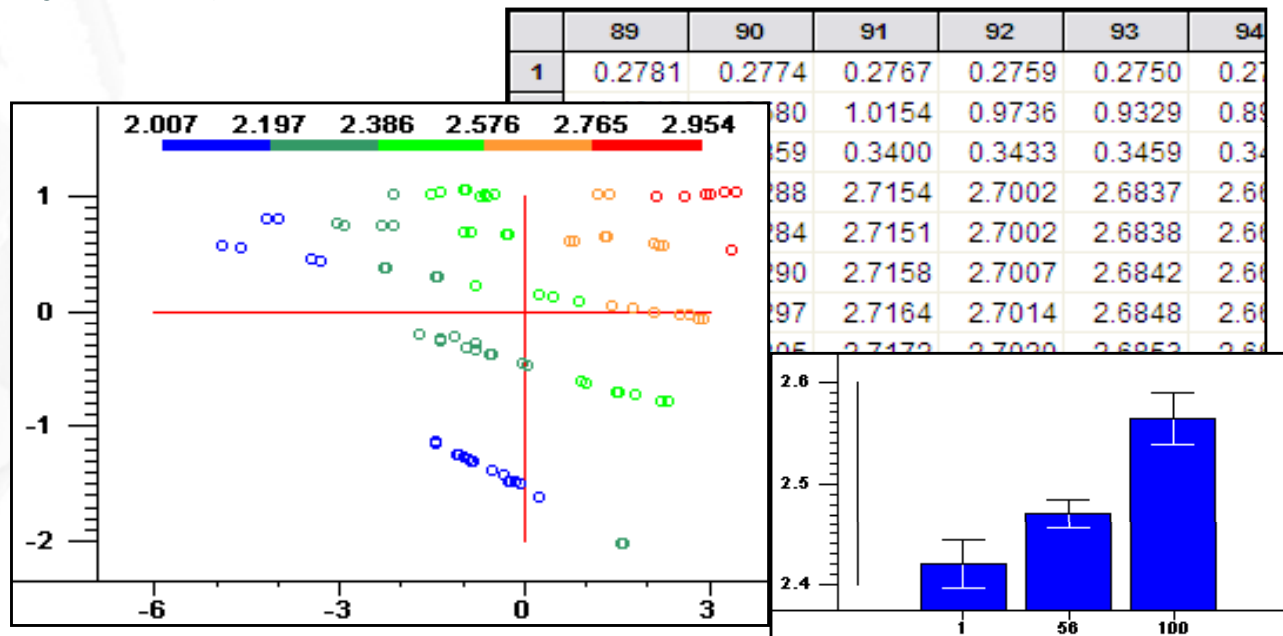
The Unscrambler®



Why multivariate data analysis?

Multivariate data analysis / chemometrics offer:

- Overview of complex problems
- Hypothesis generating tools
- Development of rapid instrumental methods
- General applicability (medicine, food, environment, economy etc.)



What is Chemometrics?

- Chemometrics:
 - application and development of mathematical and statistical methods to extract information from chemical data
- Exploratory chemometric data analysis:
 - Seeking the latent variables in data
 - Mapping correlations between variables
 - Evaluating differences between samples
 - Identifying outliers

Chemometrics in Practice

- 40% application knowledge
- 30% common sense
- 20% statistics
- 10% mathematics

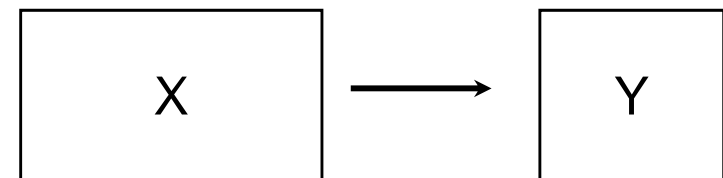
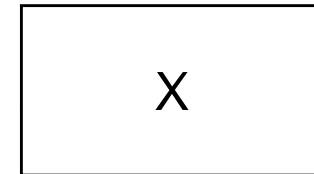


**Software
(The Unscrambler)**

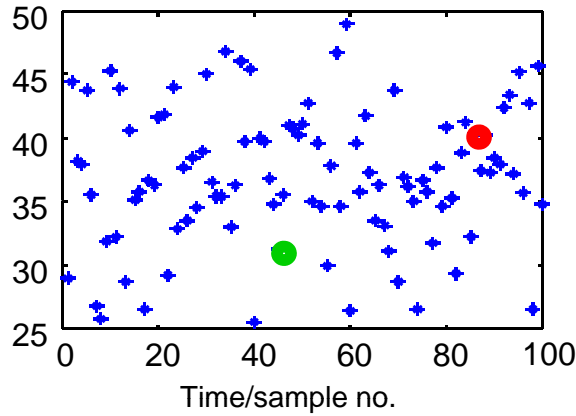
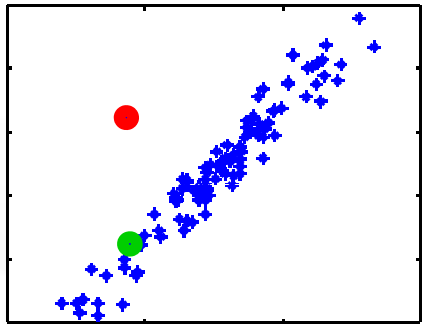
The Unscrambler®

The basic "work horses"

- ◆ Experimental design
 - ◆ Efficient experiments
- ◆ Exploratory data analysis (PCA)
 - ◆ Relations among 1 block of variables
 - ◆ Map of samples
 - ◆ Groups, patterns, outliers
- ◆ Regression methods (MLR, PCR, PLS)
 - ◆ Relations between two blocks of variables
 - ◆ Map of samples (PCR, PLS)
 - ◆ Model: $Y=XB$
 - ◆ Predict new samples



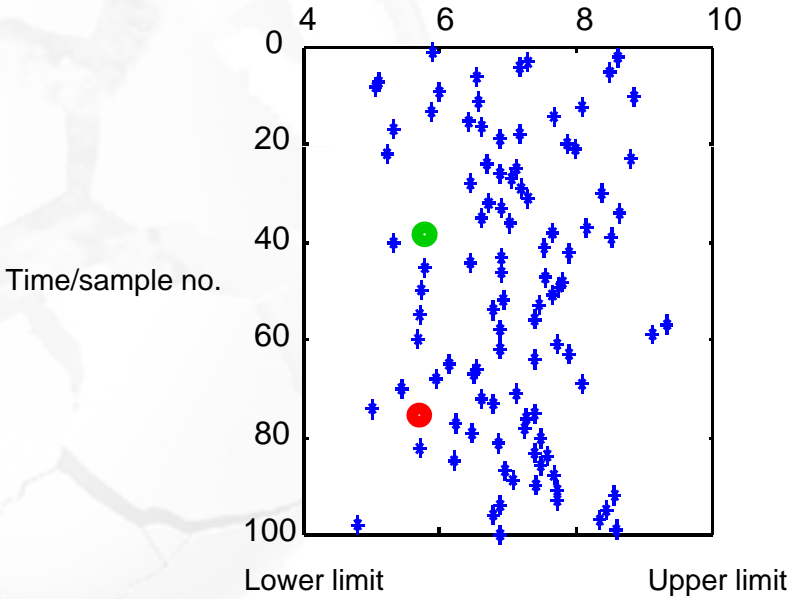
Covariance – a central point



Upper limit

Temperature

Lower limit



pH

Lower limit

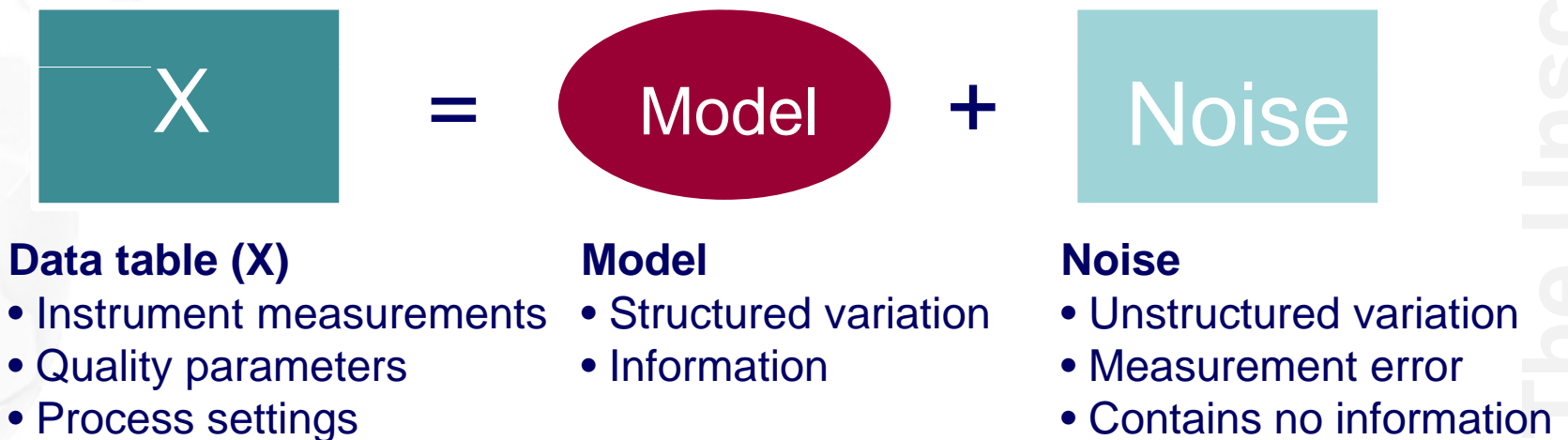
Upper limit

The Unscrambler®



Principal Component Analysis (PCA)

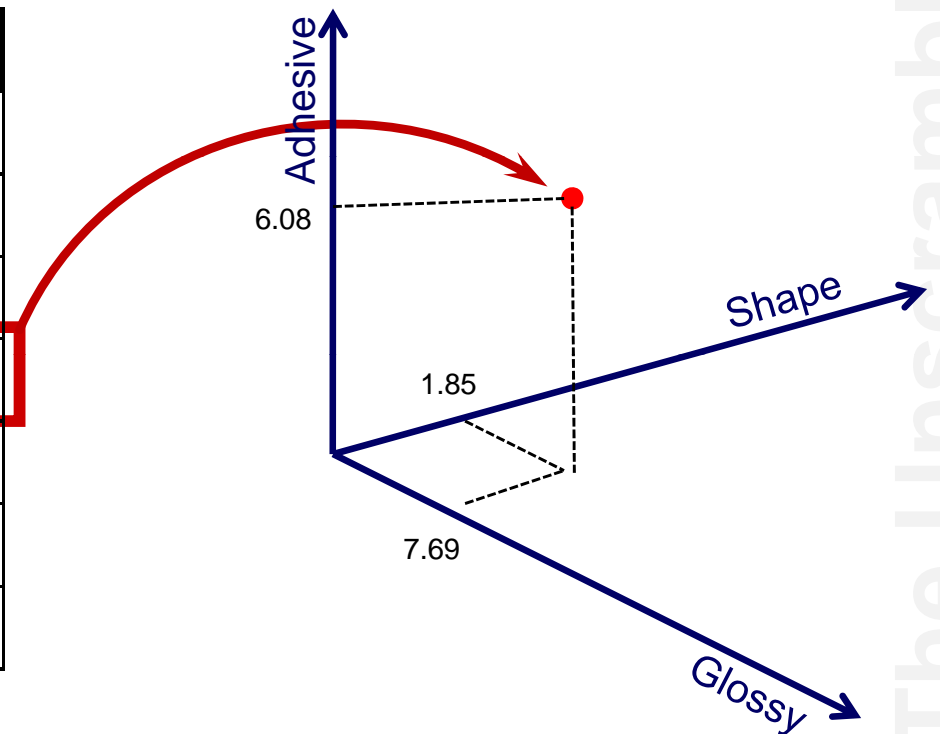
- Projection method
- Exploratory data analysis
- Extract information and remove noise
- Reduce dimensionality / Compression
- Classification and clustering



PCA (1/5)

Each row of the data table is a point in a multidimensional space

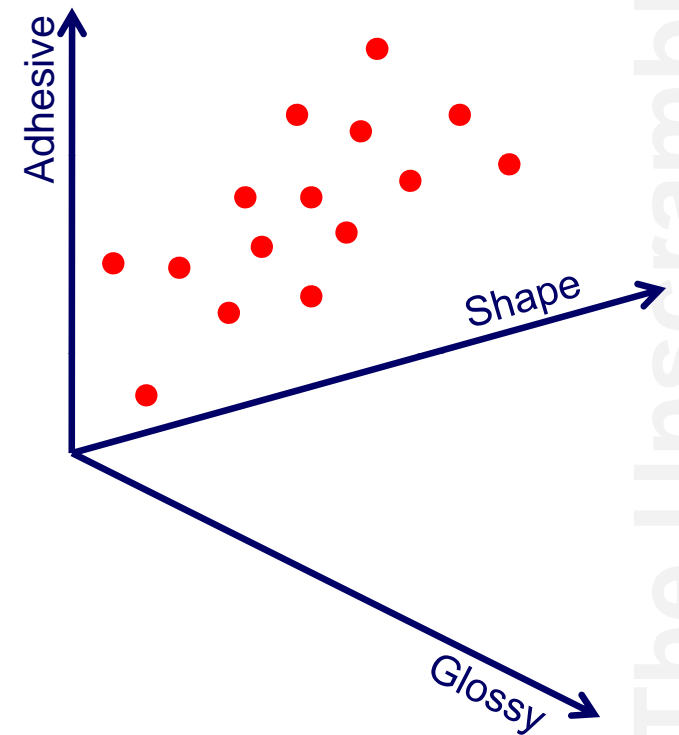
	Glossy	Shape	Adhesive
Product 1	7.08	7.08	4.15
Product 2	7.07	5.21	4.43
Product 3	7.69	1.38	5.00
Product 4	7.69	1.85	6.08
Product 5	6.47	6.87	5.27
Product 6	6.07	7.79	4.79
Product 7	6.08	8.08	5.15



PCA (2/5)

The whole data table becomes a swarm of points

- One point per sample
- Similar samples are close to each other
- Dissimilar samples are distant



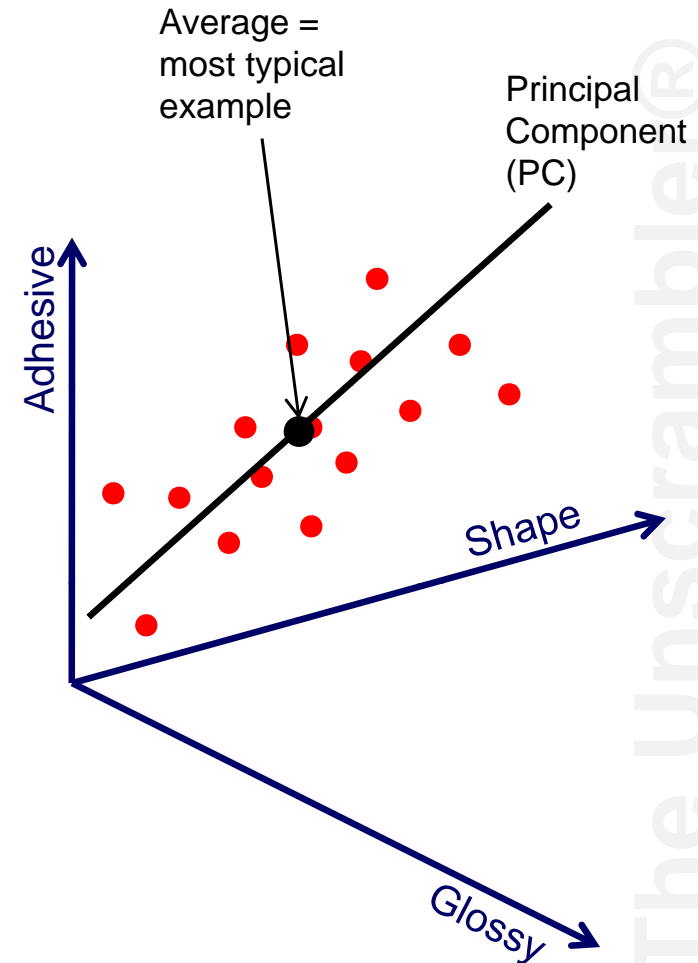
The Unscrambler®



PCA (3/5)

Variation among samples can be summarised by a straight line

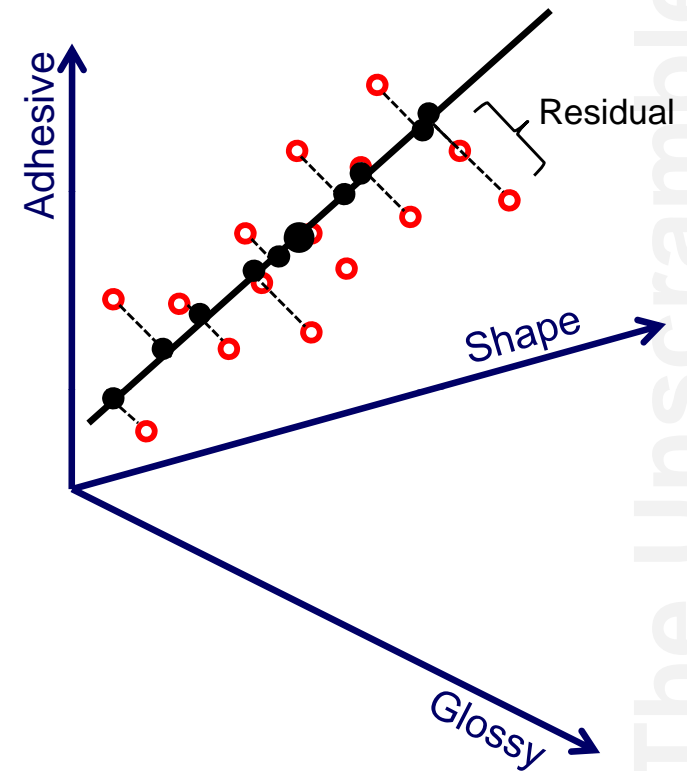
- The principal component (PC) goes through the middle of the swarm of points
- It is oriented in the direction of maximum variation
- Principal components are also called *latent variables*, or simply *components*



PCA (4/5)

Each sample is approximated by its projection onto the straight line

- Typical samples fall close to the average
- Extreme samples fall at the end of the line
- The residuals show how good the approximation is



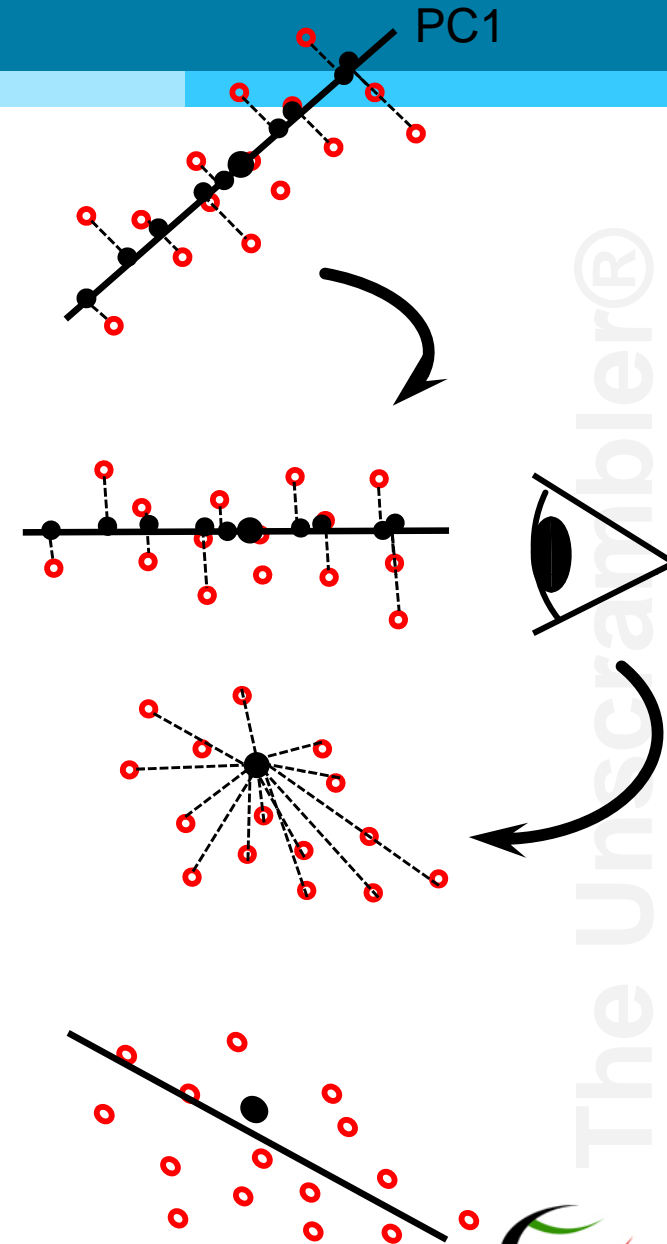
The Unscrambler®



PCA (5/5)

What about the differences that have not been summarized by the principal component?

- If we look at the samples from one end of PC1, we can see all the residuals.
- Now we can find a second straight line which describes sample differences from that new point of view.



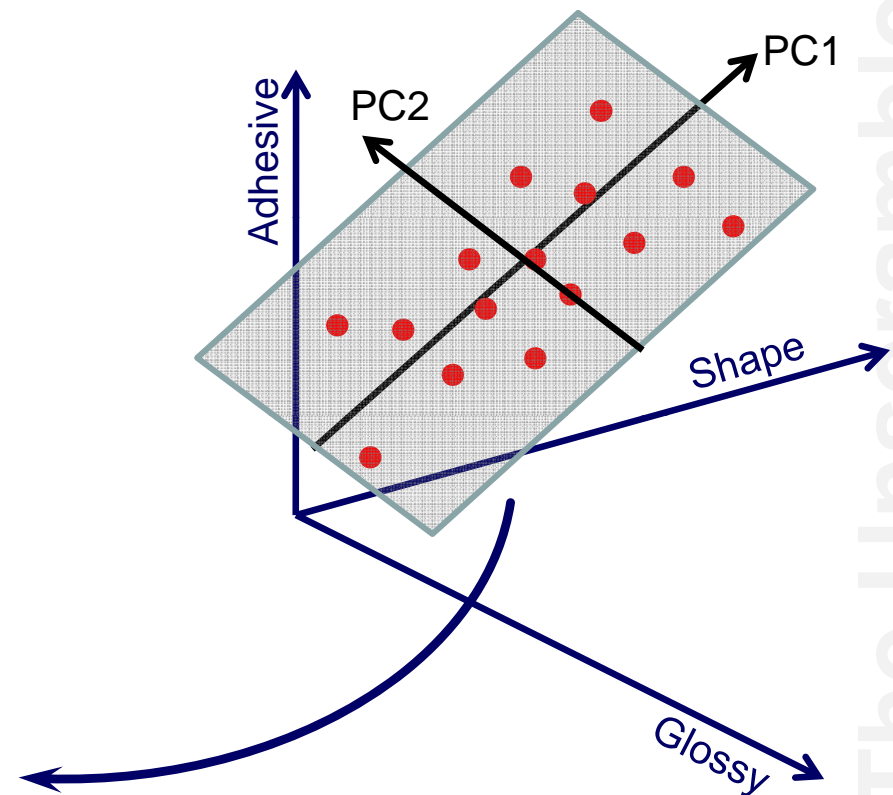
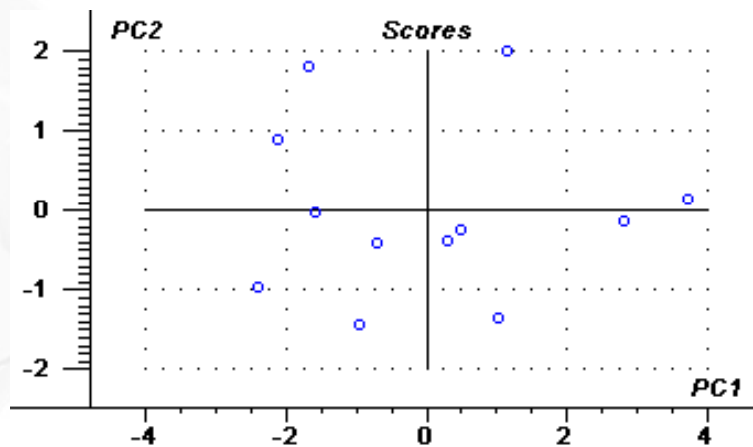
The Unscrambler®



Building a 2-Dimensional Map

With two principal components we can define a plane

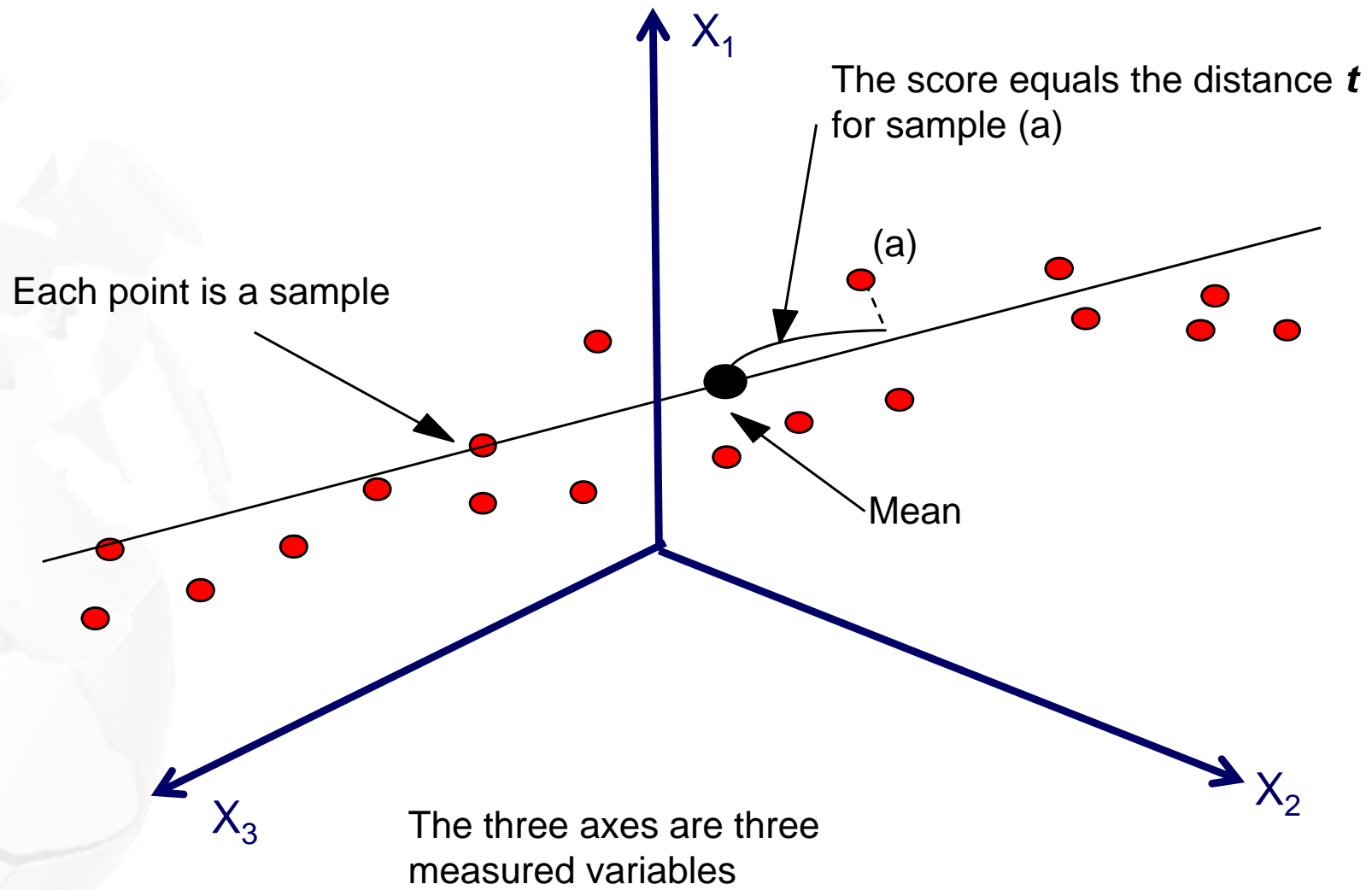
- Let us freeze the picture
- and use it as a map



The Unscrambler®

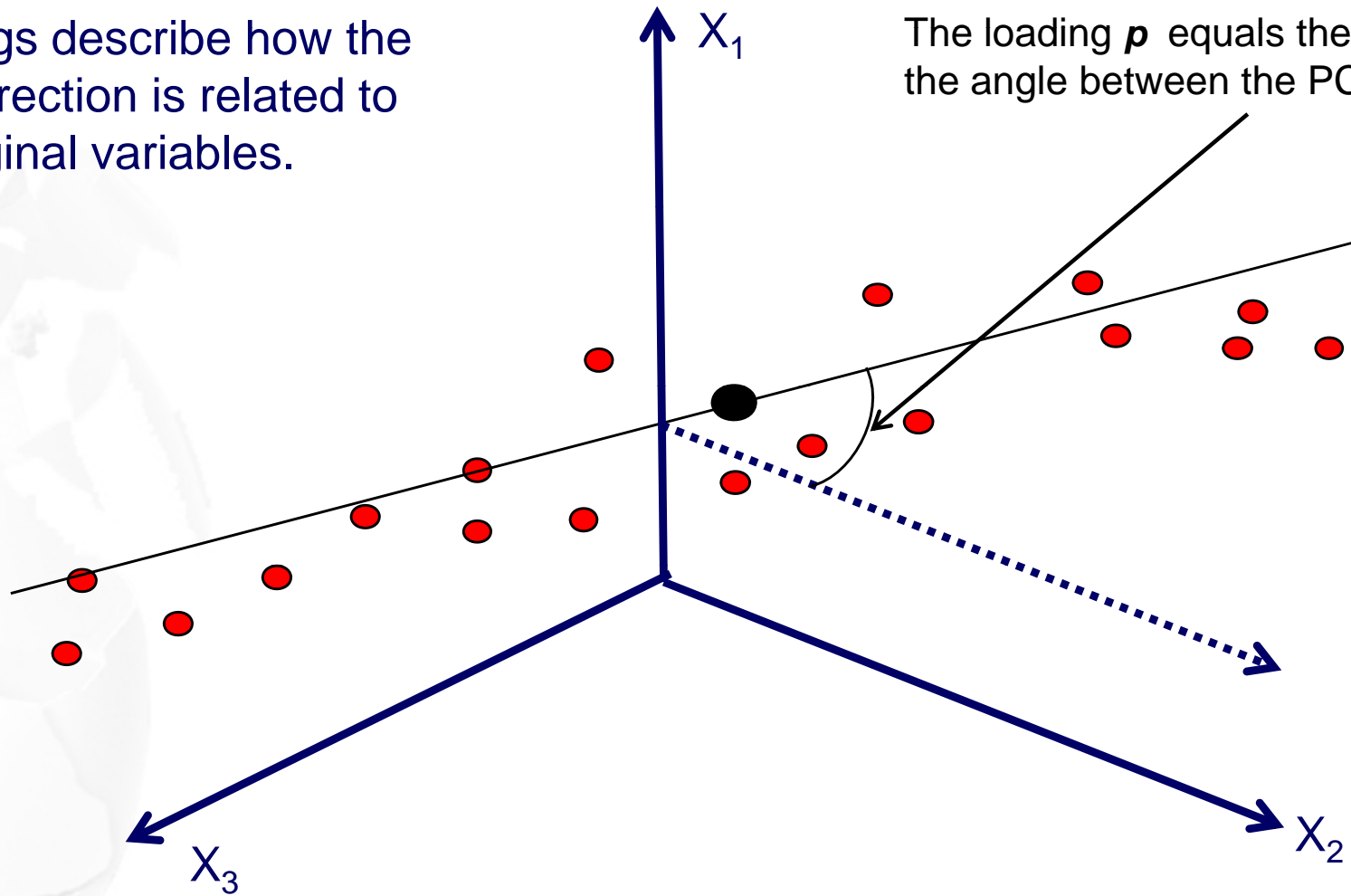


What is a Score?



What is a Loading?

Loadings describe how the PC's direction is related to the original variables.



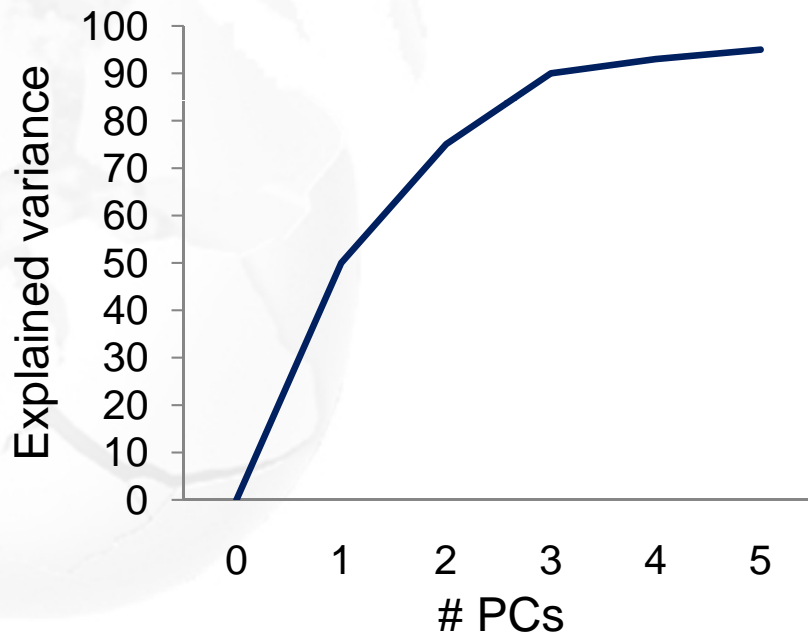
The Unscrambler®



Number of components in model

Keep adding components as long as they contain structured information

- Interpretable differences between samples
- Explains a significant amount of variation (validation!)



$$X = \text{1PC Model 50\%} + \text{Residual 50\%}$$

$$X = \text{2PC's Model 75\%} + \text{Residual 25\%}$$

$$X = \text{3PC's Model 90\%} + \text{Residual 10\%}$$

PCA Algebra

$$X = \begin{matrix} | & \text{---} p_1 \\ & \\ t_1 & \\ | & \text{---} p_2 \\ & \\ t_2 & \end{matrix} + E$$

$$X = \begin{matrix} | \\ T \\ | \end{matrix} \begin{matrix} \text{---} P \\ | \\ \text{---} \end{matrix} + E$$

$$X = TP^T + E$$

Example: Raspberry jams

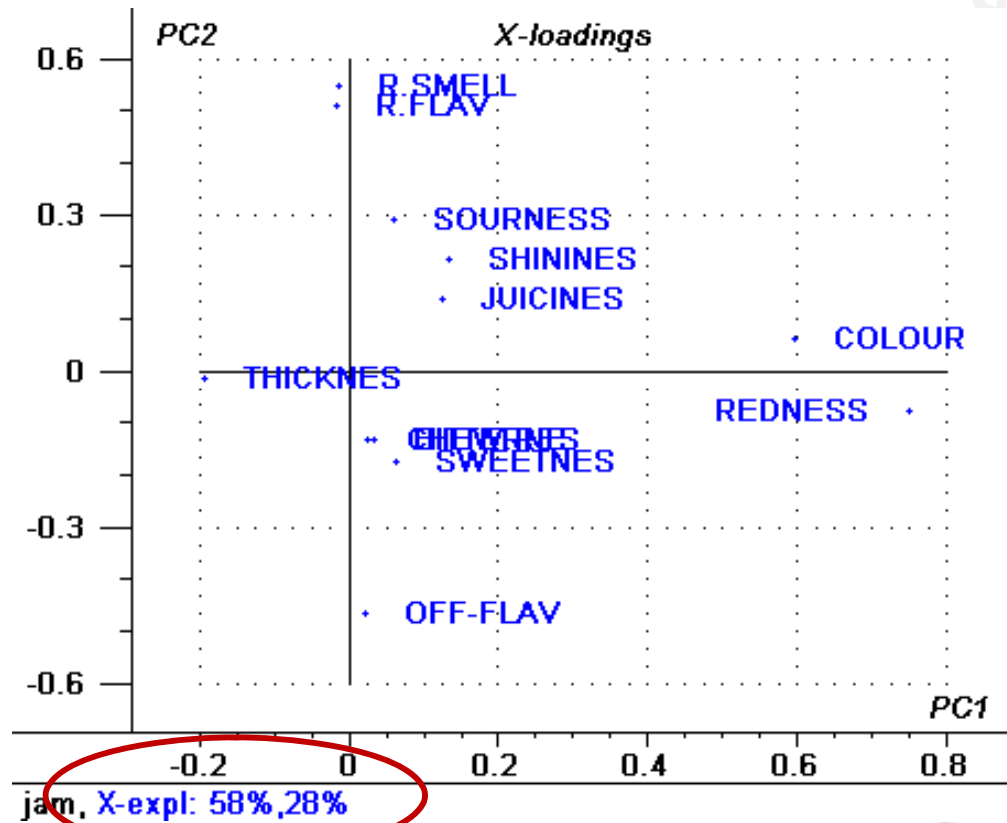
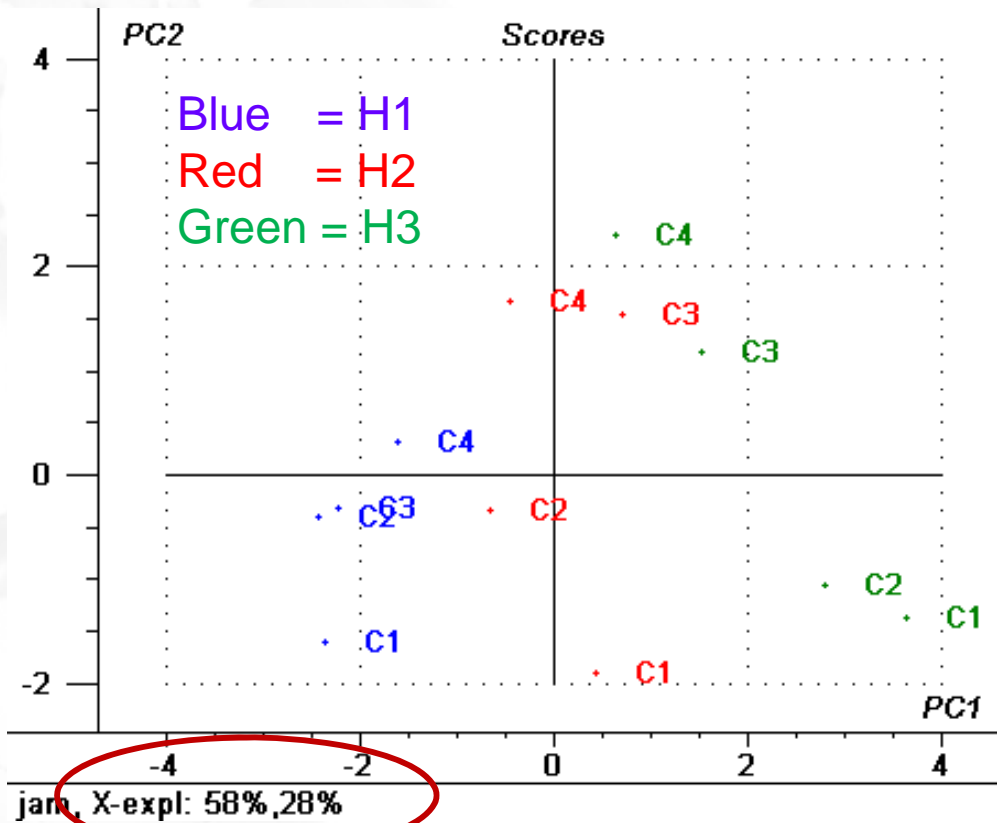
- 12 raspberry jams are made according to a factorial design
 - 4 cultivares
 - 3 harvesting times
- 12 sensory attributes have been evaluated for each jam
 - Sweetness, sourness, bitterness, redness, juiciness,...

Cultivar	Harvesting time
C1	H1
C1	H2
C1	H3
C2	H1
C2	H2
C2	H3
C3	H1
C3	H2
C3	H3
C4	H1
C4	H2
C4	H3



Example: Raspberry jams

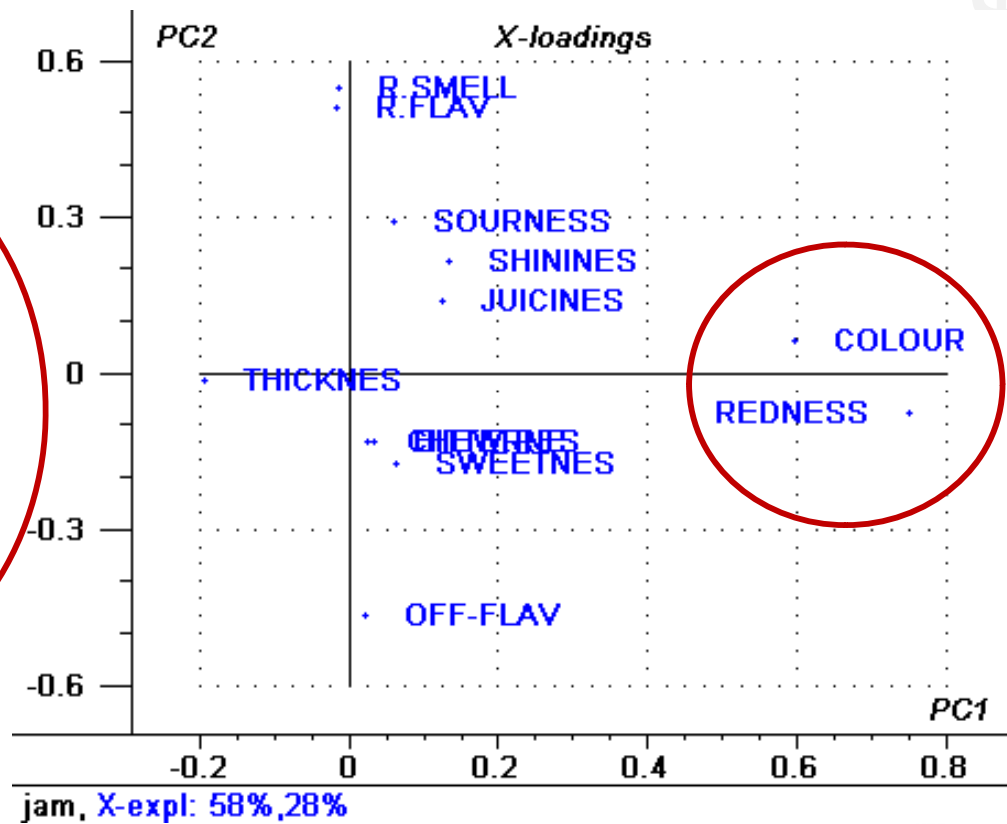
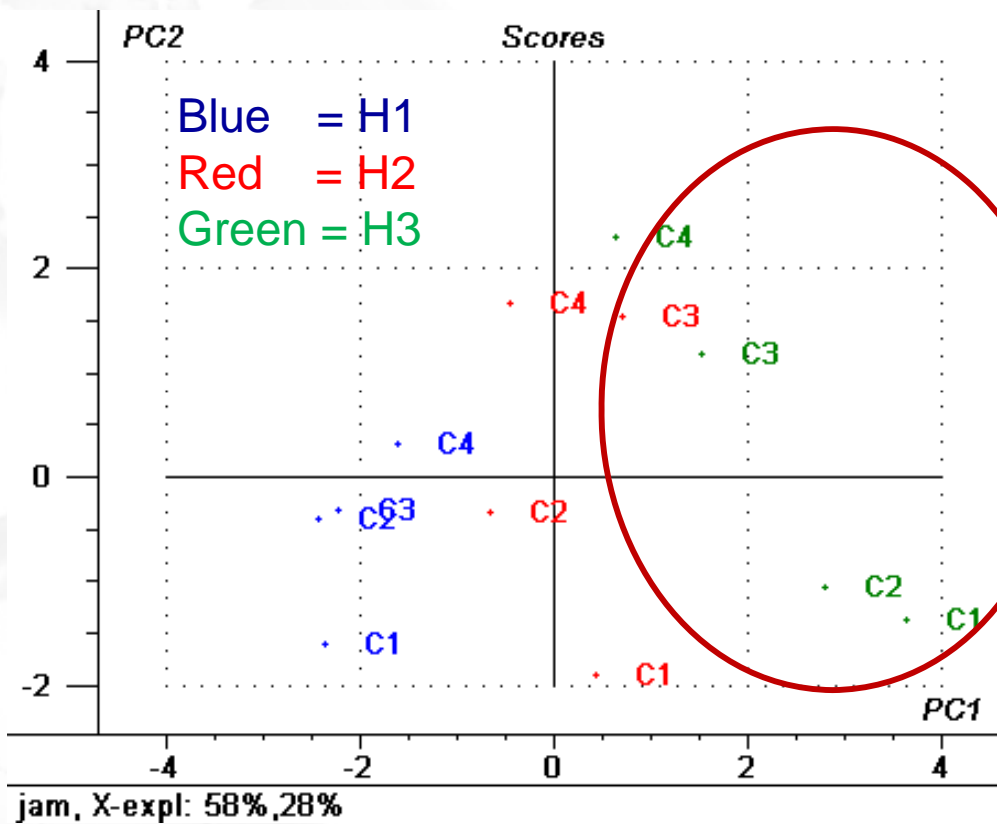
Two components explain 58% + 28% = 86% of the variation in X



er®

Example: Raspberry jams

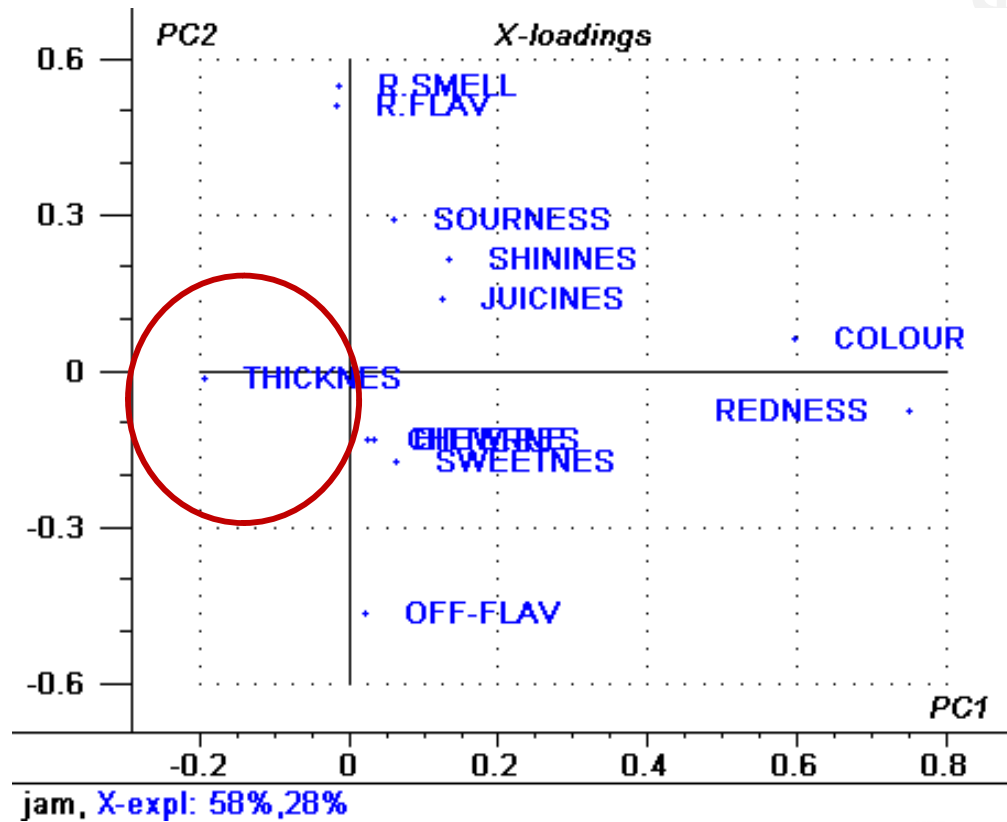
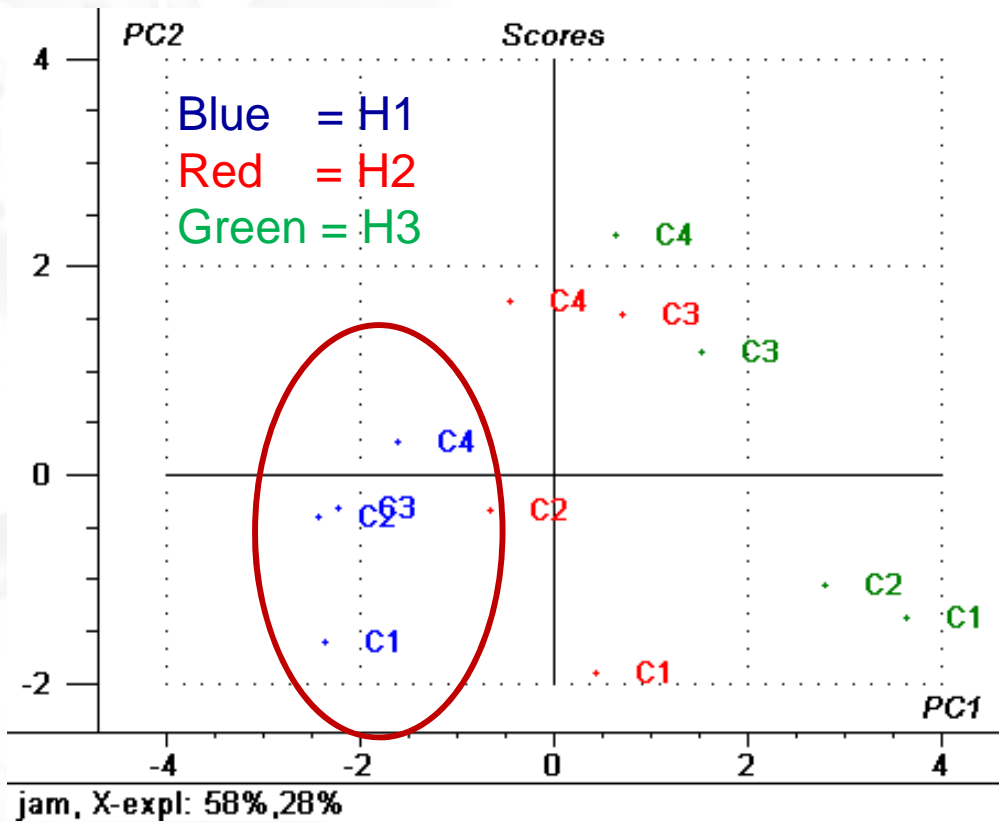
Late harvest gives a jam with high redness and colour intensity



er®

Example: Raspberry jams

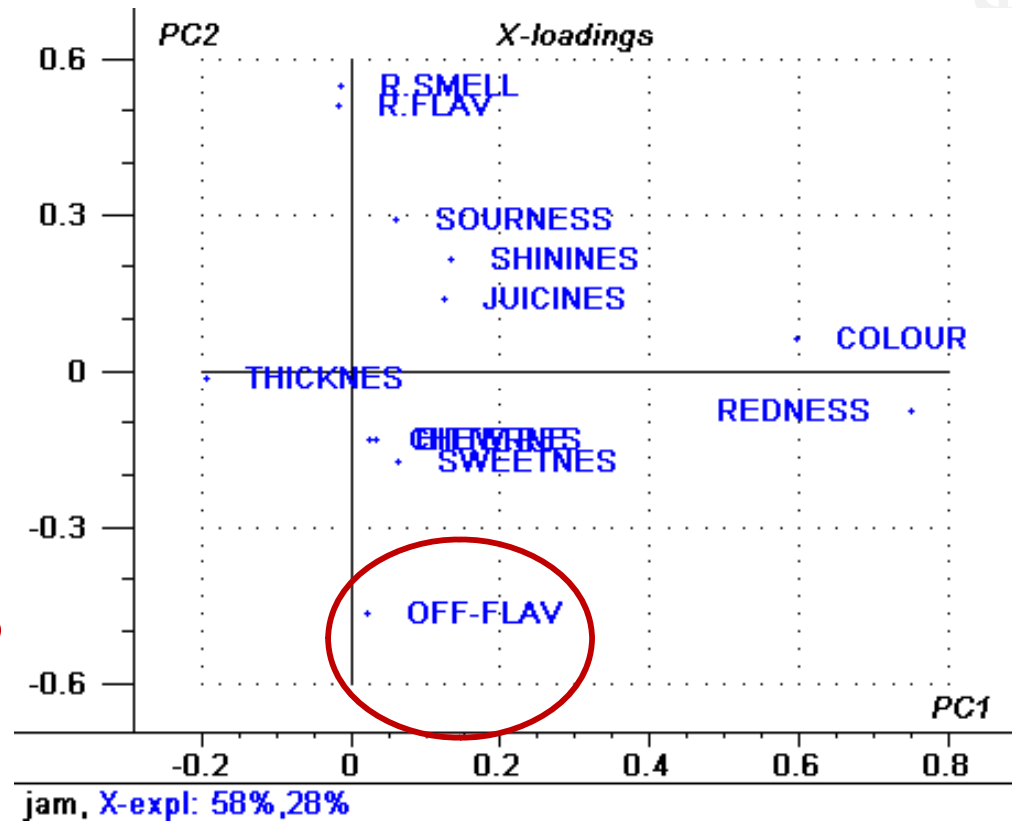
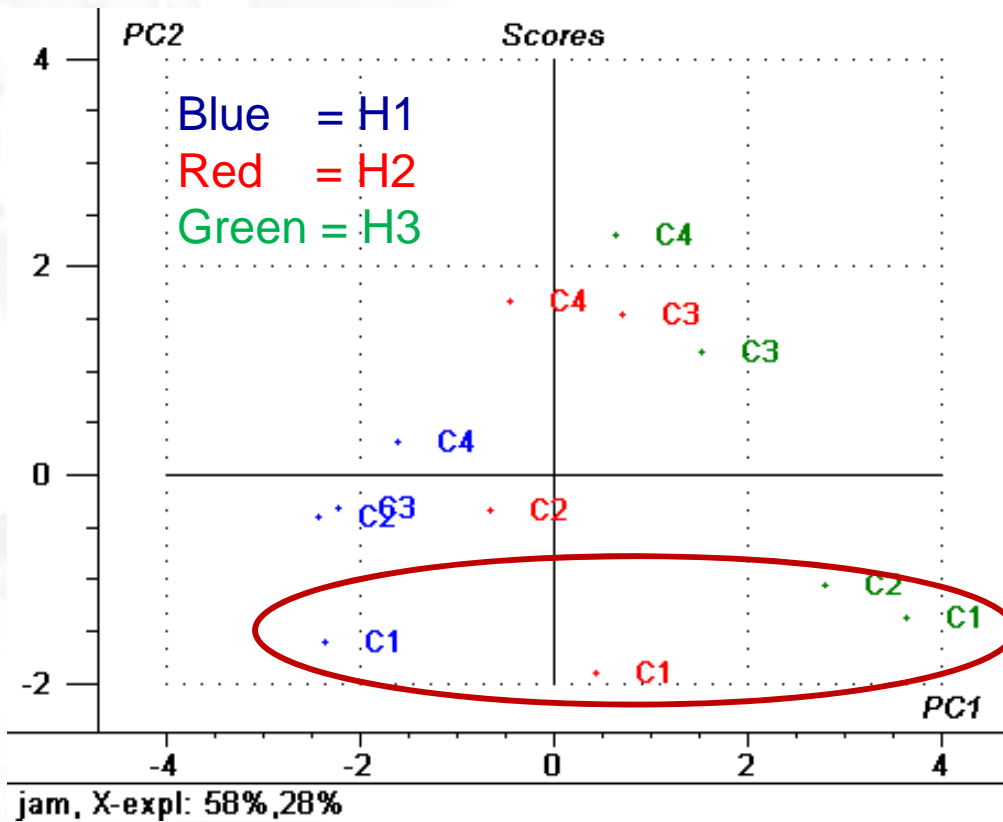
Early harvest gives a thick jam



er®

Example: Raspberry jams

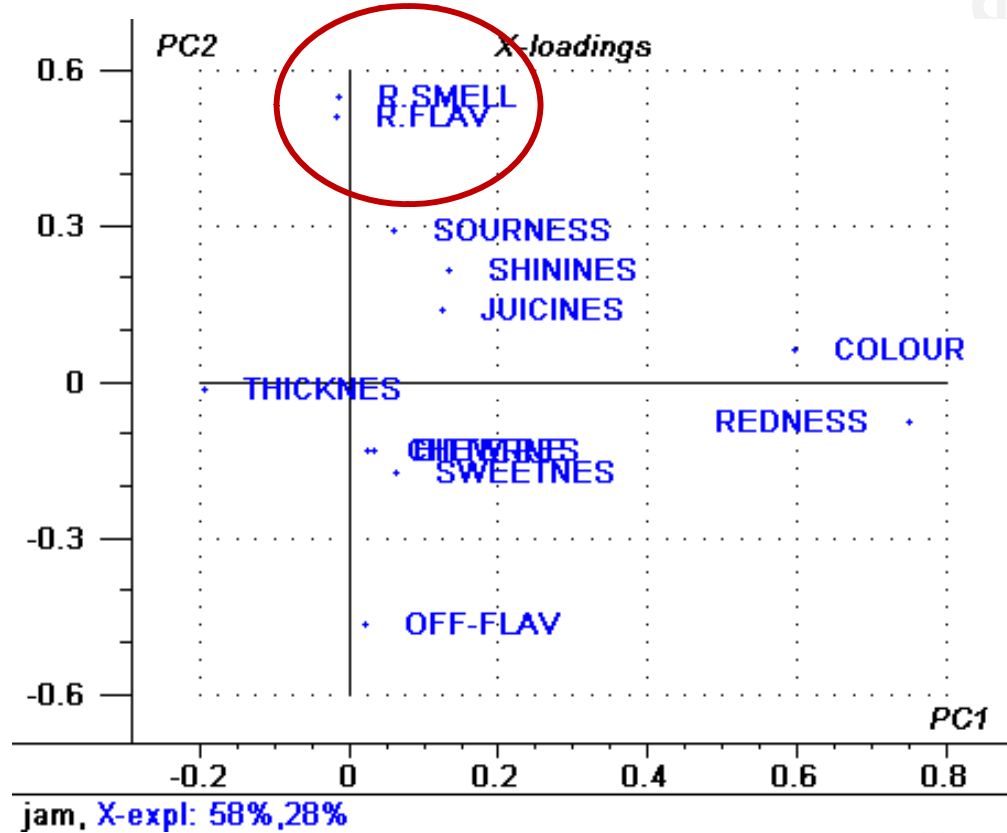
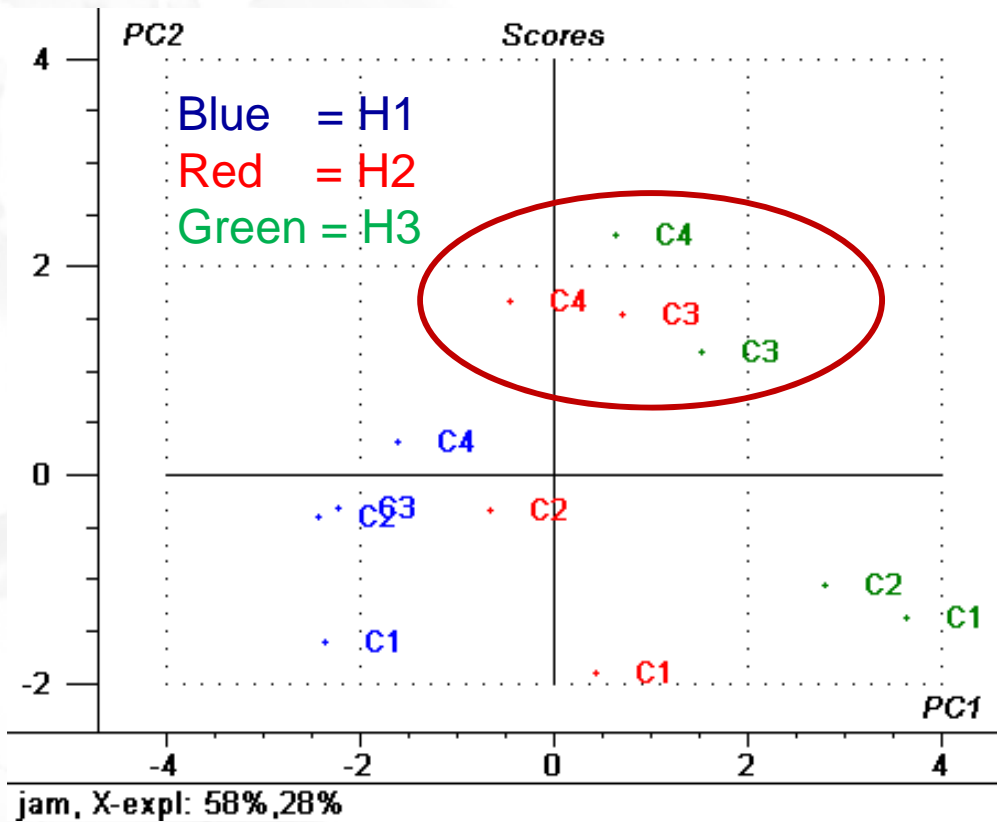
Cultivar C1 has a higher degree of off-flavour



er®

Example: Raspberry jams

Cultivar C3 and C4 has a distinct raspberry smell and flavour, if it is not harvested too early



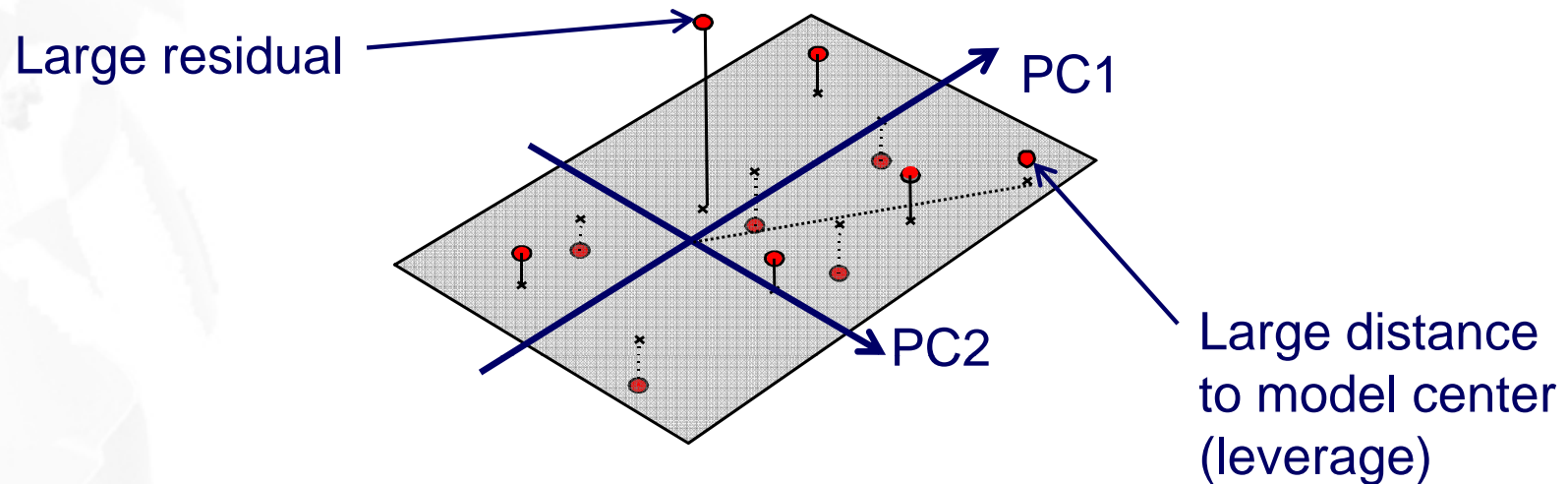
er®

Outliers

- What is an outlier?
 - An object deviating from the others
 - Can disturb the model
- Cause
 - Measurement error
 - Wrong labeling
 - Deviating product / sample
 - Noise

Outliers

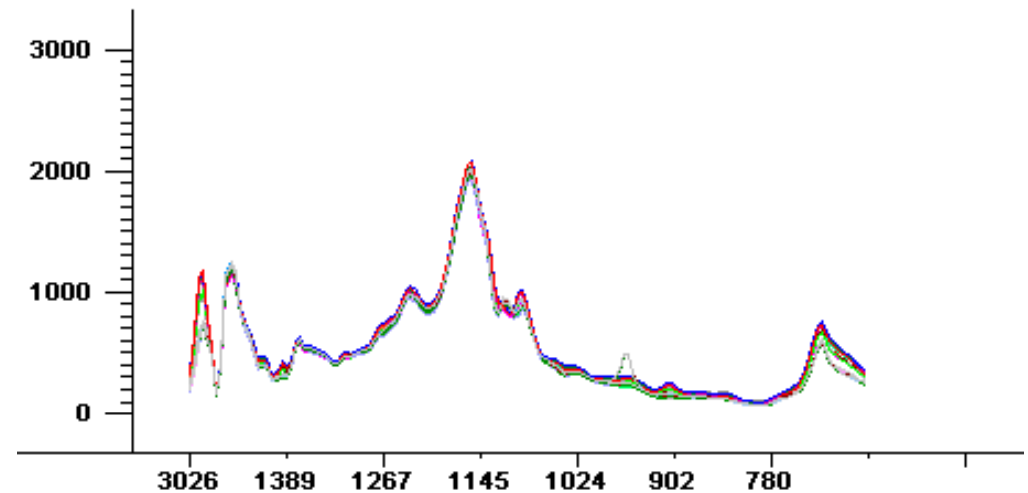
Example: sample projections onto PC1 and PC2:



- Samples with a large residual are not well described
- Samples with high leverage are influential

Example: Oils

- Spectra have been measured on 37 oil samples
 - Olive
 - Corn
 - Safflower
 - Sesame
 - Corn marg



- Can PCA be used to classify the oils?

Conclusion: Oils

- PCA was able to discriminate between the five oil types
 - Five clearly separated groups were found using PC1 and PC3
- PCA also discovered some suspicious/outlying samples
 - Wrong labelling
 - Only one sesame sample
 - One safflower sample didn't fit in (high residual)