UMETRICS
AN MKS COMPANY

Multivariate Data Analysis for Omics

September 2-3 2008

Susanne Wiklund

IID 1062

Multivariate Data Analysis and Modelling in "Omics"

Outline

Day 1

- Chapter 1
  - Introduction multivariate data analysis
  - Introduction to "omics"
  - Introduction to Principal component analysis
- Chapter 2
  - Overview of data tables
  - How PCA works
  - PCA example
  - PCA diagnostics
- Chapter 3
  - PCA for finding patterns, trends and outliers
  - PCA example
- Chapter 4
  - Data processing
  - Scaling
  - Normalisation

Day 2

- Chapter 5
  - Introduction to Orthogonal partial least squares (OPLS)
  - From PCA to OPLS-DA
  - Classification
  - Biomarker identification
  - Multiple treatments
- Chapter 6
  - Validation

Exercises

- Foods: PCA
- Rats Metabonomics 1: Metabolomics, NMR data, PCA
- Health: clinical data, PCA using paired samples
- MSMouse: Metabolomics, LC/MS data, PCA and OPLS-DA, task 2 not included, miss classification
- Genegrid I: Micro array, PCA + OPLS-DA
- Ovarian cancer: Proteomics, MS data, OPLS-DA, S-plot
- PCA vs. OPLS-DA: Metabolomics, NMR data, PCA and OPLS-DA
- GC/MS metabolomics: Resolved and integrated GC/MS data, OPLS-DA, S-plot and SUS-plot
- Rats Metabonomics 2: Metabolomics, NMR data, OPLS-DA, S-plot, SUS-plot
- Identification of bias effects in Transcriptomics data: micro array data, PCA, OPLS-DA
- Proteomics anti diabetics: Proteomics, MS data

- Underscore means that all participants should do these exercises.

Multivariate Analysis
for ”omics” data

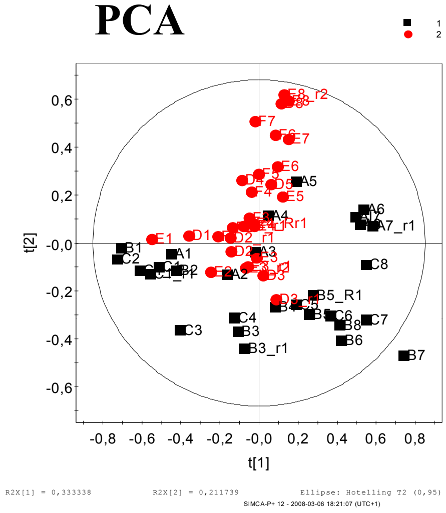
Chapter 1

Introduction

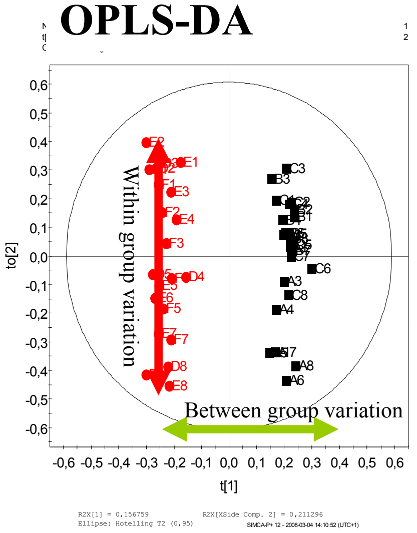
UMETRICS

General cases that will be discussed during this course

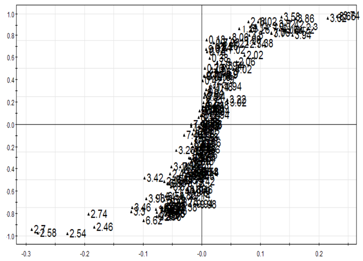
PCA



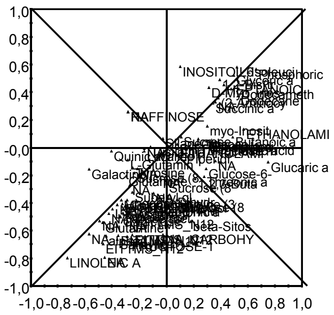
OPLS-DA



S-plot



SUS-plot



PCA
- **Patterns**
- **Trends**
- **Outlier detection**

OPLS-DA
- **Classification**
- **Potential biomarkers**
- **Multiple treatments**

UMETRICS

Outline

- Need for Multivariate Analysis
  - Example
- Measurements
  - Univariate, Bivariate, Multivariate
- Why Multivariate methods
- Introduction to Multivariate methods
  - Data tables and Notation
  - What is a projection?
  - Concept of Latent Variable
  - “Omics”
- Introduction to principal component analysis

Background

- Needs for multivariate data analysis

- Most data sets today are multivariate
  - due to

  (a) availability of instrumentation

  (b) complexity of systems and processes

- Continuing uni- and bivariate analysis is
  - often misleading ex: will be described
  - often inefficient ex: t-test on 245 variables

Multivariate Data Analysis

- Extracting information from data with multiple variables by using all the variables simultaneously.

- It's all about:
  - How to get information out of existing multivariate data

- It's much less about:
  - How to structure the problem
  - Which variables to measure
  - Which observations to measure (DoE)

Introduction to "omics"

- "omics" in the literature
  - Metabolomics
  - Metabonomics
  - Transcriptomics
  - Genomics
  - Proteomics
  - Bionomics
  - Toxicogenomics
  - And many more

- The "omics" data in this course includes
  - Metabolomics
  - Proteomics
  - Transcriptomics
- What do they have in common?
  - Last 5 letters
  - Few samples
  - Many variables
  - Measurement of all detectable species represented i.e. very complex data
  - Classification and diagnostics
  - Biomarkers
  - Explore biology

Introduction to “omics”

<u>Metabolomics</u>

“*comprehensive analysis of the whole metabolome under a given set of conditions*”[1]

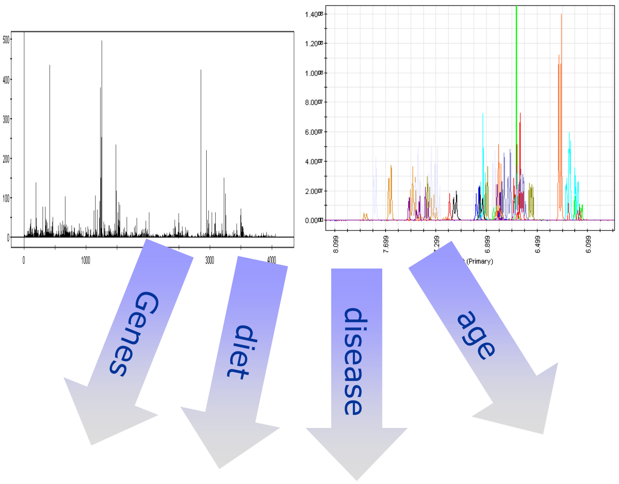
<u>Metabonomics</u>

”*the quantitative measurement of the dynamic multiparametric metabolic response of living systems to pathophysiological stimuli or genetic modification*” [2]

1. Fiehn, O., et.al Metabolite profiling for plant functional genomics. *Nature Biotechnology.* 2000;18:1157-1161.
2. Nicholson, J. K., et.al 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica.* 1999;29:1181-1189.

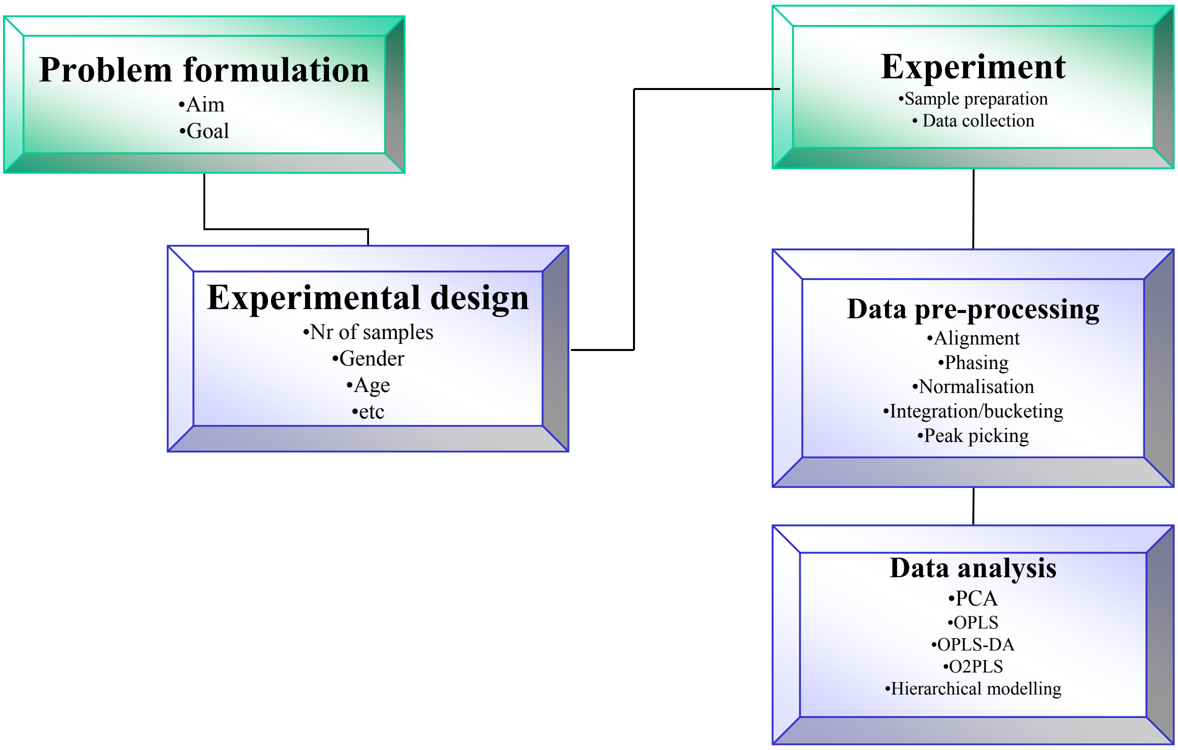
Objectives in “Omics”

- Study organisms as integrated systems
  - Genes
  - Proteins
  - metabolic pathways
  - cellular events
- Extracts and distil information on
  - Genes
  - Disease
  - Physiological state
  - Diet
  - Biological age
  - Nutrition



- Create new diagnostic tools
- One major goal is to extract biomarkers and understand the interplay between molecular and cellular components

“Omics” workflow



Problem formulation
- Aim
- Goal

Experimental design
- Nr of samples
- Gender
- Age
- etc

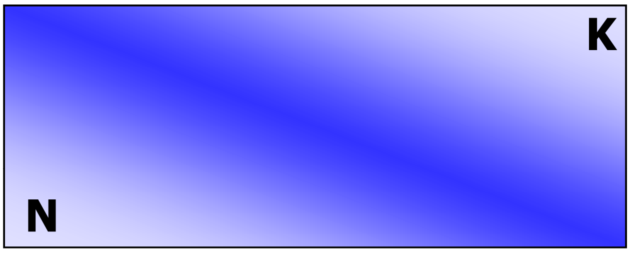
Experiment
- Sample preparation
- Data collection

Data pre-processing
- Alignment
- Phasing
- Normalisation
- Integration/bucketing
- Peak picking

Data analysis
- PCA
- OPLS
- OPLS-DA
- O2PLS
- Hierarchical modelling

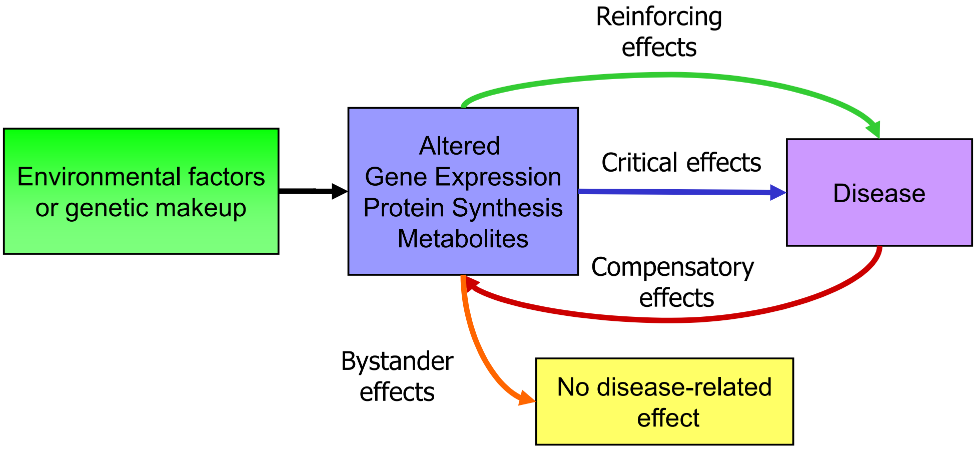
Today's Data

- GC/MS, LC/MS, NMR spectrum or genechip
  - c. 10,000 peaks for Human urine
- Problems
  - Many variables
  - Few observations
  - Noisy data
  - Missing data
  - Multiple responses
- Implications
  - High degree of correlation
  - Difficult to analyse with conventional methods
- Data ≠ Information
  - Need ways to extract information from the data
  - Need reliable, predictive information
  - Ignore random variation (noise)
- **Multivariate analysis** is the tool of choice



Causality vs Correlation

- Perturbation of a biological system causes myriad changes, only some will be directly related to the cause
  - Typically we find a population of changes with statistical methods
  - May be irrelevant or even counter-directional
  - Further biological evidence always required



Reinforcing effects

Environmental factors or genetic makeup

Altered Gene Expression Protein Synthesis Metabolites

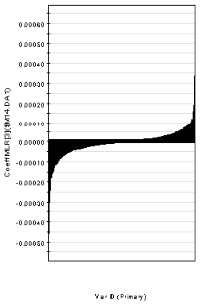
Critical effects

Disease

Compensatory effects

Bystander effects

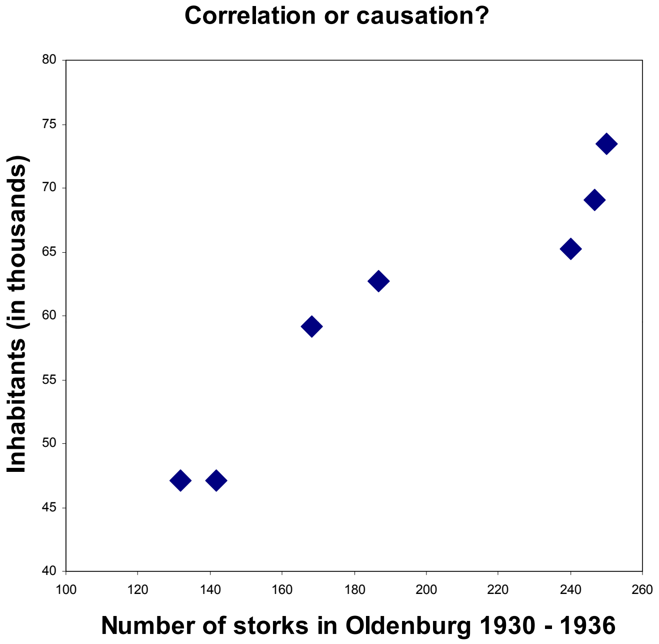
No disease-related effect



Correlation and Causality

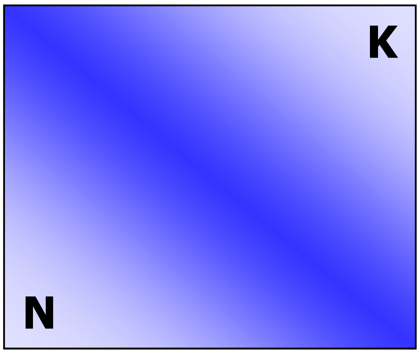
Although the two variables are correlated, this does not imply that one causes the other!

Real but non-causal, or spurious ?

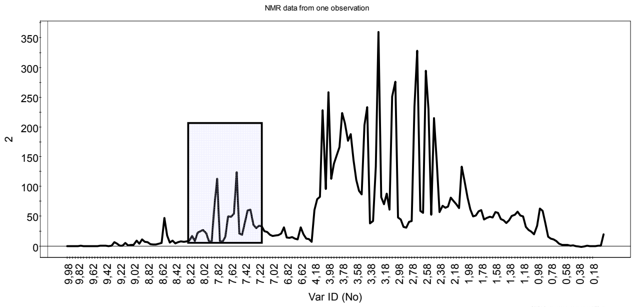


Data with many Variables

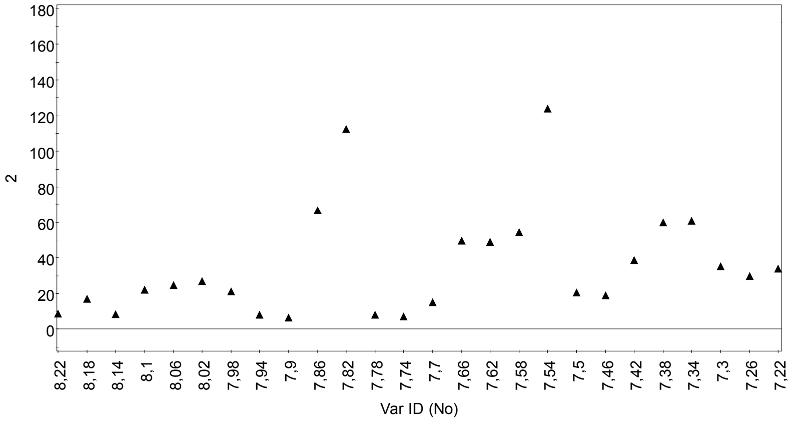
- Multivariate
  - More than 6 variables
- N Observations
  - Humans, rats, plants
  - trials, time points
- K Variables
  - Spectra, peak tables
- **Most systems are characterised by 2-6 underlying processes yet we measure thousands of things**



Observations and spectroscopic variables



SIMCA-P+ 12 - 2008-07-08 13:26:13 (UTC+1)



SIMCA-P+ 12 - 2008-07-08 13:37:31 (UTC+1)

•Each sample spectrum is one observation

•Each data point in the spectrum will represent one variable

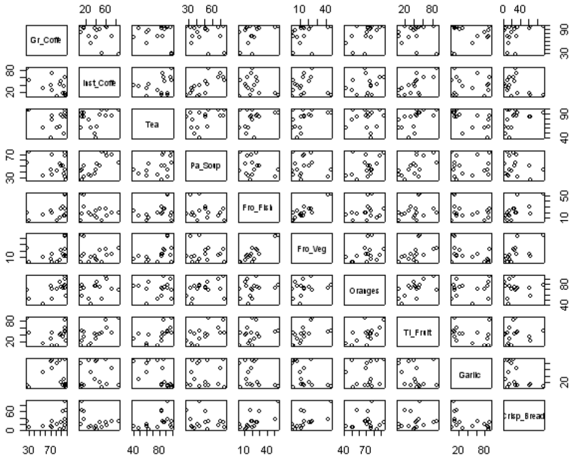
•Variables can also be resolved and integrated, in that case each integral will create a variable

Types of Data in “omics”

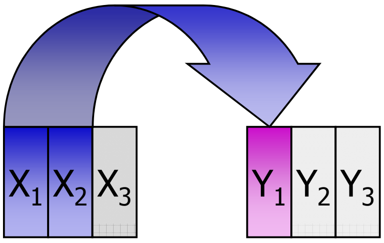
Field	Observations (N)	Variables (K)
Metabolomics	Biofluids, plant extracts, tissue samples	**Spectra from:** ^{1}H NMR, ^{1}C NMR ^{1}H-^{13}C NMR, GC/MS, LC/MS, UPLC/MS
Proteomics	Tissue Samples	2D Gels Electrophoresis/MS
Genomics/transcriptomics	Tissue Samples	Micro arrays, Fluorescence probes
Chromatography	Columns, Solvents, Additives, Mixtures	Physical Properties, Retention Times

Poor Methods of Data Analysis

- Plot pairs of variables
  - Tedious, impractical
  - Risk of spurious correlations
  - Risk of missing information



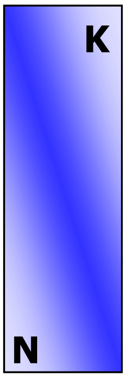
- Select a few variables and use MLR
  - Throwing away information
  - Assumes no ‘noise’ in X
  - One Y at a time



Development of Classical Statistics – 1930s

- Multiple regression
- Canonical correlation
- Linear discriminant analysis
- Analysis of variance

Tables are long and lean

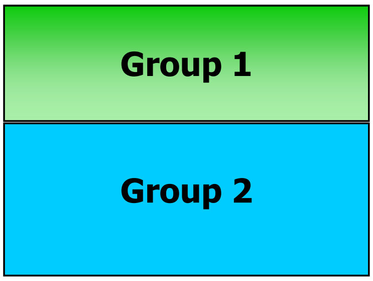


Assumptions:

- Independent X variables
- Precise X variables, error in Y only
- Many more observations than variables
- Regression analysis one Y at a time
- No missing data

Risks with Classical Methods

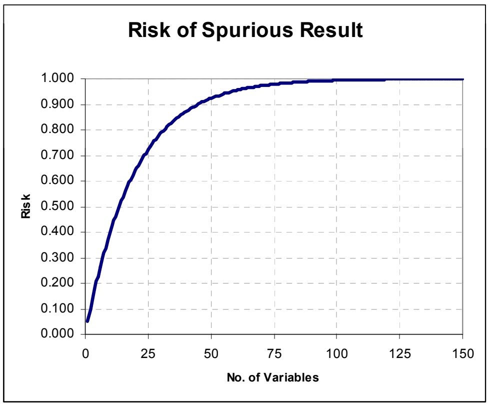
- Comparing two groups (t-test)
- Typically 5% significance level used
  - Type I errors: false positives, spurious results
  - Type II errors: false negatives, risk of not seeing the information
- Type I Risk = $1 - 0.95^K$



Group 1

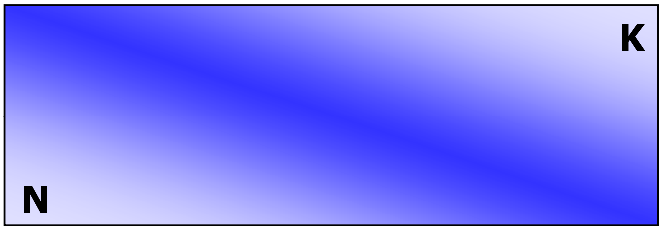
Group 2

K	5	10	60	100
Risk	23%	40%	95%	99.4%



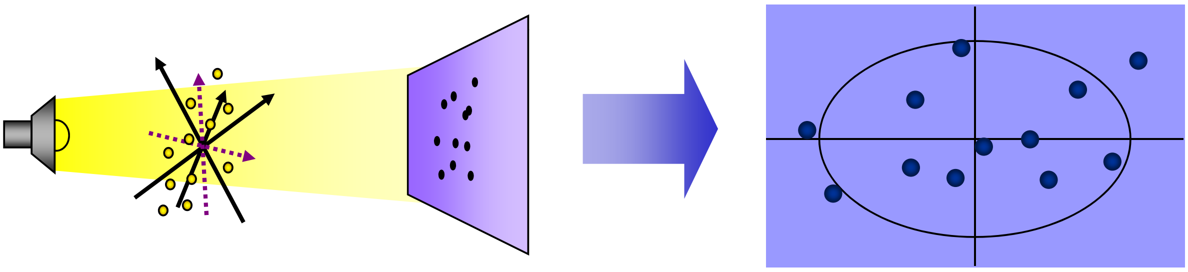
Research in 21st Century

- Experimental costs, ethics, regulations => few observations
- Instrumental & electronics revolution => many variables
- Chemometrics: short & wide data tables



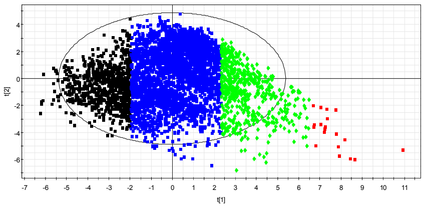
A Better Way

- Multivariate analysis by Projection
  - Looks at ALL the variables together
  - Avoids loss of information
  - Finds underlying trends = “latent variables”
  - More stable models



Why MVDA by Projections (PCA & OPLS) ?

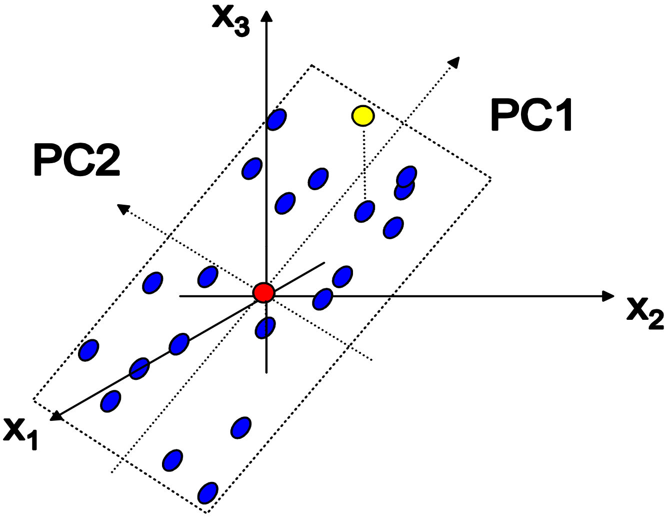
- Deals with the dimensionality problem
- Handles all types of data tables
  - Short and wide, N >> K
  - Almost square, N ≈ K
  - Long and lean, N << K
- Handles correlation
- Copes with missing data
- Robust to noise in both X and Y
- Separates regularities from noise
  - Models X and models Y
  - Models relation between X and Y
  - Expresses the noise
- Extracts information from all data simultaneously
  - Data are not the same as information
- Results are displayed graphically



What is a Projection?

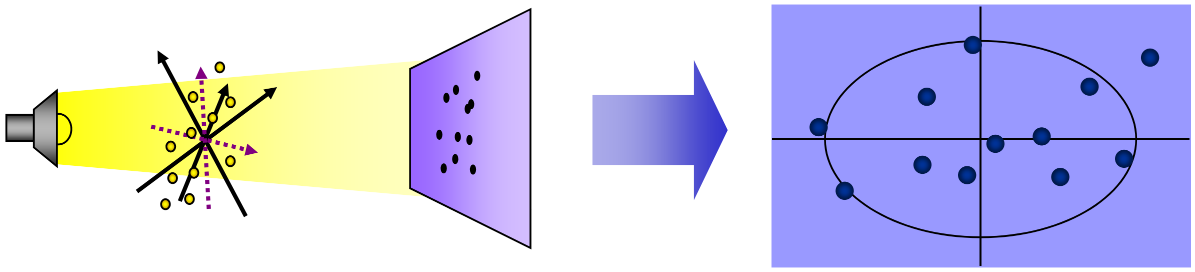
➔ Reduction of dimensionality, model in latent variables!

- Algebraically
  - Summarizes the information in the observations as a few new (latent) variables
- Geometrically
  - The swarm of points in a K dimensional space (K = number of variables) is approximated by a (hyper)plane and the points are projected on that plane.

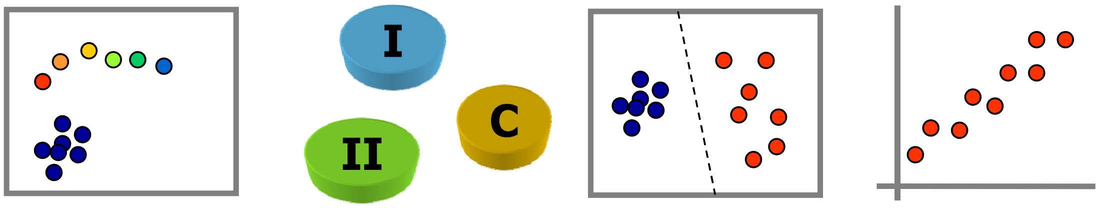


What is a Projection?

- Variables form axes in a multidimensional space
- An observation in multidimensional space is a point
- Project points onto a plane



Fundamental Data Analysis Objectives



Overview	Classification	Discrimination	Regression
Trends Outliers Quality Control Biological Diversity Patient Monitoring	Pattern Recognition Diagnostics Healthy/Diseased Toxicity mechanisms Disease progression	Discriminating between groups Biomarker candidates Comparing studies or instrumentation	Comparing blocks of omics data Metab vs Proteomic vs Genomic Correlation spectroscopy (STOCSY)
PCA	**SIMCA**	**PLS-DA** **OPLS-DA**	**O2-PLS**

Summary

- Data 2008
  - Short wide data tables
  - Highly correlated variables measuring similar things
  - Noise, missing data

- Poor methods of analysis
  - One variable at a time
  - Selection of variables (throwing away data)

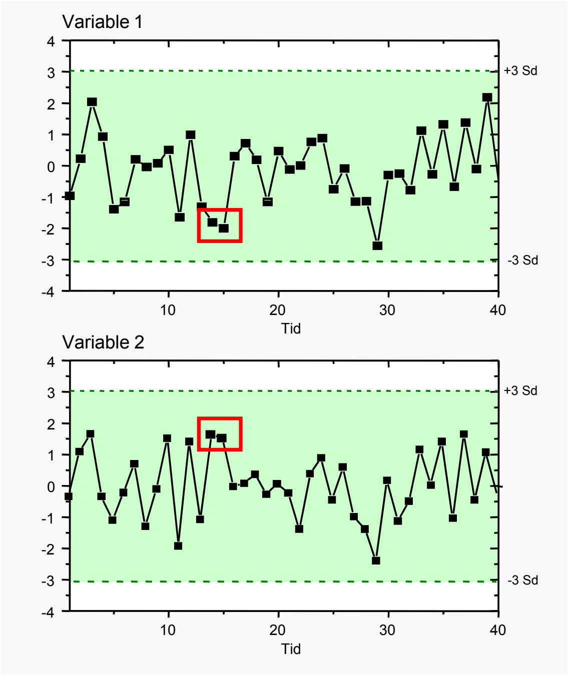
- Fundamental objectives
  - Overview & Summary
  - Classification & Discrimination
  - Relationships

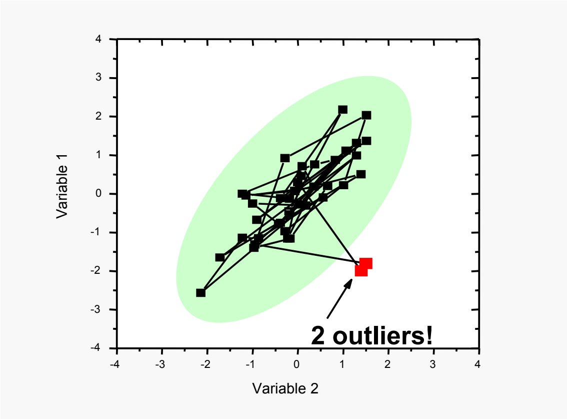
- Multivariate methods use redundancy in data to:
  - Reduce dimensionality
  - Improve stability
  - Separate signal from noise

Principal Components Analysis (PCA)

The foundation of all latent variable projection methods

Correlation between Variables





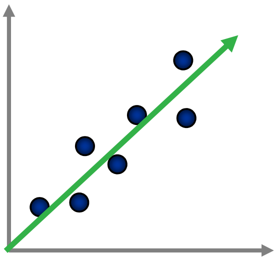
- The information is found in the correlation pattern - not in the individual variables!

Principal Components Analysis

- **Data visualisation and simplification**
  - Information resides in the ***correlation structure*** of the data
  - Mathematical principle of projection to lower dimensionality

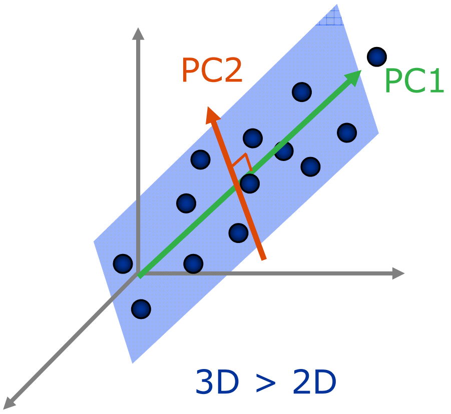
2 Variables

V1	V2
1	1.3
2	2.3
3	2.7
4	3.9
…	…



Many Variables

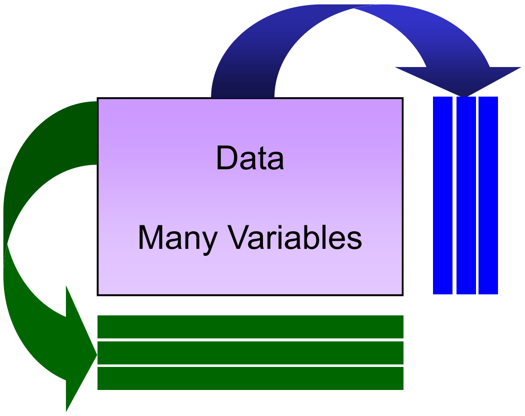
V1	V2	V3	Vn
1	1.3	0.4	
2	2.3	1.2	
3	2.7	2.1	
4	3.9	4.6	
…	…	…	…



3D > 2D

PCA Simplifies Data

- PCA breaks down a large table of data into two smaller ones
- Plots of scores and loadings turn data into pictures
- Correlations among **observations** and **variables** are easily seen



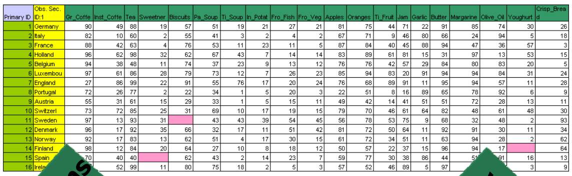
SCORES

- Summarise the observations
- Separates signal from noise
- Observe patterns, trends, clusters

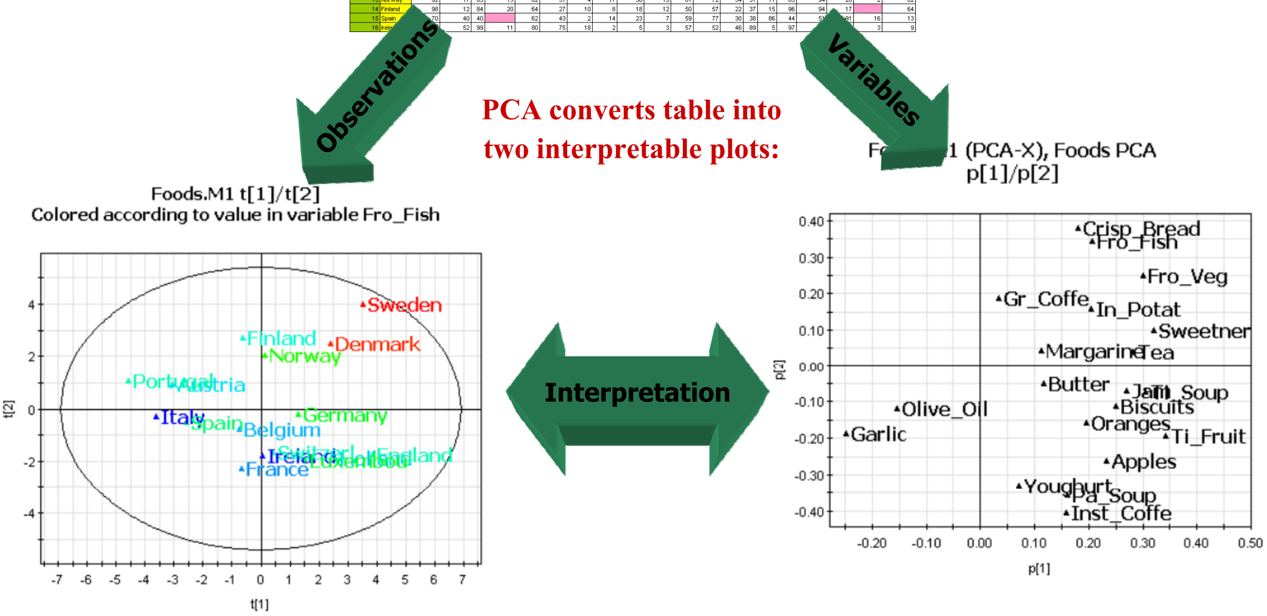
LOADINGS

- Summarise the variables
- Explain the position of observations in scores plot

PCA Converts Tables to Pictures



Primary ID	Obs. Sec. ID:1	Gr_Coffe	Inst_Coffe	Tea	Sweetner	Biscuits	Pa_Soup	Ti_Soup	In_Potat	Fro_Fish	Fro_Veg	Apples	Oranges	Ti_Fruit	Jam	Garlic	Butter	Margarine	Olive_Oil	Youghurt	Crisp_Bread
1	Germany	90	49	88	19	57	51	19	21	27	21	81	75	44	71	22	91	85	74	30	26
2	Italy	82	10	60	2	55	41	3	2	4	2	67	71	9	46	80	66	24	94	5	18
3	France	88	42	63	4	76	53	11	23	11	5	87	84	40	45	88	94	47	36	57	3
4	Holland	96	62	98	32	62	67	43	7	14	14	83	89	61	81	15	31	97	13	53	15
5	Belgium	94	38	48	11	74	37	23	9	13	12	76	76	42	57	29	84	80	83	20	5
6	Luxembou	97	61	86	28	79	73	12	7	26	23	85	94	83	20	91	94	94	84	31	24
7	England	27	86	99	22	91	55	76	17	20	24	76	68	89	91	11	95	94	57	11	28
8	Portugal	72	26	77	2	22	34	1	5	20	3	22	51	8	16	89	65	78	92	6	9
9	Austria	55	31	61	15	29	33	1	5	15	11	49	42	14	41	51	51	72	28	13	11
10	Switzerl	73	72	85	25	31	69	10	17	19	15	79	70	46	61	64	82	48	61	48	30
11	Sweden	97	13	93	31		43	43	39	54	45	56	78	53	75	9	68	32	48	2	93
12	Denmark	96	17	92	35	66	32	17	11	51	42	81	72	50	64	11	92	91	30	11	34
13	Norway	92	17	83	13	62	51	4	17	30	15	61	72	34	51	11	63	94	28	2	62
14	Finland	98	12	84	20	64	27	10	8	18	12	50	57	22	37	15	96	94	17		64
15	Spain	70	40	40		62	43	2	14	23	7	59	77	30	38	86	44	51	91	16	13
16	Irela	[illegible]	52	99	11	80	75	18	2	5	3	57	52	46	89	5	97	[illegible]	[illegible]	3	9



PCA converts table into two interpretable plots:

Scores plot relates to **observations**

Loadings plot relates to **variables**

PCA Example

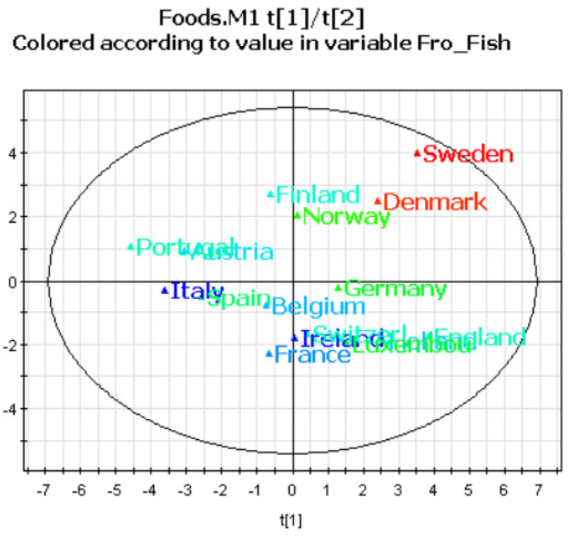
Problem: To investigate patterns of food consumption in Western Europe, the percentage usage of 20 common food products was obtained for 16 countries

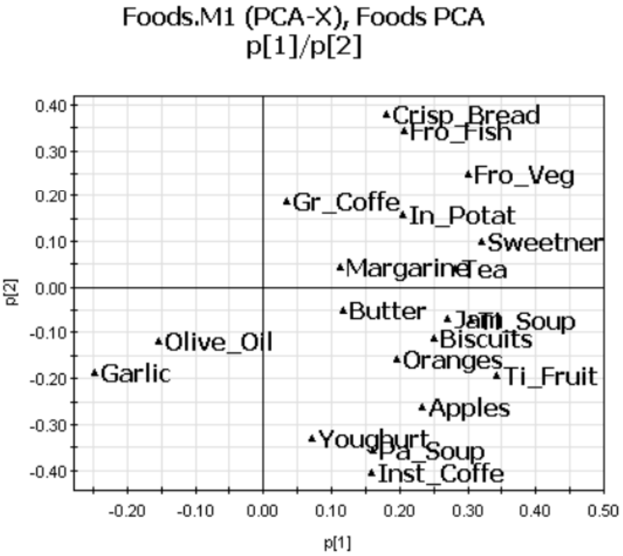
Perform a multivariate analysis (PCA) to overview data

Food consumption patterns for 16 European countries (part of the data).

COUNTRY	Grain coffee	Instant coffee	Tea	Sweet-ner	Bis-cuits	Pa soup	Ti soup	In potat	Fro fish	Fro veg	Fresh apple	Fresh orange	Ti fruit	Jam	Garlic	Butter	Marg-arine
Germany	90	49	88	19	57	51	19	21	27	21	81	75	44	71	22	91	85
Italy	82	10	60	2	55	41	3	2	4	2	67	71	9	46	80	66	24
France	88	42	63	4	76	53	11	23	11	5	87	84	40	45	88	94	47
Holland	96	62	98	32	62	67	43	7	14	14	83	89	61	81	15	31	97
Belgium	94	38	48	11	74	37	23	9	13	12	76	76	42	57	29	84	80
Luxembou	97	61	86	28	79	73	12	7	26	23	85	94	83	20	91	94	94
England	27	86	99	22	91	55	76	17	20	24	76	68	89	91	11	95	94
Portugal	72	26	77	2	22	34	1	5	20	3	22	51	8	16	89	65	78
Austria	55	31	61	15	29	33	1	5	15	11	49	42	14	41	51	51	72
Switzerl	73	72	85	25	31	69	10	17	19	15	79	70	46	61	64	82	48
Sweden	97	13	93	31		43	43	39	54	45	56	78	53	75	9	68	32
Denmark	96	17	92	35	66	32	17	11	51	42	81	72	50	64	11	92	91
Norway	92	17	83	13	62	51	4	17	30	15	61	72	34	51	11	63	94
Finland	98	12	84	20	64	27	10	8	18	12	50	57	22	37	15	96	94
Spain	70	40	40		62	43	2	14	23	7	59	77	30	38	86	44	51
Ireland	30	52	99	11	80	75	18	2	5	3	57	52	46	89	5	97	25

General PCA Example - Foods


Foods.M1 t[1]/t[2]
Colored according to value in variable Fro_Fish


Foods.M1 (PCA-X), Foods PCA
p[1]/p[2]

Observations

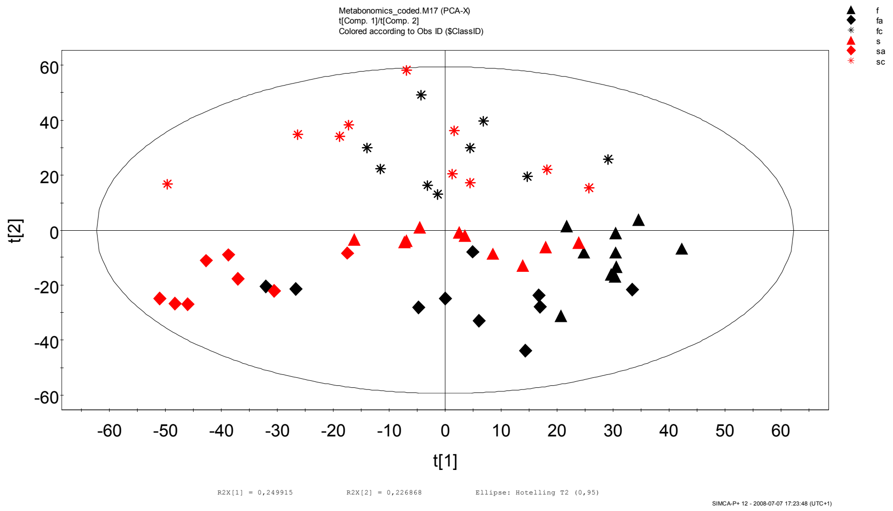
Scores plot

Variables

Loadings plot

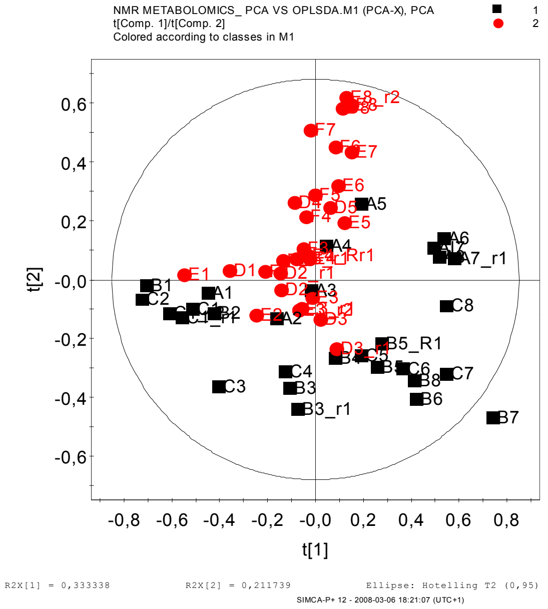
PCA to overview 1

- Example: Toxicity study of rats
- Two different types of rats and two different types of drugs were used
  - aim: identify trends and biomarkers for toxicity
- PCA useful to identify outliers, biological diversity and toxicity trends



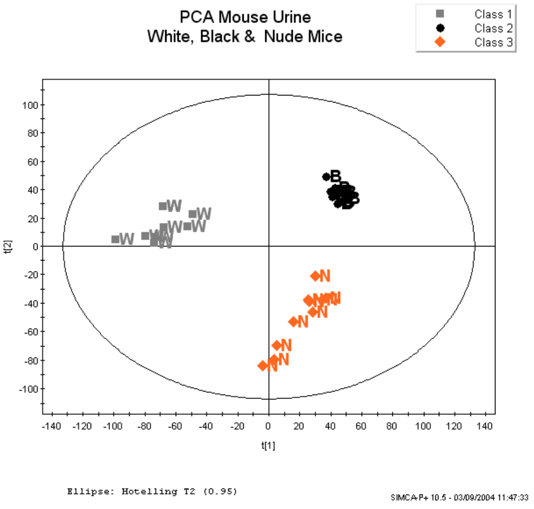
PCA for Overview 2

- Example: HR/MAS ^{1}H NMR study from poplar plants
  - Aim: biomarkers to explore biology
- Scores plot shows poplar samples from two different types one wild type and the other transgenic
- Interpretation of PCA scores shows patterns and trends



PCA for Overview 3

- Genetic study of mice
  - Black, White, Nude
  - Mass Lynx data
- PCA useful for QC of biological results:
  - Biological diversity
  - Outlier detection
  - Finding trends



Data courtesy of Ian Wilson and Waters Corporation Inc

Summary

- Data is not Information
- Information lies in correlation structure
- Projection methods explain correlation structure among all variables
- PCA provides graphical overview – natural starting point for any multivariate data analysis
- PCA gives
  - Scores: summary of observations
  - Loadings: summary of variables

Multivariate Analysis for ”omics” data

Chapter 2

Overview of Data Tables:

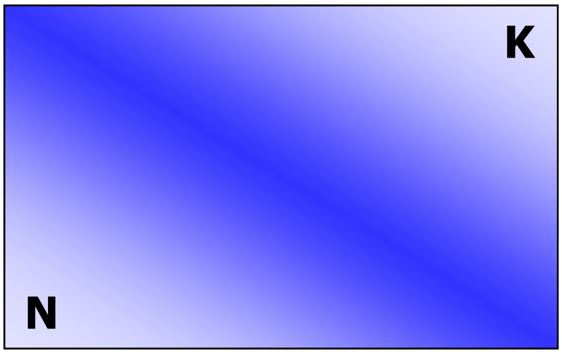
Principal Components Analysis (PCA)

Contents

- Notations
- Scaling
- Geometric interpretation
- Algebraic interpretation
- Example
- PCA diagnostics

Notation

- N Observations
  - Humans
  - Plants
  - Other individuals
  - Trials
  - Etc
- K Variables
  - Spectra
  - Peak tables
  - Etc



Notation

N = number of observations

K = number of variables

A = number of principal components

ws = scaling weights

$t_1, t_2,\dots, t_A$ scores (forming matrix T)

$p_1, p_2,\dots, p_A$ loadings (forming matrix P)

Key Concepts with Multivariate Methods

1. Data must be scaled or transformed appropriately

2. Data may need to be ‘cleaned’
   - Outliers
     - Interesting
     - But they can upset a model
     - Must detect, investigate and possibly removed

3. Need to determine how well the model fits the data.

4. Fit does not give Predictive ability!
   - Model information not noise – avoid overfit
   - Need to estimate predictive ability

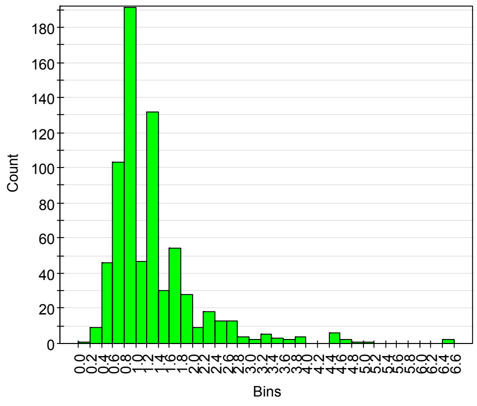
Data Pre-Processing - Transformations

If the data are not approximately normally distributed, a suitable transformation might be required to improve the modelling results

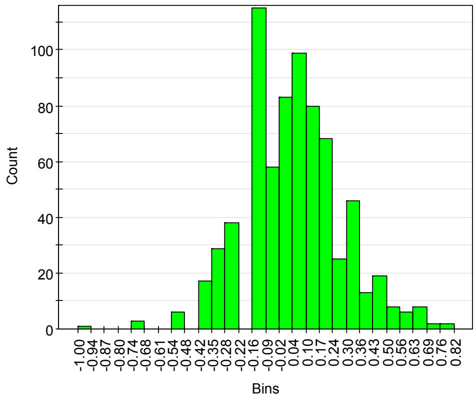
- Before transformation
  – skew distribution

- After log-transformation
  – More close to normal distribution

cuprum.DS1 cuprum
Histogram of DS1.kNi



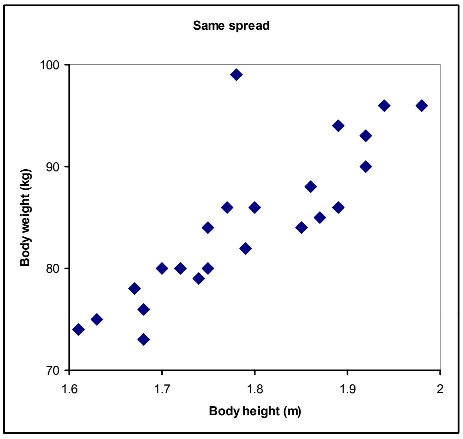
cuprum.M1 (PCA-X), PCA for overview log-transform
Histogram of M1.XVar(kNi)(trans)



Scaling Example - Height vs Weight

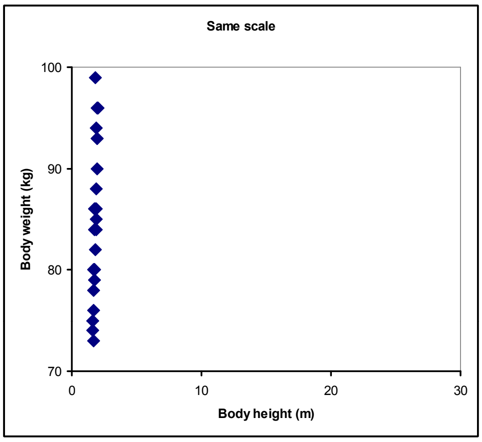
Data for 23 individuals (22 players + referee in a football match)

Height (m)	**1.8**	**1.61**	**1.68**	**1.75**	**1.74**	**1.67**	**1.72**	**1.98**	**1.92**	**1.7**	**1.77**	**1.92**
Weight (kg)	**86**	**74**	**73**	**84**	**79**	**78**	**80**	**96**	**90**	**80**	**86**	**93**
Height (m)	**1.6**	**1.85**	**1.87**	**1.94**	**1.89**	**1.89**	**1.86**	**1.78**	**1.75**	**1.8**	**1.68**	
Weight (kg)	**75**	**84**	**85**	**96**	**94**	**86**	**88**	**99**	**80**	**82**	**76**	



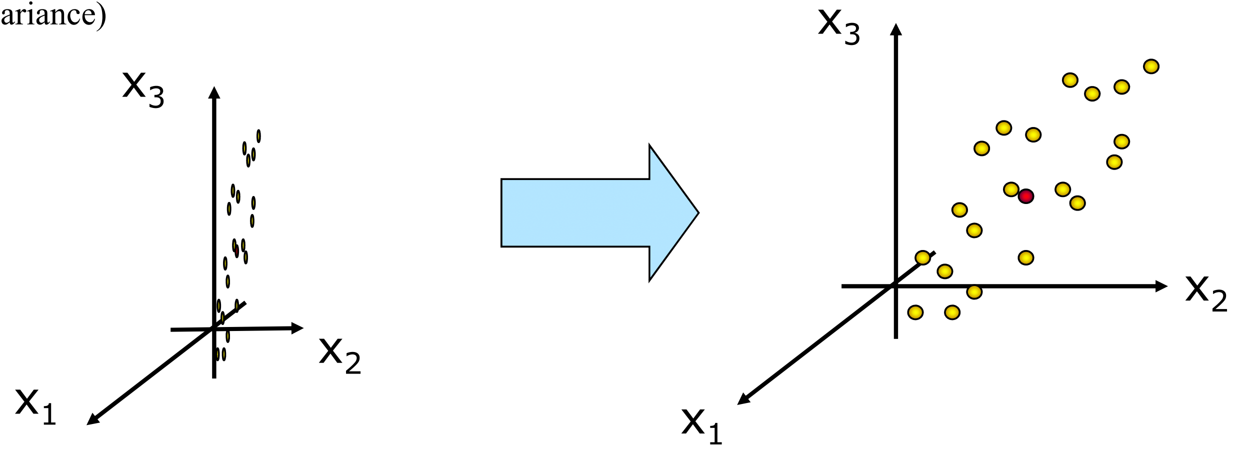
Left: scaled

Right: unscaled, outlier is not so easy to spot!



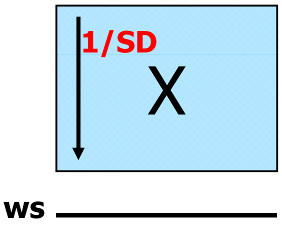
Data Pre-Processing - Scaling

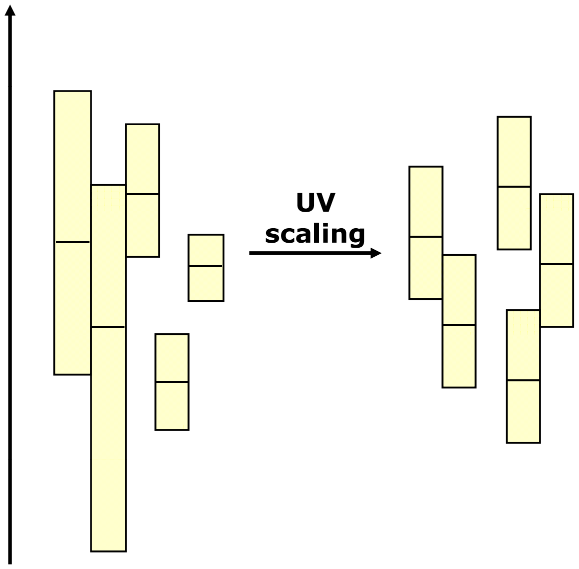
- **Problem:** Variables can have substantially different ranges
- Different ranges can cause problems for modelling and interpretation
- Defining the length of each variable axis i.e. the SD
- Default in SIMCA: To set variation along each axis to one (unit variance)



Unit Variance Scaling (UV)

- PCA is scale dependent
  - Is the size of a variable important?





- Scaling weight is 1/SD for each variable i.e. divide each variable by its standard deviation
- Variance of scaled variables = 1

Summary

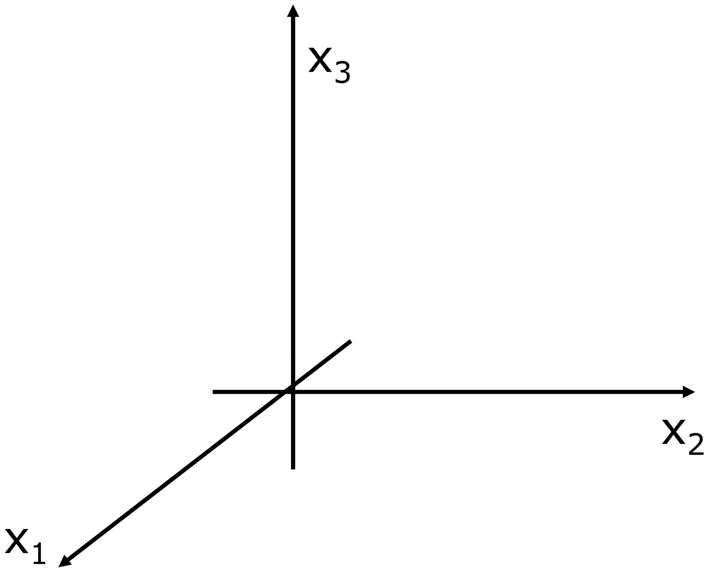
- Variables may need to be transformed prior to analysis to make them more normally distributed
- Results are scale dependent – which scaling is appropriate?
  - (will come back to this in chapter 4)
- Default is UV scaling – all variables given equal weight
- Not usually recommended with spectroscopic data where no scaling is the norm
- Compromise is Pareto scaling which is commonly used in metabonomic studies (Chapter 4)

Multivariate Analysis for ”omics” data

How PCA Works

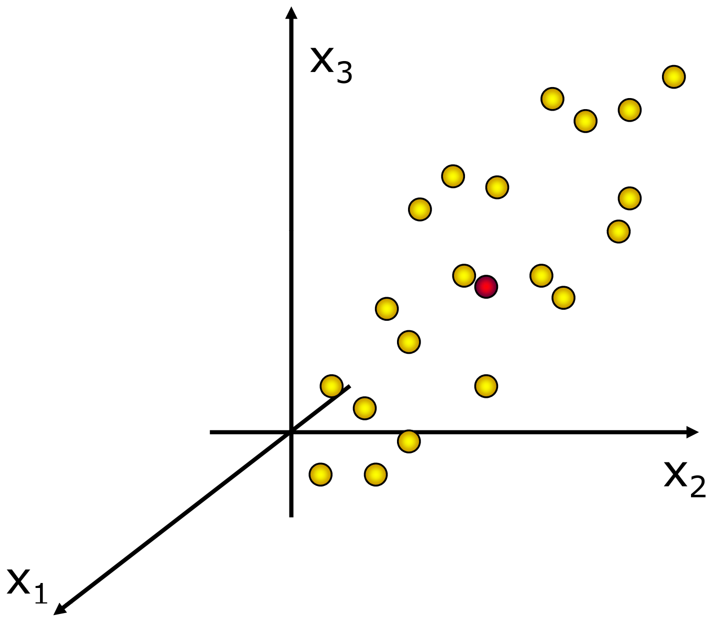
UMETRICS

PCA - Geometric Interpretation



- We construct a space with K dimensions – 3 shown for illustration
- Each variable is an axis with its length determined by scaling, typically unit variance

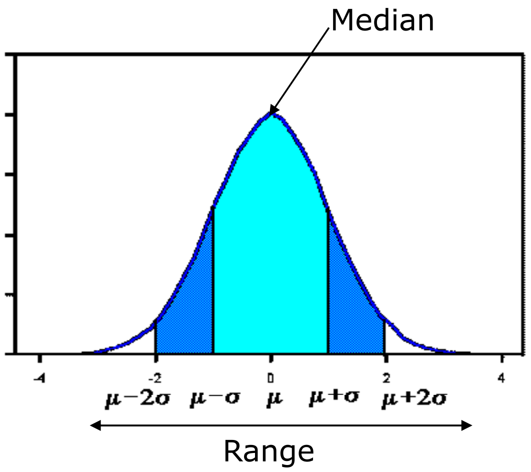
PCA - Geometric Interpretation



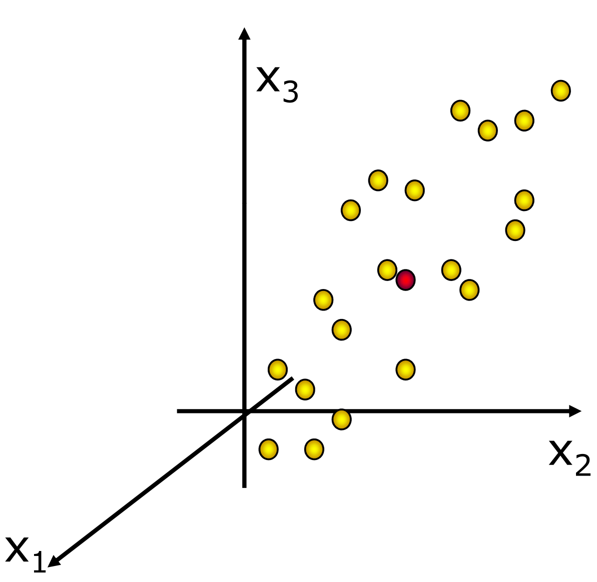
- Each observation is represented by a point in K-dimensional space
- Hence, the data table X is a swarm of points in this space

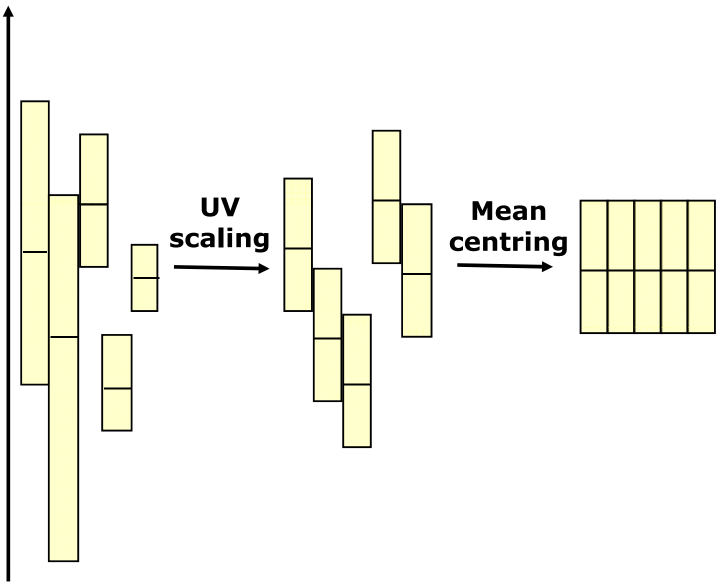
Data: Measurements made on a system

- Each variable are represented by
  - average = avg = $\Sigma x_i / N$
  - median (middle point)
  - SD = s = $[\Sigma(x_i - \text{avg})^2/(N-1)]^{1/2}$
  - Range: largest - smallest value
  - Variance = $s^2 = SD^2 = \Sigma(x_i - \text{avg})^2/(N-1)$



PCA – Mean Centring





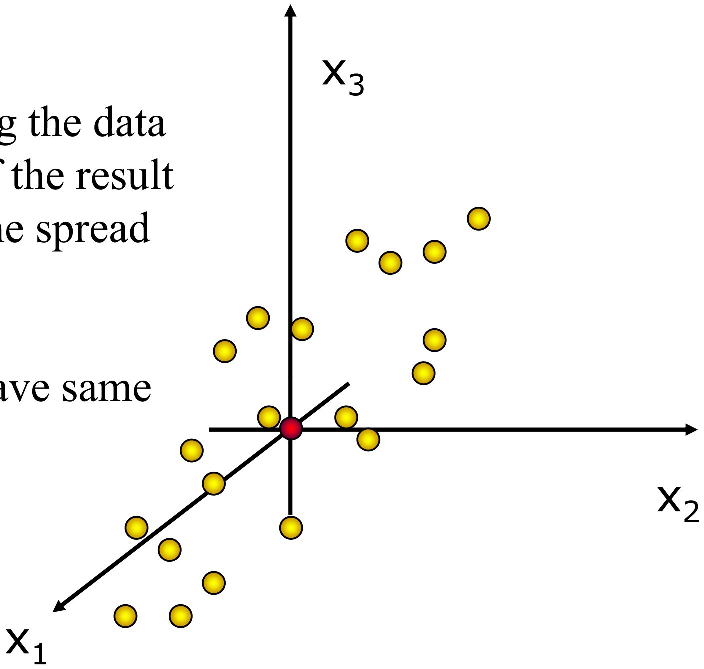
First we calculate the average (mean) of each variable, which itself is a point in K-space, and subtract it from each point

UMETRICS 8/15/2008

PCA - Mean Centring

When mean centring the data the interpretation of the result will be relative to the spread around the mean

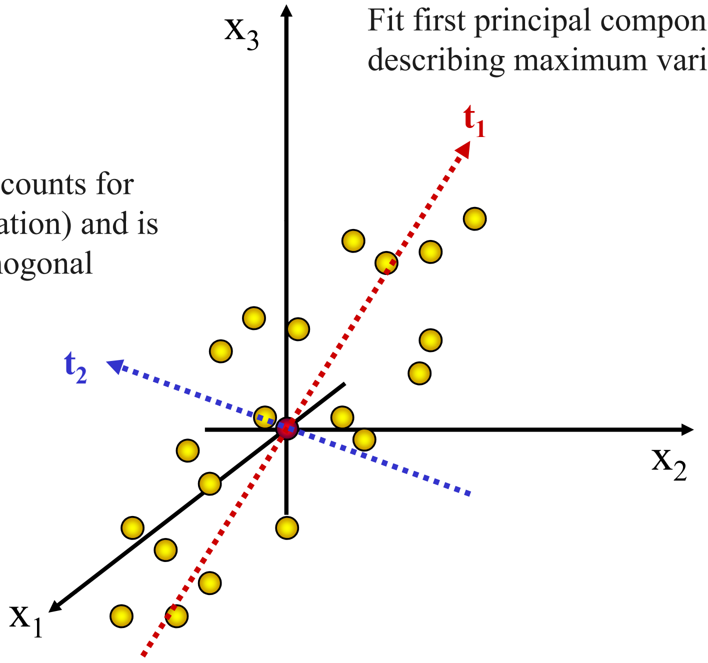
All variables will have same reference point



The mean-centring procedure corresponds to moving the co-ordinate system to the origin

UMETRICS 8/15/2008

PCA - Geometric Interpretation

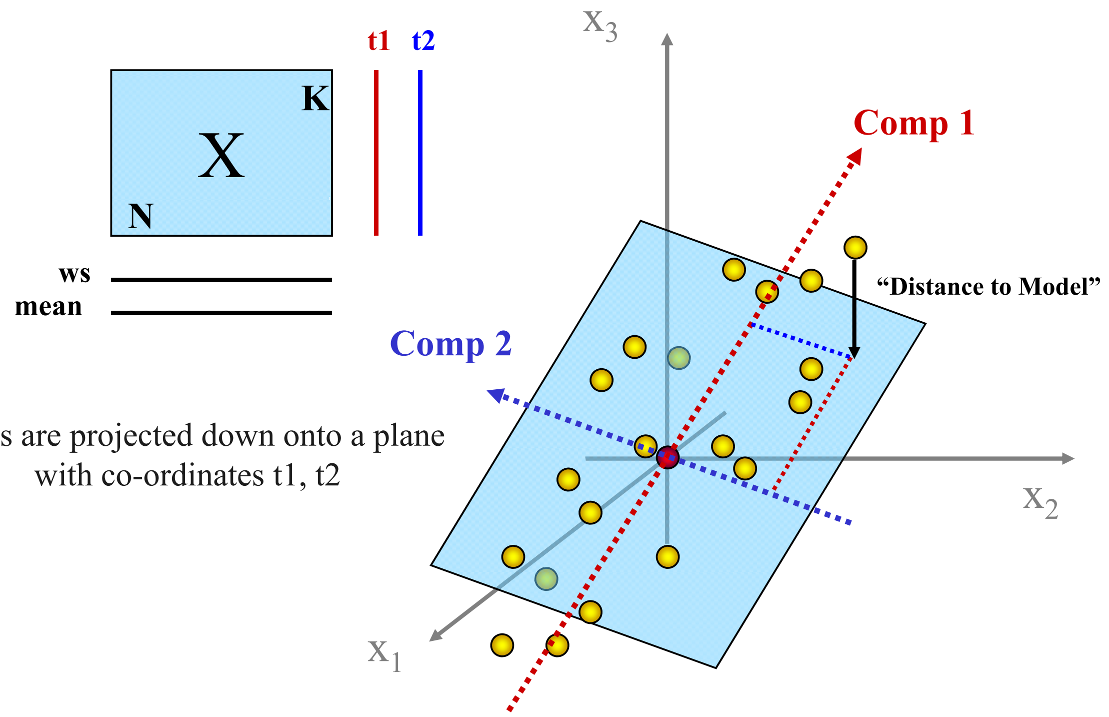


Fit first principal component (line describing maximum variation)

Add second component (accounts for next largest amount of variation) and is at right angles to first - orthogonal

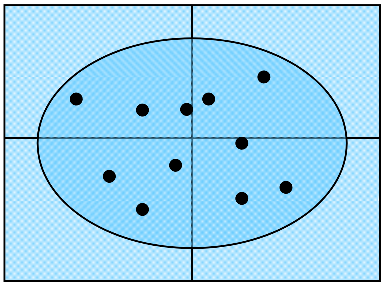
Each component goes through origin

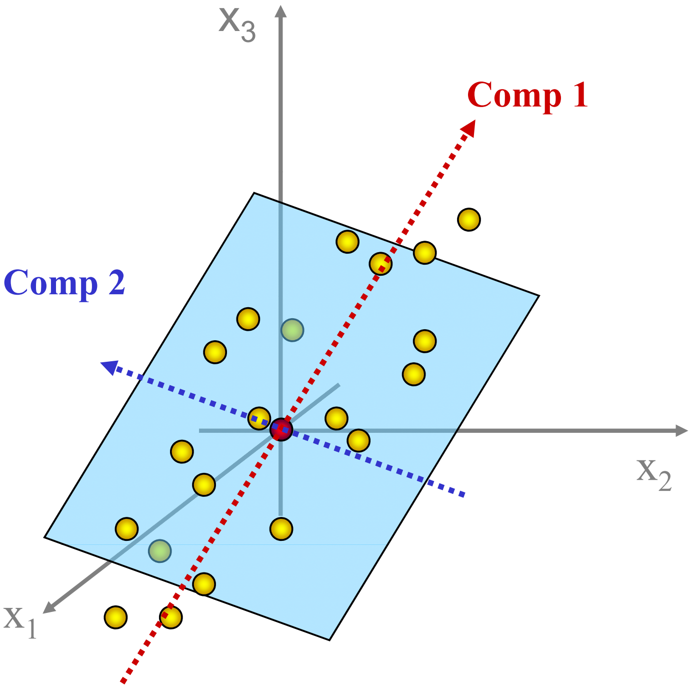
PCA - Geometric Interpretation



Points are projected down onto a plane with co-ordinates t1, t2

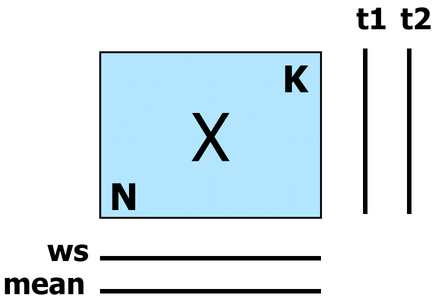
Projection onto a Plane





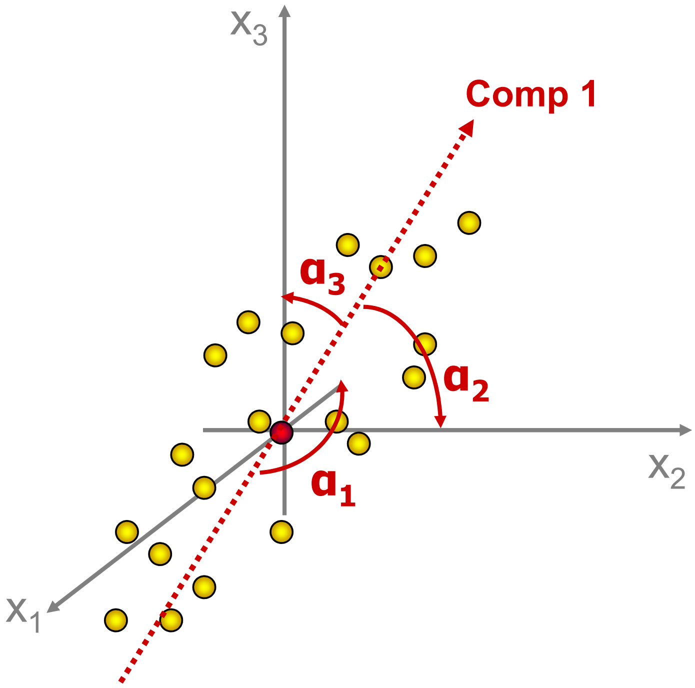
Plane is then extracted for viewing on computer screen

Loadings

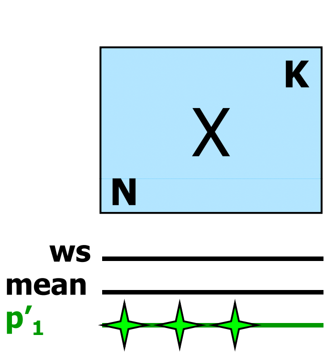


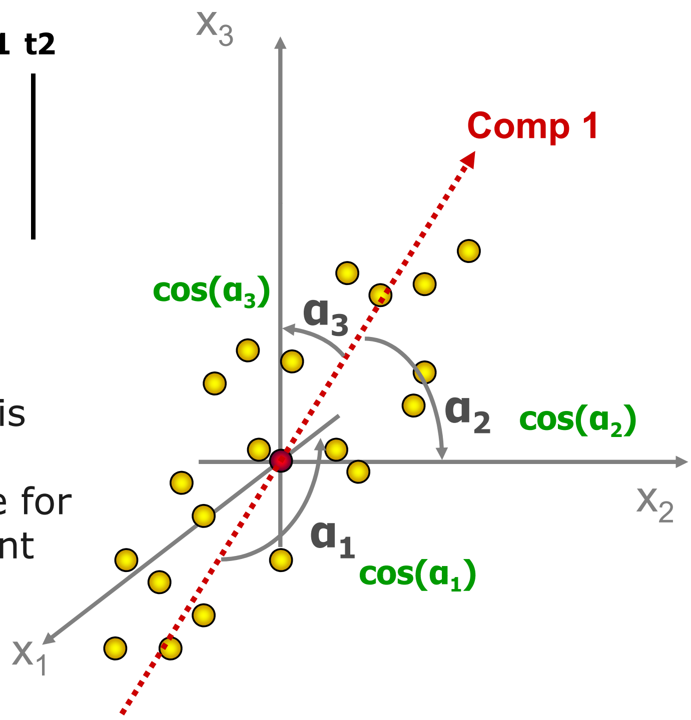
How do the principal components relate to the original variables?

Look at the angles between PCs and variable axes



Loadings





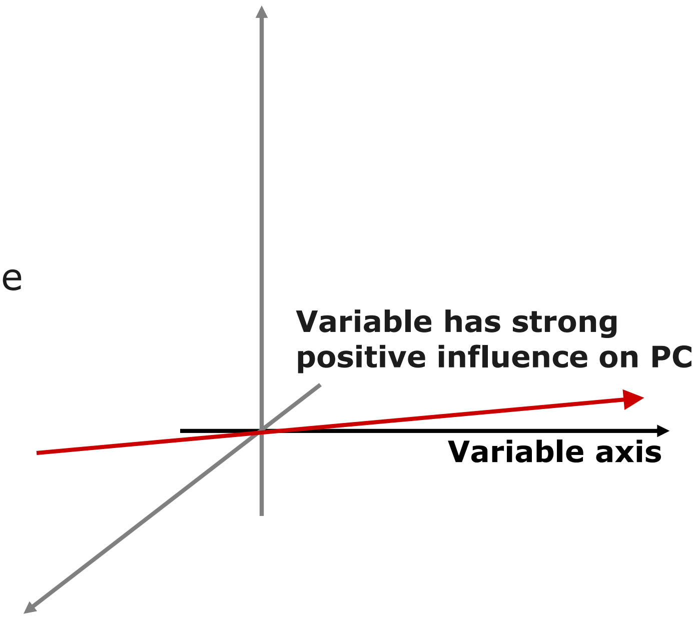
Take cos(**α**) for each axis

Loadings vector p' - one for each principal component

One value per variable

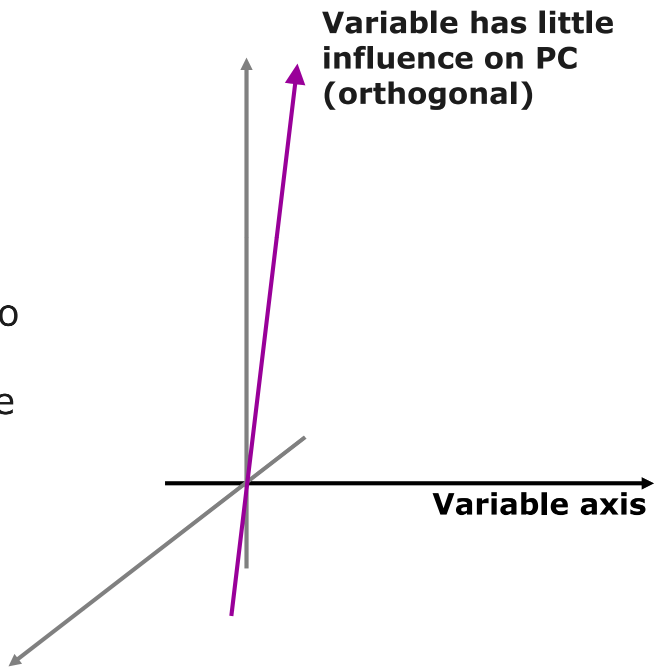
Positive Loading

If component lines up with variable axis the loading will be close to 1 showing strong influence => cos(0) = 1



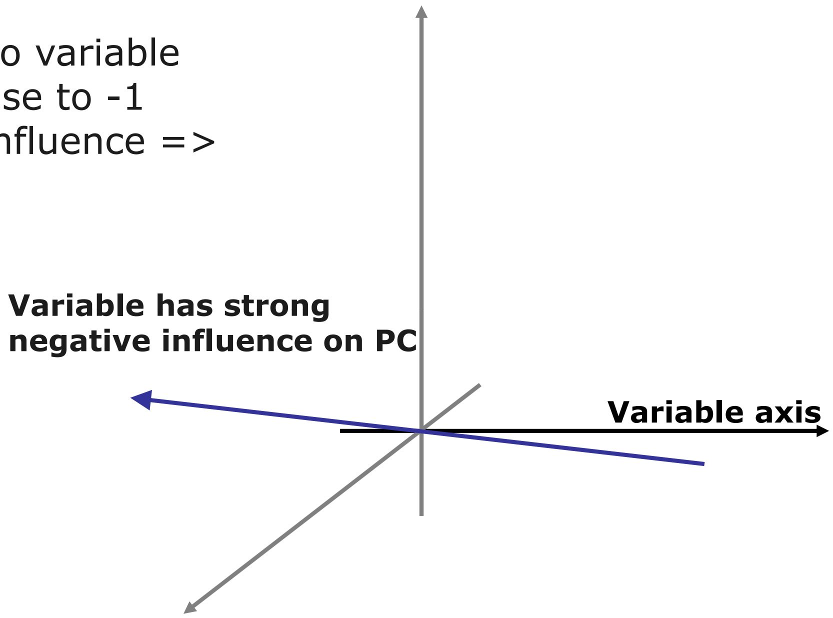
Zero Loading

If component is at right angles to variable axis the loading will be close to 0 showing little influence => cos(90) = 0



Negative Loading

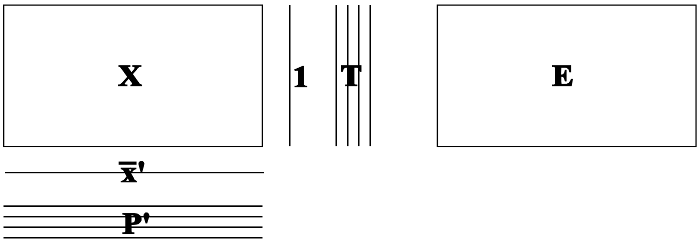
If component is opposite to variable axis the loading will be close to -1 showing strong negative influence => cos(180) = -1



Algebraic interpretation of scores

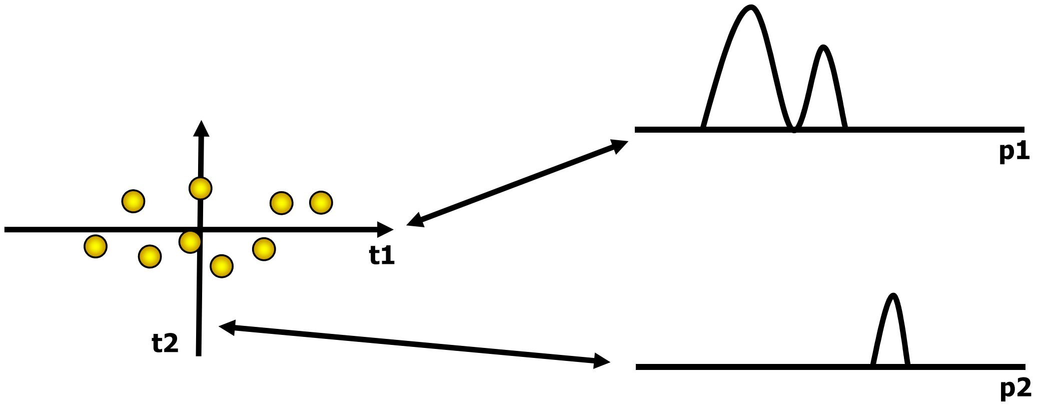
- The **scores** $\mathbf{t}_{ia}$ (comp. **a**, obs. **i**) are the places along the lines where the observations are projected
- The scores, $\mathbf{t}_{ia}$, are *new variables* that best **summarize** the old ones; **linear combinations** of the old ones with coefficients $\mathbf{p}_{ak}$
- Sorted on importance, $\mathbf{t}_1$, $\mathbf{t}_2$, $\mathbf{t}_3$,...

$$\mathbf{X} = \mathbf{1} * \bar{\mathbf{x}}' + \mathbf{T} * \mathbf{P}' + \mathbf{E}$$

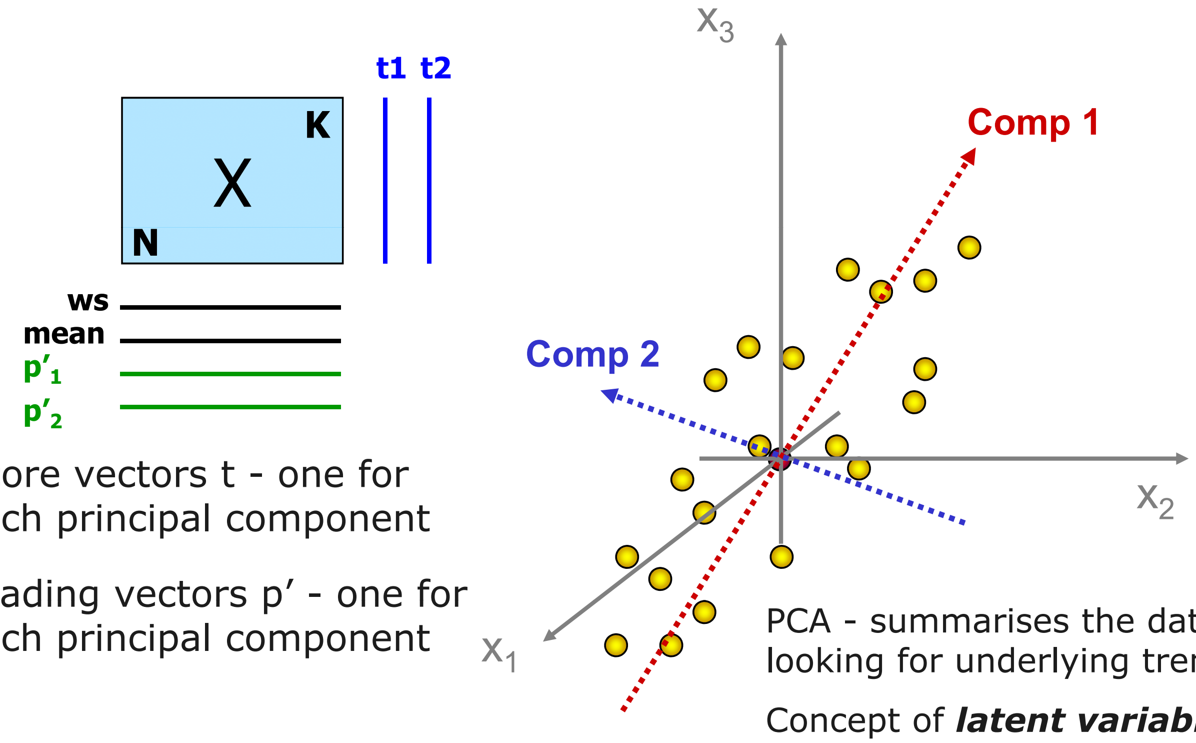


PCA interpretation

- Direction observed in t1 can be explained by looking at corresponding p1
- Direction observed in t2 can be explained by looking at corresponding p2



Summary 1



Score vectors t - one for each principal component

Loading vectors p' - one for each principal component

PCA - summarises the data by looking for underlying trends

Concept of ***latent variables***

Summary 2

- The scores, t_i, are new variables that summarise the original ones
- The scores are sorted in descending order of importance, t_1, t_2, t_3 etc
- Typically, 2-5 principal components are sufficient to summarise a data table well
- The loadings, p_k, express how the original variables relate to the scores - scores are linear combinations of the original variables
- The principal components define a new co-ordinate system describing the variation in the data

Multivariate Analysis
for ”omics” data

A PCA Example

UMETRICS

PCA Example - FOODS

PCA for Overview

Problem: To investigate food consumption patterns in Western Europe, the percentage usage of 20 common food products was obtained for 16 countries

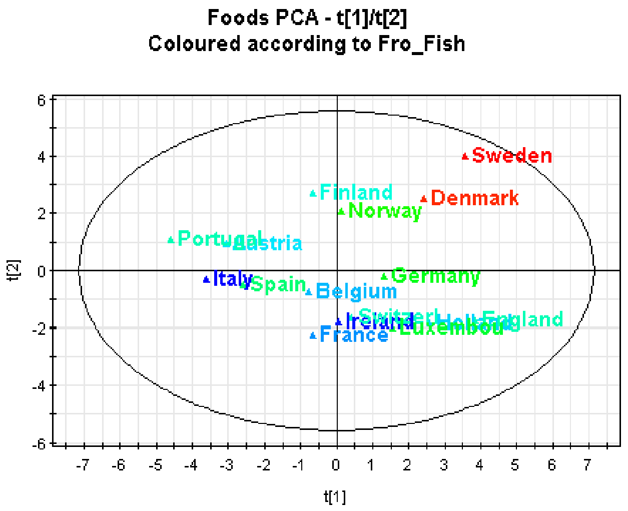
Perform a multivariate analysis (PCA) to overview data

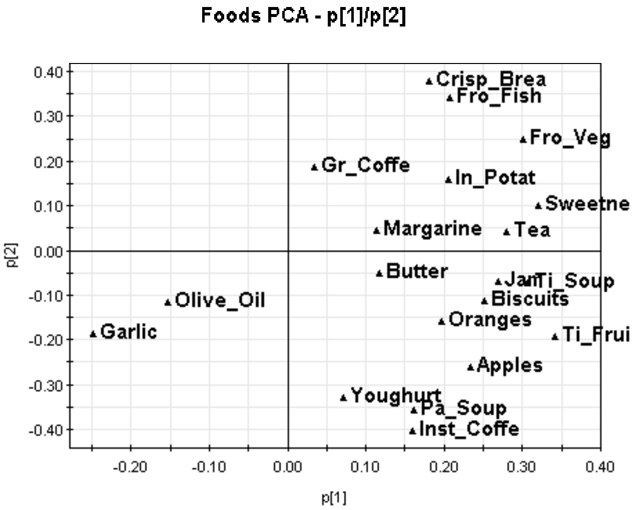
Food consumption patterns for 16 European countries (part of the data).

COUNTRY	Grain coffee	Instant coffee	Tea	Sweet-ner	Bis-cuits	Pa soup	Ti soup	In potat	Fro fish	Fro veg	Fresh apple	Fresh orange	Ti fruit	Jam	Garlic	Butter	Marg-arine
Germany	90	49	88	19	57	51	19	21	27	21	81	75	44	71	22	91	85
Italy	82	10	60	2	55	41	3	2	4	2	67	71	9	46	80	66	24
France	88	42	63	4	76	53	11	23	11	5	87	84	40	45	88	94	47
Holland	96	62	98	32	62	67	43	7	14	14	83	89	61	81	15	31	97
Belgium	94	38	48	11	74	37	23	9	13	12	76	76	42	57	29	84	80
Luxembou	97	61	86	28	79	73	12	7	26	23	85	94	83	20	91	94	94
England	27	86	99	22	91	55	76	17	20	24	76	68	89	91	11	95	94
Portugal	72	26	77	2	22	34	1	5	20	3	22	51	8	16	89	65	78
Austria	55	31	61	15	29	33	1	5	15	11	49	42	14	41	51	51	72
Switzerl	73	72	85	25	31	69	10	17	19	15	79	70	46	61	64	82	48
Sweden	97	13	93	31		43	43	39	54	45	56	78	53	75	9	68	32
Denmark	96	17	92	35	66	32	17	11	51	42	81	72	50	64	11	92	91
Norway	92	17	83	13	62	51	4	17	30	15	61	72	34	51	11	63	94
Finland	98	12	84	20	64	27	10	8	18	12	50	57	22	37	15	96	94
Spain	70	40	40		62	43	2	14	23	7	59	77	30	38	86	44	51
Ireland	30	52	99	11	80	75	18	2	5	3	57	52	46	89	5	97	25

UMETRICS

PCA Example - FOODS





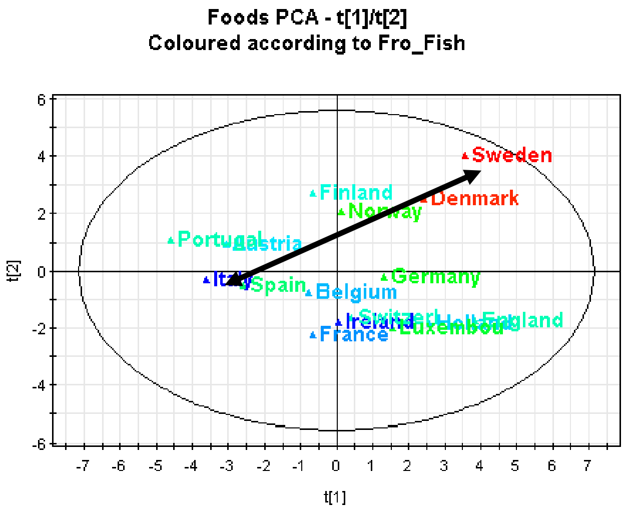
Observations

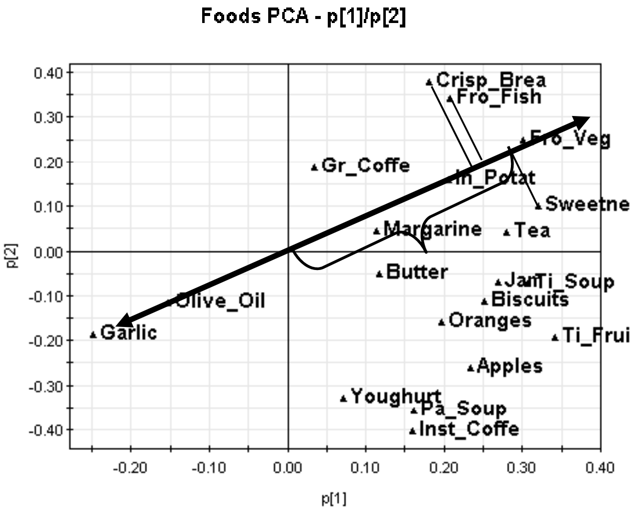
A=3

Variables

What type of information can be seen?
Any groupings?

PCA Example - FOODS



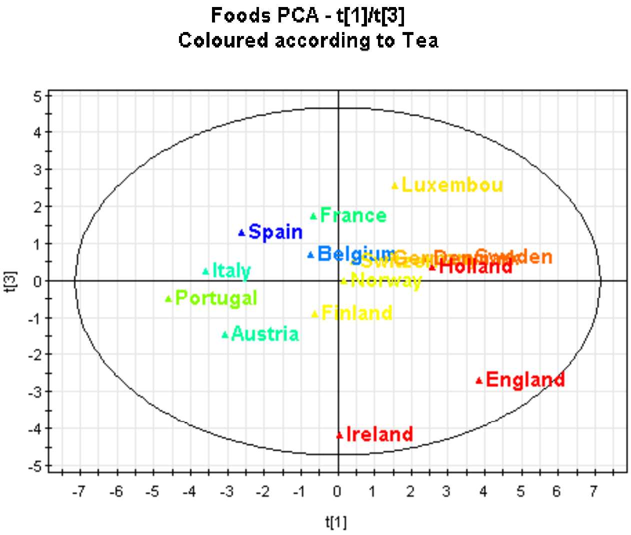


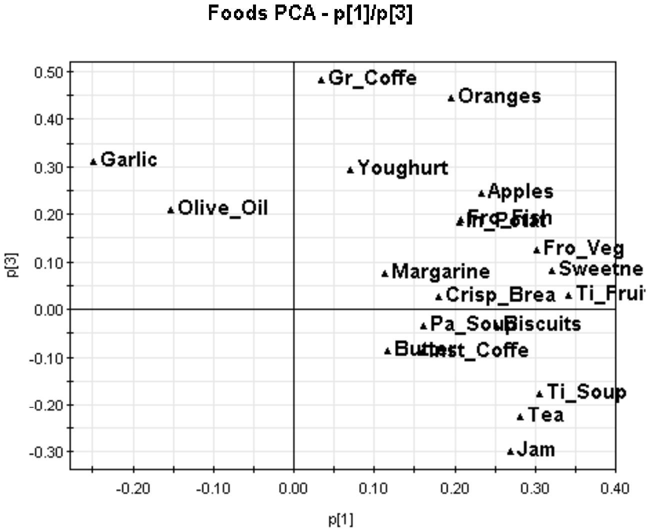
Observations

Variables

Why are Italy and Spain different from Sweden and Denmark?

PCA Example - FOODS



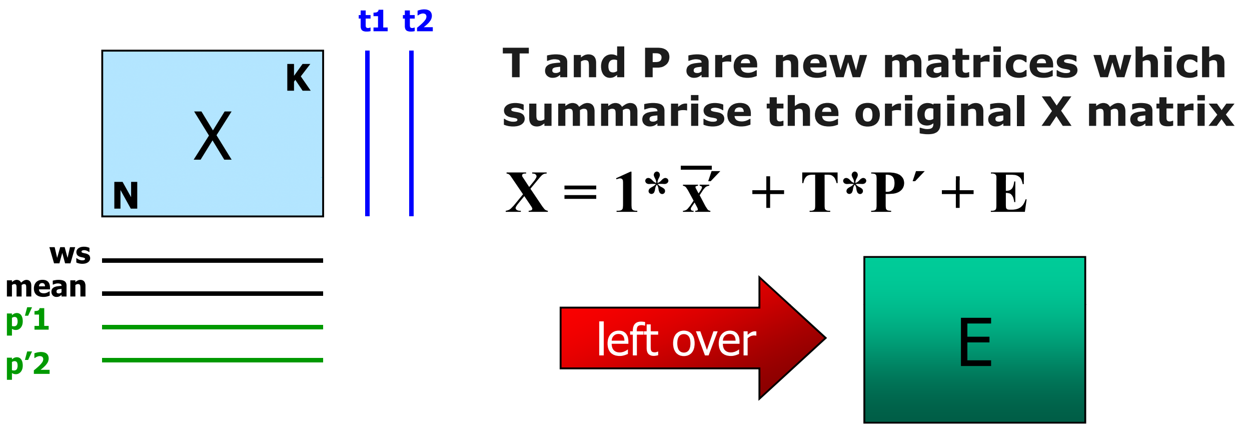


Observations

Variables

In the third component Ireland and England are different from the other countries

Summary 3



T and P are new matrices which summarise the original X matrix

$$X = 1 * \bar{x}' + T * P' + E$$

What's left over is the **residual** (or error) **matrix**

This contains the ***unexplained*** variation

The better the model the smaller the errors

Multivariate Analysis for ”omics” data

PCA Diagnostics

How good is our model?

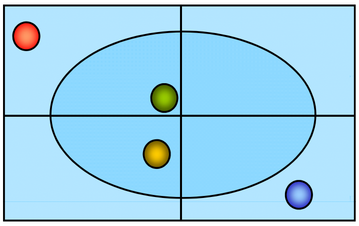
PCA - Diagnostics

- **Observation** diagnostics
  - strong and moderate outliers
  - groups
  - trends
- **Variable** diagnostics
  - correlation
  - contribution
  - which variables are well explained
- **Model** diagnostics
  - fit (R^2)
  - predictive ability (Q^2), cross-validated

Observation Diagnostics

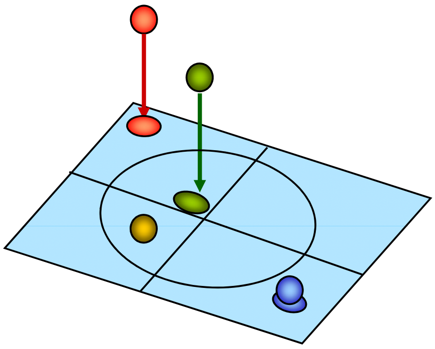
Strong outliers:

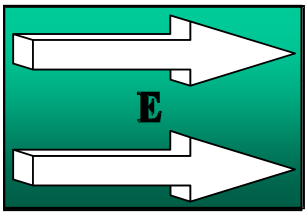
- Found in scores
- Detection tool: Hotelling’s T^2
  – defines “normal” area in score plots



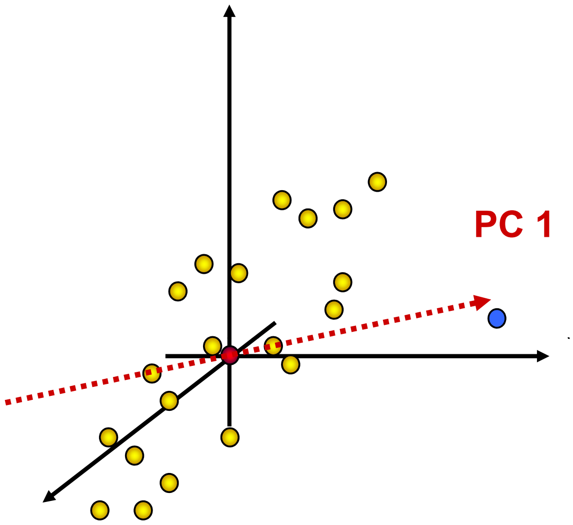
Moderate outliers:

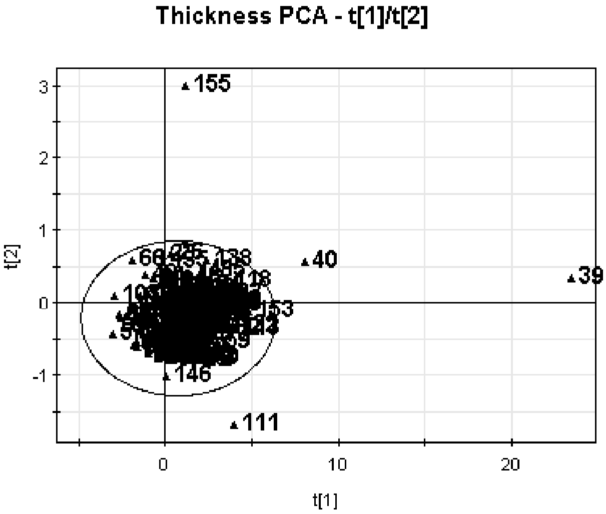
- Found in observation residuals
- Detection tool: DModX (distance to model)
- Summing and squaring residual matrix row-wise





Strong Outliers

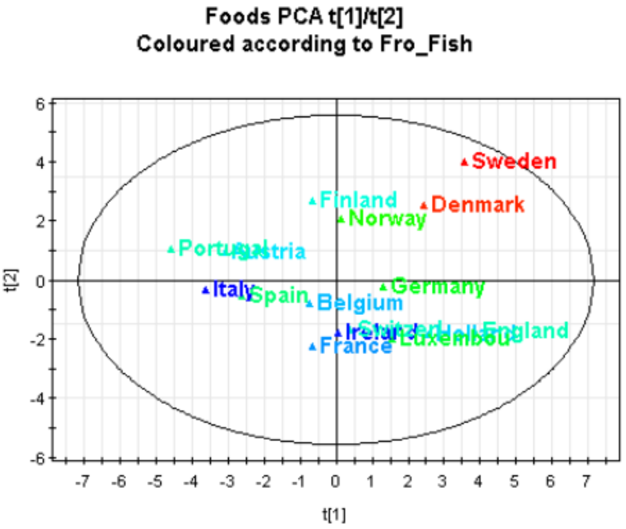


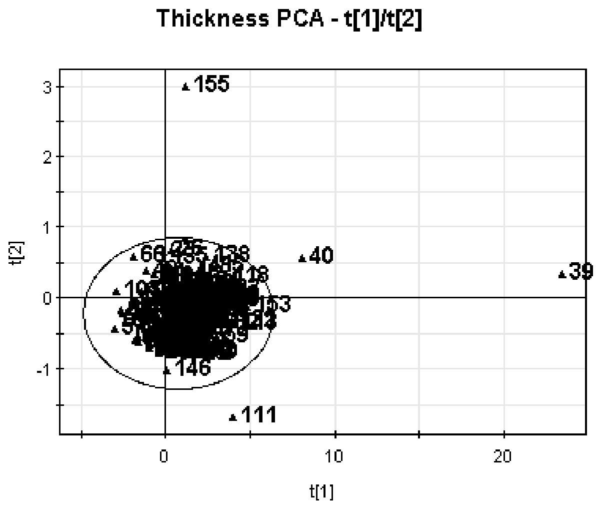


- **Outliers** are serious, interesting and easy to find
- **Strong outliers** are seen in score plots

Strong Outliers - Hotelling's T^2

- Hotelling’s T^2 is multivariate generalisation of Student’s t-distribution
- It provides a tolerance region for the data in a two-dimensional score plot, e.g., t_1/t_2



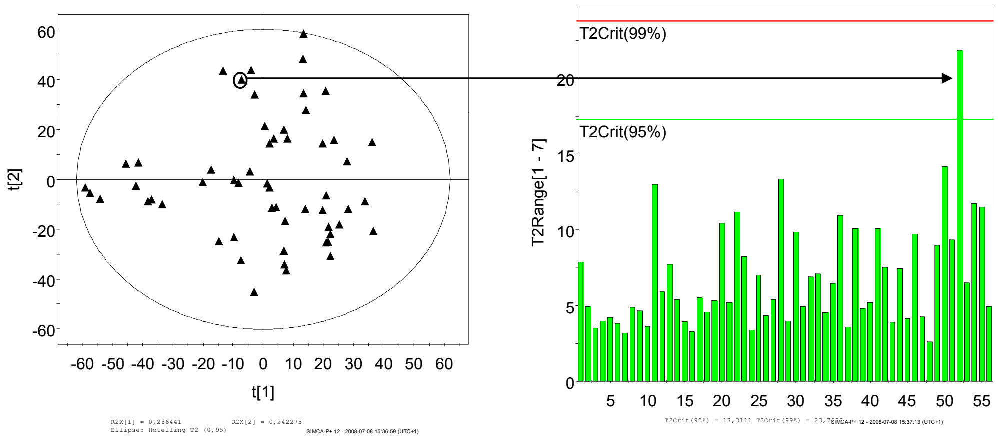


Strong Outliers - Hotelling's T^2

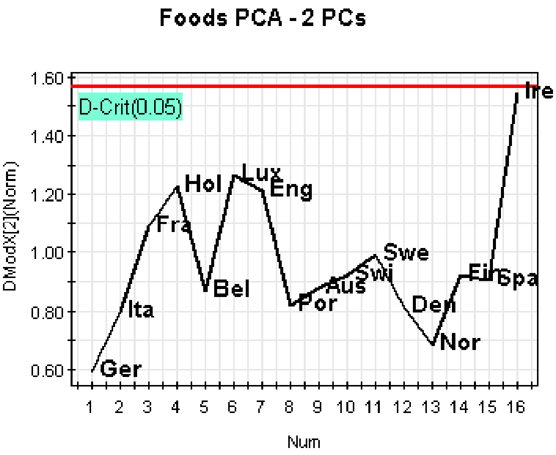
- With two components T^2 is easily visualized in the scores plot
- For more than two components look at the hotellings T^2 range plot

Score plot

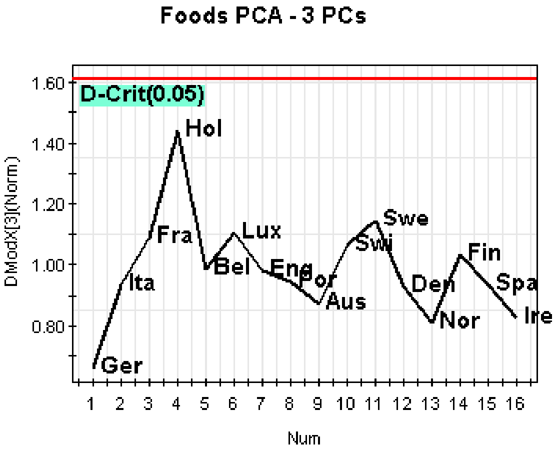
Hotellings T^2 range



Moderate Outliers (DModX)



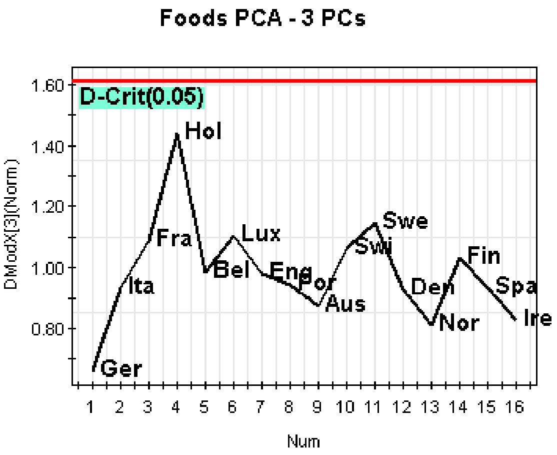
A=2

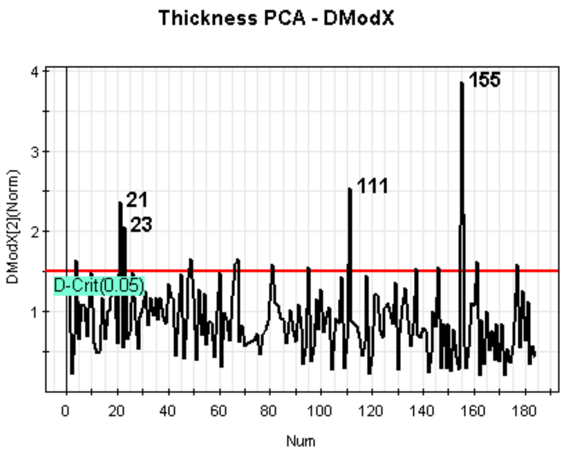


A=3

- DModX shows the distance to the model plane
- Ireland is modelled well by the third component

Moderate Outliers (DModX)



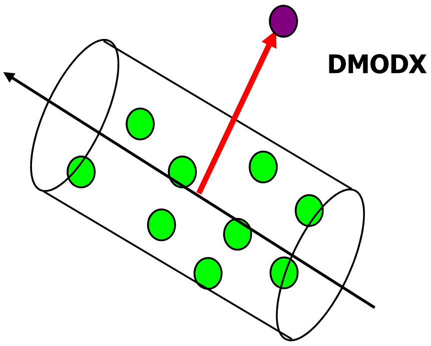


No moderate outliers

Four moderate outliers

Moderate Outliers (DModX)

- DModX shows the distance to the model plane for each observation

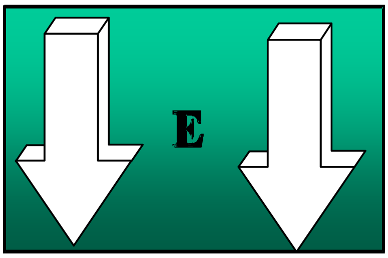


Variable Diagnostics

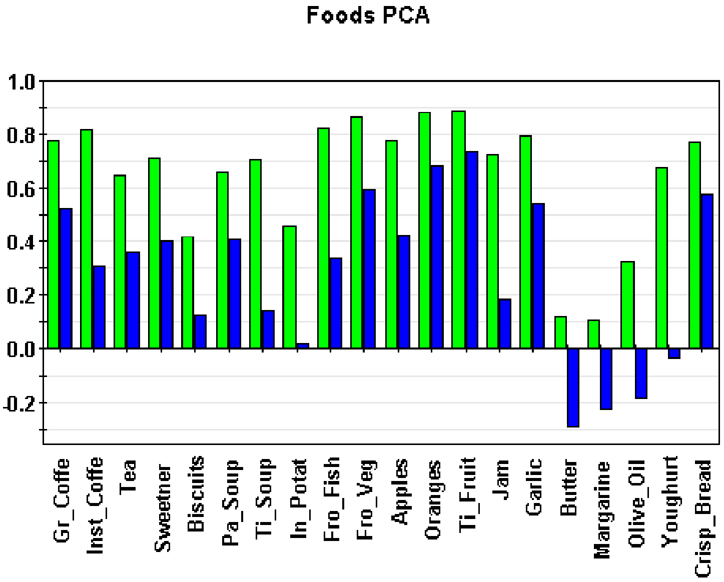
- The residuals also tell us how well each variable is modelled (R^2 value from 0 to 1)
  - Residuals of E matrix pooled column-wise

- $RSS_k = \Sigma$ (observed - fitted)^2 for variable k

- $R^2_k = 1 - RSS_k / SSX_k$



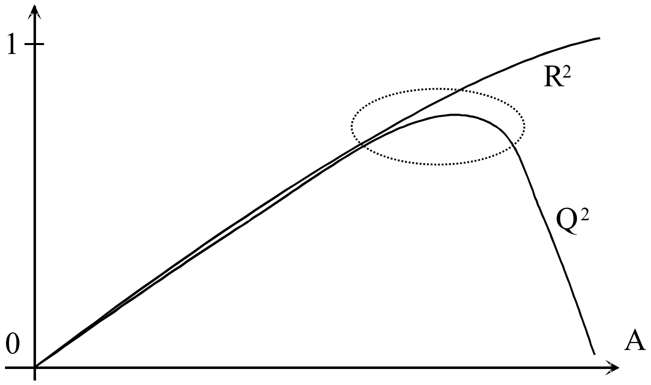
Variable Diagnostics – R^2/Q^2



- R^2 and Q^2 tell us which variables are well explained and which are not

Model Diagnostics - Validity vs Complexity

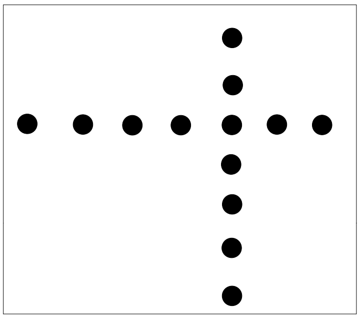
- Trade-off between fit and predictive ability
- **Question:** How can we determine the appropriate number of principal components for a particular model?
- **Answer:** cross-validation which simulates the true predictive power of a model.



R^2 estimates goodness of fit
Q^2 estimates goodness of prediction

Cross-Validation

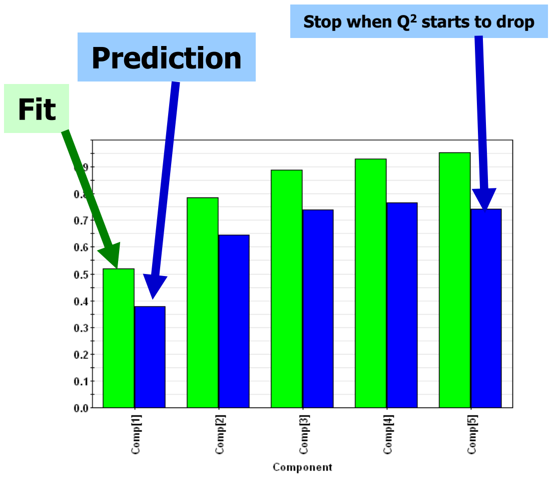
- Data are divided into G groups (default in SIMCA-P is 7) and a model is generated for the data devoid of one group
- The deleted group is predicted by the model ⇒ partial PRESS (Predictive Residual Sum of Squares)
- This is repeated G times and then all partial PRESS values are summed to form overall PRESS
- If a new component enhances the predictive power compared with the previous PRESS value then the new component is retained



- PCA cross-validation is done in two phases and several deletion rounds:
  - first removal of observations (rows)
  - then removal of variables (columns)

Model Diagnostics

- **Fit or R^2**
  - Residuals of matrix E pooled column-wise
  - Explained variation
  - For whole model or individual variables
  - RSS = Σ (observed - fitted)^2
  - R^2 = 1 - RSS / SSX
- **Predictive Ability or Q^2**
  - Leave out 1/7th data in turn
  - 'Cross Validation'
  - Predict each missing block of data in turn
  - Sum the results
  - PRESS = Σ (observed - predicted)^2
  - Q^2 = 1 − PRESS / SSX



Stop when Q^2 starts to drop

Prediction

Fit

Model diagnostics - Evaluation of R^2 and Q^2

- R^2 is always larger than Q^2
- High R^2 and Q^2 values are desirable
- The difference between R^2 and Q^2 should not be too large
- $Q^2 = 0.5$ - good model (typical for metabonomics)
- $Q^2 = 0.9$ – excellent model (typical for calibration)

Summary of Diagnostics

1. Data must be scaled appropriately

2. Outliers
   - Can upset a model
   - Investigate

3. How well does the model fit the data?
   - Study residuals
   - Look at R^2
   - Fit tells you little about predictive power

4. Predictive ability
   - Model information not noise – avoid overfit
   - Cross-validation helps determine number of components
   - Q^2 estimates predictive ability
   - True predictive ability known only from new data

PCA Summary

- PCA models the correlation structure of a dataset
- Data table X is approximated by a least squares (hyper)-plane + residuals (E)
- Large tables of data are distilled down to a few interpretable plots
- Observations are represented by scores
- Scores are linear combinations of the original variables with weights defined by the loadings
- Strong outliers are detected from hotellings T^2
- Moderate outliers are detected from DModX

Multivariate Analysis
for ”omics” data

Chapter 3 – PCA for overview of “omics” data
Finding groups, trends and outliers

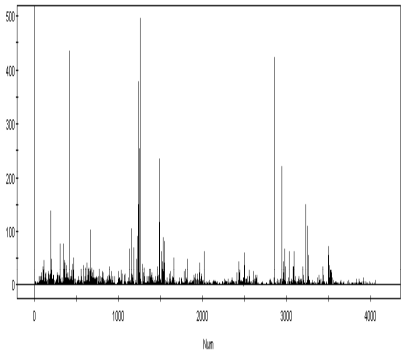
UMETRICS

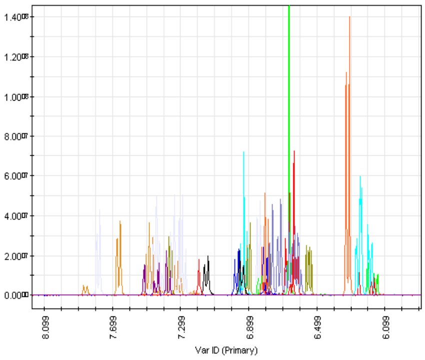
Outline

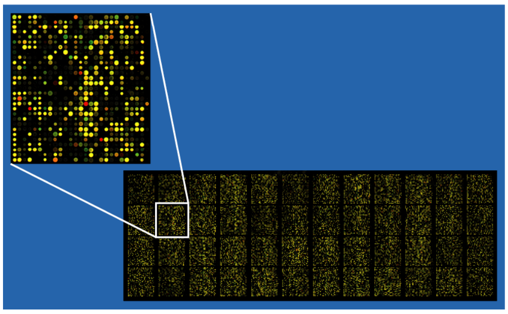
- How “omics” data is displayed
- PCA “omics” example

“Omics” data

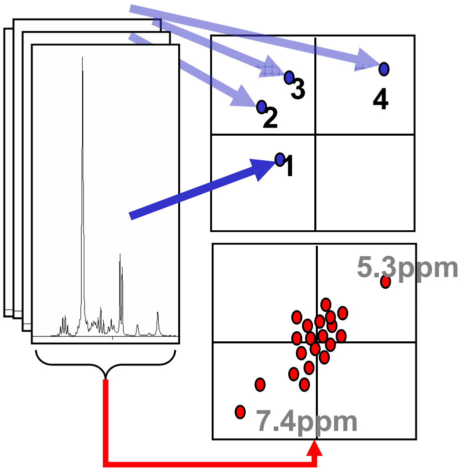
- LC-MS, GC-MS, UPLC-MS or NMR spectrum
- Microarray technology e.g. transcriptomics
- Want to compare spectra from different samples
- Look for groupings (Control vs. Treated)
- Find out which spectral features differ between treatment groups







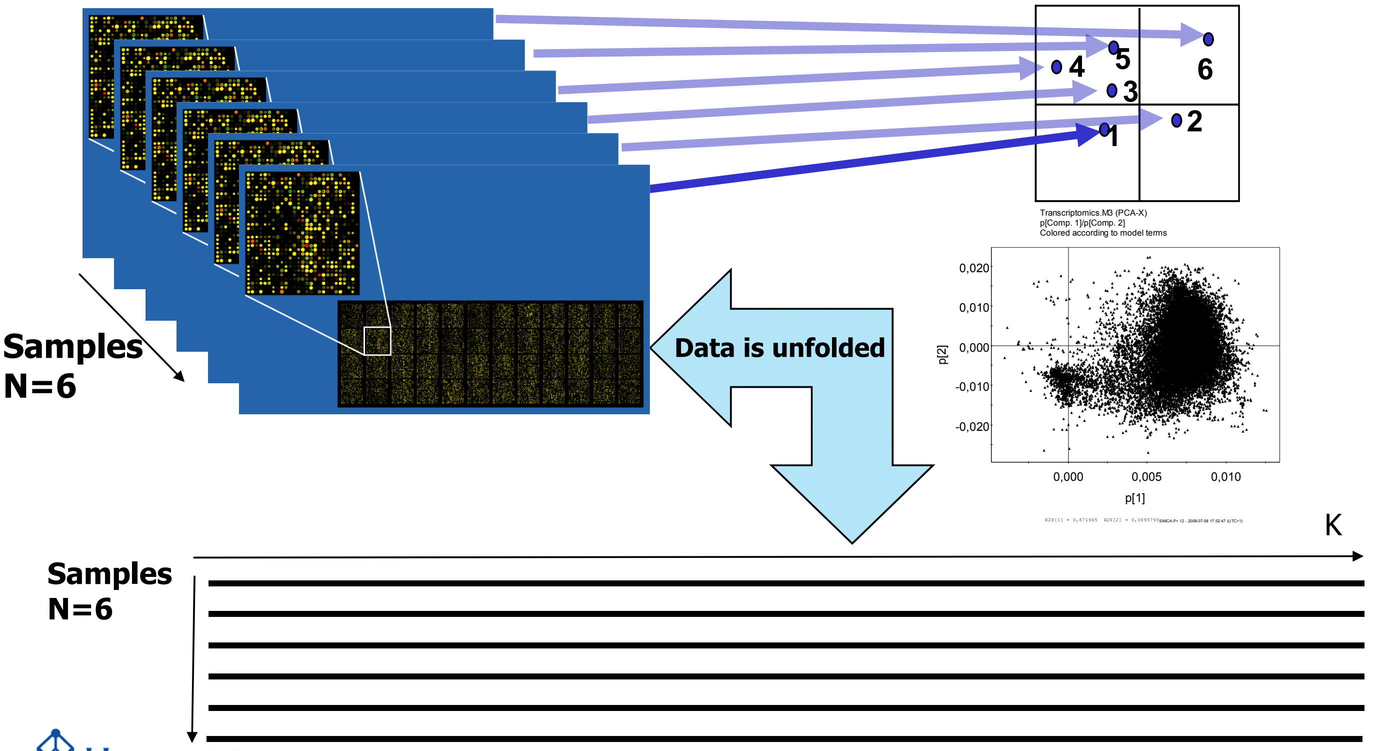
How Spectral Information is displayed



- Spectrum (observation) becomes a point in PCA **Scores plot**
- Variables (ppm or m/z) shown in PCA **Loadings Plot**

- Using plots together allows trends in the sample spectra to be ***interpreted*** in terms of chemical shift

How microarray information is displayed



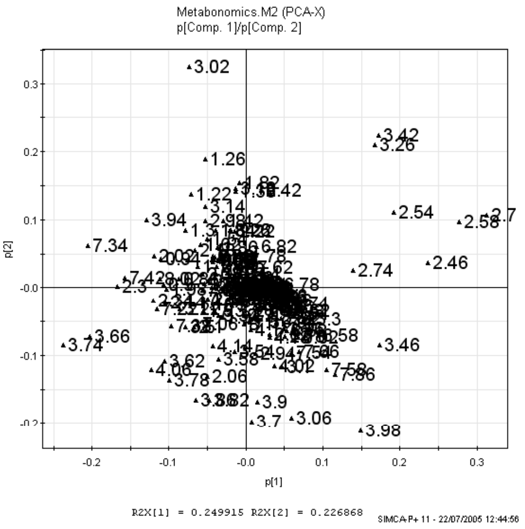
Samples
N=6

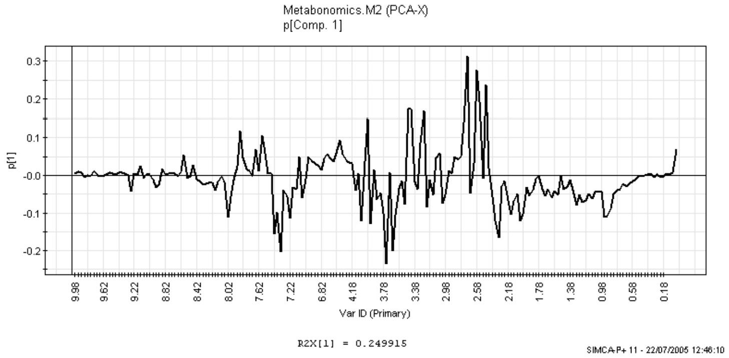
Data is unfolded

Samples
N=6

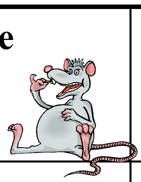
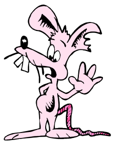
Loadings line plot for spectra

- When looking at spectra Loadings Line plot more informative than scatter plot
- More closely resembles a spectrum





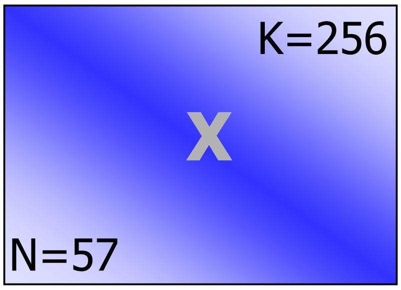
Example - Using PCA to examine trends

NMR: K = 194 variables	**Sprague Dawley** 	**Fisher**
Control 	S 10	F 10
Amiodarone **Renal toxicity**	SA 8	FA 10
Chloroquine **Hepatic toxicity**	SC 10	FC 9

Prior to PCA

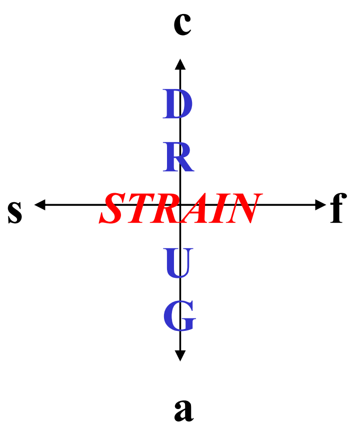
- NMR data pre-processes before import to SIMCA
- Data was centred and pareto scaled after import to SIMCA
- PCA analysis applied for overview and trends

NMR data collected for each sample



PCA to overview 1

- Two first components $R^2X = 0.48$ $Q^2X = 0.38$

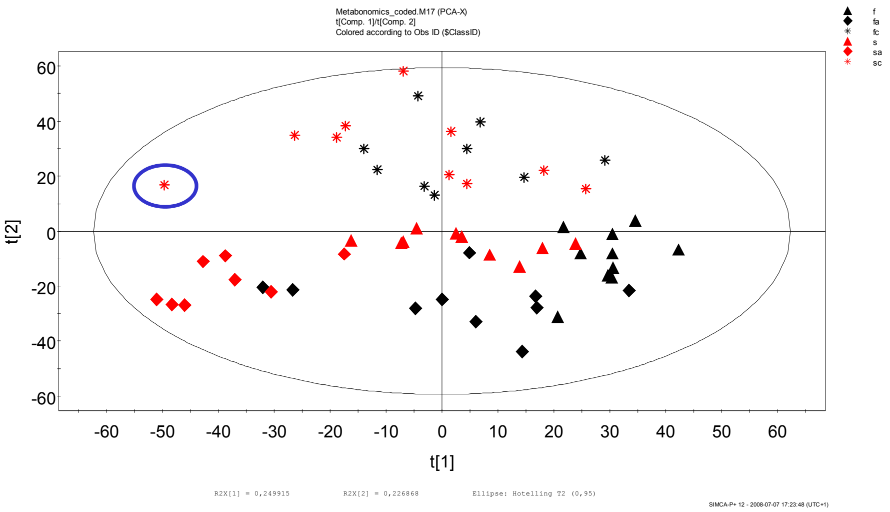


c

D
R
U
G

s STRAIN f

a

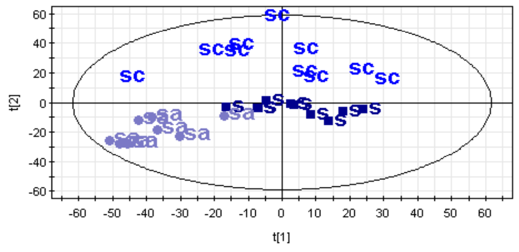


- One outlier, rat 27, encircled
  - Measurement error ?
  - Handling/environmental differences ?
  - Slow responder ?

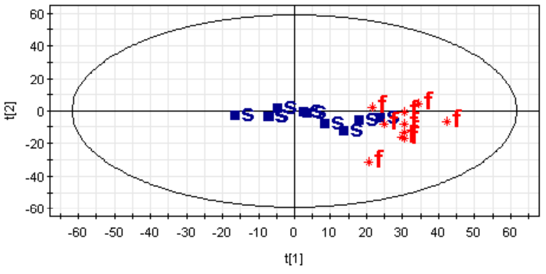
PCA for overview 1

- Model on all rats
  - –only some rats plotted
- View trends
  - Strain
  - Drug

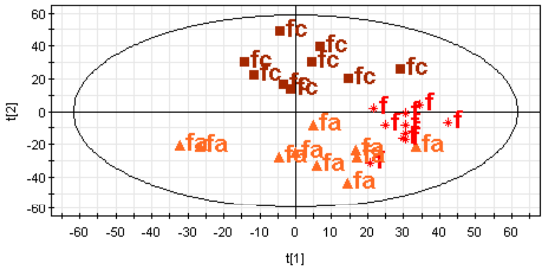
Effect of Drug on SD



Effect of Strain



Effect of Drug on F

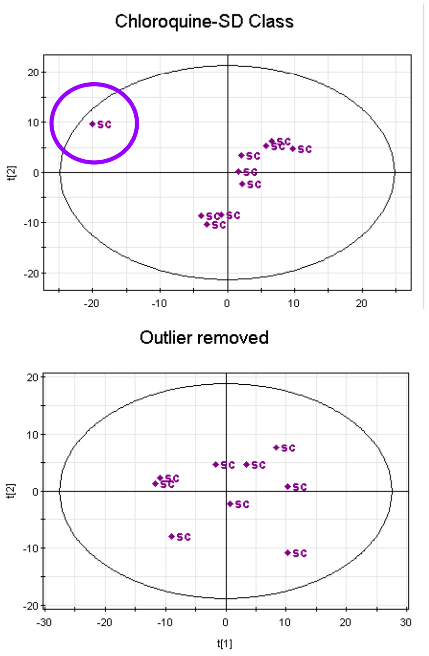


SIMCA-P+ 10.0 - 10/06/2003 16:13:47

SIMCA-P+ 10.0 - 10/06/2003 16:16:54

PCA for outlier detection

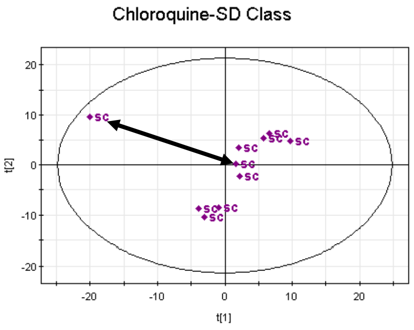
- Biggest variation in the data (first component) is caused by one sample
  - –The rest of all samples seems fairly tight
- Outliers may seriously disturb a model
  - – Incorrect values
  - –Technical problems
  - – Pre-processing error
  - – Transcription error / miss-labelling
  - – Good way to validate transcriptions
- Investigate!
- Class models cannot be built with outliers!
  - – need 'tight' classes

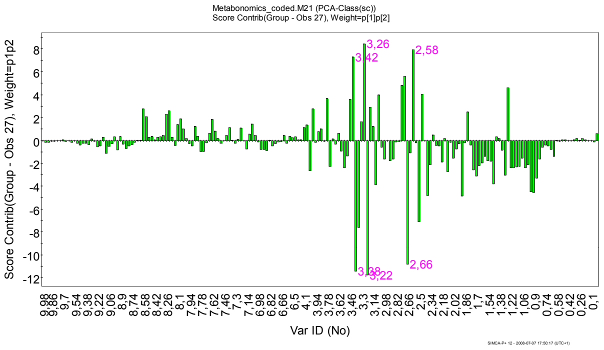


SIMCA-P+ 10.0 - 05/11/2003 10:55:53

PCA Contribution plot reveals differences

- How is SC rat 27 different from a "normal" SC-rat?
- Chemical shift regions 3.42, 3.26, 2.58, 3.38, 3.22 and 2.66

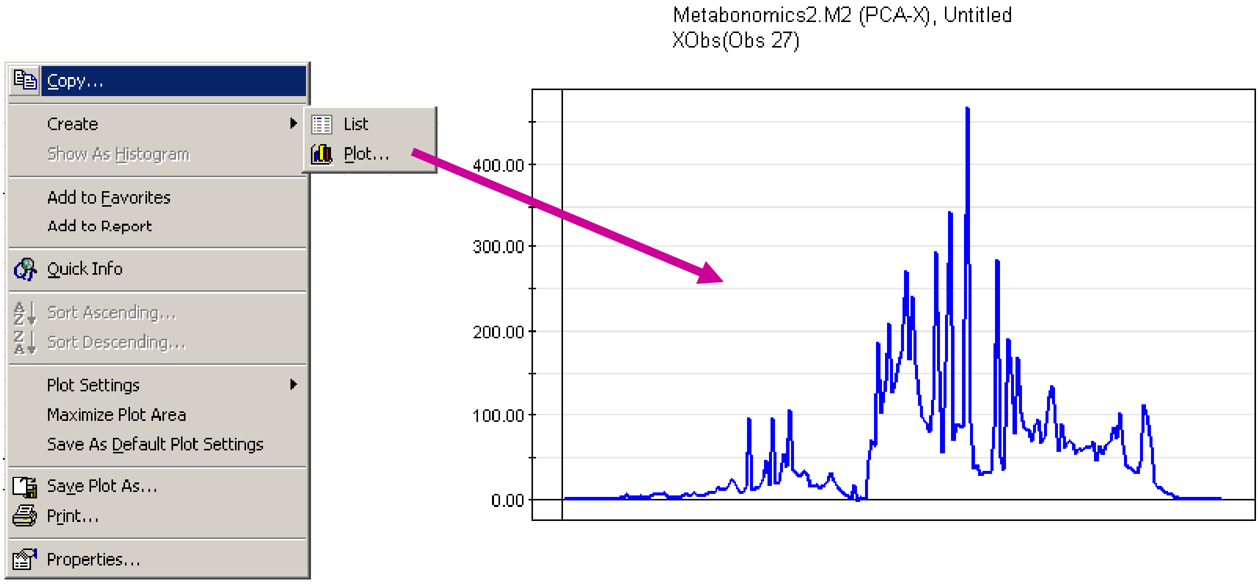




Double click on obs. of interest
The contribution plot from obs. to average will be displayed

Spectrum display - X Obs Plot

- Plot X obs direct from Scores Plot or List:

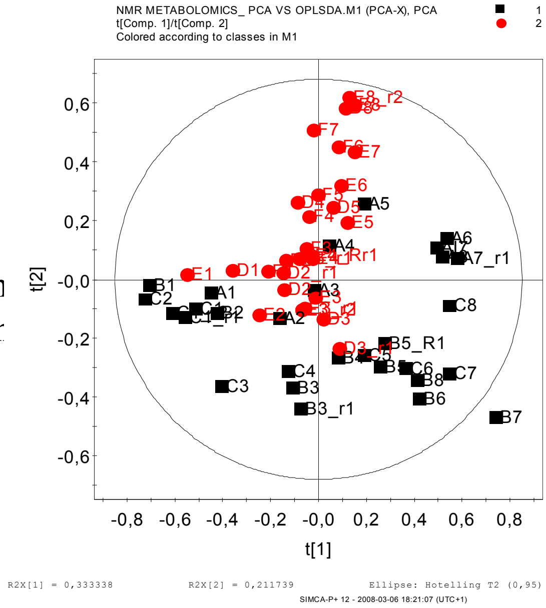


PCA for Overview 2

- Example: HR/MAS ^{1}H NMR study from poplar plants
  - Aim: biomarkers to explore biology
  - Ripening studies, source of origin etc

- Scores plot shows samples from two differen poplar (hybrid aspen) types; one wild type ar the other transgenic poplar

- Interpretation of scores shows patterns and trends

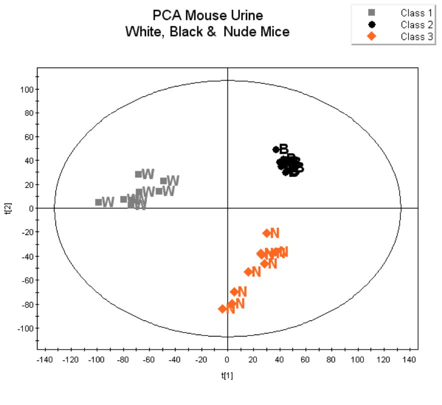


PCA for Overview 3

- Mouse Genetic Study
  - Black, White, Nude
  - Mass Lynx data

- PCA useful for QC of biolc results:
  - Biological diversity
  - Outlier detection
  - Finding trends

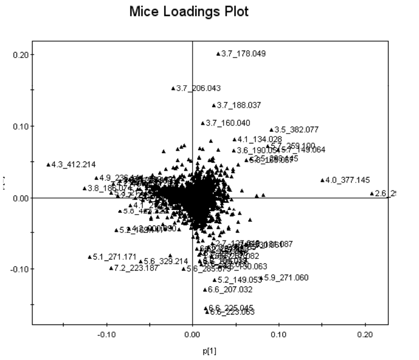
PCA Mouse Urine
White, Black & Nude Mice



Ellipse: Hotelling T2 (0.95)

SIMCA-P+ 10.5 - 03/09/2004 11:47:33

Mice Loadings Plot



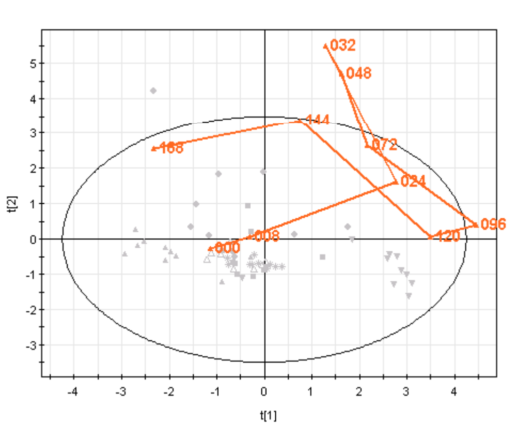
SIMCA-P+ 10.5 - 28/07/2004 10:35:54

Data courtesy of Ian Wilson and Waters Corporation Inc

PCA can examine time trends

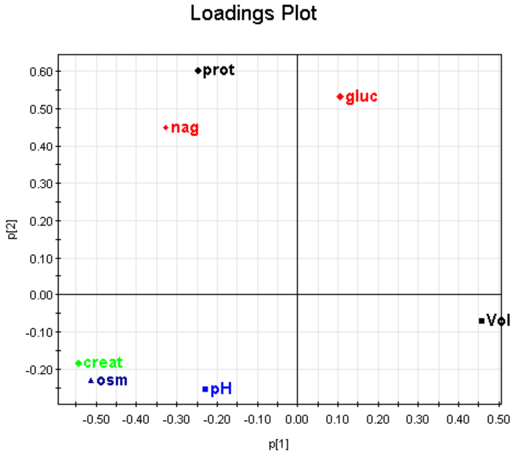
- Does animal recover?
- Examine trajectory in scores plot before during and after exposure
- Here we show clinical data rather than NMR spectra

Nephrotoxin - 7 day period



SIMCA-P+ 10.0 - 11/06/2003 15:09:35

Loadings Plot



SIMCA-P+ 10.0 - 11/06/2003 15:11:50

PCA Summary

- PCA is used to provide an overview of a data table to reveal:
  - dominating variables
  - trends
  - patters: outliers, groups, clusters
  - similarities / dissimilarities

- Classification: a new observation is considered similar to the training set if it falls within the tolerance volume of the model (DModX)
  - This type of PCA analysis is called SIMCA but is not included in this course

Multivariate Analysis for ”omics” data

Chapter 4
Data Processing

Contents

- Naming of observations
- Practical data processing
  - Pre-processing
  - Scaling and Normalisation
  - Special case for PCA

The importance of good names

- Keep it simple

- Easiest way
  - Separate name for each attribute
  - The **“Atomic Principle”** of good database practice
  - May combine on import to SIMCA or use multiple secondary ID’s

- Example
  - Animal Treatment Day
  - 12 Control 1

UMETRICS

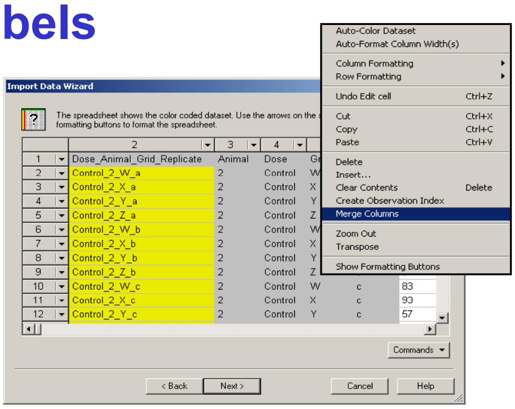
8/15/2008

Handling data labels

- Possibility to Merge columns
  - Create combined observation IDs by concatenation
  - Choice of order and separating character

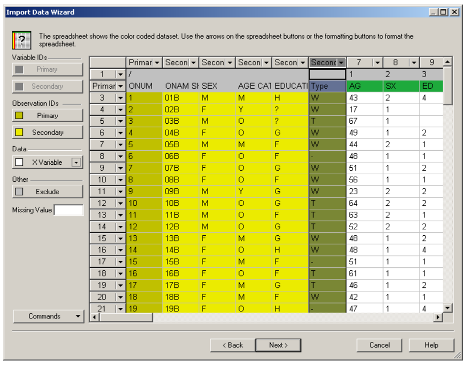
- Or use multiple secondary ID’s
  - May have unlimited secondary ID’s

- Keep names short
  - Easier to see in plots


Import Data Wizard

The spreadsheet shows the color coded dataset. Use the arrows on the s formatting buttons to format the spreadsheet.

	2	3	4			
1	Dose_Animal_Grid_Replicate	Animal	Dose	Gr		
2	Control_2_W_a	2	Control	W		
3	Control_2_X_a	2	Control	X		
4	Control_2_Y_a	2	Control	Y		
5	Control_2_Z_a	2	Control	Z		
6	Control_2_W_b	2	Control	W		
7	Control_2_X_b	2	Control	X		
8	Control_2_Y_b	2	Control	Y		
9	Control_2_Z_b	2	Control	Z		
10	Control_2_W_c	2	Control	W	c	83
11	Control_2_X_c	2	Control	X	c	93
12	Control_2_Y_c	2	Control	Y	c	57


Import Data Wizard

The spreadsheet shows the color coded dataset. Use the arrows on the spreadsheet buttons or the formatting buttons to format the spreadsheet.

	Primar	Secon	Secon	Secon	Secon	Second	7	8	9
1	/						1	2	3
Primar	ONUM	ONAM SI	SEX	AGE CA1	EDUCATI	Type	AG	SX	ED
3	1	01B	M	M	H	W	43	2	4
4	2	02B	F	Y	?	W	17	1	
5	3	03B	M	O	?	T	67	1	
6	4	04B	F	O	G	W	49	1	2
7	5	05B	M	M	F	W	44	2	1
8	6	06B	F	O	F	-	48	1	1
9	7	07B	F	O	G	W	51	1	2
10	8	08B	F	O	F	W	56	1	1
11	9	09B	M	Y	G	W	23	2	2
12	10	10B	M	O	G	T	64	2	2
13	11	11B	M	O	F	T	63	2	1
14	12	12B	M	O	G	T	52	2	2
15	13	13B	F	M	G	W	48	1	2
16	14	14B	F	O	H	W	48	1	4
17	15	15B	F	M	F	-	51	1	1
18	16	16B	F	O	F	T	61	1	1
19	17	17B	F	M	G	T	46	1	2
20	18	18B	F	M	F	W	42	1	1
21	19	19B	F	O	H	-	47	1	4

UMETRICS

8/15/2008

Important considerations

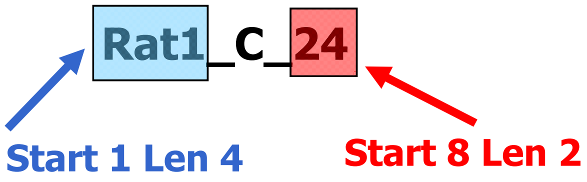
- Length issues

  Numbers have preceding 0’s
  - i.e. 01, 02   NOT 1 , 2

  Treatments have same alphanumeric length
  - Control = C or Con   High Dose = H or Hds

  Time is in consistent units
  - Hours or Days

- Why?
  - Selective masking on plots
  - Start Character, Length



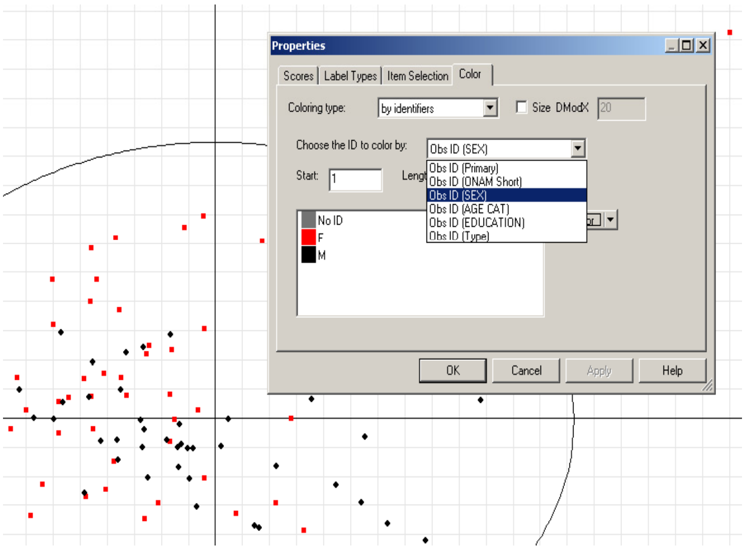
Rat1_C_24

Start 1 Len 4

Start 8 Len 2

Secondary ID’s

- Ability to colour plots by secondary ID’s



Properties

Coloring type: by identifiers

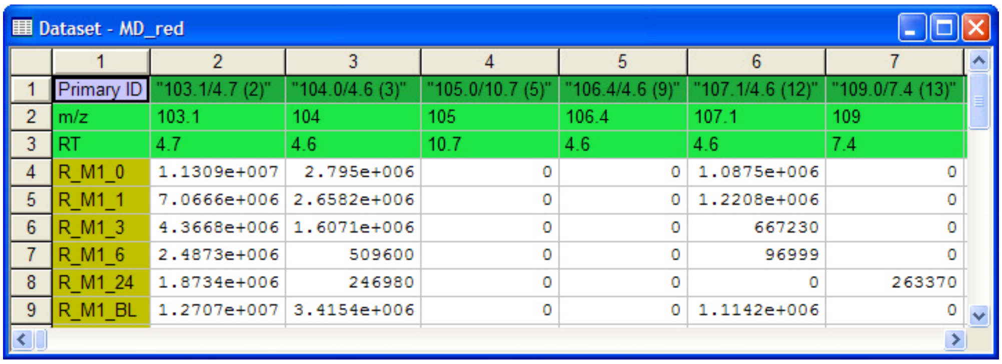
☐ Size DModX 20

Choose the ID to color by: Obs ID (SEX)

Start: 1

Using Secondary Variable IDs

- Also possible to have secondary variable ID’s
  - Here we see MS variables split into Time and Mass



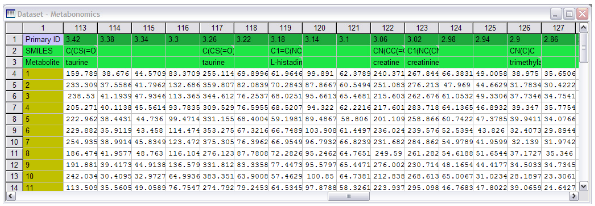
Dataset - MD_red

	1	2	3	4	5	6	7
1	Primary ID	"103.1/4.7 (2)"	"104.0/4.6 (3)"	"105.0/10.7 (5)"	"106.4/4.6 (9)"	"107.1/4.6 (12)"	"109.0/7.4 (13)"
2	m/z	103.1	104	105	106.4	107.1	109
3	RT	4.7	4.6	10.7	4.6	4.6	7.4
4	R_M1_0	1.1309e+007	2.795e+006	0	0	1.0875e+006	0
5	R_M1_1	7.0666e+006	2.6582e+006	0	0	1.2208e+006	0
6	R_M1_3	4.3668e+006	1.6071e+006	0	0	667230	0
7	R_M1_6	2.4873e+006	509600	0	0	96999	0
8	R_M1_24	1.8734e+006	246980	0	0	0	263370
9	R_M1_BL	1.2707e+007	3.4154e+006	0	0	1.1142e+006	0

Metabolite Assignment via Secondary ID’s

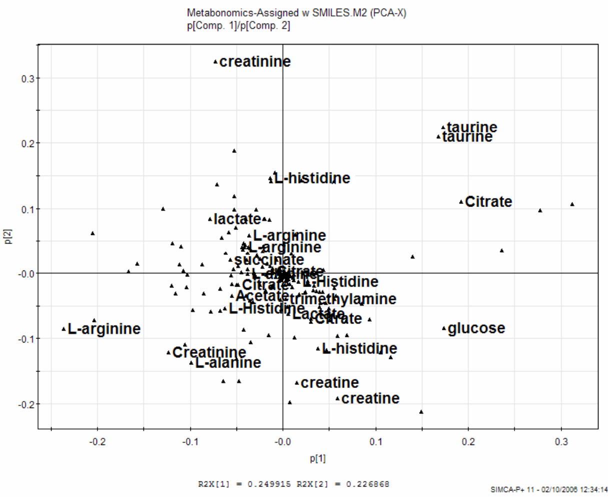
- In SIMCA-P+ 12 you may add assignments to peaks using a secondary ID

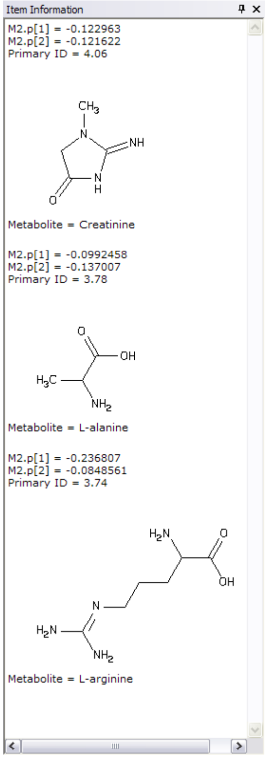
- With SMILES plugin it is also possible to show chemical structures



Dataset - Metabonomics

	1	113	114	115	116	117	118	119	120	121	122	123	124	125	126	127
1	Primary ID	3.42	3.38	3.34	3.3	3.26	3.22	3.18	3.14	3.1	3.06	3.02	2.98	2.94	2.9	2.86
2	SMILES	C(CS(=O)				C(CS(=O)		C1=C(NC			CN(CC(=	C1(NC(CN			CN(C)C	
3	Metabolite	taurine				taurine		L-histadin			creatine	creatinine			trimethyla	
4	1	159.789	38.676	44.5709	83.3709	255.114	69.8996	61.9646	99.891	62.3789	240.371	267.844	66.3831	49.0058	38.975	35.6506
5	2	233.309	37.5586	41.7962	132.686	359.807	82.0839	70.2843	87.8667	60.5494	251.083	276.213	47.969	44.6629	31.7834	30.4222
6	3	238.53	41.1939	47.9346	113.365	344.612	76.2537	68.0251	95.6613	65.4681	215.603	262.676	61.0532	49.3306	37.7346	34.7541
7	4	205.271	40.1138	45.5614	93.7835	309.529	76.5955	68.5207	94.322	62.2216	217.601	283.718	64.1365	46.8932	39.347	35.7754
8	5	222.962	38.4431	44.736	99.4714	331.155	68.4004	59.1981	89.4867	58.806	201.109	258.866	60.7422	47.3785	39.9411	34.0766
9	6	229.882	35.9119	43.458	114.474	353.275	67.3216	66.7489	103.908	61.4497	236.024	239.576	52.5394	43.826	32.4073	29.8944
10	7	254.935	38.9914	45.8349	123.472	375.305	76.3962	66.9549	96.7932	66.8239	231.682	284.862	54.9789	41.9599	32.139	31.9742
11	8	186.474	41.9577	48.763	116.104	276.123	87.7808	72.2826	95.2462	64.7651	249.59	261.282	54.6188	51.6544	37.1727	35.346
12	9	191.881	39.4173	44.9138	136.579	331.812	83.3358	77.4473	95.5797	65.4471	276.002	230.714	48.1654	44.4177	34.5033	34.7345
13	10	242.034	30.4095	32.9727	64.9936	383.351	63.9008	57.4629	100.85	64.7381	212.838	268.613	65.0067	31.0234	28.1897	23.3061
14	11	113.509	35.5605	49.0589	76.7547	274.792	79.2453	64.5345	97.8788	58.3261	223.937	295.098	46.7683	47.8022	39.0659	24.6427





Item Information

M2.p[1] = -0.122963
M2.p[2] = -0.121622
Primary ID = 4.06

Metabolite = Creatinine

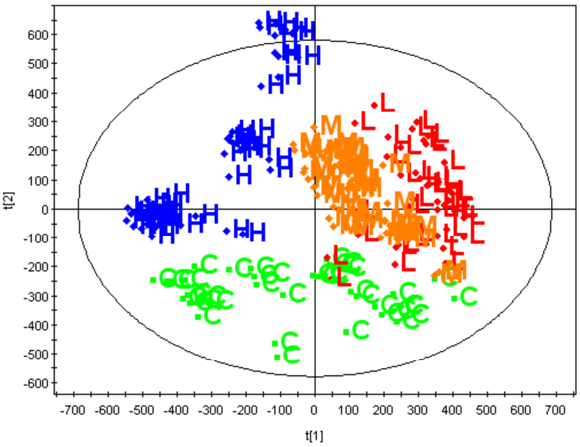
M2.p[1] = -0.0992458
M2.p[2] = -0.137007
Primary ID = 3.78

Metabolite = L-alanine

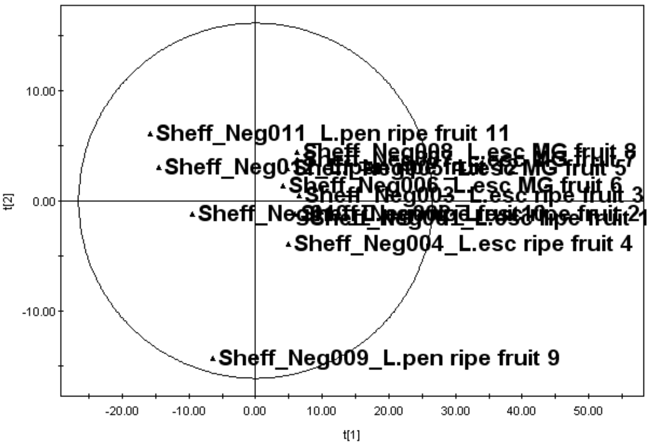
M2.p[1] = -0.236807
M2.p[2] = -0.0848561
Primary ID = 3.74

Metabolite = L-arginine

The benefits of good naming!



VS.



UMETRICS

Practical Data Processing

Pre-processing, Scaling and Normalisation

UMETRICS

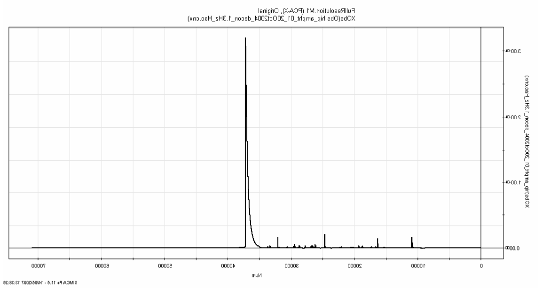
Quality in = Quality out

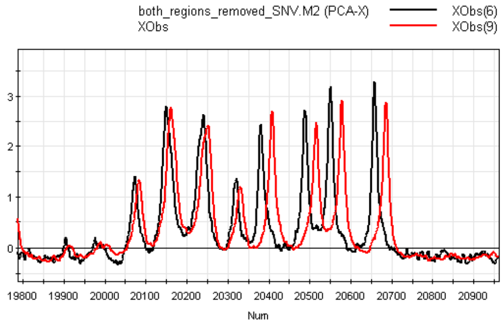
- Quality of analysis depends on quality of spectra

- Pre-processing required
  - –Type of pre-processing is depending on the **type of data**

- Typical problems in spectroscopic data
  - *–Water peak (NMR)*
  - *– Baseline problems*
  - *–Alignment of peaks shifts*
    - Chromatography problems
    - pH sensitive peaks (NMR)
    - Salt sensitive peaks (NMR)
  - *–Variation in concentration (normalisation)*
  - –Correct assignment of standard ????
  - –Phasing / Shimming (NMR data)
  - – Temperature effects
  - –Artefacts
    - column bleeding
    - ghost peaks
  - –Noise

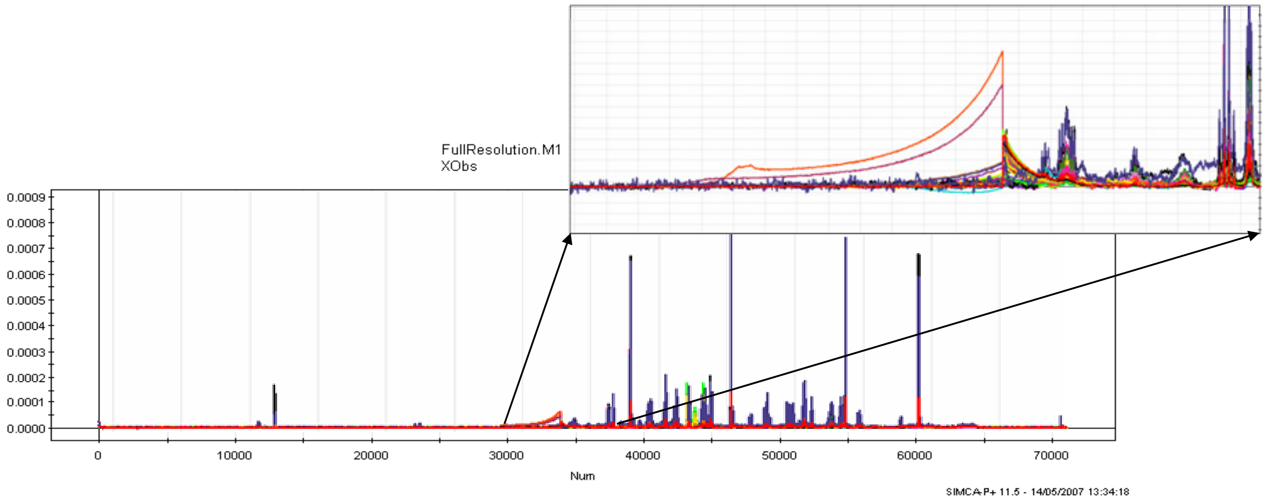
- High quality data required!





Problem No 1 – Water peak

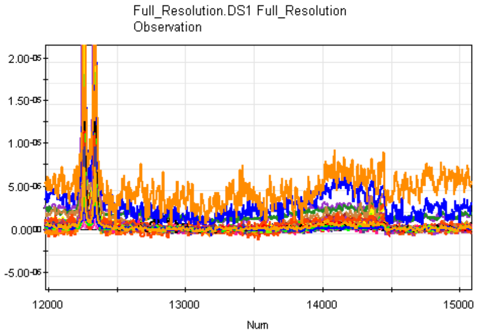
- If water peak incorrectly suppressed/removed then normalisation will ruin the data completely!
- Other known artefacts should also be removed



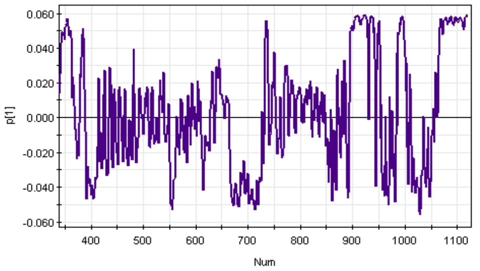
Problem No 2 – Baseline shifts

- Much more care needed to align baselines for metabonomic studies compared with routine NMR for structural identification

- Diagnostics: Baseline
  - Find quiet part of spectrum (i.e. 10ppm)
  - UV Scale
  - Examine p1 Loadings plot
  - Non zero loadings indicates problem

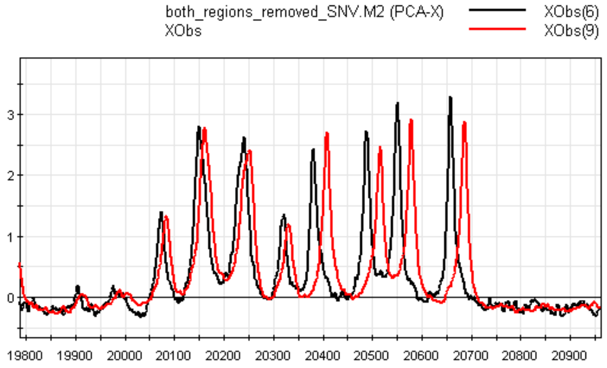


UV Loadings

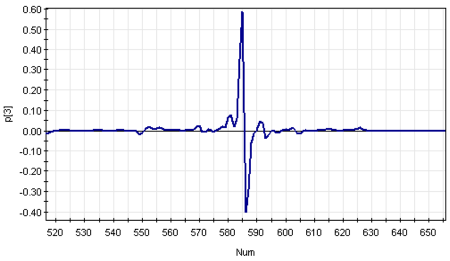


Problem No 3 –Peak shift (or Alignment)

- For NMR data
  - Variation in pH
  - Metal ions
- Very difficult to deal with afterwards
  - Careful buffering of samples
  - Consistent sample preparation
- For chromatography data
  - Variation in retention time between samples
- Diagnostics for Peak shifts
  - Examine loadings line plots
  - Look for sawtooth effect
  - indicative of pH shifts



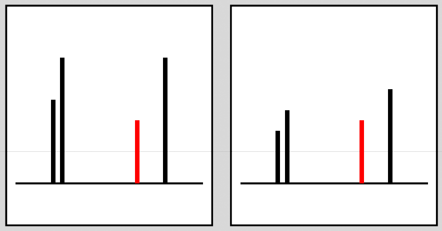
Loadings



Normalisation

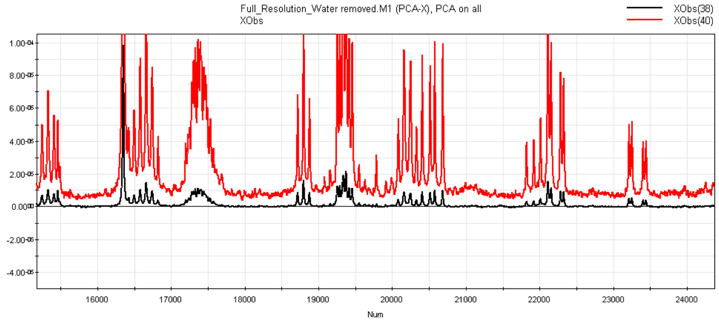
- Strength of ‘spectra’ is different across samples
  - i.e. urine varies in concentration
- Need to remove sample-sample variability
- Could ideally be solved by the addition of an internal standard.
  - Often difficult with metabonomic Urine samples
  - Impossible in MS unless using isotopic labelling
- Normalisation approaches are:
  - To an internal std
  - To peaks always present in approximately same concentration
  - To baseline ‘noise’
  - To total signal
  - To magnitude of PCA score 1 (“eigenvalue scaling” - Sqrt(t))
  - Probabilistic Quotient Normalisation

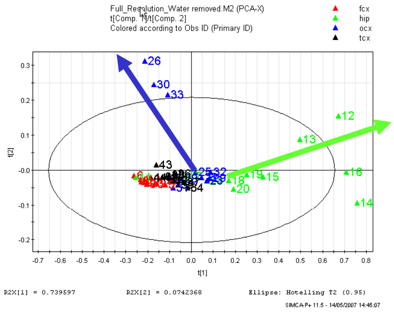
Internal Standard

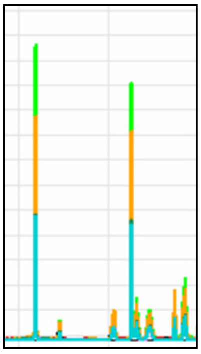


Further problems to watch for

- Normalisation problems
  - Differences in concentration between samples
  - Large amplitude spectra with enhanced noise
- Linear trends in scores plot
  - Check baseline
  - Check normalisation





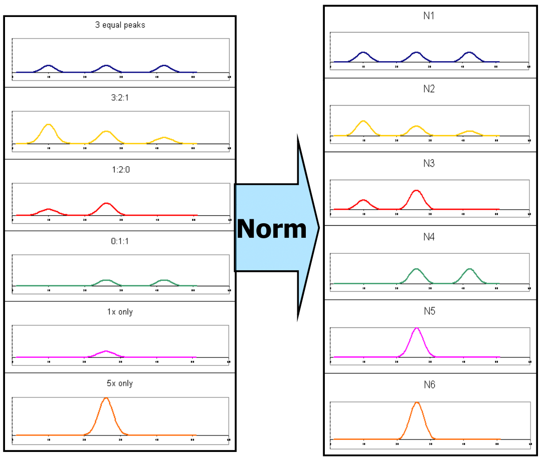


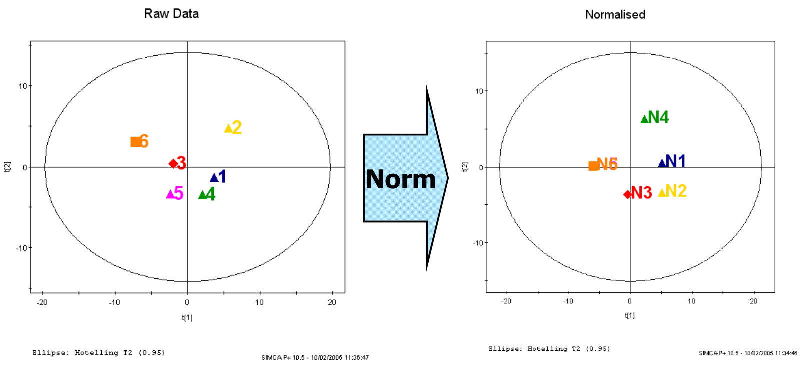
Normalisation methods

- Integral Normalisation
  - Divide each element by the sum of the spectrum
  - Often multiplied by 100
- Vector Length Normalisation
  - Divide each element of spectrum by its length
  - Length = Sqrt (x1^2 + x2^2 + xn^2)
- Probabilistic Quotient Normalisation (chapter 6)
  - Finds the most common scale factor between spectra
  - Divide each spectrum by this scale factor

The problem of 'Closure'

- BEWARE: Normalisation can introduce problems
  - "Constant Sum Problem" or "Closure"
  - Variables become correlated as everything adds up to 100%
  - If prominent peak absent then other peaks increase in apparent importance
  - May reverse the direction of trends! = disaster





Pre-processing software

- Many different software exist
- Some are mentioned in chapter 6

Scaling of "omics" data

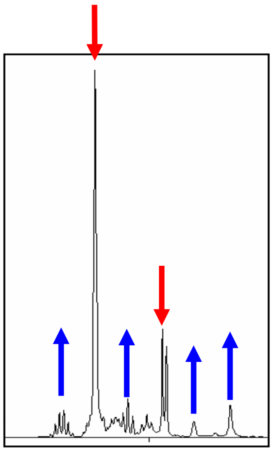
- Many choices
  - Centering (Ctr)
  - Unit variance (UV)
  - Unit variance none (UVN, no centering)
  - Pareto (Par)
  - Pareto none (ParN, no centering)
  - None
- For metabolomics data Par has shown to be a good alternative
  - A golden intermediate between UV and ctr
  - Today the S-plot only works for centered or pareto scaled data

Pareto Scaling (Par)

- What happens if big features dominate but we know medium features are also important?
  - CTR (mean centre only) Medium features overwhelmed by big
  - UV (mean centre and autoscale) Blow up baseline noise

- Answer is Pareto scaling
  - Divide each variable by the square root of its SD
  - Intermediate between no scaling (Ctr) and UV
  - Up weights medium features without inflating baseline noise.

  - Generally the preferred option
    - NMR & MS metabonomics
    - Gene chip & proteomics data

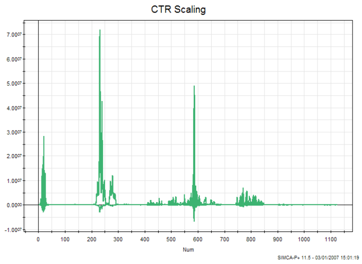


$$UV = \frac{x-\bar{x}}{SD}$$

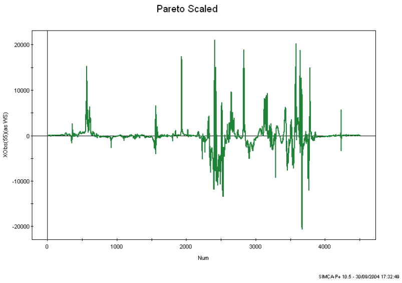
$$Par = \frac{x-\bar{x}}{\sqrt{SD}}$$

Effect of Scaling

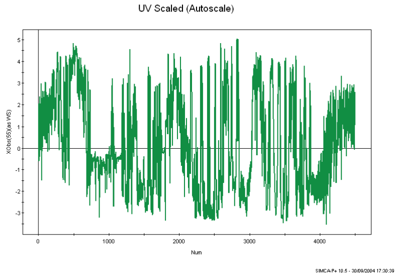
- NMR Spectrum Ctr, Par, UV



CTR Scaling

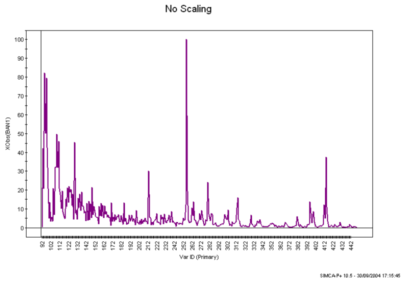


Pareto Scaled

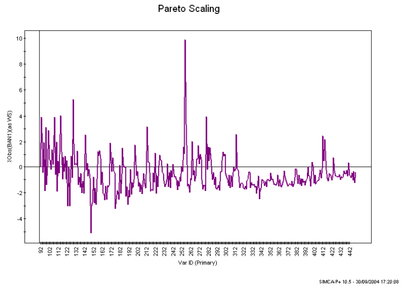


UV Scaled (Autoscale)

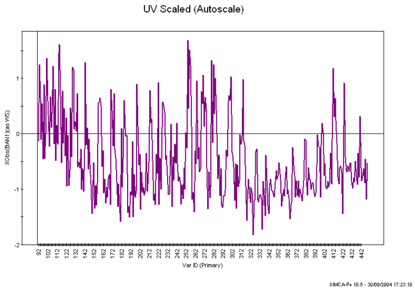
- Mass Spectrum Ctr, Par, UV



No Scaling



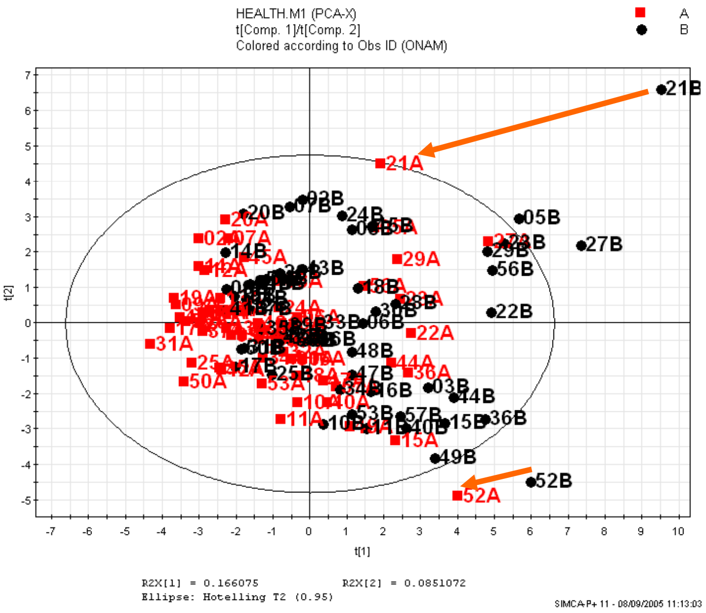
Pareto Scaling



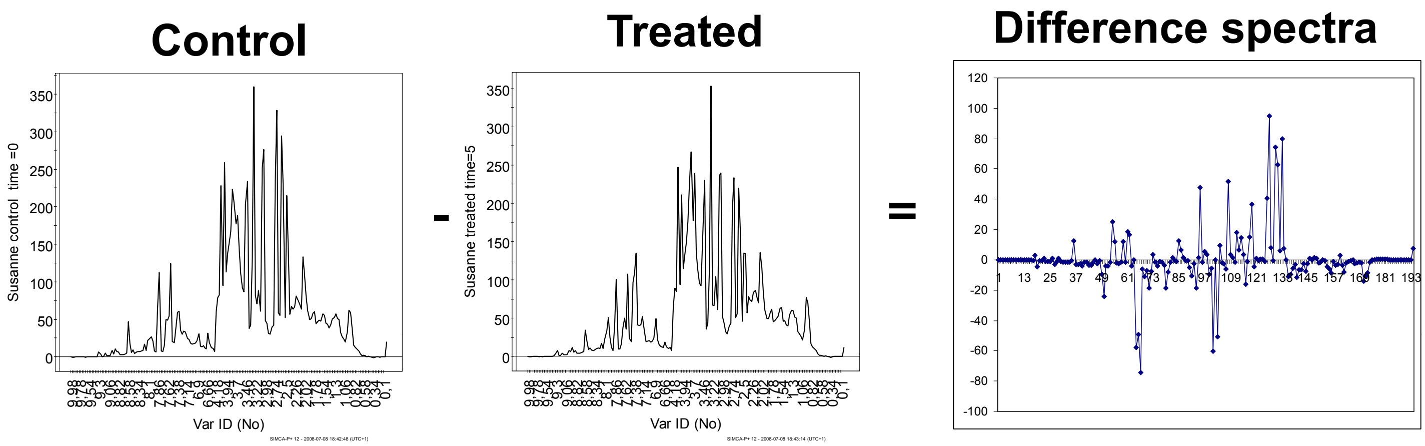
UV Scaled (Autoscale)

Special Scaling: Clinical metabonomic data

- Data consists of peak tables of metabolites from HPLC/GC or bacteria counts etc.
- Often Before and After treatment data is available (paired controls)
- Often better to look at the table of *differences*
- Example HEALTH
  - Patients subjected to a physiotherapy treatment



Difference spectra



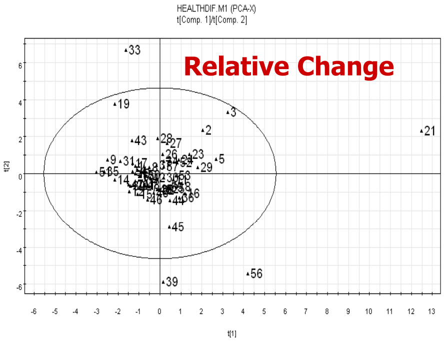
Control

Treated

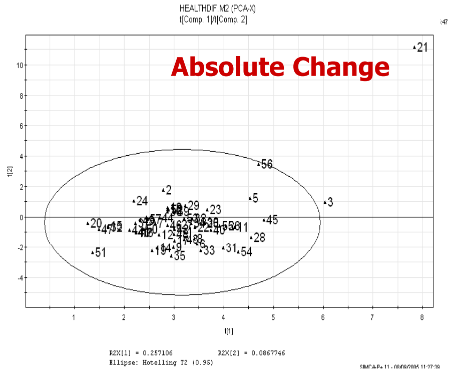
Difference spectra

No centering but Scaling

- Sometimes mean-centering will remove the interesting effect
- Here we need to use UVN, no centering but scaling
- Scores plot is no longer centered, the patients that have changed the most move further along t1
- If there were no treatment effect then subjects would cluster around the origin



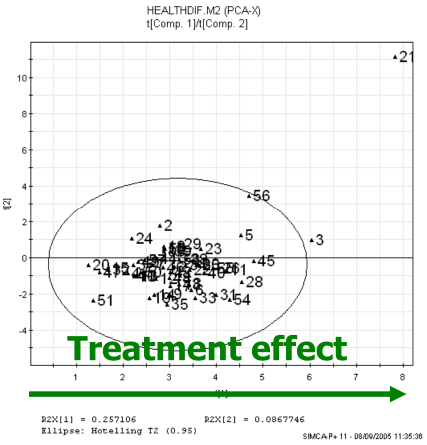
R2=0.19
Q2=-.13

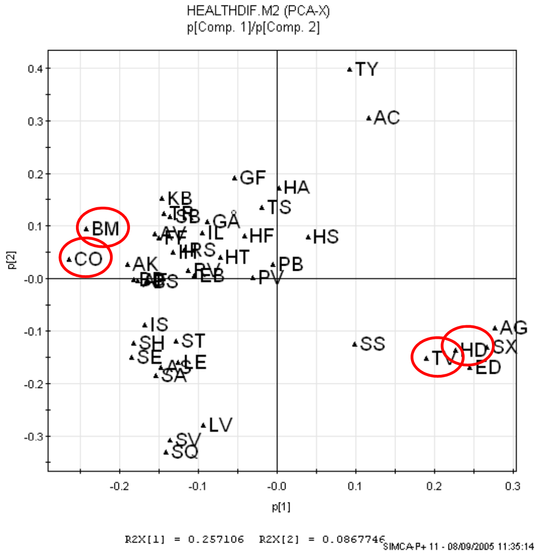


R2=0.34
Q2=0.13

HEALTH Example

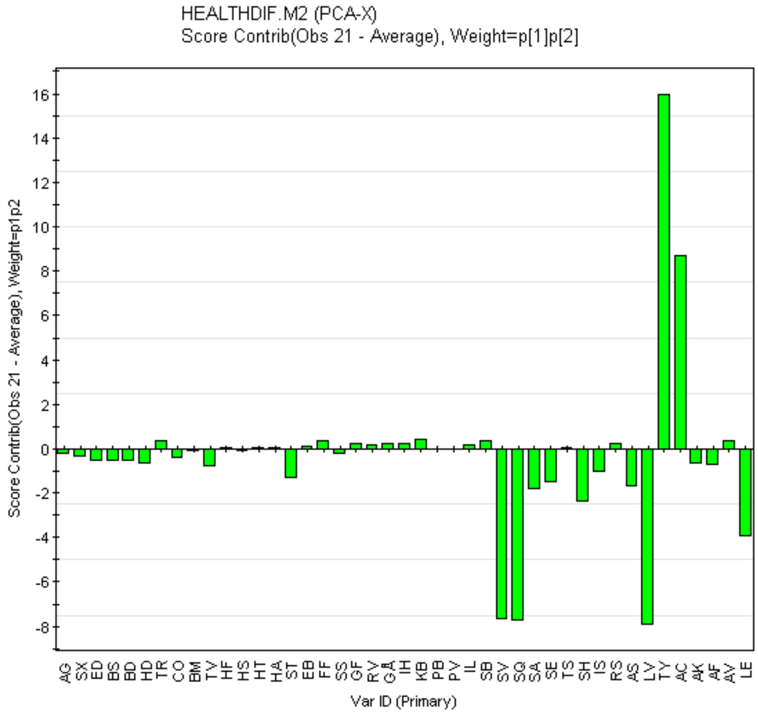
- Loading plot shows
  - Reduced cholesterol (CO) and body-mass-index (BM)
  - Increased physical fitness (TV) and HDL blood lipids (HD)





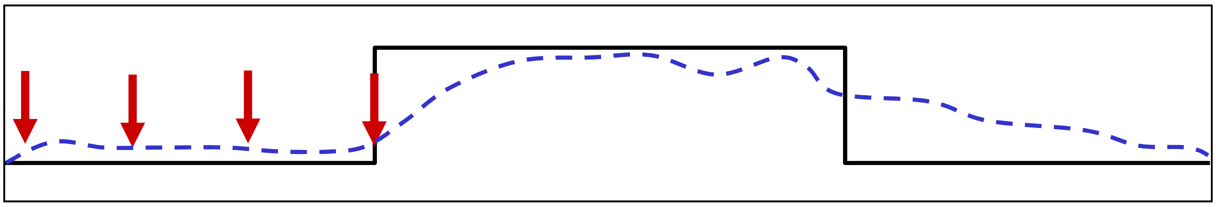
HEALTH Example: Subject 21

- Contribution plot of subject 21
- TY represents difficulty in breathing (subject breaths more easily after treatment)



Normalisation by Control subtraction

- Some time based studies, especially human, involve a control and treatment period on the same individual



- To reduce individual variability sometimes it is possible to subtract the averaged Control period of each individual so that in effect each individual becomes their own control.

Conclusions

- Collect information about data and use it as secondary ID
- Choose a good naming scheme at the outset
- Quality of analysis depends on quality of spectra
- How to pre-process the data depends on the type of data
- Beware of artefacts due to data pre-processing
- Scaling is important
- Beware of complications

Multivariate workflows for “omics” analysis

Chapter 5
Classification by OPLS-DA

Outline

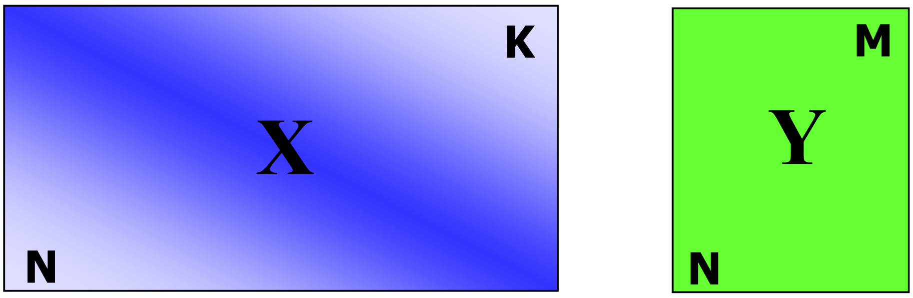
- Notations and Abbreviations
- Short history
- From PCA to OPLS
- Why OPLS-DA?
- OPLS-DA
  - The Method
  - Diagnostics
- Example 1 OPLS in classification

OPLS Notation

- N Observations
  - Humans
  - Rats
  - Plants
  - Analytical replicates
  - Trials (experimental runs)

- K X-Variables
  - NMR
  - GC/MS
  - LC/MS
  - UPLC/MS
  - etc

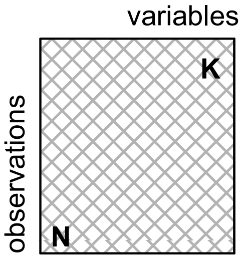
- M Y-Variables
  - Class information
  - Treatments
  - Time
  - Dose



In this course we will only work with M=1

Model notations

- N = number of observations
- K = number of X-variables
- M = number of Y-variables (here M=1)
- A = number of components



• t1	Predictive X-scores
• to1, to2,..., to(A-1)	Orthogonal X-scores
• u1	Y-scores

• p1	Predictive X-loadings
• po1, po2,..., po(A-1)	Orthogonal X-loadigns

History

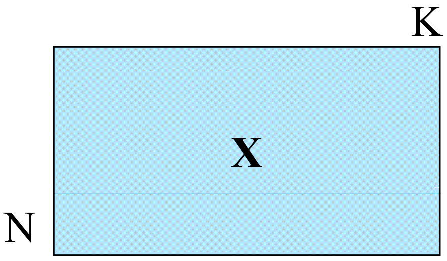
- In early 1920 Herman Wold developed the NIPALS algorithm that is used in partial least squares to latent structures, PLS
- The NIPALS algorithm was simplified by Svante Wold (Hermans son) in early 1980
- Statistical diagnostic tools and improved strategies for interpretation have been developed by co-workers ever since
- The first commercial available software of SIMCA year 1987 by Umetrics
- PLS is a well established regression and prediction method
  - Useful in most multivariate regression problems including correlated variables
  - E.g. multivariate calibration
  - Classification of wood species using NIR
  - QSAR-quantitative structure activity relationship
  - Many more examples
- OPLS is an extension of PLS which has proved to be very useful when good interpretation is important
  - “omics” data
  - NIR data

Wold, H., Estimation of principal components and related models by iterative least squares, *Multivariate Analysis* (Ed., Krishnaiah, P. R.), Academic Press, NY, pp. 391-420 (1966).

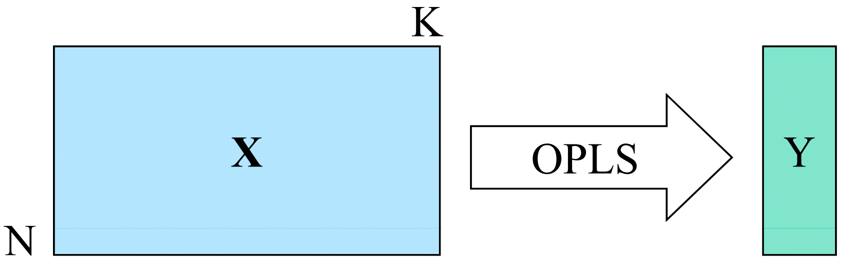
Orthogonal Partial Least Squares - OPLS ®

- The OPLS is a modification of the conventional NIPALS PLS algorithm
- OPLS was developed by Trygg and Wold, 2002
- Johan Trygg got the Elsevier Chemometrics Award 2008
- OPLS is a new way to decompose the PLS solution into
  (a) components correlated (predictive) to Y and
  (b) components unique in X but uncorrelated (orthogonal) to Y
- OPLS® Registered Trade marked and patented since august 2001
- O2PLS® Registered Trade marked and patented

From PCA to OPLS



Unsupervised
PCA on **X** will find the maximal variation in the data. PCA is the basis of all multivariate modelling.



Supervised
OPLS is a prediction and regression method that finds information in the **X** data that is related to known information, the **Y** data.

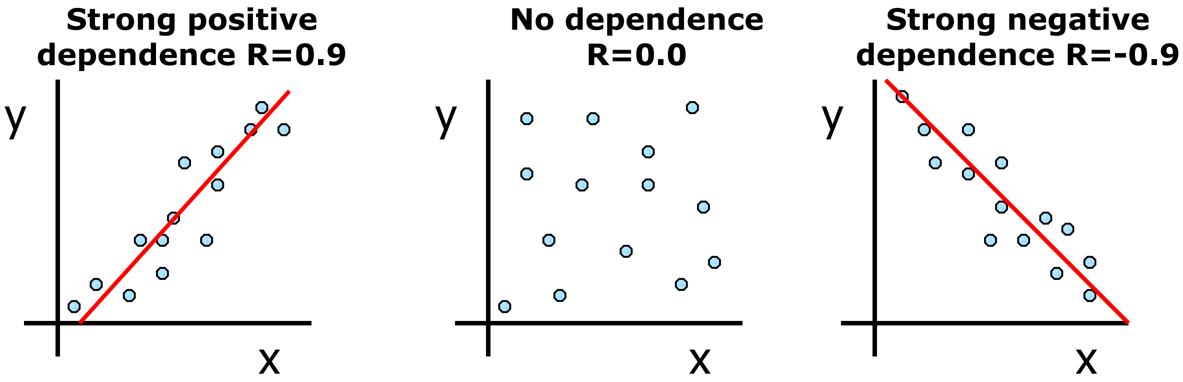
What is OPLS?

- OPLS is a **regression and prediction method**
  - Regression- how do things vary together?
  - Prediction-how well the **known** information is predicted
- Regression relates one or more X-variables to one or more Y variables

Method	X-Variables	Y-Variables
Linear Regression	1	1
Multiple Linear Regression	<N	1
Orthogonal Partial Least Squares	Many	Many

Dependence between Variables

- **Correlation** and **Covariance** are measures of
  - How things vary *TOGETHER*
  - Either positive (both variables increase together)
  - Or negative (one increases while the other decreases)
  - Correlation coefficient summarises dependence of two variables and lies in the range –1 to +1



Strong positive dependence R=0.9

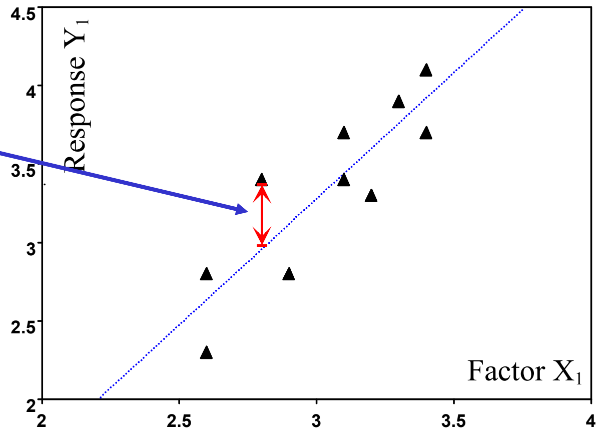
No dependence R=0.0

Strong negative dependence R=-0.9

Linear Regression

- Linear relationship between a variable X_1 and a response Y_1
- The deviation between the actual and the fitted value is known as the *residual*
- Least squares analysis minimizes the sum of squares of the residuals
- Goodness of fit: $R^2 = 1 - SS_{res}/SS_{tot.corr}$
  - 1 denotes perfect model
  - 0 corresponds to no model at all
  - 0.75 indicates decent model

$Y_1 = -1.54 + 1.61X_1 + e$; $R^2 = 0.75$



$$y = mx + c + e$$

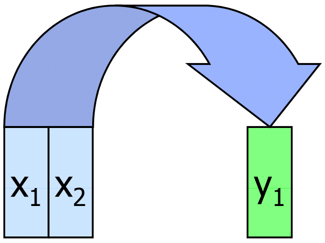
Coefficient

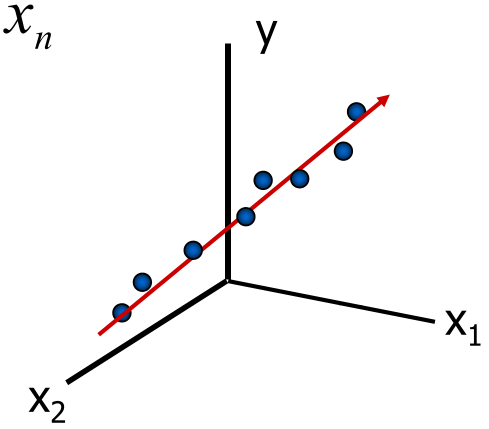
Multiple Linear Regression

- Linear relationship between many X-variables and a single response Y_1

$$y = c + b_1 x_1 + b_2 x_2 + b_3 x_3 \ldots . b_n x_n$$

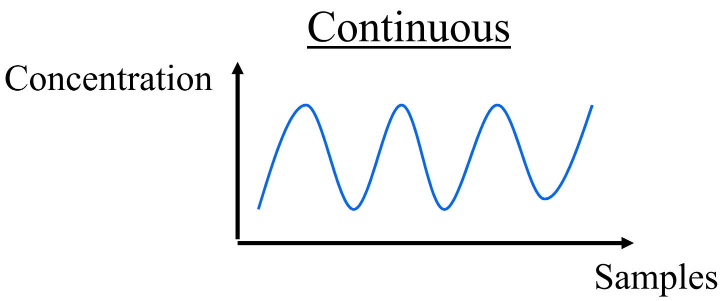
- Suitable for a few X-variables
  - X-variables should be independent
  - Must be more observations than variables

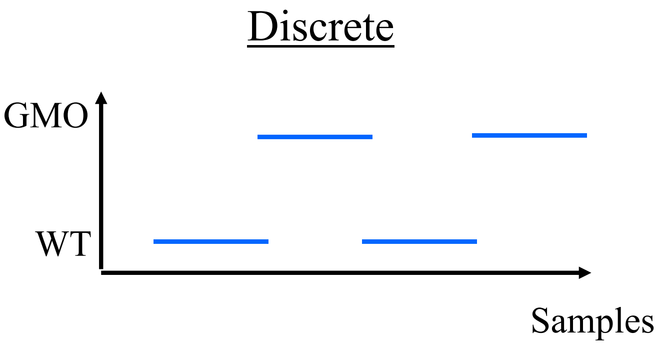




What is a Y variable?

- Y variable contains information about the sample (extract, tissue, urin etc)
  - Measured information
  - Known information
- Can be continuous
  - Time
  - Concentration
- Can be discrete
  - Wild type (WT) or Genetically modified (GMO)
  - Male or female
  - Control or treated





Discrete/categorical variables

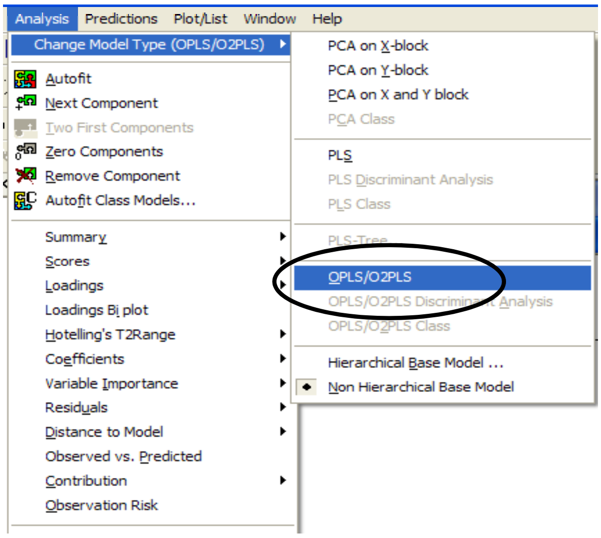
- When working with discrete variables the method is called OPLS-discriminant analysis (DA)
  - Useful in classification studies
  - Biomarker identification

- When working with continuous variables the method is called OPLS
  - Can also use PLS

- This course will cover applications based on one (1) discrete Y variable
  - Theory for continuous and discrete Y variables are the same

OPLS/O2PLS

- In SIMCA-P+12, OPLS and O2PLS use same algorithm
- To simplify the theory and applications
  - OPLS=single **Y**
  - O2PLS=multiple **Y**



Multivariate workflows for "omics" analysis

Classification by OPLS-DA

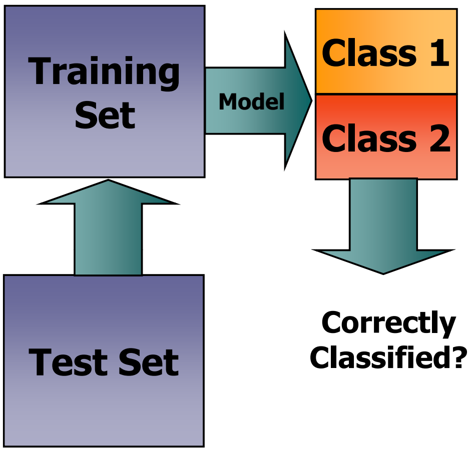
UMETRICS

Classification Models

- Two steps:

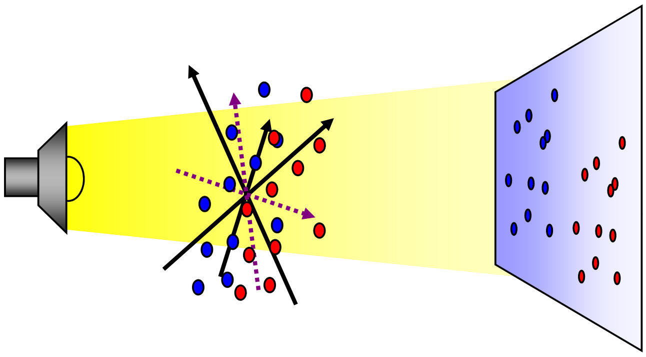
1. Train a model on Representative data
2. Test the model using new data

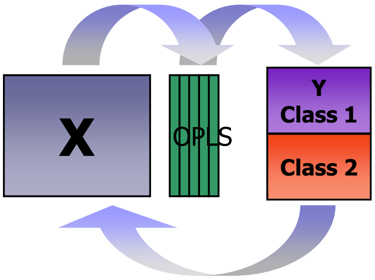
- New data must try to capture all possible experimental variation to ensure robustness
- Model judged on classification success
- Step 2 is (AMAZINGLY) often forgotten!

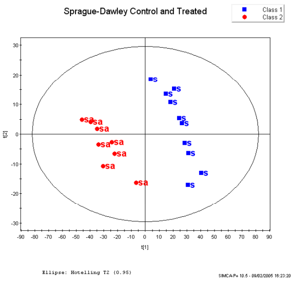


Separating Groups OPLS-DA

- OPLS-DA relies on a projection of X as does PCA
- OPLS-DA is a classification method using same algorithm as OPLS
  - Maximum Separation Projection
  - Guided by known class information
  - Easiest to interpret with 2 classes
  - Extendable to more classes
- **Interpretation** is the advantage for OPLS-DA:
  - Shows which variables responsible for class discrimination
- Omics Applications
  - Predictions (diagnostics)
  - Biomarkers in metabonomics, proteomics and genomics



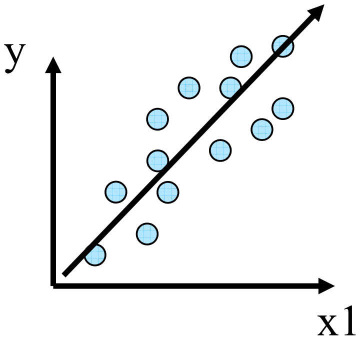




OPLS "Language"

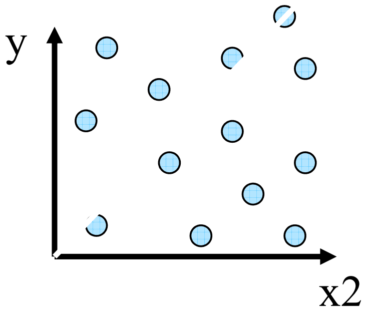
- Predictive variation = correlated variation between X and Y
- Orthogonal variation = Uncorrelated variation between X and Y

Predictive variation



Correlation(X,Y)>0

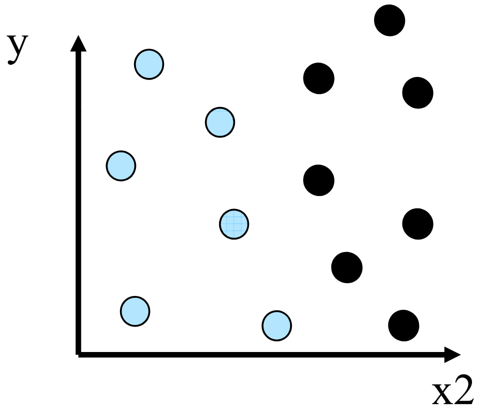
Orthogonal variation



Correlation(X,Y)=0

Orthogonal variation

- Although no correlation between X and Y, OPLS finds other types of systematic variation
- Random variation will not be found by OPLS, this part is left in the residuals
- This variation is important information for the total understanding of the studied biological system



Known variation

- Time trends
- Gender
- Growth conditions (plants)

Unknown variation

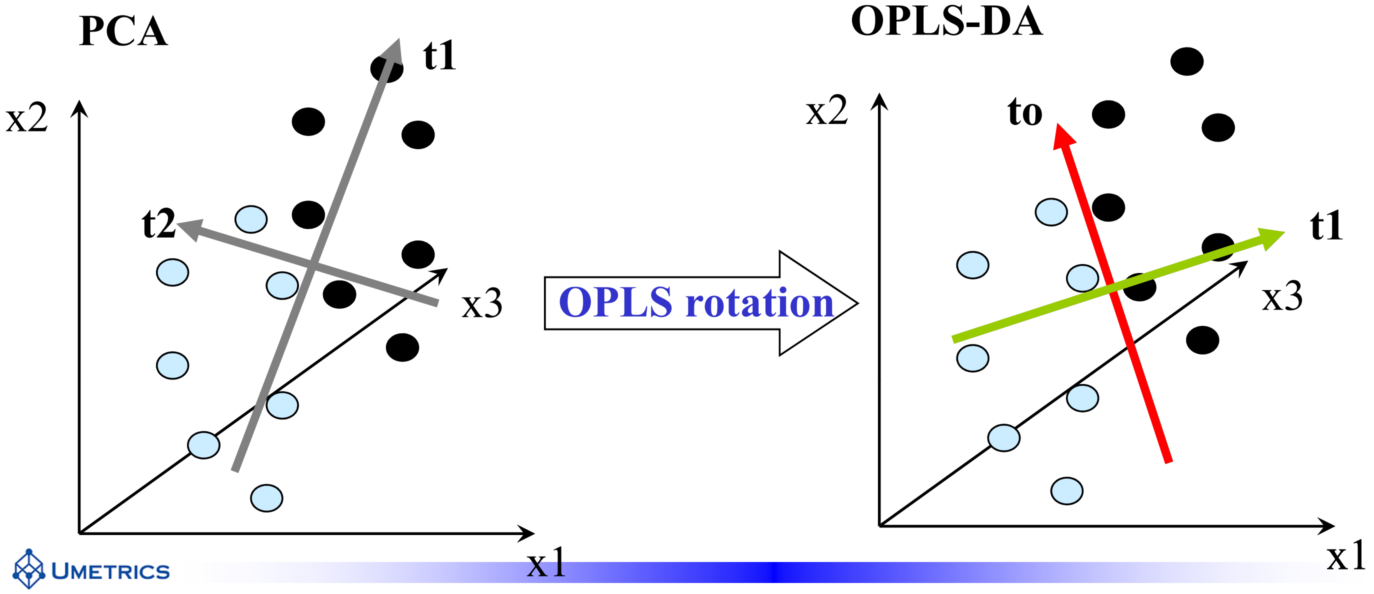
- Instrumental problems
- Sampling problem
- Sample management
- Life style (humans)

The importance of knowledge about orthogonal variation

- Increased understanding of all variation in the studied system will
  - Improve interpretation
  - Reduce the possibility of misleading interpretation
  - Improve the biological interpretation
  - Improve experimental procedures in the future
    - Design of experiment
    - Improve normalisation
    - Improve animal handling
    - Standardize diet for humans
- Use all known information about the data in the analysis
  - Take notes about all things that happens during the experimental and pre-processing procedure

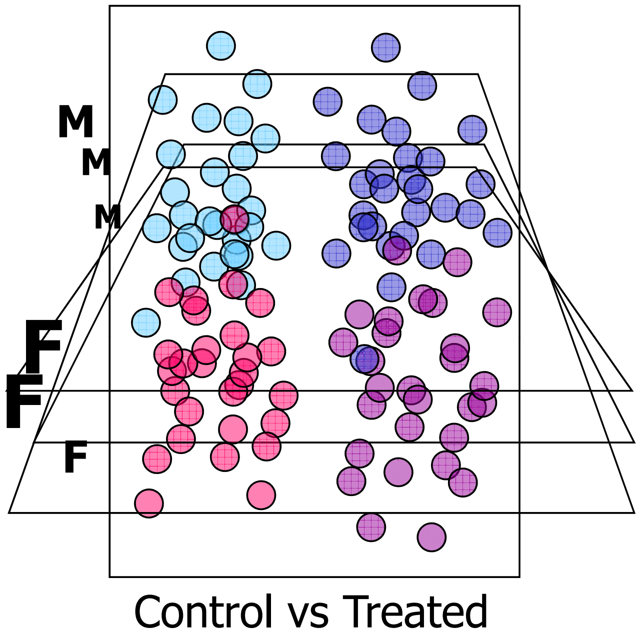
Why not only PCA?

- OPLS will focus the predictive information in one component and the other systematic information will be found in higher components
- This facilitates interpretation
- We still need PCA to look at trends and identify outliers!



OPLS-DA can cope with unwanted variation

- Often the effect we are looking for is masked by other unwanted variation
- OPLS is able to rotate the projection so that the model focuses on the effect of interest
- Here we want to focus on **control vs treated** but **gender** is the bigger influence on X
- OPLS causes a rotation so that the first OPLS component shows the between class difference

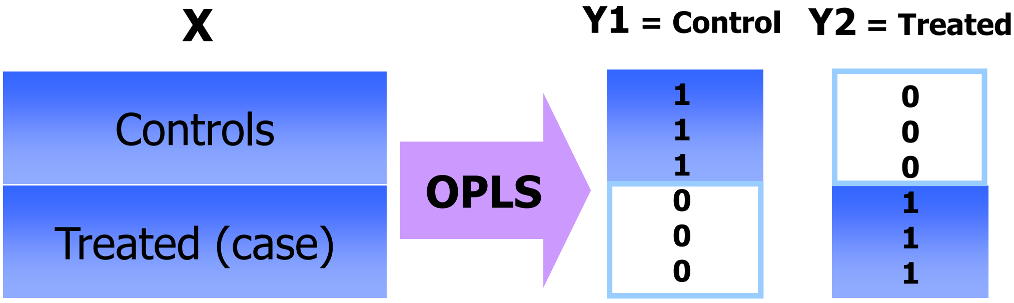


How to make an OPLS-DA model 1

- There are two alternatives how to do OPLS-DA in SIMCA-P

1. Use the OPLS/O2PLS-DA function

- SIMCA uses a binary variable for Y which represents class membership (discrete variable)
- In SIMCA a Dummy Y variable is assigned when you define a class ($DA1 or $DA2)
- Select OPLS/OPLS-DA for modelling
- Predictions have a value between 0 and 1 depending on class membership
- This alternative is faster than the second



X

Y1 = Control

Y2 = Treated

Controls

Treated (case)

OPLS

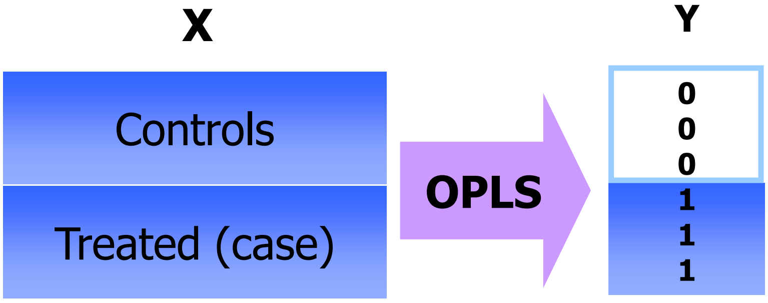
1 1 1 0 0 0

0 0 0 1 1 1

How to make an OPLS-DA model 2

2. Use the OPLS/O2PLS function

- Create a binary Y vector in e.g. excel and paste it in to the work sheet
- Always choose Y=1 for treated and Y=0 for controls to designate **class belonging**
  - Possible to assign class membership during import or after import
- This will simplify interpretation
  - Positive loadings mean up regulated
  - Negative loading mean down regulated
- Select OPLS/O2PLS using the created Y (1 and 0) as the response
- Predictions then give value between 0 and 1 depending on membership
- The significant advantage with method 2: **easier to compare different models**



X

Y

Controls

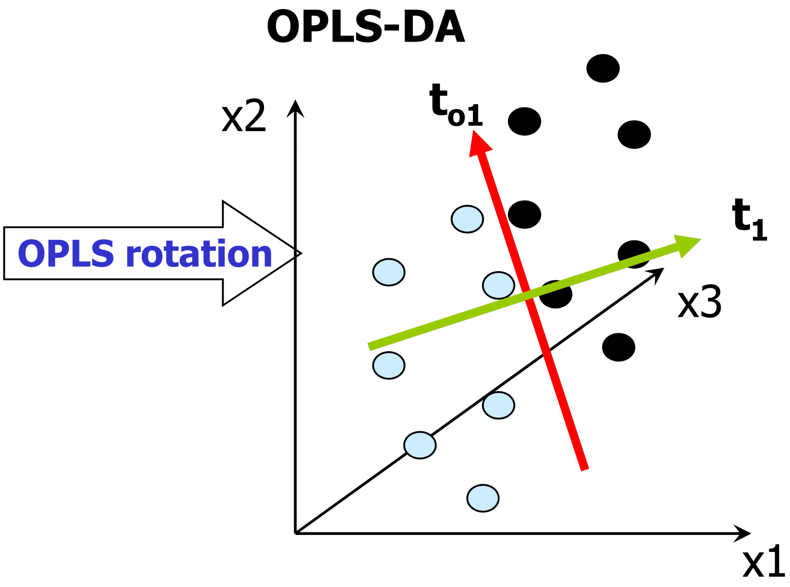
Treated (case)

OPLS

0 0 0 1 1 1

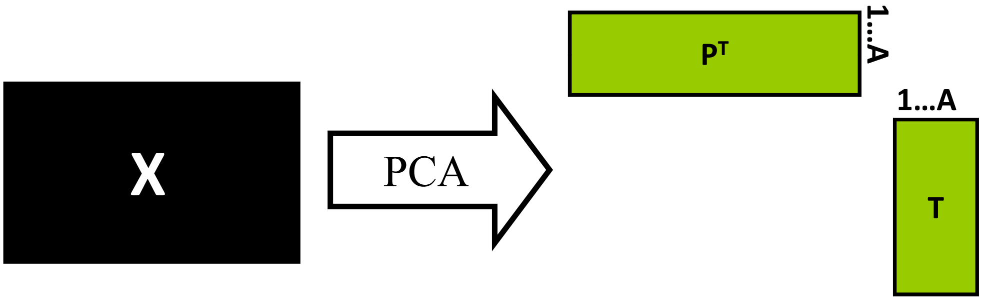
OPLS-DA Geometric Interpretation

- OPLS-DA finds the variation in X that is correlated the Y variable
- This is done by a rotation towards the direction of Y
- At the same time OPLS-DA finds components that are uncorrelated to Y but systematic in X
- As in PCA, the data should first be scaled and centred
- Each observation is represented by a point in multi-dimensional space



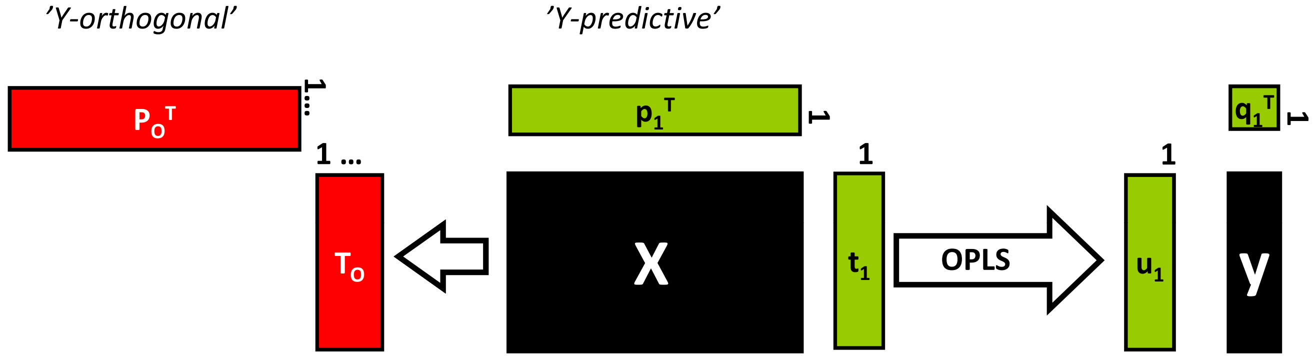
Recall from PCA

- PCA compress the **X** data block into **A** number of orthogonal components
- Variation seen in the score vector **t** can be interpreted from the corresponding loading vector **p**



PCA Model $\mathbf{X} = t_1p_1^T + t_2p_2^T + \ldots + t_Ap_A^T + E = TP^T + E$

OPLS with single Y / modelling and prediction



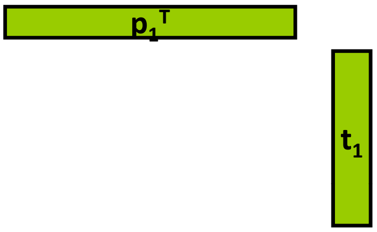
OPLS Model

$$\begin{cases} X = t_1 p_1^T + T_O P_O^T + E \\ Y = t_1 q^T{}_1 + F \end{cases}$$

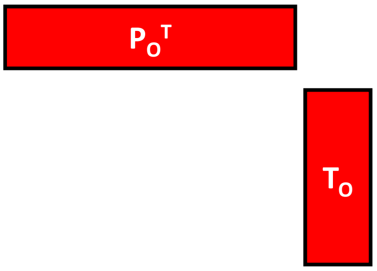
OPLS with single y / interpretation

Few vectors to keep in mind for interpretation

- What's correlated between X→Y?
  - Look at the Y-predictive vector i.e. t_1 and the corresponding p_1

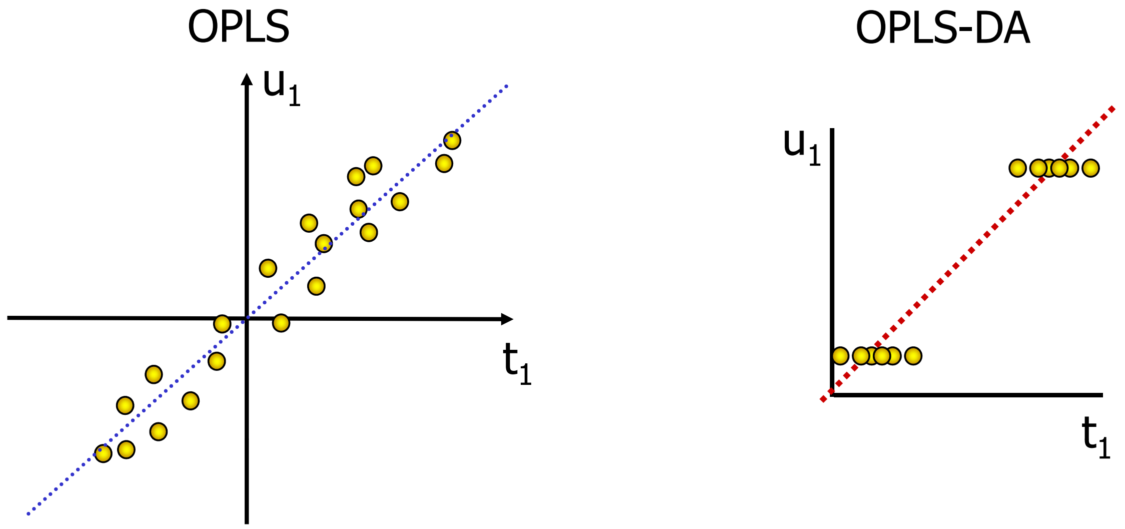


- What is seen in the uncorrelated vectors, X⊥Y?
  - Unique systematic variation in X
  - Look at the Y-orthogonal vectors i.e. T_o and the corresponding P_o



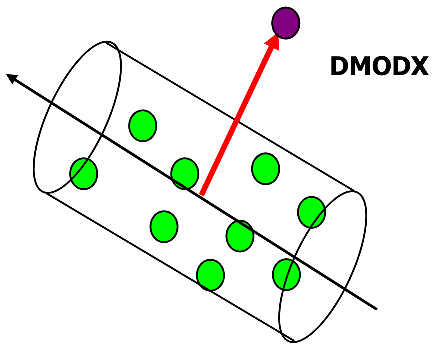
OPLS Inner Relation

- The inner relation between X and Y is seen in the score plot between t_1 and u_1

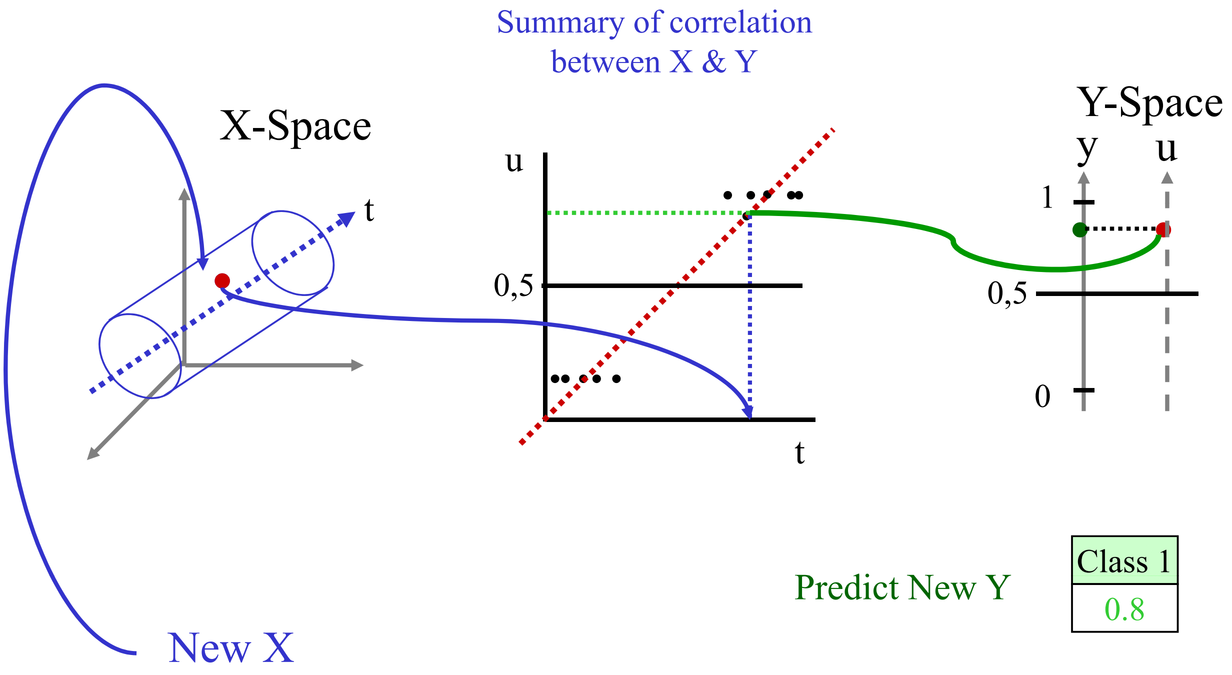


Recall PCA and DModX

- Same rules as for PCA
- DModX shows the distance to the model plane for each observation
- Use DModX to detect moderate deviating samples
- High DModX= uncertain prediction



OPLS-DA - Predictions



Class 1
0.8

Summary of OPLS 1

- OPLS will improve model visualization and interpretation
  - Separates data into predictive and uncorrelated information
    - Improved diagnostics (will be explained)
  - Improved visualization tools
    - Score plot t[1] vs to
    - Loading plot p[1] and p(corr)[1], p(corr)o
    - S-plot
    - SUS-plot

- Concept of uncorrelated information
  - Experimental problem(s)
    - Life style (humans)
    - Growth conditions (plants)
    - Instrument failures

Summary OPLS 2

- **Explanation for single Y M=1**
- The first predictive OPLS component is a line in the X-space with maximum co variation and correlation between X and Y. The direction of the predictive component can be found in t_1 and p_1
- The additional orthogonal OPLS components are lines in the X-space which are uncorrelated to Y. The direction of these orthogonal components can be found in T_o and P_o
- Easier interpretation of score plot (to be demonstrated)
- Easier identification/interpretation of putative bio markers (to be demonstrated)
- Identification/interpretation of uncorrelated information (to be demonstrated)
- More transparent interpretation of model diagnostics (to be demonstrated)
- Works well for omics data

OPLS or OPLS-DA is NOT

- A method that gives better prediction than PLS
  - Models are identical so predictions are identical
  - Q2 is different between OPLS and PLS but that is due to different techniques of cross-validation
- A pre processing method

Multivariate workflows for “omics” analysis

OPLS-DA model diagnostics

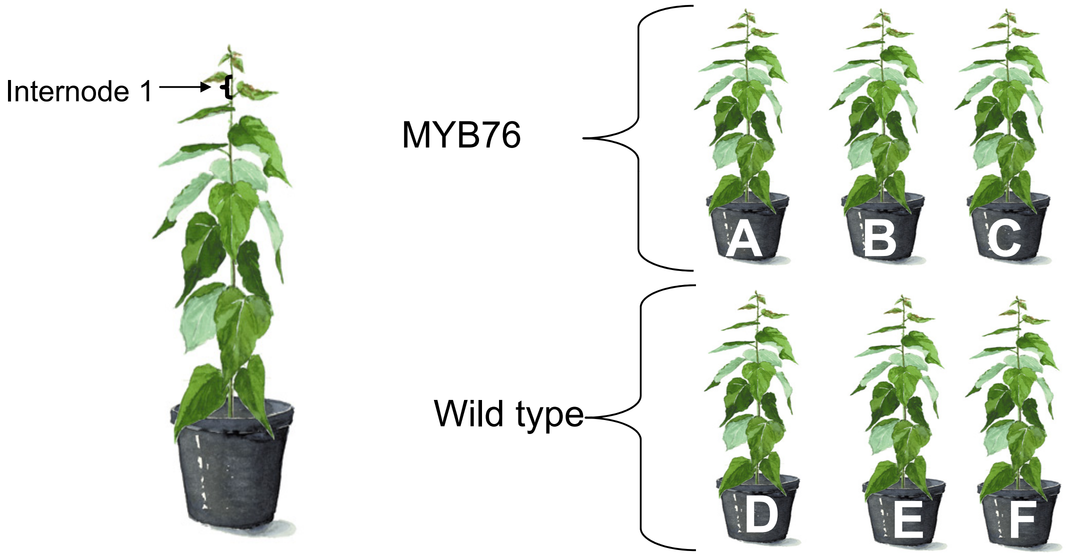
UMETRICS

Example 1 PCA compared to OPLS-DA

Plant metabolomics

- **Samples:** Transgenic aspen
  - Wild type, WT
  - MYB76
- **Data:** High resolution magic angle spinning, HR/MAS, 1H NMR spectroscopy
  - 500MHz spectrometer
- **Objective:** To detect metabolic differences between
  - a) Wild type poplar (control group) and Transgenic, MYB76 modified poplar

Example 1 Sampling

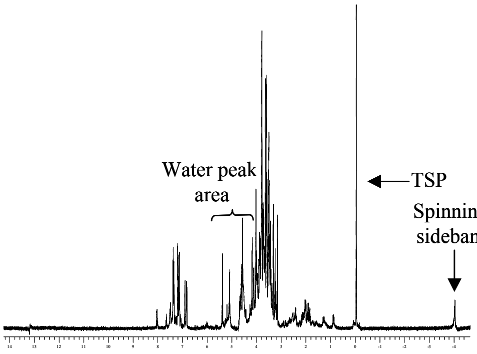


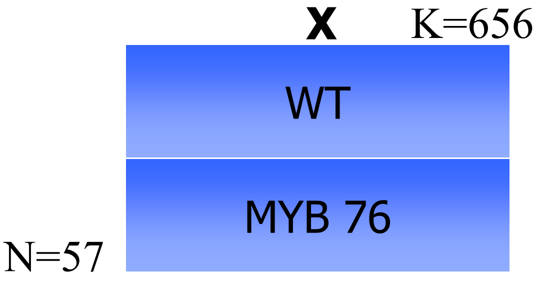
Samples collected by the internodes of poplar plants
N_{tot}=57 samples (3*8*2 + analytical replicates)

Example 1 Data pre-treatment

Data reduction
Bucketing width 0.02 ppm
Removal of water peak,
TSP and Spinning Sidebands

Normalisation
variation in concentration
is removed



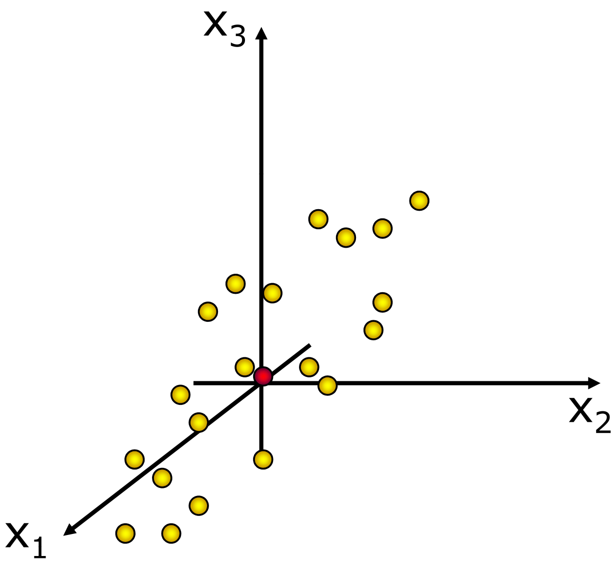


Two models

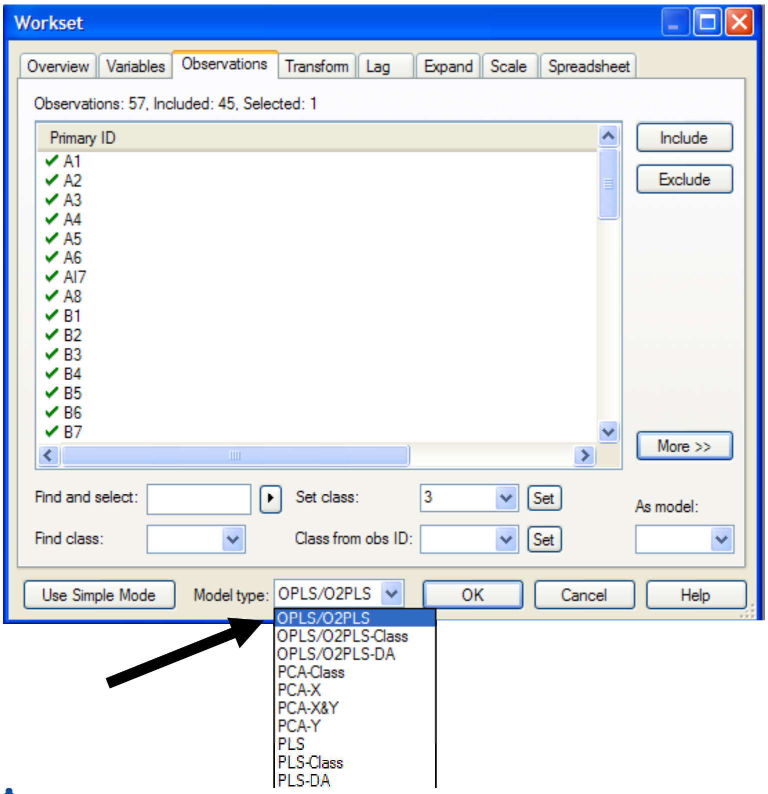
1) PCA
2) OPLS-DA

Scaling of Variables

- Same rules as for PCA
- Default in SIMCA: To centre and set variation along each axis to one (unit variance)
- For spectroscopic data pareto scaling is a good choice
  - –Minimise the influence of noise and artefacts
  - –Spectral line shapes is maintained



Example 1 Model settings in SIMCA-P+ 12



Observations: 57, Included: 45, Selected: 1

Primary ID
✓ A1
✓ A2
✓ A3
✓ A4
✓ A5
✓ A6
✓ AI7
✓ A8
✓ B1
✓ B2
✓ B3
✓ B4
✓ B5
✓ B6
✓ B7

Find and select:
Set class: 3
Find class:
Class from obs ID:
As model:
Model type: OPLS/O2PLS

Workset

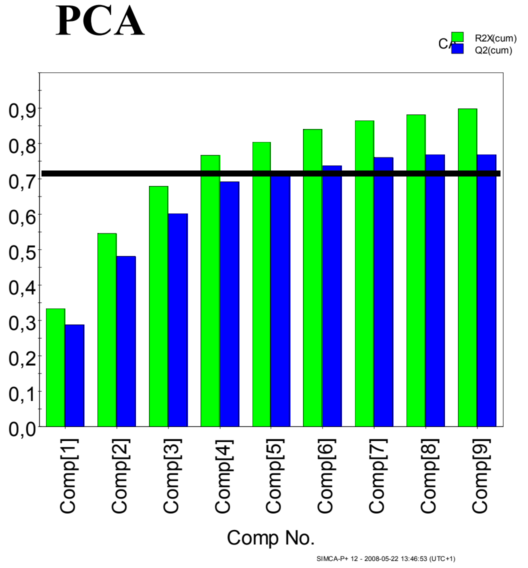
- Set scaling (par)
- Define the response vector Y (WT=0, MYB76=1)
- Set model type to OPLS/O2PLS

Same data for PCA model

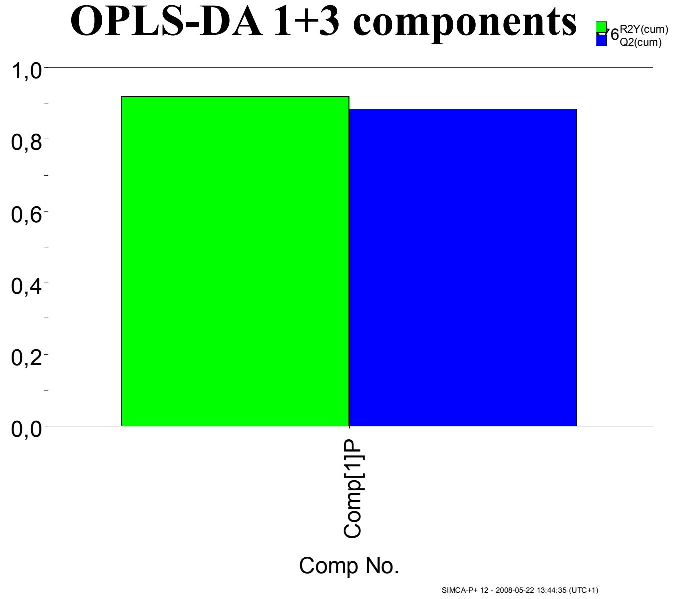
Example 1 Nr of components

- Always use cross-validation to decide number of components
  - Not always easy

PCA

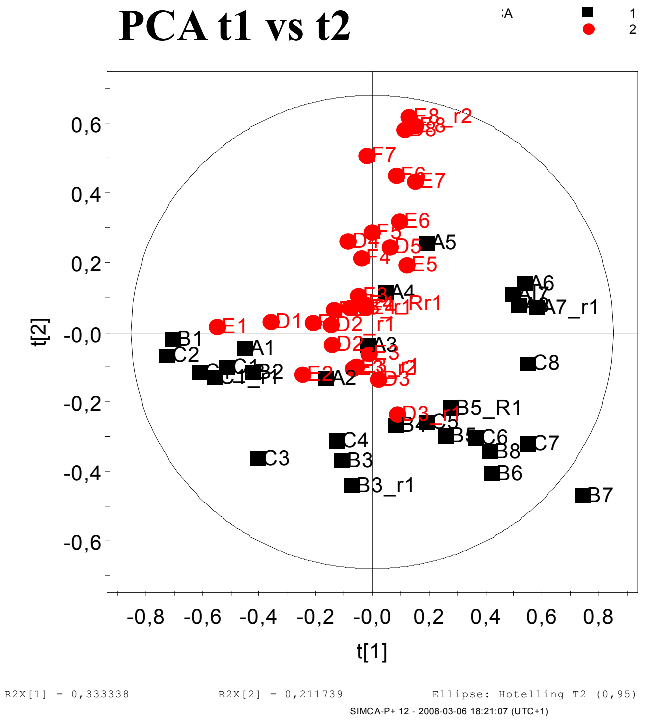


OPLS-DA 1+3 components

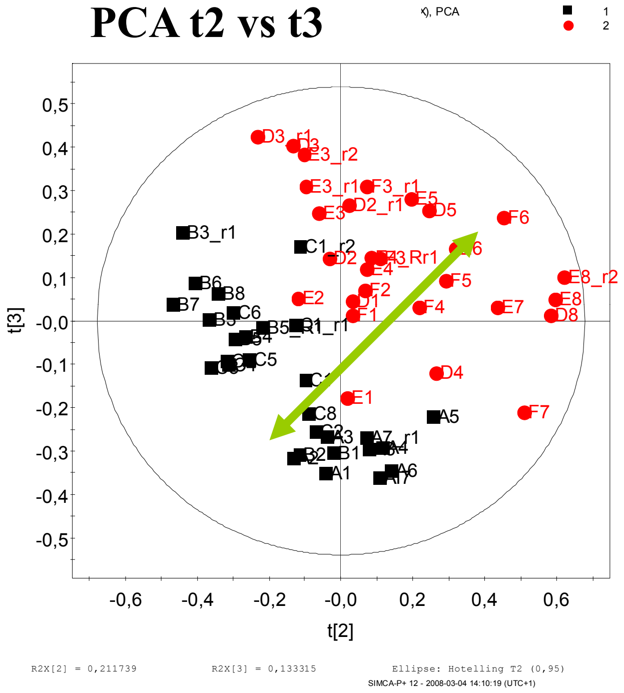


Example 1 Model interpretation of scores

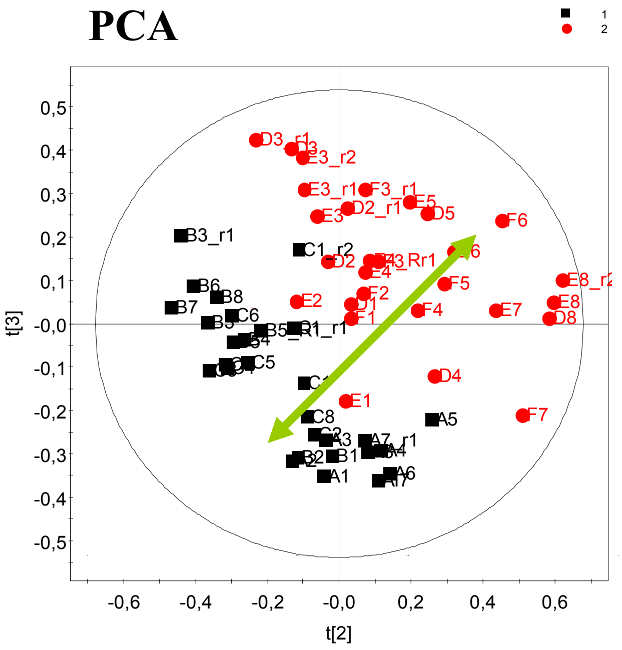
PCA t1 vs t2



PCA t2 vs t3



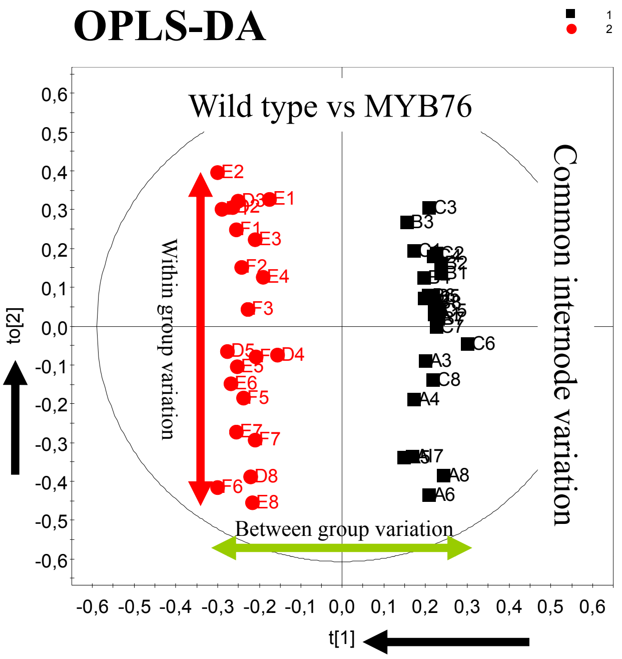
Example 1 Model interpretation of scores



R2X[2] = 0,211739 R2X[3] = 0,133315 Ellipse: Hotelling T2 (0,95)

SIMCA-P+ 12 - 2008-03-04 14:10:19 (UTC+1)

SIMCA-P+ 12 - 2008-03-06 18:21:07 (UTC+1)



R2X[1] = 0,156759 R2X[XSide Comp. 2] = 0,211296
Ellipse: Hotelling T2 (0,95)

SIMCA-P+ 12 - 2008-03-04 14:10:52 (UTC+1)

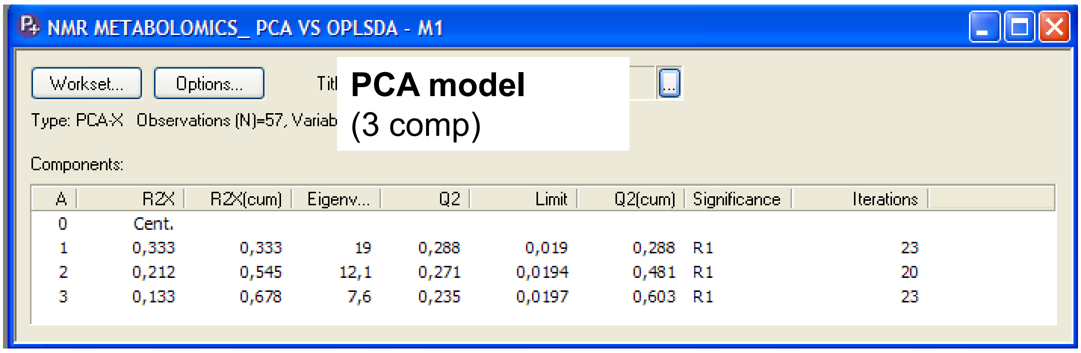
Example 1 Model Diagnostics

- **Questions**

1. **How good is the separation between the two plants?**
2. **How much variation in X is *related* to the separation of the two plants?**
3. **How much of the variation is related to common internode variation?**

- **Can not answer these questions due to PCA mixes both types of variation. Interpretation issues.**

Example 1 Model diagnostics PCA



PCA model
(3 comp)

Type: PCA-X Observations (N)=57, Variab

Components:

A	R2X	R2X(cum)	Eigenv...	Q2	Limit	Q2(cum)	Significance	Iterations
0	Cent.							
1	0,333	0,333	19	0,288	0,019	0,288	R1	23
2	0,212	0,545	12,1	0,271	0,0194	0,481	R1	20
3	0,133	0,678	7,6	0,235	0,0197	0,603	R1	23

Fit or R2

- •Explained variation
- •For whole model or individual variables

R2X(cum) = 1- RSS/SSX$_{tot.corr}$=1-unexplained variation

RSS=$\sum$(obs-pred)^2

Prediction (Cross-validation) Q^2

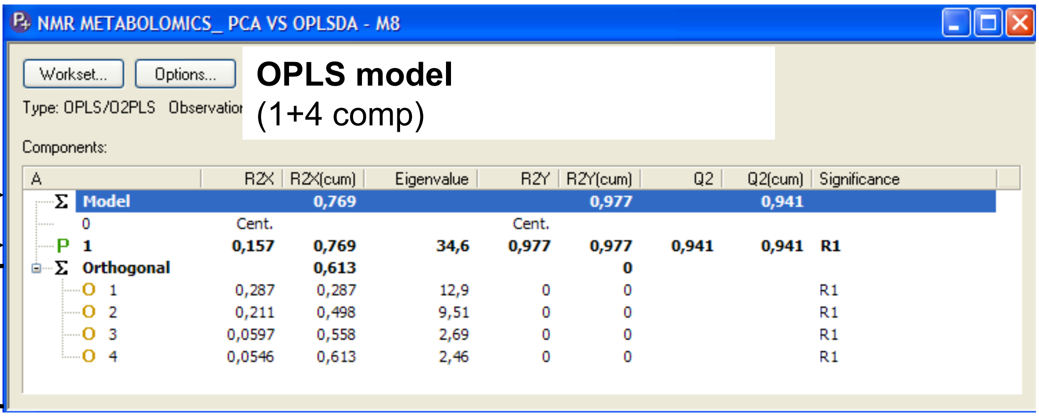
- •Predictive variation
- •Leave out data in turn
- •Predict each missing block of data in turn
- •Sum the results

Q^2 = 1 - PRESS/SSX$_{tot.corr}$

PRESS=$\sum$(obs-pred)^2 → Cross validation

Example 1 Model Diagnostics OPLS-DA

SIMCA-P+ 12



OPLS model
(1+4 comp)

Type: OPLS/O2PLS Observatior

Components:

Model summary

Predictive variation

Orthogonal variation
X⊥Y

A	R2X	R2X(cum)	Eigenvalue	R2Y	R2Y(cum)	Q2	Q2(cum)	Significance
Σ Model		**0,769**			**0,977**		**0,941**	
0	Cent.			Cent.				
P 1	**0,157**	**0,769**	**34,6**	**0,977**	**0,977**	**0,941**	**0,941**	**R1**
Σ Orthogonal		**0,613**			**0**			
O 1	0,287	0,287	12,9	0	0			R1
O 2	0,211	0,498	9,51	0	0			R1
O 3	0,0597	0,558	2,69	0	0			R1
O 4	0,0546	0,613	2,46	0	0			R1

R^2 and Q^2 is also of importance for OPLS and OPLS-DA models

- •R^2 is separated into predictive and orthogonal variation

Example 1 Model Diagnostics in OPLS-DA

Model Summary

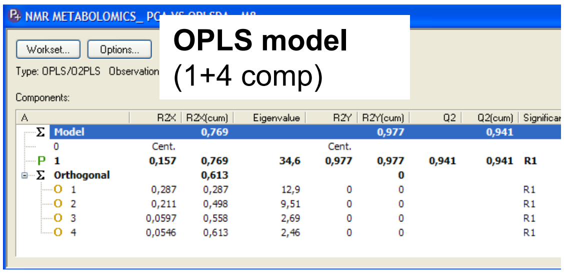
R2X(cum)=0,769. Predictive + orthogonal variation in X that is explained by the model (0,157+0,613=0,769).
R2Y(cum)= 0,977. Total sum of variation in Y explained by the model.
Q2(cum)= 0,914. Goodness of prediction, calculated by full cross validation.

P=Predictive variation, variation in X that is correlated to Y

A=1, corresponds to number of correlated components between X and Y.
R2X=0,157. This is the amount of variation in X that is correlated to Y.

O=Orthogonal variation, variation in X that is uncorrelated to Y, X⊥Y

A=4. corresponds to number of uncorrelated components. Each uncorrelated component can be interpreted individually.
R2X=Amount of variation in X that is uncorrelated to Y but with systematic variation. Each component is represented individually.
R2X(cum)=0,613. In bold is the total sum of variation in X that is uncorrelated to Y.



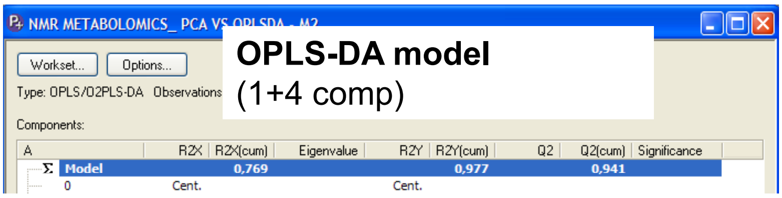
Type: OPLS/O2PLS Observation

OPLS model
(1+4 comp)

Components:

A	R2X	R2X(cum)	Eigenvalue	R2Y	R2Y(cum)	Q2	Q2(cum)	Significar
Σ Model		0,769			0,977		0,941	
0	Cent.			Cent.				
P 1	0,157	0,769	34,6	0,977	0,977	0,941	0,941	R1
Σ Orthogonal		0,613			0			
O 1	0,287	0,287	12,9	0	0			R1
O 2	0,211	0,498	9,51	0	0			R1
O 3	0,0597	0,558	2,69	0	0			R1
O 4	0,0546	0,613	2,46	0	0			R1

Example 1 Model Diagnostics in OPLS-DA

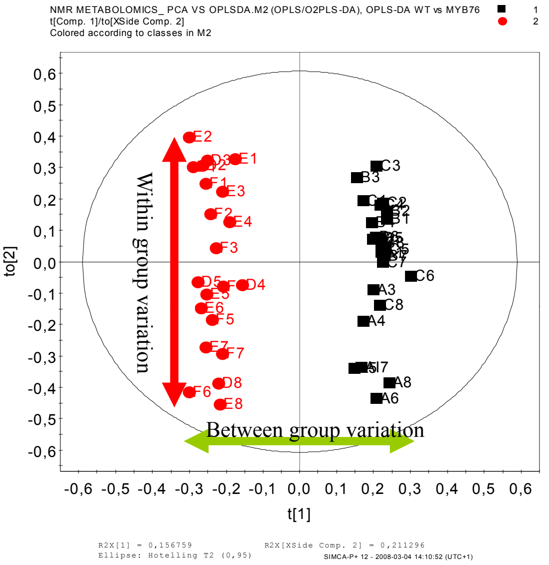


Type: OPLS/O2PLS-DA Observations

OPLS-DA model
(1+4 comp)

Components:

A	R2X	R2X(cum)	Eigenvalue	R2Y	R2Y(cum)	Q2	Q2(cum)	Significance
Σ Model		0,769			0,977		0,941	
0	Cent.			Cent.				

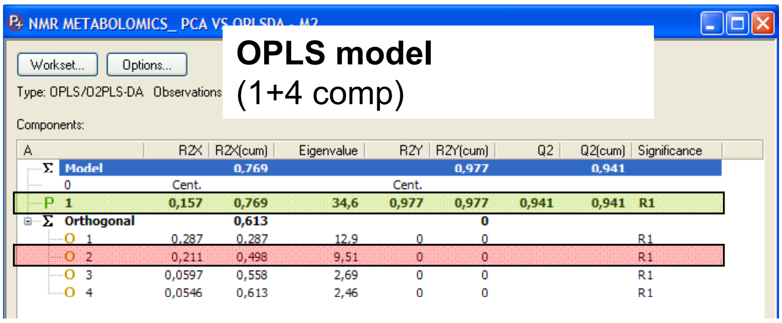


How good is the separation between the two plants?

R2Y(cum)=0,977
Q2Y(cum)=0,941

- **The higher R2Y and Q2 the better separation between WT and MYB76**
- **If R2Y(cum) and Q2(cum) differ more than 0.3 be careful**

Example 1 Model Diagnostics in OPLS-DA

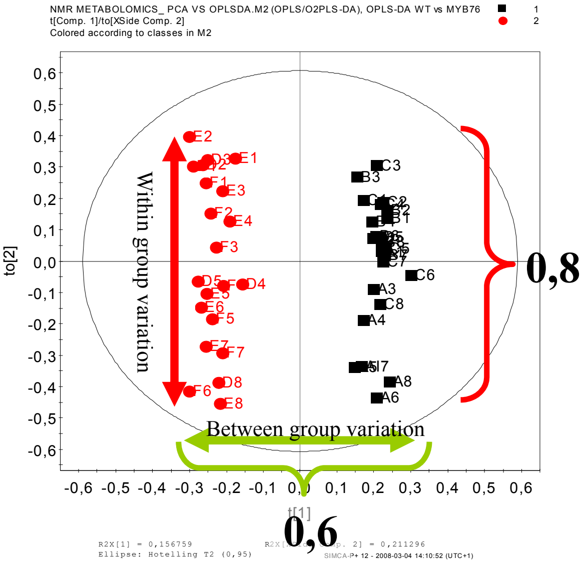


Type: OPLS/O2PLS-DA Observations

Components:

A	R2X	R2X(cum)	Eigenvalue	R2Y	R2Y(cum)	Q2	Q2(cum)	Significance
Σ Model		0,769			0,977		0,941	
0	Cent.			Cent.				
P 1	0,157	0,769	34,6	0,977	0,977	0,941	0,941	R1
Σ Orthogonal		0,613			0			
O 1	0,287	0,287	12,9	0	0			R1
O 2	0,211	0,498	9,51	0	0			R1
O 3	0,0597	0,558	2,69	0	0			R1
O 4	0,0546	0,613	2,46	0	0			R1

OPLS model
(1+4 comp)



How much variation in X is *related* to the separation of the two plants?
R2X(1)=0,157→15,7% → 0,6

How much of the variation is related to common internode variation?
R2X(o2)=0,211→21,1% → 0,8

Example 1 Answers to questions

1. **How good is the separation between the two plants?**
   R2Y(cum)=0,977
2. **How much variation in X is *related* to the separation of the two plants?**
   R2X(1)=0,157→15,7%
3. **How much of the variation is related to common internode variation?**
   R2X(o2)=0,211→21,1%

OPLS model diagnostics can answer all questions

Summary

- OPLS is a rotation of the model plane
  - No magic, pure mathematics
- OPLS separates predictive variation from orthogonal variation
  - Predictive variation = Correlated variation between X and Y
  - Orthogonal variation = systematic variation in X uncorrelated to Y
- Facilitates model interpretation
- OPLS makes the diagnostics more transparent

Multivariate workflows for “omics” analysis

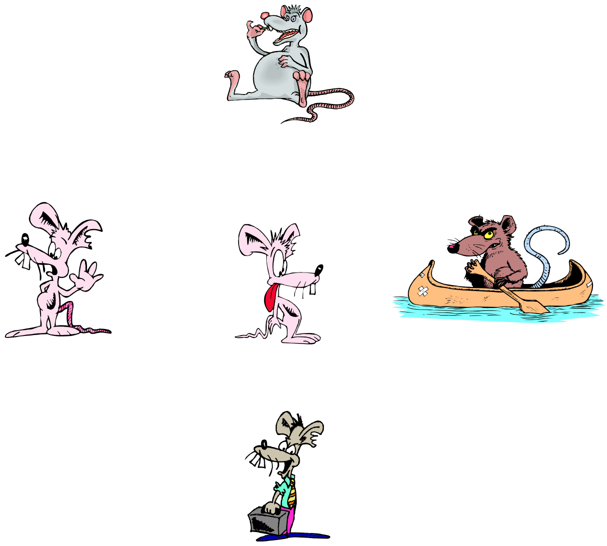
OPLS-DA for biomarker identification

Outline

- Multiple groups
- Useful tools in biomarker identification
  - S-plot
  - SUS-plot
- Example: GC/MS metabolomics
- Balanced models

Multiple groups

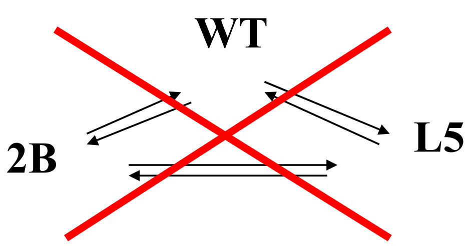
- **Detailed information is easiest to interpret with 2 classes**
  - OPLS-DA **loadings** are difficult to interpret with >2 classes
- **It is still possible to compare more than 2 classes**
  - A solution to the problem will be provided

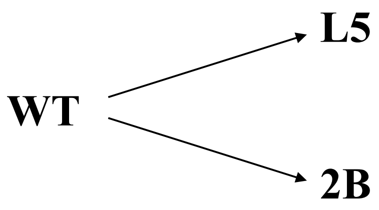


Problem formulation “omics”

- Omics rarely have a multi-group problem
- Omics problems often have one (1) Control vs. several treated
  - **Wild type vs. number of genetically modified (plant science)**
  - Control vs. treated
  - Control vs. time point 1,2,3,4

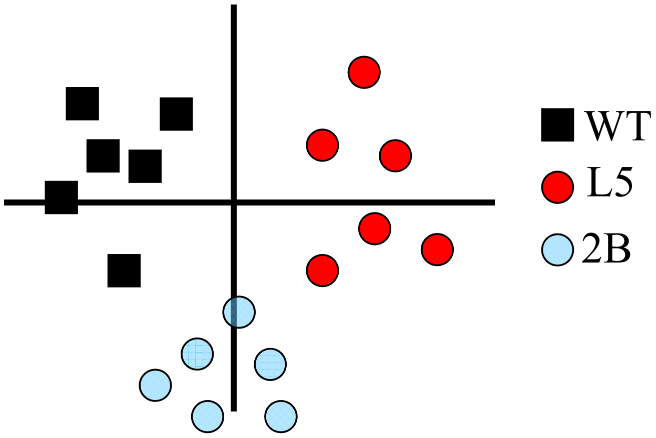
- The multivariate evaluation should therefore be performed so that we compare each treated group vs the control.



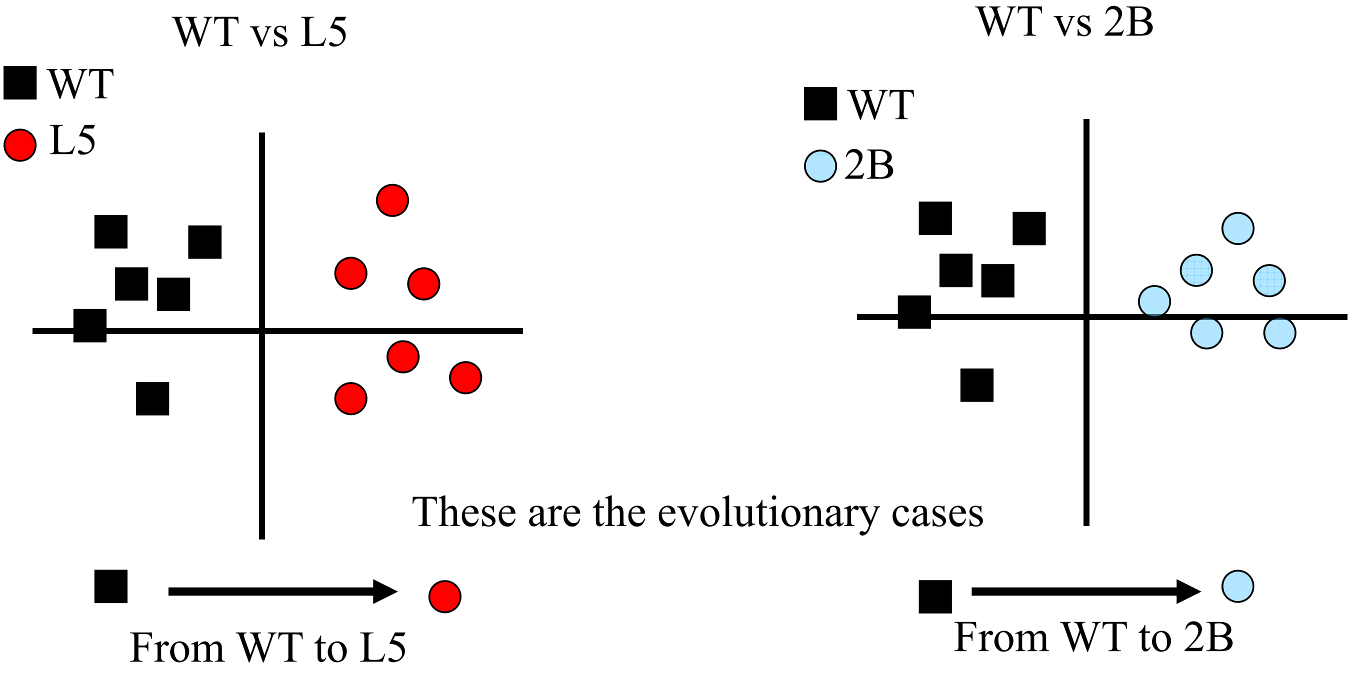


Why not more than 2 classes?

- If ONLY score overview is wanted than 3 to many classes is OK
- The reference (the zero) in a 3 class model is the average of all 3 classes
- In this example the loading interpretation would be VERY difficult
- For biomarker identification use only 2 classes at time



Better to do separate models



These are the evolutionary cases

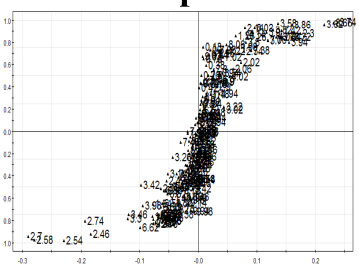
From WT to L5

From WT to 2B

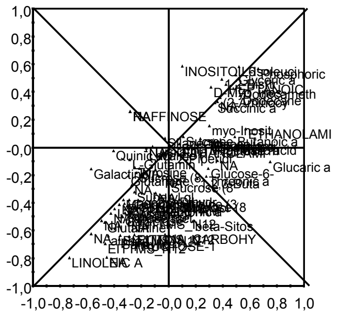
Useful tools in biomarker identification

- S-plot for the extraction of putative bio-markers
- Loading plot with jack-knifed confidence intervals
- SUS-plot to detect Shared and Unique Structures when many classes are compared to a common reference

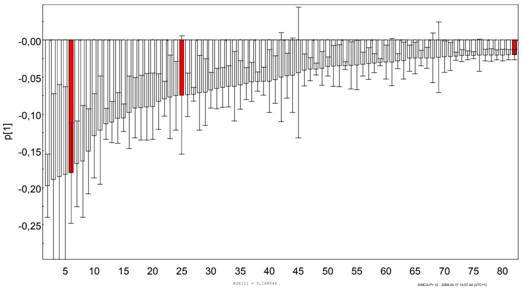
S-plot



SUS-plot



Loading plot

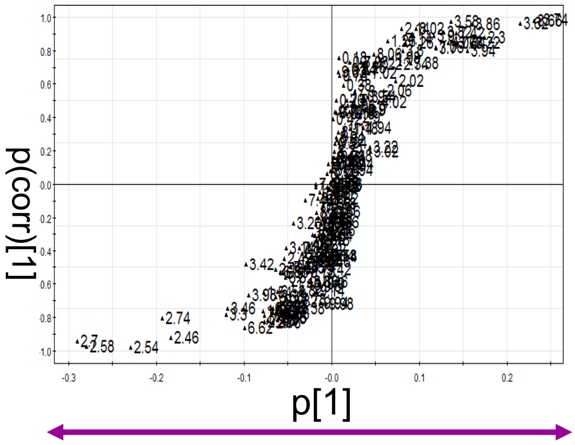


S-plot

Putative biomarker identification

- Visualisation of the OPLS-DA loadings
- Combines the modelled covariance and modelled correlation from the OPLS-DA model in a scatter plot
- If the data, **X**, has variation in peak intensities, this plot will look like an S
- The p_1-axis will describe the magnitude of each variable in X
- The $p(corr)_1$ -axis represents the reliability of each variable in X
- $p(corr)_1$ – axis always between ±1

Reliability



Variable Magnitude

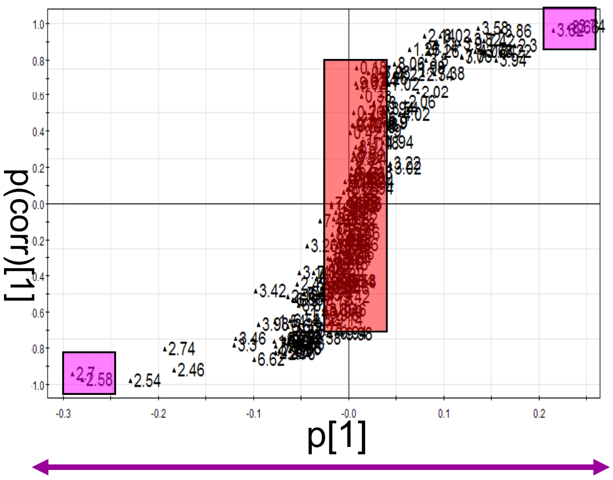
S-plot

- Why is this of interest?
- Good overview of the data and model
- Peaks with low magnitude/intensity are close to the noise level
  - High risk for spurious correlations
- Ideal biomarker have high magnitude and high reliability
  - Smaller risk for spurious correlations

SIGNAL TO NOISE

Modelled correlation

Reliability



Variable Magnitude
Modelled co-variation

S-plot

Notations used in SIMCA-P+ 12

Variable magnitude→ $Cov(\mathbf{t}_1, \mathbf{X}) = \dfrac{\mathbf{t}_1^T \times \mathbf{X}}{N-1} = \mathbf{p}[1]$

Modelled co-variation

Reliability→ $Corr(\mathbf{t}_1, \mathbf{X}) = \dfrac{Cov(\mathbf{t}_1, \mathbf{X})}{\sigma_{\mathbf{t}_1}\sigma_{\mathbf{X}}} = \dfrac{\mathbf{p}[1]}{\sigma_{\mathbf{t}_1}\sigma_{\mathbf{X}}} = \mathbf{p}(corr)[1]$

Modelled correlation

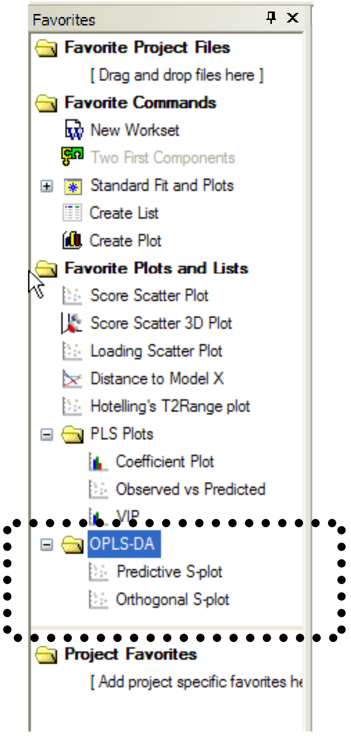
p[1] are also called model loadings

p(corr)[1] variation related to variable magnitude is removed

σ_{t1}= Standard deviation of t_1

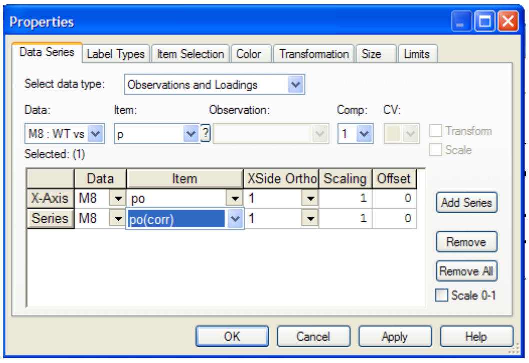
σ_X= Standard deviation of each X variable

How to do S-plot in SIMCA-P+12



Favorites
- Favorite Project Files
  - [Drag and drop files here]
- Favorite Commands
  - New Workset
  - Two First Components
  - Standard Fit and Plots
  - Create List
  - Create Plot
- Favorite Plots and Lists
  - Score Scatter Plot
  - Score Scatter 3D Plot
  - Loading Scatter Plot
  - Distance to Model X
  - Hotelling's T2Range plot
  - PLS Plots
    - Coefficient Plot
    - Observed vs Predicted
    - VIP
  - OPLS-DA
    - Predictive S-plot
    - Orthogonal S-plot
- Project Favorites
  - [Add project specific favorites he

- Predictive component
  - Go to *favorites/OPLS-DA* and select *predictive S-plot*
  - Change the axis
- Orthogonal components
  - Go to *favorites/OPLS-DA* and select *predictive S-plot*
  - Change the axis under *properties*
  - Ad this plot to *favorites* under the name *orthogonal S-plot*



Properties

Data Series | Label Types | Item Selection | Color | Transformation | Size | Limits

Select data type: Observations and Loadings

Data: M8 : WT vs Item: p Observation: Comp: 1 CV:

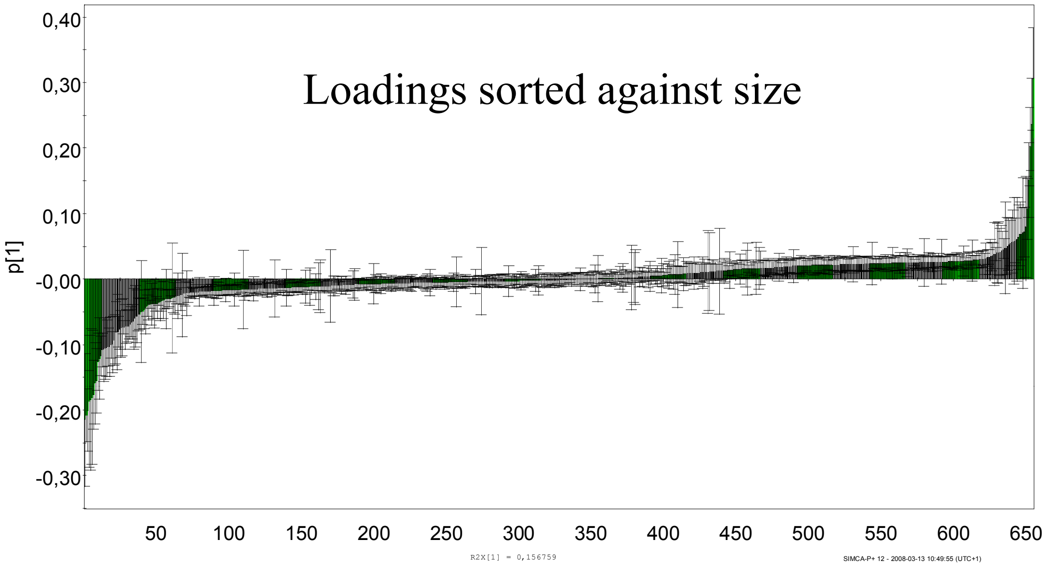
□ Transform □ Scale

Selected: (1)

	Data	Item	XSide Ortho	Scaling	Offset
X-Axis	M8	po	1	1	0
Series	M8	po(corr)	1	1	0

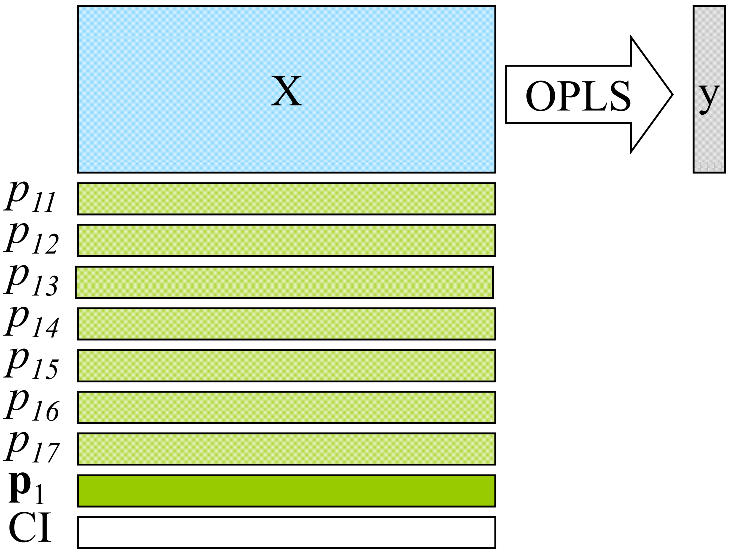
S-plot and loading plot

- The extraction of putative biomarkers from the S-plot could be combined with the jack-knifed confidence intervals seen in the loading plot



Jack-knifed confidence interval

- Loading confidence interval (CI) calculated with help from cross validation
- Example: 7-fold CV are used



The confidence interval reflects the variable stability and uncertainty

$\mathbf{p}_1=(p_{11}+p_{12}+\ldots+p_{17})/7$

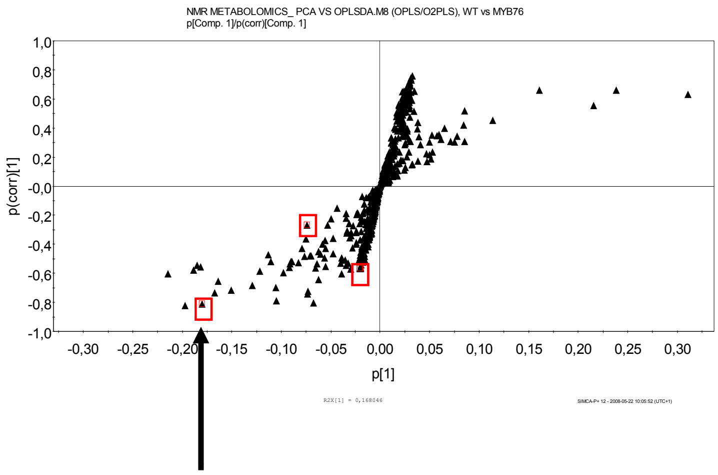
CI=SE*t(α,*df*)

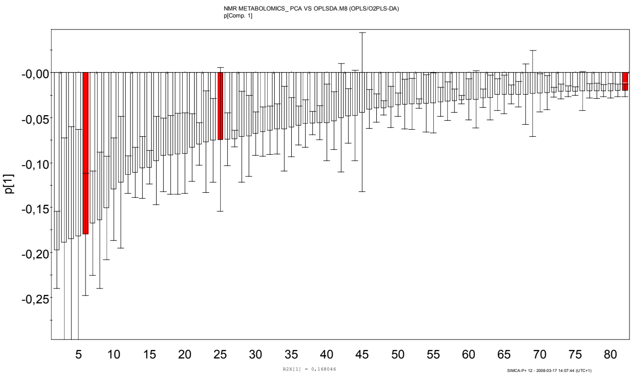
SE=SD/√df - calculated from the cross validated loadings

t(α,*df*) - by default α=0.05 and df=number of CV rounds (here 7)

S-plot and loading plot

- The combination of S-plot and loading plot interpretation can easily be done interactively in SIMCA-P+12



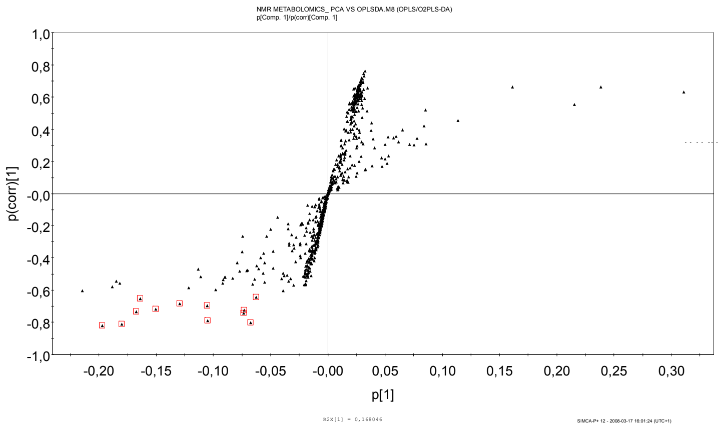


Marked variables in the S-plot

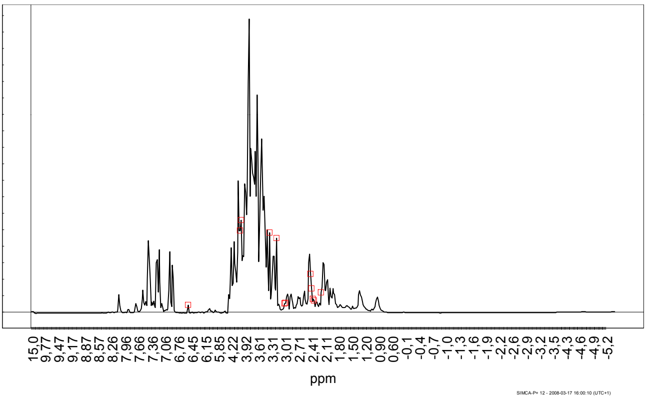
Same will be marked in the loading plot including the confidence intervals

S-plot for NMR data

- If NMR data is modelled you could also work interactively with the S-plot and the average NMR spectrum

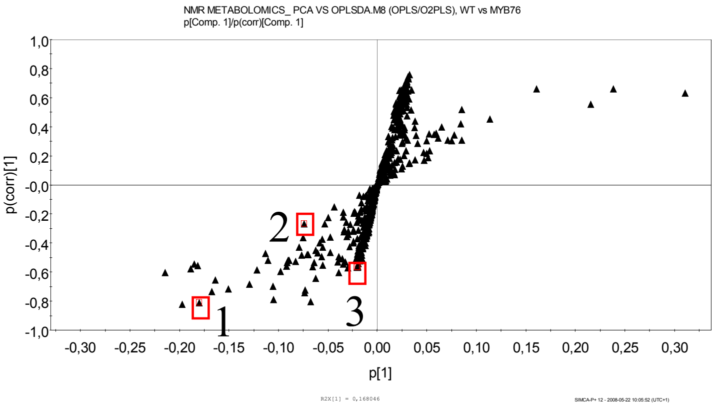


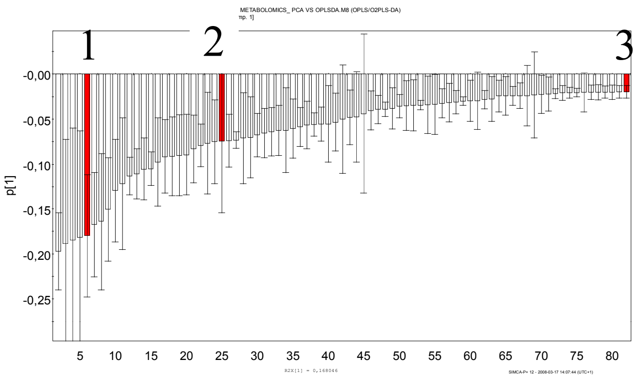
Average spectrum



S-plot and loading plot for putative biomarker identification

- Three general cases
  1. Ideal case or highly likely-High magnitude/high reliability
  2. Unlikely case-High magnitude/low reliability
  3. Difficult case-Low magnitude/high reliability

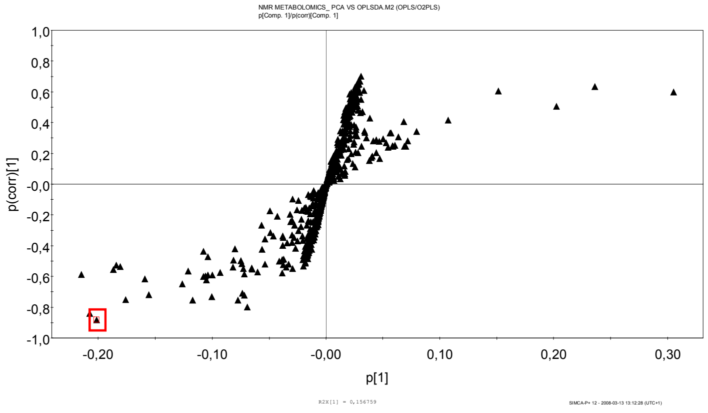


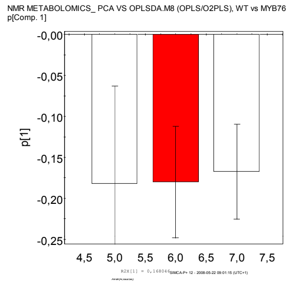


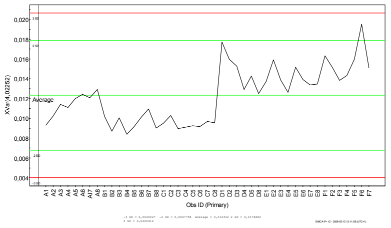
CHECK THE RAW DATA!

1. Ideal case

- Putative biomarker with high magnitude and high reliability



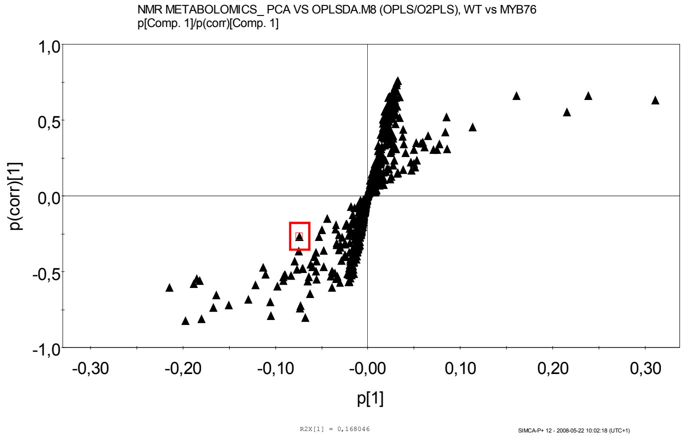


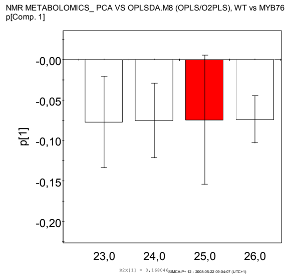


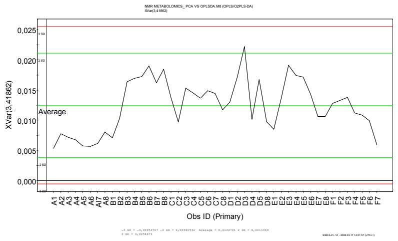
- Double-click on the variable in the S-plot and the raw data plot will appear
- Statistically significant variable

2. Unlikely case

- Putative biomarker with high magnitude and low reliability



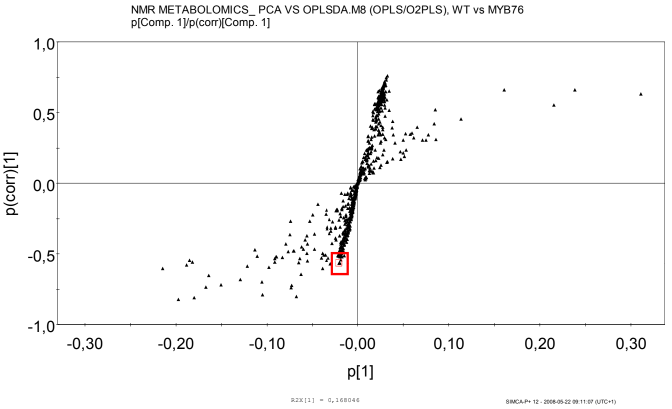




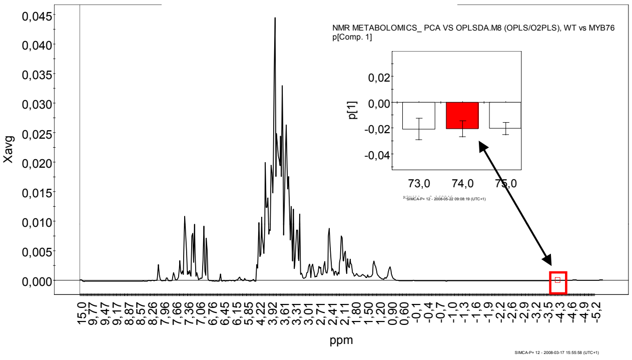
- Not statistically significant variable

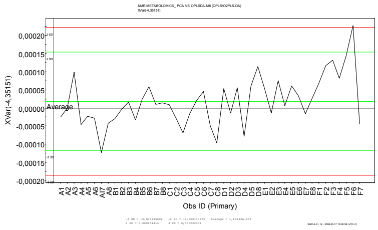
3. Difficult case

- Putative biomarker with low magnitude and medium reliability



Average spectrum

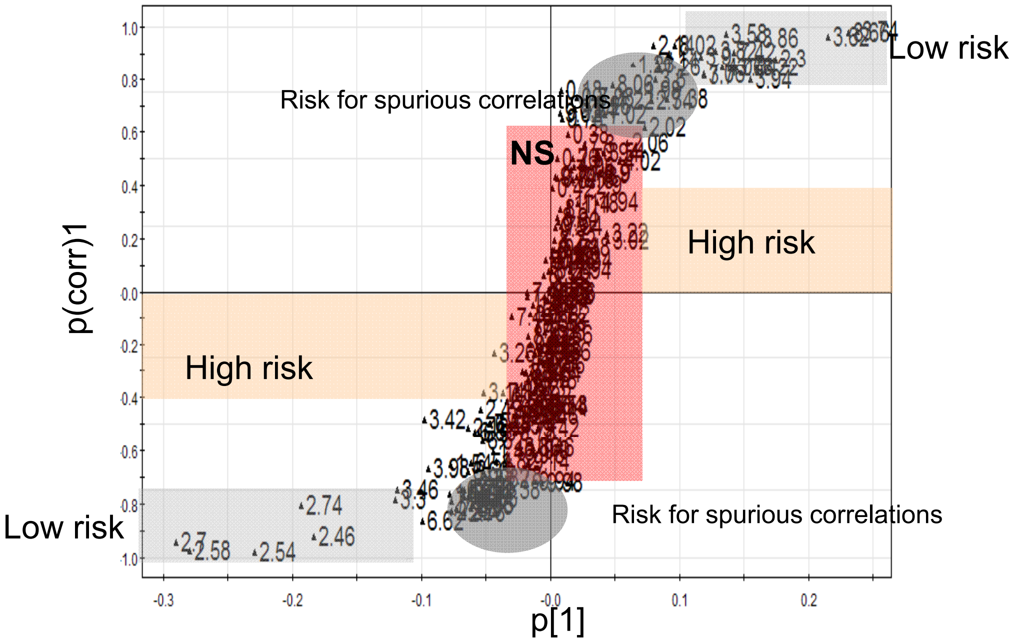




- Not significant
  - Variable in the noise

How to choose cut offs in the S-plot

No simple answer but some thumb rules are helpful.



Reference: Cohen, J., What I Have Learned (So Far). American Phychologist 1990, 45, (12), 1304-1312

Why not exact cut of limits

- Problem
  - More samples stabilize variability
    - Smaller p(corr)1 will be statistically true
    - This is the nature of variability
- Some people like to divide correlations into
  - Small -0,3 - (-)0,1 and 0,3 - 0,1
  - Medium -0,5 - (-)0,3 and 0,5 - 0,3
  - Large -1 - (-)0,5 and 1 - 0,5
  - Ref: Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.)
- p(corr)1 can be divided in a similar way
- Do not be strict, think about context and purpose

Example 2 Multi class Metabolomics

- Samples: Scraped xylem from transgenic aspen
  - Wild type, WT, N=10
  - L5, down regulated *pttPM1* gene, N=7
  - 2B, up regulated *pttPM1* gene, N=9
- Data: Resolved and integrated metabolites from Gas Chromatography Mass Spectrometry, GC/MS
- Objective
  - Class separation
  - Identify putative biomarkers
  - Identify Shared and unique structures
  - How to interpret uncorrelated variation

Reference:

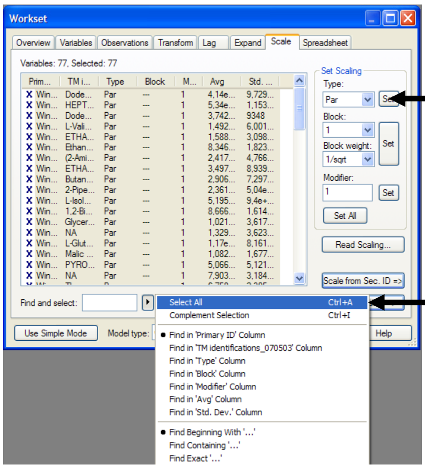
Wiklund et. al Analytical Chemistry **2008,** 80, 115-122

Visualization of GC/TOF-MS-Based Metabolomics Data for Identification of Biochemically Interesting Compounds Using OPLS Class Models

Example 2 Model settings

Workset

- Set scaling (par)
- Define the response vector Y
  - Model 1 Y=0 for Wild type Y=1 for L5
  - Model 2 Y=0 for Wild type Y=1 for 2B
- Exclude all samples from the other class
- Set model type to OPLS/O2PLS
  - Make two models



Workset

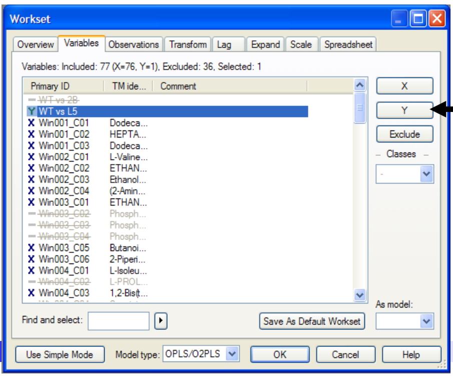
Variables: 77, Selected: 77

Prim...	TM i...	Type	Block	M...	Avg	Std. ...
X Win...	Dode...	Par	---	1	4,14e...	9,729...
X Win...	HEPT...	Par	---	1	5,34e...	1,153...
X Win...	Dode...	Par	---	1	3,742...	9348
X Win...	L-Vali...	Par	---	1	1,492...	6,001...
X Win...	ETHA...	Par	---	1	1,588...	3,098...
X Win...	Ethan...	Par	---	1	8,346...	1,823...
X Win...	(2-Ami...	Par	---	1	2,417...	4,766...
X Win...	ETHA...	Par	---	1	3,497...	8,939...
X Win...	Butan...	Par	---	1	2,906...	7,297...
X Win...	2-Pipe...	Par	---	1	2,361...	5,04e...
X Win...	L-Isol...	Par	---	1	5,195...	9,4e+...
X Win...	1,2-Bi...	Par	---	1	8,666...	1,614...
X Win...	Glycer...	Par	---	1	1,021...	3,617...
X Win...	NA	Par	---	1	1,329...	3,623...
X Win...	L-Glut...	Par	---	1	1,17e...	8,161...
X Win...	Malic ...	Par	---	1	1,082...	1,677...
X Win...	PYRO...	Par	---	1	5,066...	5,121...
X Win...	NA	Par	---	1	7,903...	3,184...

Set Scaling
Type: Par
Block: 1
Block weight: 1/sqrt
Modifier: 1

2 choose par

1 select all



Workset

Variables: Included: 77 (X=76, Y=1), Excluded: 36, Selected: 1

Primary ID	TM ide...	Comment
— ~~WT vs 2B~~		
Y WT vs L5		
X Win001_C01	Dodeca...	
X Win001_C02	HEPTA...	
X Win001_C03	Dodeca...	
X Win002_C01	L-Valine...	
X Win002_C02	ETHAN...	
X Win002_C03	Ethanol...	
X Win002_C04	(2-Amin...	
X Win003_C01	ETHAN...	
— ~~Win003_C02~~	Phosph...	
— ~~Win003_C03~~	Phosph...	
— ~~Win003_C04~~	Phosph...	
X Win003_C05	Butanoi...	
X Win003_C06	2-Piperi...	
X Win004_C01	L-Isoleu...	
— ~~Win004_C02~~	L-PROL...	
X Win004_C03	1,2-Bis(t...	

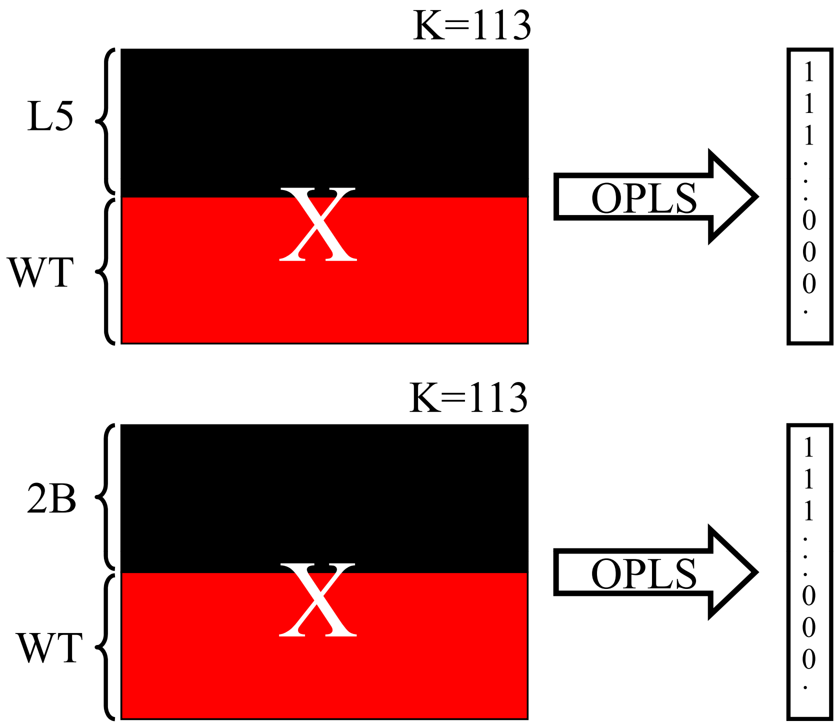
Model type: OPLS/O2PLS

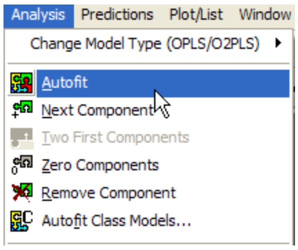
Set Y

Example 2 Multi class Metabolomics

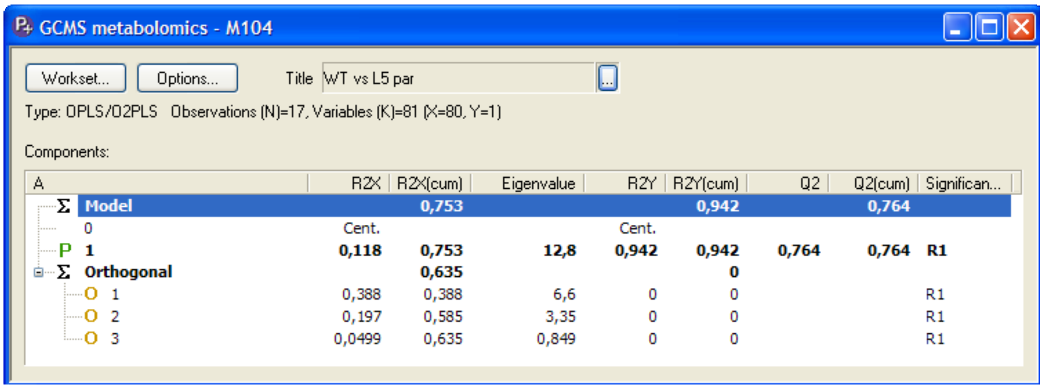
Data table of resolved and integrated metabolite from GC/MS profiles

Fit Two OPLS models one for each transgen





Example 2 Model 1 WT vs L5

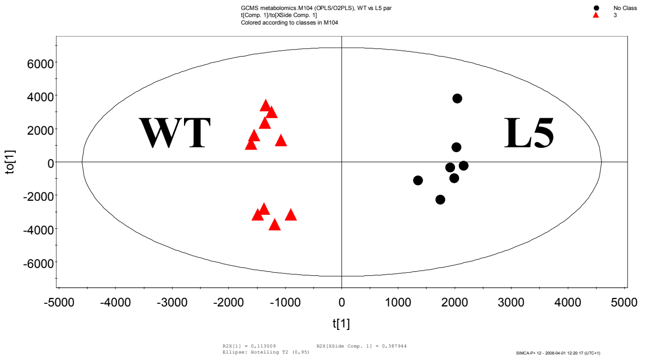


Title WT vs L5 par

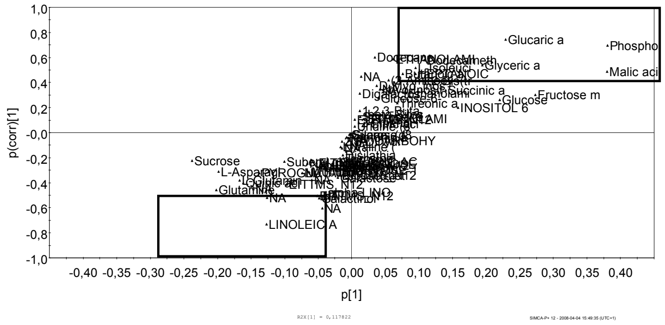
Type: OPLS/O2PLS Observations (N)=17, Variables (K)=81 (X=80, Y=1)

Components:

A		R2X	R2X(cum)	Eigenvalue	R2Y	R2Y(cum)	Q2	Q2(cum)	Significan...
Σ	Model		0,753			0,942		0,764	
	0	Cent.			Cent.				
P	1	0,118	0,753	12,8	0,942	0,942	0,764	0,764	R1
Σ	Orthogonal		0,635			0			
O	1	0,388	0,388	6,6	0	0			R1
O	2	0,197	0,585	3,35	0	0			R1
O	3	0,0499	0,635	0,849	0	0			R1



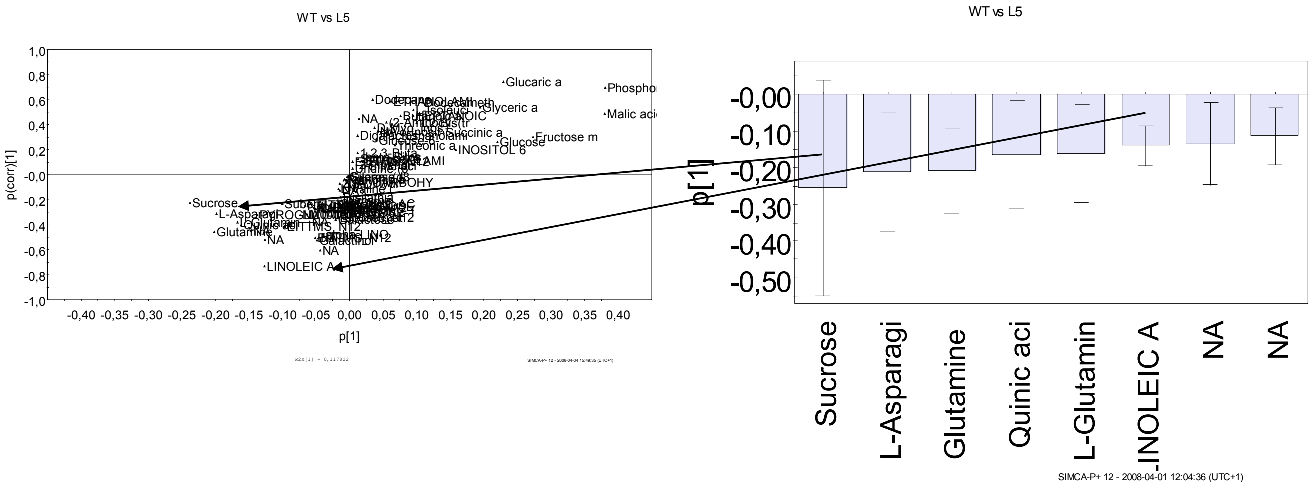
S-plot



S-plot and p1 column plot

- WT vs 5

High confidence interval=uncertain
Low confidence interval=Statistically significant



Shared and unique structure-SUS

- Comparing biomarkers from two sets (two models)
  - Which biomarkers vary in the same direction in both models= **shared structure**
  - Which biomarkers vary in a unique direction= **unique structure**
- Plotting $p(corr)_1$ from both models

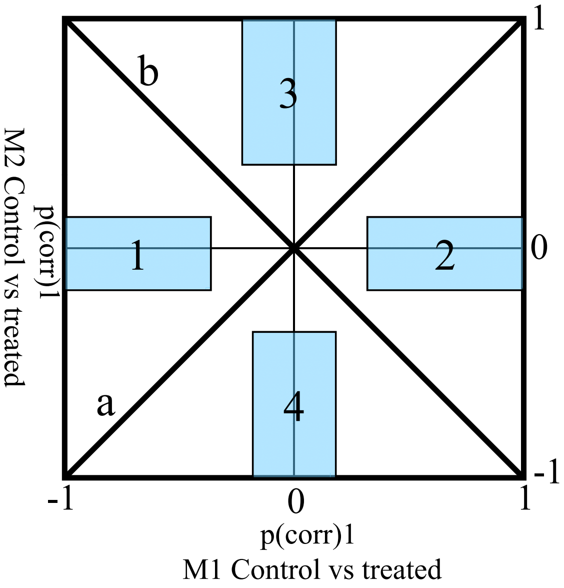
Metabolites found on:

Diagonal a=Shared structure both classes up or down

Diagonal b=Shared structure but in opposite directions.

1 and 2=Unique for M1

3 and 4=Unique for M2

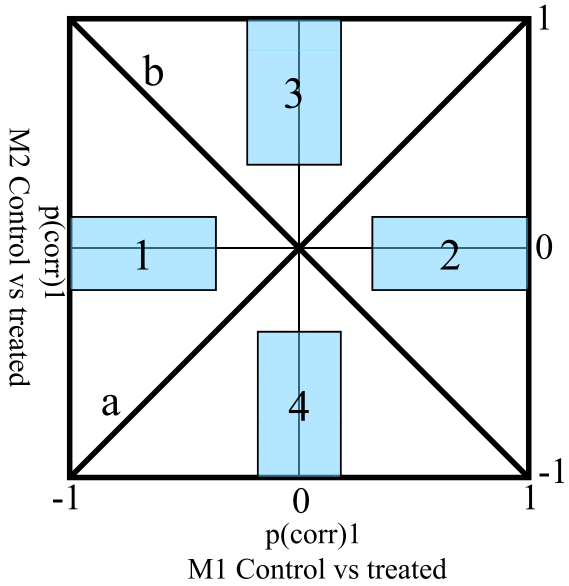


Shared and unique structure-SUS

- **a**=Shared structure both classes up or down
  - Closer to -1 and +1 the more reliable
  - Many on this diagonal implicates same effect
- **b**=Shared structure but in opposite directions.
  - Closer to -1 and +1 the more reliable
  - Many on this diagonal implicates opposite effect

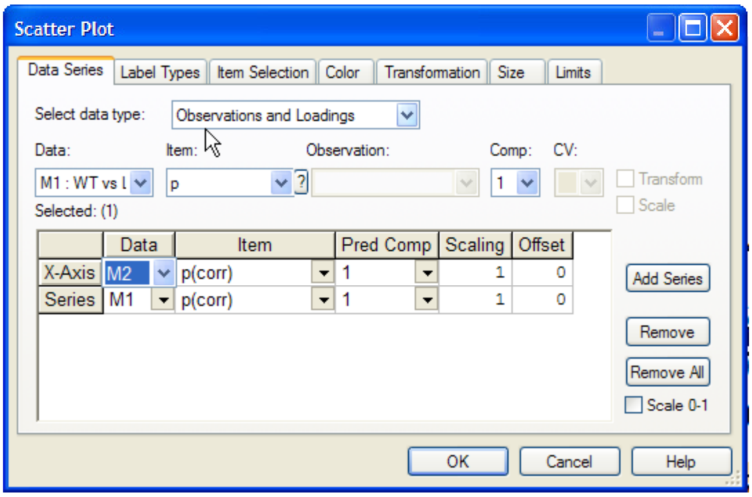
- **1 and 2**=Unique for M1
  - Biomarkers differs for control and treated 1
  - Biomarkers are similar for control and treated 2

- **3 and 4**=Unique for M2
  - Biomarkers differs for control and treated 2
  - Biomarkers are similar for control and treated 1



SUS-plot in SIMCA-P+12

- Under *Plot/lists/scatter plot*
- Select the two models to compare



Select data type: Observations and Loadings

Data: M1 : WT vs l

Item: p ?

Observation:

Comp: 1

CV:

□ Transform

□ Scale

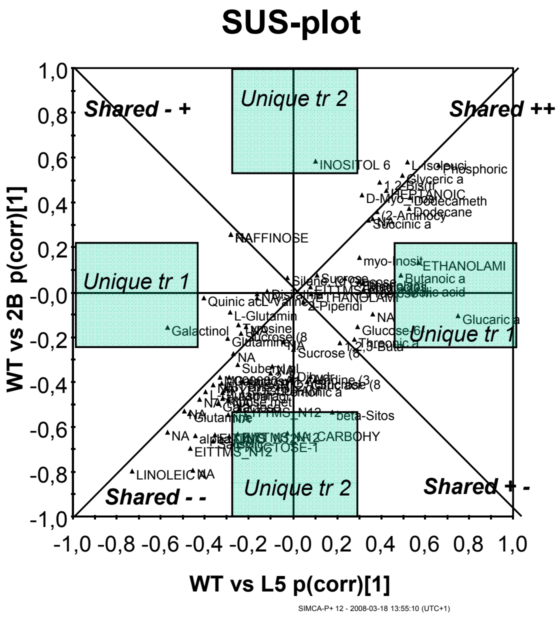
Selected: (1)

	Data	Item	Pred Comp	Scaling	Offset
X-Axis	M2	p(corr)	1	1	0
Series	M1	p(corr)	1	1	0

□ Scale 0-1

Example 2 SUS-plot

- To ad line in plot go to power point

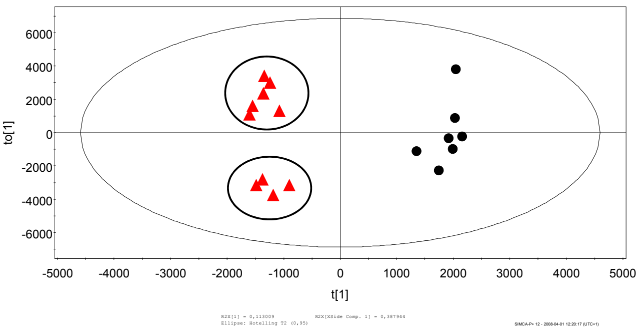


More than 3 groups

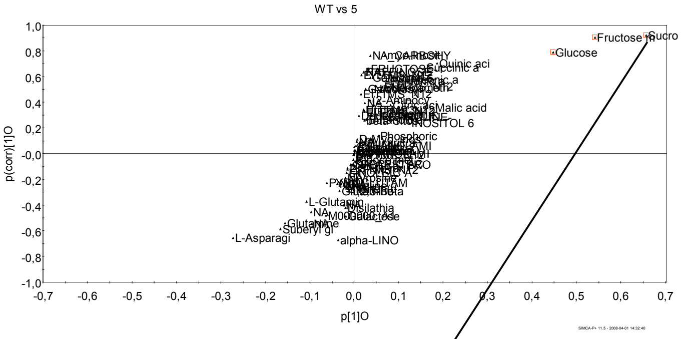
- SUS-plot works well with 2 models i.e. 3 groups
- More than 2 models gets a little bit more complicated
- 3-4 models pair wise SUS-plots
- Alternatively colour the SUS-plot by different models
- More than 4 models
  - Sorted and coloured exel lists
  - Clustering analysis

Example 2 Orthogonal variation

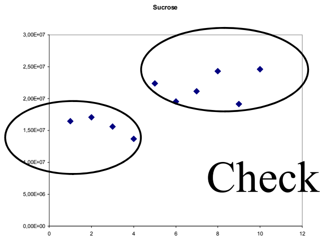
- Why is there a small class separation within the WT



S-plot orthogonal vectors



Sucrose



Check the raw data

Orthogonal S-plot
Identifies the unknown effects in X

What to report about biomarkers

- All groups have their own way of reporting results
  - Here are some suggestions
- S-plot
- SUS-plot
  - If more than two classes
- Effect size between two classes based on model
  - p(corr)1
  - Confidence interval (very important but unfortunately very often ignored)
- Effect size between two classes based on raw data
  - ratio between control and treated
  - Cohen's *d* → *d*=(M1-M2)/σ (M1=mean for group 1 and M2 mean for group 2, σ pooled std for both groups*)*
  - P-values from t-test

S-plot vs. other methods

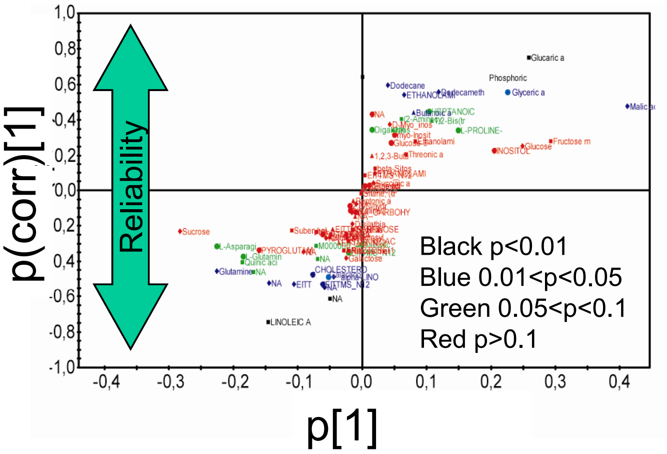
The S-plot demonstrates influence of both magnitude and reliability

Alternative methodology

1. Student’s t-test
   – Focus only on reliability (assumes t-distribution)

2. Volcano plot (common in transcriptomics)
   – t-test between two groups (reliability)
   – Plot with -log(p-value) vs. log2(fold change)

3. Permutation test
   – Focus only on reliability (no distribution assumption)
   – Test the stability of the result
   – e.g. gives a p-value on a correlation



Conclusions

- The S-plot is an easy way to visualize an OPLS classification model. It has mainly been used to filter out putative biomarkers from “omics” data e.g. NMR, GC/MS and LC/MS metabolomics data.
- The S-plot can be done using PCA or PLS-DA ONLY if a clear class separation is seen in the first component in the score plot.
- In the S-plot both magnitude (intensity) and reliability is visualised.
- We can obtain a list of potential biomarkers which are statistically significant.
- These biomarkers are **statistically significant**, but **not necessarily biochemically significant**.
- The SUS-plot is a useful tool to identify shared and unique structures from multiple classes

Multivariate workflows for “omics” analysis

Chapter 6

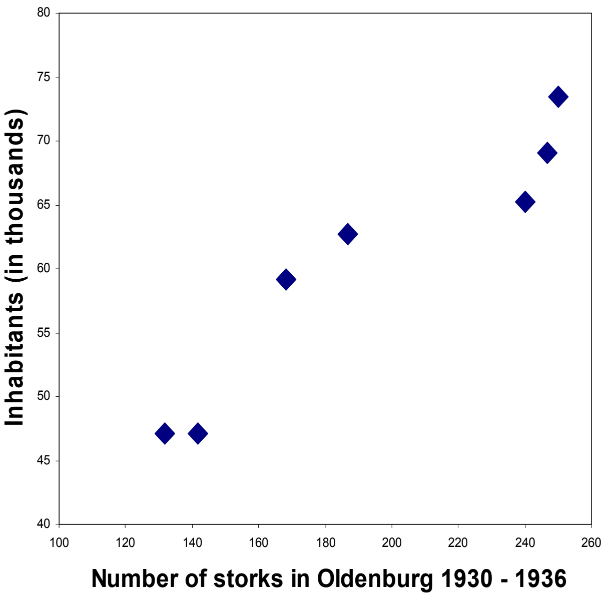
Model validation

Recall correlation vs. causation

Although the two variables are correlated, this does not imply that one causes the other!

Real but non-causal, or spurious ?

Correlation or causation?

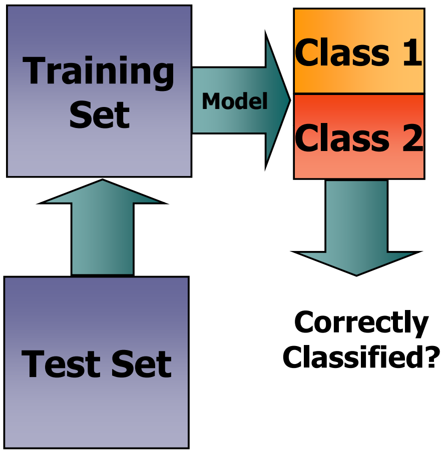


Classification Models

- Two steps:

1. Train a model on Representative data
2. Test the model using new data

- New data must try to capture all possible experimental variation to ensure robustness
- Model judged on classification success
- Step 2 is (AMAZINGLY) often forgotten!



Training Set

Model

Class 1

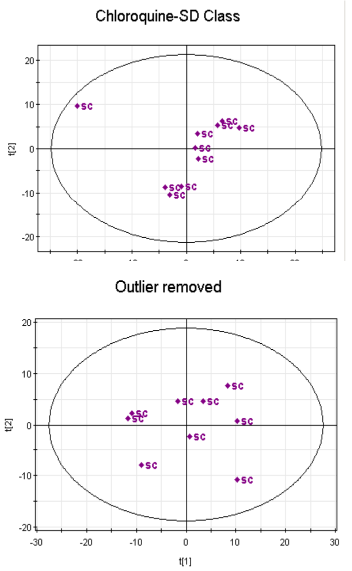
Class 2

Test Set

Correctly Classified?

Validation of Classification Models

- Always start by evaluating the training set
  - PCA of individual classes
  - PCA of all classes
  - Plot scores
  - look for patterns and trends
- Outliers may <u>seriously</u> disturb a model
  - Try to explain why
    - Incorrect values
    - Pre-processing errors
  - Remove if motivated
- Make OPLS classification model
  - Training set
- Validate model
  - Internal
  - external



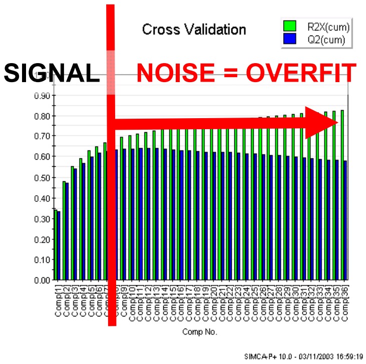
Chloroquine-SD Class

Outlier removed

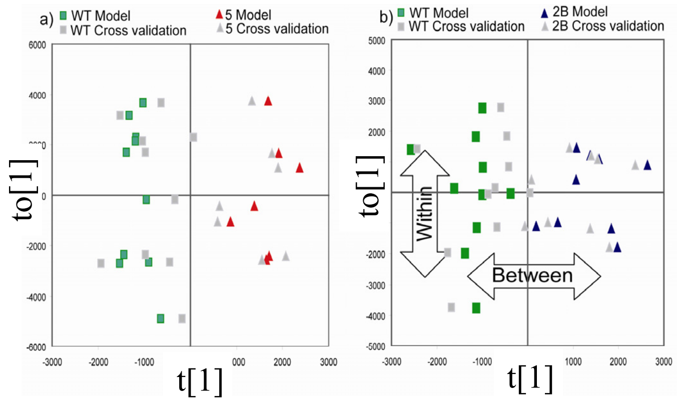
SIMCA-P+ 10.0 - 05/11/2003 10:55:53

Validation in SIMCA-P+12

1. Internal Validation
   - Cross validation Q^2 (Default 1/7)
     - OPLS full CV
     - Detects Over fit
   - Cross validated score plot
   - Outlier detection
   - Distance to model, DModX



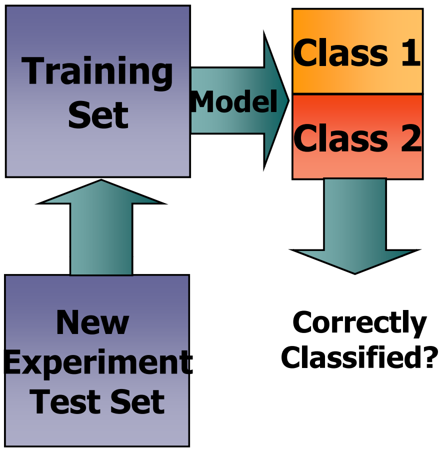
2. External Validation-test set
   - Classification (OPLS-DA)
     - Misclassification list
     - Prediction list
     - Distance to model, DModX
   - OPLS
     - Regress Obs vs Pred = Q^2 ext
     - RMSEP
     - Prediction list



Validation of Classification Models

External validation

- Ideal case
  - Repeat investigation from scratch
  - New day, new personnel
  - New individuals
  - (New spectrometer)
- Aim
  - Capture as many sources of variation as possible
  - Ensure method is robust
    - Is the classification the same?
    - Do you find the same biomarkers?


Training Set

Model

Class 1

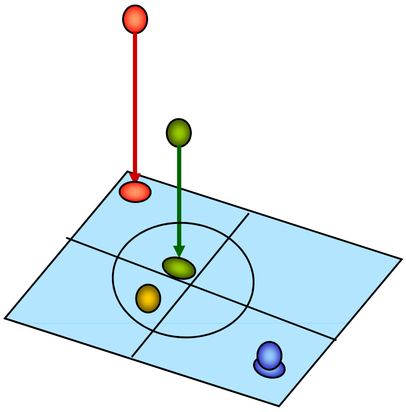
Class 2

New Experiment Test Set

Correctly Classified?

Distance to Model, DModX

- A key concept in multivariate data modelling
- How far is an observation from the model space?
  - i.e. what is unexplained by the model?
  - Residuals (what is left over)
- Used to assess similarity to training set
  - Find outliers
  - Assign class memberships
  - Warn when extrapolating

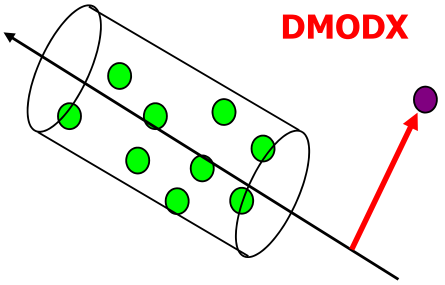


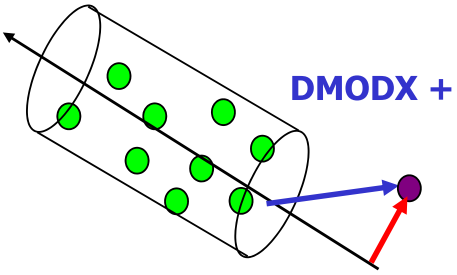
DMODX when Classifying

Two types of DMODX when making predictions

1. **DMODX** For detecting outliers
   - Called "regular" in SIMCA-P
   - Used with training set

2. **DMODX** + for Classification
   - Takes into account length of "beer can"
   - Used for prediction sets





Cross-validated score plot

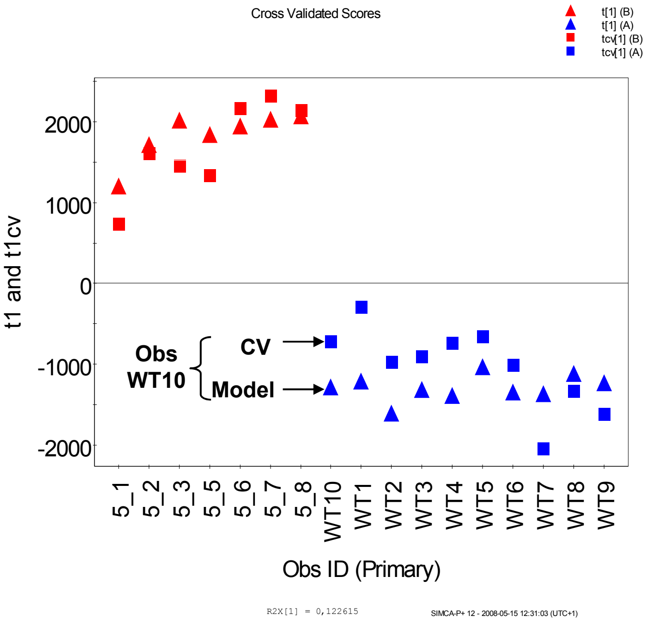
- For each observation there is one score value from the model (t1) and one from cross validation (t1cv)
- Visualize the prediction uncertainty for each observation

🟥 t1 class 5

🔺 t1cv class 5

🟦 t1 reference WT

▲ t1cv reference WT

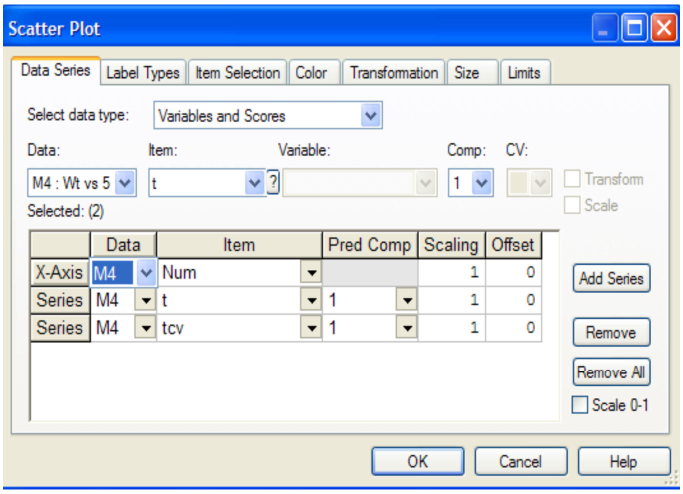


R2X[1] = 0,122615

SIMCA-P+ 12 - 2008-05-15 12:31:03 (UTC+1)

CV-Score plot

- Make this plot under *plot/list/scatter plot*



Select data type: Variables and Scores

Data: M4 : Wt vs 5

Item: t

Variable:

Comp: 1

CV:

□ Transform

□ Scale

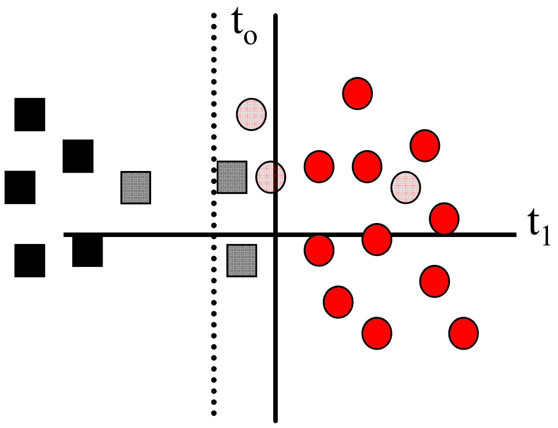
Selected: (2)

	Data	Item	Pred Comp	Scaling	Offset
X-Axis	M4	Num		1	0
Series	M4	t	1	1	0
Series	M4	tcv	1	1	0

□ Scale 0-1

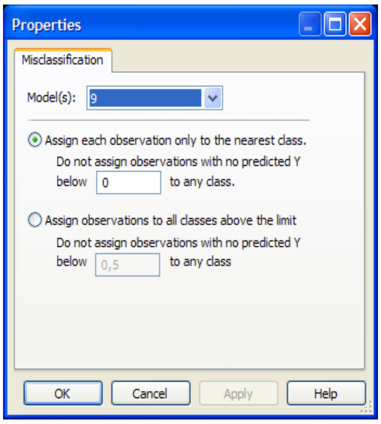
Balanced models

- Try to avoid different number of samples in each class
- The reference point will not be in the centre of the model
  - Might be misleading for predictions and interpretation



Misclassification table

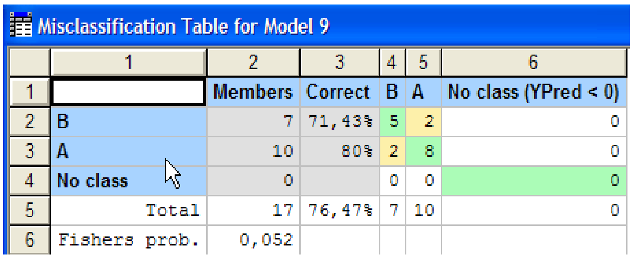
- Under *predictions/misclassification*
- Calculates the number of correct predicted
- Must define class membership in the raw data
- Must use OPLS/O2PLS-DA
- Results:
  - 71,43% of B samples correctly classified, 80% of A samples correctly classified
- Fisher’s Exact Probability gives us the likelihood of obtaining such a classification table by chance
- p=0.0027 (statistically significant because < 0.05)



Model(s): 9

● Assign each observation only to the nearest class.
Do not assign observations with no predicted Y below 0 to any class.

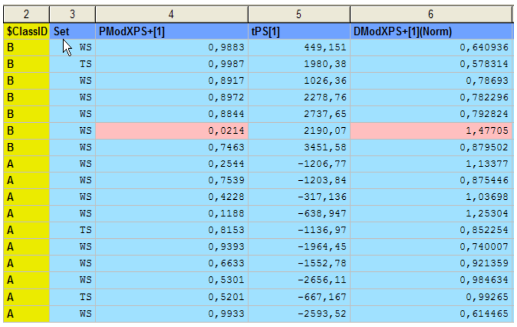
○ Assign observations to all classes above the limit
Do not assign observations with no predicted Y below 0,5 to any class


Misclassification Table for Model 9

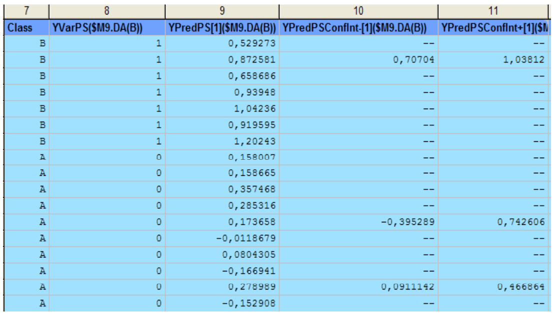
	1	2	3	4	5	6
1		Members	Correct	B	A	No class (YPred < 0)
2	B	7	71,43%	5	2	0
3	A	10	80%	2	8	0
4	No class	0		0	0	0
5	Total	17	76,47%	7	10	0
6	Fishers prob.	0,052				

Testing the success of classification

- Under *predictions/Prediction list*
- WS=observations used in the work set
- TS=Observations in the test set
- PModXPS is the probability that the observation fits the model
  - Red is a warning that the observation do not fit model
- YpredPS are the predictions from the external test set
- YpredPSConfint are the confidence intervals from YpredPS
- If no new data available leave out a proportion of the dataset
- Count correctly classified



2	3	4	5	6
$ClassID	Set	PModXPS+[1]	tPS[1]	DModXPS+[1](Norm)
B	WS	0,9883	449,151	0,640936
B	TS	0,9987	1980,38	0,578314
B	WS	0,8917	1026,36	0,78693
B	WS	0,8972	2278,76	0,782296
B	WS	0,8844	2737,65	0,792824
B	WS	0,0214	2190,07	1,47705
B	WS	0,7463	3451,58	0,879502
A	WS	0,2544	-1206,77	1,13377
A	WS	0,7539	-1203,84	0,875446
A	WS	0,4228	-317,136	1,03698
A	WS	0,1188	-638,947	1,25304
A	TS	0,8153	-1136,97	0,852254
A	WS	0,9393	-1964,45	0,740007
A	WS	0,6633	-1552,78	0,921359
A	WS	0,5301	-2656,11	0,984634
A	TS	0,5201	-667,167	0,99265
A	WS	0,9933	-2593,52	0,614465



7	8	9	10	11
Class	YVarPS($M9.DA(B))	YPredPS[1]($M9.DA(B))	YPredPSConfInt-[1]($M9.DA(B))	YPredPSConfInt+[1]($M
B	1	0,529273	--	--
B	1	0,872581	0,70704	1,03812
B	1	0,658686	--	--
B	1	0,93948	--	--
B	1	1,04236	--	--
B	1	0,919595	--	--
B	1	1,20243	--	--
A	0	0,158007	--	--
A	0	0,158665	--	--
A	0	0,357468	--	--
A	0	0,285316	--	--
A	0	0,173658	-0,395289	0,742606
A	0	-0,0118679	--	--
A	0	0,0804305	--	--
A	0	-0,166941	--	--
A	0	0,278989	0,0911142	0,466864
A	0	-0,152908	--	--

More sophisticated classification statistics

- A simple approach is "% correctly classified"

- A more in depth method is to count True and False Positives and Negatives and Calculate Sensitivity & Specificity

$$Sensitivity = \frac{100 * TP}{(TP + FN)}$$

$$Specificity = \frac{100 * TN}{(TN + FP)}$$

Trouble shooting

Why don’t I get a significant OPLS model?

1. Unlucky selection of samples during cross-validation could complicate modelling
   - Change CV settings to be balanced between classes
   - ‘Leave one out in turn’ When n < 10
   - ‘Leave two out in turn’ When n < 16
   - Build model on 2/3 predict 1/3 when n =12, 15

2. Many variables + very few samples + a lot of orthogonal variation
   - Learn from the orthogonal variation
   - Try pre-processing the data
   - If possible reduce the amount of pure noise
   - If possible do more experiments

3. Worst case, no predictive variation exist
   - Learn from previous results and try again

Summary

- Always start the analysis by evaluating the raw data
  - PCA of individual classes
  - PCA of all classes
  - look for trends and patterns
- Outlier detection is very important
- Validate models using
  - Full cross validation
  - Cross validated score plots
  - Prediction list
- Most importantly is external validation
  - Repeat investigation from scratch
  - Same biomarker?
  - Same classification?

Summary

- Many different methods for classification

- The advantage with OPLS-DA is that the method is good both for classification and interpretation of the data
  - Biomarkers
  - Diagnostics
  - Predictions

- All classification methods must be tested
  - DModX
  - CV score-plot
  - Misclassification table
  - ROC curves

Multivariate workflows for "omics" analysis

Concluding remarks

and some additional useful slides

Remeber

- Always start to evaluate the data by PCA
- Continue with OPLS-DA and biomarker identification
- Correlation≠Causation
- Biomarkers selected by ANY type of statistical method will ONLY be statistically significant, not necessarily biologically significant

Types of classification methods

- OPLS-DA (Orthogonal PLS - Discriminant Analysis) Multivariate equivalent of PLS. Works with many groups but the interpretation is difficult with more than 2 groups. Interpretation is facilitated by OPLS-DA
- PLS-DA (PLS - Discriminant Analysis) Multivariate equivalent of LDA, works with many X's and less than 6 groups and 'tight' classes
- SIMCA (Soft Independent modelling by class analogy) Multivariate pattern recognition method for many groups and asymmetric "one class" problems
- LDA (Linear Discriminant Analysis) works only for a few independent X's and less than 6 groups
- Other methods (not discussed here)
  - PCA-LDA (Exactly equivalent to PLS-DA but takes 2 steps)
  - KNN (Nearest Neighbours), SVM (Support vector machines)
  - NN (Neural Networks), PDF (Probability density functions)

Classification

- In SIMCA-P+ there are two recommended multivariate options:

- OPLS-DA
  - Predictions works for many groups
  - Interpretation is easy with 2 groups
  - OPLS model built on class membership (0 or 1)
  - Maximum separation projection
  - Shows *differences between groups as a whole*

- SIMCA (Soft Independent Modelling by Class Analogy, not described in this course)
  - For many groups
  - Local PCA model for each class
  - Good for “fuzzy” groups with overlap
  - Lacks information of why groups differ

For reference: Alternative Classification Measures

- Many alternative measures of classification used by the “Machine Learning” community

- Kappa Statistic
  - Compares classification vs chance
  - 100% = perfect 0% = Chance
  - Quite a nice statistic

$$Kappa = \frac{p(correct) - p(expected)}{1 - p(expected)}$$

- MCC

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN) \times (TN + FP) \times (TN + FN)}}$$

- F-measure
  - It is a harmonic mean
  - Recall = Sensitivity
  - Precision = Predictive power of positive test (% Correct "in" class)

$$Fmeasure = 2 * \frac{Precision \times Recall}{Precision + Recall}$$

For reference: Confusion Matrix

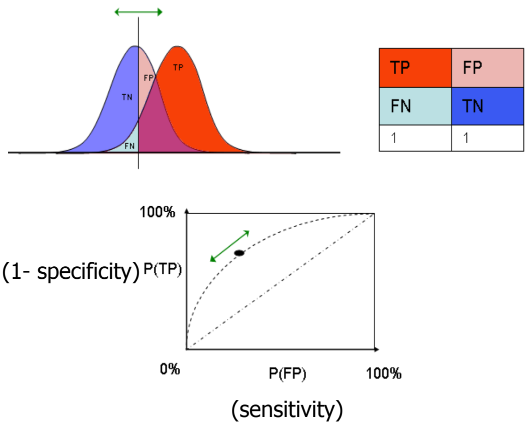
	Predicted Active	Predicted Inactive	Row Total	% Correct	
Active	TP	FN	TP + FN	$\frac{100*TP}{(TP + FN)}$	Sensitivity
Inactive	FP	TN	FP + TN	$\frac{100*TN}{(TN + FP)}$	Specificity
Column total	TP + FP	FN + TN			
% Correct	$\frac{100*TP}{(TP+FP)}$	$\frac{100*TN}{(FN+TN)}$		$\frac{100*(TP+TN)}{(TP+TN+FP+FN)}$	
	Predictive Power of positive test	Predictive Power of negative test			Total % predicted correct

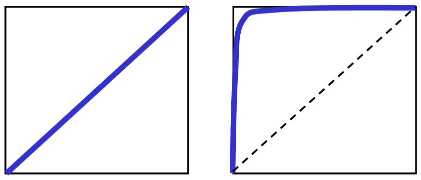
aka. 'Recall'

aka. 'Precision'

ROC Curves

- Receiver Operating Characteristic (ROC curve)
- Graphical plot of the fraction of TP (sensitivity) to fraction of FP (1-specificity) for a binary classifier system as its discrimination threshold is varied
- For SIMCA an ROC curve may be made by selecting different probability cut-offs
- For OPLS-DA (see later) an ROC curve may be made by moving the discriminating threshold

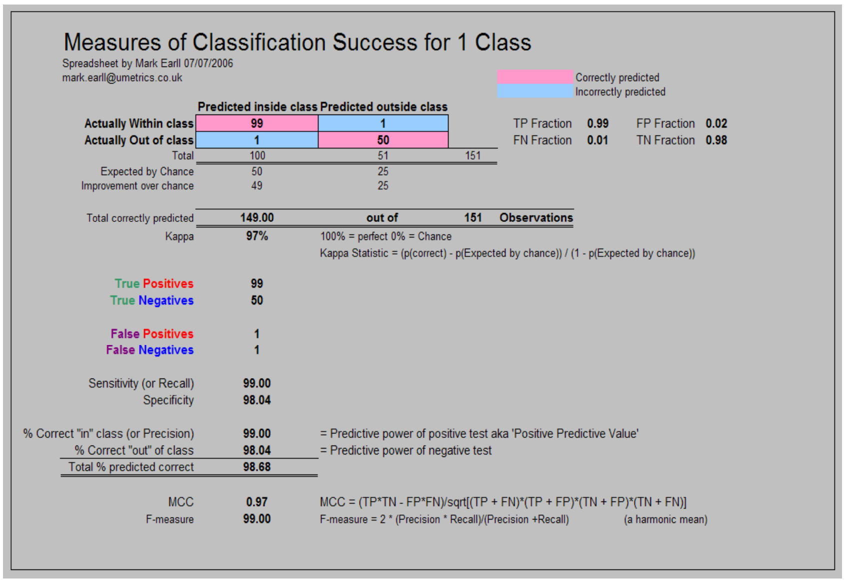




Random predictor   Perfect predictor

Example of classification stats

- Classification stats are easily encoded in Excel
- Which measure you use determined by personal preference or community!



Measures of Classification Success for 1 Class

Spreadsheet by Mark Earll 07/07/2006
mark.earll@umetrics.co.uk

🟥 Correctly predicted
🟦 Incorrectly predicted

	Predicted inside class	Predicted outside class	
Actually Within class	99	1	
Actually Out of class	1	50	
Total	100	51	151
Expected by Chance	50	25	
Improvement over chance	49	25	

TP Fraction	0.99	FP Fraction	0.02
FN Fraction	0.01	TN Fraction	0.98

Total correctly predicted 149.00 out of 151 Observations

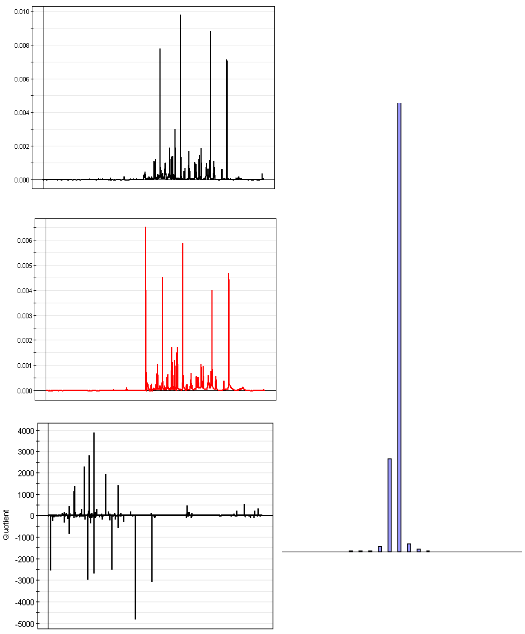
Kappa 97% 100% = perfect 0% = Chance
Kappa Statistic = (p(correct) - p(Expected by chance)) / (1 - p(Expected by chance))

True Positives	99	
True Negatives	50	
False Positives	1	
False Negatives	1	
Sensitivity (or Recall)	99.00	
Specificity	98.04	
% Correct "in" class (or Precision)	99.00	= Predictive power of positive test aka 'Positive Predictive Value'
% Correct "out" of class	98.04	= Predictive power of negative test
Total % predicted correct	98.68	
MCC	0.97	MCC = (TP*TN - FP*FN)/sqrt[(TP + FN)*(TP + FP)*(TN + FP)*(TN + FN)]
F-measure	99.00	F-measure = 2 * (Precision * Recall)/(Precision +Recall) (a harmonic mean)

UMETRICS

Probabilistic Quotient Normalisation

- Basis
  - Assumes the intensity of a majority of signals is a function of dilution only
  - Divide the spectrum by a reference, the most probable quotient is used as a normalisation factor
  - i.e. the most common value
- Procedure
  - Integral Normalisation
  - Calculate Reference spectrum (mean or median of controls)
  - Divide spectrum by reference spectrum
  - Calculate median quotient
  - Divide all variables of spectrum by this median value



Reference: Probablistic Quotient normalisation as a robust method to account for dilution of complex biological mixtures
Dieterle F, Ross A, Schlotterbeck,G, Senn H, Analytical Chemistry 2006 Jul 1;78(13):4281-90

UMETRICS

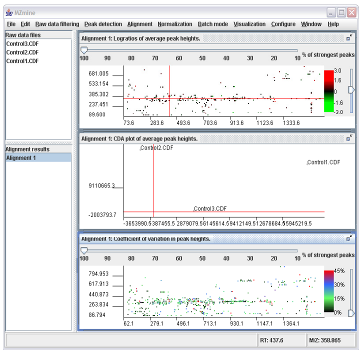
NMR pre-processing software

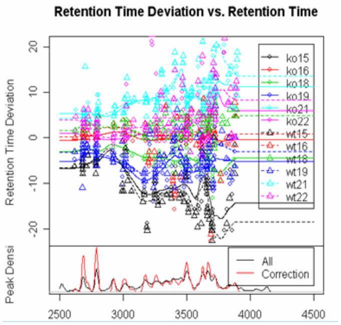
- Brucker Amix2
  - Binning and Referencing
  - Exports CSV
- ACD NMR SpecManager
  - Standard binning
  - Intelligent bucketing (recognises peaks) BUT need to do all test and training sets together!
- Chenomx Targeted Profiling
  - Extensive NMR library of known metabolites
  - Semi-automated synthesis of mixture spectrum from reference spectra
  - Frequency flexibility for shifting peaks
  - Output is in the form of a peak table, akin to chromatography
- R algorithms
  - Many open source routines becoming available
- A number of algorithms appearing for peak alignment
  - Polynomial time warping
  - Semi-parametric warping (Eilers 2007)

UMETRICS

Alignment of LC-MS data

- Alignment by mass matching or time windowing
- Several software packages available:
  - Commercial:
    - Marker-Lynx (Waters)
    - Metalign (Plant Research International)
    - ACD IntelliXtract
    - ABI Metabonomics macro
  - Open source:
    - Java based “MZmine” (VTT, Finland)
    - R-Based XCMS (Scripps)
    - Sashimi project
    - Open MS (C++ proteomics MS)





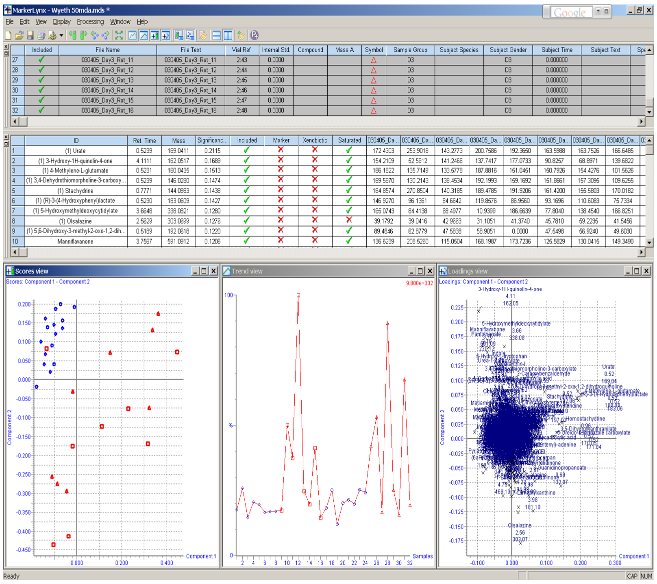
C.A. Smith, E.J. Want, G.C. Tong, A. Saghatelian, B.F. Cravatt, R. Abagyan, and G. Siuzdak.
Metlin XCMS: Global metabolite profiling incorporating LC/MS filtering, peak detection, and novel non-linear retention time alignment with open-source software.
53rd ASMS Conference on Mass Spectrometry, June 2005, San Antonio Texas

URL’s
http://mzmine.sourceforge.net/index.shtml
http://sashimi.sourceforge.net/software.html
http://metlin.scripps.edu/download/

UMETRICS

Waters Marker Lynx

- Add-in to Waters Mass Lynx software
- Finds main peaks and aligns based on m/z
- Thresholding function
- Includes basic PCA
- Has data export to SIMCA-P and EZInfo for more detailed analysis



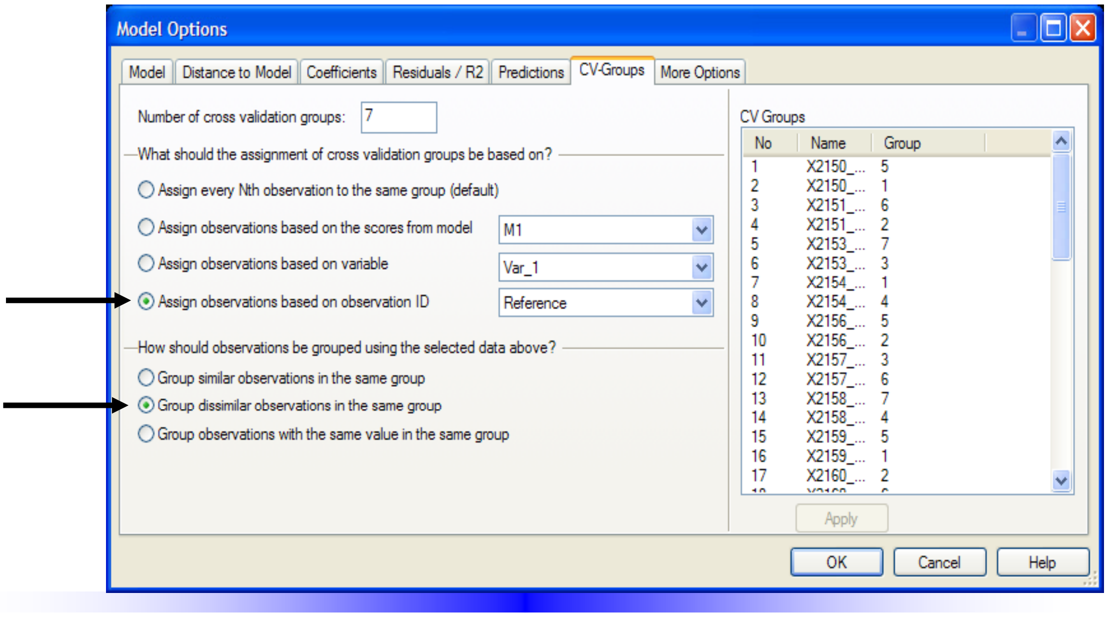
MarkerLynx - Wyeth 50mda.mds *

	Included	File Name	File Text	Vial Ref.	Internal Std.	Compound	Mass A	Symbol	Sample Group	Subject Species	Subject Gender	Subject Time	Subject Text
27	✓	030405_Day3_Rat_11	030405_Day3_Rat_11	2:43	0.0000			△	D3		D3	0.000000	
28	✓	030405_Day3_Rat_12	030405_Day3_Rat_12	2:44	0.0000			△	D3		D3	0.000000	
29	✓	030405_Day3_Rat_13	030405_Day3_Rat_13	2:45	0.0000			△	D3		D3	0.000000	
30	✓	030405_Day3_Rat_14	030405_Day3_Rat_14	2:46	0.0000			△	D3		D3	0.000000	
31	✓	030405_Day3_Rat_15	030405_Day3_Rat_15	2:47	0.0000			△	D3		D3	0.000000	
32	✓	030405_Day3_Rat_16	030405_Day3_Rat_16	2:48	0.0000			△	D3		D3	0.000000	

	ID	Ret. Time	Mass	Significanc...	Included	Marker	Xenobiotic	Saturated	030405_Da...	030405_Da...	030405_Da...	030405_Da...	030405_Da...	030405_Da...	030405_Da...	030405_Da...
1	(1) Urate	0.5239	169.0411	0.2115	✓	✗	✗	✓	172.4303	253.9018	143.2773	200.7586	192.3650	163.5988	163.7526	166.6485
2	(1) 3-Hydroxy-1H-quinolin-4-one	4.1111	162.0517	0.1689	✓	✗	✗	✓	154.2109	52.5912	141.2466	137.7417	177.0733	90.8257	68.8971	139.6822
3	(1) 4-Methylene-L-glutamate	0.5231	160.0435	0.1513	✓	✗	✗	✓	166.1822	135.7149	133.5778	187.8816	151.0451	150.7926	154.4276	101.5626
4	(1) 3,4-Dehydrothiomorpholine-3-carboxy...	0.5239	146.0280	0.1474	✓	✗	✗	✓	169.5870	130.2143	138.4534	192.1993	159.1692	151.8661	157.3095	109.6255
5	(1) Stachydrine	0.7771	144.0983	0.1438	✓	✗	✗	✓	164.8574	270.8504	140.3185	189.4785	191.9206	161.4200	155.5803	170.0182
6	(1) (R)-3-(4-Hydroxyphenyl)lactate	0.5230	183.0609	0.1427	✓	✗	✗	✓	146.9270	96.1361	84.6642	119.8576	86.9560	93.1696	110.6083	75.7334
7	(1) 5-Hydroxymethyldeoxycytidylate	3.6648	338.0821	0.1280	✓	✗	✗	✓	165.0743	84.4138	68.4977	10.9399	186.6639	77.8040	138.4540	166.8251
8	(1) Olsalazine	2.5629	303.0699	0.1276	✓	✗	✗	✗	39.1792	39.0416	42.9663	31.1051	41.3740	45.7810	59.2235	41.5456
9	(1) 5,6-Dihydroxy-3-methyl-2-oxo-1,2-dih...	0.5189	192.0618	0.1220	✓	✗	✗	✓	89.4846	62.8779	47.5838	58.9051	0.0000	47.5498	56.9240	49.6030
10	Manniflavanone	3.7567	591.0912	0.1206	✓	✗	✗	✓	136.6239	208.5260	115.0504	168.1987	173.7236	125.5829	130.0415	149.3490

Balanced cross-validation in SIMCA-P+12 (advanced)

- Use unfitted data set
  - Will not work for a model that is already fitted
- Go to *work set/model options*



Model Options

Number of cross validation groups: 7

What should the assignment of cross validation groups be based on?

- ○ Assign every Nth observation to the same group (default)
- ○ Assign observations based on the scores from model: M1
- ○ Assign observations based on variable: Var_1
- ● Assign observations based on observation ID: Reference

How should observations be grouped using the selected data above?

- ○ Group similar observations in the same group
- ● Group dissimilar observations in the same group
- ○ Group observations with the same value in the same group

CV Groups

No	Name	Group
1	X2150_...	5
2	X2150_...	1
3	X2151_...	6
4	X2151_...	2
5	X2153_...	7
6	X2153_...	3
7	X2154_...	1
8	X2154_...	4
9	X2156_...	5
10	X2156_...	2
11	X2157_...	3
12	X2157_...	6
13	X2158_...	7
14	X2158_...	4
15	X2159_...	5
16	X2159_...	1
17	X2160_...	2

MVDA-Exercise FOODS

The European food consumption pattern

Background

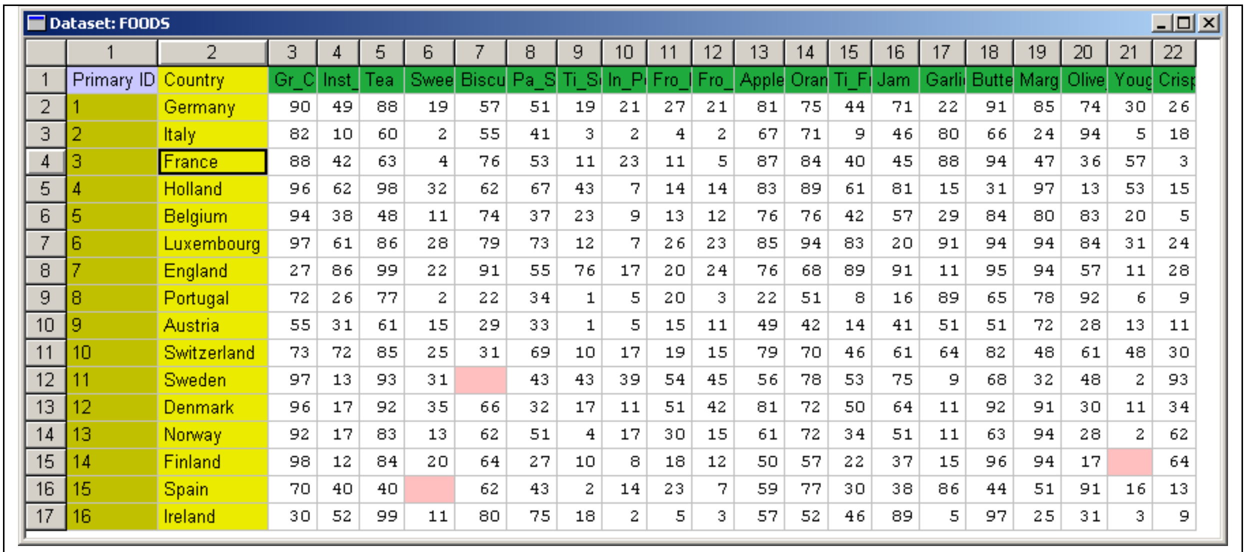
Data were collected to investigate the consumption pattern of a number of provisions in different European countries. The purpose of the investigation was to examine similarities and differences between the countries and the possible explanations.

Objective

You should learn how to initiate a new project in SIMCA, import data and make the first projections. You should also be able to explain why there are groupings in the plots. Data characteristics that differentiate Portugal and Spain from Sweden and Denmark should be discussed.

Data

The data set consists of 20 variables (the different foods) and 16 observations (the European countries). The values are the percentages of households in each country where a particular product was found. For the complete data table, see below. This table is a good example of how to organise your data.



Dataset: FOODS

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
1	Primary ID	Country	Gr_C	Inst_	Tea	Swee	Biscu	Pa_S	Ti_S	In_P	Fro_	Fro_	Apple	Oran	Ti_F	Jam	Garli	Butte	Marg	Olive	Youg	Crisp
2	1	Germany	90	49	88	19	57	51	19	21	27	21	81	75	44	71	22	91	85	74	30	26
3	2	Italy	82	10	60	2	55	41	3	2	4	2	67	71	9	46	80	66	24	94	5	18
4	3	France	88	42	63	4	76	53	11	23	11	5	87	84	40	45	88	94	47	36	57	3
5	4	Holland	96	62	98	32	62	67	43	7	14	14	83	89	61	81	15	31	97	13	53	15
6	5	Belgium	94	38	48	11	74	37	23	9	13	12	76	76	42	57	29	84	80	83	20	5
7	6	Luxembourg	97	61	86	28	79	73	12	7	26	23	85	94	83	20	91	94	94	84	31	24
8	7	England	27	86	99	22	91	55	76	17	20	24	76	68	89	91	11	95	94	57	11	28
9	8	Portugal	72	26	77	2	22	34	1	5	20	3	22	51	8	16	89	65	78	92	6	9
10	9	Austria	55	31	61	15	29	33	1	5	15	11	49	42	14	41	51	51	72	28	13	11
11	10	Switzerland	73	72	85	25	31	69	10	17	19	15	79	70	46	61	64	82	48	61	48	30
12	11	Sweden	97	13	93	31		43	43	39	54	45	56	78	53	75	9	68	32	48	2	93
13	12	Denmark	96	17	92	35	66	32	17	11	51	42	81	72	50	64	11	92	91	30	11	34
14	13	Norway	92	17	83	13	62	51	4	17	30	15	61	72	34	51	11	63	94	28	2	62
15	14	Finland	98	12	84	20	64	27	10	8	18	12	50	57	22	37	15	96	94	17		64
16	15	Spain	70	40	40		62	43	2	14	23	7	59	77	30	38	86	44	51	91	16	13
17	16	Ireland	30	52	99	11	80	75	18	2	5	3	57	52	46	89	5	97	25	31	3	9

Tasks

Task 1

Create a new project in SIMCA by importing the data from FOODS.XLS (*File/New*). Make sure that the entire data set has been imported: 16 observations and 20 variables. Are there any missing values in the data set?

Task 2

Analyse the data table according to the following procedure: Run PCA on the data set with all observations and variables included. Compute three principal components with *Analysis|Autofit*. Look at the score plots found under *Analysis|Scores|Scatter plot* for t_2 vs. t_1 and t_3 vs. t_1. Are there detectable groupings? Change the plot mark to the observation name with the right mouse button using *Properties|Label Types|Use Identifier*. Produce the corresponding loading plots: p_2 vs. p_1 and p_3 vs. p_1, using *Analysis|Loadings|Scatter plot*. Which variables are responsible for the groupings?

Task 3

Projection models are robust. Make a new PC model (*Workset|New as Model*) and see what happens with the model structure if you remove an influential observation like Sweden. Also remove an influential variable, for example garlic. Compare the results with those from Task 2.

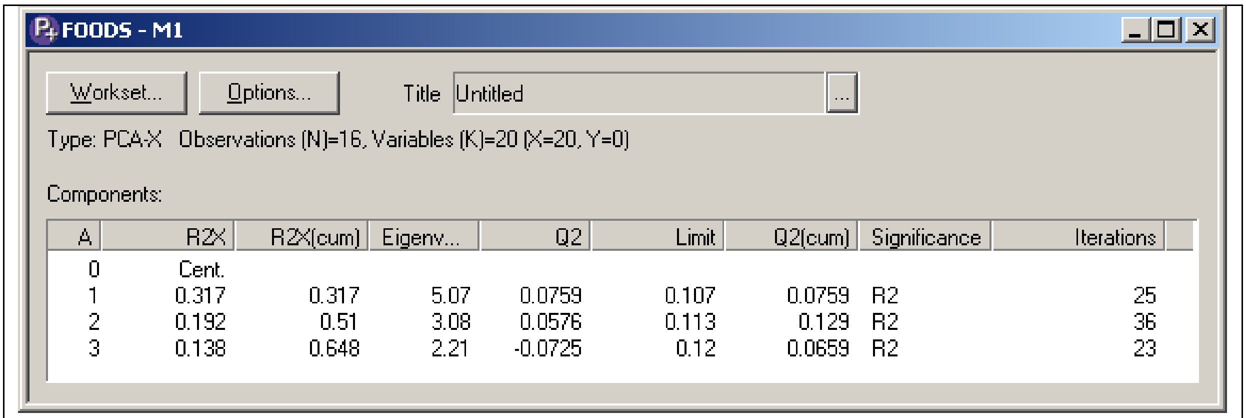
Solutions to FOODS

Task 1

There were 3 missing values.

Task 2

A three component PC model was computed:

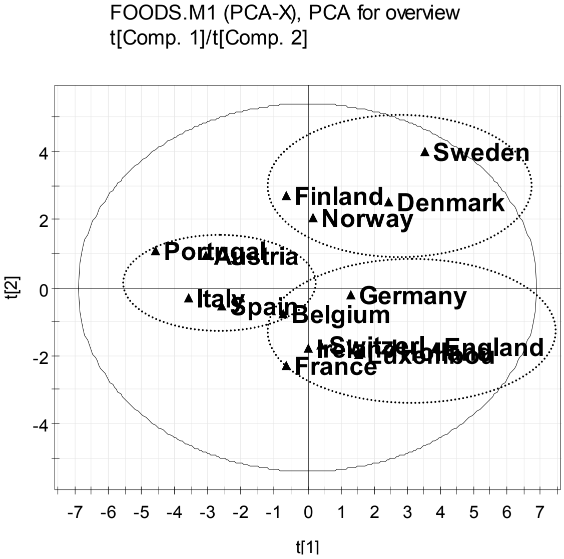


Title Untitled

Type: PCA-X Observations (N)=16, Variables (K)=20 (X=20, Y=0)

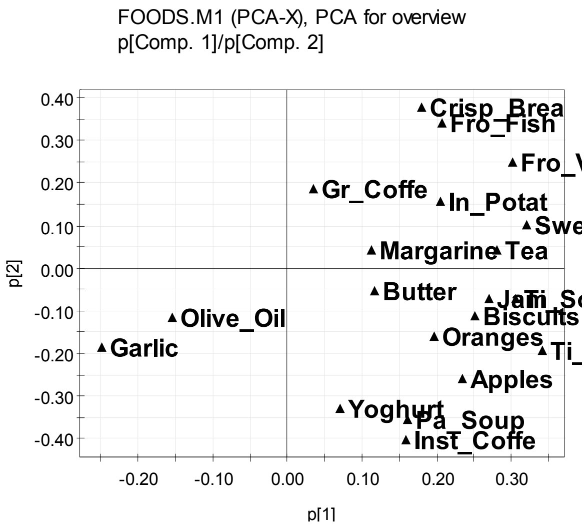
Components:

A	R2X	R2X(cum)	Eigenv...	Q2	Limit	Q2(cum)	Significance	Iterations
0	Cent.							
1	0.317	0.317	5.07	0.0759	0.107	0.0759	R2	25
2	0.192	0.51	3.08	0.0576	0.113	0.129	R2	36
3	0.138	0.648	2.21	-0.0725	0.12	0.0659	R2	23



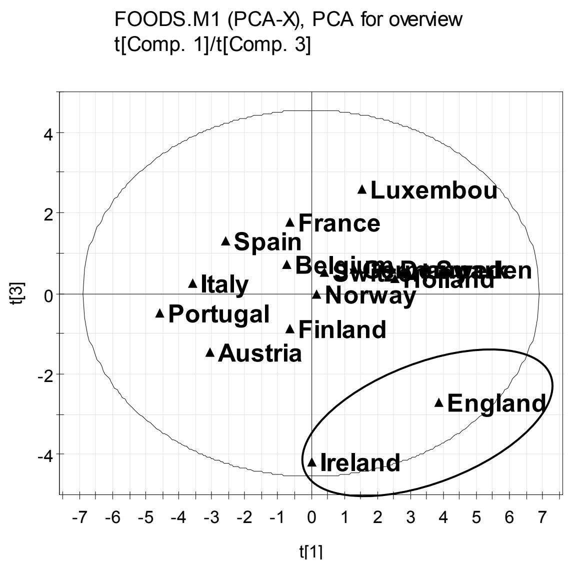
The two first components position the central European countries in the lower central region of the score plot. The southern Mediterranean countries are found in the left-hand region and the Scandinavian countries in the upper right-hand portion of the score plot.

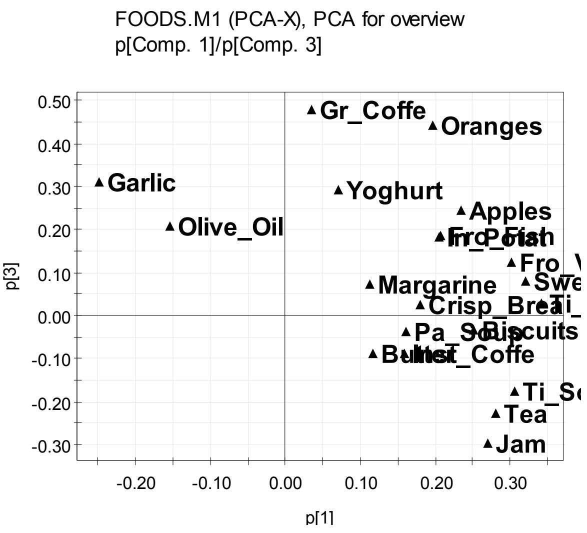
(PLEASE NOTE: The ellipses drawn around the groups are for illustration and are not available within SIMCA-P)



The corresponding loading plot shows garlic and olive oil in one discriminating group of variables. These two provisions are often consumed in the Mediterranean countries. Crisp bread and frozen fish are eaten extensively in the Scandinavian countries while the central European countries drink instant coffee and eat powder soup (Pa_soup).

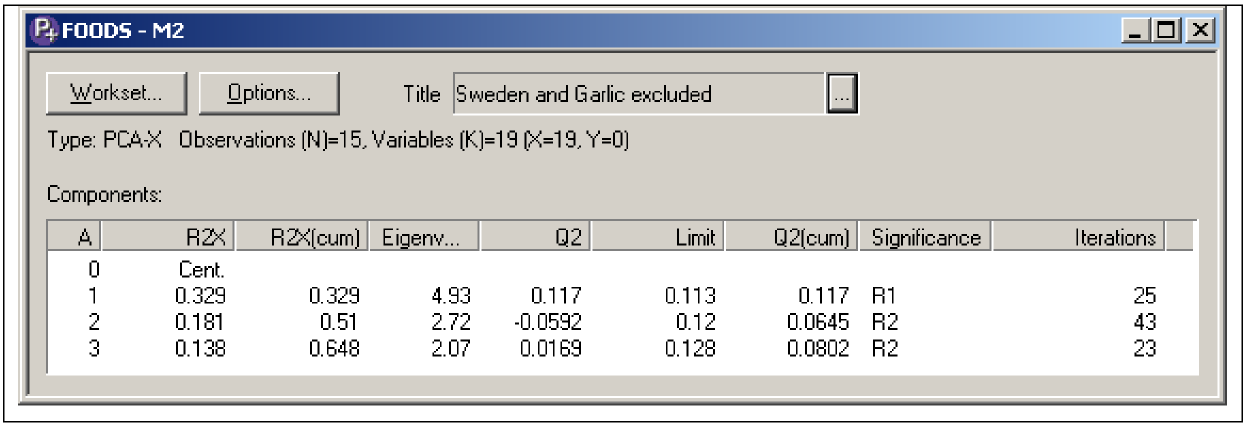
The third component separates England and Ireland from the rest of Europe. We can see the presence of the tea and jam habit, as well as the limited consumption of ground coffee, garlic, and olive oil on these islands.





Task 3

A new model was made with Sweden and Garlic excluded.



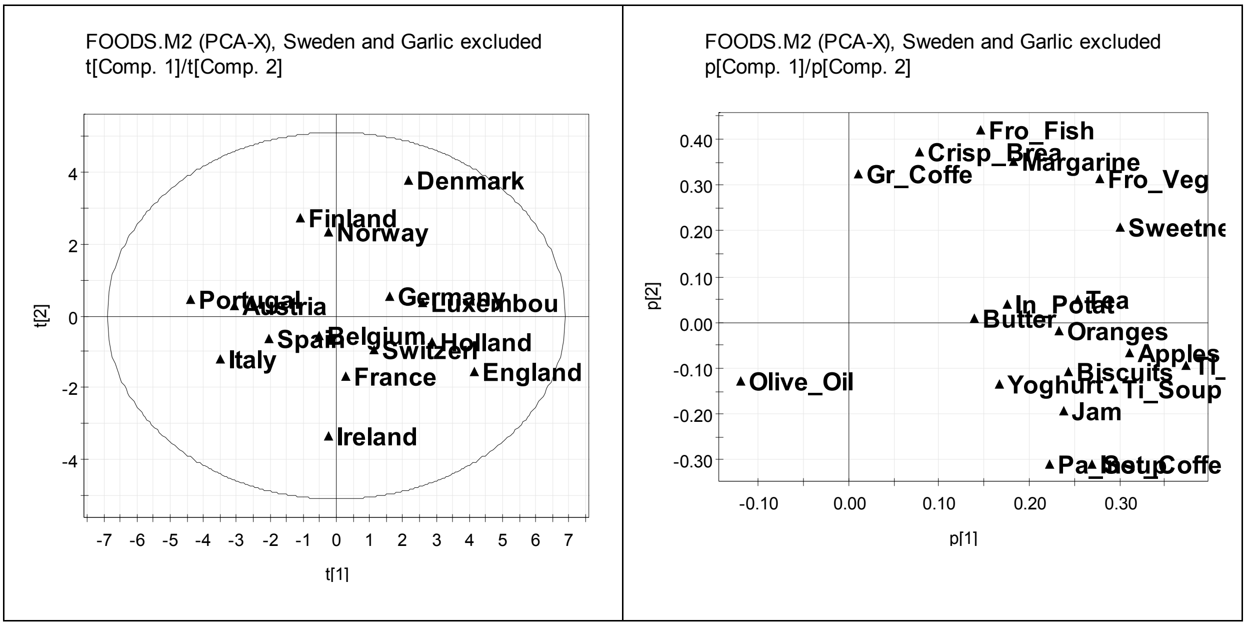
Title: Sweden and Garlic excluded

Type: PCA-X Observations (N)=15, Variables (K)=19 (X=19, Y=0)

Components:

A	R2X	R2X(cum)	Eigenv...	Q2	Limit	Q2(cum)	Significance	Iterations
0	Cent.							
1	0.329	0.329	4.93	0.117	0.113	0.117	R1	25
2	0.181	0.51	2.72	-0.0592	0.12	0.0645	R2	43
3	0.138	0.648	2.07	0.0169	0.128	0.0802	R2	23

We here show plots pertaining to the two first components.



FOODS.M2 (PCA-X), Sweden and Garlic excluded
t[Comp. 1]/t[Comp. 2]

FOODS.M2 (PCA-X), Sweden and Garlic excluded
p[Comp. 1]/p[Comp. 2]

Despite removing what seemed to be a dominating observation and an influential variable, the pictures obtained in Task 3 are very similar to those of Task 2. This is because the information removed (Sweden & Garlic) was not unique. Similar information is expressed by many variables and many observations because of the correlation pattern among them.

Conclusions

Groupings among the observations in a data set are often found in the first score plot. These groupings can be explained by investigating the corresponding loading plot. The main differences between, on one hand, Portugal and Spain, and, on the other, Sweden and Denmark, are high consumption of frozen food and crisp bread in the Scandinavian countries, and high consumption of olive oil and garlic in the Mediterranean countries.

MVDA-Exercise METABONOMICS

A Metabonomic Investigation of Phospholipidosis

Background

Metabolites are the products and by-products of the many complex biosynthesis and catabolism pathways that exist in humans and other living systems. Measurement of metabolites in human biofluids has often been used for the diagnosis of a number of genetic conditions, diseases and for assessing exposure to xenobiotics. Traditional analysis approaches have been limited in scope in that emphasis was usually placed on one or a few metabolites. For example urinary creatinine and blood urea nitrogen are commonly used in the diagnosis of renal disease.

Recent advances in (bio-)analytical separation and detection technologies, combined with the rapid progress in chemometrics, have made it possible to measure much larger bodies of metabolite data [1]. One prime example is when using NMR in the monitoring of complex time-related metabolite profiles that are present in biofluids, such as, urine, plasma, saliva, etc. This rapidly emerging field is known as Metabonomics. In a general sense, metabonomics can be seen as the investigation of tissues and biofluids for changes in metabolite levels that result from toxicant-induced exposure. The exercises below describe multivariate analysis of such data, more precisely ^{1}H-NMR urine spectra measured on different strains of rat and following dosing of different toxins.

Objective

The example data set deals with male rats treated with the drugs chloroquine or amiodarone, both of which are known to induce phospholipidosis, here coded as “c” or “a”. The drugs were administered to two different strains of rat, i.e., Sprague-Dawley and Fischer, here coded as “s” or “f”. Sprague-Dawley rats represent a standard laboratory animal model whereas Fishers rats are more susceptible to certain types of drug exposure and hence it is easier to detect drug effects. The experimental objective was to investigate whether ^{1}H-NMR data measured on rat urine samples could be used to distinguish control rats and animals subject to toxin exposure. The objective of this exercise is to shed some light on how PCA may be used in state-of-the-art Metabonomics. This exercise will continue with OPLS-DA for biomarker identification and with comparing of multiple treatments.

Data

In total, the data set contains N = 57 observations (rats) and K = 194 variables (^{1}H-NMR chemical shift region integrals). The observations (rats) are divided in six groups (“classes”):

• Control Sprague-Dawley (s), 10 rats,	**“s”**
• Sprague-Dawley treated with amiodarone (sa), 8 rats	**“sa”**
• Sprague-Dawley treated with chloroquine (sc), 10 rats	**“sc”**
• Control Fisher (f), 10 rats	**“f”**
• Fisher treated with amiodarone (fa), 10 rats	**“fa”**
• Fisher treated with chloroquine (fc), 9 rats	**“fc”**

The urine ^{1}H NMR spectra were reduced by summation of all the data points over a 0.04 ppm region. Data points between 4.5- 6.0 ppm, corresponding to water and urea resonances, were excluded, leaving a total of 194 NMR spectral regions as variables for the multivariate modelling. A more elaborate account of the experimental conditions are found in [2]. We are grateful to Elaine Holmes and Henrik Antti of Imperial College, London, UK, for giving us access to this data set.

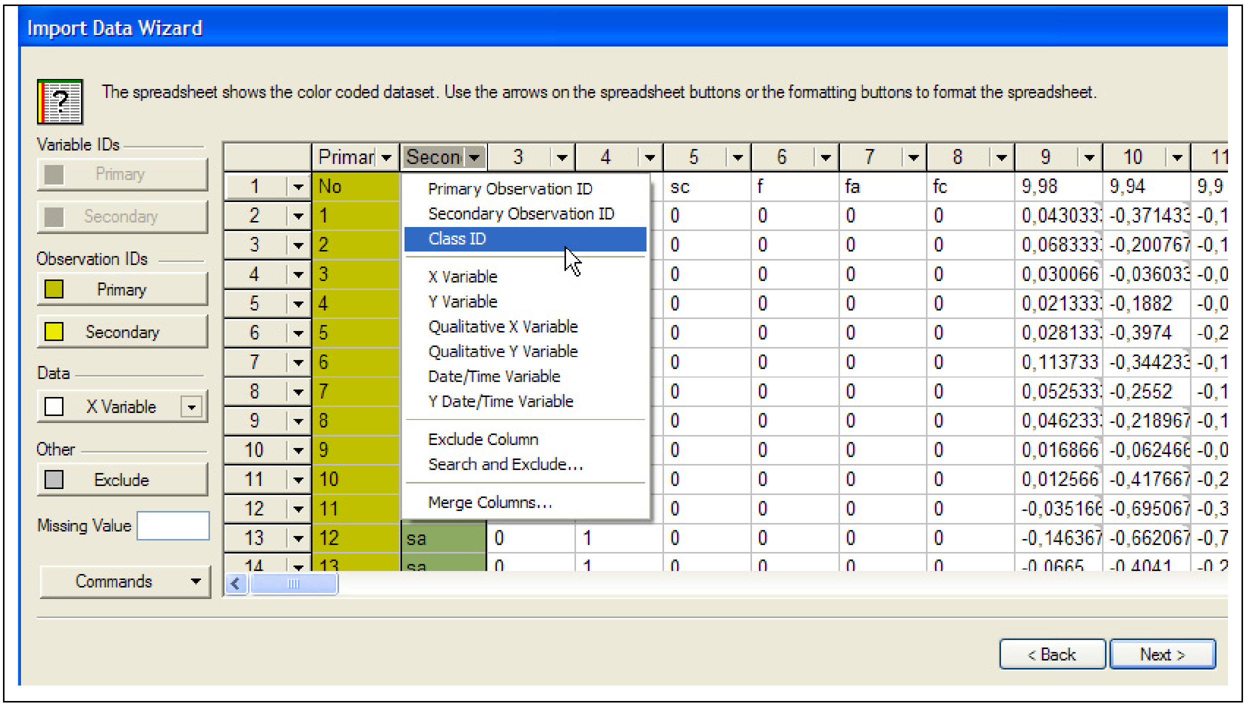
1) Nicholson, J.K., Connelly, J., Lindon, J.C., and Holmes, E., Metabonomics: A Platform for Studying Drug Toxicity and Gene Function, Nature Review, 2002; 1:153-161. 2) J.R. Espina, W.J. Herron, J.P. Shockcor, B.D. Car, N.R. Contel, P.J. Ciaccio, J.C. Lindon, E. Holmes and J.K. Nicholson. Detection of in vivo Biomarkers of Phospholipidosis using NMR-based Metabonomic Approaches. Magn. Resonance in Chemistry 295: 194-202 2001.

Tasks

Task 1

Create a new project in SIMCA by importing the data from *File/New* and select *METABONOMICS_coded.DIF.* The second column in the data set is labelled *ClassID*. Assigns this column to a *Class ID* and specify the length of the identification as 2. Accept the Primary Observation ID.

To define a Primary Variable ID, select the first row then select Primary Variable ID. This first row is equivalent to the chemical shift regions in the NMR-spectra. Column 3 to 8 are Y variables describing the different classes. These Y variables will not be used in this exercise but in next exercise using OPLS-DA for biomarker identification. These columns should be excluded in all tasks.



The spreadsheet shows the color coded dataset. Use the arrows on the spreadsheet buttons or the formatting buttons to format the spreadsheet.

	Primar	Secon	3	4	5	6	7	8	9	10	11
1	No				sc	f	fa	fc	9,98	9,94	9,9
2	1				0	0	0	0	0,043033:	-0,371433	-0,1
3	2				0	0	0	0	0,068333:	-0,200767	-0,1
4	3				0	0	0	0	0,030066	-0,036033	-0,0
5	4				0	0	0	0	0,021333:	-0,1882	-0,0
6	5				0	0	0	0	0,028133:	-0,3974	-0,2
7	6				0	0	0	0	0,113733	-0,344233	-0,1
8	7				0	0	0	0	0,052533:	-0,2552	-0,1
9	8				0	0	0	0	0,046233:	-0,218967	-0,1
10	9				0	0	0	0	0,016866	-0,062466	-0,0
11	10				0	0	0	0	0,012566	-0,417667	-0,2
12	11				0	0	0	0	-0,035166	-0,695067	-0,3
13	12	sa	0	1	0	0	0	0	-0,146367	-0,662067	-0,7
14	13	sa	0	1	0	0	0	0	-0,0665	-0,4041	-0,2

Press *Next*, and verify the entire data set has been imported: 57 observations (rats) and 194 variables (chemical shift region integrals). Are there any missing values in the data set? Press *Finish*. When *Class ID* is defined, SIMCA will identify these different classes after import of data. SIMCA will automatically generate separate work sets for each class. These work sets will be under model CM1.

Task 2

Generally, when working with spectral data it is recommended to work with non-scaled ('Ctr') data. However a disadvantage of not scaling is that only those chemical shift regions with large variation in signal amplitude will be seen. Pareto-scaling can be seen as a compromise between UV-scaling and no scaling as it enhances the contribution from medium sized features without inflating the noise from 'quiet' areas of the spectrum. For NMR data Pareto-scaling and mean-centering are a good choice for over viewing the information in the data set using PCA.

To Pareto-scale and mean-center the data, follow these steps: *Work set/Edit*, select CM1 the *Scale* tab, and mark all the variables. Under *Set Scaling* select all variables and *"Par"*, press *Set* (By default "Par" scaling automatically mean-centers the data). Press *OK.* Exclude all class variables i.e. column 3-8. Now the data is ready to fit the principal component model.

Compute an overview of each class in the data set. Are the groups homogenous, can you detect any outliers?

Task 3

Compute an overview PCA model on the entire data set. Create the necessary scores-, loadings, and DModX-plots and interpret the model. What do you see? Are there any groupings consistent with strain of rat? Toxin exposure? Are there any outliers?

Task 4

It should be noted that other comparisons might be made rather than just "s" with "sa". Other ways of focusing on drug effects are to compare "f" ⇒ "fa", "f" ⇒ "fc", and "s" ⇒ "sc". However, there are also other aspects of the data analysis, which may reveal interesting information. For example, a comparison made between "f" ⇒ "s" would indicate rat differences and perhaps diet differences. And looking at "fa" ⇒ "sa" and "fc" ⇒ "sc" might suggest species dependent drug effects.

You may experiment with any of these combinations.

There is no solution provided to this task.

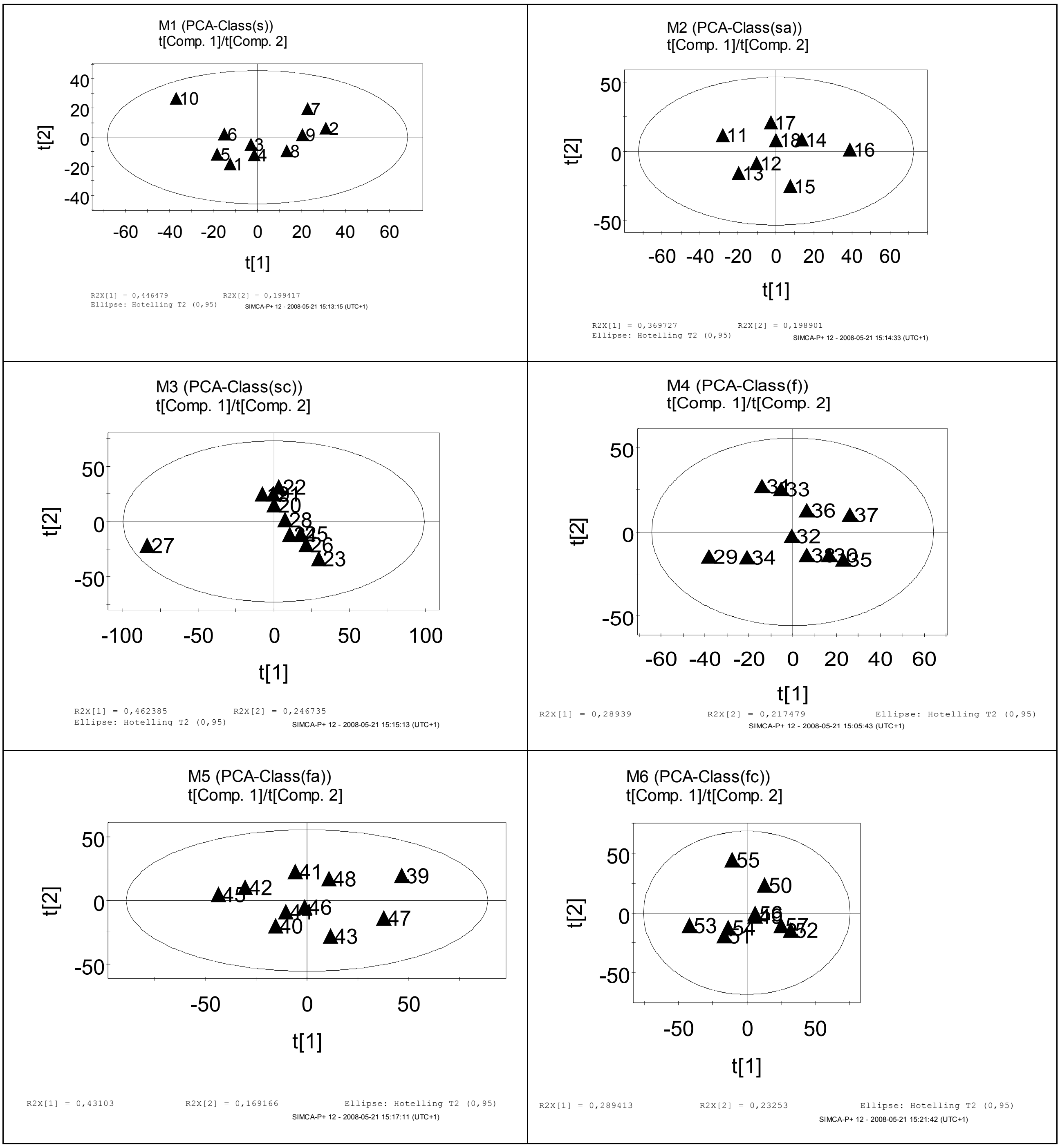
Solutions to METABONOMICS

Task 1

There are no missing data.

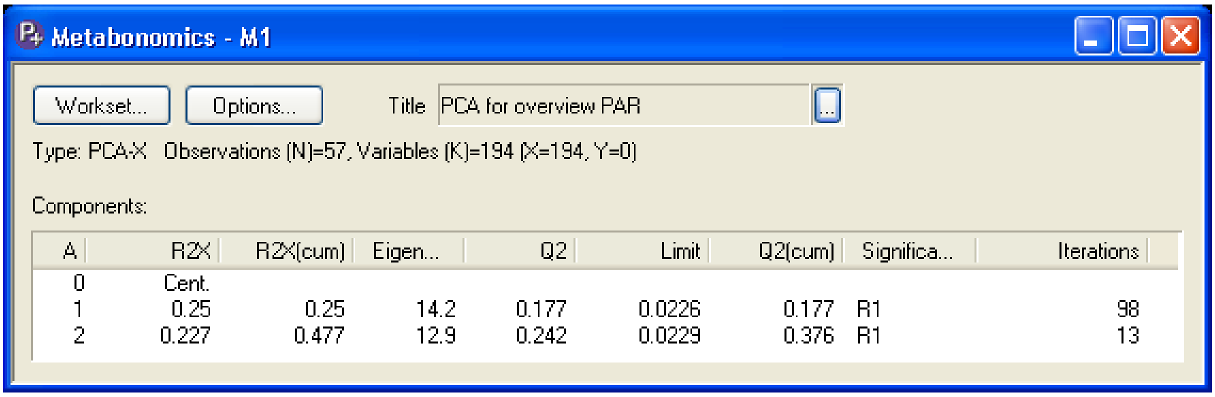
Task 2

The SC class showed one potential outlier nr 27.



Task 3

For an overview model, usually only the two first components are extracted. In this case, these showed the performance statistics $R^2X = 0.48$ and $Q^2X = 0.38$.



Metabonomics - M1

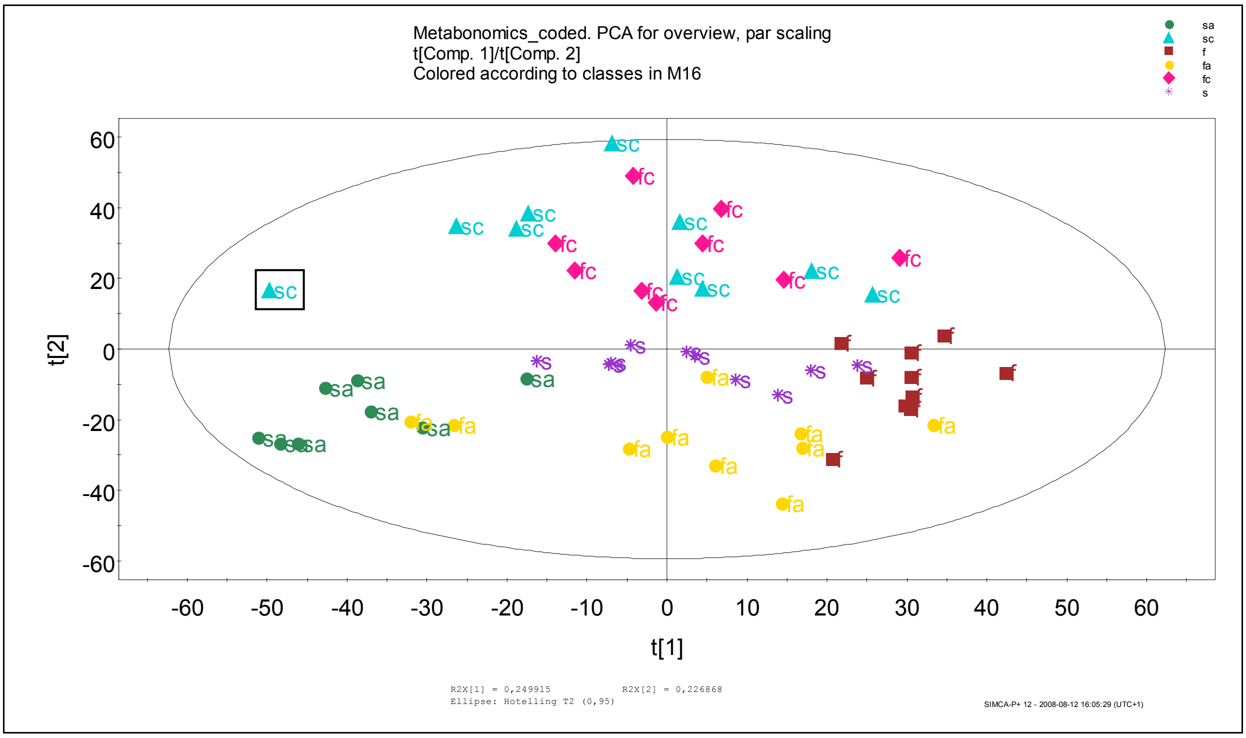
Title: PCA for overview PAR

Type: PCA-X   Observations (N)=57, Variables (K)=194 (X=194, Y=0)

Components:

A	R2X	R2X(cum)	Eigen...	Q2	Limit	Q2(cum)	Significa...	Iterations
0	Cent.							
1	0.25	0.25	14.2	0.177	0.0226	0.177	R1	98
2	0.227	0.477	12.9	0.242	0.0229	0.376	R1	13

The plot below shows the scores of these two components.

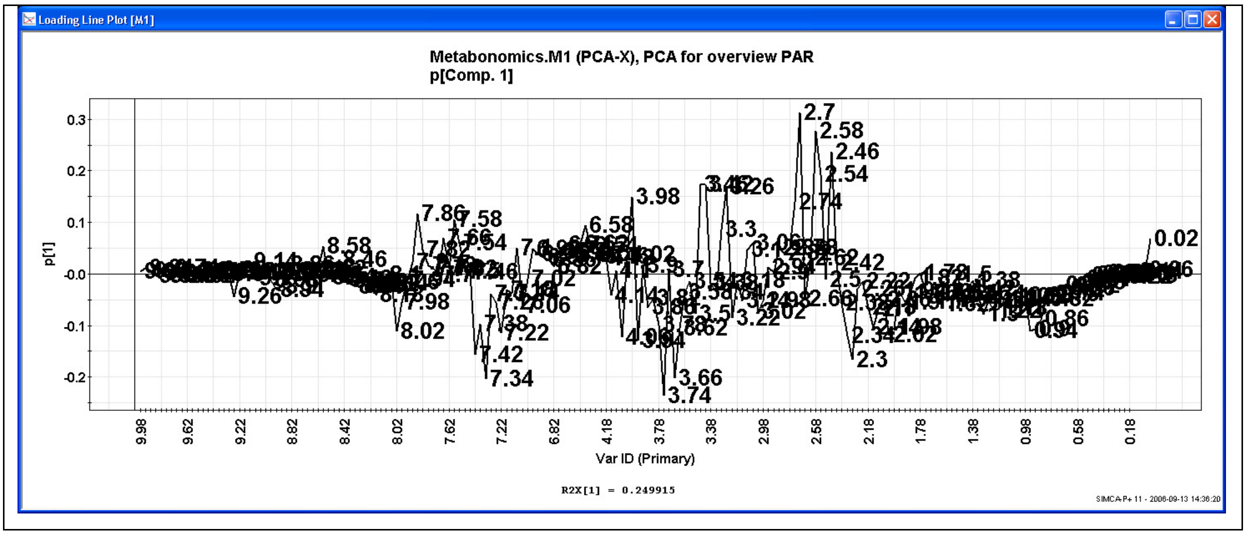


We can see that all the chloroquine-treated animals are positioned in the top part of the plot, whereas the majority of the amiodarone-treated rats are found in the bottom part. All controls are located in the central, predominantly right-hand, part of the plot. Hence, the second principal component reflects differences in the effect of the two drugs.

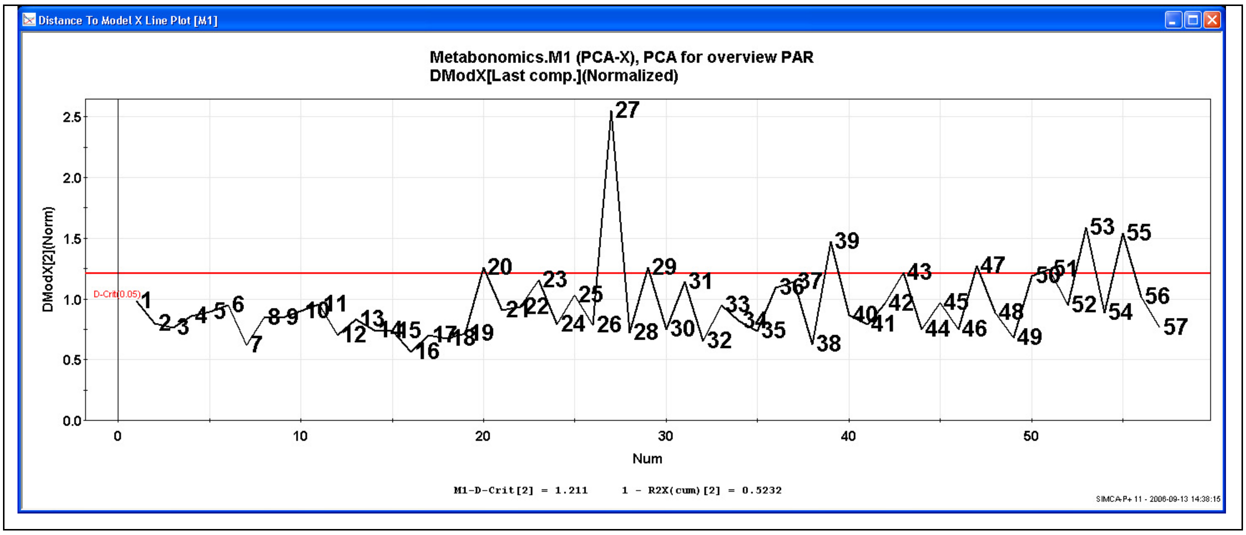
As seen, this score plot is not identical to the original one. We may take advantage of the ClassID to modify this plot regarding color, markers, etc. To accomplish this, right-click in the plot and choose *properties/Label Types tab*, Use *Identifier/obsID($ClassID)*, and press *Apply*. Next you select the *Color tab/Coloring type/By Identifier/color by ClassID*. Then you can assign any colour to the classes.

Going back to the interpretation of the score plot, an interesting discovery is that the "f"-groups tend to be "right-shifted" along the first principal component in comparison with the corresponding "s"-groups. This makes us interpret the first PC as a "difference-between-type-of-rat"-scale.

In order to interpret the scores we use the loadings. The next figure displays a line plot of the first loading spectrum. This spectrum highlights the various chemical shift regions contributing to the formation of the first score vector. For instance, the Fischer rats generally tend to have higher peaks at chemical shifts 2.46, 2.54, 2.58, 2.70 etc., and lower peaks at shifts 2.30, 3.66, 3.74, and 7.34., etc., regardless of chemical treatment. If a similar loading spectrum is plotted for the second loading vector, it is possible to identify which spectral variables reflect the major differences in NMR data following exposure to either amiodarone or chloroquine.



Moreover, it is interesting to examine the model residuals (see DModX plot below). The DModX plot reveals one very different "sc"-rat with a DModX-value exceeding the critical distance by a factor of 2. When tracing this information back to the previous score plot, we realize that this animal is the remotely positioned sc-rat (marked with the open frame). This is an observation with unique NMR-data and its spectrum should be more carefully inspected to understand where the differences arise. These differences could be due to some very interesting change in metabolic pattern, or be due to experimental variation in the handling of the rats, or perhaps a data transfer error. One way to pinpoint the likely cause for this discrepancy in DModX is through the loading plot or a contribution plot, but that option is not further exploited here.



It is obvious from the above PCA model that the observations (rats) are grouped according to treatment in the score plot. However, knowledge related to class membership is not used to find the location of the principal components. The PC-model is calculated to approximate the observations as well as possible. It must be realized that PCA finds the directions in multivariate space that represent the largest sources of variation, the so-called principal components. However, it is not necessarily the case that these maximum variation directions coincide with the maximum separation directions among the classes. Rather, it may be that other directions are more pertinent for discriminating among classes of observations (here: NMR spectra or rats).

It is in this perspective that a PLS or OPLS/O2PLS based technique, called PLS discriminant analysis (PLS-DA) or orthogonal-PLS-DA (OPLS/O2PLS-DA), becomes interesting. These methods are described in the next exercise. These methods make it possible to accomplish a rotation of the projection to give latent variables that focus on class separation ("discrimination"). The method offers a convenient way of explicitly taking into account the class membership of observations even at the problem formulation stage. Thus, the objective of PLS-DA and OPLS/O2PLS-DA is to find a model that separates classes of observations on the basis of their X-variables. This model is developed from a training set of observations of known class membership.

Conclusions

This example shows the power NMR data in combination with multivariate statistics to capture differences between groups of rats. As a rule, it is always a good idea to commence any data analysis with an initial overview PCA of the entire data set. This will indicate groups, time trends and outliers. Outliers are observations that do not conform to the general correlation structure. One clear outlier was identified among the "sc"-rats.

By way of example we have also shown how groupings spotted by an initial PCA, may be studied further on a more detailed basis. Then techniques like OPLS/O2PLS-DA and SIMCA are very useful. OPLS/O2PLS-DA will be described in another exercise.

In this exercise, we have focused on the differences between two classes, i.e. the "s" and "sc"-rats. This is an analysis that wills pick-up the drug-related effects of the chloroquine treatment. In order to find out exactly which variables (i.e., chemical shift regions) carry the class discriminatory power one may consult plots of PCA, PLS-loadings, OPLS/O2PLS-loadings or contribution plots. A few of these possibilities were hinted at throughout the exercise.

MVDA-Exercise HEALTH

Analysis of Data from a HealthCare Centre

Background

A number of patients at a healthcare centre volunteered as part of an investigation into the efficacy of a certain treatment programme. Various parameters were measured on each patient both before and after receiving the treatment.

Objective

The objective of the study was to investigate whether the treatment was effective or not. The objective of the SIMCA-P investigation is:

(i) to learn how to handle paired comparison data,

(ii) to highlight some different scalings that may be appropriate in this context.

Data

57 patients were included in the survey. Measurements were taken before and after their stay at the centre.

Secondary observation ID's

Before=B, After=A visit to hospital
Sex: Male=M, Female=F
Age category: Young/Middle/Old Y=≤25, M=26-46, O=≥47
Education: F =Elementary School G =Upper Secondary School V=Nursing School H =University
Type of ailment: W=Weakness group, T=Tightness group, Unselected = -

Variable Definitions

Variable 1

1	Y	describing before=0 and after treatment=1

Variables 2–10 (General background data)

2	AG	Age
3	SX	Sex, Male=2, Female=1
4	BS	Systolic bp
5	BD	Diastolic bp
6	HD	High density lipoproteins
7	TR	Triglycerides
8	CO	Cholesterol
9	BM	BMI
10	TV	Test Value

Variables 10–43 are physiotherapy test values, relating to physical strength, balance, walking ability, breathing, etc.

Tasks

Task 1

Import the data file HEALTH.XLS. Select ONAM, SEX, AGE CAT, EDUCATION and TYPE as secondary ID's. Overview the data using PCA. Exclude the first variable column i.e. Y. How do the patients react to the treatment? How are the variables grouped? What patterns does the PCA model describe? (Hint: use the secondary ID's and colour by identifiers)

Task 2

In order to see if the treatment had any effect, we will analyse the data using OPLS/O2PLS. For simplicity, omit the background variables age (AG) sex (SX) and education (ED) from the X-block. Include the Y variable. Analyse the data using OPLS/O2PLS. Investigate scores, loadings and residuals and try to explain what differentiates the "Before" and "After" classes.

Task 3

Given that the observations are paired, an alternative way of analysing the data would be to form a difference table summarising the changes in the patients after treatment. Analysing the data in this way tends to focus on treatment effects rather than variation in the absolute values of the variables.

Import HEALTHDIF.XLS. Again, omit the background variables age (AG) sex (SX) and education (ED) from the X-block.

Calculate a PCA model to overview the difference data using *Two First Components*. (NB: Autofit finds no significant components so the model needs to be forced to fit) Review the fit and interpret the model.

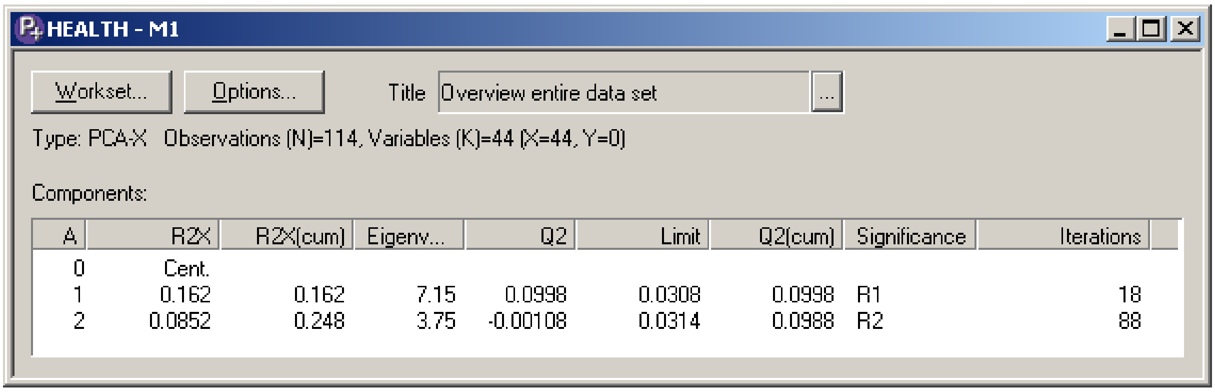
Task 4

Refit the model from Task 3 after changing the scaling to UVN. Review the fit and interpret the model. Explain why the results differ from the previous model?

Solutions

Task 1

A two component model was obtained.



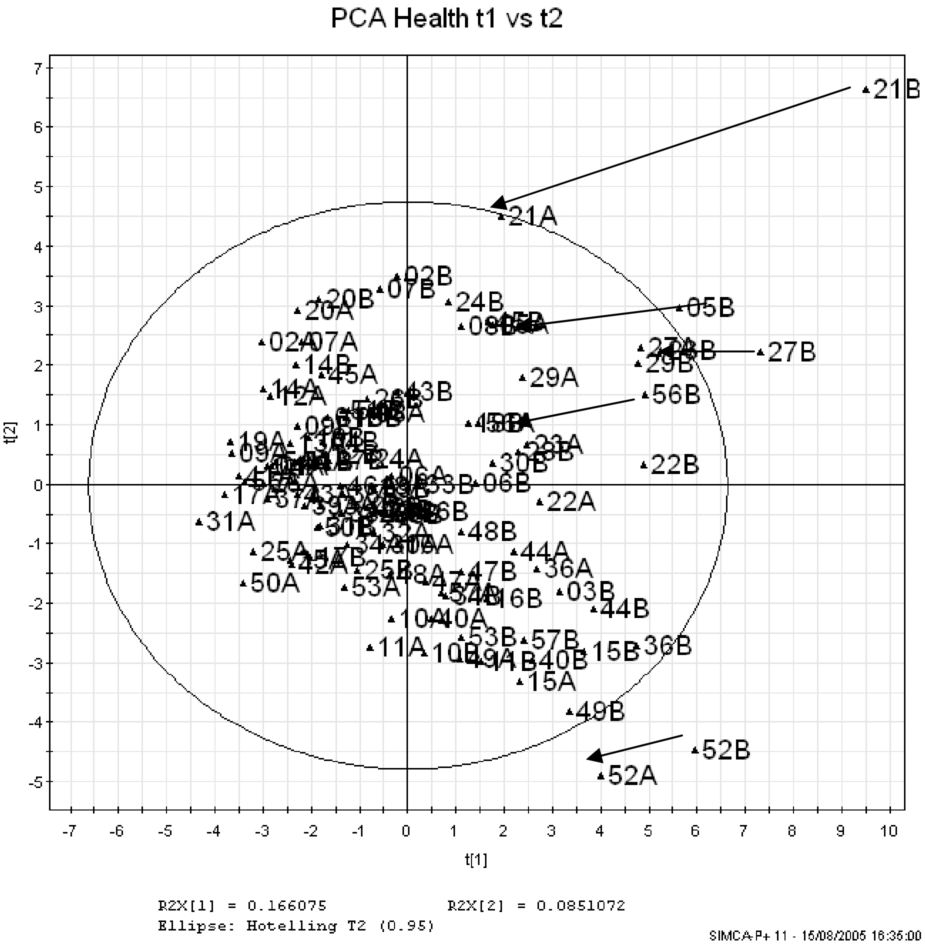
Title: Overview entire data set

Type: PCA-X Observations (N)=114, Variables (K)=44 (X=44, Y=0)

Components:

A	R2X	R2X(cum)	Eigenv...	Q2	Limit	Q2(cum)	Significance	Iterations
0	Cent.							
1	0.162	0.162	7.15	0.0998	0.0308	0.0998	R1	18
2	0.0852	0.248	3.75	-0.00108	0.0314	0.0988	R2	88

We can see from the score plot that the patients tend to move to the left along the first principal component after receiving the treatment. This has been highlighted in the plot by drawing arrows connecting the Before and After points for some patients. The length of each arrow indicates the impact of the treatment for that patient. The patients are also separated along the second component. One group of patients tends to move from the top right-hand corner towards the centre and the other group from the bottom right-hand corner towards the centre.



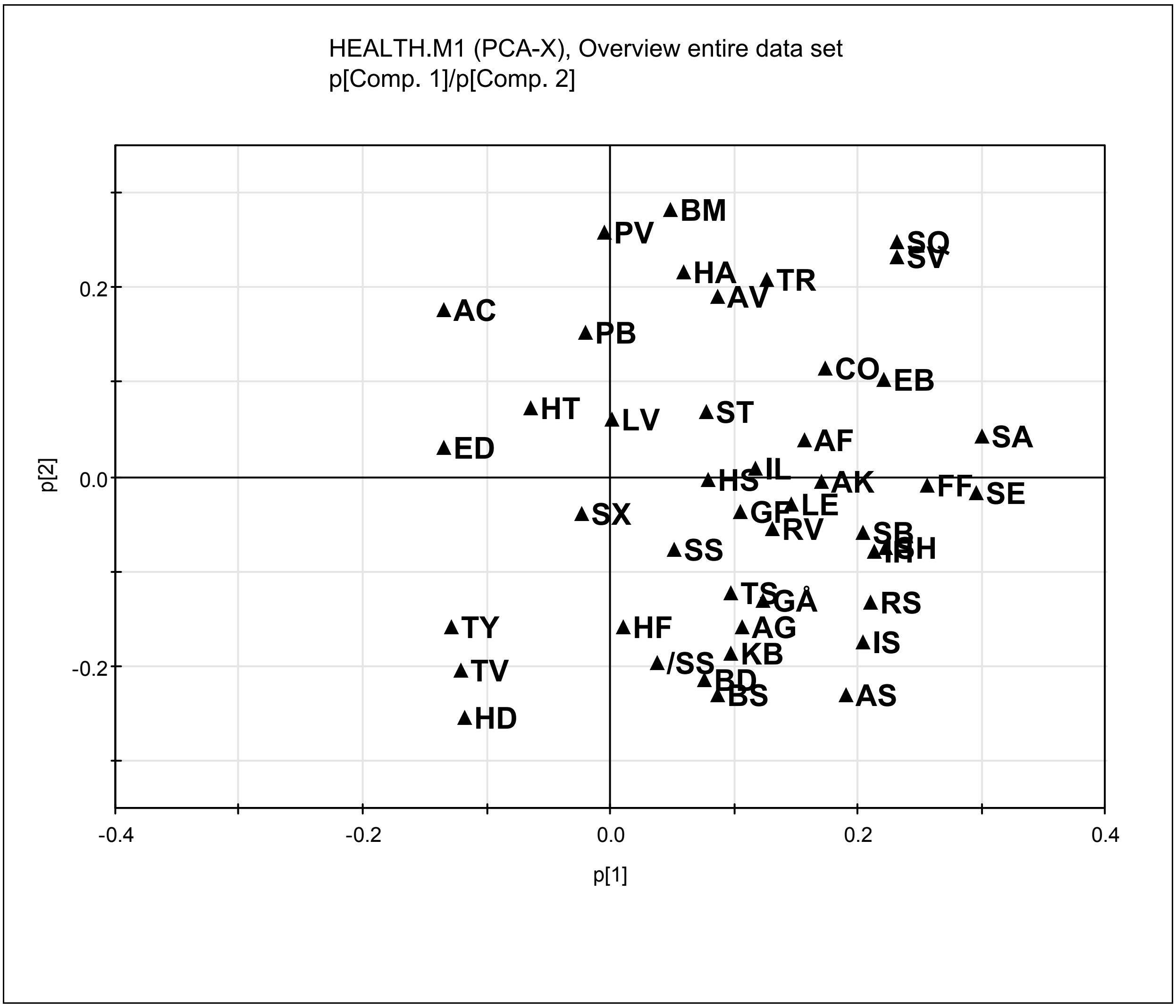
The loading plot shows that the movement from right to left reflects a return to health in that patients (e.g. 21) appear to benefit from the treatment. We can also see that level of education (ED) appears to correlate with this propensity to recover. Patients having the worst test values are generally those with the lowest education level. (This trend may also be seen in the scores plot by *Colouring by Identifiers ObsID (EDUCATION)*)

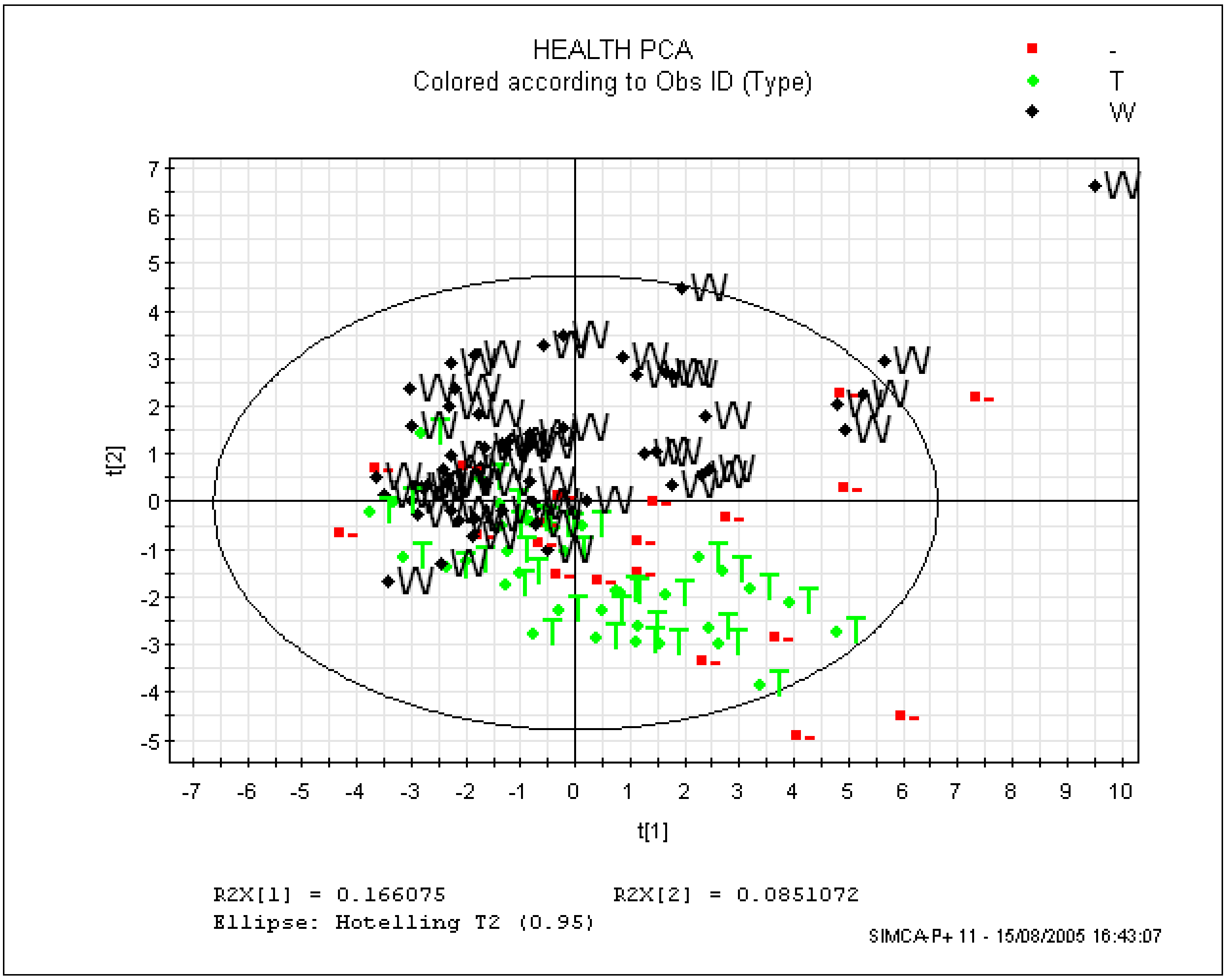
Further, in the top right-hand quadrant we find variables related to body weakness (SQ, SV, SA) and so we can assume that patient 21, for example, scores badly on these assessments. In the bottom right-hand quadrant, we find variables related to muscle tightness (SH, IS, RS, and AS). These groupings, together with the score plot above, suggest that there are two types of patients:

(i) those that migrate from the top right towards the centre ("weakness group") and

(ii) those that migrate from the bottom right towards the centre ("tightness group").

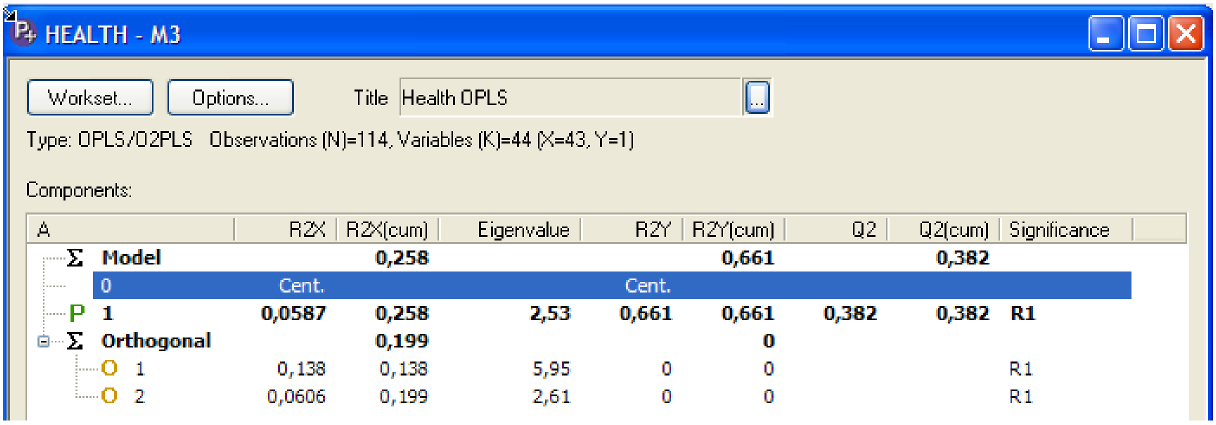
This may be seen more clearly by colouring and naming the observations in the scores plot by the 'Type' secondary ID.





Task 2

A 1+2 component OPLS model was obtained. Only 5,9% of the variation in the data is due to the treatment. The uncorrelated information is 20% of the variation in the data.



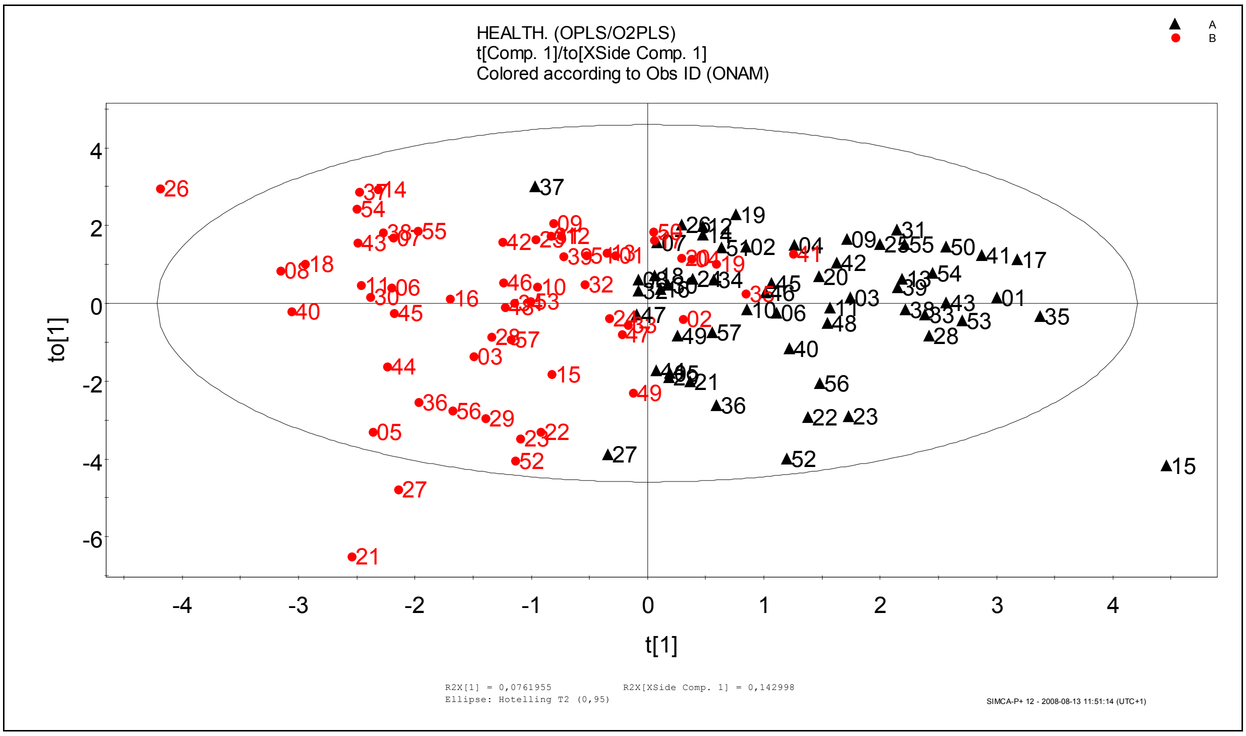
Title: Health OPLS

Type: OPLS/O2PLS   Observations (N)=114, Variables (K)=44 (X=43, Y=1)

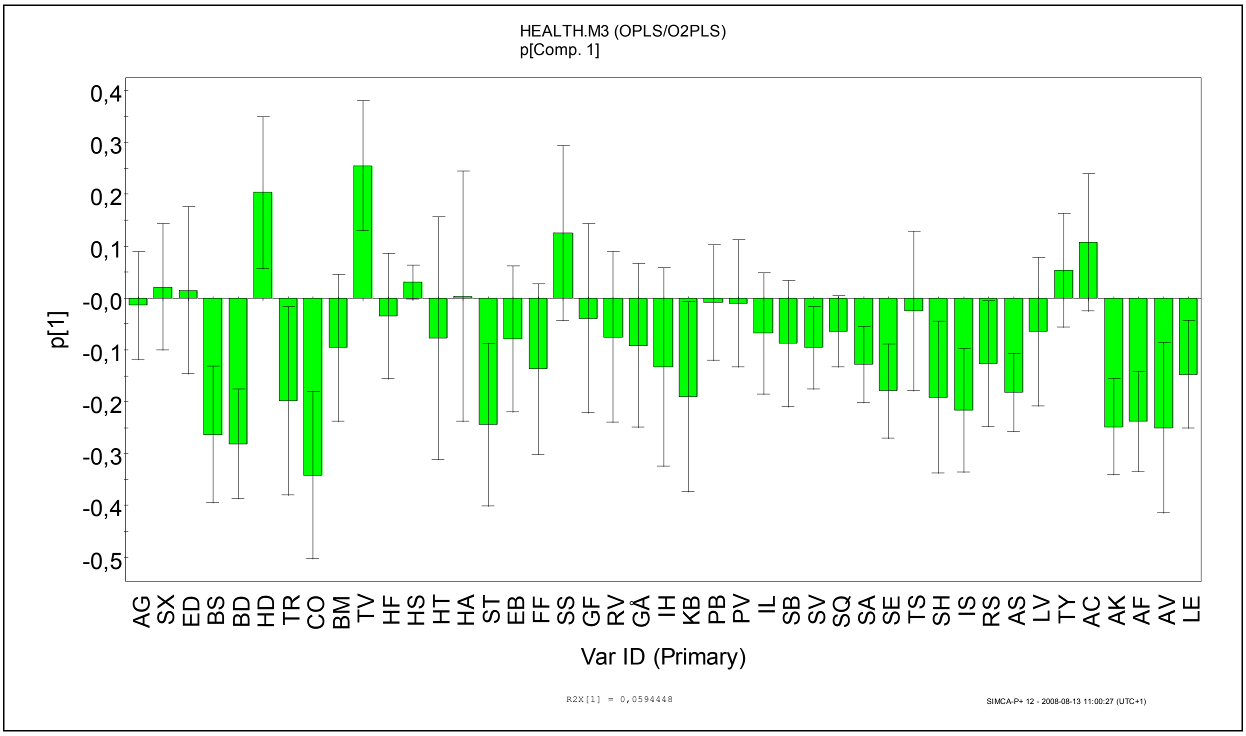
Components:

	A	R2X	R2X(cum)	Eigenvalue	R2Y	R2Y(cum)	Q2	Q2(cum)	Significance
Σ	Model		0,258			0,661		0,382	
	0	Cent.			Cent.				
P	1	0,0587	0,258	2,53	0,661	0,661	0,382	0,382	R1
Σ	Orthogonal		0,199			0			
O	1	0,138	0,138	5,95	0	0			R1
O	2	0,0606	0,199	2,61	0	0			R1

In the t_1/t_{o1} score plot (below) there is separation between the Before (dots) and After (triangles) treatment classes. Some overlap is also seen.

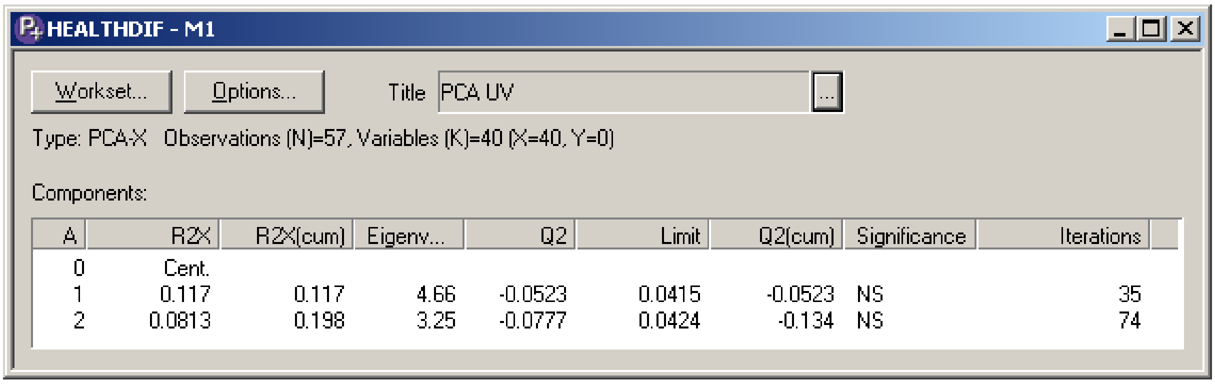


The loading plot (below) reveals how the various tests contribute to the separation of the classes. For example, the cholesterol level (CO) has clearly decreased and the physical test value (TV) has clearly increased as a result of staying at the care centre.



Task 3

A two-component PCA model was obtained for the table of differences. Notice the very low Q2, in this case we will accept this tentative model in order to get an overview of the data.



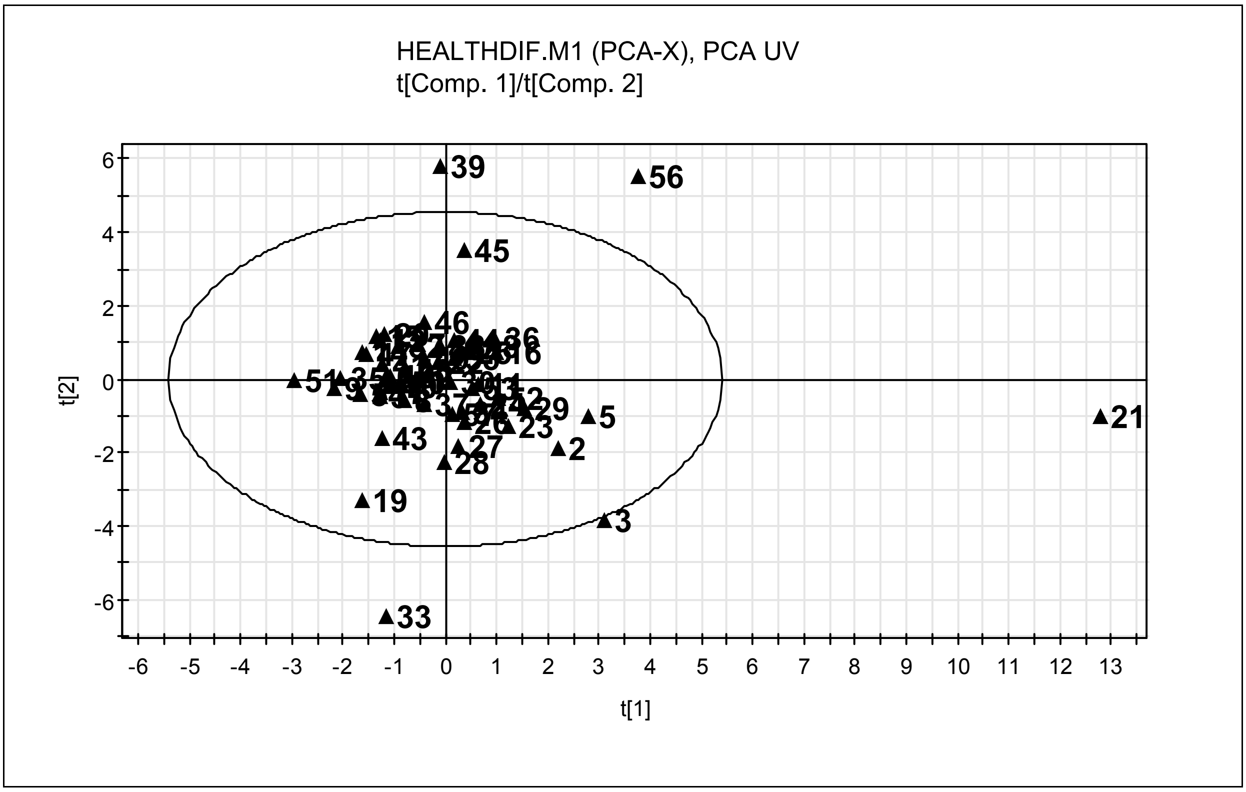
Title PCA UV

Type: PCA-X Observations (N)=57, Variables (K)=40 (X=40, Y=0)

Components:

A	R2X	R2X(cum)	Eigenv...	Q2	Limit	Q2(cum)	Significance	Iterations
0	Cent.							
1	0.117	0.117	4.66	-0.0523	0.0415	-0.0523	NS	35
2	0.0813	0.198	3.25	-0.0777	0.0424	-0.134	NS	74

The t_1/t_2 score plot (below) confirms that patient 21 has undergone the largest change.

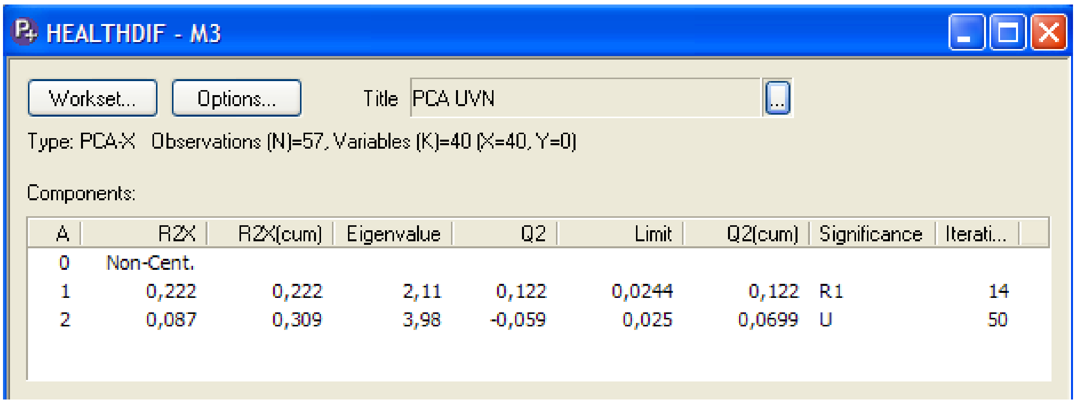


This model reflects the variation in the effect of the treatment on the patients. If all patients had experienced the same changes in health they would all be located near the centre of the score plot.

Note that because UV scaling has been used to pre-process the data for this model, the treatment effects have been eliminated due to mean-centring. This is also reflected in the low Q2 obtained. It might be more instructive, therefore, to repeat the analysis without mean-centring.

Task 4

After changing to UVN scaling, a two-component model was again obtained. However, only the forst component is significant. The second component was added for visualisation purposes and should not be over interpreted.



HEALTHDIF - M3

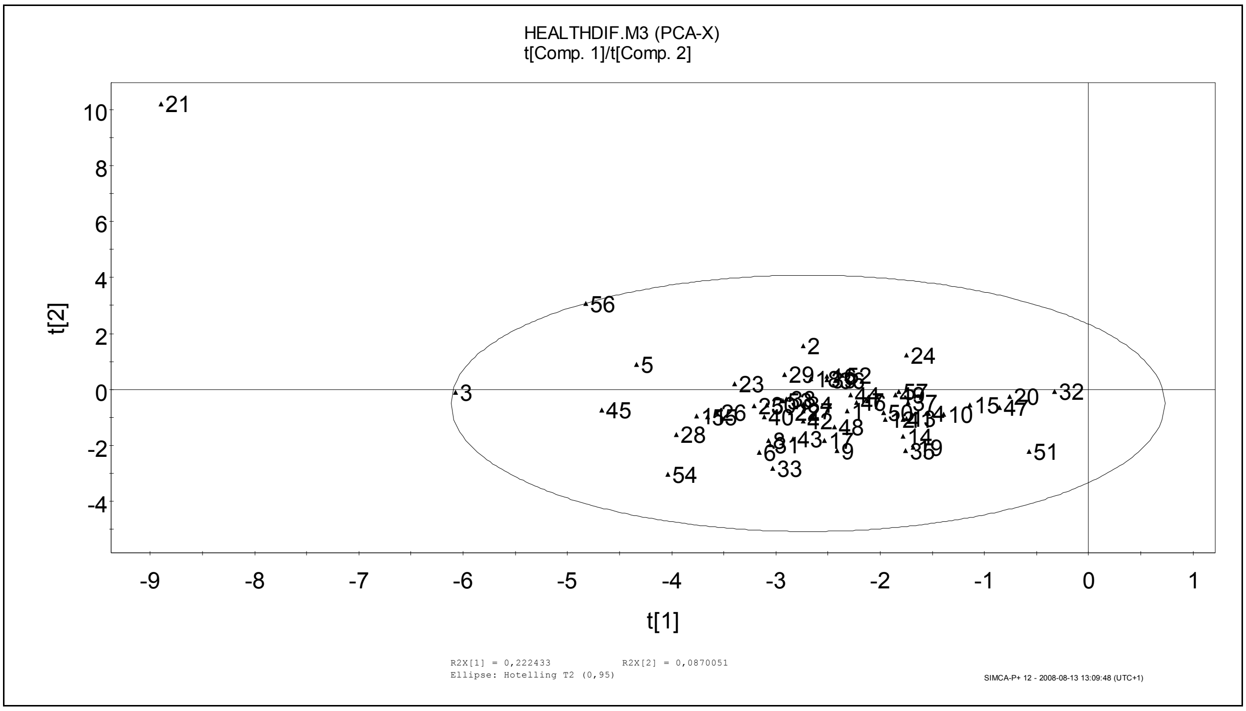
Workset... Options... Title PCA UVN

Type: PCA-X Observations (N)=57, Variables (K)=40 (X=40, Y=0)

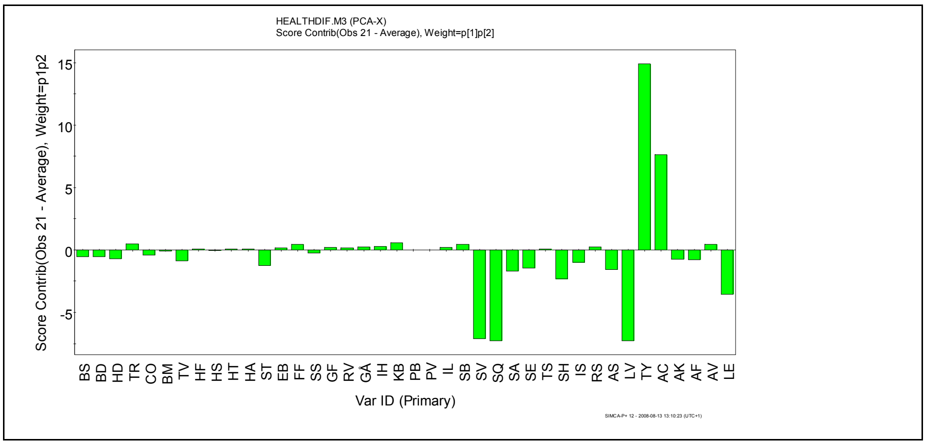
Components:

A	R2X	R2X(cum)	Eigenvalue	Q2	Limit	Q2(cum)	Significance	Iterati...
0	Non-Cent.							
1	0,222	0,222	2,11	0,122	0,0244	0,122	R1	14
2	0,087	0,309	3,98	-0,059	0,025	0,0699	U	50

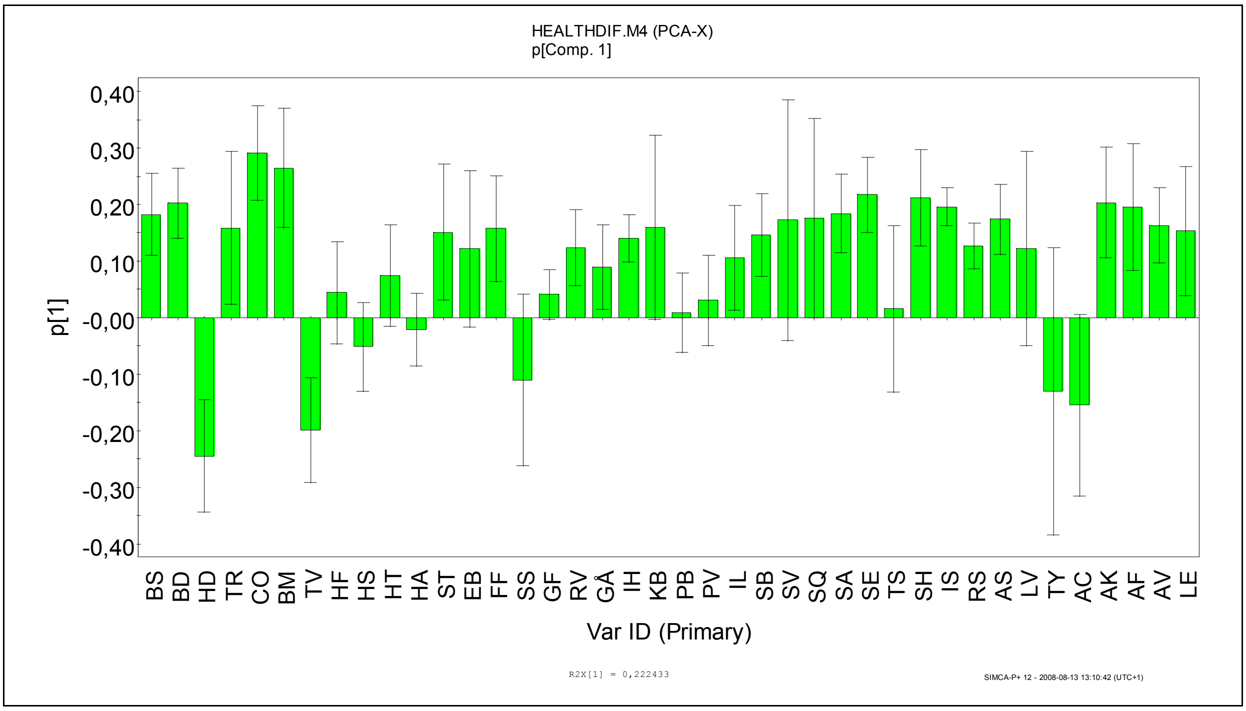
The t_1/t_2 score plot is shown below. If there were no treatment effects, all the patients would cluster around the centre of the score plot. Here, however, we find that every single patient has shifted to the left along the t_1-axis with patient 21 clearly being the most susceptible to the treatment.



Contribution plots can be constructed to interpret which variables make observation 21 extreme in the t_1/t_2 plane. The contribution plot (below left) contrasts patient 21 (extreme change) with patient 32 (minimal change). The largest shift is in variable TY, which reflects difficulty in breathing. We conclude that the main reason for patient 21's change in health is due to a significant improvement in his/her breathing ability.



A more general interpretation is provided by the loading plot (below), bearing in mind that all the patients, with the exception of 21, shift along the first component only. Also note that the observations are located at the negative side of the score plot. This means that all positive p1 loadings have decreased after treatment and all negative p1 have increased after treatment. The plot suggests that e.g. cholesterol level (CO) and body-mass index (BM) have decreased as a result of the stay at the care centre whilst physical fitness (TV) and high-density lipoproteins (HD) have simultaneously increased.



Note: The direction of a PCA can not be determined. However, to simplify the interpretation it is sometimes convenient to flip the axes in the scores and loadings. If the x-axis in the score plot and loading plot is inverted in this example the interpretation would be: all samples have shifted to the **right** side of the score plot, all **positive** p1 loadings have **increased** after treatment and all **negative** p1 have **decreased** after treatment. This does not change the result, only simplify the interpretation. To do this: right click on the plot; choose *plot settings/axis/General/values in reverse order*.

Conclusions

When analysing paired data of the type “before-and-after-treatment”, it is recommended that you look at tables of differences. The analysis of differences highlights the relative change in the numerical values of the variables, rather than on changes in their absolute values. Moreover, by avoiding centring the data and only scaling them, it is possible to focus the analysis towards the effect of the treatment, as seen in the final task.

MVDA-Exercise MS Metabonomics

Using Mass Spectroscopic Metabonomics Data

Background

The data come from a study of genetic variation in Mice. Three genetically distinct strains of mice Black, White and Nude were studied by obtaining Liquid Chromatography - Mass Spectroscopy (LC-MS) data on mouse urine samples.

Objective

The objective of this exercise is to find potential biomarkers, which result from genetic differences in mice. To do this the data must be examined to see if the animals~~y~~ may be separated into clusters. If groupings or clusters are identified then discriminant analysis can be used to determine which variables lead to the separation (and hence which are potential biomarkers). It is important in any data analysis to firstly get an overview of the data, especially in classification procedures to ensure classes are tight and do not contain outliers. Finally the models must be validated to ensure predictive ability.

Data

LC-MS data are~~is~~ three-way ~~dimensional~~ in nature with Time, Absorbance and Mass dimensions. In this case the data have been unfolded in a time-wise way. Every time a chromatographic peak is detected a set of masses is produced so that each variable is a time followed by a mass i.e. 3.2_245.67 (Time_Mass).

The data were produced by a Waters Mass-Lynx system where the export of data is done according to tunable parameters for peak selection. By default the data are sorted into mass order, which jumbles up the time information. In this case we found it clearer to sort the data in Excel so that each time-window (peak) is together in the table.

The data consist of 29 observations and 4145 x-variables and one Y variable including class information Nude mice=0, white and black=1. ~~so i~~

NB. ~~In order to maintain secrecy the data have been altered slightly to disguise the masses and retention times. This was necessary as potentially commercially useful biomarkers could be identified.~~For commercial reasons the data have been disguised.

Acknowledgement

We are grateful to Dr. Ian Wilson of AstraZeneca, Dr.Rob Plumb, Dr Chris Stumpf and Dr Jennifer Granger of Waters Micromass for the use of data generated by the Waters Marker-Lynx Mass Spectroscopy system.

Tasks

Task 1

Create a new project in SIMCA-P by selecting *File / New* and clicking on MSMouseT.txt. Set columns 1 and 2 to *Primary and Secondary Observation ID* respectively. Set row 2 to be the *Primary variable ID* and row 1 to be the *Secondary variable ID*. Click *Next~~, Next~~* ~~until the dataset is imported~~ and finish the import procedure. Open the data set and right click on the table go to *properties/observations* and define the three classes; this will be useful for OPLS/O2PLS-DA and misclassification table.

Select *Workset New*, click on the scale tab, select all variables and set the scaling to *Par.* Make sure you click the *Set* button. Under the observations tab right click the column heading to change the observation label to the *secondary ID* and turn off the *primary ID*. Using the *find* function use wildcards to set White (W*) as Class 1, Black (B*) as Class 2 and, Nude (N*) as Class 3.

Click OK. (Due to the size of the dataset you may experience a delay). Answer OK to the message about variance.

We will now perform a PCA on the whole dataset to get an overview. Change the Model Type to *PCA on X-Block* (SIMCA-P defaults to class models when classes have been defined). Click on *Autofit*. How many components does SIMCA-P find? How many of these are sensible? Reduce the number of components to this number by using the *Remove Component* function.

Important: For metabonomics applications ~~where~~ there are so many variables, you should set the *Plot labels limit* to about 500 under *View / General Options / More Options* to prevent long delays while waiting for the plots to draw.

Examine the scores and loadings plots. Are the classes of mice separable? Are there any outliers? Are there any trends in the data? Classes should also be scrutinized for outliers by separate PCA models (results not shown).

Task 2

We will now try a classification using the SIMCA method. Select *Workset / New As Model 1*. Click *OK* and *OK* again to the dialog box. SIMCA-P will now have defaulted to *PCA-Class(1)*. Go to *Analysis/Autofit Class Models* but specify 2 components in each case. A local 2 component PCA model will be built for each class. Name each model White, Black, Nude.

Examine the Scores and Loadings plots for each class. Comment upon the R2 Q2 values obtained.

Use the *Coomans Plot* and the *Classification list* under the *Predictions Menu*. What is the advantage and disadvantage of using a Coomans plot over the Classification list~~.~~? Comment on the classification obtained.

Task 3

OPLS Discriminant A~~a~~nalysis will now be carried out to search for potential biomarkers. For this exercise we will focus only on the difference between the Nude and the Black strains of mice. For this we will need to exclude the white mice. Select *Workset / New*. ~~g~~Goto the Observations tab and exclude all white mices, go to the scale tab and set scaling to *par,* remember *select all variables* and press *set*, choose the last column to Y. Select *model type OPLS/O2PLS*

Autofit a model, a two component model will be obtained.

Interpret the model by looking at the model diagnostics and score plot. Extract potential biomarkers by using the S-plot and the *p* loading plot (use column), which change the most?. Interpret the orthogonal component by using the S-plot from the orthogonal components.

(*TIP*: use the right click function to sort the loadings (*Sort Ascending*) and use the zoom tool to inspect them in more detail.)

If time allows: compare the nude mice with the white ice and compare the results with nude vs. black mice. Use the SUS-plot to compare models.

Task 4

To be able to trust the model it must be first tested to check its predictivity. Ideally this should be in the form of newly collected data but if this is not available the only option is to leave some data out of the model and make predictions as to group membership and check that the same coefficients are produced minus some of the data.

The problem with most metabonomics data is that there tend to be low numbers of observations (mainly for ethical reasons). When the dataset is below 10 observations per class "a leave one out in turn" validation should be carried out. Above 15 samples it is possible to build the model on 10 observations and predict 5. Above 20 observations the data may be split in half.

In this case (9 observations in smallest group), although a "leave one out in turn" validation would be most appropriate, we will attempt to split the dataset in half, for speed reasons. Leaving out half is a very severe test. (These instructions will also apply to a "leave one out in turn" validation, just the number of exclusions and the number of times to repeat the analysis will change.)

Select *Workset / New.* G~~g~~o to the Observations tab and exclude all white mice and alternate observations leaving 5 observations in each of the two classes (nude and black class). Scale the data using *par*. Define the response variable Y and re-run the OPLS/O2PLS model. Compare this model to the model when all of the data were~~as~~ used.

Produce an S-plot plot and compare the list of variables with that for the full dataset model in Task 3. Does this model reach the same conclusions? Can you suggest some Potential Biomarkers for further study?

To make the misclassification table the OPLS/O2PLS-DA function must be used. Make a new model *work set/new* exclude the white mice and the response Y, use same scaling as in previous model. Select *model type/OPLS/O2PLS-DA*. Fit a 1+1 component model. Go to the *Predictions Menu / Specify Prediction Set / Specify* Select only the Nude and Black mice that were excluded. Click OK. Make a *misclassification table* and a *prediction list*. Does the model correctly predict the type of Mouse?

Task 5 (Advanced Further Study)

Pareto scaling has been found to give the best results with MS data as it down-weights the largest and up-weights the medium peaks without inflating baseline noise. Discuss the disadvantages of either UV scaling or no scaling (Ctr).

Consider and discuss whether there is a role for Hierarchical modeling with LC-MS and GC-MS data using time windows.

(NB: There is no solution given for this task)

Solutions

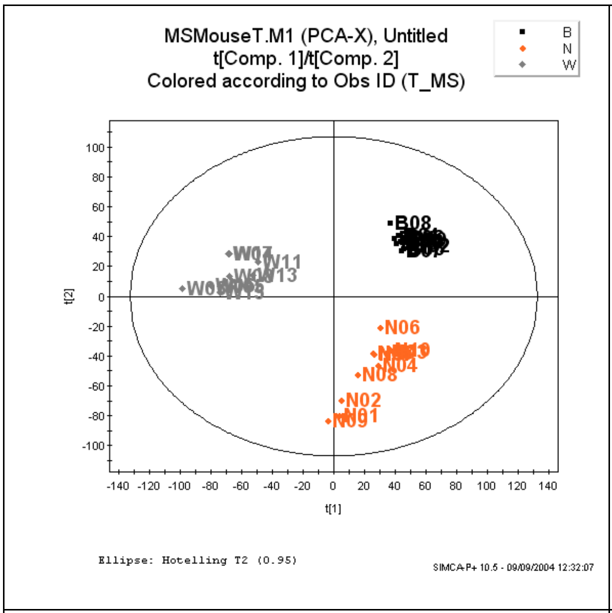
Solution Task 1

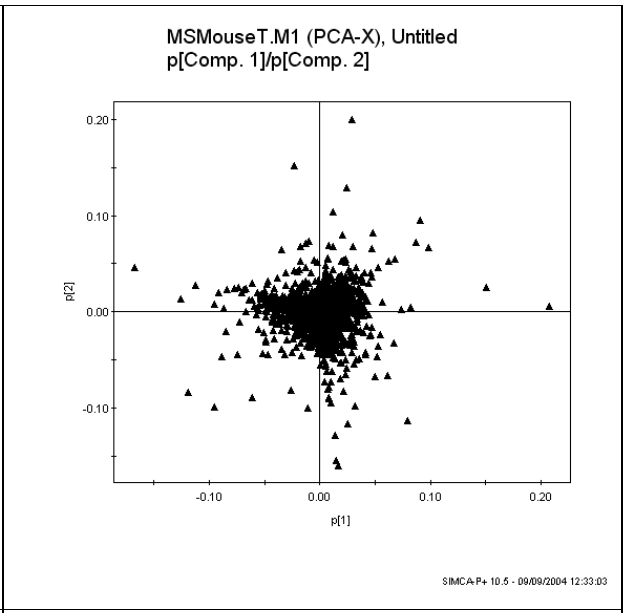
After 3 components R2=0.48 Q2=0.28. These figures are typical of PCA for “biological” type data.

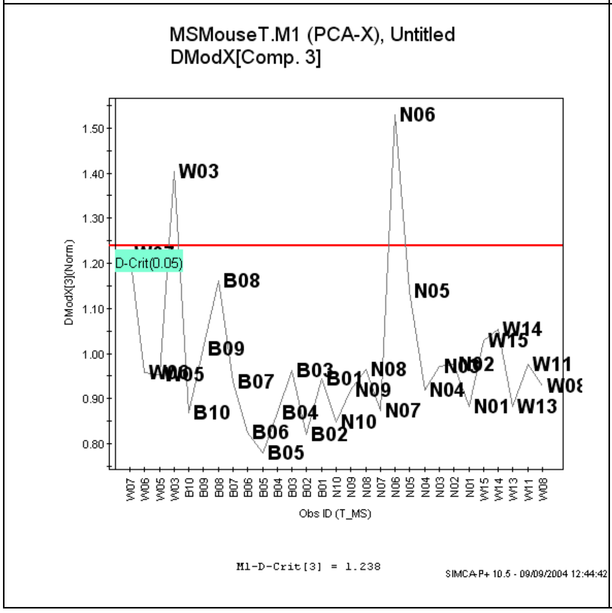
The scores plot (t1/t2) clearly shows separation between the mouse types. The black mice are the most consistent whereas the white and nude mice appear to show a slight trend. Trends in PCA plots are always interesting to investigate. If a trend is seen there is often a reason behind it and if it can be identified and fixed it leads to better procedures in future experimentation. “Biological Variation” when detected should be random.

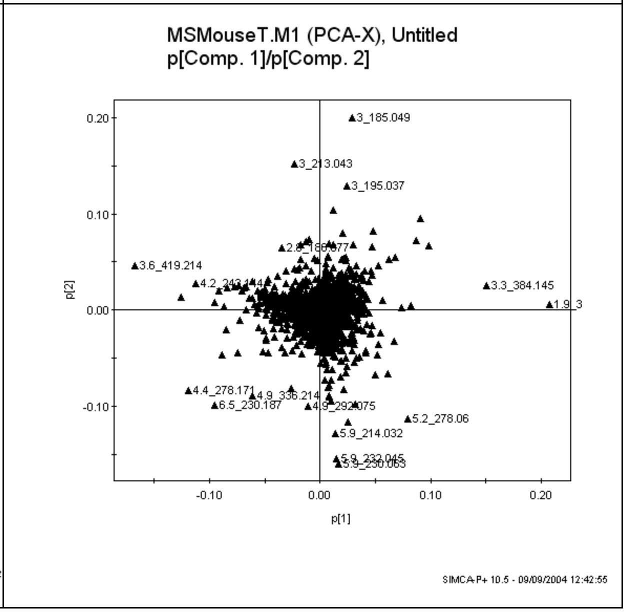
The loadings plot shows the variables responsible for the differences in the mice groupings. As we have set the label limit under general options no labels are shown. To turn on the labels for the relevant variables highlight them with the selection tool and use the drop down box on the menu bar.

The DM~~m~~odX plot after 3 components shows W03 and N06 slightly outside Dcrit, but since they do not appear as outliers in the scores plot they are left in the model.

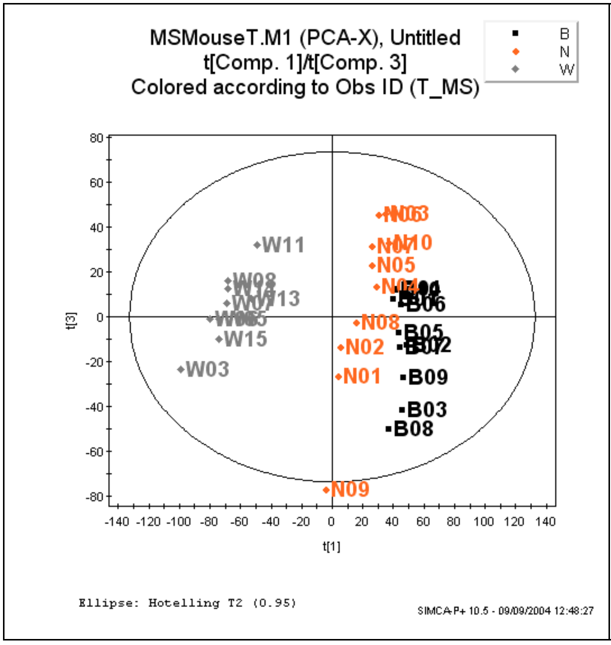


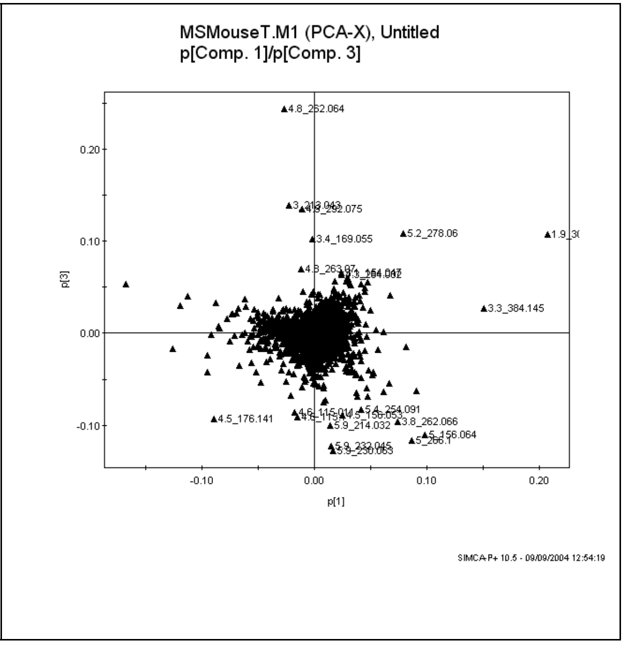






The Scores plot for components t1/t3 show mainly a within class trend. Plotting the loadings shows the variables in the top and bottom of the plots are responsible. In situations like this the chemical identity of these points could be found and viewed in the light of the animal handling records to see if the trend is related.





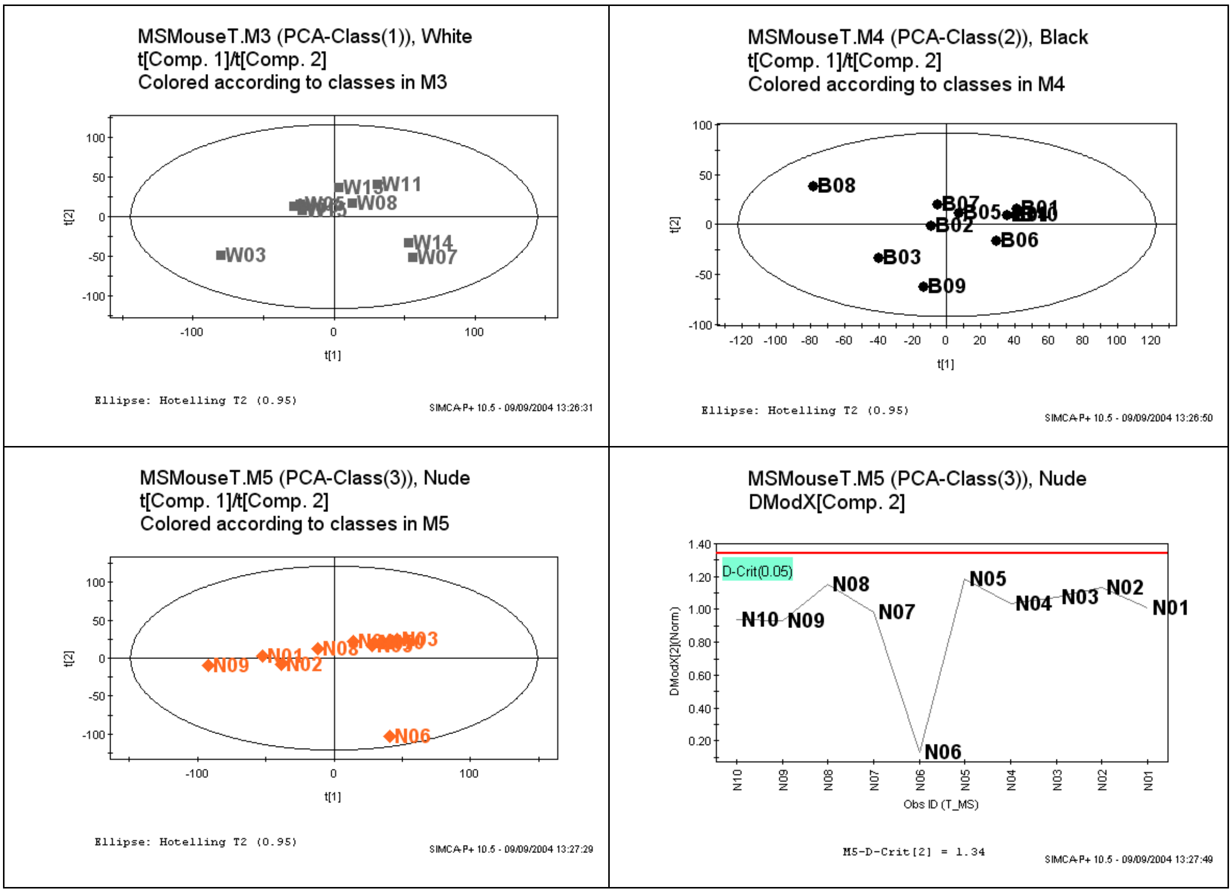
Solution Task 2

The following models should be obtained:

Observations (N) = 29, Variables (K) = 4145

No.	Model	Type	A	R2X	R2Y	Q2(cum)	Title
2	M2	PCA-Class(1)		2	0.464	-0.0332	White
3	M3	PCA-Class(2)		2	0.415	-0.108	Black
4	M4	PCA-Class(3)		2	0.49	-0.0205	Nude

In each case very poor models result. The reason for this is that each class is very uniform so that PCA struggles to find a trend within each class. This is the ideal situation for classification where “tight” classes are required for good results. The model for the Nude mice shows observation N06 slightly away from the rest of the group but it lies within Hotelling's T2 and DM~~m~~od X so it is kept within the model.



MSMouseT.M3 (PCA-Class(1)), White
t[Comp. 1]/t[Comp. 2]
Colored according to classes in M3

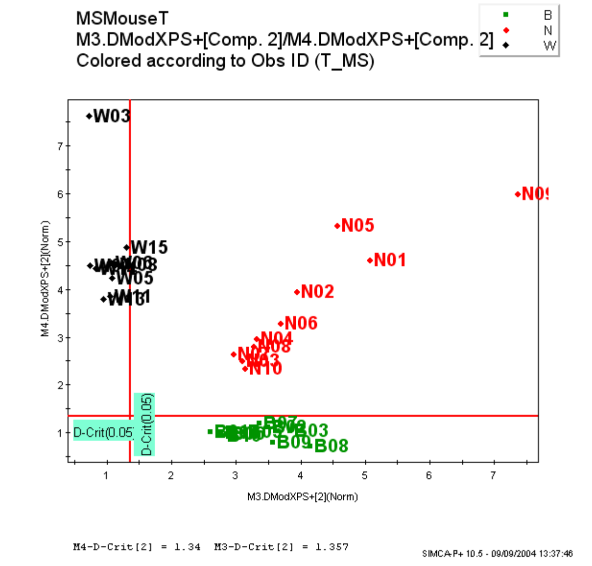
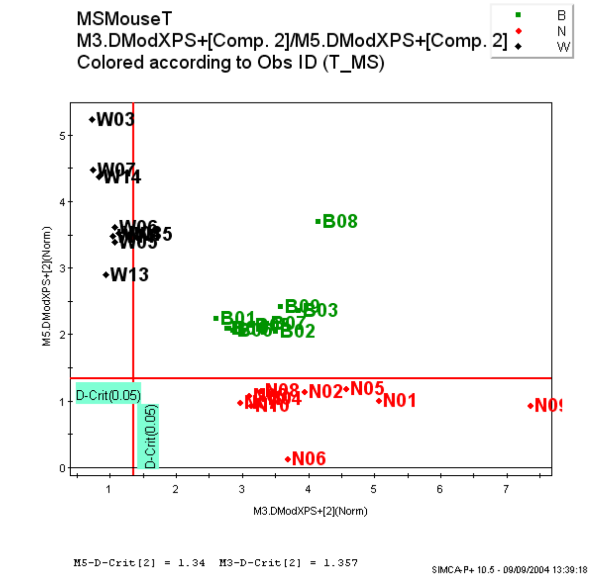
MSMouseT.M4 (PCA-Class(2)), Black
t[Comp. 1]/t[Comp. 2]
Colored according to classes in M4

MSMouseT.M5 (PCA-Class(3)), Nude
t[Comp. 1]/t[Comp. 2]
Colored according to classes in M5

MSMouseT.M5 (PCA-Class(3)), Nude
DModX[Comp. 2]

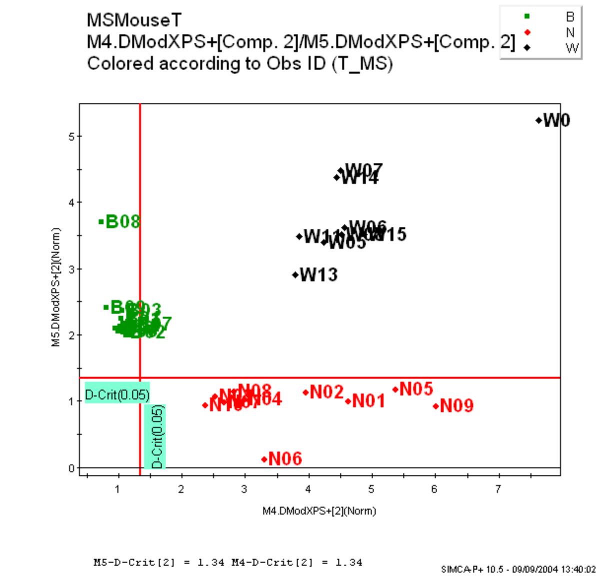
There are three possible combinations of Coomans' Plots using three models; White vs. Black; White vs. Nude and Black vs Nude. In each case classification works perfectly, however this is using the data used to train the model.

NB: For a rigorous test, one of the mice of each group should be left out and a new model built. The mice left out should then be predicted. This is then repeated 10 times so that each mouse is left out once and the process repeated. Tabulating the results will then show the true Predictivity of the model.

<table>
<tr>
<td>White vs Black
Black 10 100% correct
White 9 100% correct
Other 10 100% correct
Both 0</td>
<td>
</td>
</tr>
<tr>
<td>White vs Nude
White 9 100% correct
Nude 10 100% correct
Other 10 100% correct
Both 0</td>
<td>
</td>
</tr>
</table>

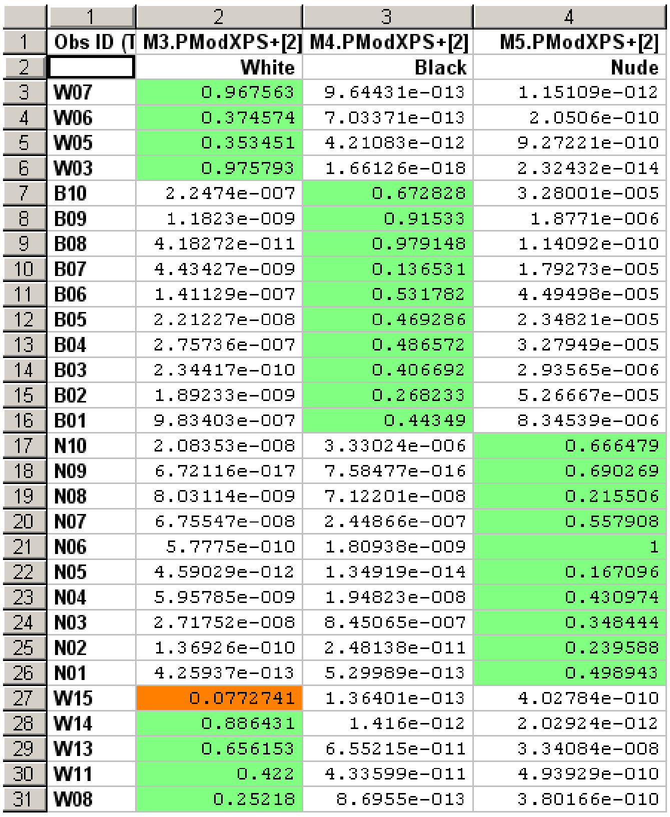
Black vs Nude

Black	10	100% correct
Nude	10	100% correct
Other	9	100% correct
Both	0	



The advantage of the Coomans' plot is the four diagnostic regions, which is useful when predicting new data. The disadvantage is that it only does binary comparisons, so in cases with more than one class specific model the number of plots required increases.

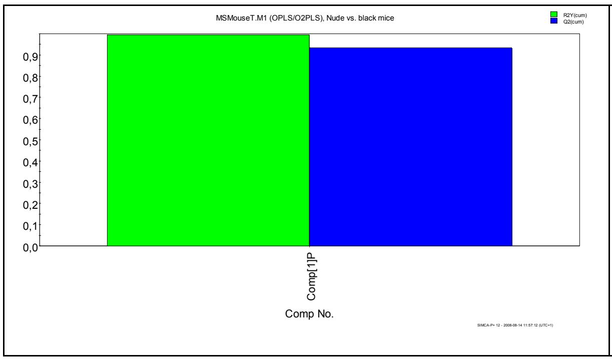
For many class SIMCA models the Classification list may be used. Using the data that the model was built upon, the classification is perfect. As mentioned above the true predictivity of the model should be determined either using new data or, failing this, ~~data~~ left out data.

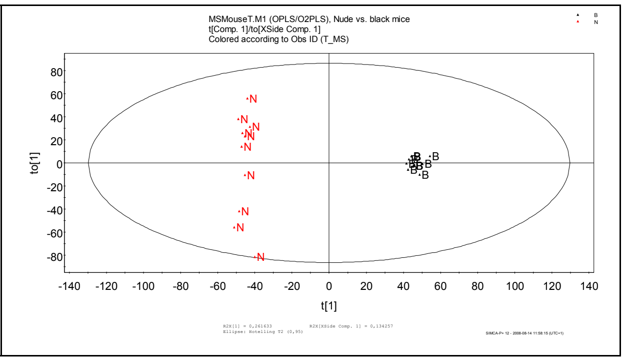


	1	2	3	4
1	Obs ID (T	M3.PModXPS+[2]	M4.PModXPS+[2]	M5.PModXPS+[2]
2		White	Black	Nude
3	W07	0.967563	9.64431e-013	1.15109e-012
4	W06	0.374574	7.03371e-013	2.0506e-010
5	W05	0.353451	4.21083e-012	9.27221e-010
6	W03	0.975793	1.66126e-018	2.32432e-014
7	B10	2.2474e-007	0.672828	3.28001e-005
8	B09	1.1823e-009	0.91533	1.8771e-006
9	B08	4.18272e-011	0.979148	1.14092e-010
10	B07	4.43427e-009	0.136531	1.79273e-005
11	B06	1.41129e-007	0.531782	4.49498e-005
12	B05	2.21227e-008	0.469286	2.34821e-005
13	B04	2.75736e-007	0.486572	3.27949e-005
14	B03	2.34417e-010	0.406692	2.93565e-006
15	B02	1.89233e-009	0.268233	5.26667e-005
16	B01	9.83403e-007	0.44349	8.34539e-006
17	N10	2.08353e-008	3.33024e-006	0.666479
18	N09	6.72116e-017	7.58477e-016	0.690269
19	N08	8.03114e-009	7.12201e-008	0.215506
20	N07	6.75547e-008	2.44866e-007	0.557908
21	N06	5.7775e-010	1.80938e-009	1
22	N05	4.59029e-012	1.34919e-014	0.167096
23	N04	5.95785e-009	1.94823e-008	0.430974
24	N03	2.71752e-008	8.45065e-007	0.348444
25	N02	1.36926e-010	2.48138e-011	0.239588
26	N01	4.25937e-013	5.29989e-013	0.498943
27	W15	0.0772741	1.36401e-013	4.02784e-010
28	W14	0.886431	1.416e-012	2.02924e-012
29	W13	0.656153	6.55215e-011	3.34084e-008
30	W11	0.422	4.33599e-011	4.93929e-010
31	W08	0.25218	8.6955e-013	3.80166e-010

Solution Task 3

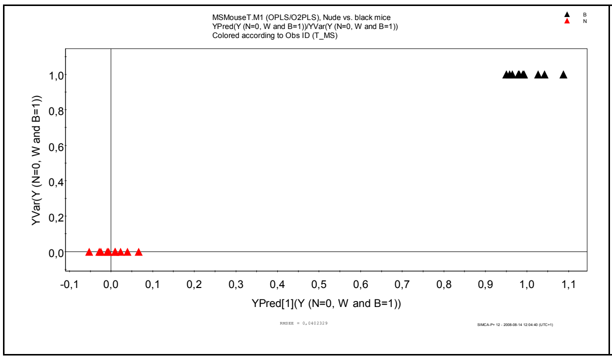
The 1+1 component OPLS/O2PLS gives a~~n~~ model with R2Y=0.99 and Q2=0.93, the predictive variation is 26% and the orthogonal variation is 13% of the total variation in the data. The black mice are in the upper section the Black below.

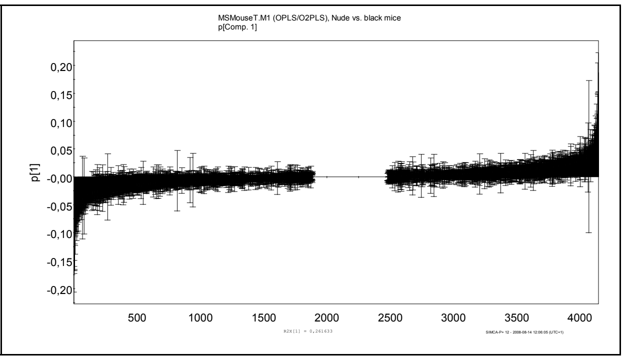


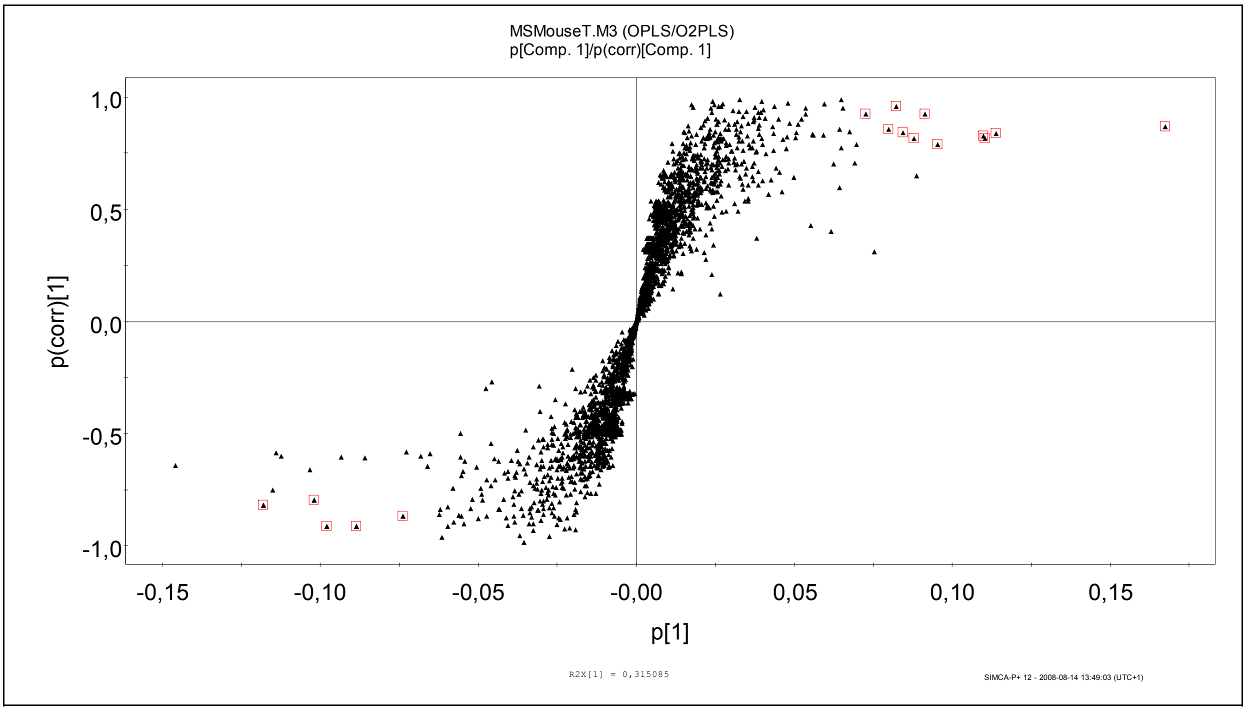


The Observed vs P~~p~~redicted plot shows complete separation of the two groups. In cases where OPLS classification fails there will be an overlap of the two clusters over the 0.5 mark on the X axis.

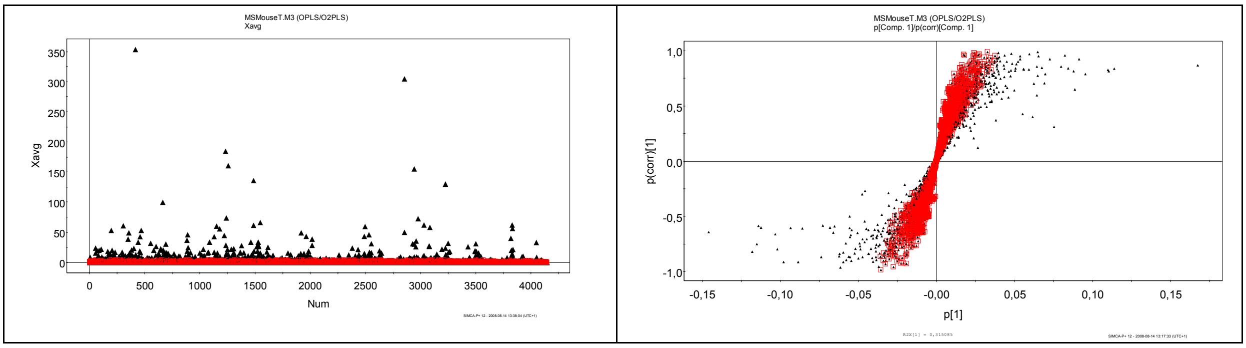
The S-plot and the p1 loading plot show the variables that are responsible for the group separation. The plot as first plotted will be difficult to interpret as it is in the same order as the dataset and with the 95% jacknifed confidence intervals present. Using the sort ascending function and the zoom tool it is possible to identify the most increasing and decreasing potential biomarkers.



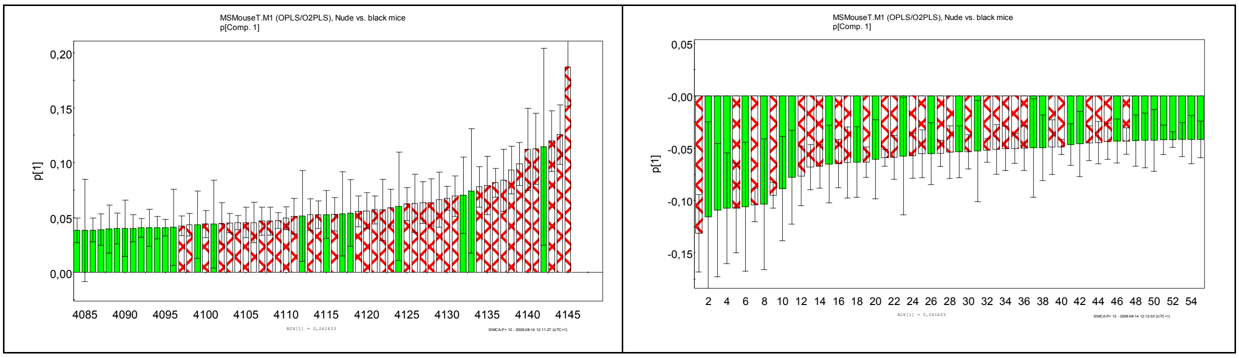




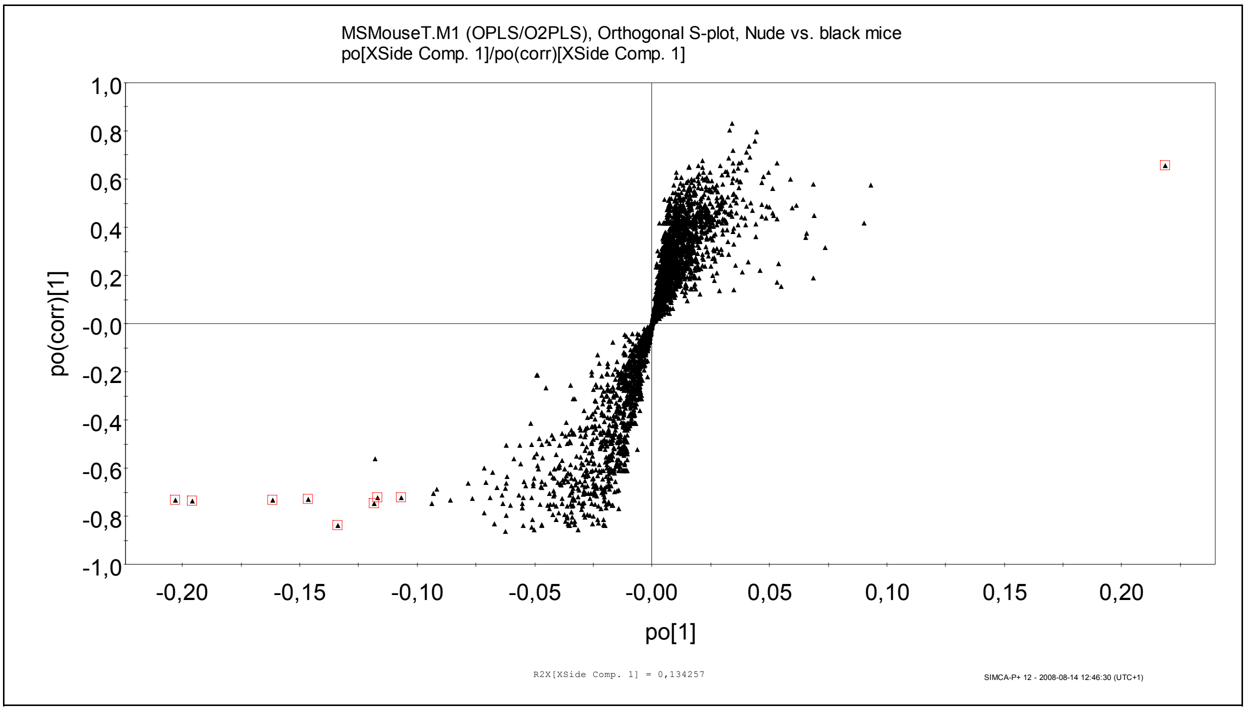
Hint: make an *x-avg* plot under *plot/list* and block mark all signals in the baseline. This will help in the decision of cut off limits (threshold) in the p1 direction.



A list of potential biomarkers can be obtained by right clicking on the S-plot (the selected metabolites must be marked). This list can be saved or copied to excel for further investigation.

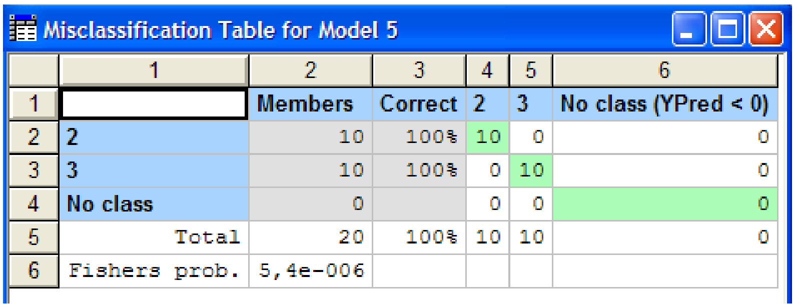


In the score plot it is seen that systematic variation in the nude mice cause the orthogonal component. In the S-plot from the orthogonal components, po1 vs. po(corr)1, it can be interpreted which metabolites that cause this variation.


MSMouseT.M1 (OPLS/O2PLS), Orthogonal S-plot, Nude vs. black mice
po[XSide Comp. 1]/po(corr)[XSide Comp. 1]
1,0
0,8
0,6
0,4
0,2
-0,0
-0,2
-0,4
-0,6
-0,8
-1,0
po(corr)[1]
-0,20
-0,15
-0,10
-0,05
-0,00
0,05
0,10
0,15
0,20
po[1]
R2X[XSide Comp. 1] = 0,134257
SIMCA-P+ 12 - 2008-08-14 12:46:30 (UTC+1)

Solution Task 4 ~~Solution~~

OPLS/O2PLS on the subset of the data has correctly predicted the White and Black mice with high confidence. This is a very good result considering we have split the dataset in half. In practice trusting a model built on only 5 observations would be potentially hazardous.

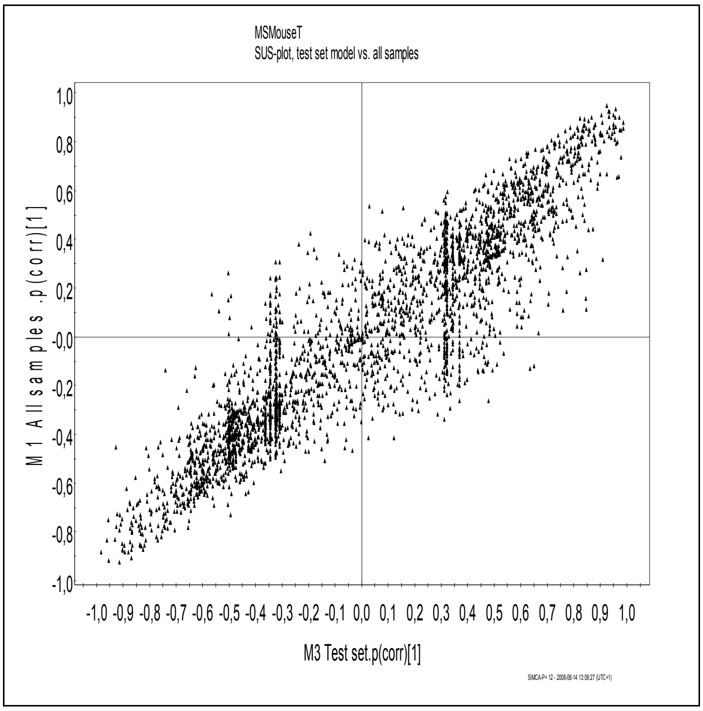


Misclassification Table for Model 5

	1	2	3	4	5	6
1		Members	Correct	2	3	No class (YPred < 0)
2	2	10	100%	10	0	0
3	3	10	100%	0	10	0
4	No class	0		0	0	0
5	Total	20	100%	10	10	0
6	Fishers prob.	5,4e-006				

To get a more detailed prediction make a prediction list.

Comparing the correlations with the previous model it is useful to look at the SUS-plot. Some differences are observed but there are a number of correlations which remain consistent. These are prime candidates for further investigation as “Potential Biomarkers”.



Discussion and Conclusions

The aim of this exercise is to show how Mass ~~-~~Spectroscopy-Liquid Chromatography data can be handled. Data are~~is~~ unfolded time-~~-~~wise to give paired variables of RetentionTime_Mass. Data must be transposed on import if it has been processed in Excel due to Excel's 256-column limitation.

PCA is a good technique for over-viewing trends, groupings and outliers in MS data. If a trend is spotted it is worth investigating the cause.

SIMCA is a technique for recognition of classes, useful in cases of incomplete resolution and with many classes. The Cooman~~'~~s' plot is useful where new data may have observations not belonging to either class defined by the two models.

OPLS/O2PLS in classification studies (OPLS/O2PLS-DA) is a maximum separation projection of data and is most useful when dealing with two classes as it shows which variables are responsible for class separation. In this way potential biomarkers may be found by looking at the most positive and negative loadings in the S-plot.

Validation of models of Metabonomic data is essential to prove the predictive ability~~ of the model~~. In the ideal case, new data not available during the model building process ~~is~~are predicted and evaluated. In cases where new data ~~is~~are not available the dataset must be split into training and test sets. The way this should be done depends on the number of Observations. Suggestions are as follows:

<10	Leave one out,	repeat 9 times (10 in all)
15	Build model on 10 predict 5	repeat 2 times (3 in all)
20	Build model on 10 predict 10	repeat once (2 in all)

Checks should be made that not only the class memberships are predicted correctly but also that the same variables (potential biomarkers) are found important in each model.

MVDA-Exercise GeneGrid

Gaining a visual overview of gene chip data

Background

Gene Chip Array data are becoming increasingly common within the Bio-Pharmaceutical and agrochemical industries. By studying which genes are either up or down regulated it is hoped to be able to gain an insight into the genetic basis of disease. Gene chips are composed of short DNA strands bound to a substrate. The genetic sample under test is labelled with a fluorescent tag and placed in contact with the gene chip. Any genetic material with a complimentary sequence will bind to the DNA strand and be shown by fluorescence.

From a data analysis point of view gene chip data are very similar to spectroscopic data. Firstly the data often have a large amount of systematic variation and secondly the large numbers of genes across a grid are analogous to the large number of wavelengths in a spectrum. If gene grid data are plotted versus fluorescent intensity we get a 'spectrum' of gene expression. The one critical difference between gene data and spectroscopy is that in spectroscopy the theory is known and peaks may be interpreted. In gene expression analysis the function of the genes and the number of genes expected to change is largely unknown, given the current level of understanding.

There are several experimental techniques to remove both within slide and between slide systematic variations. These include running paired slides using different dyes (dye swap), normalising to genes with constant expression (so-called housekeeping genes i.e. beta actin), the addition of standard synthetic DNA controls, and using different concentrations of the same gene in order to normalise the fluorescence readings.

Objective

The objectives of this study are to gain an overview of the gene chip data, investigate systematic variation between experiments and treatment groups and finally to determine which genes have changes in their expression between treatment groups.

Data

The data come from a toxicity study where the gene expression profiles for different doses of a toxin are investigated. The aim is to be able to recognise which genes are changing in response to a particular toxin so that these changes may be used to screen new drugs for toxicity in the future.

The gene grids used are composed of 1611 gene probes on a slide (or chip).

4 different doses are given, Control, Low, Medium, High.

5 animals are used per dose (some missing - 17 in total).

Controls	Animals	2, 3, 4, 5
High Dose	Animals	31, 32, 33, 34
Medium Dose	Animals	26, 27, 28, 29
Low Dose	Animals	21, 22, 23, 24, 25

4 grid repeats per slide – W X Y Z (also called spots) with 3 replicates per animal.

12 measurements in total (17 x 12 = 204 - 8) = 196 observations. (2 grid repeats missing).

It is useful to adopt a systematic naming procedure in SIMCA-P to allow selective masking of the plot labels at a later stage in the analysis. For this study the following naming scheme was used;

Example name: **C02aX**

Position 1	Dose	C,L,M,H
Position 2-3	Animal number	2,3,4,,,,,25
Position 4	GRID	a,b,c
Position 5,6	Repeat	W, X,Y, Z

Tasks

Task 1

Open SIMCA-P and create a new project by *File\New*. Select Genegrid_RAW.txt. Highlight the first Row and select *primary* observation ID. Define the second column as *ClassID* using length 1. A new column will be generated including class information. Ensure column 3 is included as data (highlight column and select *Data*). The file contains 196 observations and 1611 variables.

Open the dataset by selecting *Dataset\Edit\GenegridRAW* (or clicking the dataset icon on the toolbar).

We will use quick info to examine the 'gene spectrum'. *Dataset\QuickInfo\Observations*. Right click on observation 1 (column1, row2). Using the up and down keys on the keyboard scroll down the observations. Is there anything strange about the gene expression pattern in any of the observations? Make a note of the observations that look different or unusual.

Task 2

During import of data, SIMCA automatically prepared 4 different datasets for PCA on individual classes. All of these classes are placed under *CM1*. Change the scaling to **Par scaling** under *workset/edit/CM1/*and click the *scale* tab (Don't forget to press 'Set'), all classes will now be scaled as the selected. Exclude the last variable column called class model response, this variable will be used in exercise 5. Run PCA on all classes. Extract 2 first components in all classes by pressing select 2 components press set and OK. Display the observation names to display the animal numbers only (*start 2 length 2*). What can you observe about the repeats? How does the ***between experiment*** variation compare with the ***between animal*** variation? Is it valid to take an average for the gene expression of each group or does this lead to a loss of information?

Task 3

Select *Workset/New as model/CM1*. Change model type to PCA-X to include all observations in the same model. Fit the model using *Analysis/Two First Components*. Look at the score scatter plot. Right click on the plot and choose properties, under the *Label Types* tab select *Use Identifier* Obs ID Primary *Start* 1 *Length* 3. Do you see any outliers by Hotellings T2? Plot the DModX plot under *Analysis\Distance to Model*, showing the observation names. Make a note of the more serious outliers in the DModX plot. How do they compare with the outliers seen in the PCA-class and Quick Info plots? What can you say about the treatment groups?

Task 4

Prepare to fit a new model by choosing *Workset\New as Model 1.* Remove animal number 28 and observations C02aW,X,Y,Z; C02cY; C04aY; M29bX; L21aX; L23aY and refit the model (choosing *Analysis\Change Model Type\PCA on X Block*). Again fit the model using *Analysis\Two First Components*. Look at the score scatter plot. Under the *Label Types* tab select *Use Identifier* Obs ID Primary *Start* 1 *Length* 1. Is there a trend going from Control to Low, Medium and High?

Contribution plots are useful in the interpretation of the model. Use the default plot action tool on the toolbar to create a contribution plot (i.e., firstly click on a point in the control group and then click on a point in the high group). The resulting plot shows the difference in gene expression between the two selected observations. Use the zoom tool to zoom the X axis until you can see the individual genes.

Task 5

To observe the gene changes that occur when going from one group to another OPLS Discriminant analysis (OPLS-DA) is a good method to use. Select *Work set\New*, set scaling to pareto, exclude all previous detected outliers and use only the control and high dose class in the model (i.e. exclude the other two classes). Select the last column, class model response, to be Y. Under *work set* change the model type to OPLS/O2PLS.

Fit the model using with two components. Interpret the score plot for $t_1/t_{o1}$and the model diagnostics. Is there a good separation between the groups? To see the gene changes between these groups go to favourites and choose *S-plot*. Combine the evaluation of the S-plot with the loading plot found under *analysis/loadings/column plot.* Right click the loading plot and select *Sort ascending*. Use the zoom tool to examine the plot in more detail. Which are the genes which are most up-regulated and which are the most down-regulated?

Task 6 - Discussion Section (No solution given)

If time permits run OPLS/O2PLS on the other classes (compare control vs. low and medium dose in two separate modles) and compare shared and unique structure from the different models by using the SUS-plot.

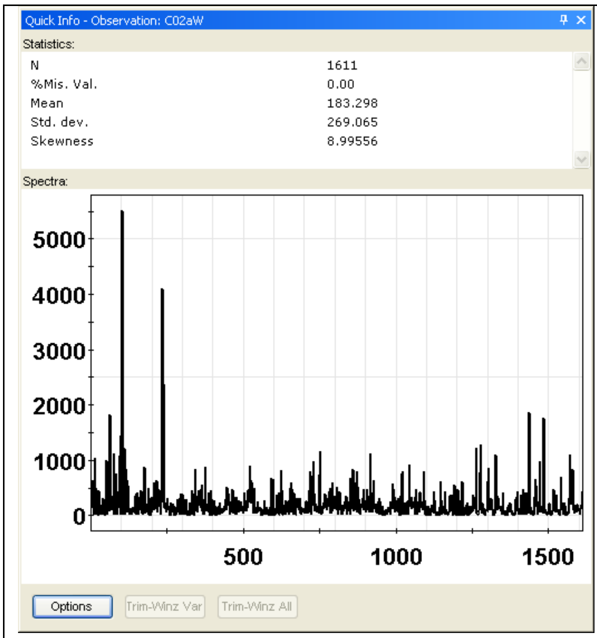
It is possible to transform the data. The data exhibit skewness, which can be made more normal by applying a Log transform (a natural choice for Fluorescence data). How does this affect the analysis?

To remove some of the systematic variation the data may be normalised to either the height of the biggest peak (akin to Mass spectra) or the total signal strength. This may be achieved in Excel by using the MAX or SUM functions and then dividing each value by the result. How does this affect the results? Are there normalisation techniques that can be used at the experimental stage?

Solutions to GeneGrid

Task 1

Using Quick Info Observations some of the observations look distinctly different. The first four observations C02aW,X,Y,Z have a noisy appearance. C02bX is more typical of the majority of observations. C02cY has a single signal which totally dominates the ‘gene spectrum’; possibly an extraneous object on the slide is causing this high point. Observations C04aY; M29bX; L21aX; L23aY are odd, as well as all observations from Animal 28 which have a very noisy appearance.

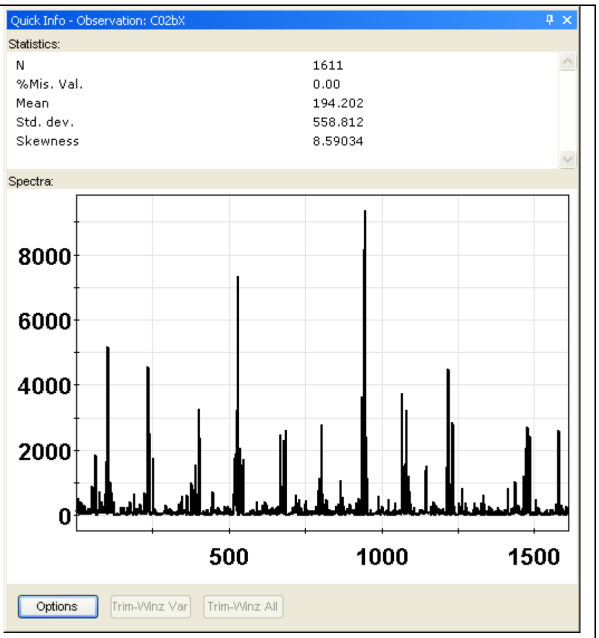


Quick Info - Observation: C02aW

Statistics:

```
N              1611
%Mis. Val.     0.00
Mean           183.298
Std. dev.      269.065
Skewness       8.99556
```

C02aW

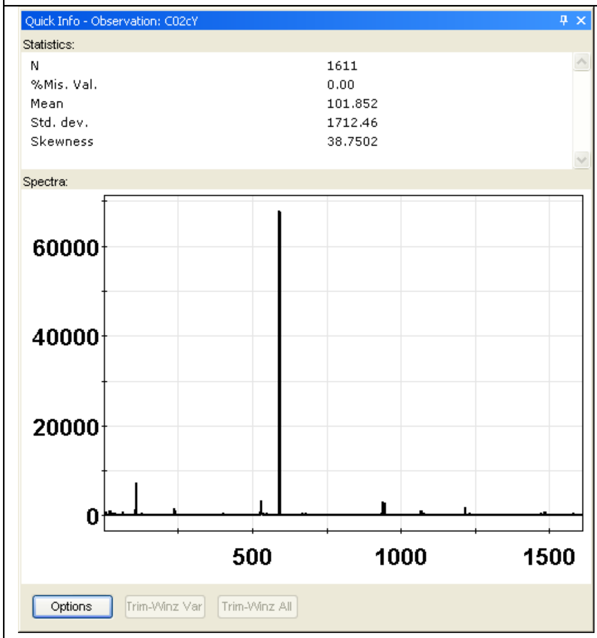


Quick Info - Observation: C02bX

Statistics:

```
N              1611
%Mis. Val.     0.00
Mean           194.202
Std. dev.      558.812
Skewness       8.59034
```

C02bX

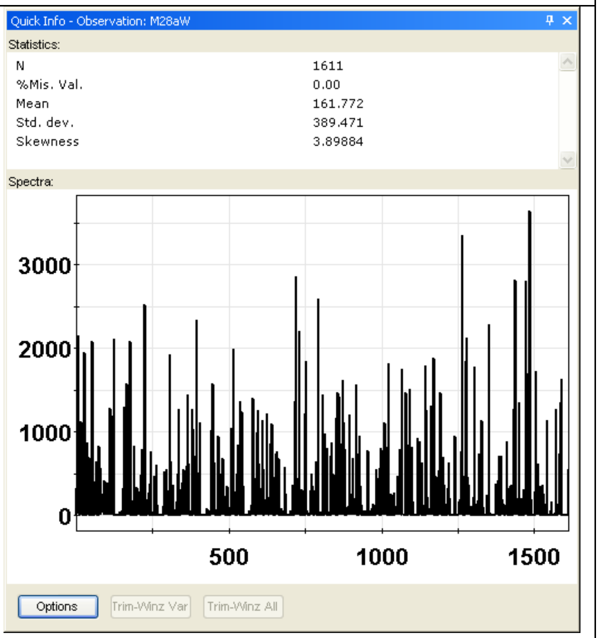


Quick Info - Observation: C02cY

Statistics:

```
N              1611
%Mis. Val.     0.00
Mean           101.852
Std. dev.      1712.46
Skewness       38.7502
```

C02cY



Quick Info - Observation: M28aW

Statistics:

```
N              1611
%Mis. Val.     0.00
Mean           161.772
Std. dev.      389.471
Skewness       3.89884
```

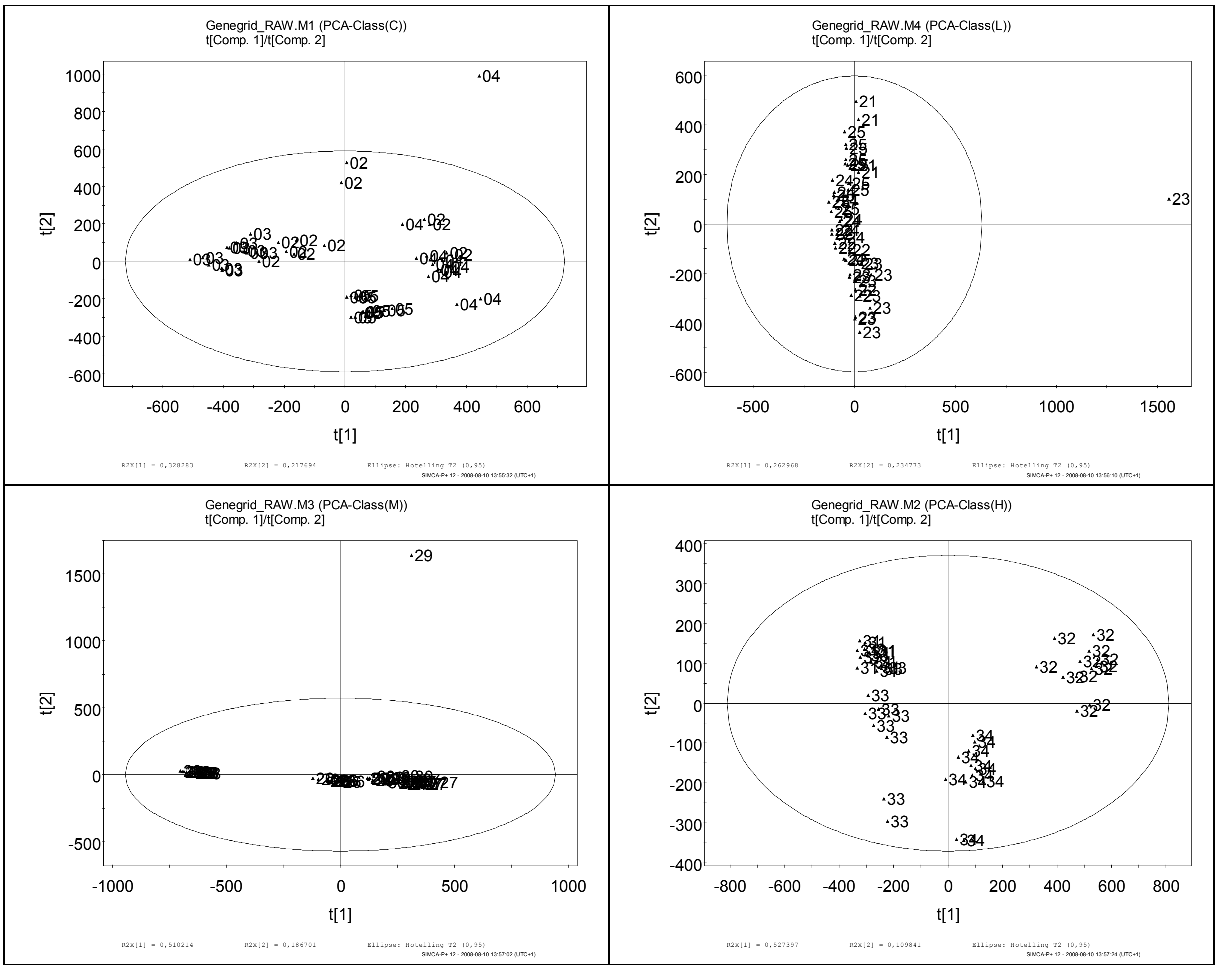
M28aW

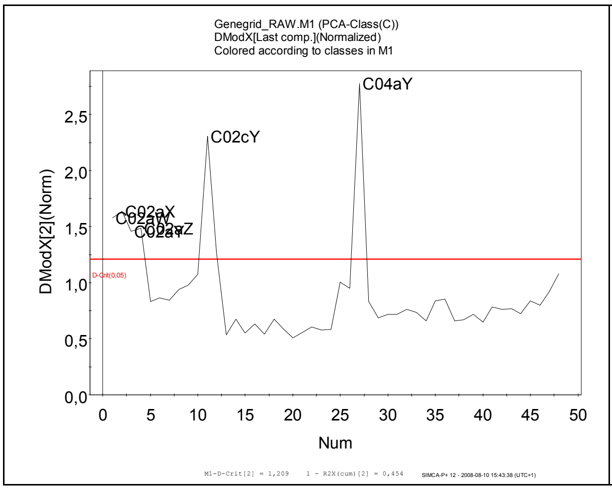
Task 2

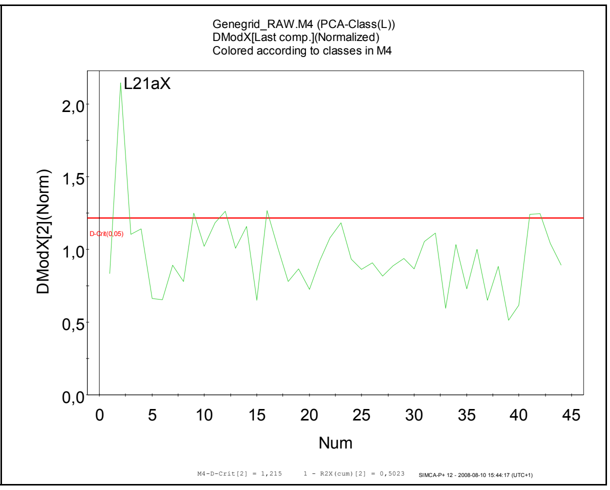
PCA on the individual classes yields the following models:

Controls	M1	PCA-Class(C)	A=2	R^2=0.546	Q^2= 0.33
Low	M4	PCA-Class(L)	A=2	R^2=0.498	Q^2= 0.2
Medium	M3	PCA-Class(M)	A=2	R^2=0.697	Q^2= 0.44
High	M2	PCA-Class(H)	A=2	R^2=0.637	Q^2= 0.59

There are several deviating samples that can be visualized in the score and DModX plots from each class. The extreme samples are C02cY; C04aY; L21aX; L23aY; M29bX; and all samples for animal 28. Additional possible outliers are C02aW,X,Y,Z. These samples are not as obvious as the other extreme samples, but by looking at the raw data (use quick info on the primary data set) it is seen that these samples profile deviate from the other samples. When these outliers are removed it is clearly seen that each treatment group exhibits a clustering of repeats for each animal (plots not shown). This shows that the *between animal* variation is greater than the *between experiment* repeats. This point to genetic variability in the animals, which is information that would be lost when averaging the treatment groups. Averaging by animal would be a way of data reduction without loosing too much information.

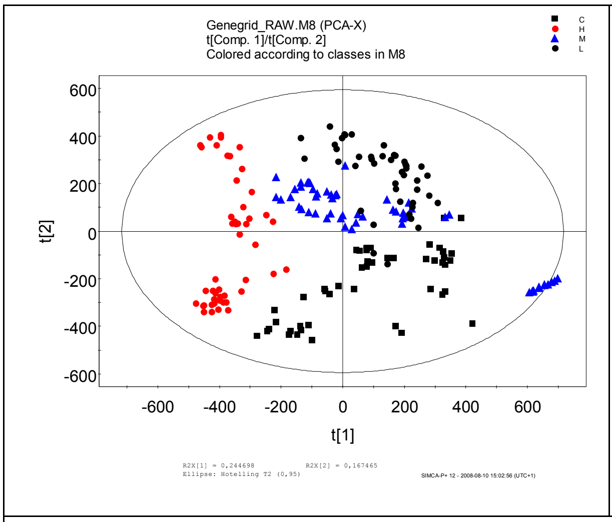


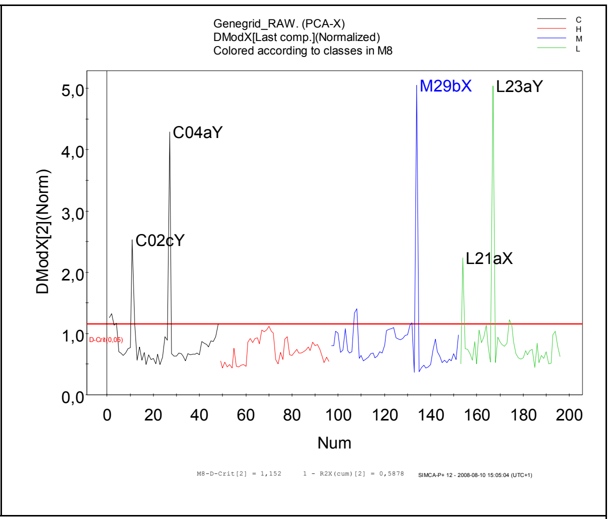


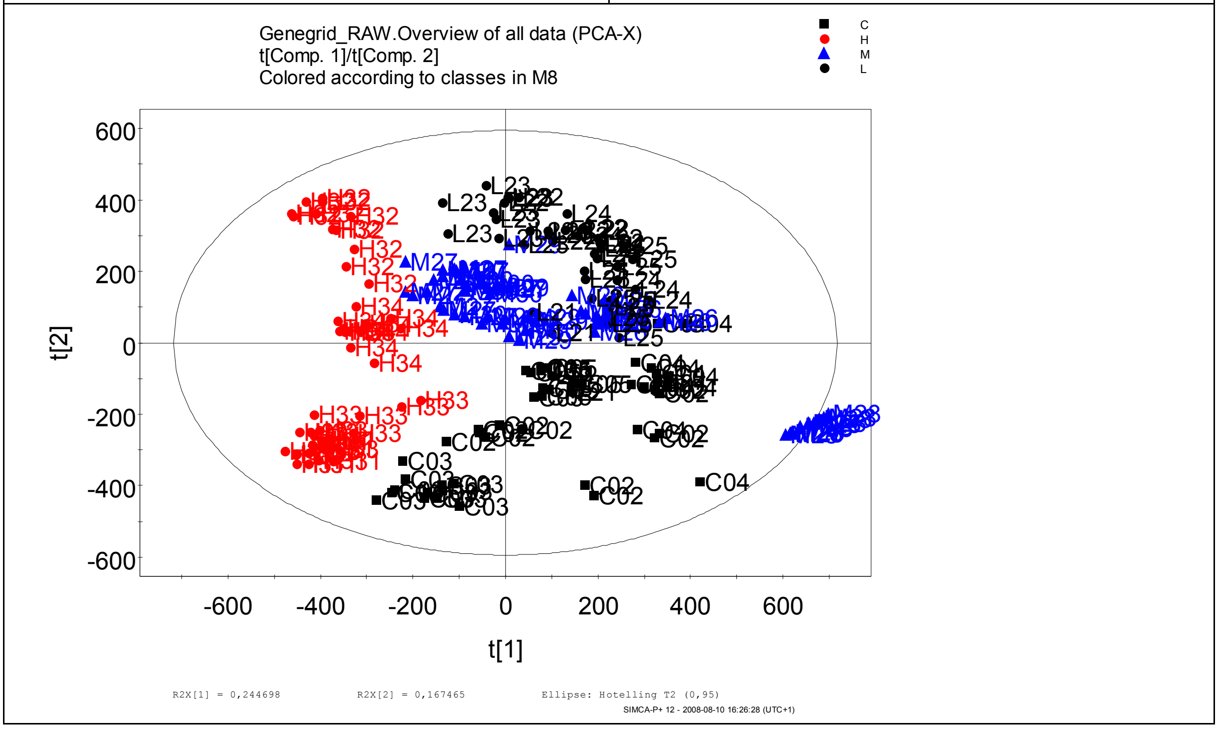


Task 3

PCA on all samples with 2 components gives $R^2 = 0.41$ and $Q^2 = 0.39$, a weak model but useful for visualisation. Hotellings T2 shows the odd behaviour of observations from animal 28. The DModX plot identifies the same outliers as those found 'bye eye' in PCA-class models. Looking at the score plot the four treatment groups show some clustering but there is also a degree of overlap.



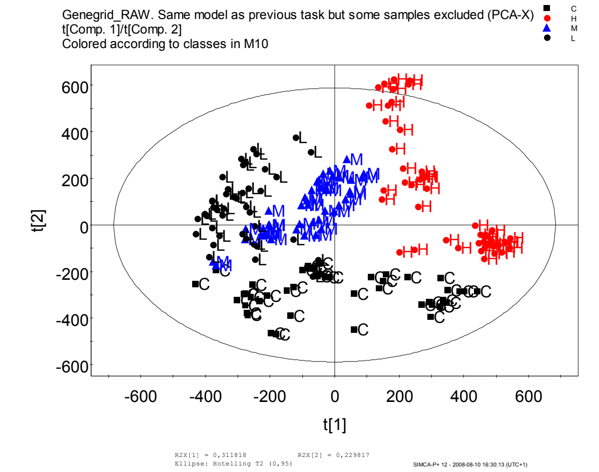




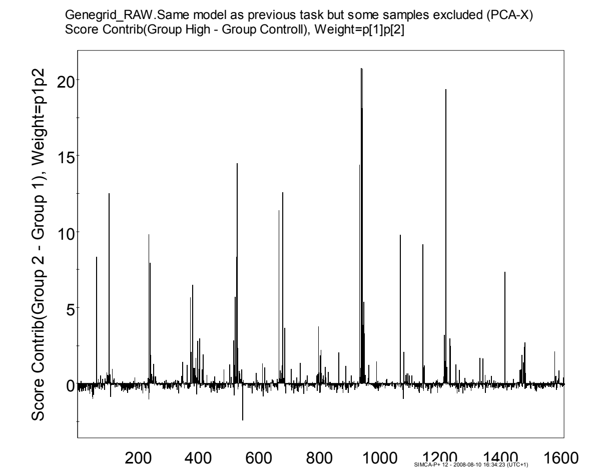
Before removing outliers it is good practice to investigate them fully to see if there is any useful information to be gained in their unusual behaviour. Animal 28 obviously is deviating from the rest in the treatment group and should be investigated. As mentioned in Task 1, observation C02cY has one spot which is dominating indicating perhaps a faulty chip or contamination. The outliers highlighted by the DModX-plot are dramatically different from the majority and so it seems reasonable to remove them.

Task 4

The updated PCA model with 2 components gives a better model by cross validation, $R^2 = 0.55$ and $Q^2 = 0.53$. The score plot is showing definite groupings with a small overlap. Differences in the gene expression between the observations are displayed in the contribution plot.



Score scatter plot

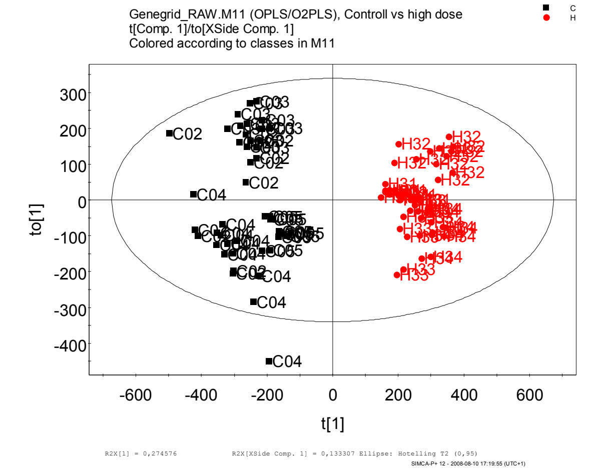


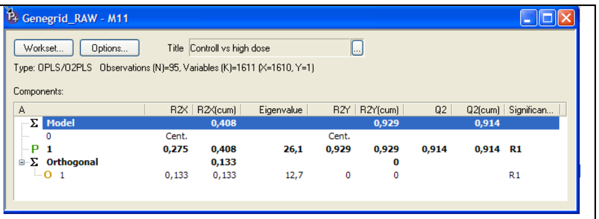
Contribution plot

To simplify the interpretation of discriminating genes it is recommended to continue the analysis with OPLS/O2PLS between two classes at time.

Task 5

OPLS-DA between Control and High dose gives a strong model with R^2=0.929 and Q^2=0.916, the predictive variation, t1, corresponds to 27.5% of all variation in the data and the uncorrelated variation, to1 (orthogonal variation), corresponds to 13.3%. The plot shows complete separation of the two groups. The S-plot shows the extent to which each gene is either up or down regulated when going from control to high dose. The loading plot shown below is sorted in ascending order. Zooming in on the plot also shows the jack-knifing confidence intervals.



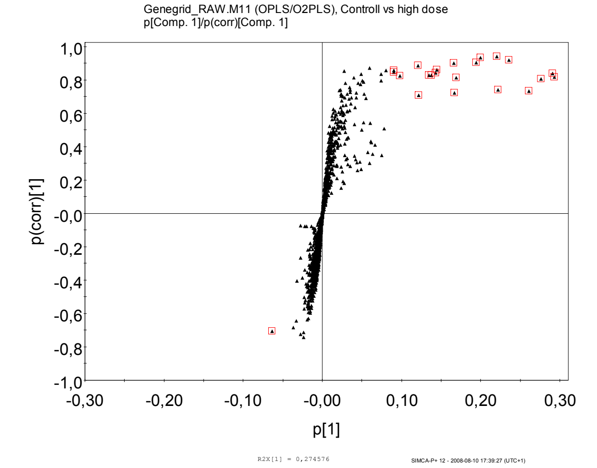


Title Controll vs high dose

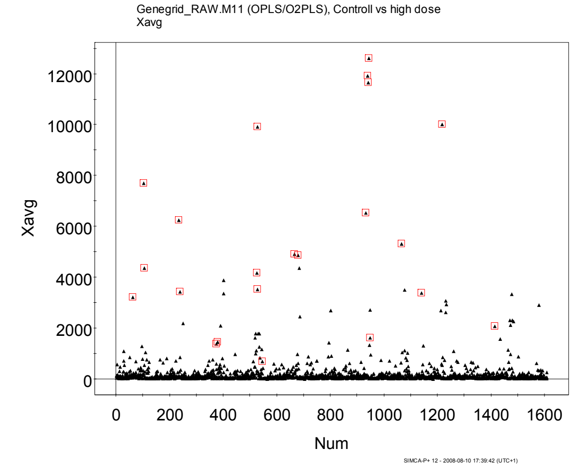
Type: OPLS/O2PLS Observations (N)=95, Variables (K)=1611 (X=1610, Y=1)

Components:

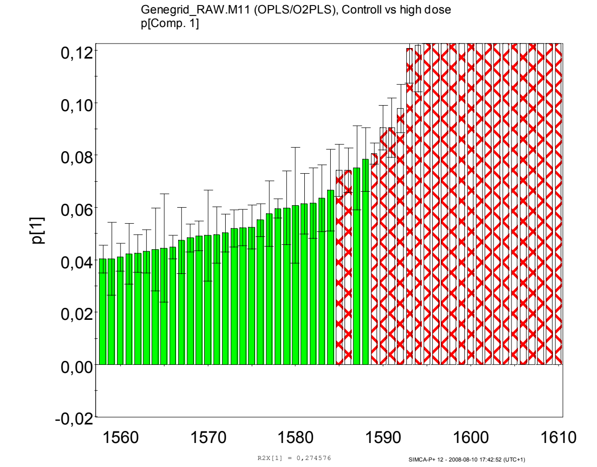
A	R2X	R2X(cum)	Eigenvalue	R2Y	R2Y(cum)	Q2	Q2(cum)	Significan...
Σ Model		0,408			0,929		0,914	
0	Cent.			Cent.				
P 1	0,275	0,408	26,1	0,929	0,929	0,914	0,914	R1
Σ Orthogonal		0,133			0			
O 1	0,133	0,133	12,7	0	0			R1



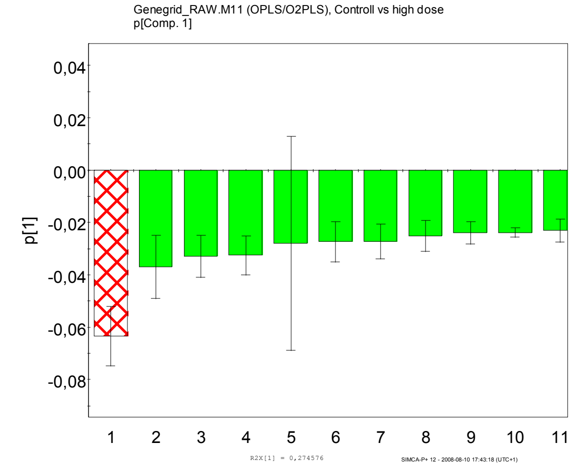
Most up and down-regulated genes.



X average plot where all selected genes from the S-plot are marked to clarify the distance to the baseline. Biomarkers close to the baseline are highly uncertain.



Zoomed region in from the P1 loading plot. Up-regulated genes are visualized.



Zoomed region from the P1 loading plot. Down-regulated genes are visualized.

Conclusions

Analytical bioinformatics data can be visualised quickly in SIMCA-P. PCA gives an overview of the data and highlights experimental variations and outliers. PCA contribution plots may be used to see which genes have altered relative to another observation. PLS-Discriminant analysis can be used to determine the differences in gene expression between treatment groups. The data may be scaled or transformed in order to optimise the separation between treatment groups or to focus on the number of genes that change.

MVDA-Exercise Ovarian Cancer

Proteomic classification of patients with ovarian cancer

Background

New techniques for early diagnosis of ovarian cancer could have a major effect on women's health. Current methods have a relatively poor positive predictive success rate of just 10-20%. In this study, proteomic spectra of blood serum were investigated as potential indicators of pathological changes in the ovaries.

We are indebted to the Food and Drug Administration's Center for Biologics Evaluation and Research for making their data available via the website:

http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp

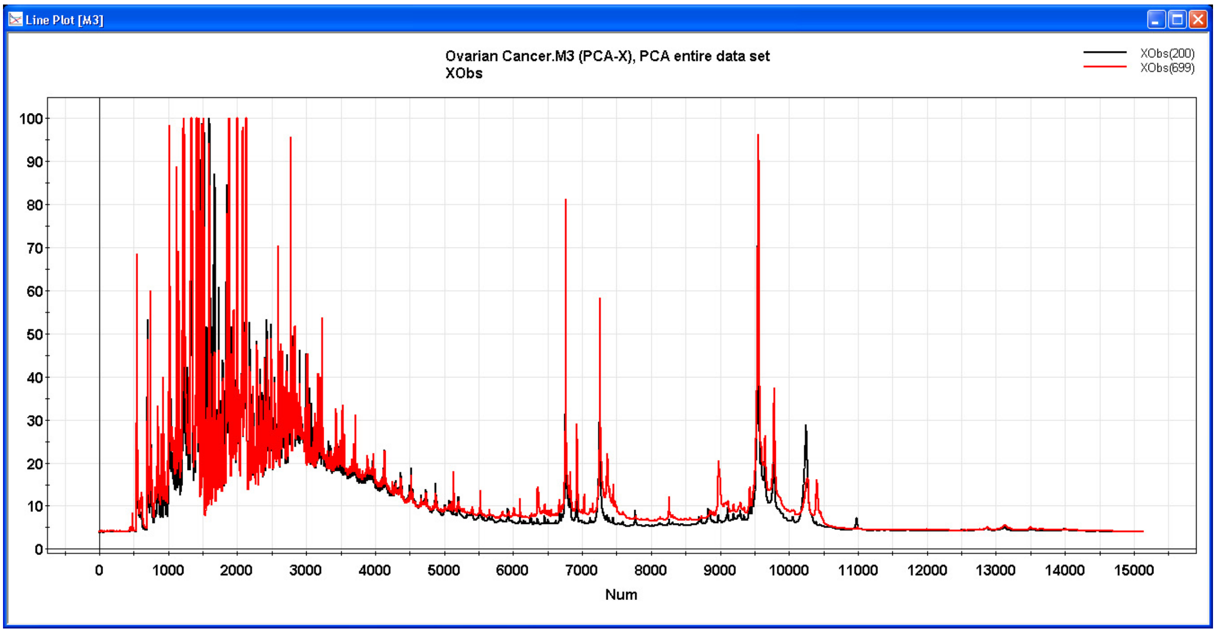
The data analysed here (8-7-02) is unpublished work based on low resolution surface-enhanced laser desorption and ionisation time-of-flight (SELDI-TOF) mass spectroscopy. The original dataset contains spectra for 91 unaffected women and 162 patients with biopsy-proven ovarian cancer. Of the latter group, only the first 100 numbered patients were included in this exercise.

Objective

The objective of this exercise is to assess how well proteomics spectra can discriminate between ovarian cancer patients and unaffected women (controls).

Data

The dataset contains MS spectra for 191 individuals, 91 unaffected women (controls) and 100 ovarian cancer patients. The spectra consist of 15154 M/Z values. The spectra from a randomly chosen control (200) and cancer patient (699) are shown below for illustration.



Tasks

Task 1

Open SIMCA-P and import *Ovarian Cancer.dif.*

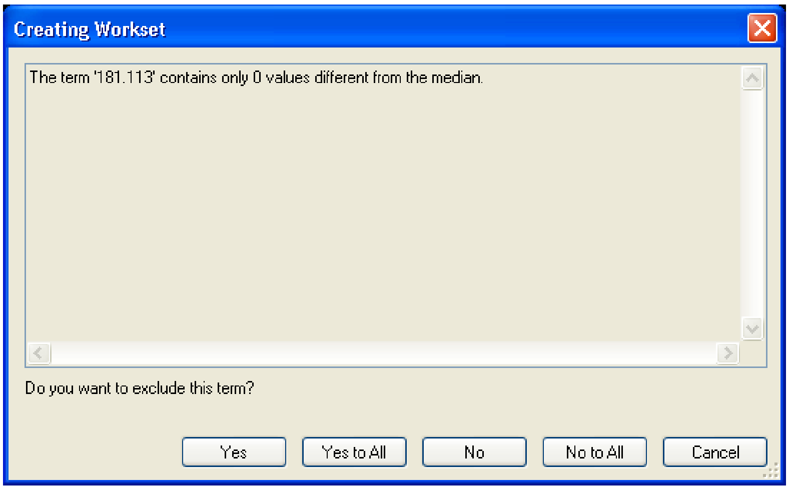
Mark column 1 as the primary observation ID and column 2 as ClassID and 3 as secondary observation ID. Assign the fourth column (labelled "DA") as Y variable. When assigning Class ID you will in the next step see the number of observations in each class. This feature will also simplify PCA for individual classes.

Mark row 1 as the primary variable ID, this contains the M/Z values. Check that the data set contains 191 observations, 15155 variables and no missing values. After import SIMCA-P+12 will automatically generate two data sets under *CM1* for PLS-class. However this is not what should be done in the first exercise.

To change this, go to *Workset/Edit CM1* and change the scaling to Pareto by selecting all the variables, highlighting *Par* and clicking on *Set*. When choosing *CM1* for edit, all sub datasets will be edited simultaneously. Pareto scaling works well for MS spectra as it offers a compromise between no scaling and unit variance scaling. Change model type to PCA-class.

SIMCA-P will now be ready to fit a PCA model on both classes.

SIMCA-P will ask you whether you wish to exclude a few variables that have zero variance. Accept the exclusion by clicking on Yes to all. These variables are constant for all samples and are therefore of no interest.

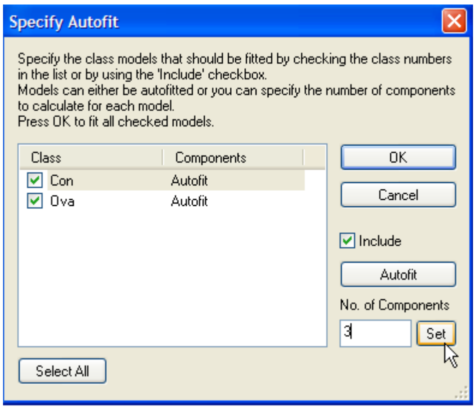

Creating Workset

The term '181.113' contains only 0 values different from the median.

Do you want to exclude this term?

Task 2

Build separate PCA models of each class by *specify autofit* to 3 components.


Specify Autofit

Specify the class models that should be fitted by checking the class numbers in the list or by using the 'Include' checkbox.
Models can either be autofitted or you can specify the number of components to calculate for each model.
Press OK to fit all checked models.

Class	Components
☑ Con	Autofit
☑ Ova	Autofit

☑ Include

No. of Components: 3

Remember to press *set* followed by *OK to* fit the PCA model.

Are there any outliers in either class?

Task 3

Make an PCA overview with all 191 samples together by marking *CM1* and selecting *Analysis/Change Model Type/PCA on X-block.* Fit the model with three principal components by selecting *Analysis/Two First Components+ next component*. Plot the scores. How well are the two groups separated? Are there any outliers?

Task 4 OPLS-DA

A training set was selected for each class using the principles of multivariate design. A 4^3 factorial design embracing 64 combinations of the first three principal components of each class was used. Only samples that corresponded to points in the design were selected with a limit of one sample per design point. This resulted in a training set of 43 controls and 54 cancer patients as follows:

Controls:

182, 183, 184, 185, 188, 193, 195, 198, 199, 208, 209, 212, 217, 218, 222, 223, 224, 225, 226, 230, 231, 233, 236, 240, 241, 242, 245, 246, 248, 250, 251, 252, 254, 257, 258, 259, 260, 261, 263, 265, 266, 274, 281

Cancer Patients:

601, 602, 605, 606, 613, 614, 618, 620, 621, 623, 624, 626, 628, 629, 630, 631, 632, 636, 638, 640, 641, 642, 643, 644, 647, 651, 652, 653, 655, 659, 661, 662, 663, 664, 666, 667, 669, 671, 679, 681, 683, 688, 689, 691, 693, 694, 701, 702, 704, 705, 706, 707, 708, 710

The remaining 48 controls and 46 cancer patients will form the test set to assess the true predictive power of the models.

Do the training sets defined above constitute a diverse and representative subset of each class?

Create a new workset containing only the training samples by excluding all samples except those listed above. Set Y under *workset*/*variables/*select DA and press Y, set scaling to *pareto*. Select *OPLS/O2PLS* and fit the model. Plot the scores.

How well are the classes separated? The separation can be visualized using *Analysis/Observed vs. Predicted* which is based on all the components. The control samples should all have predicted values above 0.5 and the ovarian cancer samples predicted values less than 0.5. The cross-validated score plot is also recommended for evaluation of the predictive ability.

Find the most important potential biomarkers for discrimination between control and ovarian cancer. Use the S-plot and the loading plot with confidence interval.

Right-click on the loading plot and choose *Sort Ascending* on the values. If you wish, remove the confidence intervals, right-click on the plot go to *Properties* and select *Confidence level None*.

Make a list of potential biomarkers. In the S-plot, mark the most important biomarkers, right-click on the plot and select *Create List*. The masses with the largest p1 and largest p(corr)1 are the most important for classification purposes and could be used to provide biomarkers of the disease.

Task 5

Use the 94 test samples to validate the model built in Task 4. Define the prediction set using *Predictions/Specify Predictionset/Complement Workset*. This will bring in the samples not used to train the model.

With an OPLS-DA model: List the predicted values for the test set using *Predictions/Y Predicted/List*.

With a conventional OPLS model using the Y-variable: Plot observed versus predicted to visualize the model results.

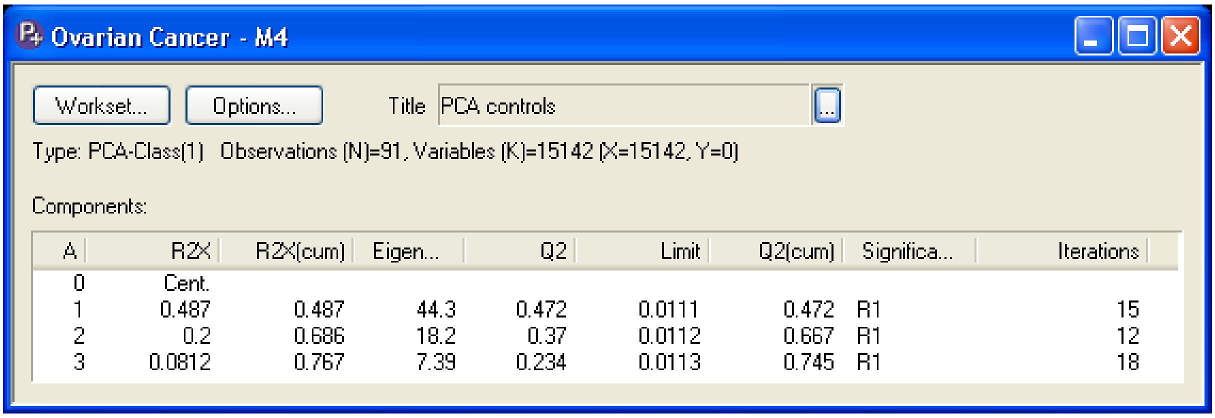
How many samples are correctly classified?

Summarize the predictions using the misclassification list. To be able to do this table in SIMCA-P+12 you must do an OPLS/O2PLS-DA model instead of using the Y response for discrimination. Under *workset/new as model* (select the OPLS model used in previous results)/*OPLS/O2PLS-DA*. Fit the model using same number of components as previous model, then both models will be identical. Select the test set under *predictions/complement workset*. Make the misclassification table under *predictions/misclassification*.

Solutions to Ovarian Cancer

Task 2

The two class models are summarised below.



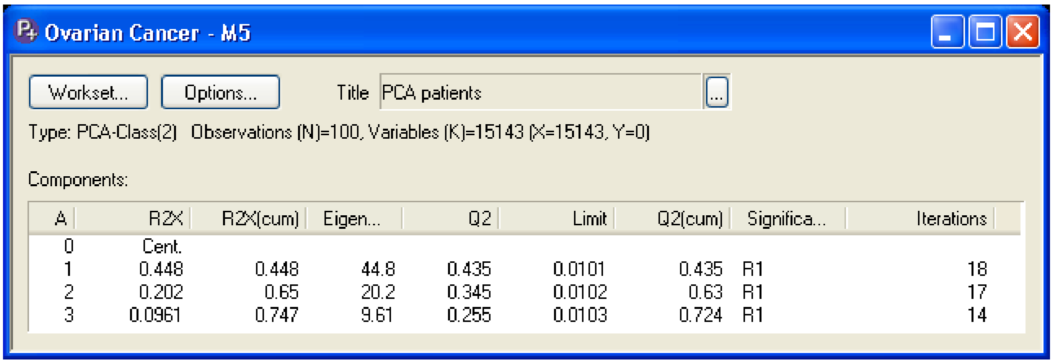
Ovarian Cancer - M4

Workset... Options... Title PCA controls

Type: PCA-Class(1) Observations (N)=91, Variables (K)=15142 (X=15142, Y=0)

Components:

A	R2X	R2X(cum)	Eigen...	Q2	Limit	Q2(cum)	Significa...	Iterations
0	Cent.							
1	0.487	0.487	44.3	0.472	0.0111	0.472	R1	15
2	0.2	0.686	18.2	0.37	0.0112	0.667	R1	12
3	0.0812	0.767	7.39	0.234	0.0113	0.745	R1	18



Ovarian Cancer - M5

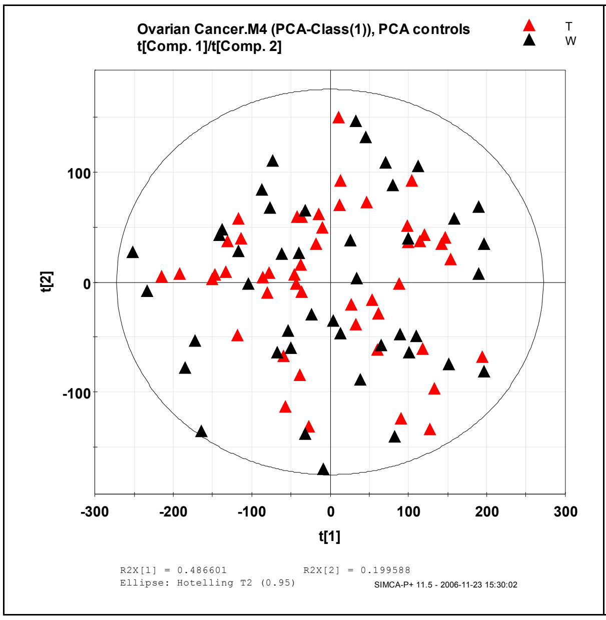
Workset... Options... Title PCA patients

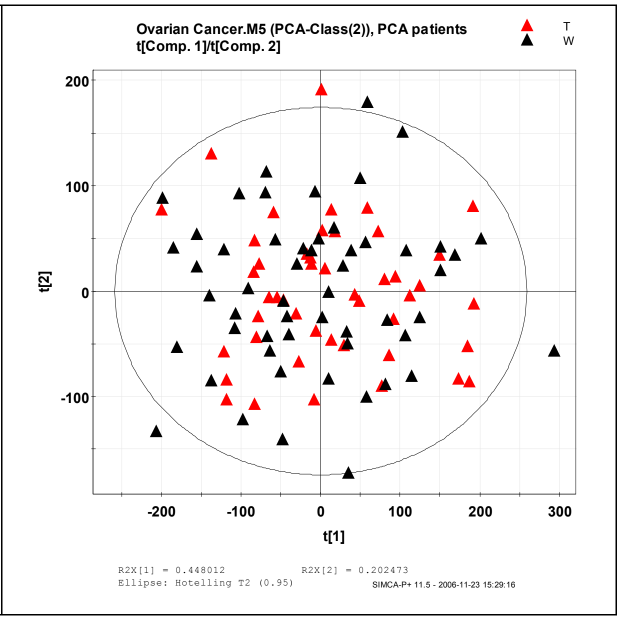
Type: PCA-Class(2) Observations (N)=100, Variables (K)=15143 (X=15143, Y=0)

Components:

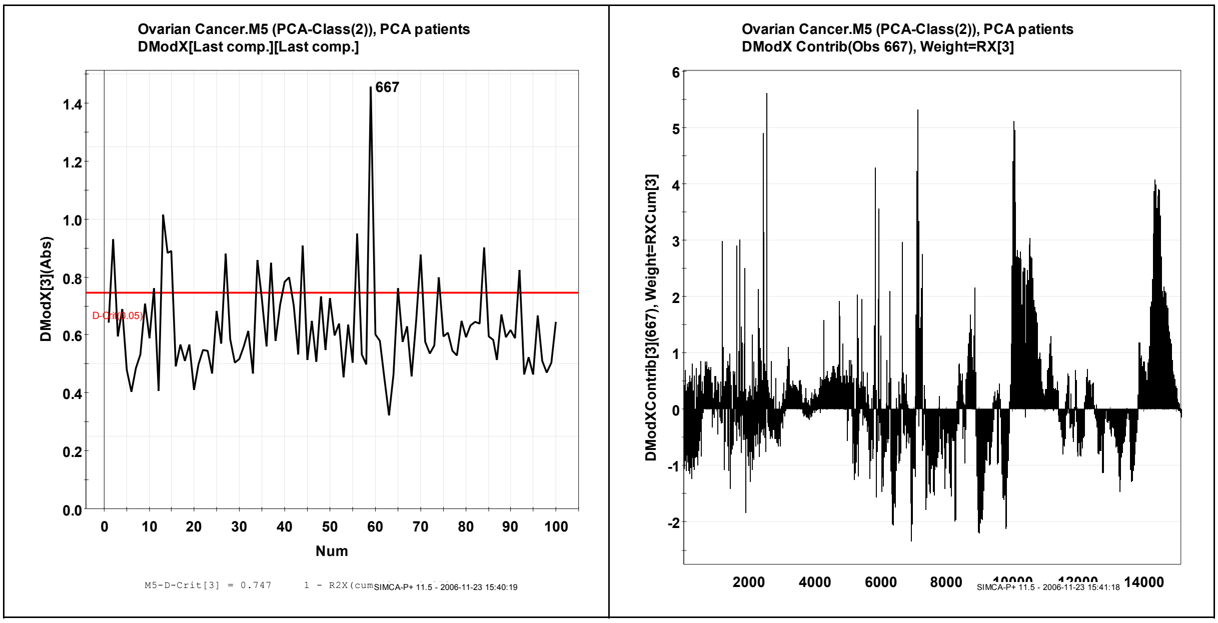
A	R2X	R2X(cum)	Eigen...	Q2	Limit	Q2(cum)	Significa...	Iterations
0	Cent.							
1	0.448	0.448	44.8	0.435	0.0101	0.435	R1	18
2	0.202	0.65	20.2	0.345	0.0102	0.63	R1	17
3	0.0961	0.747	9.61	0.255	0.0103	0.724	R1	14

Plots of t1vs t2 for each class are shown below. In these plots, the samples are color-coded according to work set (W) and test set (T) membership. This confirms that the training sets are truly representative of each class.



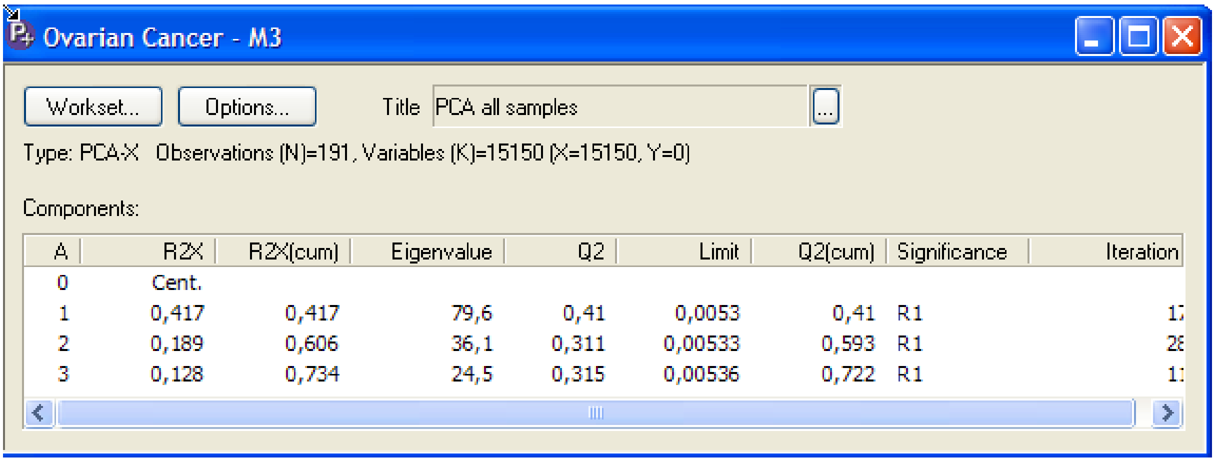


There are no strong outliers in either class. However, individual 667 has a rather high DModX in the ovarian cancer class. The corresponding contribution plot suggests that this sample has somewhat higher spectral values for some of the larger masses.



Task 3

An overview model of all 191 samples with three principal components is given below.



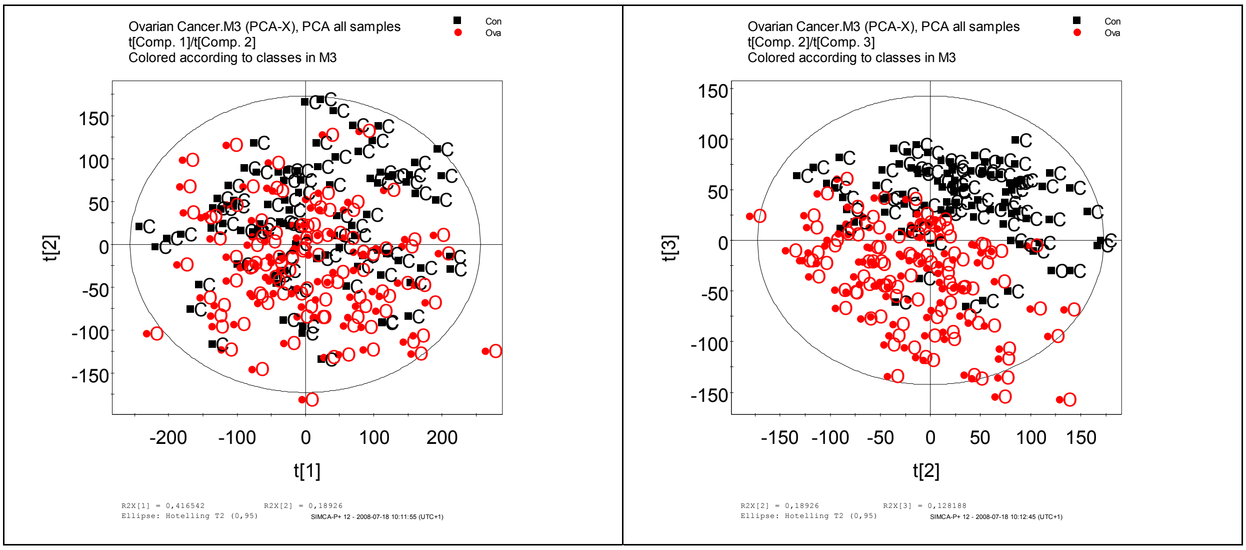
Title PCA all samples

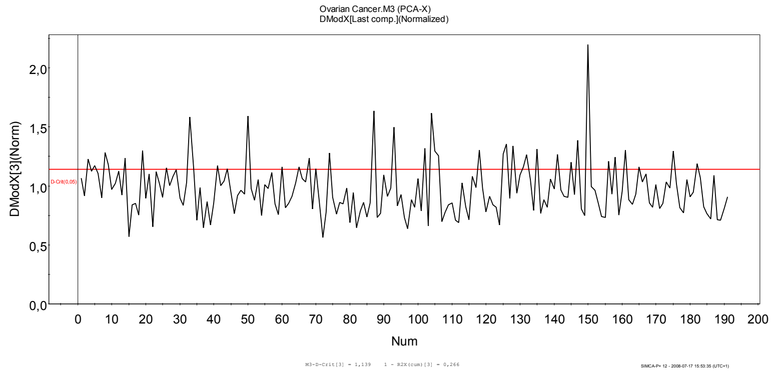
Type: PCA-X   Observations (N)=191, Variables (K)=15150 (X=15150, Y=0)

Components:

A	R2X	R2X(cum)	Eigenvalue	Q2	Limit	Q2(cum)	Significance	Iteration
0	Cent.							
1	0,417	0,417	79,6	0,41	0,0053	0,41	R1	17
2	0,189	0,606	36,1	0,311	0,00533	0,593	R1	28
3	0,128	0,734	24,5	0,315	0,00536	0,722	R1	11

The scores plot shows some separation of the controls (C) and ovarian cancer patients (O) along the second and third principal component, although the two groups clearly overlap. There are no strong outliers. There are a few moderate outliers, mainly from the cancer group.

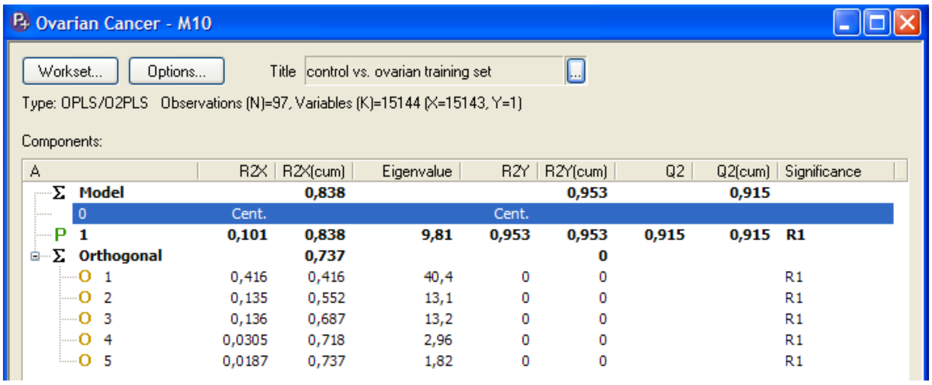




Task 4

OPLS-DA gives six (1 predictive + 5 orthogonal) components with R2Y=0.95 and Q2=0.92. The plot of t1 vs. to1 indicates a separation of the two classes. This separation must be verified by Q2, the cross-validated score plot and the Observed vs. Predicted plot based on all six components. The plots are important as they make the predictive ability more transparant.

There is complete separation of the two classes with all ovarian cancer samples to the left of 0.5 and all controls to the right of 0.5.

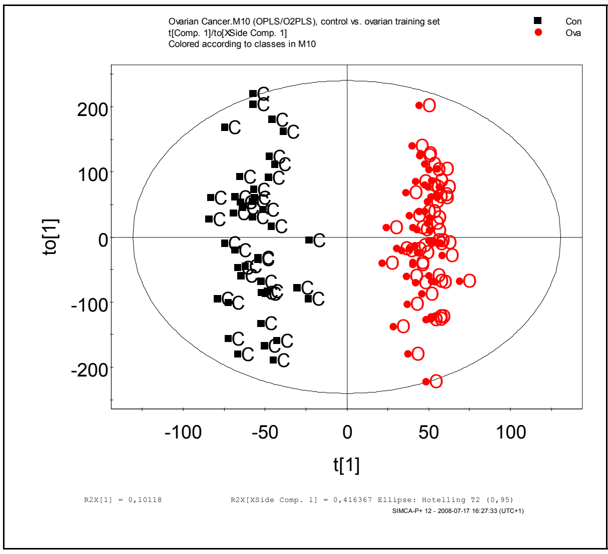


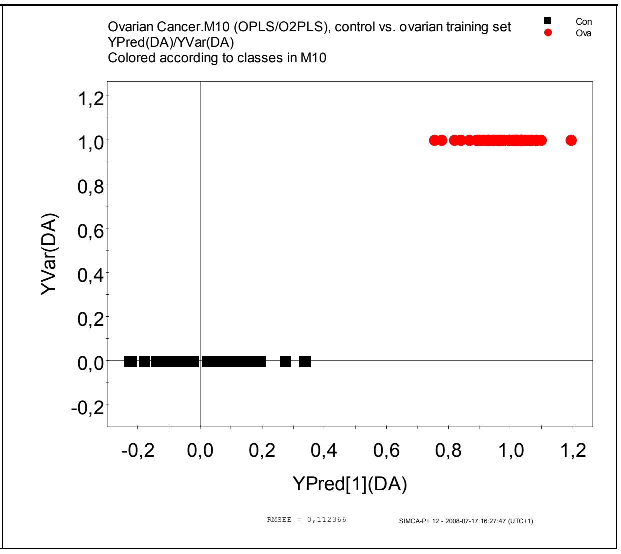
Title control vs. ovarian training set

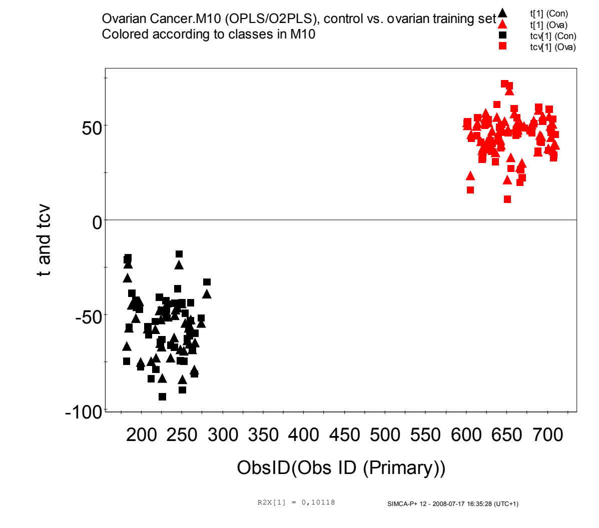
Type: OPLS/O2PLS Observations (N)=97, Variables (K)=15144 (X=15143, Y=1)

Components:

A	R2X	R2X(cum)	Eigenvalue	R2Y	R2Y(cum)	Q2	Q2(cum)	Significance
Σ **Model**		**0,838**			**0,953**		**0,915**	
0	Cent.			Cent.				
P **1**	**0,101**	**0,838**	**9,81**	**0,953**	**0,953**	**0,915**	**0,915**	**R1**
Σ **Orthogonal**		**0,737**			**0**			
O 1	0,416	0,416	40,4	0	0			R1
O 2	0,135	0,552	13,1	0	0			R1
O 3	0,136	0,687	13,2	0	0			R1
O 4	0,0305	0,718	2,96	0	0			R1
O 5	0,0187	0,737	1,82	0	0			R1

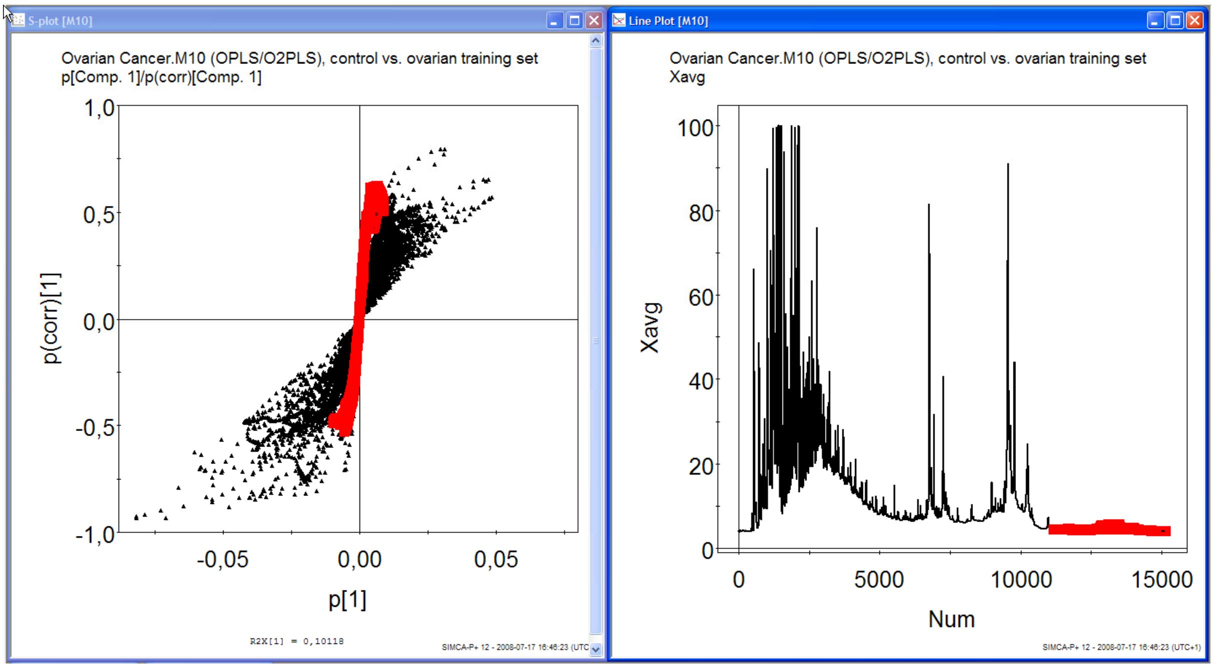


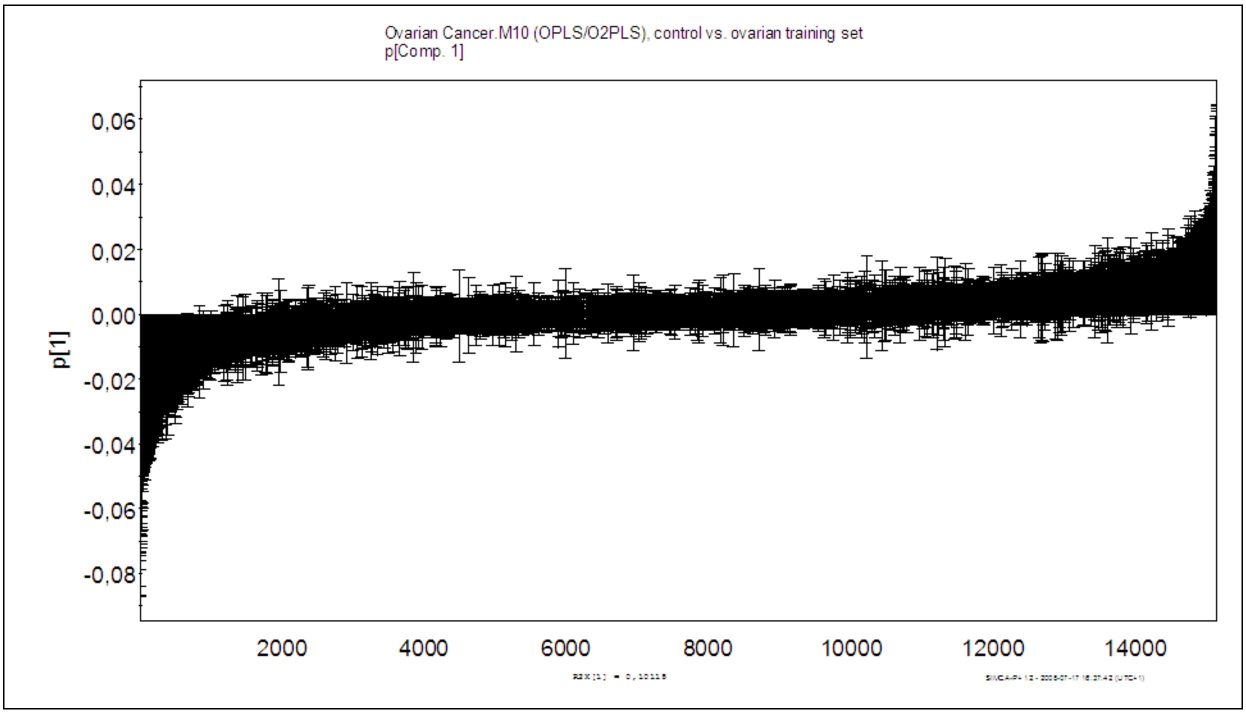




In the CV-score plot it is seen that all samples were predicted to its own class during cross validation.

To extract potential biomarkers, the S-plot was evaluated together with the X average plot and the sorted loading plot with the 95% confidence intervals, shown below. By marking the baseline in the Xave plot it is easy to visualize the same region in the S-plot. All potential biomarkers close to this line are highly uncertain and should not be considered significant although they have a high p(corr)1 value.


S-plot [M10]
Ovarian Cancer.M10 (OPLS/O2PLS), control vs. ovarian training set
p[Comp. 1]/p(corr)[Comp. 1]
p(corr)[1]
1,0
0,5
0,0
-0,5
-1,0
-0,05
0,00
0,05
p[1]
R2X[1] = 0,10118
SIMCA-P+ 12 - 2008-07-17 16:46:23 (UTC
Line Plot [M10]
Ovarian Cancer.M10 (OPLS/O2PLS), control vs. ovarian training set
Xavg
Xavg
100
80
60
40
20
0
0
5000
10000
15000
Num
SIMCA-P+ 12 - 2008-07-17 16:46:23 (UTC+1)

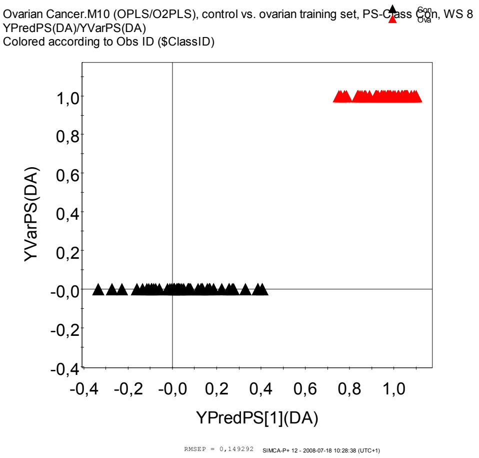


Based on the S-plot and the loading plots, the 18 most important masses are listed below. Negative p(corr)[1] refer to masses associated with down regulation in the ovarian samples and positive are associated with up-regulation in the ovarian cancer samples. As seen in the table the p(corr)[1] value are different for each M/Z. It is wise to divide the effect size (ES) into small medium and large ES if many potential biomarkers appear.

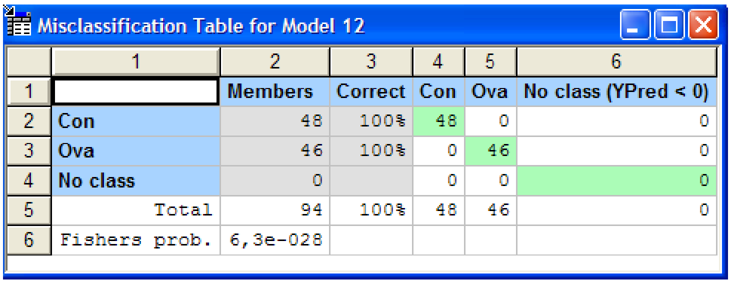
M/Z	**Loading p[1]**	**p(corr)[1]**
25,4014	-0,0536253	-0,778866
25,4956	-0,0562088	-0,83488
25,5899	-0,0539473	-0,845497
25,6844	-0,0502256	-0,824039
221,862	-0,0661238	-0,792194
244,66	-0,070749	-0,931777
244,952	-0,081619	-0,933421
245,245	-0,0816744	-0,926634
245,537	-0,078534	-0,915735
245,83	-0,0728345	-0,898917
246,122	-0,0643447	-0,877963
246,415	-0,0538228	-0,857674
434,297	0,0274748	0,722155
434,686	0,0308307	0,771015
435,075	0,0313782	0,794713
435,465	0,0296506	0,795315
435,854	0,0261589	0,7835
436,244	0,0218106	0,752144

Task 5

All 94 members of the test set are correctly classified, see obs/pred plot and classification table below.



The misclassification table will calculate the number of correctly classified samples.



Misclassification Table for Model 12

	1	2	3	4	5	6
1		Members	Correct	Con	Ova	No class (YPred < 0)
2	Con	48	100%	48	0	0
3	Ova	46	100%	0	46	0
4	No class	0		0	0	0
5	Total	94	100%	48	46	0
6	Fishers prob.	6,3e-028				

This table summarizes all predictions from the test set. The table indicates that all samples were correctly predicted by the model.

Controls	Con Pred	Ova Pred	Ovarian	Con Pred	Ova Pred
Con 181	1.00	0.00	Ova 604	0.03	0.97
Con 186	0.89	0.11	Ova 608	0.08	0.92
Con 189	0.98	0.02	Ova 609	-0.10	1.10
Con 190	1.08	-0.08	Ova 610	-0.06	1.06
Con 191	0.84	0.16	Ova 612	0.05	0.95
Con 192	0.97	0.03	Ova 615	0.03	0.97
Con 194	1.11	-0.11	Ova 617	0.04	0.96
Con 196	0.83	0.17	Ova 619	0.16	0.84
Con 197	0.72	0.28	Ova 622	-0.08	1.08
Con 200	1.27	-0.27	Ova 625	0.24	0.76
Con 201	0.98	0.02	Ova 627	0.15	0.85
Con 202	0.95	0.05	Ova 633	-0.05	1.05
Con 204	1.13	-0.13	Ova 634	0.06	0.94
Con 205	0.93	0.07	Ova 635	-0.06	1.06
Con 207	0.60	0.40	Ova 639	-0.02	1.02
Con 210	0.96	0.04	Ova 646	0.08	0.92
Con 211	0.97	0.03	Ova 648	-0.09	1.09
Con 214	0.92	0.08	Ova 654	-0.05	1.05
Con 215	0.99	0.01	Ova 656	0.01	0.99
Con 216	1.08	-0.08	Ova 657	0.13	0.87
Con 220	1.09	-0.09	Ova 658	0.15	0.85
Con 221	1.34	-0.34	Ova 660	-0.04	1.04
Con 227	0.87	0.13	Ova 665	0.13	0.87
Con 228	1.09	-0.09	Ova 668	0.01	0.99
Con 229	0.93	0.07	Ova 670	0.07	0.93
Con 234	0.87	0.13	Ova 672	-0.05	1.05
Con 235	0.73	0.27	Ova 673	0.25	0.75
Con 237	0.61	0.39	Ova 674	-0.06	1.06
Con 239	0.87	0.13	Ova 675	-0.09	1.09
Con 243	0.74	0.26	Ova 676	0.08	0.92
Con 244	1.01	-0.01	Ova 677	-0.06	1.06
Con 247	1.02	-0.02	Ova 678	0.22	0.78
Con 253	1.23	-0.23	Ova 680	0.05	0.95
Con 255	0.62	0.38	Ova 682	0.00	1.00
Con 256	0.87	0.13	Ova 686	-0.05	1.05
Con 262	0.59	0.41	Ova 687	0.16	0.84
Con 264	1.16	-0.16	Ova 692	-0.02	1.02
Con 267	1.00	0.00	Ova 696	0.11	0.89
Con 269	0.92	0.08	Ova 697	0.21	0.79
Con 270	0.78	0.22	Ova 698	0.01	0.99
Con 271	0.97	0.03	Ova 699	0.00	1.00
Con 273	0.81	0.19	Ova 703	-0.04	1.04
Con 275	0.67	0.33	Ova 709	0.02	0.98
Con 276	1.12	-0.12	Ova 711	0.05	0.95
Con 277	1.06	-0.06	Ova 712	0.03	0.97
Con 278	0.99	0.01	Ova 713	-0.04	1.04
Con 279	0.74	0.26			
Con 280	0.83	0.17			

Conclusions

The use of proteomics data to discriminate between ovarian cancer patients and unaffected women works extremely well with this dataset from the Food and Drug Administration's Center for Biologics Evaluation and Research. A model was built on 43 controls and 54 ovarian cancer samples. This was used to correctly classify a test set of 48 controls and 46 ovarian cancer samples. The application of chemometric techniques like OPLS Discriminant Analysis clearly has a major role to play in new research areas such as proteomics, genomics and metabonomics.

Generation of the mass spectra used in this study requires just a small sample of blood serum that can be obtained with a pin-prick. This highlights the potential of proteomics as a screening tool for diseases such as ovarian cancer in the general population which could yield a quantum leap in terms of quality of life.

MVDA-Exercise METABOLOMICS with OPLS

Comparing PCA with OPLS in Metabolomics

Background

A gene encoding a MYB transcription factor, with unknown function, *PttMYB76*, was selected from a library of poplar trees for metabolomic characterization of the growth process in Poplar trees.

Objective

The objective of this exercise is to shed some light on how PCA and OPLS-DA may be used in state-of-the-art Metabolomics. In particular, the objectives are to:

- Demonstrate how PCA can be used to look at patterns and trends
- Demonstrate the strength of OPLS-DA compared to PCA
- Describe the model diagnostics of an OPLS model

Data

In total, the data set contains N = 57 observations, 6 trees devided into segments of 8 by the internode of the tree plus analytical replicates and K = 655 variables (^{1}H-NMR chemical shift regions bucket with 0.02ppm). The internode represents the growth direction of a plant. Internode 1 is the top of the plant and 8 is the bottom. The observations (trees) are divided in two groups ("classes"):

- MYB76 poplar plant (A*i*, B*i*, C*i*)
- Wild type Poplar plant (D*i*, E*i*, F*i*)

The name settings A, B, C corresponds to MYB76 plants and D, E, F to the wild type (control) plants. The *i* after the letter corresponds to the internode number of the plant. The last 12 experiments in the data set are analytical replicates i.e. samples that was run two times in the spectrometer. The analytical replicates are marked with a r1 or r2 after the internode number.

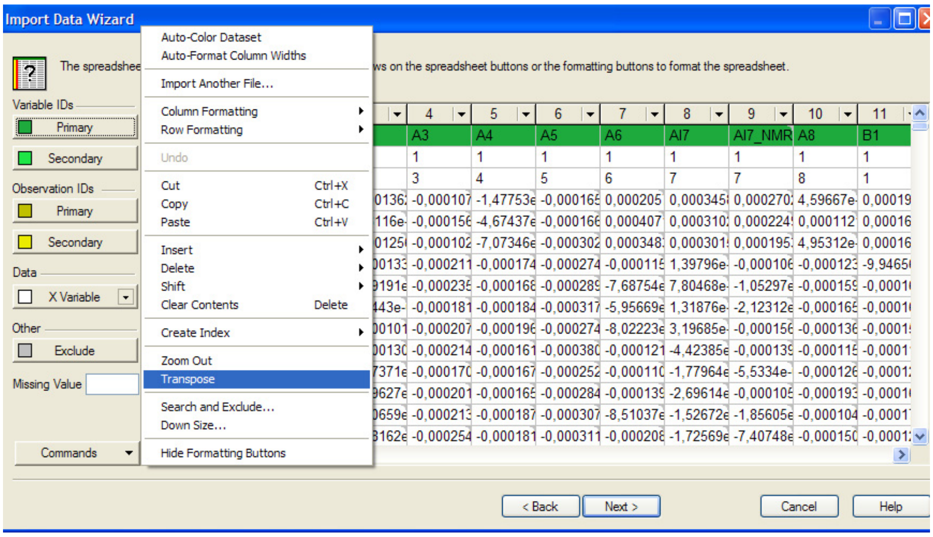
The plant material were analyzed by a 500 MHz NMR spectrometer equipped with a HR/MAS probe. The ^{1}H NMR spectra were reduced by summation of all the data points over a 0.02 ppm region. Data points between 4.2- 5.6 ppm, corresponding to water resonances, were excluded, leaving a total of 655 NMR spectral regions as variables for the multivariate modelling. A more elaborate account of the experimental conditions is found in [1].

1) S. Wiklund et.al A new metabonomic strategy for analysing the growth process of the poplar tree. Plant Biotechnology Journal 2005 3 pp 353-362

Tasks

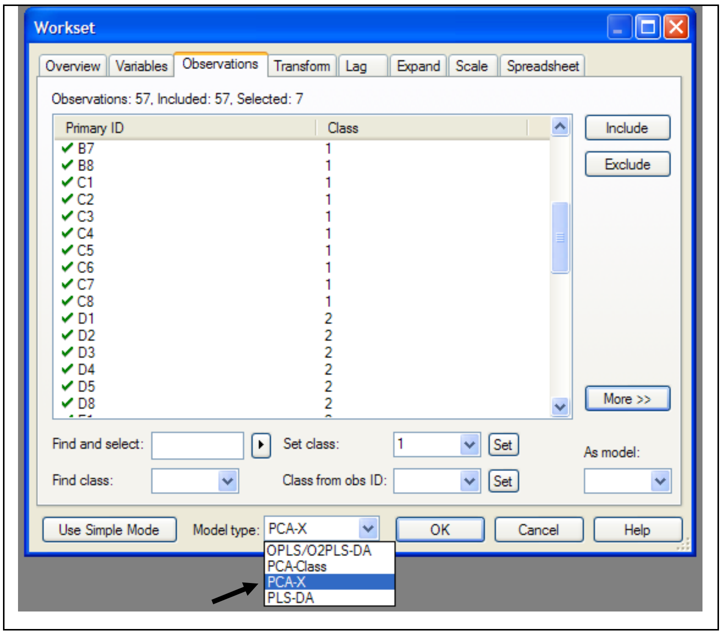
Task 1 (PCA)

Import the file NMR METABOLOMICS_PCA vs OPLSDA.xls and create a SIMCA-P project. The imported file must be transposed before saving the project. In the Import Data Wizard, go to *commands/transpose*, as demonstrated in the figure. Mark the first row and select *primary* variable id. Make sure that the first column is marked as primary observation IDs. In the second column you can see that the data has been extended to designate the different classes, this column will be used as a response vector y in OPLS-DA regression. Mark this as Y in the data set. It is recommended to create this discriminating vector although it is possible to define classes in SIMCA. The simple reason why to do this is due to a risk that SIMCA might flip the vectors in different models and this could confuse interpretation of multiple classes. When choosing creating a class vector where the control is 0 and treated is 1, it is guarantied that the vectors will not flip and comparing data from multiple models will be less complicated.



	4	5	6	7	8	9	10	11
	A3	A4	A5	A6	AI7	AI7_NMR	A8	B1
	1	1	1	1	1	1	1	1
	3	4	5	6	7	7	8	1
0136	-0,000107	-1,47753e	-0,000165	0,000205	0,000345	0,000270	4,59667e-	0,00019
116e-	-0,000156	-4,67437e	-0,000166	0,000407	0,000310	0,000224	0,000112	0,00016
0125	-0,000102	-7,07346e	-0,000302	0,000348	0,000301	0,000195	4,95312e-	0,00016
00133	-0,000211	-0,000174	-0,000274	-0,000115	1,39796e-	-0,000106	-0,000123	-9,9465
9191e	-0,000235	-0,000168	-0,000289	-7,68754e	7,80468e-	-1,05297e	-0,000159	-0,0001
443e-	-0,000181	-0,000184	-0,000317	-5,95669e	1,31876e-	-2,12312e	-0,000165	-0,0001
00101	-0,000207	-0,000196	-0,000274	-8,02223e	3,19685e-	-0,000156	-0,000136	-0,0001
00130	-0,000214	-0,000161	-0,000380	-0,000121	-4,42385e	-0,000139	-0,000115	-0,0001
7371e	-0,000170	-0,000167	-0,000252	-0,000110	-1,77964e	-5,5334e-	-0,000126	-0,0001
9627e	-0,000201	-0,000165	-0,000284	-0,000139	-2,69614e	-0,000105	-0,000193	-0,0001
0659e	-0,000213	-0,000187	-0,000307	-8,51037e	-1,52672e	-1,85605e	-0,000104	-0,0001
3162e	-0,000254	-0,000181	-0,000311	-0,000208	-1,72569e	-7,40748e	-0,000150	-0,0001

The first task is to make a PCA overview of the data. Before any modelling is done, change scaling to par (pareto) and define two classes (A,B,C=1 and D,E,F=2). Set model type to PCA, see figure below. All these settings are done in the *workset* menu.



Observations: 57, Included: 57, Selected: 7

Primary ID	Class
✓ B7	1
✓ B8	1
✓ C1	1
✓ C2	1
✓ C3	1
✓ C4	1
✓ C5	1
✓ C6	1
✓ C7	1
✓ C8	1
✓ D1	2
✓ D2	2
✓ D3	2
✓ D4	2
✓ D5	2
✓ D8	2

Set class: 1

Model type: PCA-X

Interpret the PCA model. Create the scores and DModX-plots. What do you see? Are there any groupings consistent with the different plants? Internode variation? Any outliers? What about the analytical replicates?

Hint: Colour the plot by classes and name the observations by their primary IDs.

Task 2 (Comparing PCA to OPLS-DA)

We will now compare PCA to OPLS-DA. Under *workset / new as model 1* (if model 1 is the PCA model). Exclude the analytical replicates and select the class variable as Y. Change the model type to OPLS/O2PLS. *Auto fit* a model with 1+4 components. Compute the corresponding PCA model. Plot scores and compare the results from the PCA model to the OPLS-DA model. What can bee seen in the first OPLS-DA component? What can bee seen in the orthogonal components?

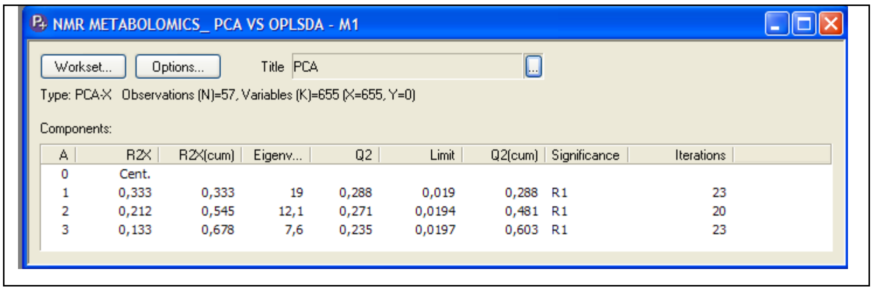
Task 3 (Diagnostics of OPLS/O2PLS-DA model)

How good is the model based on predictive ability? How much of the total variation in the data corresponds to the separation between the wild type and MYB76 plants? How much of the variation in the data is systematic but uncorrelated (orthogonal) to the separation between the classes? How much of the variation in the data is noise?

SOLUTIONS to METABOLOMICS with OPLS

Task 1 (PCA)

Interpretation of the first and second component, t1 and t2, indicates an internode variation along t1. This common internod variation will deviate for the two plants at higher internode numbers, this is seen in t2. With three components the WT and MYB76 class will separate. It is also seen that the analytical replicates are quite stable compared to internode variation and differences between the two classes.



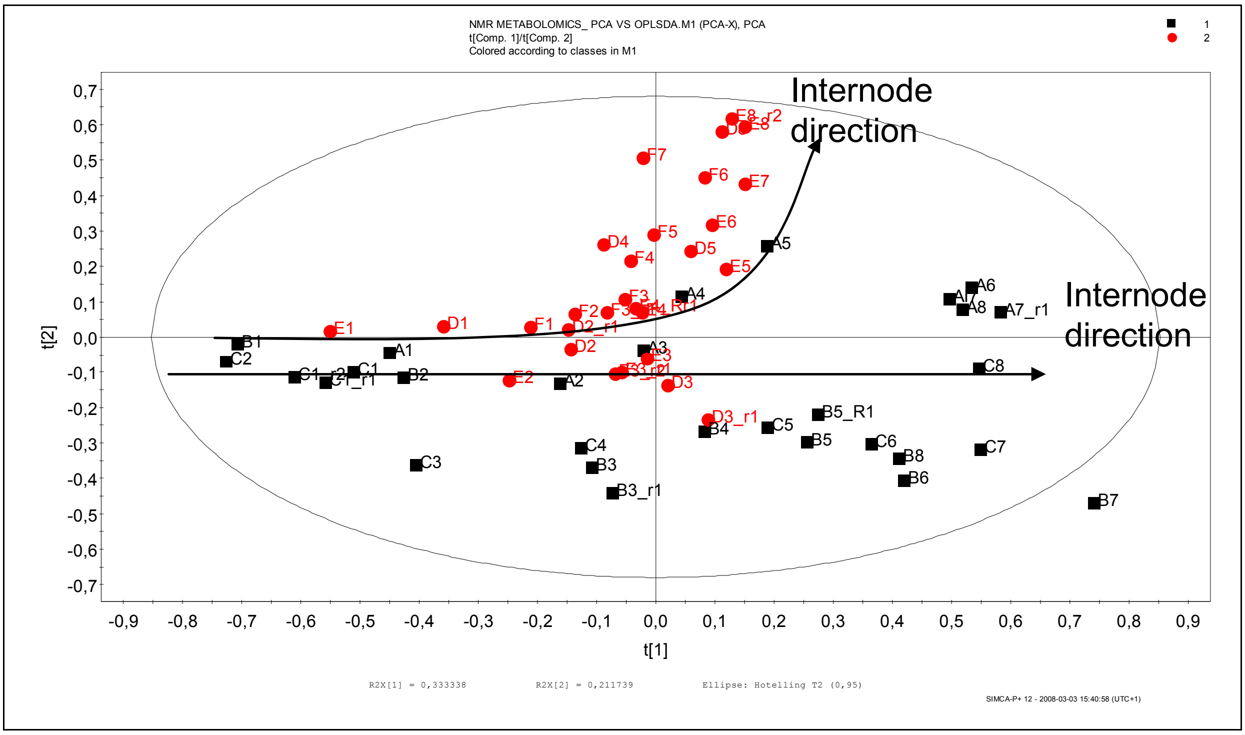
NMR METABOLOMICS_ PCA VS OPLSDA - M1

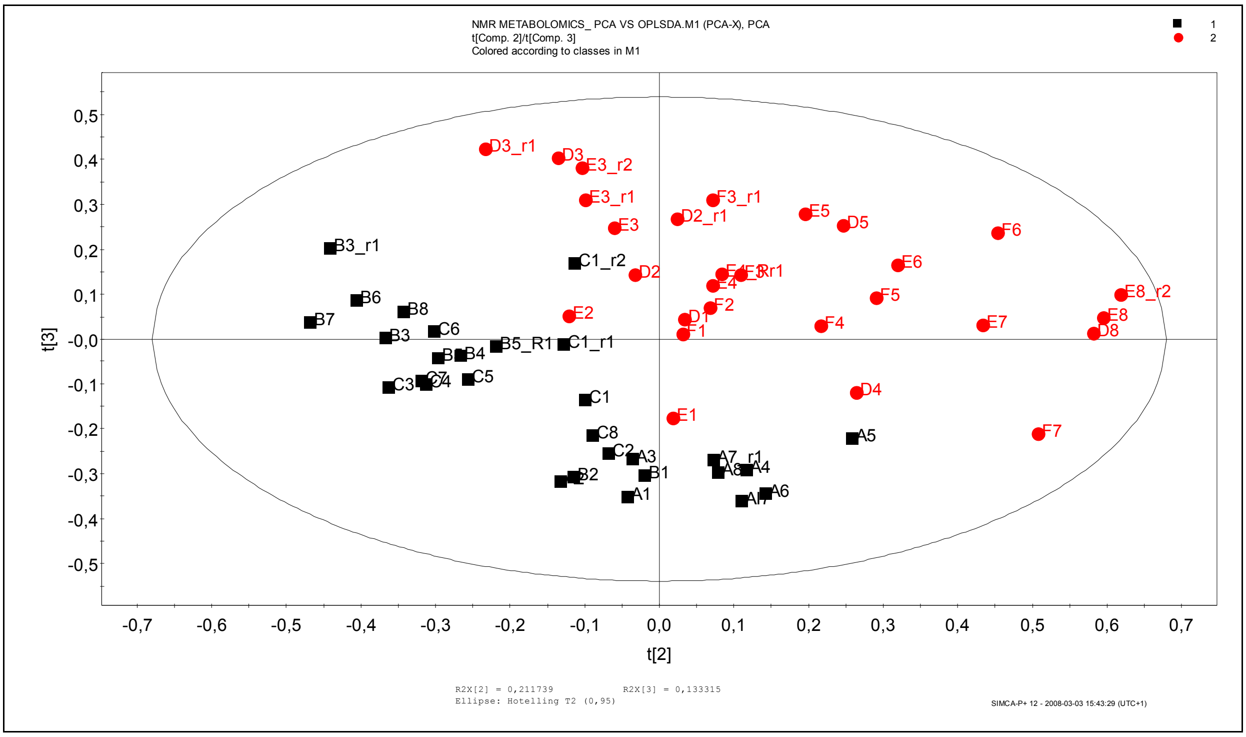
Title PCA

Type: PCA-X Observations (N)=57, Variables (K)=655 (X=655, Y=0)

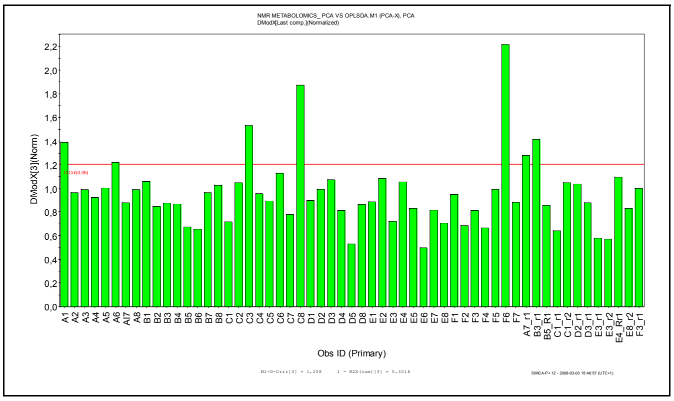
Components:

A	R2X	R2X(cum)	Eigenv...	Q2	Limit	Q2(cum)	Significance	Iterations
0	Cent.							
1	0,333	0,333	19	0,288	0,019	0,288	R1	23
2	0,212	0,545	12,1	0,271	0,0194	0,481	R1	20
3	0,133	0,678	7,6	0,235	0,0197	0,603	R1	23





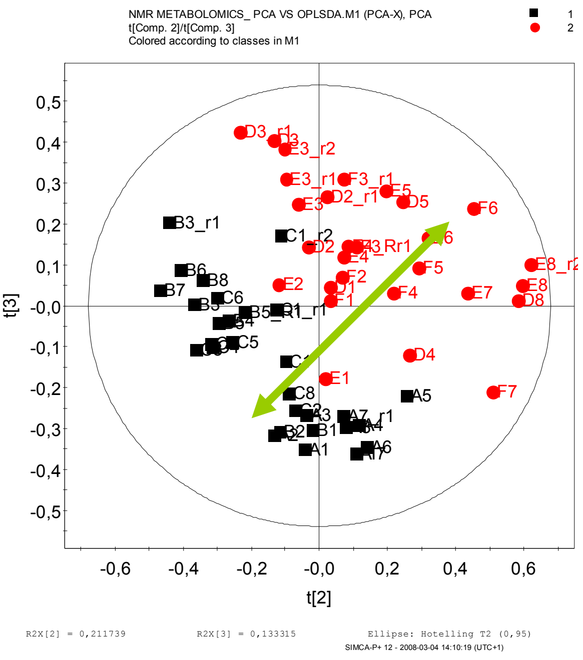
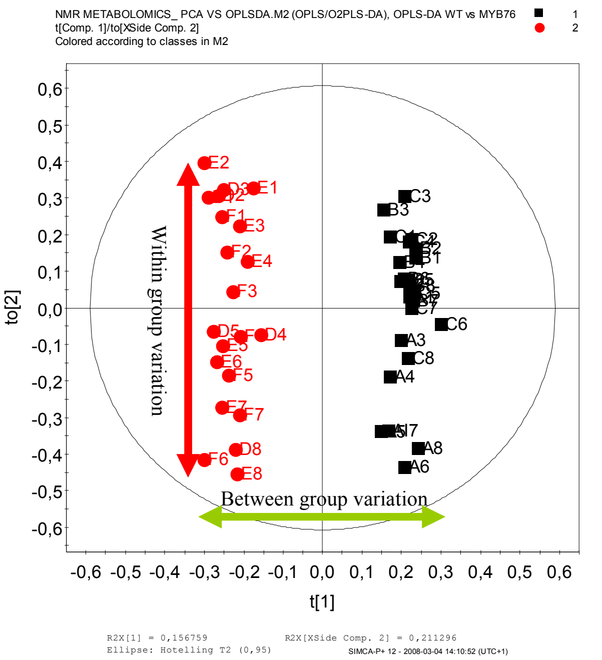
The DModX plot indicates that a few observations are outside the model limits. However these observations are only moderate outliers and will therefore remain in the model.



Task 2 (Comparing PCA to OPLS-DA)

A basic requirement to be able to interpret an OPLS-DA model is that we get a reasonably good OPLS-DA model with a good Q2. This is in fact the basic requirements for all prediction modelling. In this example we got a Q2 of 0,941 which is very high.

The advantage with the OPLS-DA model is that the between group variation (class separation) is seen in the first component and within group variation will be seen in the orthogonal components. From the plots below we see that the OPLS-DA model is a rotated PCA model.

PCA	OPLS-DA
	

The difference between PCA and OPLS-DA is clearly visualized in the two plots above. In the PCA model the difference between WT and MYB76 is seen in a combination of two component, t2 and t3. In the OPLS-DA model the difference between WT and MYB76 is seen in the first component, t1. The common internode variation is visualized in the second orthogonal component, to2.

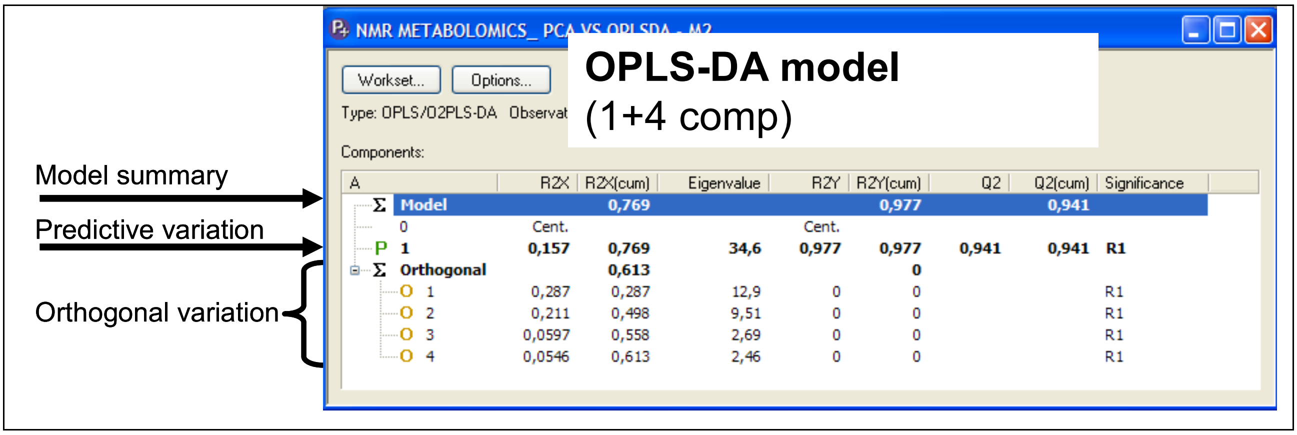
The simple reason why this is seen is because this is the nature of the OPLS/O2PLS algorithm. The algorithm will rotate the plane and separate correlated variation (in this example the two classes) from uncorrelated variation between X and y. Uncorrelated variation is also called orthogonal variation and is not related to the observed response y.

Because OPLS concentrates the between group variation (class separation) into the first component the interpretation of the loading vectors will also be simplified.

Technical Note: As OPLS rotates the first score vector t1 when additional components are computed the t1 vs. to1 plot changes when you add additional components to the model. Make sure that the model is optimized by using cross validation. Do NOT optimize the model by visualizing the class separation from the score plot.

Task 3 (Diagnostics of OPLS-DA model)

OPLS-DA diagnostics are also separated into predictive and orthogonal variation. To answer the questions in this task we need to understand all numbers in the model overview window seen in the figure below.



Type: OPLS/O2PLS-DA Observat

Components:

A	R2X	R2X(cum)	Eigenvalue	R2Y	R2Y(cum)	Q2	Q2(cum)	Significance
Σ Model		0,769			0,977		0,941	
0	Cent.			Cent.				
P 1	0,157	0,769	34,6	0,977	0,977	0,941	0,941	R1
Σ Orthogonal		0,613			0			
O 1	0,287	0,287	12,9	0	0			R1
O 2	0,211	0,498	9,51	0	0			R1
O 3	0,0597	0,558	2,69	0	0			R1
O 4	0,0546	0,613	2,46	0	0			R1

Model Summary

R2X(cum) is the sum of predictive + orthogonal variation in X that is explained by the model, 0,157+0,613=0,769. Can also be interpreted as 76,9% of the total variation in X.

R2Y(cum) is the total sum of variation in Y explained by the model, here 0,977.

Q2(cum) is the goodness of prediction, here 0,914.

Predictive variation=variation in X that is correlated to Y

A corresponds to the number of correlated components between X and Y. If only one response vector is used then A is always 1.

R2X is the amount of variation in X that is correlated to Y, here 0,157.

Orthogonal variation=variation in X that is uncorrelated to Y

A corresponds here to the number of uncorrelated (orthogonal) components. Each orthogonal component is represented and can be interpreted individually.

R2X is the amount of variation in X that is uncorrelated to Y. Each component is represented individually.

R2X(cum) In bold is the total sum of variation in X that is uncorrelated to Y, here 0,613.

Answers to questions

How good is the model based on predictive ability?

Q2=0,914

How much of the total variation in the data corresponds to the separation between the wild type and MYB76 plants?

Predictive variation between X and Y R2X=0,157 which is 15,7% of the total variation in X.

How much of the variation in the data is systematic but uncorrelated (orthogonal) to the separation between the classes?

R2X(cum)=0,613 or 61,3% of the total variation in the data.

How much of the variation in the data is noise?

This is the amount of variation that can not be explained by the model.

Noise=total variation in the data - predictive variation - orthogonal variation

Noise=1- 0,157 - 0,613=0,23 → 23%

Conclusions

- OPLS will rotate the model plane towards the direction of Y
- The rotation separates correlated (predictive) variation from uncorrelated (orthogonal) variation between X and Y.
- In OPLS-DA studies with two classes, the predictive component, t1, will describe the differences between two groups and the orthogonal components will describe systematic variation in the data that is not correlated to Y.
- The separation of predictive and orthogonal components will facilitate interpretation of metabolomics data in terms of model diagnostics and also for biomarker identification. The later will be described in another example.
- OPLS-DA in Metabolomics studies allows the user to mine complex data and provides information which allows us to propose intelligent hypotheses.

GC/MS Metabolomics with OPLS

OPLS in multi class metabolomics

Background

This exercise is the study of genetically modified poplar plant by GC/MS metabolomics. Two modifications are investigated i.e. up regulation and down regulation within the *PttPME1* gene. This gene is involved in the production of pectin methyl esterase, PME, which is an enzyme that de-esterifies methylated groups within pectin. Pectin is big complex polymer and will not be analyzed by this type of technique. Never the less, the metabolic profile was of interest as both lines indicated several symptoms of oxidative stress response.

Objective

The objective of this exercise is to shed some light on how OPLS may be used in state-of-the-art multi class Metabolomics. In particular, the objectives are to:

- demonstrate how to extract putative biomarkers from the S-plot
- demonstrate how to provide with statistical evidence to extracted biomarkers
- demonstrate how to extract information that is unrelated (orthogonal) to the modelled response, **y**
- demonstrate how to compare multiple classes by the use of an SUS-plot

Data

In total, the data set contains N = 26 observations (plants) and K = 80 variables (resolved and integrated GC/MS profiles by the use of H-MCR [1]). The observations (plants) are divided in three groups ("classes"):

• Control Wild type plant, 10 plants	**"WT"**
• *PttPME1* down regulated poplar, L5, 7 plants	**"L5"**
• *PttPME1* up regulated poplar, 9 plants	**"2B"**

The GC/MS data are three-way by nature with time, absorbance and mass dimensions. In this example the three way data have been pre-processed by Hierarchical Multivariate Curve Resolution, H-MCR [1]. H-MCR resolves the chromatic profiles, calculates the area of the resolved profiles and generates the corresponding mass spectrum. The resolved mass spectrum can be subjected to a library search and the compound can thus be identified. A more elaborate account of the experimental conditions is found in referens [2].

1) Jonsson et. al Journal of Proteome Research ***2006,*** *5,1407-1414, Predictive metabolite profiling applying hierarchical multivariate curve resolution to GC-MS data-A potential tool for multi-parametric diagnosis*

2) Wiklund et. al Analytical Chemistry ***2008,*** *80, 115-122, Visualization of GC/TOF-MS-Based Metabolomics Data for Identification of Biochemically Interesting Compounds Using OPLS Class Models*

Tasks

Task 1

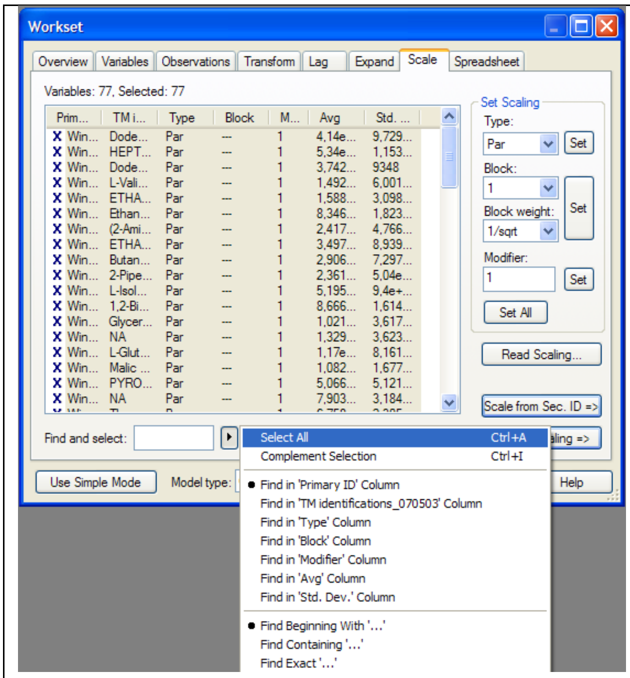
Select *File New* and choose GCMS metabolomics_all.xls. Choose the tab *GCMS metabolomics Xylem*. On import set the first row to *Primary Variable ID,* second column to *Secondary Variable ID* and first column to *Primary Observation ID.* Two columns have been appended to the data to designate the different classes. One column for WT vs. 2B and the other for WT vs. L5. The WT plants are 0 and the modified plants are 1 in both Y vectors. Set both vectors to Y. The last two columns in the data called *CV 2B* and *CV L5* is neither X or Y variables. These two columns will be used in task 5 in order to balance the exclusion of observations during cross validation. REMEMBER to exclude these two columns in all modeling.

Before we make any OPLS modelling it is recommend that an overview PCA for each class is performed. This is done to ensure that no outlier exists in the data. Exclude the response vectors WT vs. L5 and WT vs. 2B before executing a PCA model. If outliers exists in the data these should be checked historically for any explainable reason. Also check if the pre processing of the raw data is correct.

Task 2 (Contrasting two classes)

We will now make two class models where in each model two classes are contrasted. In model 1 select all WT and L5 observations and exclude observations from 2B. Define the WT vs L5 variable as the single Y-variable under *workset/variables* and exclude variable WT vs 2B also remember to exclude CV2B and CVL5. Pareto scale all variables. Scaling is performed under *workset /scale*. In SIMCA the default scaling is unit variance (UV). Change this to pareto (par), *select all/par/set*. Exclude all samples from class 2B under *workset/observations*. Under *workset/model type* select *OPLS/O2PLS*. Autofit the model.

Do the same thing for WT vs 2B. Make sure that you use exactly the same variables and scaling in both models. This can be done under *workset/new as model* (select the model with WT vs L5). Exclude all samples from L5 and include all samples from 2B. Compute the corresponding OPLS/O2PLS model for WT vs 2B. Compare the results from the two models.



Variables: 77, Selected: 77

	Prim...	TM i...	Type	Block	M...	Avg	Std. ...
X	Win...	Dode...	Par	---	1	4,14e...	9,729...
X	Win...	HEPT...	Par	---	1	5,34e...	1,153...
X	Win...	Dode...	Par	---	1	3,742...	9348
X	Win...	L-Vali...	Par	---	1	1,492...	6,001...
X	Win...	ETHA...	Par	---	1	1,588...	3,098...
X	Win...	Ethan...	Par	---	1	8,346...	1,823...
X	Win...	(2-Ami...	Par	---	1	2,417...	4,766...
X	Win...	ETHA...	Par	---	1	3,497...	8,939...
X	Win...	Butan...	Par	---	1	2,906...	7,297...
X	Win...	2-Pipe...	Par	---	1	2,361...	5,04e...
X	Win...	L-Isol...	Par	---	1	5,195...	9,4e+...
X	Win...	1,2-Bi...	Par	---	1	8,666...	1,614...
X	Win...	Glycer...	Par	---	1	1,021...	3,617...
X	Win...	NA	Par	---	1	1,329...	3,623...
X	Win...	L-Glut...	Par	---	1	1,17e...	8,161...
X	Win...	Malic ...	Par	---	1	1,082...	1,677...
X	Win...	PYRO...	Par	---	1	5,066...	5,121...
X	Win...	NA	Par	---	1	7,903...	3,184...

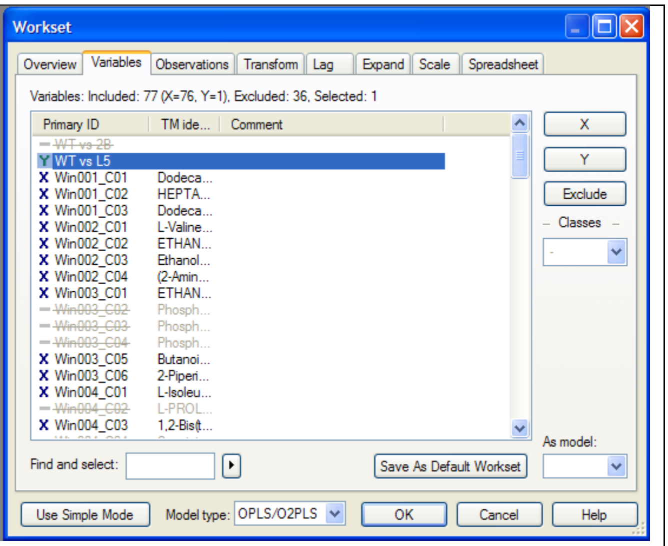
Type: Par

Block: 1

Block weight: 1/sqrt

Modifier: 1

Choose *select all/par* and press *set.*

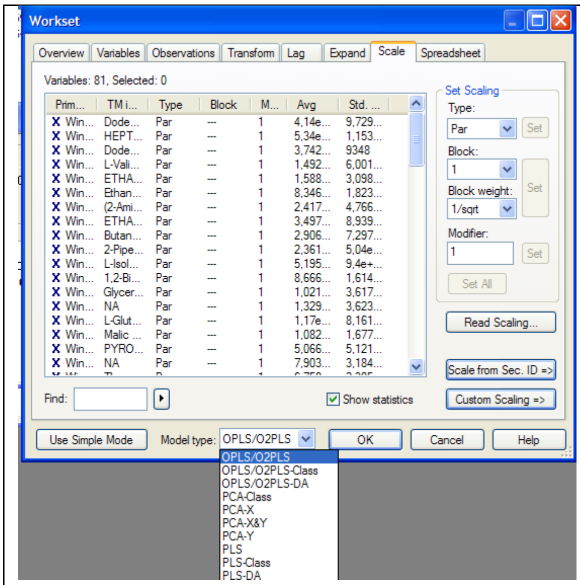


Variables: Included: 77 (X=76, Y=1), Excluded: 36, Selected: 1

	Primary ID	TM ide...	Comment
—	~~WT vs 2B~~		
Y	WT vs L5		
X	Win001_C01	Dodeca...	
X	Win001_C02	HEPTA...	
X	Win001_C03	Dodeca...	
X	Win002_C01	L-Valine...	
X	Win002_C02	ETHAN...	
X	Win002_C03	Ethanol...	
X	Win002_C04	(2-Amin...	
X	Win003_C01	ETHAN...	
—	~~Win003_C02~~	Phosph...	
—	~~Win003_C03~~	Phosph...	
—	~~Win003_C04~~	Phosph...	
X	Win003_C05	Butanoi...	
X	Win003_C06	2-Piperi...	
X	Win004_C01	L-Isoleu...	
—	~~Win004_C02~~	L-PROL...	
X	Win004_C03	1,2-Bis(t...	

Model type: OPLS/O2PLS

Choose WT vs L5 as Y. Exclude WT vs 2B and CV2B and CVL5.



Variables: 81, Selected: 0

Prim...	TM i...	Type	Block	M...	Avg	Std. ...
X Win...	Dode...	Par	---	1	4,14e...	9,729...
X Win...	HEPT...	Par	---	1	5,34e...	1,153...
X Win...	Dode...	Par	---	1	3,742...	9348
X Win...	L-Vali...	Par	---	1	1,492...	6,001...
X Win...	ETHA...	Par	---	1	1,588...	3,098...
X Win...	Ethan...	Par	---	1	8,346...	1,823...
X Win...	(2-Ami...	Par	---	1	2,417...	4,766...
X Win...	ETHA...	Par	---	1	3,497...	8,939...
X Win...	Butan...	Par	---	1	2,906...	7,297...
X Win...	2-Pipe...	Par	---	1	2,361...	5,04e...
X Win...	L-Isol...	Par	---	1	5,195...	9,4e+...
X Win...	1,2-Bi...	Par	---	1	8,666...	1,614...
X Win...	Glycer...	Par	---	1	1,021...	3,617...
X Win...	NA	Par	---	1	1,329...	3,623...
X Win...	L-Glut...	Par	---	1	1,17e...	8,161...
X Win...	Malic ...	Par	---	1	1,082...	1,677...
X Win...	PYRO...	Par	---	1	5,066...	5,121...
X Win...	NA	Par	---	1	7,903...	3,184...

Type: Par

Block: 1

Block weight: 1/sqrt

Modifier: 1

Find:

☑ Show statistics

Model type: OPLS/O2PLS

Set model type to OPLS/O2PLS

Specific questions:

Can you separate WT from L5 and WT from 2B by OPLS classification models?

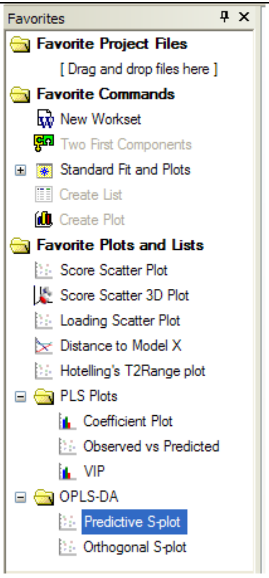
How much of the variation in **X** is related to the separation between WT and L5, WT and 2B?

How much of the variation in **X** is systematic but uncorrelated to the separation between WT and L5, WT and 2B?

Can you see any patterns in the score plot except for the separation between the classes?

Task 3 (Identifying putative biomarkers between Control and Treated)

Identify putative biomarkers by using the S-plot (p[1] vs p(corr)[1]) and the loading column plot (p[1]) with confidence intervals. Use the two plots interactively. The creation of an S-plot from the predictive components can be performed under *favourites/OPLS-DA/predictive S-plot*. The plot settings should be slightly changed. The y-axis (p(corr)[1]) should vary between ±1. Right click the mouse and select *plot settings/axes*. Change the y-axis to ±1 and the x-axis to be symmetric around 0. The S-plot is only applicable if the data are pareto or ctr scaled.



S-plot from the predictive component.

Specific questions:

Why is the S-plot important?

What does the confidence interval seen in the loading column plot (p[1]) mean?

What are the default settings in SIMCA for the column loadings confidence intervals (what is α by default)?

In what regions should you be careful when extracting putative biomarkers from the S-plot?

Why is it not recommended to only use the column plot with confidence intervals for putative biomarker identification?

Identify the specific pattern seen in the orthogonal components by the orthogonal S-plot.

Task 4 (Investigate if the same biomarkers appear in two models)

In order to compare the outcome from two models the *shared and unique structure* plot, SUS-plot, is useful. This plot is done from *plot list/scatter plots/observations and loadings* select the p(corr)[1] vector from both models. Both axes in this plot should vary between ±1. Right click the mouse and select *plot settings/axes*. Change the both the x and y-axis to ±1.

Specific questions:

In what regions will you find shared information?

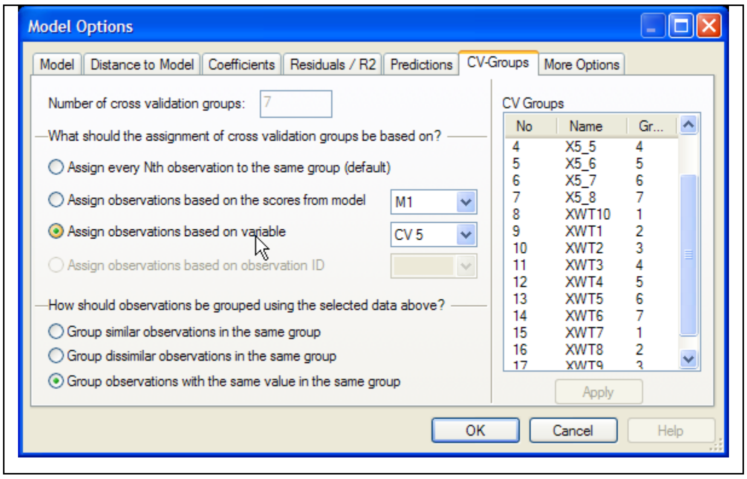
In what regions will you find up regulated and unique information for L5?

In what regions will you find up regulated and unique information for 2B?

Task 5 (Change the sample exclusion during cross-validation)

Balanced models are important in "omics" studies. Often the number of samples in different classes is unequal and this could make interpretation misleading. One alternative is to change the exclusion of samples during CV to be more balanced between classes. This task is only for teaching, no solutions are provided.

Make a new model with same settings as in previous tasks, both for WT vs L5. Go to *workset/model options/CV-groups*. Select *assign observation based on variable* and choose *CV5,* finally *group observations with the same value in the same group* and press *apply.*



Number of cross validation groups: 7

What should the assignment of cross validation groups be based on?

○ Assign every Nth observation to the same group (default)

○ Assign observations based on the scores from model M1

● Assign observations based on variable CV 5

○ Assign observations based on observation ID

How should observations be grouped using the selected data above?

○ Group similar observations in the same group

○ Group dissimilar observations in the same group

● Group observations with the same value in the same group

CV Groups

No	Name	Gr...
4	X5_5	4
5	X5_6	5
6	X5_7	6
7	X5_8	7
8	XWT10	1
9	XWT1	2
10	XWT2	3
11	XWT3	4
12	XWT4	5
13	XWT5	6
14	XWT6	7
15	XWT7	1
16	XWT8	2
17	XWT9	3

Do the same thing with WT vs 2B but select the CV2B vector in cross validation. Do you get same results as the default CV settings?

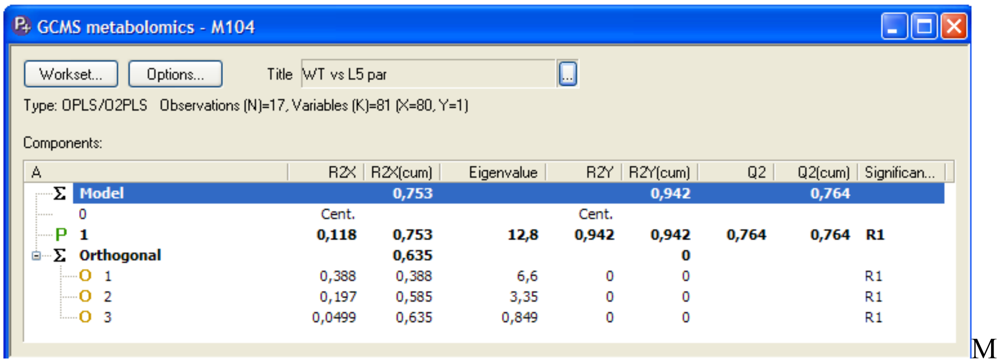
SOLUTIONS to GC/MS metabolomics

Task 1

No obvious outliers were found in the three classes.

Task 2 (Contrasting two classes)

A basic requirement to answer all the questions in this task is that we get a reasonably good OPLS-DA model with a good Q2. To answer the questions look into the model overview window.



GCMS metabolomics - M104

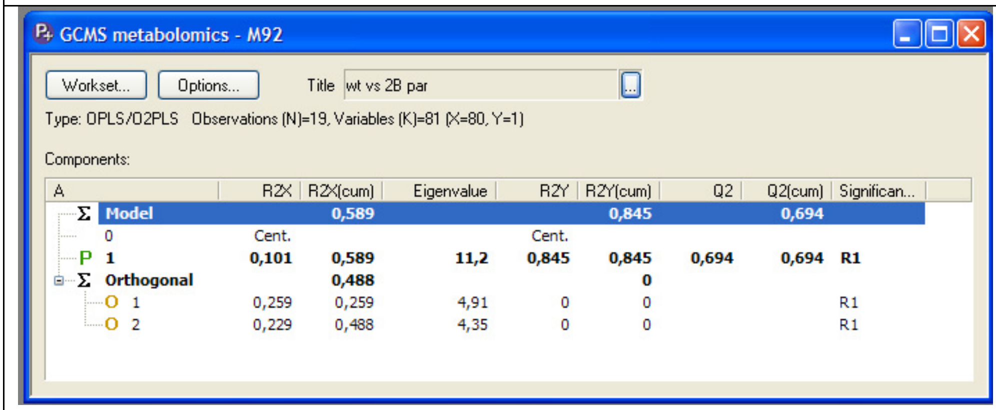
Title: WT vs L5 par

Type: OPLS/O2PLS Observations (N)=17, Variables (K)=81 (X=80, Y=1)

Components:

A	R2X	R2X(cum)	Eigenvalue	R2Y	R2Y(cum)	Q2	Q2(cum)	Significan...
Σ Model		0,753			0,942		0,764	
0	Cent.			Cent.				
P 1	0,118	0,753	12,8	0,942	0,942	0,764	0,764	R1
Σ Orthogonal		0,635			0			
O 1	0,388	0,388	6,6	0	0			R1
O 2	0,197	0,585	3,35	0	0			R1
O 3	0,0499	0,635	0,849	0	0			R1

Model WT vs L5



GCMS metabolomics - M92

Title: wt vs 2B par

Type: OPLS/O2PLS Observations (N)=19, Variables (K)=81 (X=80, Y=1)

Components:

A	R2X	R2X(cum)	Eigenvalue	R2Y	R2Y(cum)	Q2	Q2(cum)	Significan...
Σ Model		0,589			0,845		0,694	
0	Cent.			Cent.				
P 1	0,101	0,589	11,2	0,845	0,845	0,694	0,694	R1
Σ Orthogonal		0,488			0			
O 1	0,259	0,259	4,91	0	0			R1
O 2	0,229	0,488	4,35	0	0			R1

Model WT vs 2B

Answers:

Can you separate WT from L5 and WT from 2B by OPLS classification models?

WT vs L5: R2Y=0,942, Q2Y=0,764, good class separation and high predictive ability.

WT vs 2B: R2Y=0,845, Q2Y=0,694, good class separation and high predictive ability.

Better separation between WT vs L5 than WT vs 2B

How much of the variation in **X** is related to the separation between WT and L5, WT and 2B?

WT vs L5: predictive component R2X=0,118→11,8%

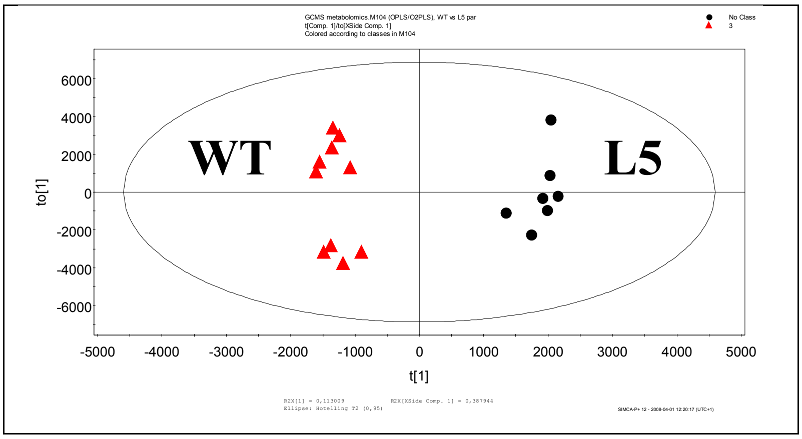
WT vs 2B: predictive component R2X=0,101→10,1%

How much of the variation in **X** is systematic but uncorrelated to the separation between WT and L5, WT and 2B?

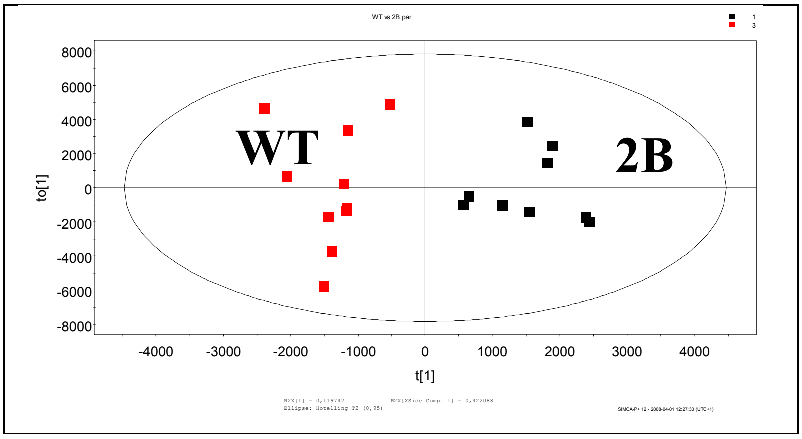
WT vs L5: R2X=0,635→63,5%

WT vs 2B: R2X=0,488→48,8%

Can you see any patterns in the score plot except for the separation between the classes?



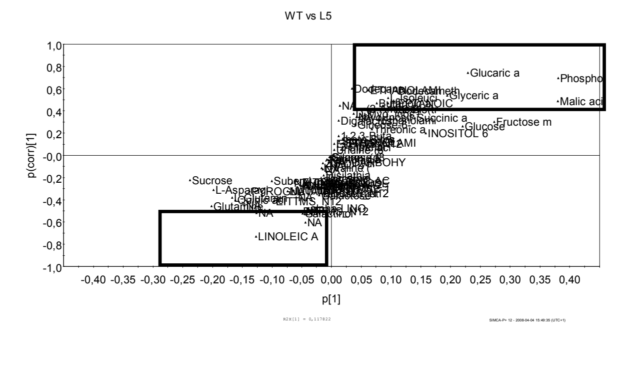
A clear separation between the WT and L5 can be seen in the first OPLS component, t[1]. This visualized separation must be combined with a high Q2 for a good class separation. Only a clear separation in the score plot is NOT a valid class separation. It is also seen in the first orthogonal score vector, to[1], that the WT class separates into two sub classes.



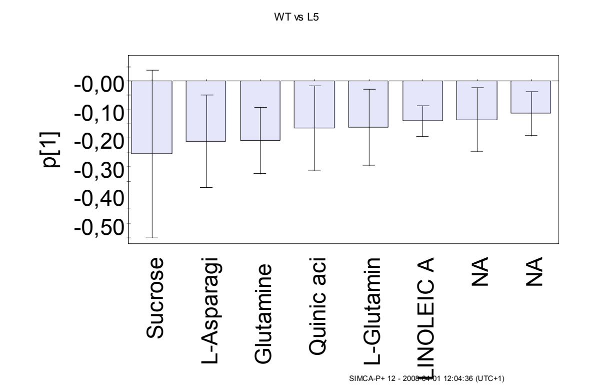
A separation between the WT and 2B can be seen in the first OPLS component, t[1]. The class separation seen in model WT vs L5 is not as clear in this model. The reason why this separation is not as clear in this case is due to the impact from class 2B.

Task 3 (Identifying putative biomarkers between Control and Treated)

The predictive S-plot is a good way to identify putative biomarkers. The column plot of p[1] is also of relevance since it is able to provide the confidence intervals of each loading value.



WT vs L5



Loading plot with confidence intervals

Hint: sort the loading plot to ascending values.

It is clearly seen that e.g sucrose is highly uncertain as a putative biomarker. In the S-plot sucrose has a high magnitude but a low reliability. This is confirmed from the loading plot p[1] where the confidence limit crosses zero. An interesting putative biomarker is linoleic acid which both has high magnitude and high reliability and the confidence interval is also low. Other putative biomarkers are glucaric acid, phosphoric acid and malic acid. All of these are of interest for further investigations. Remember that these putative biomarkers are only statistically significant.

Hint: Make a list of all interesting biomarkers by marking those of interest in the S-plot and then right click on the plot and *create/list*.

Why is the S-plot important?

The S-plot is an easy way to understand metabolomics data and is used to filtering out putative biomarkers.

In this plot both magnitude (intensity) and reliability is visualised. In spectroscopic data the peak magnitude is important as peaks with low magnitude are close to the noise level and are thus of higher risk for spurious correlation. The meaning of high reliability means high effect and lower uncertainty for putative biomarker.

What does the confidence interval seen in the loading column plot (p[1]) mean?

The confidence interval reflects the variable uncertainty and is directly correlated to the reliability. The confidence interval is very useful in reporting the significance of a putative biomarker.

What are the default settings in SIMCA for the column loadings confidence intervals (what is α by default)?

SIMCA will by default set the α to 0,05. The meaning of this is that a metabolite with a confidence interval which does not cross 0 is by 95% statistically safe. An alternative interpretation is: the probability of making the wrong decision from is 5%. The default setting can be changed.

In what regions should you be careful when extracting putative biomarkers from the S-plot?

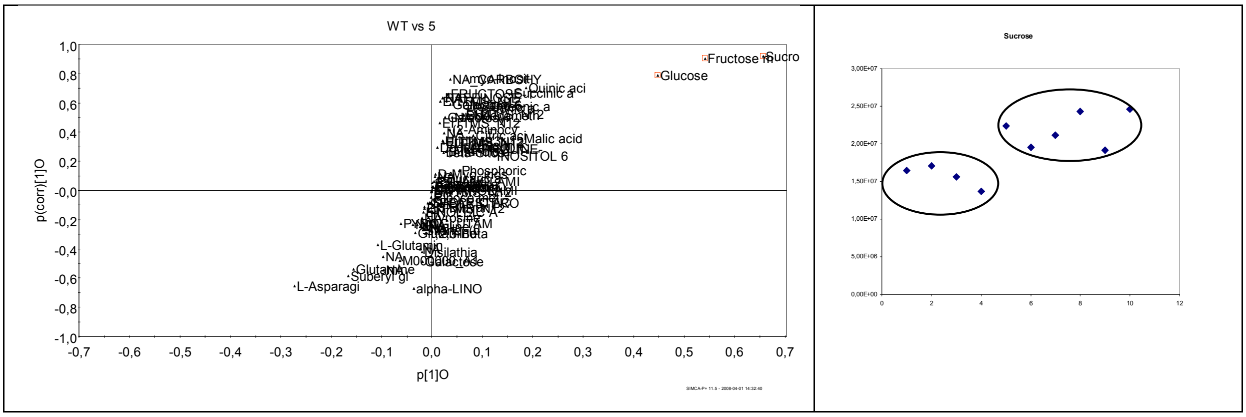
Metabolites with: low reliability plus low magnitude and high reliability plus low magnitude are uncertain. The importance of magnitude depends on how the spectral data was pre-processed. Always remember to go back and check the raw data.

Why is it not recommended to only use the column plot with confidence intervals for putative biomarker identification?

The signal to noise overview is much easier to see in the S-plot. The combination of both plots also makes it easier to extract putative biomarkers.

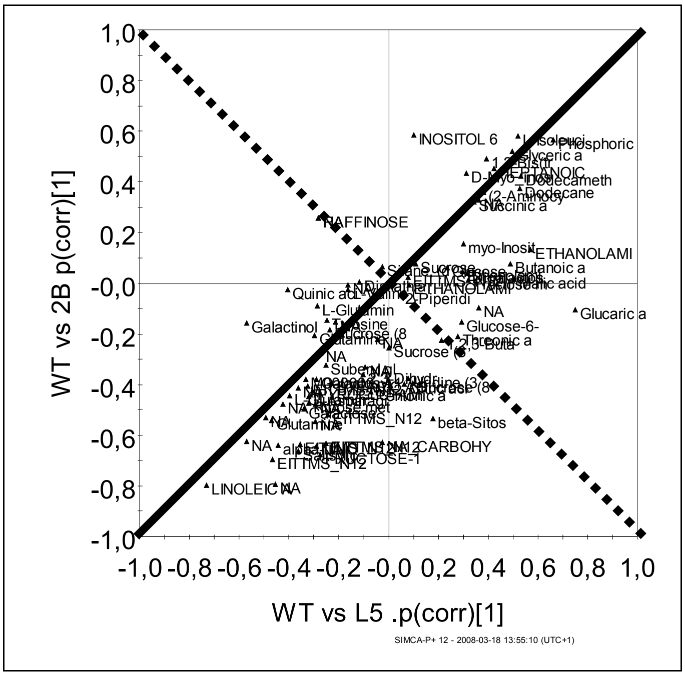
Identify the specific pattern seen in the orthogonal components by the orthogonal S-plot.

The cause to separation seen in the WT class can be identified in the S-plot from the orthogonal components i.e. p[1]o vs p(corr)[1]o.



Here it is seen that the sugars sucrose, fructose and glucose are the main cause to the first orthogonal component. By double clicking on sucrose in the S-plot, the raw data plot appears. In the plot above only the WT sample are seen, these have also sorted and plotted in excel to clarify the result.

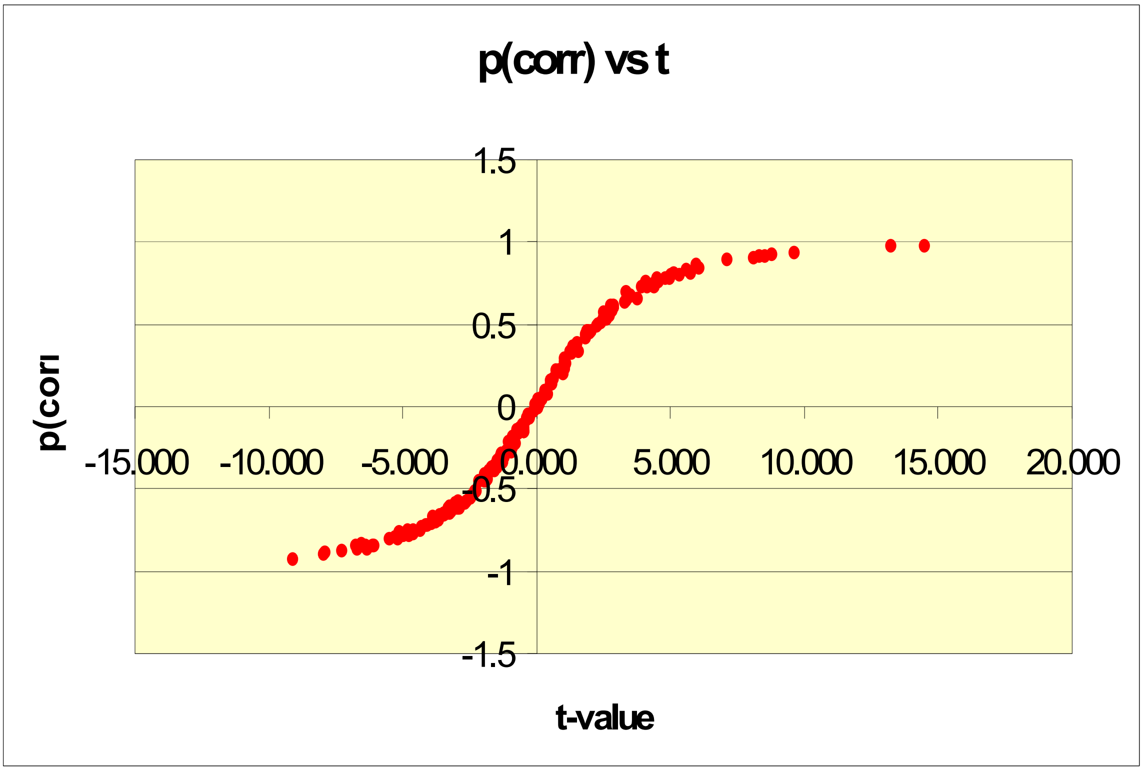
Task 4 (Investigate if the same biomarkers appear in two models)



It is visualized in the SUS-plot that many metabolites have the same effect on both L5 and 2B. Al metabolites on the diagonal have same effect in both plants e.g. linoleic acid (down regulated) and phosphoric acid (up-regulated). The broken line have same effect but in opposite directions in both transgenic plants. Some of the unique metabilites found in L5 was glucaric acid, malic acid, ethanolamine, butanoic acid. These were up regulated in L5. Quinic acid was also found as a unique down regulated metabolite in L5, but this metabolite was also highly uncertain. In 2B it was found that inositol was up regulated and several unassigned metabolites were found down regulated.

P(corr) is directly related to Students t, see illustration below.

Using Students t by itself means that we are only looking at high effects and neglecting the peak magnitude. It is well known that in spectroscopic data the signal to noise is highly important. This is also a reason why the YELLOW area in the S-Plot is dangerous and prone to spurious results.



Conclusions

- The S-plot can be done using PCA or PLS-DA ONLY if a clear class separation is seen in the first component in the score plot. If not the vectors p[1] and p(corr)[1] will be confounded by variation that is NOT related to class separation which will lead to an misleading interpretation.
- We obtain a list of potential biomarkers which are statistically significant and which separate one class from another.
- These biomarkers are **statistically significant**, but **not necessarily biochemically significant**.
- They may have biochemical significance and they may be the biomarkers we are interested in, however, this must be established through extensive testing.
- Metabonomics/Metabolomics allows the user to mine complex data and provides information which allows us to propose intelligent hypotheses.
- OPLS-DA is an excellent tool for this purpose.

MVDA-Exercise METABONOMICS with OPLS

Using OPLS in metabonomics

This is a follow-up exercise to the exercise named METABONOMICS. It is recommended that new users of SIMCA-P+ go through that exercise first.

Background

Metabolites are the products and by-products of the many complex biosynthesis and catabolism pathways that exist in humans and other living systems. Measurement of metabolites in human biofluids has often been used for the diagnosis of a number of genetic conditions, diseases and for assessing exposure to xenobiotics. Traditional analysis approaches have been limited in scope in that emphasis was usually placed on one or a few metabolites. For example urinary creatinine and blood urea nitrogen are commonly used in the diagnosis of renal disease.

Recent advances in (bio-)analytical separation and detection technologies, combined with the rapid progress in chemometrics, have made it possible to measure much larger bodies of metabolite data [1]. One prime example is when using NMR in the monitoring of complex time-related metabolite profiles that are present in biofluids, such as, urine, plasma, saliva, etc. This rapidly emerging field is known as Metabonomics. In a general sense, metabonomics can be seen as the investigation of tissues and biofluids for changes in metabolite levels that result from toxicant-induced exposure. The exercises below describe multivariate analysis of such data, more precisely ^{1}H-NMR urine spectra measured on different strains of rat and following dosing of different toxins.

Objective

The objective of this exercise is to shed some light on how PCA, PLS-DA and OPLS-DA may be used in state-of-the-art Metabonomics. In particular, the objectives are to:

- demonstrate the strength of OPLS-DA compared with PLS-DA;
- demonstrate how the results of OPLS-DA can be used to investigate if there are species differences when the rats are given the different drugs.

Data

In total, the data set contains N = 57 observations (rats) and K = 194 variables (^{1}H-NMR chemical shift regions). The observations (rats) are divided in six groups ("classes"):

•	Control Sprague-Dawley (s), 10 rats,	**"s"**
•	Sprague-Dawley treated with amiodarone (sa), 8 rats	**"sa"**
•	Sprague-Dawley treated with chloroquine (sc), 10 rats	**"sc"**
•	Control Fisher (f), 10 rats	**"f"**
•	Fisher treated with amiodarone (fa), 10 rats	**"fa"**
•	Fisher treated with chloroquine (fc), 9 rats	**"fc"**

The urine ^{1}H NMR spectra were reduced by summation of all the data points over a 0.04 ppm region. Data points between 4.5- 6.0 ppm, corresponding to water and urea resonances, were excluded, leaving a total of 194 NMR spectral regions as variables for the multivariate modelling. A more elaborate account of the experimental conditions are found in [2]. We are grateful to Elaine Holmes and Henrik Antti of Imperial College, London, UK, for giving us access to this data set.

1) Nicholson, J.K., Connelly, J., Lindon, J.C., and Holmes, E., Metabonomics: A Platform for Studying Drug Toxicity and Gene Function, Nature Review, 2002; 1:153-161. 2) J.R. Espina, W.J. Herron, J.P. Shockcor, B.D. Car, N.R. Contel, P.J. Ciaccio, J.C. Lindon, E. Holmes and J.K. Nicholson. Detection of in vivo Biomarkers of Phospholipidosis using NMR-based Metabonomic Approaches. Magn. Resonance in Chemistry 295: 194-202 2001.

Tasks

Task 1

Import the file Metabonomics_coded.xls. Mark the second column as *ClassID* and choose the length 2. This will assign the different classes automatically.

As you can see 6 columns have been appended to the data to designate the different classes. The composition of these new class variables is seen in the figure below. Set these columns as Y variables.

Before we make any detailed modelling it is recommend that an overview PCA model of the entire data set is computed. Such a model was shown in the METABONOMICS exercise.

Task 2 (Contrasting two classes)

We will now contrast two classes, the s and sa classes. Select observations 1-18 and Pareto scale the X-variables. Define the sa variable as the single Y variable. Exlude the other five 1/0 variables. Calculate a PLS-DA model with two components. Compute the corresponding OPLS-DA model. Plot scores and loadings and compare the results of the two models.

Task 3 (Identifying potential biomarkers between Control and Treated)

Make S-plot and column plot of the models´ loadings in order to identify potential biomarkers.

Task 4 (Investigate if the same biomarkers appear in two rat strains)

Compute multiple OPLS models contrasting two classes and try to elucidate whether the same biomarkers appear as important in the different cases.

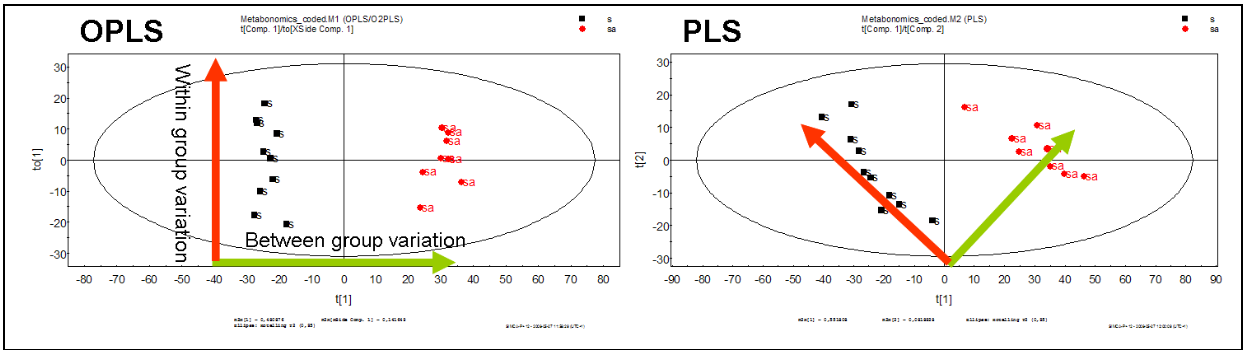
Task 5 (validate the OPLS-DA models by using cross-validated score plot)

In the ideal case, the external test set is prepared at a different time using a different set of samples. However, in this exercise this does not exist and the second best alternative will be used instead. The alternative way is to look at the cross-validated score plot which indicates the sample prediction uncertainty.

SOLUTIONS to METABONOMICS with OPLS

Task 2

The PLS and OPLS models are identical as they give the same prediction with the same number of components. The advantage with the OPLS model is that concentrates the between group variation (class separation) into the first component. From the plots below we see that the OPLS model is a rotated PLS model.

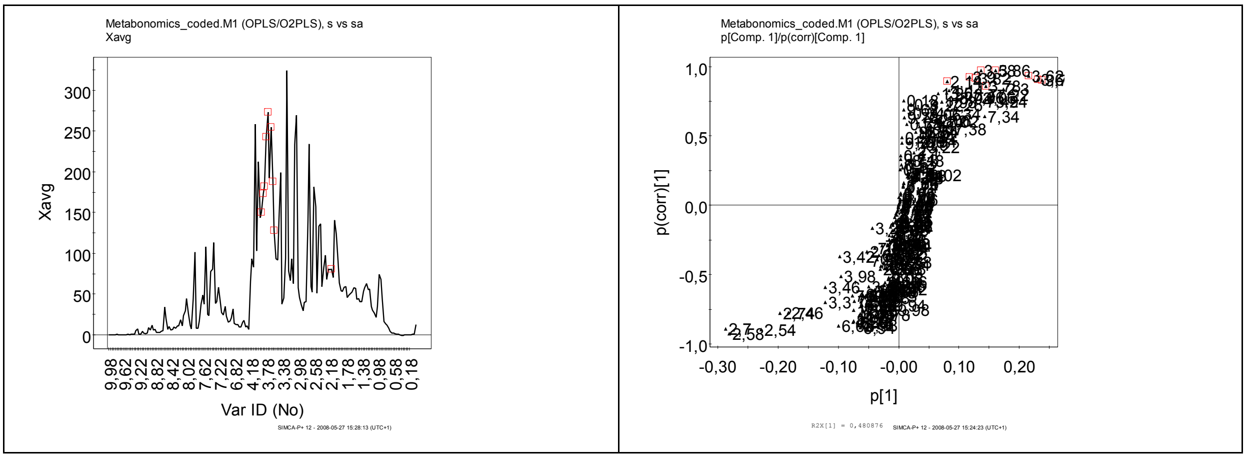


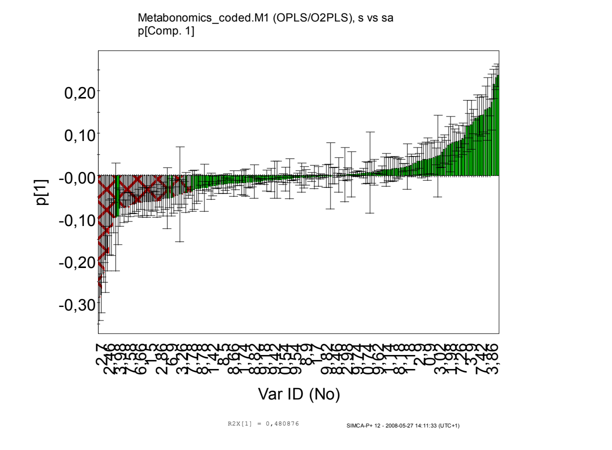
Technical Note: As OPLS rotates the first score vector t1 when additional components are computed the t1 vs. t2 plot changes when you add additional components to the model. Later t vectors t2, t3 etc are not rotated when a new component is added. The current example has only a small change but larger changes may be the case. In the current example t2 has also changed direction, which can happen but has no consequence.

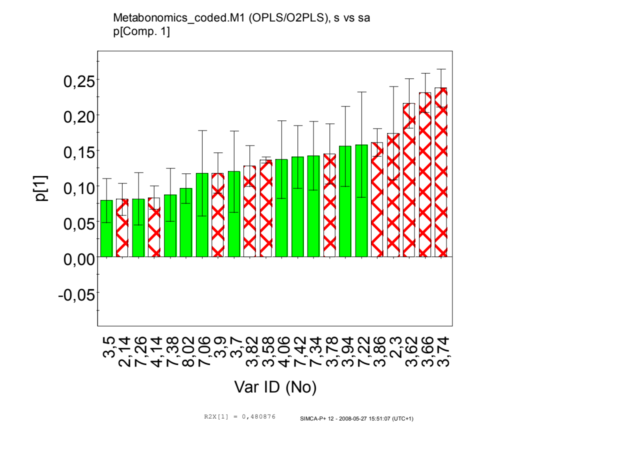
Because OPLS concentrates the between group variation (class separation) into the first component the t1 vs t2 plot visually improves the class separation for each component.

Task 3

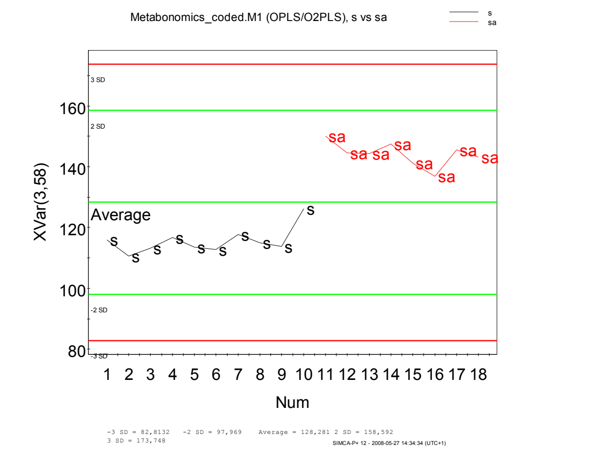
By plotting the S-plot, the x-average plot and the column plot is a good way to identify potential biomarkers. The advantage of using the Xavg plot is that many NMR spectroscopist like to resemble the results in the original NMR shape as it help identifying the selected variables. The marked signals reveal down regulated metabolites in the sa group. With NMR applications the line plot representation is also prevalent because this format also looks like the NMR spectrum. However, the column plot format is of relevance since it is able to provide the confidence intervals of each loading value.

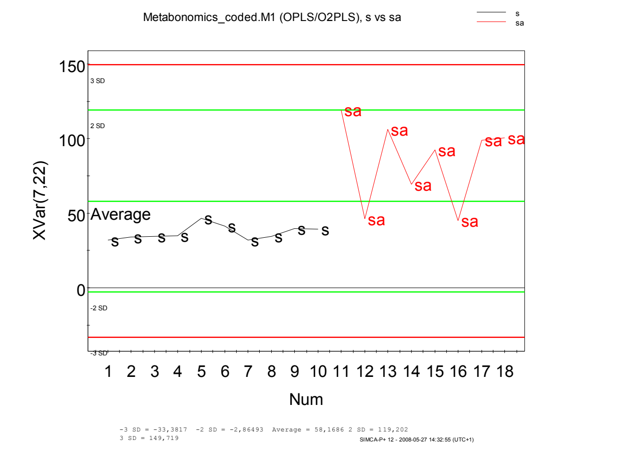






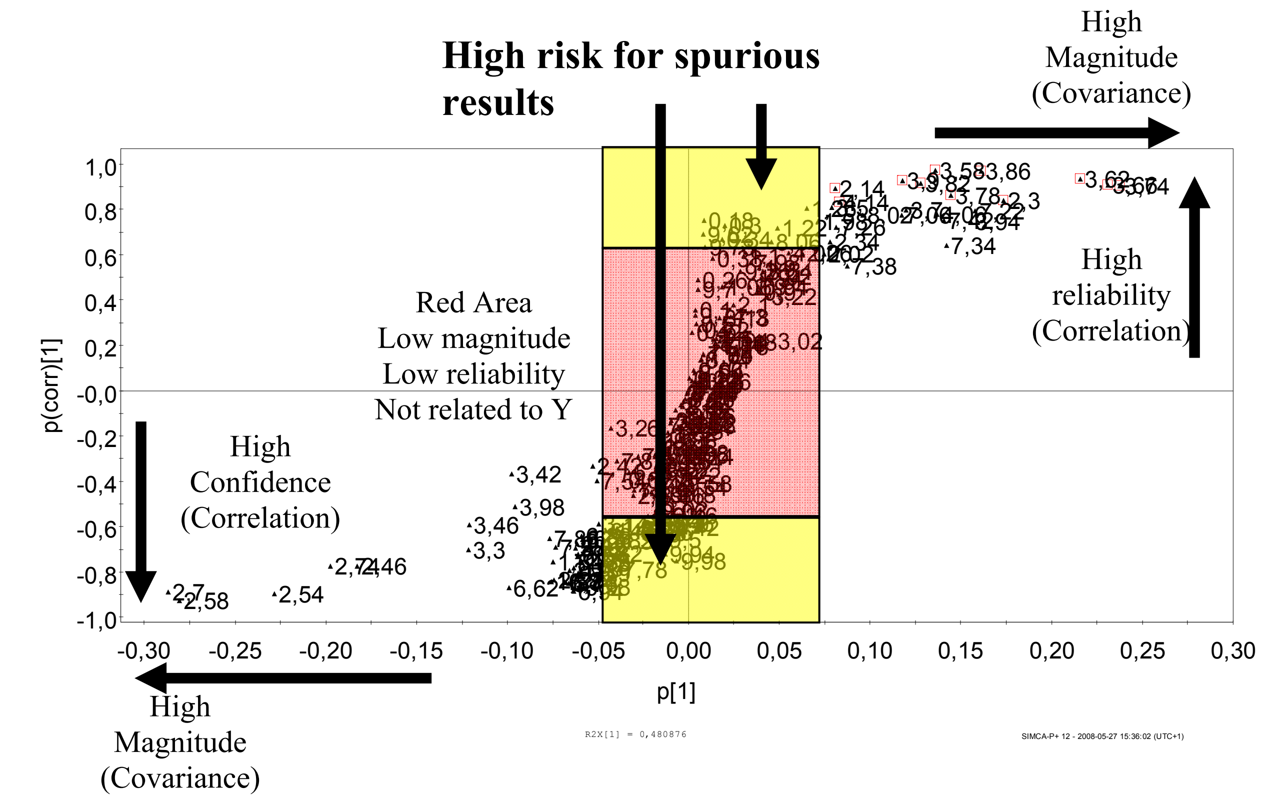
Sorting of the column plot followed by zooming onto the positive end gives us the potential biomarkers that have increased from the controls to the treated animals. Zooming at the other end gives us the biomarkers that have decreased going from controls to treated animals. The confidence intervals indicate how trustworthy the results are. Shift 7.22 –in the upper right plot -- has a large confidence interval. If we double click on that column we get the lower right plot which shows a large variation in that variable. Shift 3.58, in the upper right plot -- has a small confidence interval. If we double click on that column we get the lower left plot which shows a small variation within the groups and a large variation between the groups in that variable.

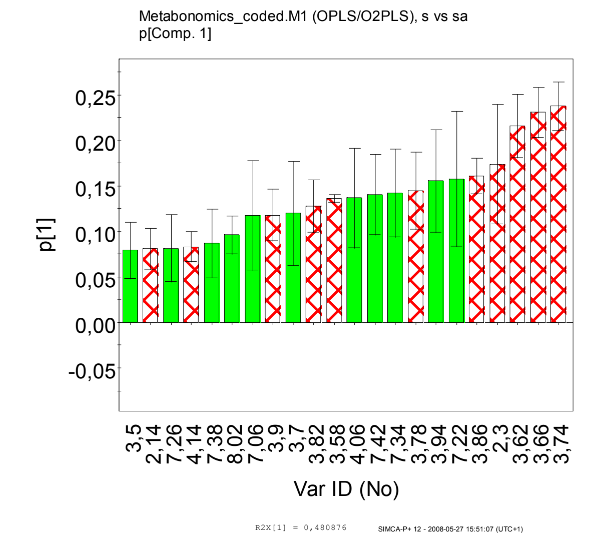




The S-plot, i.e., p*1 vs p(corr) -- only applicable if the data are pareto or ctr scaled -- visualises the information from the loading (p*1) plot and its confidence limits in another way, resulting in an easy to understand plot that can be used to filtering out potential biomarkers.

In the plot below the highlighted potential biomarkers (variables) have a p(corr) above 0.82. This means that in the plot variables with larger confidence intervals are not included and therefore remain green. The Red area has potential biomarkers that have Low magnitude and a Low reliability and therefore they are not related to Y, i.e. not affected by the treatment. The yellow area has a high pcorr but low influence on the model. This is the area where there is a high risk for spurious correlations.





Task 4

To accomplish this task, we run 4 OPLS-DA models in parallel:

M1 Control Sprague-Dawley (s), 10 rats, vs treated with **amiodarone** (sa), 8 rats "sa"

M2 Control Sprague-Dawley (s), 10 rats, vs treated with **chloroquine** (sc), 10 rats "sc"

M3 Control Fisher (f), 10 rats "f" vs treated with **amiodarone** (fa), 10 rats "fa"

M4 Control Fisher (f), 10 rats "f" vs treated with **chloroquine** (fc), 9 rats "fc"

Solution A

Make an Splot for M1 and one for M3

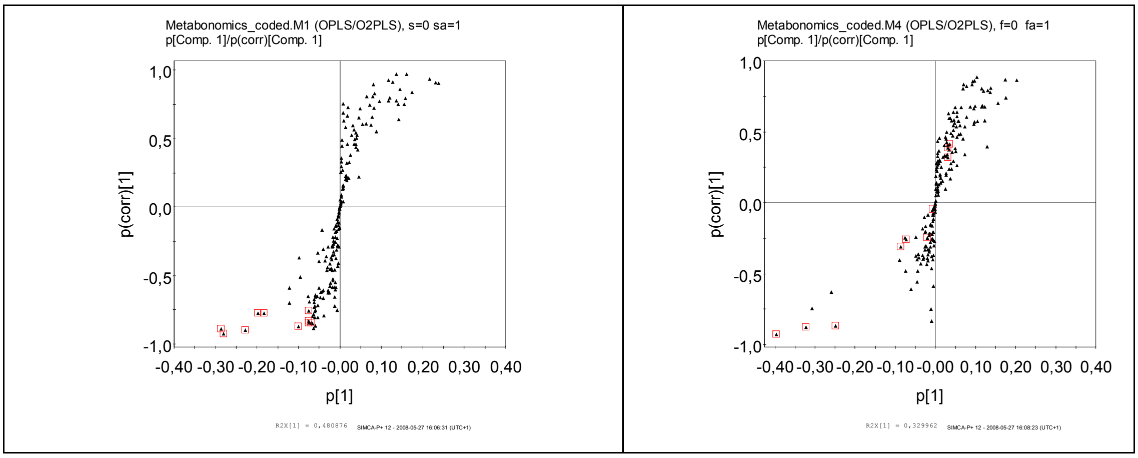
M1 Control Sprague-Dawley (s), 10 rats, "s" vs treated with **amiodarone** (sa), 8 rats "sa"

M2 Control Sprague-Dawley (s), 10 rats, vs treated with **chloroquine** (sc), 10 rats "sc"

M3 Control Fisher (f), 10 rats "f" vs treated with **amiodarone** (fa), 10 rats "fa"

M4 Control Fisher (f), 10 rats "f" vs treated with **chloroquine** (fc), 9 rats "fc"

Use the plot facility in SIMCA and mark the high magnitude/high reliability biomarker in one plot and see where they appear in the other. This is one way of illustrating which biomarkers that are the same in both rat strains. Please realise that SIMCA will scale the plots differently so to get the best plot you need to rescale the axis in both plots so that they are identical and symmetrical.



Solution B

M1 Control Sprague-Dawley (s), 10 rats, “s” vs treated with **amiodarone** (sa), 8 rats “sa”

M2 Control Sprague-Dawley (s), 10 rats, vs treated with **chloroquine** (sc), 10 rats “sc”

M3 Control Fisher (f), 10 rats “f” vs treated with **amiodarone** (fa), 10 rats “fa”

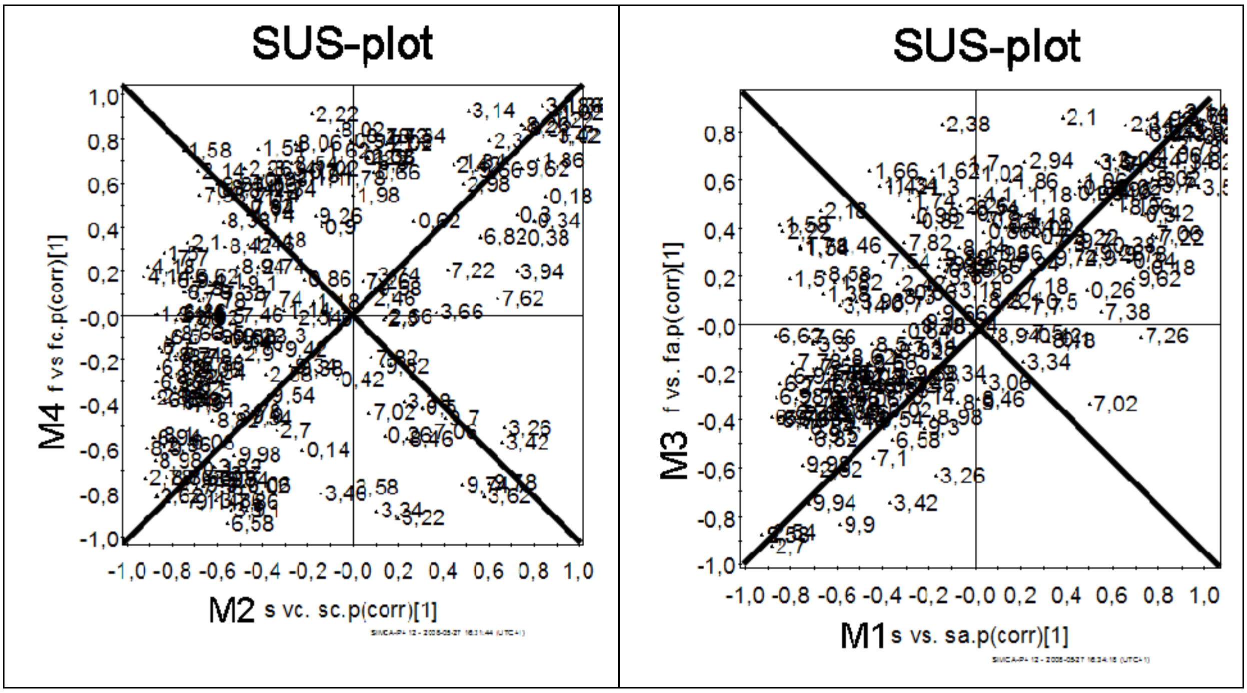
M4 Control Fisher (f), 10 rats “f” vs treated with **chloroquine** (fc), 9 rats “fc”

Plot p(corr)1 for M1 vs. p(corr)1 for M3 and p(corr)1 for M2 vs p(corr)1 for M4.

1.98 in the right plot, which is in the upper right corner increases after treatment with amiodarone in both rat strains.

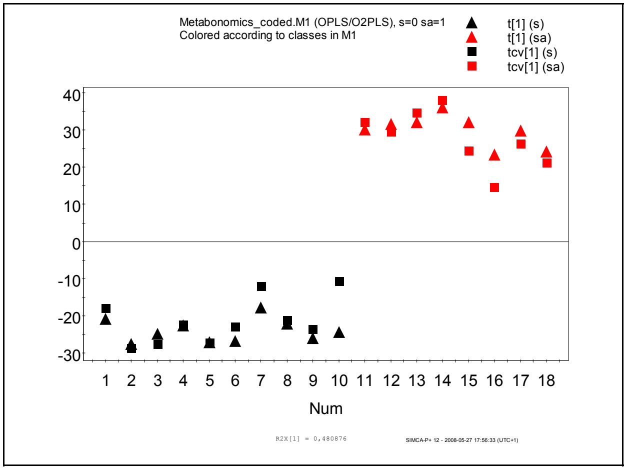
3.26 in the left plot increases in Sprague-Dawley but decreases in Fisher rats after treatment with chloroquine.

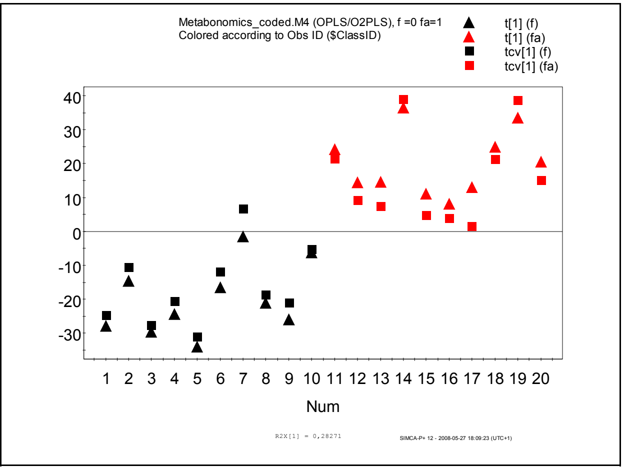
3.26 in the right plot decreases in Fisher rats and is constant in Sprague-Dawley after treatment with amiodarone.



Task 5

Model validation should ideally be performed using an external test set. However, as OPLS uses full cross validation one alternative approach is to look at the cross validated score plot. This plot visualizes the stability for each observation in the model. In the two plots below it is seen that the model with s vs. sa rats are much more stabile than the model with f vs. fa rats.





Statistical significance vs. biochemical significance

- We obtain a list of potential biomarkers which are statistically significant and which separate one class from another.
- These biomarkers are statistically significant, but not necessarily biochemically significant.
- They may have biochemical significance and they may be the biomarkers we are interested in, however, this must be established through extensive testing.
- Metabonomics/Metabolomics allows the user to mine complex data and provides information which allows us to propose intelligent hypotheses.
- OPLS-DA is an excellent tool for this purpose.

Identification of bias effects in transcriptomics data

OPLS-DA to find information about the uncorrelated variation

Background

The study of gene functions and behaviours are routinely performed by using dual-channel cDNA microarrays. This technique will simultaneously quantify the expression levels of tens of thousands of mRNA species most commonly as cDNA after reverse transcription. This technique has proven to be highly useful in functional genomics studies.

The experimental procedure contains the following steps:

1. cDNA probes from a library are attached on a solid surface at pre-defined positions.
2. RNA samples are reversed-transcribed to cDNA. These are labelled with fluorescent dyes and allowed to hybridize to the probes. In two channel microarray, two RNA samples (often reference and treated) are labelled with different florophores e.g. Cy5 and Cy3 and measured together on the same surface.
3. Superfluous material is washed away
4. Fluorescence signals are generated by laser-induced exitations of the residual probes. These signals are assumed to be proportional to the expression levels of the RNA species in the sample.

During the experimental data generation there are several steps where unwanted sources of systematic variation may be introduced. Some of the most common sources of systematic variation are:

- *Array bias*-caused by offset between two analytical replicate using different arrays.
- *Dye bias*-caused by slightly different physical properties between different dyes.
- *Spatial bias*-reflecting regions on the microarray surface with stronger or weaker signals than others.

The data set used in this exercise is called *H8k* and is comprised of 26 two channel cDNA microarrays. The experimental design is a traditional dyeswap design containing a treated sample and a reference sample measured using technical replication

Objective

The objective of this exercise is to shed some light on what type of information that can be extracted from the orthogonal component.

Data

In total, the data set contains N = 52 (2*26) observations and K = 19199 variables elements on the array. The observations are divided in two groups ("classes") designated 1 for treated and 0 for non-treated rats [1]. The data also contains observation information about array and dye and variable information about different blocks on the array. This information will be useful when analysing the data.

1) Bylesjö et. al BMC Bioinformatics ***2007****,85:207 doi:10.11/1471-2105-8-207,Orthogonal projections to latent structures as a strategy for microarray data normalization*

Tasks

Task 1 (PCA)

Select *File New* and choose Transcriptomics.xls. On import set the first row to *Primary Variable ID,* second, third and fourth row to *Secondary Variable ID*, first column to *Primary Observation ID* and second, third and fourth column . One column have been appended to the data to designate the different classes. The references samples are 0 and the treated rats are 1. Set this vector to Y. The last column in the data called *CV variable* is neither a X or Y variable. This column will be used in cross-validation in order to balance the exclusion of observations during cross validation. REMEMBER to exclude this column in all modeling.

Start the analysis by an overview PCA for each class. Make a PCA of the entire X block using both classes. Can you detect any outliers, patterns or trends? If outliers exists in the data these should be checked historically and removed if there is a good reason.

Task 2 (Contrasting two classes)

We will now make a class model using OPLS. Define the class variable as the single Y-variable under *workset/variables* and exclude *CV variable*. Pareto scale all variables. Scaling is performed under *workset /scale*. In SIMCA the default scaling is unit variance (UV). Change this to pareto (par), *select all/par/set*. Under *workset/model type* select *OPLS/O2PLS*. Before auto fitting the model the selection of samples of cross-validation must be changed. This is done in order to balance the classes during CV. Go to *workset/model options/CV-groups.* Select *assign observation based on variable* and choose *CV variable,* finally *group observations with the same value in the same group* and press *apply.* Auto fit the model.

Specific questions:

Can you separate the control from the treated by OPLS-DA?

How much of the variation in **X** is related to the separation between controls and treated?

How much of the variation in **X** is systematic but uncorrelated to the classes?

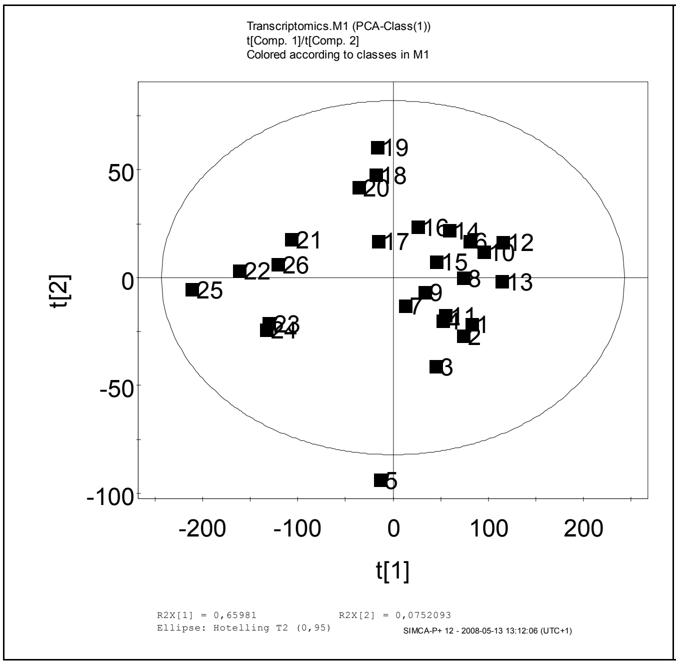
Task 3 (Identifying the variation seen in the orthogonal components)

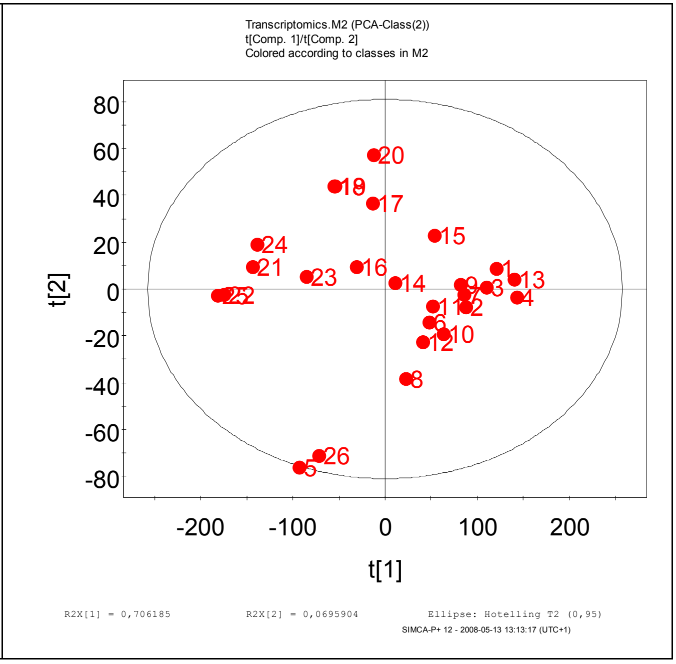
Interpret the orthogonal components. Make plots from the orthogonal components to, po and colour the plots by the extra information given in the secondary observation name and secondary variable name. Can you understand what is seen in the orthogonal components?

SOLUTIONS to OPLS with Transcriptomics data

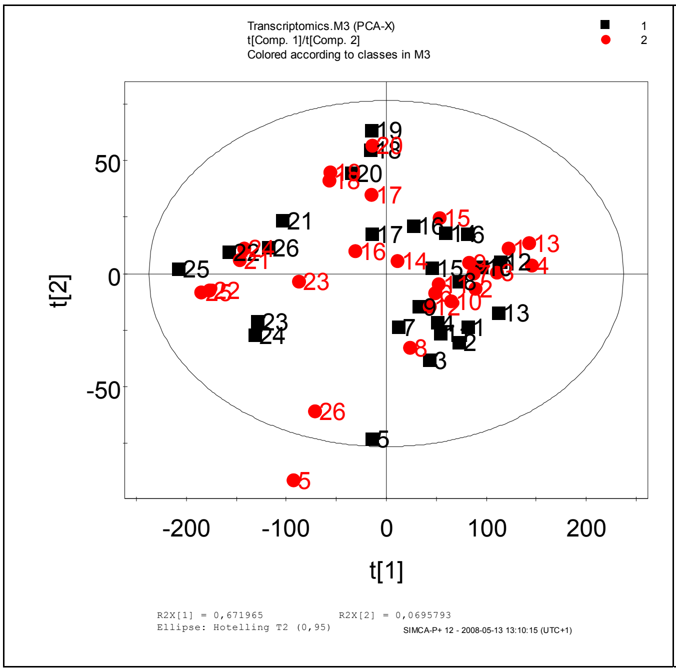
Task 1 (PCA)

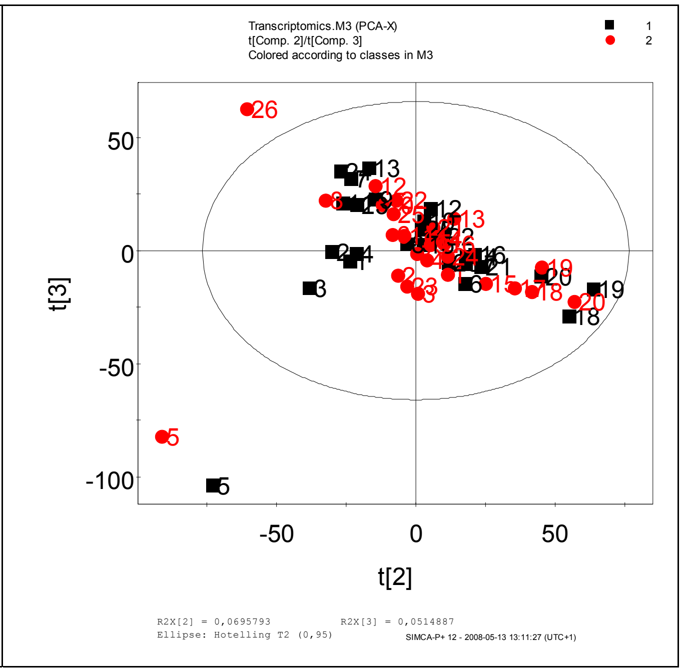
Tree outlier can be seen in the two groups i.e. nr 5 in both classes and nr 26 in class 2.





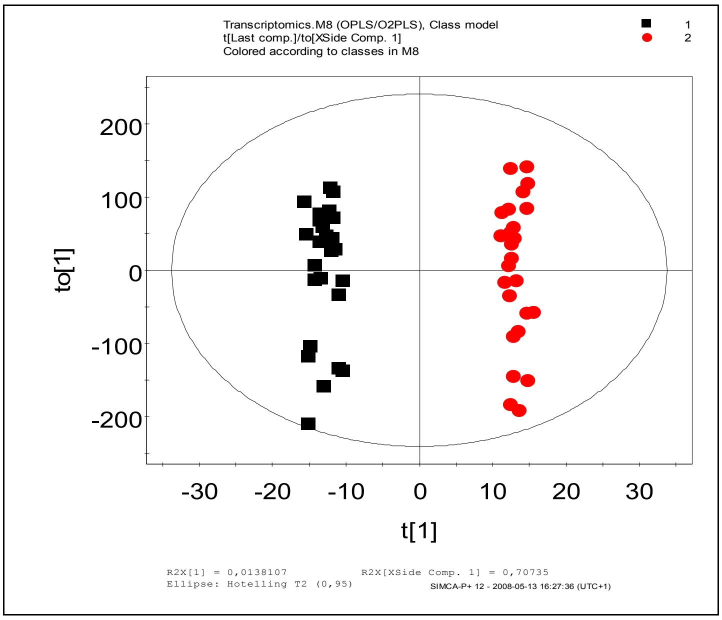
Same outliers were also clearly seen in the PCA from all classes. Nr 5 is an outlier in both classes i.e. originates from the same array. This array was traced back to the experimental work and it was verified that it was caused by an error during the experimental work. For this reason array 5 was excluded prior to the OPLS-DA analysis.





Task 2 (Contrasting two classes)

In the score plot the first component, t1, represents the variation caused by class separation. Class one represents the reference samples and class 2 represents the treated samples.



Answers to questions:

Can you separate the control from the treated by OPLS-DA?

Reference vs. treated: R2Y=0,989, Q2Y=0,96, good class separation and high predictive ability.

How much of the variation in **X** is related to the separation between controls and treated?

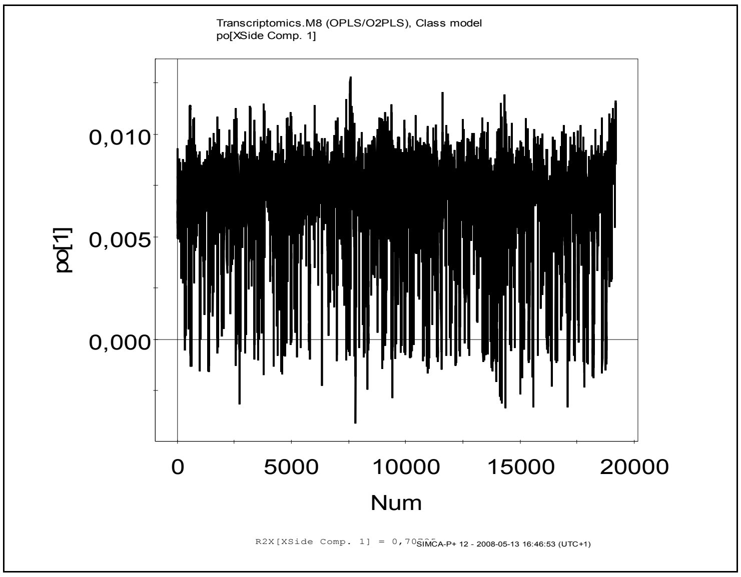
predictive component R2X=0,0138→1,38%

How much of the variation in **X** is systematic but uncorrelated to the classes?

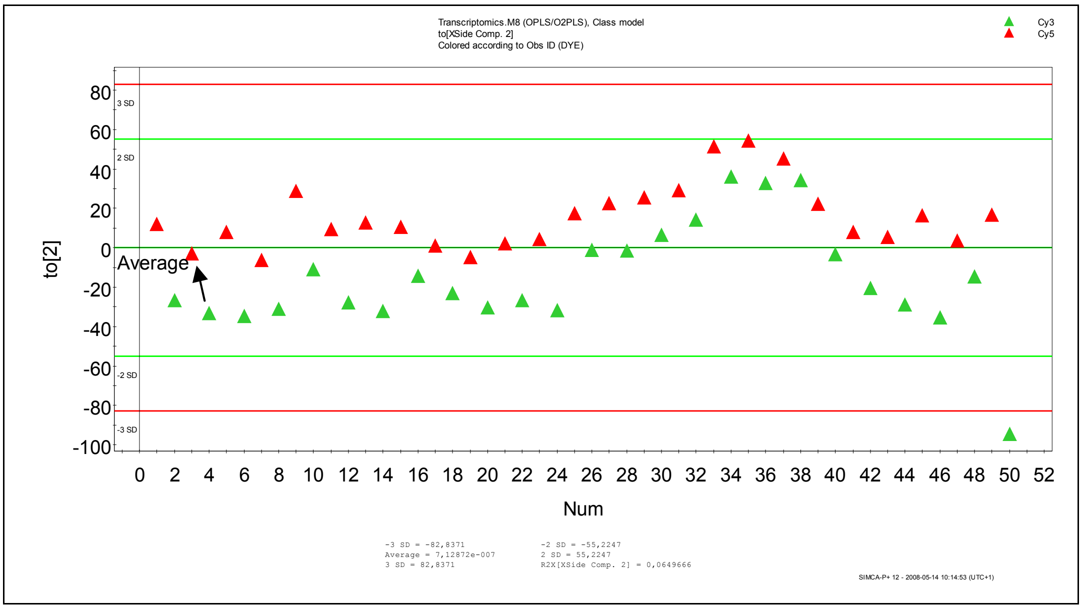
Orthogonal component R2X=0,799→79,9%

Task 3 (Identifying the variation seen in the orthogonal components)

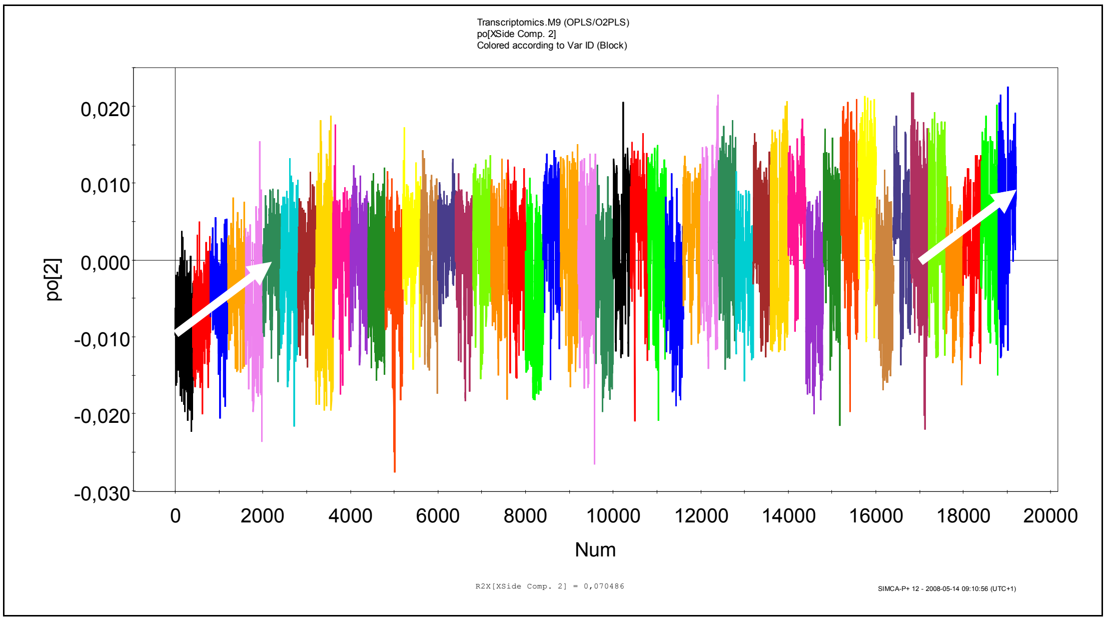
From the scores plot from to1 it could not explained what caused this variation. However the corresponding loading plot, po1, are not centred around 0 which clearly explain that this variation is caused by a baseline shift. This is called *array bias*.



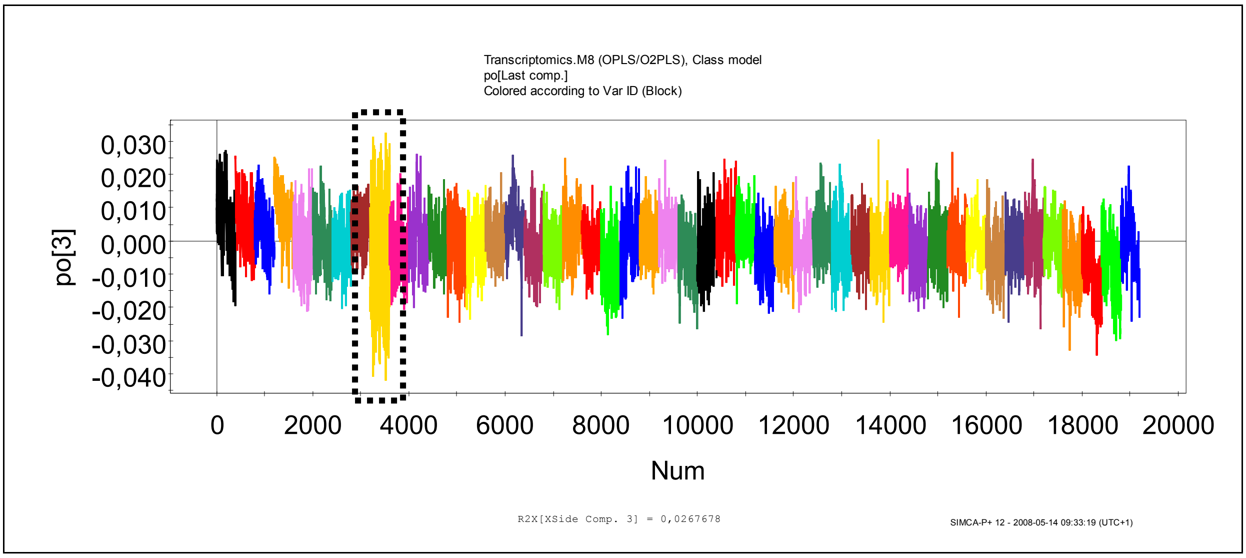
In the score plot from the second orthogonal component, to2, it is clearly seen that there is a systematic effect of the two different dyes (Cy3 and Cy5) on the same array. Each array is represented by one Cy3 and one Cy5 observation. In the score plot these two observations are slightly tilted, indicated by the arrow in the plot below. The effect of dyes seen in to2 is also confounded by print tip groups which can be visualized in the corresponding loading vector, po2.



The corresponding loadingplot, po2, can be coloured by the different blocks to highlight the dye effect which mainly can be visualised in the print tip groups.



The loadings from the third orthogonal component indicates *spatial bias* i.e. regions on the surface with stronger or weaker signals than others. Block 9 is one region with higher variability i.e. stronger signals.



Conclusions

Normally in transcriptomics studies the bias effects i.e. *array bias, dye bias and spatial bias* are removed prior data analysis. The focus on this exercise is to highlight the useful information that can be found in the OPLS orthogonal components. The seen information should be used to learn about the data and to improve future studies and if possible make better pre-processing to remove the identified information.